TRANSFER LEARNING BASED WILDLIFE RECOGNITION FOR TELE-OBSERVATION IN FIELD OCCLUSION ENVIRONMENT

[1,2]*Hongpeng Wang,* [1,2]*Yulin Song,* [3]*Sheng Li,* [1,2]*Wan Dai,* [1,2]*Jingtai Liu*

[1]College of Artificial Intelligence, Nankai University, Tianjin, 300353, China
[2]Tianjin Key Laboratory of Intelligent Robotics, Tianjin, 300353, China
[3] School of Life Sciences, Peking University, Beijing, 100871, China

ABSTRACT

Timely and credible species recognition of wildlife at their habitats is helpful for ecological monitoring. Thanks to the powerful feature extraction ability of convolutional neural networks(CNN), the performance of species identification has been significantly improved. However the CNN based approach still does not achieve their full potential. The cluttered backgrounds and rich feature changes of wild environment bring great challenges to wildlife recognition. In this paper, we propose a novel approach to improve the anti-occlusion ability of CNN model, which is achieved by training a improved anti-occlusion loss function. The anti-occlusion constraint in our proposed loss function works to reduce the distance of feature expression before and after a sample has been occluded. We validated the effect of occlusion based on public dataset cifar-10, which confirms that its necessary to improve the anti-occlusion ability of CNN model. Based on single-labeled training dataset and generalization dataset, comprehensive comparative evaluation proves that our proposed loss can effectively improve the generalization ability of CNN model during the identification of wildlife.

***Index Terms*—** customized loss function, anti-occlusion, wildlife monitoring, transfer learning, CNN

1. INTRODUCTION

The growing population and urbanization activities have continually threatened the survival of wildlife. Timely and credible species recognition of wildlife at their habitats becomes more and more necessary for ecological monitoring. There are some citizen scientists organizations dedicated to collect images of wildlife. While at the habitats of wildlife(such as nature reserves), ”Camera-Trap” is becoming an increasingly popular monitoring tool due to its effectiveness and reliability. Camera-Trap is a kind of motion-triggered device that can capture images and videos of wildlife, which is placed at a fixed location by reserchers and replacement regularly. However, manually handling such continuous and numerous imagery data is very cumbersome for scientists, both time consuming and monotonous. Therefore, modern technologies for wildlife identification can not only improve the efficiency of ecological monitoring, but also reduce the artificial burden.

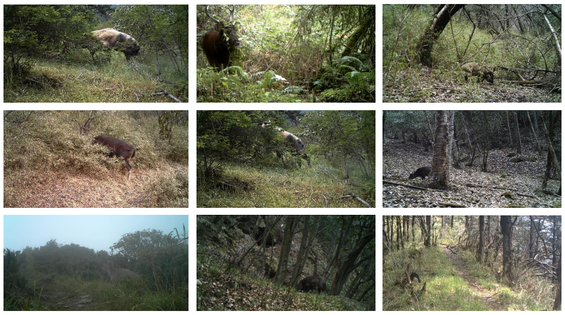

Fig. 1. Display of imagery data with object occlusion problem. It shows that it’s necessary to improve the anti-occlusion ability of CNN model.

Unlike the classification task based on public datasets, wildlife identification is facing with many challenges due to various conditions: (1) Object occlusion caused by lush plants, shown as Fig. 1. (2) Wild animals have similar body shape and body color which cause high between-class similarity. (3) It’s hard to obtain large number of labeled samples.

Wildlife identification attracts more and more researchers due to the needs of wildlife protection and ecological monitoring. Guobin Chen et al.[1] first apply CNN in recognition of camera-trap imagery at 2014, they build a five layers CNN model and got better performance than traditional bag of visual words method. In addition, Nguyen et al. demonstrated the feasibility of a deep learning approach towards constructing scalable automated wildlife monitoring system[2] by identifying three most common animals (bird, rat and bandicoot) in Australia with a 5 layers CNN. Jason et al.[3] evaluated five detection components against their own wildlife

Thanks to National Natural Science Foundation of China(Grant No. 61375087) and Key Program of Natural Science Foundation of Tianjin (Grant No. 15JCZDJC31200) for funding.

dataset, and aiming to increase the reliability and automation of animal censusing studies.

2. RELATED WORK

Object recognition has been actively studied for the last few decades[4, 6, 5, 7]. Although the recognition based on the public dataset reaches the saturation level, it has to be acknowledged that the public dataset has the advantages of rich sample and high resolution. While the two conditions are difficult to satisfy for small sample datasets.

Transfer learning both can preserve the model's recognize ability and adapt to the characteristics of the new dataset. On the basis of the high-capacity CNN architectures, we can obtain representations of own datasets through transfer learning. Junwei Han et al.[8] finetune pre-trained architecture to introduce rotation-invariant and fisher discriminative into CNN model, which improves the efficiency of objection detection. Lee et al.[9] predict the quality of distorted videos with temporal features which composed of spatial features and handcrafted features. Object Recognition with CNN also widely used in medical field. Riel et al.[10] transfer features pretrained by CNN model for cancer detection.

On the basis of transfer learning, in order to make the model more suitable for specific task, designing custom loss function becomes the popular research direction. De Cheng et al.[11] train a CNN model with an improved triplet loss function that serves to pull the instances of the same person closer. Bailer et al.[12] introduce a thresholded loss into siamese networks for optical flow estimation, which are demonstrated performs better than existing losses.

3. CNN ANTI-OCCLUSION ABILITY TRAINING METHOD

The goal of our method is to train a CNN model with anti-occlusion ability based on customized anti-occlusion loss function. This is achieved by finetune our loss with artificial occlusion dataset on the basis of those exising high-performance CNN architectures.

In this section, we present the proposed anti-occlusion method in details. We first present the improbed anti-occlusion loss function in details, then we describe the overall framework of our anti-occlusion method.

3.1. Improved Anti-occlusion Loss Function

In the multi-classification task based on neural network, the commonly used loss function is the cross-entropy loss function. In order to adapt the network to our artificial occlusion dataset, we add the anti-occlusion constraint based on the cross-entropy loss function. At the same time, in order to reduce the over-fitting, we add the L2 regularization constraint.

The final loss function J_{AB} (the letter AB means anti-blocking) is consists of three parts as shown in equation(1):

$$J_{AB} = L(X_{AB}, Y) + \underbrace{\lambda_1 B(X, T_{AB}X)}_{anti-blocking-constraint} + \underbrace{\lambda_2 R(W_{AB})}_{L2-regularization} \tag{1}$$

where $T_{AB}=\{T_{b1}, T_{b2}, \cdots, T_{bn}\}$ indicates n occlusion degrees of the original picture correspond to n values of BF(as discribed in section 3.2 , $X_{AB} = \{X, T_{AB}X\}$ means the collection of entire artificial occlusion dataset, Y is the label of corresponding X and W_{AB} means the parameters of CNN model. λ_1 and λ_2 are two weight coefficients.

The first part is shown in equation(2):

$$L(X_{AB}, Y) = -\frac{1}{N} \sum_{x_i \in X_{AB}} \sum_{c=1}^{C} y_c^{(x_i)} \log y_{c\_}^{(x_i)} \tag{2}$$

where N means the total number of samples, C is the number of classes, $y^{(x_i)}$ is the number c element of the label vector y corresponding to x_i and $y\_^{(x_i)}$ is the number c element of inference vector$y_$ corresponding to x_i.

The second part is shown in equation(3):

$$B(X, T_{AB}X) = \frac{1}{2N} \sum_{x_i \in X_{AB}} \left\| O_k^{(x_i)} - \frac{1}{n_B} \sum_{T_b \in T_{AB}} O_k^{(T_b x_i)} \right\|^2 \tag{3}$$

where O_k is the output of layer k and $k \in \{fc6, fc7\}$ means our anti-blocking constraint works on two layers: fc6 and fc7, and look at the details, the anti-blocking constraint is used to calculate the difference between the feature of the original image x_i and the average feature of the images after occlusion.

The third part is shown in equation(4):

$$R(W_{AB}) = \frac{1}{2} \sum_{x_i \in X_{AB}} \|W_{AB}\|^2 \tag{4}$$

where W_{AB} means the parameters of CNN model.

3.2. Transfer Learning Based Traning Framework

The framework of the proposed CNN anti-occlusion training method is illustrated in Fig. 2.

Firstly, in order to make full use of existing high-capacity CNN architectures, we obtained the parameters of several layers by pre-trained with ImageNet[14] dataset. In our experiment, the architecture we pretrained is VGG16[5], we transfer the parameters of it's first 13 layers; Secondly, our anti-occlusion dataset is generated on the bias of wildlife dataset

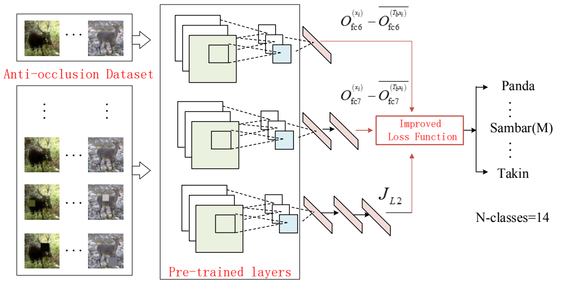
Anti-occlusion Dataset
Pre-trained layers
$O_{\mathrm{fc6}}^{(x_i)} - \overline{O_{\mathrm{fc6}}^{(T_b x_i)}}$
$O_{\mathrm{fc7}}^{(x_i)} - \overline{O_{\mathrm{fc7}}^{(T_b x_i)}}$
J_{L2}
Improved Loss Function
Panda
Sambar(M)
Takin
N-classes=14

Fig. 2. Illustration of our proposed training method. There are three main contributions: (1) Artificial occlusion treatment. (2) Transfer learning. Finetune a pre-trained convnet to reduce overfitting. (3) Customize loss function. Introducing anti-occlusion constraint to reduce the distance of feature expression before and after a sample has been occluded.

Far-eco(will illustrated in section 4.1.2 with different degrees of occlusion factor(as discribed in section 4.1.1; Thirdly, the improved anti-occlusion loss function we designed consists of three parts (will described detailly in Section 3.1, which are derived from the three fully connected layers.

The steps of the training process are detailed in algorithm1.

Algorithm 1 CNN anti-occlusion training method

```
Require: anti-occlusion dataset $(X_{AB}, Y)$ and pre-trained parameters;
Ensure: $W_{AB}$: parameters of trained CNN;
Initialize the parameters of full connected layers;
repeat
    compute standard loss with formula (2);
    compute anti-blocking constraint with formula (3);
    compute L2 regularization with formula (4);
    compute the finally loss with formula (1) ;
    update parameters $W_{AB}$;
until (epoches > threshold)
```

4. EXPERIMENTS

4.1. Verification of the Impact of Occlusion on Recognition

Before comparing the effects of different loss functions, we explore the impact of occlusion based on public dataset. Firstly, we define a occlusion factor ”BF” which expressing the degree of occlusion. Then, we process the train set of cifar-10[13] with various degrees of artificial occlusion. By comparing the accuracy of CNNs trained with different occlusion training sets, we can observe the effect of occlusion on recognition.

4.1.1. Definition of Occlusion Factor

We define the occlusion degree BF as the percentage of pixels covered by artificial occlusion in the total number of pixels.

When the image size is N*N and the artificial occlusion size is n*n, the occlusion degree BF is expressed as equation(5):

$$BF = \frac{n^2}{N^2} \times 100\% \tag{5}$$

4.1.2. Artificial Occlusion Dataset

Based on the fact that the occlusion comes from backgrounds, we select a random pixel as background in the image, then fill in the selected range with the background pixel. The pixel size N of Cifar-10[13] dataset is 32, value BF of our artificial occlusion dataset is 0-100%. The 33 groups of training sets train a classical 5 layers CNN model separately. The architecture of the CNN network including 2 convolutional layers and 2 max-pooling layers and finally a 14-way fully connected layer. The hyper-parameter of filters in convolutional layers is 5x5x6 and 5x5x12. The two max-pooling are both with the size of 2x2.

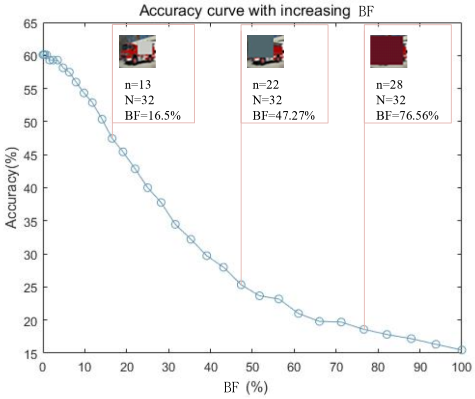

Fig. 3. Verification of the effect of occlusion based on public dataset. Taking a truck image as example, red box shows 3 artificial occlusion result with different BF. The curve shows the variety of recognition accuracy with the increase of occlusion degree.

4.1.3. Verification Results

Shown as Fig. 3, we can draw the following conclusions: 1)As the occlusion area increases, the recognition rate gradually decreases. 2)Before the occlusion area accounts for 50%, the recognition accuracy decreases linearly until about 1/3 of zero occlusion. 3)When the occlusion area is 100%, the recognition accuracy remain 15% rather than 0%.

4.2. Wildlife Dataset

In order to measure the generalization ability of models, we train on Far-eco dataset and test on generalization dataset to obtain generalization accuracy.

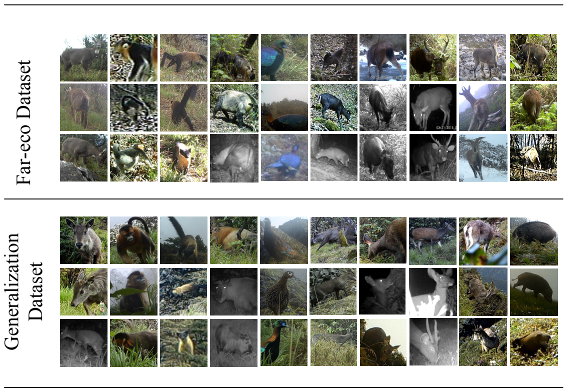

Fig. 4. A miniature of 10 species of the Far-eco dataset and generalization dataset. Each column corresponds to the same species.

Far-eco dataset We collected and annotated a standard camera-trap dataset of 14 species in the southwest of China[15], which contains 20600 images. These image data were collected from 24 camera sites in four nature reserves in China. Each image includes only one animal.

Generalization dataset The generalization dataset is very diffferent with Far-eco, which includs 4874 images of 11 species. There are 3 species that exist in Far-eco but not in the generalization dataset because they did not appear at the second location.

Artificial occlusion Far-eco Preprocess with occlusion as described in Section 4.1.2, making Far-eco an artificial occlusion dataset$X_{AB} = \{X, T_{AB}X\}$. During the training process, the occlusion level BF will affect the characteristic expression of x_i. We set the occlusion processing BF as 1%-20%.

4.3. Comparision of CNN based Approaches

We designed two sets of experiments for comparison, in which there are two differents: 1) How large is the input image size; 2) Which loss function we optimize.

As shown in Table 1, we have prepared datasets with 3 kinds of image size, some have been occluded and others are not. There are four different datasets in all, which are datasetA-D. As shown in Table 2, there are 3 kinds of loss functions where J_L is the classical cross-entropy loss function, J_{L2} is cross-entropy loss function with $L2$ regularization, and $J_{AB}$is our proposal improved anti-occlusion loss.

Table 1. Datasets.

Occlusion Image-size	Yes	No
32*32	–	Dataset-B
64*64	Dataset-A	Dataset-C
128*128	–	Dataset-D

Table 2. Loss functions.

Symbol	Expression
J_L	$J_L = L(X, Y)$
J_{L2}	$J_{L2} = L(X, Y) + \lambda R(W)$
J_{AB}	$J_{AB} = L(X_{AB}, Y) + \lambda_1 B(X, T_{AB}X) + \lambda_2 R(W_{AB})$

Therefore, combining the variables in the two tables, we designed 5 experiments. In experiment4, λ=0.001. In experiment5, $\lambda 1$=0.001 and $\lambda 2$=0.005. Experiment1-3 optimize the J_L loss function, and the input image sizes of them are respectively 32*32, 64*64, and 128*128. As shown in table3, the generalization accuracy of experiment 2 was the highest in the three sets of experiments, which explain that the input size of 64*64 is more appropriate for this architecture. Therefore, in experiments 4 and 5 we uniformly used a dataset with the input size of 64*64. As shown in table3, experment5 obtain the highest generalization accuarcy, proving our proposed anti-occlusion constraint can effectively improve the generalization ability of CNN model for wildlife recognition.

Table 3. Experiments.

Model	test-accuracy(%) on Far-eco	generalization -accuracy(%)
Dataset-B with J_L	75	26.77
Dataset-C with J_L	92	37.81
Dataset-D with J_L	90	35.85
Dataset-C with J_{L2}	91	39.12
Dataset-A with J_{AB} (ours)	95	**45.62**

5. CONCLUSION AND FEATURE WORK

In this paper, we present a novel loss function with anti-occlusion constraint which outperforms compared with other existing losses on the challenging camera-trap dataset. On our generalization dataset which is collected from different regions and times with Far-eco, the CNN model trained by our proposed method can achieve the generalization accuracy of 45.62%, which will be helpful for ecological monitorning. In the future work, we will strive to extract more universal feature expressions of wild animals, which will pave the way for further species analysis and ecological monitoring.

6. REFERENCES

[1] Guobin Chen, Tony X. Han, Zhihai H, Roland Kays, and Tavis Forrester. ”Deep convolutional neural network based species recognition for wild animal monitoring.” Image Processing (ICIP), 2014 IEEE International Conference on. IEEE, 2014. pp. p. 858-862.

[2] Hung Nguyen, Sarah J. Maclagan, Tu Dinh Nguyen, Thin Nguyen, Paul Flemons, Kylie Andrews, Euan G. Ritchie and Dinh Phung. ”Animal recognition and identification with deep convolutional neural networks for automated wildlife monitoring.” In: Data Science and Advanced Analytics (DSAA), 2017 IEEE International Conference on. IEEE, 2017. pp. 40-49.

[3] Jason Parham, Charles Stewart, Jonathan Crall and Daniel Rubenstein. ”An Animal Detection Pipeline for Identification.” 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018. pp. 1075-1083.

[4] Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton, ”Imagenet classification with deep convolutional neural networks.” Advances in neural information processing systems. 2012. pp. 1097-1105.

[5] Simonyan, Karen, and Andrew Zisserman. ”Very deep convolutional networks for large-scale image recognition.” arXiv preprint arXiv:1409.1556 (2014).

[6] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke and Andrew Rabinovich. ”Going deeper with convolutions.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. pp. 1-9.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. ”Deep residual learning for image recognition.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. pp. 770-778.

[8] Cheng, Gong, Peicheng Zhou, and Junwei Han. ”RIFD-cnn: Rotation-invariant and fisher discriminative convolutional neural networks for object detection.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. pp. 2884-2893.

[9] Ahn, Sewoong, and Sanghoon Lee. ”Deep Blind Video Quality Assessment Based on Temporal Human Perception.” 2018 25th IEEE International Conference on Image Processing (ICIP). IEEE, 2018. pp. 619-623.

[10] Sjors van Riel, Fons van der Sommen, Sveta Zinger, Erik J. Schoony Peter and H.N. de With. ” Automatic Detection of Early Esophageal Cancer with CNNS Using Transfer Learning.” 2018 25th IEEE International Conference on Image Processing (ICIP). IEEE, 2018. pp. 1383-1387.

[11] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang and Nanning Zheng. ”Person re-identification by multichannel parts-based cnn with improved triplet loss function.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. pp. 1335-1344.

[12] Bailer, Christian, Kiran Varanasi, and Didier Stricker. ”CNN-based patch matching for optical flow with thresholded hinge embedding loss.” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017. pp. 2710-2719.

[13] Alex Krizhevsky and Geoffrey E. Hinton. ”Learning multiple layers of features from tiny images.” (Vol. 1, No. 4, p. 7). Technical report, University of Toronto.

[14] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K. and Fei-Fei, L. ”Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition.” 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2009. pp. 248-255.

[15] Yulinsong, Hongpeng Wang, Sheng Li, Fulai Xu and Jingtai Liu. ”CNN based Wildlife Recognition with Super-pixel Segmentation for Ecological Surveillance.” 2018 IEEE Conference on on CYBER Technology in Automation, Control, and Intelligent Systems(CYBER). IEEE, 2018.