Article

DJAN: Deep Joint Adaptation Network for Wildlife Image Recognition

Changchun Zhang [1,2,3] and Junguo Zhang [1,2,3,*]

1 School of Technology, Beijing Forestry University, Beijing 100083, China; zhangchangchun@bjfu.edu.cn
2 State Key Laboratory of Efficient Production of Forest Resources, Beijing 100083, China
3 Key Laboratory of National Forestry and Grassland Administration on Forestry Equipment and Automation, Beijing 100083, China
* Correspondence: zhangjunguo@bjfu.edu.cn

Simple Summary: Identifying wildlife species is crucial in various wildlife monitoring tasks. In this paper, a wildlife image recognition approach is implemented based on deep learning with a joint adaptation network. This paper presents a joint adversarial learning approach and a cross-domain local and global representation learning approach. Utilizing the two approaches, a Deep Joint Adaptation Network model for wildlife image recognition is designed. The proposed model can yield high accuracy in wildlife image recognition and is beneficial to improve the generalization ability in complex environments. Our research is of the utmost importance for wildlife recognition and wildlife biodiversity monitoring.

Abstract: Wildlife recognition is of utmost importance for monitoring and preserving biodiversity. In recent years, deep-learning-based methods for wildlife image recognition have exhibited remarkable performance on specific datasets and are becoming a mainstream research direction. However, wildlife image recognition tasks face the challenge of weak generalization in open environments. In this paper, a Deep Joint Adaptation Network (DJAN) for wildlife image recognition is proposed to deal with the above issue by taking a transfer learning paradigm into consideration. To alleviate the distribution discrepancy between the known dataset and the target task dataset while enhancing the transferability of the model's generated features, we introduce a correlation alignment constraint and a strategy of conditional adversarial training, which enhance the capability of individual domain adaptation modules. In addition, a transformer unit is utilized to capture the long-range relationships between the local and global feature representations, which facilitates better understanding of the overall structure and relationships within the image. The proposed approach is evaluated on a wildlife dataset; a series of experimental results testify that the DJAN model yields state-of-the-art results, and, compared to the best results obtained by the baseline methods, the average accuracy of identifying the eleven wildlife species improves by 3.6 percentage points.

Keywords: transfer learning; wildlife recognition; distribution discrepancy; domain adaptation; deep learning

1. Introduction

Wildlife identification is a crucial component of monitoring and conserving biodiversity in the wild. Wildlife monitoring serves as the foundation for wildlife conservation management [1] and plays an importance role in this regard. Through collecting and analyzing data on various aspects of wildlife, such as habitat utilization, migration patterns, quantity distribution, and behavioral activities, monitoring provides valuable information [2] and insights for conservation organizations and government agencies. These data unveil key issues related to species population status, habitat quality, and habitat connectivity, which, in turn, contribute to the development of effective conservation strategies and management

plans. Thus, the identification of wildlife species is essential for species diversity detection and the conservation of rare and endangered wildlife. Currently, camera traps are the mainstream method for wildlife species monitoring [3–5]. However, manual identification of monitoring images is known to suffer from the challenges of high intensity and low efficiency. With the rapid development of artificial intelligence technology, automatic wildlife identification methods based on deep networks have demonstrated excellent performance on specific datasets [6–9]. A deep convolutional neural network for automatic identification of wildlife was proposed [10] and achieved a Top-1 accuracy of 88.9% and a Top-5 accuracy of 98.1% on the Serengeti dataset, a public wildlife dataset. Trnovszky et al. [11] achieved a recognition accuracy of 98.0% using an improved LeNet model on a dataset containing five wildlife species. Verma et al. [12] addressed the interference of cluttered scene images, which do not contain individual animals, in wildlife recognition in monitoring datasets. They utilized deep convolutional neural networks (DCNNs) to extract features from cluttered scene images and achieved cluttered scene image recognition based on VGGNet and ResNet, further improving the recognition accuracy of high-value wildlife monitoring images. Schneider et al. [13] focused on the problem of model generalization in unknown scenes and compared the performance of different deep learning methods on various data scenarios. Vargas-Felipe et al. [14] used convolutional neural networks to recognize wildlife in monitoring images, and their recognition accuracy was significantly better than that of traditional recognition methods. Schindler [15] combined instance segmentation networks with action recognition networks to detect wildlife and perform simultaneous action recognition, providing more-diversified ecological analysis data for wildlife monitoring. However, in the real world, as depicted in Figure 1, factors such as different backgrounds, varying lighting conditions, and diverse shooting scales can lead to changes in feature distributions within the same class of images [16–19]. These factors can result in suboptimal performance of existing deep learning algorithms, posing significant challenges to image recognition.
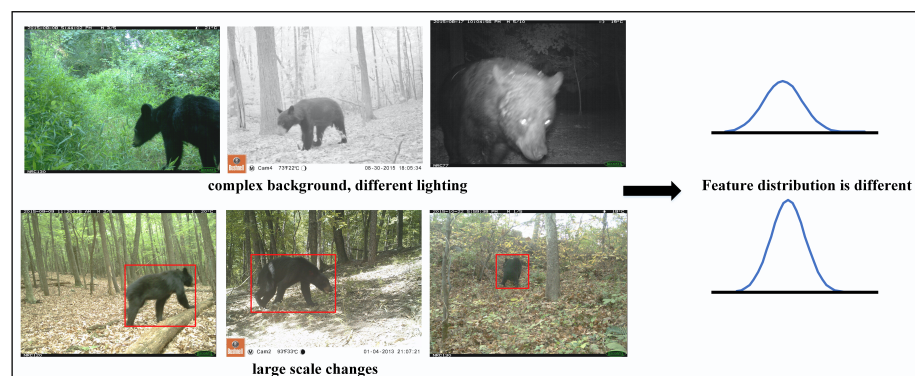

Figure 1. The characteristics of wildlife images.

With the successful application of deep domain adaptation in various fields such as pattern recognition and computer vision, the learning paradigm based on deep transfer learning updates or 'transfers' models from one domain to another, breaking the limitations of traditional deep learning models that require a large amount of labeled data as a prerequisite and also overcoming the strict requirement for data to follow the same distribution. Therefore, domain-adaptation-based wildlife image recognition has become a current research hotspot and has yielded good results. Norouzzadeh et al. [20] applied transfer learning to pre-trained convolutional neural network models such as AlexNet, NiN, VGGNet, and GoogLeNet, and compared the performance of the two models in automatically detecting animal species in CT image datasets. A transfer-learning-based wildlife image classification method [21] is proposed that is based on the Xception network, achieving an average accuracy rate of 99.01%, which is a 57.82% improvement in accuracy compared to standard convolutional neural network methods. Thangaraj et al. [22] utilized

the concept of transfer learning fine-tuning and fine-tuned mainstream network models such as DenseNet169 and Xception for animal individual recognition. Compared with other models, the InceptionResNetV2 model achieved an accuracy rate of 94.82%. By employing deep domain adaptation approaches, not only are the limitations of traditional deep learning methods requiring a large amount of labeled samples and samples following the same distribution overcome, but also data-driven optimization of recognition models with strong generalization capability can be achieved. However, the aforementioned models are still limited to optimizing DCNN models for wildlife recognition through transfer learning fine-tuning approaches, and they cannot be widely generalized and applied. The diversity of wildlife species, the similarities between classes, and the differences within classes increase the difficulty of wildlife recognition, requiring the establishment of models with strong feature extraction and generalization capabilities for wildlife recognition.

The above-mentioned transfer learning methods only utilize the idea of fine-tuning, which can partially leverage knowledge from the source domain to assist the learning of the target domain and mitigate domain differences. However, in order to further alleviate domain discrepancies and enhance the model's generalization performance, some researchers have proposed a series of deep transfer learning methods. Sun et al. [23] proposed the Correlation Alignment (CORAL) method, which utilizes a linear transformation to align the second-order statistical information between the source domain sample distribution and the target domain sample distribution, aiming to minimize domain differences. However, CORAL relies on linear transformations and cannot be trained in an end-to-end style. To address such a limitation, Sun et al. [24] further extended the CORAL algorithm and introduced the Deep Correlation Alignment (DCORAL) algorithm. DCORAL embeds CORAL directly into a deep network, constructing a differentiable loss function to minimize cross-domain correlation differences. With the advent of Generative Adversarial Networks (GANs) [25], adversarial domain adaptation methods [26–30] have emerged, which aim to generate domain-transferable features through the adversarial game between a feature generator and a domain classifier. Inspired by the idea of GANs, a domain-adversarial neural network [31] is proposed, which involves three modules: a shared feature extractor cross-domain, a label classifier, and a domain classifier. The feature extractor and label classifier aim to minimize classification errors in the source domain, ensuring that the learned features are discriminative. At the same time, the domain classifier maximizes the domain classification error, encouraging domain-invariant feature distributions. To balance the competition between the feature extractor and label classifier with the domain classifier during training, a gradient reversal layer (GRL) is introduced. GRL works by reversing the gradients of the domain classifier's loss so that while the domain classifier aims to minimize its loss, the feature extractor maximizes its loss. This is achieved by flipping the sign of the gradients and propagating them back to the feature extractor. The gradient reversal layer enables the network to ensure domain confusion between the domains while preserving the discriminative power of the learned features.

Based on DANN, Long et al. [32] presented a joint domain adaptation network model that outperforms domain-adversarial network models on multiple image classification databases. The JDAN model considers how to match the joint activation distributions of multiple layers in the source and target domains by constraining the Joint Maximum Mean Discrepancy (JMMD). It uses a multi-layer neural network to parameterize JMMD and employs adversarial learning methods to learn discriminative features. Chen et al. [33] discovered that although adversarial domain adaptation methods enhance the transferability of features, they can decrease the discriminability of feature classes. To address this, they introduced batch spectral penalization during training to ensure that the differences between feature values are not too large, thereby preserving feature discriminability. The shared network feature extraction mechanism prevents the generation of domain-specific information for each domain. To overcome this, Tzeng et al. [34] used a weight non-sharing strategy to independently generate features for each domain. Unlike DANN [31], this method allows the feature extractor to generate more domain-specific features due to the

non-shared parameters. Volpi et al. [35] utilized a feature generator for data augmentation within the source domain feature space. They employed a domain classifier to distinguish between the generated and authentic features, ultimately aligning the distribution of the augmented data with that of the target domain.

The aforementioned methods that use cross-domain feature alignment based on marginal probability distributions overlook the effect of conditional probability distributions in improving network transfer gain. Therefore, researchers have introduced class information to align the conditional probability distributions cross-domain. A multi-adversarial domain adaptation method is proposed [36] to capture the multi-modal structure of data. For a K-class classification problem, this method introduces K domain discriminators, where each domain discriminator matches the samples from the source and target domains with the same label. Specifically, it performs soft classification of the data based on probabilities generated by the class classifier. When the label space of the target domain is a subset of the label space of the source domain, i.e., in the case of partial domain adaptation, it needs to select a subset of source domain samples from the shared label space. There are also some methods that, although they do not employ multiple domain discriminators, achieve similar effects by introducing class information and are therefore classified as multi-adversarial methods. A conditional domain adversarial networks is presented [37], which combines instance weights with adversarial feature learning to capture the multi-modal structure hidden under complex data distributions, leading to effective knowledge transfer. However, during the process of translating technology into practical applications, existing methods have two limitations. Firstly, they rely on single-domain adaptation and class information, neglecting the relationship between adversarial domain adaptation and domain alignment. Secondly, they fail to capture the long-term relationship between local and global feature representations, which would help better understand the overall structure and relationships within images. To address these challenges, a Deep Joint Adaptation Network (DJAN) is proposed in this paper, as shown in Figure 2. Specifically, it incorporates class information into the domain adversarial network to explore more fine-grained transferable features. Additionally, it further exploits more transferable features across domains via considering correlation alignment. It is worth noting that the domain-transferable features learned in this process exhibit strong transferability for different images and different transfer tasks, demonstrating the generalization capability of this method. It effectively improves the accuracy of wildlife image recognition while enhancing the usability and scalability of the approach. In the learning process, a transformer unit is proposed to capture the long-term relationship between local and global feature representations, achieving optimal transfer effects. In summary, this paper presents three main contributions:

- A correlation alignment constraint and the strategy of conditional adversarial training are proposed to enhance the capability of individual domain adaptation modules.
- Combining the correlation alignment constraint and the strategy of conditional adversarial training, a transformer unit is proposed to capture the long-range relationships between the local and global feature representations, which facilitates better understanding of the overall structure and relationships within the image.
- A series of experimental results prove that our approach yields state-of-the-art results, and, compared to the best results achieved by the baseline methods, the average accuracy of identifying the eleven wildlife species improves by 4.7%.
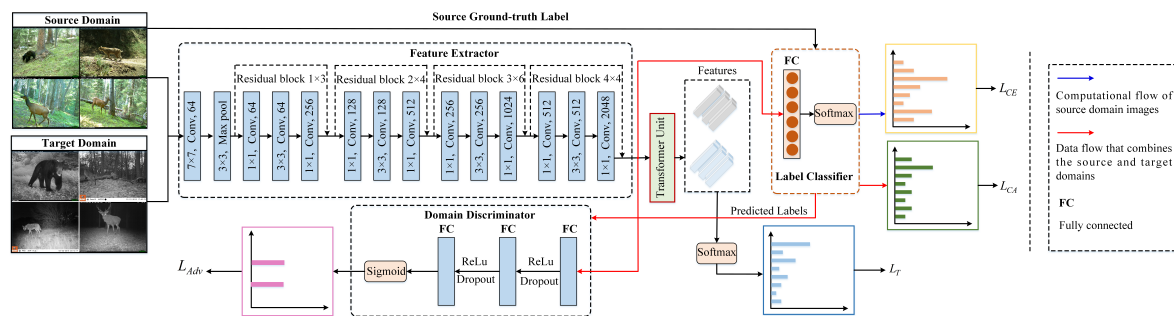
Figure 2. Structure diagram of Deep Joint Adaptation Network. Note: L_{CE} denotes the loss function of optimizing the source domain data; L_{CA} is correlation alignment loss, which is defined as the measurement between the second-order statistics information of label classifier prediction of the training and test images; L_{Adv} indicates the objective optimization function for the final domain discriminator D and feature extractor G in adversarial domain alignment; L_T represents a transformer loss.

2. Materials and Methods

2.1. Dataset

In this paper, the wildlife dataset utilized was derived from two publicly available wildlife datasets, ENA24 [38] and NACTI [39]. These datasets were used to construct two distinct datasets, which are respectively abbreviated as **ES** and **NS** with different distributions but consistent class spaces. They contain 25,591 images of 11 different wildlife species (as shown in Figure 3 and Table 1). These images were captured utilizing infrared camera traps placed in different locations and at different times, resulting in variations in backgrounds, animal poses, and other factors. These distribution differences mainly arise from variations in lighting conditions, camera angles, backgrounds, vegetation, and colors. In this study, each image in the wildlife dataset was resized to a unified size of 224×224 before training the model. The dataset was then divided into source and target domains. Based on the optimization criteria of transfer learning models, two transfer learning wildlife recognition tasks were constructed: **ES→NS** and **NS→ES**.

Table 1. Statistics of the wildlife image dataset.

	bear	bobcat	coyote	deer	fox	raccoon
ES	2870	455	1814	4336	1866	1376
NS	867	328	181	2357	427	142
	squirrel	striped skunk	virginia opossum	wild boar	wild turkey	
ES	1386	1323	110	1893	815	
NS	927	295	715	819	289	

2.2. Joint Adversarial Learning

In this section, a joint adversarial learning method combining conditional adaptation learning and correlation domain alignment is proposed to promote fine-grained and highly transferable feature generation.

Figure 3. Images of 11 wild animals. Note: 1. bear, 2. bobcat, 3. coyote, 4. deer, 5. fox, 6. raccoon, 7. squirrel, 8. striped skunk, 9. Virginia opossum, 10. wild boar, and 11. wild turkey.

2.2.1. Conditional Adaptation Learning

Deep learning has enjoyed tremendous success in the field of image processing, and one important reason is its ability to confuse domain discriminators and achieve domain alignment by learning new features across domains. Therefore, we first focus on how to learn more transferable features across domains. The Domain-Adversarial Neural Networks (DANN) method [31] can generate domain-transferable feature representations via a minimax game strategy. Specifically, the feature extractor G and the label classifier C minimize the classification error L_{CE} in the source domain, ensuring that the features generated by the deep network are discriminative. At the same time, the feature extractor G maximizes the domain adversarial classification error L_{CE}, making the feature distribution domain-transferable. This mechanism establishes a competitive relationship between the feature extractor G, the label classifier C, and the domain discriminator D. During the backpropagation optimization process, the gradient from the domain discriminator D to the feature extractor G needs to be multiplied by a negative constant. The aforementioned domain-adversarial network provides domain-transferable feature information, which is beneficial for training an effective domain adaptation network model and achieving better unsupervised domain adaptation for image classification. However, dataset bias [40] persists in domain-specific feature and classifier layers, and adversarial adaptation on specific layers alone is insufficient to alleviate dataset bias. Additionally, when domain samples contain complex multimodal structures, domain-adversarial neural network models may not effectively capture the multimodal structure [41] of the samples for fine-grained cross-domain feature distribution alignment. This is known as the mode collapse problem in Generative Adversarial Networks. Incorporating category information predicted

by classifiers into adversarial domain adaptation helps address the challenges faced by domain-adversarial neural network models.

For this purpose, a reverse focal loss is applied to the domain discriminator D to focus on easily discriminable samples, i.e., samples that are difficult to match. As a result, the objective optimization function for the final domain discriminator D and feature extractor G in adversarial domain alignment is described by Equation (1).

$$L_{Adv} = -\frac{1}{n_s} \sum_{i=1}^{n_s} \exp(D(\phi(\mathbf{h}_i^s))) \log(D(\phi(\mathbf{h}_i^s))) - \frac{1}{n_t} \sum_{j=1}^{n_t} \exp\left(1 - D\left(\phi\left(\mathbf{h}_j^t\right)\right)\right) \log\left(1 - D\left(\phi\left(\mathbf{h}_j^t\right)\right)\right) \tag{1}$$

where $\exp(D(\cdot))$ represents the reverse focal weights for each sample in the source domain, and $\exp(1 - D(\cdot))$ denotes the reverse focal weights for each sample in the target domain. The connecting variable $\mathbf{h} = (\mathbf{f}, \mathbf{g})$ is set, where $\mathbf{g}$ represents the results of the label classifier, and $\mathbf{f}$ represents the features generated by the feature extractor.

2.2.2. Correlation Domain Alignment

The deep domain transfer model obtained through domain-adversarial networks lacks alignment based on the relevant information of two-domain features. It only captures the second-order statistical information of the unexplored samples and includes irrelevant background information in the learned features. The selected cross-domain feature alignment is irreplaceable due to its second-order statistical characteristics, and it contributes significantly to the final prediction. Therefore, in this paper, we introduce a correlation alignment constraint mechanism that aligns the classifier output features of two-domain samples based on second-order statistical information to promote the accuracy of wildlife recognition. Specifically, the calculation formula for the correlation alignment of features output by the label classifiers for the cross-source and target domain is computed by Equation (2).

$$L_{CA} = \frac{1}{4d^2} \|C_S - C_T\|_F^2 \tag{2}$$

where d indicates the dimension of the features, $\| \cdot \|_F^2$ denotes the Frobenius norm of a matrix, and C_S and C_T indicate the correlation matrices of the source and target domain samples, respectively. Their definitions are as follows:

$$C_S = \frac{1}{N_S - 1}\left(B_S^{\mathrm{T}} B_S - \frac{1}{N_S}\left(1^{\mathrm{T}} B_S\right)^{\mathrm{T}}\left(1^{\mathrm{T}} B_S\right)\right) \tag{3}$$

$$C_T = \frac{1}{N_T - 1}\left(B_T^{\mathrm{T}} B_T - \frac{1}{N_T}\left(1^{\mathrm{T}} B_T\right)^{\mathrm{T}}\left(1^{\mathrm{T}} B_T\right)\right) \tag{4}$$

where N_S and N_T denote, respectively, the batch sizes of the source and target domain samples. B_S and B_T indicate the feature representations output by the label classifier C, and 1^{T} is a row vector consisting of all ones. The features B_S and B_T can be gradient computed respectively using the chain rule based on Equations (5) and (6).

$$\frac{\partial L_{CA}}{\partial B_S^{ij}} = \frac{1}{d^2(N_S - 1)}\left(\left(B_S^{\mathrm{T}} - \frac{1}{N_S}\left(1^{\mathrm{T}} B_S\right)^{\mathrm{T}} 1^{\mathrm{T}}\right)^{\mathrm{T}}(C_S - C_T)\right)^{ij} \tag{5}$$

$$\frac{\partial L_{CA}}{\partial B_T^{ij}} = \frac{1}{d^2(N_T - 1)}\left(\left(B_T^{\mathrm{T}} - \frac{1}{N_T}\left(1^{\mathrm{T}} B_T\right)^{\mathrm{T}} 1^{\mathrm{T}}\right)^{\mathrm{T}}(C_S - C_T)\right)^{ij} \tag{6}$$

where B_S^{ij} represents the j-th dimension of the i-th source sample, and B_T^{ij} indicates the j-th dimension of the i-th target sample.

2.3. Cross-Domain Local and Global Representation Learning

By utilizing joint adversarial learning as described above, the model can extract the transferable features to reduce the domain discrepancy cross-domain. However, in the hard sample, the decision boundary of the model fails to distinguish them and may mistake these samples. In order to mitigate this issue, cross-domain local and global representation learning is proposed to build the relationship between samples and boost the discriminant ability of the decision boundary. Motivated by the prosperity of transformers in natural language processing and image processing, which aims to build the relationships of the long-range sequence and improve the transferability of features, a transformer transfer loss is designed to construct the relationship of the local and global representations via the transformer unit, as shown in Figure 4.
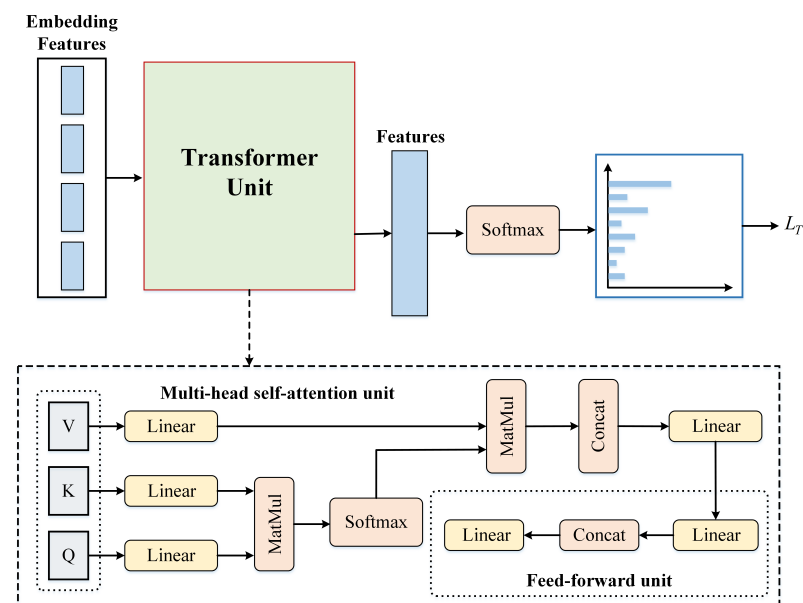

Figure 4. Detailed structure demonstration of the transformer loss.

Specifically, for the local feature X_L and global feature X_G generated by the feature extractor G, we utilize the transformer unit to construct a transformer loss, establishing a relationship with the long-range sequence to enhance the transferability of the features and to facilitate better understanding of the overall structure and relationships within the image. The global feature X_G and the local feature X_L in each wildlife image can be identified as different patches, which can be termed as token embeddings. All these token embeddings are fed into the transformer unit: the specific operations performed by these embedding features in the transformer unit can be found in Figure 4's dashed-box section. Then, we can obtain features $\mathbf{Z}$ generated based on the local feature X_L and global feature X_G through the transformer unit. Finally, a transformer loss is defined as follows:

$$L_T = -\sum_{i=1}^{\mathbf{C}} y_i(\log(\text{Softmax}((\mathbf{Z})))) \tag{7}$$

where $\mathbf{C}$ represents the number of the class in the wildlife dataset, and y_i is the ground-truth label regrading the ith class.

2.4. Optimization of Deep Joint Adaptation Network

We follow similar steps as [42] to train the source-domain discriminative model $C(G(\mathbf{x}))$ for the wildlife image recognition task via optimizing the parameters of the

feature extractor G and the label classifier C. Therefore, the objective function of optimizing the source domain data is:

$$L_{CE} = \min_{G,C} \frac{1}{n_s} \sum_{i=1}^{n_s} L_C(C(G(x_i)), y_i) \tag{8}$$

where $L_C(\cdot)$ is a cross-entropy loss function.

In conclusion, by performing joint adversarial learning and cross-domain local and global representation learning in Equations (1), (2), and (7), we can effectively improve the transferability of features and boost the generalization ability so as to learn the benefit decision boundary. Therefore, joining adversarial learning and cross-domain local and global representation learning—that is, the Deep Joint Adaptation Network—the total loss function can be formulated as follows:

$$L = L_{CE} + \alpha * L_{Adv} + \beta * L_{CA} + \gamma * L_T \tag{9}$$

where α, β, and γ denote the hyperparameters; these hyperparameters are used to control the degree of influence between different constraints.

3. Results

3.1. Implementation Details

This section introduces the details of the experiments: (1) We used a 50-layer ResNet [43] as the base network model and conducted experiments on the wildlife dataset. (2) All experiments in this chapter were conducted using the PyTorch framework, and the base network models were pre-trained on the ImageNet dataset. (3) We used mini-batch stochastic gradient descent with a momentum of 0.9. The learning rate strategy described in [42] was employed, where the learning rate $\eta_p = \frac{\eta_0}{(1+ap)^b}$, and p linearly varies from 0 to 1 during the training phase. The variable η_0 is the initial learning rate, which starts at zero for the task-specific fully connected layers, and the learning rate for the convolutional layers is 10 times higher than that of the fully connected layers (i.e., the learning rate for the convolutional layers is 0.001, and for the fully connected layers, it is 0.01). The progressive strategy for the domain discriminator is as described in [37], which is $\frac{2}{1+\exp(-\varepsilon p)} - 1$ multiplied by α, with α linearly increasing from 0 to 1 and ε set to 10. (4) During the training and testing process, the batch size is set to 128. Additionally, the software and hardware employed in all experiments is reported in Table 2.

Table 2. The hardware and software configuration of the experiment.

Test Environment	Type
Operating System	Ubuntu 16.04
Framework	PyTorch 1.4.0
CPU	Intel(R) Core(TM) i9-10900K CPU @ 3.7G Hz
GPU	GeForce RTX 3090Ti (24 G/Nvidia)
RAM	32 G
Programming Language	Python 3.6

3.2. Evaluation Metrics

To evaluate the effectiveness of our DJAN model for wildlife image recognition, **Accuracy** is used as the evaluation metric and is calculated by Equation (10).

$$\textbf{Accuracy} = \frac{|x^t : x^t \in X^t \wedge \hat{y}^t = y^t|}{|x^t : x^t \in X^t|} \tag{10}$$

where y^t represents the true class label of the sample x^t, which is unknown during the learning phase, and $\hat{y}^t$ represents the predicted class label made by the label classifier for

the sample x^t in the target domain. Furthermore, this paper also utilizes three evaluation metrics, namely **Precision**, **Recall**, and the **F1** score, to assess the effectiveness of our method. These evaluation metrics are defined respectively by Equations (11)–(13).

$$\textbf{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{11}$$

$$\textbf{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{12}$$

$$\textbf{F1} = \frac{2 * \textbf{Precision} * \textbf{Recall}}{\textbf{Precision} + \textbf{Recall}} \tag{13}$$

where TP (True Positive) indicates the number of correct identifications of a certain wildlife species, FP (False Positive) represents the number of incorrect identifications of a certain wildlife species, and FN (False Negative) is the number of incorrect identifications of a certain wildlife species as another category of wildlife. The F1 score denotes the harmonic mean based on precision and recall.

3.3. Comparison with State-of-the-Art Models

In order to reveal the effectiveness of the DJAN approach, we compare DJAN with the state-of-the-art models on the wildlife dataset; these baselines include ResNet50 [43], MMD [44], DANN [31], DCROAL [24], CDAN [37], DSAN [45], BNM [46], HAN [47], and JTN [42]. Table 3 reports the evaluation results on the wildlife dataset with two transfer tasks (i.e., **ES→NS** and **NS→ES**). For fair comparison, the results of the comparative methods are obtained by the authors through experimentation using the source code.

Table 3. Accuracy performance comparison of the DJAN method and the baselines on wildlife dataset.

Method	ES→NS	NS→ES	Avg.
ResNet50	41.4	55.4	48.4
MMD	44.0	59.2	51.6
DANN	42.8	51.6	47.2
DCORAL	44.1	55.1	49.6
CDAN	45.5	56.2	50.9
DSAN	41.5	53.7	47.6
BNM	28.0	60.1	44.0
HAN	47.2	62.0	54.6
JTN	44.4	62.5	53.5
DJAN	48.8	67.5	58.2

As can be observed from Table 3, the DJAN model overpasses all baselines in both the **ES→NS** and **NS→ES** transfer tasks. This is because DJAN can boost the generalization ability of the network and generate more transferable feature representations via constructing joint adversarial learning and cross-domain local and global representation learning for the wildlife image recognition task. In addition, among the comparison methods, the HAN method [47] achieved the highest transfer task accuracy in wild animal image recognition. It utilizes a conditional domain adaptation and correlation alignment constraint. Compared to the HAN method, our DJAN model achieved a 3.6% (58.2% vs. 54.6%) improvement in average recognition accuracy across two transfer tasks (i.e., **ES→NS** and **NS→ES**). The CDAN method [37] is a further extension of the DANN method [31] that learns cross-domain transferable features by using conditional adversarial domain adaptation. Despite that, our DJAN method still achieves a 7.3% (58.2% vs. 50.9%) higher accuracy than it. Overall, the proposed DJAN model achieved an average accuracy of 58.2% in recognizing 11 categories of wild animals, surpassing all the baseline models. This reflects that our model effectively improves the generalization capacity via employing joint adversarial

learning and cross-domain local and global representation learning for wild animal image recognition tasks.

3.4. Analysis of Prediction Results from Recognition Models

This section demonstrates that our DJAN model can correctly classify some samples, while the HAN method suffers from misclassification, as shown in Figure 5. The main reason for these misclassifications is that some images were only captured from the back or side of the target animal or the target was in a dark environment where its species-specific image features were not displayed. Therefore, it is difficult for the model to correctly identify them. This once again confirms the effectiveness of our proposed method for improving the transfer performance of the model and consequently enhancing its predictive accuracy by employing conditional adaptation learning, correlation domain alignment, and cross-domain local and global representation learning for wildlife image recognition.
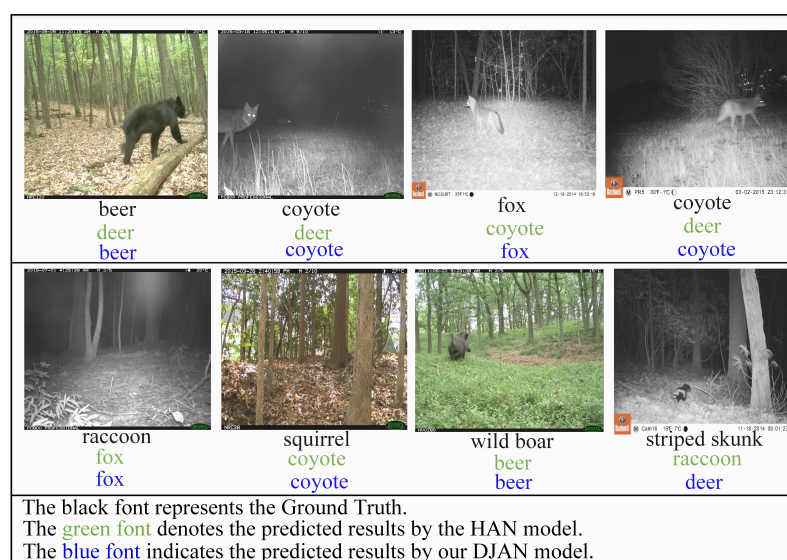

The black font represents the Ground Truth.
The green font denotes the predicted results by the HAN model.
The blue font indicates the predicted results by our DJAN model.

Figure 5. The prediction result comparison of our DJAN and HAN methods on transfer task **ES→NS** for wild animal images.

3.5. Analysis of Model Performance Using Different Evaluation Criteria

To comprehensively evaluate the performance of our method in wildlife image recognition, we conducted experimental tests using four evaluation metrics: **Accuracy**, **Precision**, **Recall**, and **F1** score, on the transfer task **ES→NS**. The experimental results are shown in Table 4. The trends are consistent with the results in Table 3. Our method achieved the highest wildlife image recognition accuracy, precision, recall, and F1 score, with an average accuracy of 48.8%, precision of 0.43, recall of 0.40, and F1 score of 0.35. Comparing to the best results from the comparative methods, our method exhibited improvements of 4.4%, 0.08, 0.05, and 0.03, respectively, further validating the effectiveness of our DJAN approach.

Table 4. Performance analysis of wildlife image recognition models under different evaluation criteria.

Methods	Accuracy	Precision	Recall	F1
DCORAL	44.1	0.38	0.29	0.24
JTN	44.4	0.38	0.37	0.32
DJAN	48.8	0.46	0.42	0.35

3.6. Ablation Study

We conducted ablation experiments to verify the effectiveness of each constraint in our approach on the wildlife dataset, and the results are reported in Table 5. We

observed the following: (1) "ResNet50" indicates the results yielded by directly fine-tuning ResNet50 [43] on the wildlife dataset, while "DJAN w/o Adv" denotes the results obtained solely using the deep discriminative feature learning network. Even when only utilizing the correlation alignment adaptation learning and transformer transfer loss, our approach achieves better domain adaptation wildlife image recognition accuracy compared to the ResNet50 method. This suggests that correlation domain alignment and cross-domain local and global representation learning contribute to learning cross-domain transferable features in our approach and achieves higher transfer gains. (2) "DJAN w/o CA" presents the results obtained by removing the correlation domain alignment constraint. Our approach achieves an improvement of nine percentage points (58.2% vs. 49.2%) compared to "DJAN w/o CA". This improvement is attributed to the ability of the "DJAN w/o CA" to enhance feature transferability. (3) "DJAN w/o T" indicates the results obtained via solely utilizing conditional adaptation learning and the correlation domain alignment constraint. On **ES→NS** and **NS→ES** transfer tasks, our approach outperforms the baselines in achieving the best wildlife image recognition results. (4) Compared to our DJAN approach, the "DJAN method w/o T" model achieves an accuracy below 7.9% (50.3% vs. 58.2%). This further quantitatively validates the effectiveness of cross-domain local and global representation learning, enhancing the accuracy of domain adaptation wildlife image recognition.

Table 5. Results of ablation studies on wildlife dataset.

Methods	ES→NS	NS→ES	Avg.
ResNet50	41.4	55.4	48.4
DJAN w/o Adv	46.2	53.7	50.0
DJAN w/o CA	40.5	57.9	49.2
DJAN w/o T	43.3	57.3	50.3
DJAN	48.8	67.5	58.2

3.7. Deep Joint Adaptation Network Convergence

To investigate the convergence performance of the proposed DJAN model compared to the comparative methods, we conducted the test accuracy experiments on the transfer task **ES→NS**, as shown in Figure 6. By Figure 6, we can observe that the recognition accuracy of our DJAN method gradually increases as the number of iterations increases and then exhibits a stable trend. Moreover, we can also see that although the JTN method achieves stable convergence performance similar to our DJAN model, the convergence performance of our DJAN model is significantly superior to that of the JTN method throughout the entire convergence process. Therefore, the substantial performance improvement for the wildlife image recognition task across domains further reflects the advantages of our DJAN.

3.8. Parameter Sensitivity Analysis of the DJAN Model

As shown in Tables 6–8, we investigated the sensitivity of hyperparameters α, β, and γ on the transfer tasks **ES→NS** and **NS→ES**. The strategy we adopted was as follows: if we wanted to determine the value of parameter α, we fixed the values of parameters β and γ and obtained the optimal value for parameter α. Similarly, we followed the same process to determine the optimal values for parameters β and γ: by fixing the values of the other parameters. From Tables 6–8, we can observe the following: (1) The parameter α changes within the range of $[0.2, 0.4, 0.6, 0.8, 1.0, 1.2]$. The transfer performance of our method in this paper initially increases and then decreases. As α increases, the performance first increases and then decreases. Therefore, we set the parameter α to 1.0. (2) The parameter β is optimized within the range of $[0.2, 0.4, 0.6, 0.8, 1.0, 1.2]$, the accuracy of wildlife image recognition using our method initially stabilizes and then increases before eventually decreasing. Therefore, the value of parameter β is set to 0.25. (3) The parameter γ is optimized within the range of $[0.15, 0.25, 0.35, 0.45, 0.55, 0.65]$, the transfer gain of the DJAN method first increases steadily and then decreases. Therefore, in our DJAN method, the parameter $\gamma = 0.25$.
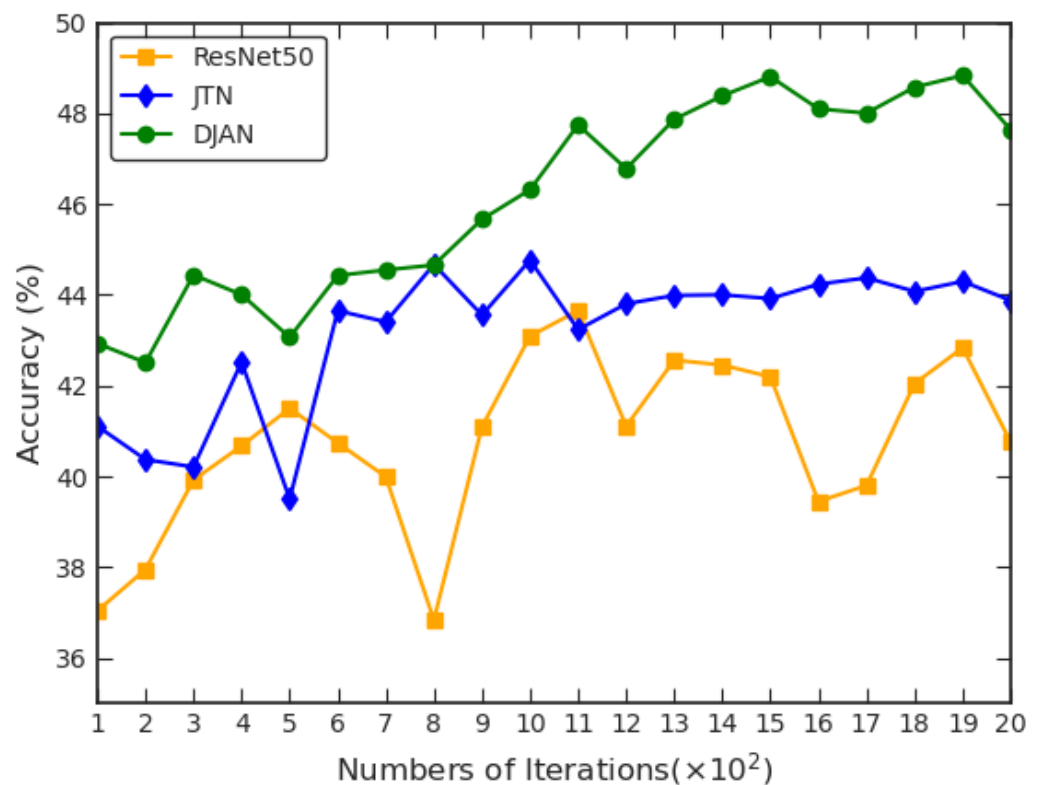
Figure 6. The curve of test accuracy based on ResNet50, JTN, and DJAN on transfer task **ES→NS**.

Table 6. The experimental comparison results of different hypermeter α weights on wildlife dataset.

α	0.2	0.4	0.6	0.8	1.0	1.2
ES→NS	41.3	43.4	45.8	41.7	48.8	46.4
NS→ES	56.9	57.1	56.5	60.2	67.5	58.0

Table 7. The experimental comparison results of different hypermeter β weights on wildlife dataset.

β	0.2	0.4	0.6	0.8	1.0	1.2
ES→NS	41.3	42.4	43.8	42.7	48.8	42.1
NS→ES	54.0	56.8	53.4	54.6	67.5	58.0

Table 8. The experimental comparison results of different hypermeter γ weights on wildlife dataset.

γ	0.15	0.25	0.35	0.45	0.55	0.65
ES→NS	43.8	48.8	40.8	39.8	42.6	40.4
NS→ES	51.5	67.5	55.1	56.1	58.8	55.1

4. Discussion

By joining adversarial learning image recognition and vision transformer study research, a Deep Joint Adaptation Network is proposed for a wildlife image recognition model that can both accurately recognition wildlife and enhance generalization ability. Different from domain adaptation wildlife image recognition, such as [20–22], the proposed method generates domain-transferable feature representations across domains by the correlation alignment constraint, conditional adversarial training, and cross-domain local and global representation learning method. The correlation alignment constraint and the strategy of conditional adversarial training improve the capability of individual domain adaptation modules. In addition, a transformer unit is utilized to capture the long-range relationships between the local and global feature representations, which facilitates better

understanding of the overall structure and relationships within the image. In the study of wildlife image recognition, some deep learning models, such as [10–12], also yielded significant recognition results. However, when these methods are confronted with datasets where the feature distribution is significantly affected by factors such as different backgrounds, lighting conditions, and varying scales of capture, the recognition performance can degrade significantly. Moreover, our method avoids the use of costly sample annotation information, which promotes the application of wildlife image recognition in real-world scenarios.

As shown in Tables 3–5, it can be observed that conditional adaptation learning and correlation domain alignment obviously further promote the ability of the model to extract more domain-transferable features. This reflects that the combination of the two could further enhance highly transferable feature learning, strengthening the complementarity between adversarial feature learning and domain correlation alignment and thus achieving higher accuracy in wildlife image recognition. However, this mechanism does not consider the long-term relationship between local and global feature representations, which can help to better understand the overall structure and relationships within the image. In this paper, with the help of the domain adaptation strategy of conditional adaptation learning and correlation domain alignment, we designed a transformer loss constraint to capture the long-range relationships between the local and global feature representations, which facilitated better understanding of the overall structure and relationships within the image. With the utilization of the transformer loss to guide the DJAN model learning cross-domain local and global representation, the DJAN model yielded an accuracy of 58.2%, which achieved a 7.9 percentage point improvement in **Accuracy** compared to utilizing conditional adaptation learning and the correlation domain alignment constraint.

The proposed method for wildlife recognition in this paper has indeed improved the recognition accuracy. However, there are still instances where certain individual wildlife species are misclassified (as shown in Figure 5). The main reason for this is the impact of the wildlife photography environment on the transferability of features and the recognition results. Therefore, it is worth considering the introduction of image enhancement algorithms to mitigate the influence of complex background factors and to improve the effectiveness of the model. Furthermore, the next step of this research involves deploying the proposed wildlife recognition model on edge devices to achieve real-time, efficient, and privacy-secure wildlife monitoring at the edge.

5. Conclusions

This study presented the Deep Joint Adaptation Network (DJAN), a novel network to address the issue of weak generalization in wildlife image recognition by leveraging the principles of transfer learning. DJAN incorporates two key components: the correlation alignment constraint and conditional adversarial training, which collectively enhance the adaptability of individual domain-adaptation modules. Additionally, transformer units are employed to capture long-range relationships between local and global feature representations, enabling a more comprehensive understanding of the image's overall structure and relationships. Experimental evaluations conducted on wildlife datasets testify of the effectiveness of the DJAN model, yielding state-of-the-art results. Compared to baseline methods, our DJAN approach achieved an average accuracy improvement of 3.6 percentage points for the classification of eleven wildlife species. These findings highlight the potential of DJAN in advancing wildlife image recognition and its application in the conservation and monitoring of diverse wildlife species in open environments. Further research can explore additional optimizations and extend the use of DJAN to enhance our understanding and preservation of wildlife biodiversity.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No data were used for the research described in the article.

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

References

1. Ji, Y.; Wei, X.; Li, D.; Feng S. A framework for assessing variations in ecological networks to support wildlife conservation and management. *Ecol. Indic.* **2021**, *155*, 110936. [CrossRef]
2. Yang, D.; Ren, G.; Tan, K.; Huang, Z.; Li, D.; Li, X.; Wang, J.; Chen, B.; Xiao, W. An adaptive automatic approach to filtering empty images from camera traps using a deep learning model. *Wildl. Soc. Bull.* **2021**, *45*, 230–236. [CrossRef]
3. Vélez, J.; McShea, W.; Shamon, H.; Castiblanco-Camacho, P.; Tabak, M.; Chalmers, C.; Fergus, P.; Fieberg, J. An evaluation of platforms for processing camera-trap data using artificial intelligence. *Methods Ecol. Evol.* **2023**, *14*, 459–477. [CrossRef]
4. Fisher, J.T. Camera trapping in ecology: A new section for wildlife research. *Ecol. Evol.* **2023**, *13*, e9925. [CrossRef]
5. Cordier, C; Smith D.; Smith Y.; Downs, C. Camera trap research in Africa: A systematic review to show trends in wildlife monitoring and its value as a research tool. *Glob. Ecol. Conserv.* **2022**, *40*, e02326.
6. Miao, Z.; Liu, Z.; Gaynor, K.; Palmer, M.; Yu. S.; Getz, W. Iterative human and automated identification of wildlife images. *Nat. Mach. Intell.* **2021**, *3*, 885–895. [CrossRef]
7. Tuia, D.; Kellenberger, B.; Beery, S.; Costelloe, B.; Zuffi, S.; Risse, B.; Mathis, A.; Mathis, M.; Langevelde, F.; Burghardt, T.; et al. Perspectives in machine learning for wildlife conservation. *Nat. Commun.* **2022**, *13*, 792. [CrossRef]
8. Petso, T.; Jamisola, R.S.; Mpoeleng, D. Review on methods used for wildlife species and individual identification. *Eur. J. Wildl. Res.* **2022**, *68*, 3. [CrossRef]
9. Roy, A.M.; Bhaduri, J.; Kumar, T.; Raj, K. WilDect-YOLO: An efficient and robust computer vision-based accurate object localization model for automated endangered wildlife detection. *Ecol. Inform.* **2023**, *75*, 101919. [CrossRef]
10. Gomez, A.; Salazar, A.; Vargas-Bonilla, J.F. Towards automatic wild animal monitoring: Identification of animal species in camera-trap images using very deep convolutional neural networks. *Ecol. Inform.* **2016**, *75*, 41. [CrossRef]
11. Trnovszky, T.; Kamencay, P.; Orjesek, R.; Benco, M.; Sykora, P. Animal recognition system based on convolutional neural network. *Adv. Electr. Electron. Eng.* **2017**, *15*, 517–525. [CrossRef]
12. Verma, G.K.; Gupta, P. Wild animal detection from highly cluttered images using deep convolutional neural network. *Int. J. Comput. Intell. Appl.* **2018**, *17*, 1850021. [CrossRef]
13. Schneider, S.; Greenberg, S.; Taylor, G.W.; Kremer, S.C. Three critical factors affecting automated image species recognition performance for camera traps. *Ecol. Evol.* **2020**, *10*, 3503–3517. [CrossRef] [PubMed]
14. Vargas-Felipe, M.; Pellegrin, L.; Guevara-Carrizales, A.; López-Monroy, P.; Escalante, H.; Gonzalez-Fraga, J. Desert bighorn sheep (*Ovis canadensis*) recognition from camera traps based on learned features. *Ecol. Inform.* **2021**, *64*, 101328. [CrossRef]
15. Schindler, F.; Steinhage, V. Identification of animals and recognition of their actions in wildlife videos using deep learning techniques. *Ecol. Inform.* **2021**, *61*, 101215. [CrossRef]
16. Yin, Y.; Yang, Z.; Hu, H.; Wu, X. Universal multi-Source domain adaptation for image classification. *Pattern Recognit.* **2022**, *121*, 108238. [CrossRef]
17. Oza, P.; Sindagi, V; Sharmini, V.V.; Patel, V. Unsupervised domain adaptation of object detectors: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, 1–24. [CrossRef]
18. Peng, S.; Zeng, R.; Cao, L.; Yang, A.; Niu, J.; Zong, C.; Zhou, G. Multi-source domain adaptation method for textual emotion classification using deep and broad learning. *Knowl.-Based Syst.* **2023**, *260*, 110173. [CrossRef]
19. Wang, S.; Wang, B.; Zhang, Z.; Heidari, A.; Chen, H. Class-aware sample reweighting optimal transport for multi-source domain adaptation. *Neurocomputing* **2023**, *523*, 213–223. [CrossRef]
20. Norouzzadeh, M.S.; Nguyen, A.; Kosmala, M.; Swanson, A.; Palmer, M.; 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] [PubMed]
21. Wang, X.; Li, P.; Zhu, C. Classification of wildlife based on transfer learning. In Proceedings of the International Conference on Video and Image Processing, Xi'an, China, 25–27 December 2020; pp. 236–240.
22. Thangaraj, R.; Rajendar, S.; Sanjith, M.; Sasikumar, S.; Chandhru, L. Automated Recognition of Wild Animal Species in Camera Trap Images Using Deep Learning Models. In Proceedings of the Third International Conference on Advances in Electrical, Sanya, China, 24–26 March 2023; pp. 1–5.
23. Sun, B.; Feng, J.; Saenko, K. Return of frustratingly easy domain adaptation. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; Volume 30.
24. Sun, B.; Saenko, K. Deep coral: Correlation alignment for deep domain adaptation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 443–450.

25. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, B.; Ozair, S.; Courville A.; Bengio, Y. Generative adversarial networks. *Commun. ACM*. **2020**, *63*, 139–144. [CrossRef]

26. He, C.; Zheng, L.; Tan, T.; Fan, X.; Ye, Z. Manifold discrimination partial adversarial domain adaptation. *Knowl.-Based Syst.* **2022**, *252*, 109320. [CrossRef]

27. Shi, H.; Huang, C.; Zhang, X.; Zhao, J.; Li, S. Wasserstein distance based multi-scale adversarial domain adaptation method for remaining useful life prediction. *Appl. Intell.* **2023**, *53*, 3622–3637. [CrossRef]

28. Fu, S.; Chen, J.; Lei, L. Cooperative attention generative adversarial network for unsupervised domain adaptation. *Knowl.-Based Syst.* **2023**, *261*, 110196. [CrossRef]

29. Luo, X.; Chen, W.; Liang, Z.; Li, C.; Tan, Y. Adversarial style discrepancy minimization for unsupervised domain adaptation. *Neural Netw.* **2023**, *517*, 216–225. [CrossRef] [PubMed]

30. She, Q.; Chen, T.; Fang, F.; Zhang, J.; Gao, Y.; Zhang, Y. Improved Domain Adaptation Network Based on Wasserstein Distance for Motor Imagery EEG Classification. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2023**, *31*, 1137–1148. [CrossRef] [PubMed]

31. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **2016**, *17*, 2096–2030.

32. Long, M.; Cao, Y.; Wang, J.; Jordan, M. Learning transferable features with deep adaptation networks. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 97–105.

33. Chen, X.; Wang, S.; Long, M.; Wang, J. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In Proceedings of the International Conference on Machine Learning, Lugano, Switzerland, 18–22 June 2019; pp. 1081–1090.

34. Tzeng, E.; Hoffman, J.; Saenko, K.; Darrell, T. Adversarial discriminative domain adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7167–7176.

35. Volpi, R.; Morerio, P.; Savarese, S.; Murino, V. Adversarial feature augmentation for unsupervised domain adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake, UT, USA, 18–22 June 2018; pp. 5495–5504.

36. Pei, Z.; Cao, Z.; Long, M.; Wang, J. Multi-adversarial domain adaptation. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 3.

37. Long, M.; Cao, Z.; Wang, J.; Jordan, M. Conditional adversarial domain adaptation. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; Volume 31

38. Yousif, H.; Kays, R.; He, Z. Dynamic programming selection of object proposals for sequence-level animal species classification in the wild. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 1–14.

39. Tabak, M.; Norouzzadeh, M.; Wolfson, D; Sweene, S.; Vercauteren, K.; Snow, N.; Halseth, J.; Salvo, P.; Lewis, J.; White, M. Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods Ecol. Evol.* **2019**, *10*, 585–590. [CrossRef]

40. Torralba, A.; Efros, A.A. Conditional adversarial domain adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Providence, RI, USA , 20–25 June 2011; pp. 1521–1528.

41. Tian, H.; Tao, Y.; Pouyanfar, S.; Chen, S.; Shyu, M. Multimodal deep representation learning for video classification. In Proceedings of the International Conference on World Wide Web, San Francisco, CA, USA, 13–17 May 2019; pp. 1325–1341.

42. Zhang, C.; Zhao, Q.; Wu, H. Deep domain adaptation via joint transfer networks. *Neurocomputing* **2022**, *489*, 441–448. [CrossRef]

43. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

44. Pan, S.J.; Kwok, J.T.; Yang, Q.; Pan, J. Adaptive localization in a dynamic WiFi environment through multi-view learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, British, 22–26 July 2007; pp. 1108–1113.

45. Zhu, Y.; Zhuang, F.; Wang, J.; Ke, G.; Chen, J.; Bian, J.; Xiong, H.; He, Q. Deep subdomain adaptation network for image classification. *IEEE Trans. Neural Netw. Syst. Learn.* **2020**, *32*, 1713–1722. [CrossRef]

46. Cui, S.; Wang, S.; Zhuo, J.; Li, L.; Huang, Q.; Tian, Q. Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3941–3950.

47. Zhang, C.; Zhao, Q.; Wang, Y. Hybrid adversarial network for unsupervised domain adaptation. *Inf. Sci.* **2020**, *514*, 44–55. [CrossRef]