International Journal of Geo-Information

MDPI

Article

An Adaptive Embedding Network with Spatial Constraints for the Use of Few-Shot Learning in Endangered-Animal Detection

Jiangfan Feng * and Juncai Li

College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; s190231193@stu.cqupt.edu.cn
* Correspondence: fengjf@cqupt.edu.cn

Abstract: Image recording is now ubiquitous in the fields of endangered-animal conservation and GIS. However, endangered animals are rarely seen, and, thus, only a few samples of images of them are available. In particular, the study of endangered-animal detection has a vital spatial component. We propose an adaptive, few-shot learning approach to endangered-animal detection through data augmentation by applying constraints on the mixture of foreground and background images based on species distributions. First, the pre-trained, salient network U2-Net segments the foregrounds and backgrounds of images of endangered animals. Then, the pre-trained image completion network CR-Fill is used to repair the incomplete environment. Furthermore, our approach identifies a foreground–background mixture of different images to produce multiple new image examples, using the relation network to permit a more realistic mixture of foreground and background images. It does not require further supervision, and it is easy to embed into existing networks, which learn to compensate for the uncertainties and nonstationarities of few-shot learning. Our experimental results are in excellent agreement with theoretical predictions by different evaluation metrics, and they unveil the future potential of video surveillance to address endangered-animal detection in studies of their behavior and conservation.

Keywords: few-shot learning; GIS; species distributions; spatial constraints

Citation: Feng, J.; Li, J. An Adaptive Embedding Network with Spatial Constraints for the Use of Few-Shot Learning in Endangered-Animal Detection. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 256. https://doi.org/10.3390/ijgi11040256

Academic Editor: Wolfgang Kainz

Received: 5 January 2022
Accepted: 2 April 2022
Published: 14 April 2022

1. Introduction

The biodiversity crisis, i.e., the worldwide loss of species and damage to ecosystems, has continued to accelerate. Studying endangered animal presence and behavior is critical in addressing environmental challenges, such as invasive species, climate, and land-use change [1]. Driven by advances in data collection and computer vision technologies for the detection and tracking of wildlife, biodiversity research is rapidly transforming into a data-rich discipline. Imaging data have become indispensable in the retrospective analysis and monitoring of endangered-animal species' presence and behavior [2]. It is infeasible to exploit image data manually for the application situation. Thus, it is necessary to incorporate uncertainty and propose automated detection methods.

Traditional GIS and spatial analysis have limitations in model complexity in addressing big data that are complicated by nature. Recent successes in deep learning have led to the use of automated computational methods to monitor endangered animals, including automatic video- and image-processing techniques for the fine-grained recognition of various categories of objects and animals [3]. However, the development of these methods is confounded by large-scale training datasets. As a result, the model exhibits severe overfitting in few-shot applications, such as endangered-animal detection, and it typically cannot work due to the extreme lack of training examples. To address this, recently, many studies have focused on few-shot object-detection models [4–6]. In general, these methods mainly include two training stages. First, meta-training [7,8] is carried out through the use of many examples from the base class so that the model can obtain the ability to generalize,

independently of the current task. Second, only a few novel-class examples are used for fine-tuning the training so as to complete the object-detection task with unknown models. However, these methods have limited applicability in context sensitivities, providing only single and straightforward scenes. In contrast, practical wild animal detection emphasizes various objects and contexts. Thus, direct applications of existing few-shot learning solutions perform unsatisfactorily when tasked with endangered-animal detection.

Here, we demonstrate a framework that proposes adapting to a dynamically changing environment for endangered-animal detection, allowing for few-shot learning to be applied in different scenarios with spatial dependency. Furthermore, it exploits a new proposal: a set of objects and environments is processed, composed, and affect each other simultaneously, instead of being recognized individually. For example, given an endangered-animal image, the module first uses the pre-trained saliency network U2-Net [9] to segment its foreground and background. Then, it uses the pre-trained image inpainting network CR-Fill [10] to repair missing parts. Finally, we mixed the foreground and background from separate images to generate new samples. While foreground–background mixture can be effective for data augmentation, the program lacks geographical semantic knowledge. This may result in many backgrounds that are anachronistic to the foreground. For example, the *Nipponia Nippon* cannot appear in the desert. Considering the environmental correlation between endangered animals and environmental factors, we thus propose a knowledge dictionary and a relation network within a spatial constraint framework. By computing relation scores in the framework, we developed an a priori spatial dependence model with the intention of limiting the mixing of foreground and background images to only those that are rational.

The proposed module is general and not limited to animal detection, and it serves as an essential building block that is flexibly usable in any architecture. Our main contributions are as follows:

(1) We construct a new image dataset dedicated to detecting endangered animals, including independent animal categories, whether they are endangered or not. The endangered species are represented in only a few samples, while the unendagered species are represented in more examples.
(2) We propose a new framework for endangered-animal detection using few-shot learning, which can be useful in unknown scenarios by augmenting the synthetically generated parts of separate images.
(3) We provide a spatial-constraints model with two main components: a knowledge dictionary and a relation network, thus enabling the avoidance of mixing foreground species with incompatible background images from a geospatial perspective.

2. Related Works

The proposed method deals with a challenging task in the process of detecting endangered animals, one which arises in data augmentation and few-shot learning, which are closely related to two streams of research in the literature, i.e., the data-driven detection of wild animals (Section 2.1) and few-shot learning (Section 2.2).

2.1. Data-Driven Detection of Wild Animals

The emerging field of data-rich science has adopted computer vision models for pattern recognition to identify numerous animal species through phenotypic appearance [11]. In practical terms, individual animal identification methods have been applied to lemurs [12], macaques [13], and chimpanzees [14]. A substantial hurdle is developing robust models to perform on highly challenging datasets, such as those with cluttered backgrounds [15]. The issue could accomplish background subtraction with thresholds and color matching [16]. Recently, deep learning has made it possible to analyze animal biometric datasets and perform complex imaging tasks, including segmentation and classification [17,18]. While these methods have made valuable contributions, they are not robust to inevitable environmental variations. The challenge is to develop models that would

work well in images with poor illumination or uncertain environments [19]. Concerning the studies using deep learning, individual reidentification has been critical to research on animal behavior in the wild [20]. Still, this alone cannot capture the full complexity of animals and environments. Existing approaches based on temporal or spatial features are limited in adaptive capability. Additionally, they require a stable and stationary environment, making it inherently challenging to detect unknown endangered animals. In practice, the essence of wild-animal detection approaches lies in the quick adaptation to novel data, thus remaining an open challenge.

In contrast, our work takes a step forward by formulating an adaptive context-aware module as a few-shot learning task and utilizes a saliency network and inpainting network for recognizing wild animals to alleviate the challenge.

2.2. Few-Shot Learning

Recent successes in artificial intelligence offer extensive benefits. However, it remains a significant challenge for computer vision approaches to learn new concepts from very few examples [21,22]. Millions of images from camera traps and other studies are difficult for researchers to process. An essential focus for early development is the composition of objects from parts using probabilistic models [23]. Advances in deep learning have made it possible to use data augmentation [24,25], augment memory [26–28], and perform meta-learning [29,30]. For example, Chen et al. [25] designed a novel image deformation network, which learns to synthesize additional images by fusing a pair of reference images. Ramalho et al. [27] showed that the APL algorithm could perform state-of-the-art baselines on few-shot classification benchmarks with a smaller memory footprint. Wang et al. [29] proposed TAFE-Nets to learn how to adapt the image representation to a new task in a meta-learning fashion. In addition, domain-agnostic few-shot recognition has been presented [30], where the domain of the testing task is unknown. However, deep learning and augmented analysis focus on available features, the encoder lacks adaptation ability, and the performance is limited. Many studies have focused on the multiscale features and relations between samples to classify few-shot learning [31–33]. After learning in various environments, a task-agnostic classifier could better capture the characteristics. Despite performance improvement, task-adaptive parameters miss the necessary details of actual data, limiting these types of methods. In surveillance scenarios, these approaches require the objects and settings to be repeatable, which is challenging for endangered-animal detection in unknown environments.

In contrast, we apply few-shot learning to a novel visual scene, i.e., endangered-animal detection, challenging current frameworks. Its purpose is to learn a robust model with a few samples to recognize novel visual compositions for endangered animals. We take advantage of the general saliency network and the inpainting network to learn an adaptive context-aware module that adapts to few-shot tasks from different domains efficiently and effectively.

3. Proposed Method

3.1. Overview

This study mainly developed a combination of data-driven and geographical knowledge-driven strategies. The main difference between our approach and previous few-shot detection methods is an improved data augmentation algorithm for the new examples that will enhance model robustness against environment variations with spatial adaption. Thus, our approach can incorporate powerful deep learning capabilities and high-level dependencies and relations from geospatial. Here, we provide an overview of our method (see Figure 1). Our approach begins with two data collection phases, during which we apply the reweighting module and few-shot detector. First, we collect data in ordinary animal categories and use these data to train a module that is a reweighted combination of features, which is a more general valuable approach in animal detection. After the module was pre-trained,

we collected data on the endangered animals and retrained the model, specializing in endangered-animal detection.

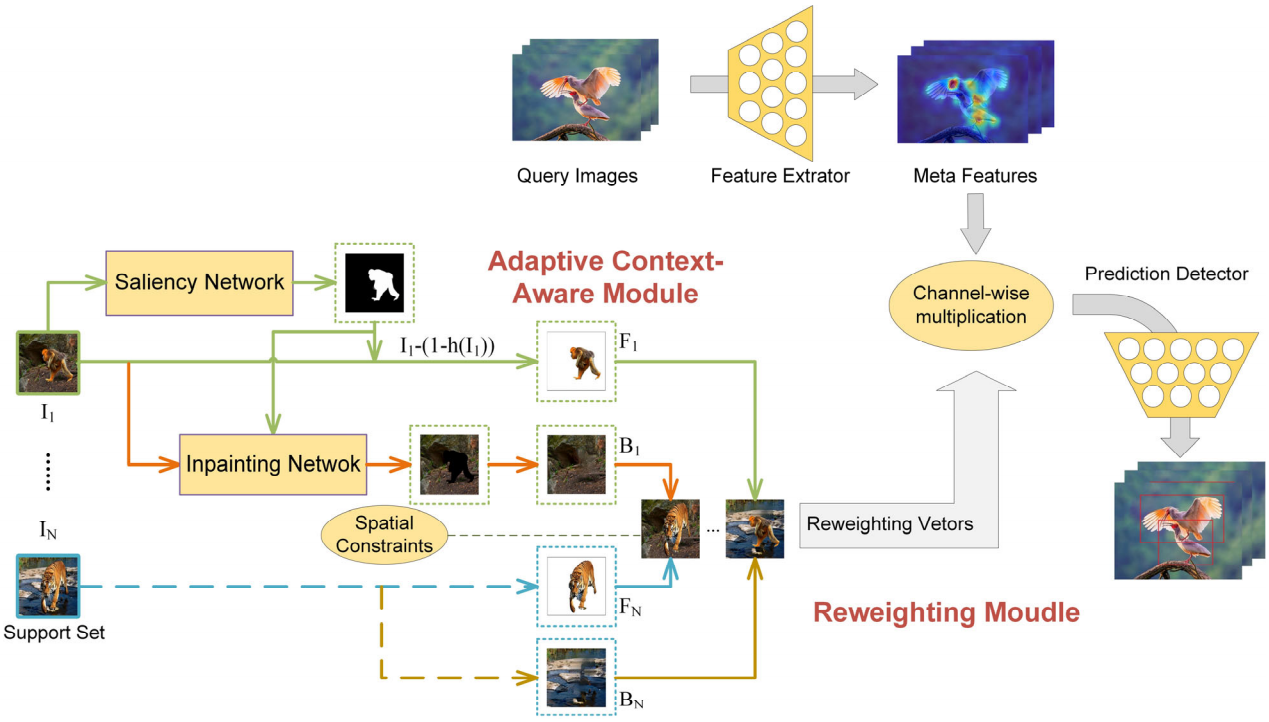


Figure 1. An overview of the proposed framework for few-shot endangered-animal detection.

3.2. Problem Definition

GIS is characterized by its distinctive spatial thinking and perspectives. It also needs to be developed, since geospatial data have been extended in the big data era. This requirement and the excellent ability of deep learning match the ever-increasing volume of biodiversity data. This study extracts the information potentially contained in the biodiversity images, improving the difficulty of applying traditional spatial analysis methods or classic machine learning models that are particularly sensitive to noisy or insufficient samples. We let a supported dataset contain a small number of endangered-animal categories within this framework, and an auxiliary dataset includes fine examples of ordinary animal categories. For a query image q including endangered animals, we aim to use the supporting dataset and supplementary dataset to learn how to automate species classification and draw their bounding boxes. Suppose the supporting dataset contains N endangered-animal species, and each species contains K samples. In that case, the problem is called the N-way K-shot detection task in endangered-animal scenes.

Inspired by the generative learning model [34], we form a scene-shared view in the new reconstruction paradigm. It would be helpful to adapt to nonstationary and uncertain environments. To reflect scene changes and uncertainties, we propose a few-shot object detection pipeline with adaptive context awareness named FRW-ACA. Our approach makes full use of the context information in the images of endangered animals. To make the process tractable, we use a feature extractor that learns how to extract meta-features from the input query image for endangered-animal detection. For adaptation to the detection task, we also use the feature reweighting module to convert several support instances in the endangered-animal category into a global vector, representing the importance or correlation of meta-features for detecting the corresponding object. The detection predictor regresses the features and directly outputs the object classification score and the bounding box.

Furthermore, the foreground objects may not match all background images for the foreground–background mixture from the geospatial aspect. The wrong mix is used in the data augmentation, leading to an unreliable model. Thus, it is essential to note that the

mixture can be interpreted as the probability that a specific foreground species matches particular background images, i.e., the spatial constraints module has been performed.

3.3. The Adaptive Context-Aware Module

The adaptive context-aware module first uses the salient network U2-Net to segment the image into foreground and background. Then, it uses the image inpainting network CR-Fill to repair the missing part of the background. Finally, we mixed the foregrounds and backgrounds from separate images to generate new samples with spatial constraints. In the case of endangered-animal detection, the general detection model is unstable in such a scene due to the uncertainty and variability of the environment. Our module is based on the pretraining network and is independent of a specific network, making the module adapt to the changing environment. Thus, it can easily embed this aware adaptive ability for novel contexts in other networks.

3.3.1. Saliency Network

As a pre-processing step, super-pixel generation algorithms [35] benefit the saliency network. We assume that each image I^m is independently represented by its n_m super pixels. Given the part descriptors $\{x_i^m\}_{i=1}^{n_m}$ for each image I^m, they are fed into a six-stage encoder and a five-stage decoder, which outputs a saliency map. In the training process, the loss is combined with the side output saliency map ℓ_{side}^m and the final fusion output saliency map ℓ_{fuse}. For each term ℓ, the cross-entropy loss is used as follows:

$$\ell = \sum_{i=1}^{m} [P_{G_i} \log P_{S_i} + (1 - P_{G_i}) \log(1 - P_{S_i})] \tag{1}$$

where P_{G_i} and P_{S_i} denote the ith super-pixel values of the ground truth and the predicted saliency probability map, respectively.

We obtain the corresponding saliency map $u(I^m)$ through the saliency network u, then the foreground:

$$F_{I^m} = I^m - (1 - u(I^m)) \tag{2}$$

3.3.2. Inpainting Network

The saliency map obtains the background with a single transition and ignores the foreground region. Therefore, to extract all of the background features in the subsequent training and avoid the instability caused by region loss, we use the image inpainting network $r(\cdot)$ to repair the missing region and obtain the complete background:

$$B_{I^m} = r(I^m, u(I^m)) \tag{3}$$

We adopt two hybrid strategies when mixing the foreground and the background [36]. The first strategy is intra-class mixing. Its corresponding foreground is combined with the same background category for an endangered-animal image to form a new image. Each image can generate an additional $K - 1$ new image through this strategy. Nevertheless, it cannot be effective under a one-shot setting because each category contains only one image under this setting.

The second strategy is inter-class mixing. Under this strategy, we can incorporate the foreground and background between different endangered-animal images. These newly developed image samples will join the supporting dataset and follow the annotation information of their foreground objects to ensure supervised training. Intuitively, the distribution of wild animals has significant geographical characteristics. Thus, we deal with yet another approach: spatial constraints. We assume that we have some prior knowledge regarding the mixed sources; here, it is in the form of background images correlated with the mixed foregrounds.

3.3.3. Spatial Constraints

Although the foreground–background mixture of separated images can generate multiple new image examples, there are many complications in integrating foreground animal species into uncertain backgrounds. For example, the wild panda lives only in the mountain bamboo forests of southwest China, and we cannot mix the panda with a desert background. Thus, this should be verified by filtering inappropriate background images. To model associations between foreground species and background selection, we created knowledge dictionaries and relation networks to form the basic framework of our spatial constraints method, which we discuss in detail. According to the knowledge dictionary and relation network, the spatial constraint function incorporates foreground objects into a priori spatial dependence models. We now describe our implementation architecture of spatial constraints.

(1) Knowledge Dictionary. The knowledge dictionary module has an image dataset that takes the name of the foreground species (species name) as input and returns a set of conceptually related background image groups as output (species distribution scenes). Figure 2 presents the architecture of the knowledge dictionary.

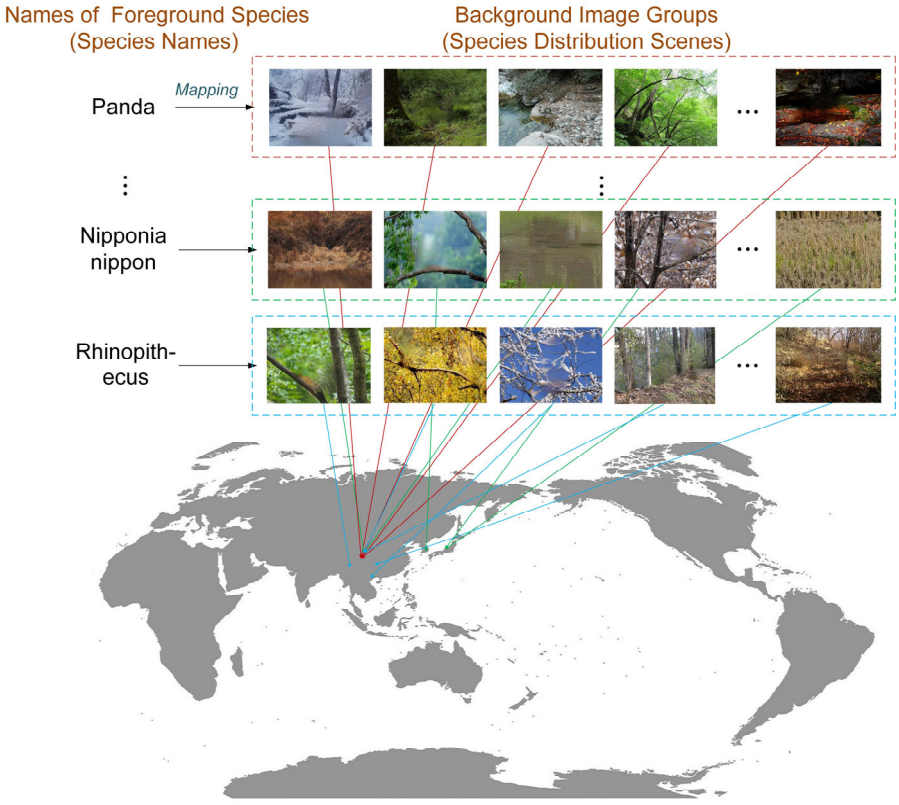


Figure 2. Architecture for the knowledge dictionary. Before data acquisition, background image groups are geotagged (GPS coordinates or GeoName). All the images are captured from an online data-sharing platform and pre-processed by the Saliency Network and Inpainting Network. Note that the number of downloaded pictures is variable in various species. After spatial sampling, the coordinates are not necessary.

The knowledge dictionary needs to balance spatial analysis's precision and computational speed. We address this problem by applying spatial sampling methods for species distribution scenes to acquire background image groups. An intuitive way to produce samples is sampled across different spatial locations considering various factors such as distribution and weather factors. For each species, we randomly select a geotagged photo from a distinct region, calculate the boundaries, and overlay the Voronoi polygons on the chosen sample. This can be easily performed with GIS software platforms.

While the Voronoi polygons partition the area, we randomly select geotagged images that mirror each region's sampling density and manually remove some samples considering weather factors, resulting in a background image group for a species.

The knowledge dictionary D is a set of mappings $F \rightarrow B$. For clarity, we will distinguish between the two classes, names of foreground species (F) and background image groups (B), where F refers to the name of endangered-animal species indexed for lookup in a dictionary, and B refers to a definition of an F in a dictionary. Let F_i be the ith species in the training foreground dataset, where $i \in \{1, 2, \ldots, N\}$. Additionally, N is the number of endangered-animal species in the training dataset and the definition of F_i in the dictionary is $B_i = \left[B_i^1, B_i^2, \ldots, B_i^J\right]$, where J refers to the number of typical backgrounds. For example, consider the following dictionary mapping: "panda" $\rightarrow \{I_j, j = 1, ..., J\}$, where $\{I_j\}$ is a background image group composed of living areas of wild pandas in different seasons. Note that J is variable in various species. In this approach, geographical distribution and season variation are applied to select typical background images of animal species to improve the robustness against appearance variations. The knowledge dictionary is simple and easy to extend.

(2) Relation Network. This paper adopted the relation network to analyze the spatial correlations between background images and species distribution scenes quantitatively. From traditional GIS and spatial analysis, the geolocation of images or determining the scene's physical location is required. However, only a tiny fraction of images are geotagged in our datasets and online datasets. Here, we suggest that the images' vision characteristics can provide valuable information on the extent of spatial similarities. Therefore, we choose to learn spatial correlation in an end-to-end manner. The relation network aims to learn a transferrable deep metric to compare the relation between background images from the Inpainting Network (Section 3.3.2) and species distribution scenes from the knowledge dictionary.

Inspired by the recent advances in relation networks trained end-to-end from scratch [37], a relation network in the spatial constraints framework is proposed (see Figure 3). Once trained, the network can produce spatial correlations by computing relation scores between background images and species without further updating. We emphasize that we do not assign geographic coordinates to images. Instead, the relation network is a Similarity-Measure-Based and End-to-End model that learns similarity using visual representations.

We assume that $c(.,.)$ uses depth concatenation to be the operator of background images and species distribution scenes, f_φ is the embedding module, and $f(\cdot)$ produces the feature maps. The inpainted background images and species distribution scenes are fed into the relation module g_ϕ, which eventually creates a scalar in the range [0, 1] representing the scene similarity between the species distribution scenes B_i and inpainted background images from the Inpainting Network. Thus, we generate relation scores $r_{i,j}$ for the relationship between the species and possible backgrounds:

$$r_{i,j} = g_\phi(c(f_\varphi(B_i), f_\varphi(b_j))) \;\; j = 1, 2, \ldots, M \tag{4}$$

where M is the number of background images from the Inpainting Network; the maximum value of M is $K - 1 + K(N - 1)$.

We use mean square error (MSE) loss (Equation (5)) to train the model, regressing the relation score $r_{i,j}$ to the ground truth: matched pairs have a similarity value of 1, and the mismatched pair has a similar value of 0.

$$\varphi, \phi \leftarrow \underset{\varphi,\phi}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{j=1}^{M} \left(r_{i,j} - 1\right)^2 \tag{5}$$

Therefore, the fused image I_u is obtained from the endangered animal and related background images, denoted by f_i and b_j. Concretely, the generated model is written in the following compact form:

$$I_u = \alpha f_i + \beta b_j \quad j = 1, 2, \ldots, c \tag{6}$$

where α and β are regularization parameters, c is the threshold and means the number of selected background scenes ordered by $r_{i,j}$. (We choose the top 40% in this paper, and the value is experimentally observed).

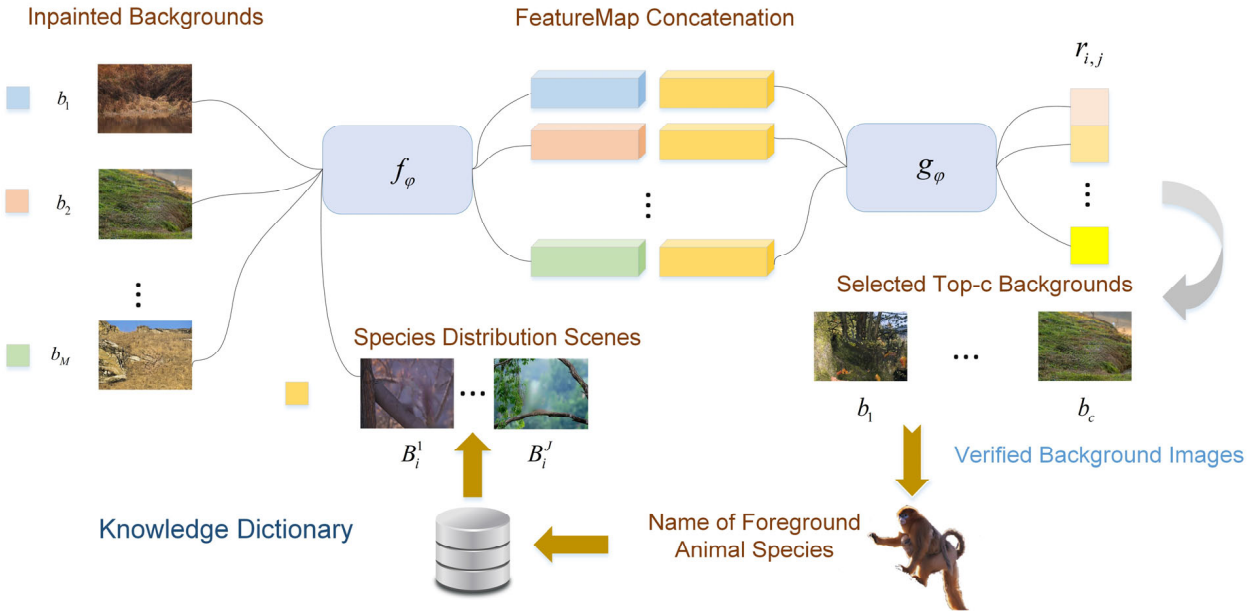


Figure 3. Architecture for the selected top-c backgrounds with knowledge dictionary and relation network. The goal is to automatically detect the relevance of the inpainted background images and species by measuring the image similarity. Note that all pictures do not assign geographic coordinates because images in the dataset often do not have geotags. While we have not chosen geotagged photos, this framework allows us to develop more general models.

3.4. Reweighting Module

We further investigated the feature reweighting module $\mathcal{M}$ taking $(I, Mask)$ as the input and embedding it into the class-specific representation:

$$\omega_i = \mathcal{M}(I_i, Mask_i) \tag{7}$$

where i represents the index of endangered-animal classes, I_i denotes the ith subset of the image dataset I, and $Mask_i$ is the ith mask. We can obtain the class-specific part F_i by:

$$F_i = F \otimes \omega_i \tag{8}$$

where $\otimes$ denotes channel-wise multiplication. The class-specific feature F_i will be transmitted to the detection predictor $\mathcal{P}$ to obtain the objectness score o_i, offsets of the object position (x, y, h, w), the corresponding classification score c_i, and the corresponding classification score c_i:

$$\{o_i, x_i, y_i, h_i, w_i, c_i\} = \mathcal{P}(F_i) \tag{9}$$

For a set of jth detection tasks $\mathcal{T}_j$, the subsequent loss will be minimized to jointly optimize the feature learner $\mathcal{D}$, reweighted module $\mathcal{M}$, and detection predictor $\mathcal{P}$:

$$\min_{\theta_D, \theta_M, \theta_P} \sum_j \mathcal{L}(\mathcal{T}_j) = \sum \mathcal{L}_{det}\left(\mathcal{P}_{\theta_P}\left(\mathcal{D}_{\theta_D}\left(I_j^q\right) \otimes \mathcal{M}_{\theta_M}(\mathcal{S}_j)\right), Mask_j^q\right) \tag{10}$$

where θ_D, θ_M, and θ_P are the parameters of the feature learner $\mathcal{D}$, the reweighting module $\mathcal{M}$, and the detection predictor $\mathcal{P}$, respectively. The supported dataset $\mathcal{S}_j$ contains N samples from different categories $\left(I_j^q, Mask_j^q\right)$ representing the image and the corresponding annotation in the query dataset to evaluate model performance. The detector loss $\mathcal{L}_{det}$ is calculated from the classification score's cross-entropy loss [34], the classification score, the bounding box regression loss, and the objectness regression loss [38].

3.5. Optimization

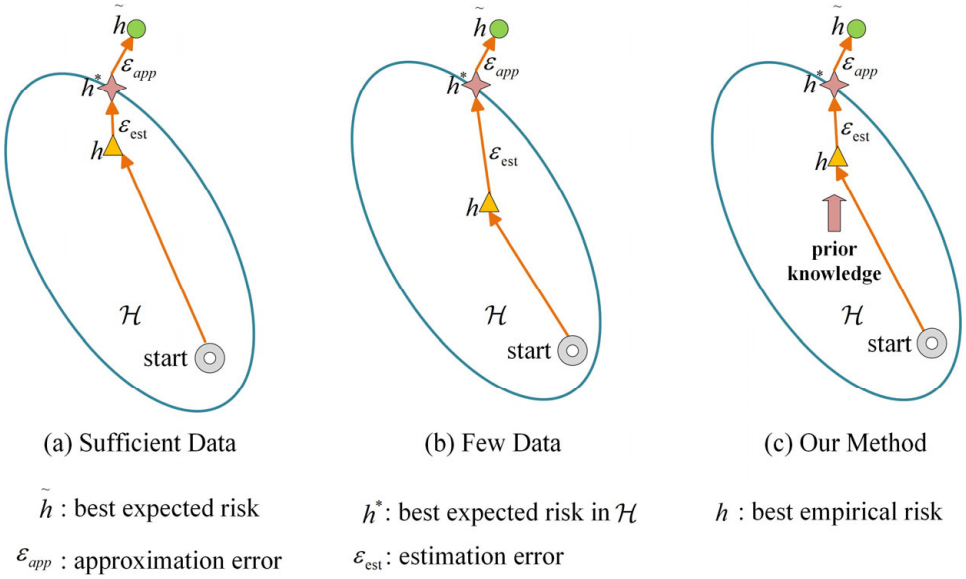
The empirical risk minimization (ERM) algorithm is a popular tool to optimize detection methods. Its basic principle is to select a classifier with the smallest value of a risk function. Given a hypothesis $h(\cdot)$ and data pairs $\{x_i, y_i\}$ by a joint probability distribution $p(x, y)$, we can transform the function optimization problem into the expected risk minimization problem:

$$R(h) = \int L(y, h(x)) dp(x, y) \tag{11}$$

For the sake of simplicity, we consider only using the empirical risk function $R_{emp}(\cdot)$ to replace the expected risk:

$$R_{\text{emp}}(h) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, h(x_i)) \tag{12}$$

In addition, the training required to process large volumes of data continues to limit the scale and depth of ERM [39]. We need to build a new optimization method for few-shot learning under the ERM paradigm. Figure 4a,b demonstrate the difference between sufficient and few samples in the machine learning process.



(a) Sufficient Data (b) Few Data (c) Our Method

$\tilde{h}$: best expected risk h^*: best expected risk in $\mathcal{H}$ h : best empirical risk

ε_{app} : approximation error ε_{est}: estimation error

Figure 4. Comparison of learning with sufficient samples, few samples, and our method (**a**–**c**).

We can use the above ACA module (mentioned in Section 3.3), which uses prior knowledge to achieve adaptive context-aware ability and dramatically improves the number of samples. Therefore, we can obtain a more accurate empirical risk minimization function $h(\cdot)$ (Figure 4c). Algorithm 1 shows the training procedure.

Algorithms 1 Detection Algorithm of the FRW-ACA Model

Input: Auxiliary set $\mathcal{A}$ and support set $\mathcal{S}$.
Output: Object position (x, y, h, w) and confidence score c.
1: Reorganize the training images $\mathcal{A}$ with multiple few-shot tasks $\mathcal{T}_j$.
2: **for** training epoch = 1 to 500 **do**.
3: Add each task into the training and adjust the θ_D, θ_M, θ_P to optimize Equation (10).
4: **end for.**
5: Generate additional images via the ACA module and add them to the $\mathcal{S}$.
6: **for** training epoch = 1 to 20 **do.**
7: Conduct the few-shot training to fine-tune the model using images $\mathcal{S}$.
8: **end for.**
9: Load the model and input the query image to get the result.

4. Experiments

In this section, we present experiments that use a new dataset constructed for the scene of endangered-animal object detection. In addition, we remove the spatial constraints of our method to demonstrate the benefits they provide.

4.1. Dataset

For evaluation purposes, the experiments were intended for few-shot endangered-animal detection in uncertain environments. We then need data from which animals, endangered or not, and uncertain backgrounds can be augmented together. However, these data are not available in the commonly used object detection datasets. There is a vast context gap between the images of these basic categories and the pictures of endangered animals. Figure 5 shows samples in the public dataset.

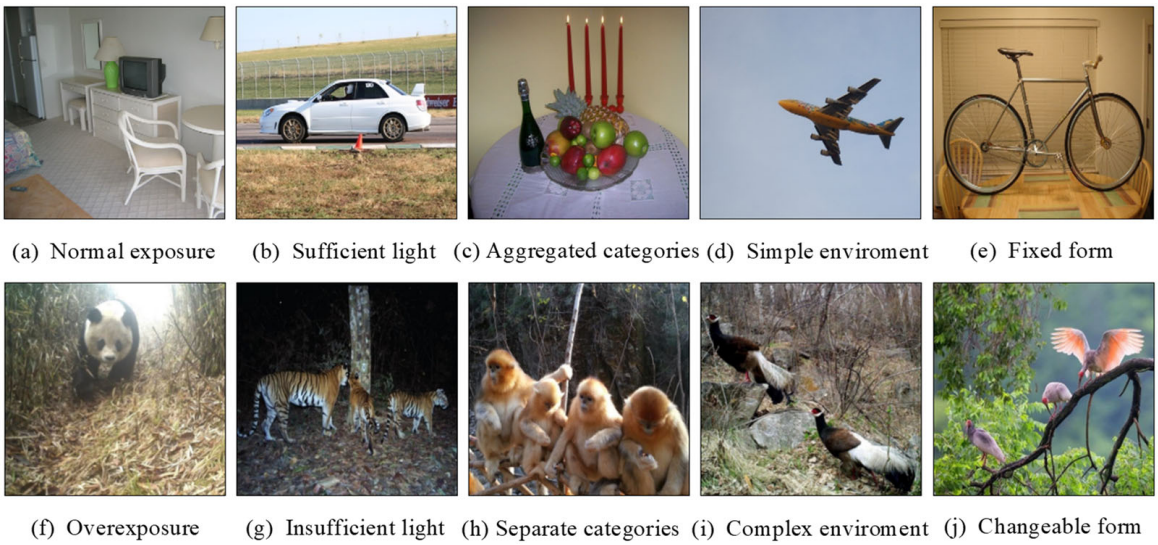


Figure 5. Example of the contextual gaps among images on the public dataset (**a**–**e**) and pictures of endangered animals (**f**–**j**).

Therefore, we constructed an EAOD (Endangered Animal Object Detection) dataset for fair evaluation. We obtained animal images from public sources such as OIDv4 [40], AwA2 [41], Flickr, and Google Images to construct our EAOD dataset. However, we cannot directly use these datasets because (1) the labeling standards of different datasets are inconsistent. For example, the same category of objects has other labels in different datasets. (2) Many images have poor labeling quality in the dataset. (3) In addition to the object detection dataset, the images from other sources do not contain the bounding box labeling required in the object detection task. Therefore, we first unified the labeling standards in EAOD to ensure that the same category objects have the same labeling name. Then, we

removed images that fit the object size, such as objects that even humans cannot accurately detect because these objects are not suitable either as base class samples or as novel class samples. Next, we carefully labeled the animal images collected according to the VOC [42] dataset format and divided the data into training and test sets according to the collection of few-shot learning.

This dataset contains 13,455 images and 20 categories, with 15 ordinary animal categories for base training. The remaining five categories are endangered animals for few-shot fine-tuning training. The frequent categories' training and test sets contain sufficient image samples for the entire detection task and are split into 8:2. The training set for endangered-animal categories uses few-shot settings (each category uses 1/2/3/5/10 objects), and the test set contains 100 images per category for evaluation.

4.2. Experimental Settings

All experiments use the same platform with Intel CPU Xeon Gold 6126*4, NVIDIA Tesla V100*2, 256G RAM, and Ubuntu 16.04. The mAP(mean Average Precision), Precision–Recall Curve (P-R Curve), and normalized AP are chosen as evaluation metrics.

It is challenging to learn a good meta feature learner $\mathcal{D}$ and a reweighted module $\mathcal{M}$ in model training due to the large gap between endangered and ordinary animal samples. Therefore, we used a two-phase learning scheme to ensure the model's generalization performance for endangered-animal categories. The first phase is base training, and we only use ordinary animal categories for training. Although many standard animal samples have been implemented, we are still faced with multiple few-shot detection learning tasks to ensure that the model performs well in object detection tasks of endangered animals. Specifically, we used a batch size of 64 to train the base class samples for 500 epochs, where the learning rate is 1×10^{-3}, momentum is 9×10^{-1}, and weight decay is 5×10^{-4}.

The second phase is the few-shot fine-tuning. Again, the samples of ordinary animal and endangered animal categories will be trained simultaneously, and each class will have only K annotated bounding boxes that can be used. This phase of the training process is the same as the first phase but requires fewer iterations because the model will take full advantage of the relevant features. Therefore, at this stage, we used a batch size of 4 to conduct 20 fine-tuning epoch training for all categories in which the learning rate was adjusted to 1×10^{-4}, and other settings were consistent with the base training stage.

4.3. Baseline

Our few-shot detector is constructed based on the detection framework of the one-stage model. Considering the most advanced one-stage models combined with few-shot learning strategies, we choose the YOLOv4 model [43] as the baseline, including the YOLOv4-joint and YOLOv4-ft, and YOLOv4-ft-full. Another baseline is a general few-shot object detection model FRW, which mainly uses the feature reweighting module to make full use of the characteristics of the base class. It extracts the generalized meta-features to detect novel objects. We also compare our method with other object detection models, such as EfficientDet [44], CenterNet [45], RetinaNet [46], and Meta R-CNN [47].

4.4. Results and Comparisons

Results of EAOD. The quantitative results are tabulated in Table 1, and the detection samples are shown in Figure 6.

Table 1. Few-shot detection performance (mAP) on the endangered-animal categories of the EAOD. (* indicates remove spatial constraints in the mixture).

Method/Shot	1-Shot	2-Shot	3-Shot	5-Shot	10-Shot
EfficientDet-ft [44]	0.07	0.34	2.61	7.52	18.74
EfficientDet-ft-full [44]	1.14	3.58	12.96	19.46	37.86
CenterNet-ft [45]	0.13	0.95	3.64	8.79	19.59
CenterNet-ft-full [45]	0.89	5.31	15.37	23.54	42.76
RetinaNet-ft [46]	0.26	1.29	4.15	7.43	21.31
RetinaNet-ft-full [46]	1.48	6.75	14.47	22.21	44.38
YOLOv4-joint [43]	0.00	0.00	0.29	0.86	19.76
YOLOv4-ft [43]	0.00	0.42	5.84	8.34	25.81
YOLOv4-ft-full [43]	1.31	7.93	17.04	24.93	49.93
FRW [34]	25.51	48.15	55.07	65.01	73.08
Meta R-CNN [47]	21.13	46.37	56.92	67.34	75.13
FRW-ACA(Intra, Ours)	-	50.67	58.97	69.70	77.01
FRW-ACA(Inter *, Ours)	42.39	53.60	60.23	70.06	78.56
FRW-ACA(Inter, Ours)	43.18	54.51	61.49	71.18	80.42

Figure 6. Detection samples of our 10-shot detection on endangered animals, including bounding boxes, categories, and confidence scores.

Table 1 shows our method with various baseline detection performances (mAP) on the EAOD dataset. To fully demonstrate the performance of all procedures in endangered-animal detection, we adopted five few-shot settings for the experiments, namely, 1-shot, 2-shot, 3-shot, 5-shot, and 10-shot. Table 1 summarizes the results, demonstrating that our method is superior to other baselines under few-shot settings, especially in the set of 1-shot. From another perspective, the results verify the robustness of our model for novel samples, which is particularly prominent in extreme few-shot scenes. The YOLOv4-joint performance is inadequate, especially in the first three-shot settings. In addition, the result shows that it is difficult to obtain a generalized model when the base samples and the novel samples are trained simultaneously in the few-shot scene, such as detecting endangered animals. It also shows that our two-phase training scheme is better, especially with spatial constraints. In addition, Figure 6 shows some qualitative results on EAOD using our FRW-ACA model with spatial dependency.

To evaluate the performance more comprehensively, we propose a comparison of P-R curves (Figure 7). Compared with other methods, the significant difference between ours and those mentioned above is that it performs adaptive actions that respond to environmental changes. Furthermore, our approach significantly improves the performance, showing the best performance on the EAOD dataset.

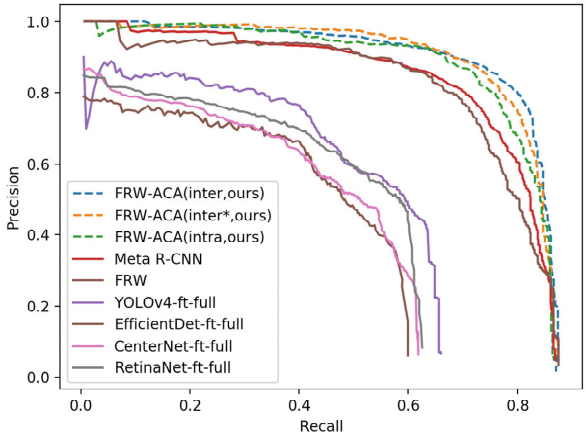


Figure 7. Precision–recall curves of our model and other baselines on the EAOD dataset. (* indicates remove spatial constraints in the mixture).

Time-Cost. Figure 8 shows the learning speed from different methods. Intuitively, our system may exhibit more time because it produces many new samples containing additional semantic information during training, which the network had never seen before. However, the result shows that our model still maintains a fast learning ability based on FRW, and only a few epoch iterations are needed to approach the convergence value. On YOLOv4, the index continuously rises, and more iterations are required to adapt the novel scene. Here, we only use the single curve of FRW-ACA because our model fully uses the knowledge learned in the base training phase. Thus, the model converges quickly in the endangered animals' training phase; there is no significant difference in learning speed among the three adaptive strategies.

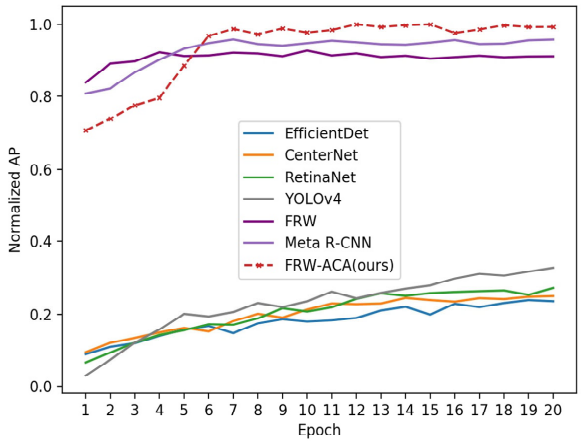


Figure 8. Comparison of learning speed. We plot the normalized AP against the number of training epochs.

Cross-dataset Evaluation. In addition, to demonstrate our model's generalization ability and adaptive context-aware ability in the object detection scene of endangered animals, we also use the dataset benchmark VOC widely used in object detection to evaluate the model on endangered-animal categories. Specifically, we first used the images in the VOC dataset for base training and then used the pictures of endangered animals for fine-tuning training under few-shot settings. Finally, we evaluated the performance of endangered-animal categories on the EAOD dataset. The number of mAP for each few-shot detection method varied somewhat due to shot numbers, as shown in Table 2; our model enables high precision with few shots.

Table 2. Few-shot detection performance (mAP) for cross-dataset evaluation. (* indicates remove spatial constraints in the mixture).

Method/Shot	1-Shot	2-Shot	3-Shot	5-Shot	10-Shot
EfficientDet-ft [44]	0.04	0.28	1.06	3.69	10.48
EfficientDet-ft-full [44]	0.12	1.37	5.52	12.16	28.63
CenterNet-ft [45]	0.09	0.52	1.48	4.72	12.09
CenterNet-ft-full [45]	0.15	2.84	7.09	17.12	30.23
RetinaNet-ft [46]	0.12	0.39	1.27	4.93	12.64
RetinaNet-ft-full [46]	0.18	2.67	7.22	16.43	33.19
YOLOv4-ft [43]	0.00	0.66	1.51	3.83	17.18
YOLOv4-ft-full [43]	0.00	4.51	10.98	22.05	42.61
FRW [34]	10.17	32.86	42.03	53.86	62.95
Meta R-CNN [47]	7.38	28.91	44.97	57.23	66.42
FRW-ACA(Intra, Ours)	-	37.32	47.28	60.56	68.27
FRW-ACA(Inter *, Ours)	26.32	38.26	50.09	60.83	71.45
FRW-ACA(Inter, Ours)	26.91	39.05	51.27	62.34	74.91

As a result, our model still has the best performance index. However, we observed that the mAP values of our model and other baselines decreased to varying degrees compared to the EAOD dataset. The result is mainly because the image categories in the VOC dataset are quite different from those of endangered-animal image categories. Therefore, the model cannot fully use the features learned in the primary stage. However, it also shows that the EAOD dataset constructed by us has better applicability in the object detection scenario of endangered animals.

4.5. Analysis

Our research provides evidence that spatial constraints might permit a more accurate data augmentation, and the images' vision characteristics can provide helpful information on the extent of spatial similarities. Furthermore, our method improves detection precision from the overall experimental results, mainly due to the adaptive context-aware module and the EAOD dataset. The adaptive context-aware module helps the framework obtain more context semantics with low samples. Then, the model has greater generalizability and adaptability in novel scenarios. Furthermore, the EAOD dataset compensates for the context gap between ordinary datasets and examples in endangered-animal scenes. As a result, the model can fully use the features learned in the primary training stage.

5. Conclusions and Future Work

The data explosion has posed challenges and opportunities to the geographical information community. GIS needs to be extended to accommodate sensors' dynamic observations, including volunteered geographic information. By incorporating the technologies emerging from geographical data science and computer vision, we can turn big data into useful information and knowledge that more efficiently serves biodiversity research.

Overall, the methodology presented here opens interesting paths to analyzing biodiversity data. Unlike entirely data-driven automation methods, our method is based on knowledge-driven and data-driven strategies. It also provides image primitives to produce a rich set of examples by incorporating contextual detail, which can be valuable for boosting learning sample efficiency, especially for endangered-animal detection. Furthermore, the adaptive embedding network can offer valuable insights into the generative learning framework for few-shot learning, leading to adaptive, more capable, and efficient processing. Such an approach should also serve as an embedding building block that is flexibly usable in any architecture.

There are some limitations to our study, notably the size of our dataset. Another limitation concerns the fact that, for individual-level recognition, our method is heavily reliant on the detection performance of the prominent parts. For example, the detector often fails to detect animal infants. In the big data era, the role of geospatial data service

needs to change from data warehouses to smart information providers [48]. Therefore, future directions to improve our framework include adopting multiple variables and data augmentation to enhance accuracy and generalizability and extending traditional GIS and spatial analysis in biodiversity research. Furthermore, few-shot detection that is distinctive exists in various domains. We envisage possible applications for our framework in, for example, wild animal detection for video surveillance, automated behavior recognition, etc.

Author Contributions: Methodology, Jiangfan Feng and Juncai Li; investigation, Jiangfan Feng; data curation, Juncai Li; validation, Juncai Li; writing—original draft preparation, Jiangfan Feng and Juncai Li; writing—review and editing, Jiangfan Feng. All authors have read and agreed to the published version of the manuscript.

Funding: The work is supported by the National Natural Science Foundation of China (41971365) and the Chongqing Research Program of Basic Science and Frontier Technology (cstc2019jcyj-msxmX0131).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Haucke, T.; Steinhage, V. Exploiting depth information for wildlife monitoring. *arXiv* **2021**, arXiv:2102.05607.
2. Caravaggi, A.; Banks, P.; Burton, C.; Finlay, C.M.V.; Haswell, P.M.; Hayward, M.W.; Rowcliffe, M.; Wood, M. A review of camera trapping for conservation behaviour research. *Remote Sens. Ecol. Conserv.* **2017**, *3*, 109–122. [CrossRef]
3. Yang, L.; Luo, P.; Change, L.C.; Tang, X. A large-scale car dataset for fine-grained categorization and verification. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
4. Ji, Z.; Liu, X.; Pang, Y.; Ouyang, W.; Li, X. Few-Shot Human-Object Interaction Recognition with Semantic-Guided Attentive Prototypes Network. *IEEE Trans. Image Process.* **2020**, *30*, 1648–1661. [CrossRef]
5. Li, X.; Wu, J.; Sun, Z.; Ma, Z.; Cao, J.; Xue, J.-H. BSNet: Bi-Similarity Network for Few-shot Fine-grained Image Classification. *IEEE Trans. Image Process.* **2020**, *30*, 1318–1331. [CrossRef] [PubMed]
6. Liu, B.; Yu, X.; Yu, A.; Zhang, P.; Wan, G.; Wang, R. Deep Few-Shot Learning for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 2290–2304. [CrossRef]
7. Gu, K.; Zhang, Y.; Qiao, J. Ensemble Meta-Learning for Few-Shot Soot Density Recognition. *IEEE Trans. Ind. Inform.* **2020**, *17*, 2261–2270. [CrossRef]
8. Ma, X.; Shahbakhti, M.; Chigan, C. Connected Vehicle Based Distributed Meta-Learning for Online Adaptive Engine/Powertrain Fuel Consumption Modeling. *IEEE Trans. Veh. Technol.* **2020**, *69*, 9553–9565. [CrossRef]
9. Qin, X.; Zhang, Z.; Huang, C.; Dehghan, M.; Zaiane, O.R.; Jagersand, M. U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognit.* **2020**, *106*, 107404. [CrossRef]
10. Zeng, Y.; Lin, Z.; Lu, H.; Patel, V.M. Cr-fill: Generative image inpainting with auxiliary contextual reconstruction. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021.
11. Norouzzadeh, M.S.; Nguyen, A.; Kosmala, M.; Swanson, A.; Palmer, M.S.; Packer, C.; Clune, J. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E5716–E5725. [CrossRef]
12. Crouse, D.; Jacobs, R.L.; Richardson, Z.; Klum, S.; Jain, A.; Baden, A.L.; Tecot, S.R. LemurFaceID: A face recognition system to facilitate individual identification of lemurs. *BMC Zool.* **2017**, *2*, 562. [CrossRef]
13. Witham, C.L. Automated face recognition of rhesus macaques. *J. Neurosci. Methods* **2017**, *300*, 157–165. [CrossRef] [PubMed]
14. Deb, D.; Wiper, S.; Gong, S.; Shi, Y.; Tymoszek, C.; Fletcher, A.; Jain, A.K. Face recognition: Primates in the wild. In Proceedings of the 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS), Long Beach, CA, USA, 22–25 October 2018.
15. Weinstein, B.G. A computer vision for animal ecology. *J. Anim. Ecol.* **2017**, *87*, 533–545. [CrossRef] [PubMed]
16. Koniar, D.; Hargaš, L.; Loncová, Z.; Duchoň, F.; Beňo, P. Machine vision application in animal trajectory tracking. *Comput. Methods Programs Biomed.* **2015**, *127*, 258–272. [CrossRef]
17. Yudin, D.; Sotnikov, A.; Krishtopik, A. Detection of Big Animals on Images with Road Scenes using Deep Learning. In Proceedings of the 2019 International Conference on Artificial Intelligence: Applications and Innovations (IC-AIAI), Vrdnik, Banja, Serbia, 30 September–4 October 2019.

18. Koochaki, F.; Shamsi, F.; Najafizadeh, L. Detecting mtbi by learning spatio-temporal characteristics of widefield calcium imaging data using deep learning. In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montreal, QC, Canada, 20–24 July 2020.
19. Schofield, D.; Nagrani, A.; Zisserman, A.; Hayashi, M.; Matsuzawa, T.; Biro, D.; Carvalho, S. Chimpanzee face recognition from videos in the wild using deep learning. *Sci. Adv.* **2019**, *5*, eaaw0736. [CrossRef]
20. Kuncheva, L. Animal reidentification using restricted set classification. *Ecol. Inform.* **2021**, *62*, 101225. [CrossRef]
21. Lai, N.; Kan, M.; Han, C.; Song, X.; Shan, S. Learning to Learn Adaptive Classifier–Predictor for Few-Shot Learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 3458–3470. [CrossRef]
22. Munkhdalai, T.; Yu, H. Meta networks. In Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, NSW, Australia, 6–11 August 2017.
23. Wong, A.; Yuille, A.L. One shot learning via compositions of meaningful patches. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Boston, MA, USA, 7–13 December 2015.
24. Hariharan, B.; Girshick, R. Low-shot visual recognition by shrinking and hallucinating features. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
25. Chen, Z.; Fu, Y.; Wang, Y.X.; Ma, L.; Liu, W.; Hebert, M. Image deformation meta-networks for one-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
26. Xu, Z.; Zhu, L.; Yang, Y. Few-shot object recognition from machine-labeled web images. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
27. Ramalho, T.; Garnelo, M. Adaptive posterior learning: Few-shot learning with a surprise-based memory module. *arXiv* **2019**, arXiv:1902.02527.
28. Kaiser, Ł.; Nachum, O.; Roy, A.; Bengio, S. Learning to remember rare events. *arXiv* **2017**, arXiv:1703.03129.
29. Wang, X.; Yu, F.; Wang, R.; Darrell, T.; Gonzalez, J.E. Tafe-net: Task-aware feature embeddings for low shot learning. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), Long Beach, CA, USA, 16–20 June 2019.
30. Wang, R.-Q.; Zhang, X.-Y.; Liu, C.-L. Meta-Prototypical Learning for Domain-Agnostic Few-Shot Recognition. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, 1–7. [CrossRef]
31. Li, H.; Dong, W.; Mei, X.; Ma, C.; Huang, F.; Hu, B.G. LGM-Net: Learning to generate matching networks for few-shot learning. In Proceedings of the 36th International Conference on Machine Learning (ICML), Long Beach, CA, USA, 9–15 June 2019.
32. Liu, Q.; Zhang, X.; Liu, Y.; Huo, K.; Jiang, W.; Li, X. Multi-Polarization Fusion Few-Shot HRRP Target Recognition Based on Meta-Learning Framework. *IEEE Sensors J.* **2021**, *21*, 18085–18100. [CrossRef]
33. Rahman, S.; Khan, S.; Porikli, F. A Unified Approach for Conventional Zero-Shot, Generalized Zero-Shot, and Few-Shot Learning. *IEEE Trans. Image Process.* **2018**, *27*, 5652–5667. [CrossRef]
34. Kang, B.; Liu, Z.; Wang, X.; Yu, F.; Feng, J.; Darrell, T. Few-shot object detection via feature reweighting. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–4 November 2019.
35. Wang, C.; Liu, Z.; Chan, S.-C. Superpixel-Based Hand Gesture Recognition with Kinect Depth Camera. *IEEE Trans. Multimed.* **2015**, *17*, 29–39. [CrossRef]
36. Zhang, H.; Zhang, J.; Koniusz, P. Few-shot learning via saliency-guided hallucination of samples. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), Long Beach, CA, USA, 16–20 June 2019.
37. Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P.H.; Hospedales, T.M. Learning to compare: Relation network for few-shot learning. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
38. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
39. Wang, Y.; Yao, Q.; Kwok, J.T.; Ni, L.M. Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv. (Csur)* **2020**, *53*, 1–34. [CrossRef]
40. Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; et al. The open images dataset v4. *Int. J. Comput. Vision* **2020**, *128*, 1956–1981. [CrossRef]
41. Xian, Y.; Lampert, C.H.; Schiele, B.; Akata, Z. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 2251–2265. [CrossRef]
42. Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2009**, *88*, 303–338. [CrossRef]
43. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
44. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2019.
45. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
46. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.

47. Yan, X.; Chen, Z.; Xu, A.; Wang, X.; Liang, X.; Lin, L. Meta r-cnn: Towards general solver for instance-level low-shot learning. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–4 November 2019.
48. Li, X.; Cheng, G.; Wang, L.; Wang, J.; Ran, Y.; Che, T.; Li, G.; He, H.; Zhang, Q.; Jiang, X.; et al. Boosting geoscience data sharing in China. *Nat. Geosci.* **2021**, *14*, 541–542. [CrossRef]