Generalize Animal Recognition CNN model with Background Removal

Zhipeng, Chen
zhc023@ucsd.edu

Guanlin, Qian
gqian@ucsd.edu

Xuzhe, Zhi
xzhi@ucsd.edu

March 15, 2023

1 Abstract

Camera traps are commonly used by ecologists to monitor wildlife biodiversity loss. However, the variation in illumination, camera angle, background, vegetation, color, and relative animal frequencies across different camera traps poses a challenge for multi-class species classification using machine learning models. In this paper, we propose a domain adaptation method using background removal to address this challenge. Our approach involves subtracting the background from the original image to highlight the animal features. We trained a convolutional neural network on images with background removed, and our model achieves a classification accuracy of 0.37 on images taken by existing camera traps and 0.35 on images taken by new camera traps. These results represent a significant improvement over the baseline model. Moreover, we preserved the original color of the animal pixels by masking the background-removed image. The average time taken for classifying one image using background removal is 0.0108 seconds, slightly longer than the baseline model. Our findings demonstrate the potential of using background removal as a domain adaptation method to improve the accuracy of wildlife biodiversity monitoring using camera traps.

2 Introduction

Camera Traps are placed in the wild to enable zoologists and environmental scientists to gather valuable information about wild animals. Once triggered by any moving object, these traps can automatically capture large quantities of images. These images are crucial for monitoring animal species count, density, and behavior in the area. Analyzing the massive image data retrieved by camera traps can be challenging. Therefore, it is beneficial to train a model to recognize and classify animals in the image.

However, a widely recognized issue in deploying such models to images captured by camera traps in different countries, locations, angles and environments

is domain adaptation. Changes in light conditions, seasons, and camera types can significantly impact image quality and result in reduced model accuracy. Moreover, training a new model can be time-consuming and costly, making it difficult to keep up with the increases of camera traps and their ever-changing surrounding environments.

To help address this challenge, we employ a preprocessing step to remove background noise from the images. The background removal ensures that the model is only trained on more relevant information, thereby improving its accuracy and adaptivity to new images captured in diverse environments. We then train the model on the preprocessed images and experiment with its ability to classify and recognize animals in images from different sources and environments. Comparing the model performance over different data subsets, we can assess the effectiveness of our approach and determine whether it can be applied to other camera trap settings.

3 Dataset

The dataset used in this study is the iWildCam dataset from WILDs collected by Stanford, which includes over 200,000 images of 182 animal species captured by camera traps deployed in more than 200 locations. The images have various widths and heights but share the same color channels, and all images were cropped to a size of 448 x 448 pixels for model input.

The images are labeled with one of 182 animal species and the domain, which specifies the identity of the camera trap that captured the image. However, the distribution of images across different domains is not even, with some domains having as few as three images and others having thousands of images. Additionally, some images are labeled as having an animal present, but it may be difficult to identify any animals in the image with the naked eye due to abnormal illumination and exposure.

To ensure the models are able to generalize to new camera deployments, the training and test sets comprise photos from disjoint sets of camera traps. All images in the testing set are from domains (locations) that are unseen in the training set. By training the model on images from a subset of camera traps and evaluating the model on images from a different subset of camera traps, we can test the model's ability to generalize across domains with different environmental conditions and animal populations.

4 Methods

For the method part, we preprocess the image data with two methods: background removal and normalization. Then we train a customized Convolutional Neural Network model to classified the processed image.

4.1 Background Removal:

Background removal is a process that involves subtracting the background image of a location from the original image, resulting in an image that highlights the animal features in the photo. In the context of a domain adaptation problem, where models are trained on one set of camera trap photos and tested on another set, background removal can be used as a method to reduce variation between camera trap deployments. Since different camera traps can have vastly different backgrounds, removing the background from images ensures that the resulting images contain only animal features that are consistent across different camera traps.

4.1.1 Background Extraction

To accomplish this, we used the mean and median pixel values of all images in a location to obtain daytime and nighttime backgrounds. The median background of each location is used as it was found to be visually sounder (less blurry as shown in Figure.1) than the mean background.

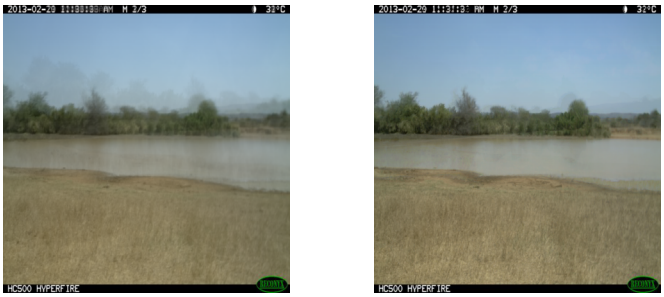


Figure 1: Mean Background vs. Median Background (location 288)

4.1.2 Background Subtraction and Masking

To preserve the original color of animal pixels after background removal, we used a masking method to filter out important pixels from the original image. For each image, a threshold is determined based on the mean and standard deviation of the norms of all the pixels in the background-removed image. Then, pixels in

the original image whose RGB norm is above this threshold are identified and masked, resulting in a filtered image that retains only the important pixel values. This method ensures that the original color of the animal pixels is retained even after the background is removed, improving the accuracy of classification models trained on these filtered images. However, it is worth noting that this method may not be effective if there is significant variation in animal coloration across different camera trap deployments.

(a) Original image

(b) Background-subtracted image

(c) Filtering mask

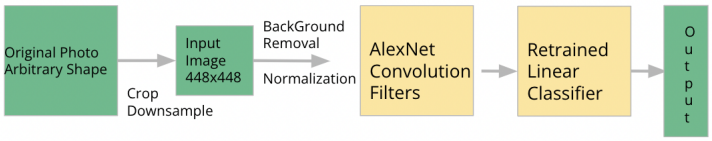
(d) Masked image

4.2 Normalization:

Normalization is a process in which the intensities of the pixels in an image are adjusted to a common scale. This is achieved by transforming the pixel intensities so that they have a mean of zero and a standard deviation of one. In our situation, we decide to use Z-score normalization to the images that have been apple with the removing background method. This process is useful in image processing because it helps to reduce the variance between images and make them more comparable to each other. We use Z-score normalization because it is a useful process in image processing that can helps to improve the performance of CNN model since it can improves image segmentation, facilitates visualization of images, and reduces the size of images for storage and transmission.

4.3 Custom Convolutional Neural Network

We will use the CNN model for the wild animal classification task. We will use the architecture similar to AlexNet with 5 learnable convolutional layers and 3 fully connected layers (Krizhevsky et al. 2012), along with max pooling and batch normalization layers between convolution layers. The nonlinear activation function used is ReLU. The image input for alexnet is 227x227 and the output layer is modified to fit the 100 animal classes of the dataset. The model is trained with gradient descent and the loss function is cross entropy loss. In the training process, we will tune the learning rate, batch size and data preprocessing methods to achieve lower loss and better accuracy.



5 Results

Overall, the model's performance on test data from unseen locations exhibits improvement with the application of background removal. The accuracy in predicting both the original test set and the balanced test increases with the use of this technique. However, it is worth noting that the "Background Mask" method necessitates a considerably longer preprocessing time and relatively higher prediction time compared to other approaches.

The iWildcam dataset possess two test datasets: one is called "id_test" and comprises images captured from camera locations identical to those of the training images; the other datasets "test" includes images taken from camera locations distinct from the ones where the training images were captured. Accuracy in classifying the "id_test" images indicates the quality of the model. Comparing the accuracy in predicting "id_test" and "test" images show the generalizability of the model. The difference is also an evaluation of the effectiveness of the background removal and normalization technique in addressing the domain adaptation problem. Table 1 displays the model's performance on various subsets of data. The table indicates that using background removal techniques leads to higher accuracy in classifying animals in both the photos taken from the same location as the training set and those from different locations.

Table 1: Compare accuracy for model generalizability

Model	Baseline Model	Model with Background Removal
"id_test" accuracy	27.2%	37.1%
"test" accuracy	23%	35.7%
balanced "test" accuracy	6%	13%

In consideration of actual application. We focused on improving the time efficiency of the background removal. In Table 2 below, we compared the average time to preprocess an image and predict an image with and without background removal. Even with background removal, the average time it takes to predict an image is 0.0108 second without GPU. This is notably shorter than the typical 0.03 seconds it takes for a camera trap to capture a frame of an image.

We have observed that our model has better prediction accuracy and general-

Table 2: Compare time efficiency of background removal

Model	Baseline Model	Model with Background Removal
preprocess	0.006 per image	0.0027 per image
predict	0.008s per image	0.0108s per image

ization performance on images captured during the daytime. As Table 3 shows, the model improves the test accuracy in predicting daytime image by 11.7% but only improves the test accuracy in predicitng nighttime image by 8.7%. Daytime images exhibit distinct coloration with varying values in each RGB channel per pixel. In contrast, images captured during low-light conditions or nighttime appear grayscale, with similar values in each RGB channel per pixel. In the masking process, we keep a pixel's original value if any RGB channel has a value that differs from the background by a certain threshold. Consequently, our background removal algorithm performs more effectively on daytime images, facilitating the identification of animal-containing regions.

Table 3: Compare accuracy of model prediction over daytime nighttime images

Model	Baseline Model	Model with Background Removal
"id_test" daytime	27.4%	39.3%
"id_test" nighttime	26.8%	34.4%
"test" daytime	23%	34.7%
"test" nighttime	23.2%	32.1%

6 Discussion

The use of background removal in domain adaptation for classification tasks in camera trap photos is a promising method for improving model accuracy by reducing the variation between camera trap deployments. Our study demonstrates that background removal can be an effective method for capturing animal features while minimizing the impact of background variation. However, there are limitations to this method, such as the risk of losing important animal features in images with small or similar-colored animals, and the need for sufficient data to extract representative background images. In addition, the filtering process used to preserve animal pixel coloration may not work well with highly variable animal coloration across camera trap deployments. Future work could explore alternative methods for addressing these limitations, such as incorporating additional information on animal coloration or using alternative techniques for identifying important pixels in filtered images.

Due to computational power limitations, we were only able to test a limited number of combinations and record the performance of our animal type classification model. However, it would be highly beneficial to further test the model

with background removal techniques, especially in other related tasks such as animal counting in each image. We also want to test background removal with more kinds of CNN model. Such tests would help to provide a more accurate assessment of the model's performance and suitability of background removal technique for diverse scenarios.

Overall, the background removal process is a valuable tool in the domain adaptation toolkit for classification tasks in camera trap photos, and further research could improve its effectiveness and potential applications.

7 References

[1] Koh, P. W., Sagawa. (2021, July 16). Wilds: A benchmark of in-the-wild distribution shifts. arXiv.org. Retrieved March 8, 2023, from https://arxiv.org/abs/2012.07421
[2] "IWildCam 2022 - FGVC9." Kaggle, https://www.kaggle.com/c/iwildcam2022-fgvc9.
[3] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105)