Recognition of Unseen Bird Species by Learning from Field Guides

Andrés C. Rodríguez[1] * Stefano D'Aronco [1] Rodrigo Caye Daudt[1] Jan D. Wegner[1,2]
Konrad Schindler[1]
[1] EcoVision Lab - Photogrammetry and Remote Sensing, ETH Zurich, Switzerland
[2] Institute for Computational Science, University of Zurich, Switzerland

Abstract

We exploit field guides to learn bird species recognition, in particular zero-shot recognition of unseen species. Illustrations contained in field guides deliberately focus on discriminative properties of each species, and can serve as side information to transfer knowledge from seen to unseen bird species. We study two approaches: (1) a contrastive encoding of illustrations, which can be fed into standard zero-shot learning schemes; and (2) a novel method that leverages the fact that illustrations are also images and as such structurally more similar to photographs than other kinds of side information. Our results show that illustrations from field guides, which are readily available for a wide range of species, are indeed a competitive source of side information for zero-shot learning. On a subset of the iNaturalist2021 dataset with 749 seen and 739 unseen species, we obtain a classification accuracy of unseen bird species of 12% *@top-1 and* 38% *@top-10, which shows the potential of field guides for challenging real-world scenarios with many species. Our code is available at* https://github.com/ac-rodriguez/zsl_billow.

1. Introduction

Fine-grained species recognition is essential for biodiversity monitoring. Identifying the species of observed animals and plants is the basis for several important biodiversity indicators, e.g., the number of different species in an area, the abundance of individual species, and their geographical distribution. Many species are locally or globally threatened by human activities, making it all the more important to monitor their distributions and support conservation efforts [10].

A bottleneck for automatic species recognition in the wild has long been the collection of enough observations. There are different modalities for automatic species recognition. Perhaps the two most prominent ones are acoustic recognition from sound recordings and visual recognition form images. While the focus of this work remains on the latter, acoustic recognition is especially relevant for bird species identification. It is a popular way to do abundance estimation and was explored in early works with Support Vector Machines [11]. Abundance estimation via sound recordings remains an active research area, where new datasets and competitions are being published [33, 38]. For visual recognition, in the last years, the cooperation of experts and nature enthusiasts has enabled the emergence of community science projects. Volunteers record and share images and locations of their observations, which experts can curate and organise to obtain large-scale databases for biodiversity monitoring. Examples include the iNaturalist [20] and eBirds [39] projects. The eBirds platform alone has accumulated >34 million images for bird species, from ≈800'000 contributors. Those databases make it possible to train automatic species recognition systems, which would be a valuable asset for scalable biodiversity monitoring.

In principle, automatic species identification can capitalise on the recent advances in computational object recognition. It now achieves human-level performance, and is far more scalable than manual labelling of images; especially in cases where specialized expertise might be needed. [1]

Provided a large volume of labelled training data, one can resort to a supervised learning scheme: A model learns to classify a specific bird species from many images of the bird of interest in many expected natural conditions and backgrounds. This usually means that a large volume of labelled images is needed for training. Due to the sheer number of species in most ecosystems, many of which are rare or at least rarely spotted, it can be extremely challenging to gather a sufficient number of training samples for every one of them. For example, the iNaturalist 2021 dataset [43] comprises 1'486 bird species, yet the Birds of the World collection [7] reports over 10'000 known bird species.

When data collection is limited, one can resort to machine learning strategies other than supervised learning that may still be able to deliver acceptable recognition results,

[1] E.g., on ImageNet computers outperform most humans when it comes to recognising different dog breeds, as well as different species of mushrooms.

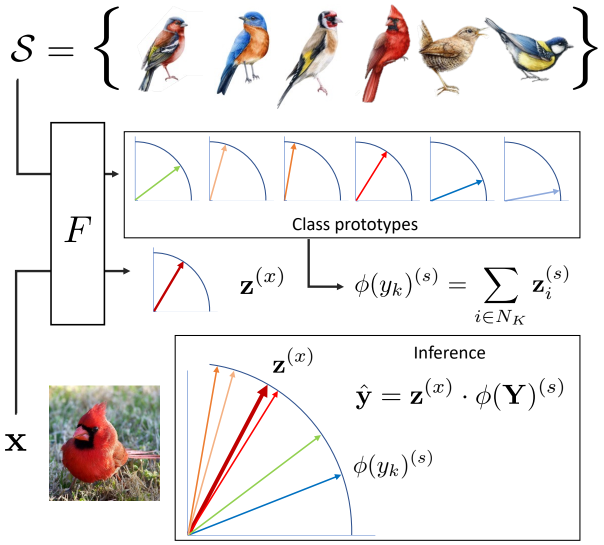
$\mathcal{S} = \{ \}$
F
Class prototypes
$\mathbf{z}^{(x)}$
$\phi(y_k)^{(s)} = \sum_{i \in N_K} \mathbf{z}_i^{(s)}$
Inference
$\mathbf{z}^{(x)}$
$\hat{\mathbf{y}} = \mathbf{z}^{(x)} \cdot \phi(\mathbf{Y})^{(s)}$
$\phi(y_k)^{(s)}$
$\mathbf{x}$

Figure 1. Zero-shot learning with field guides via prototype alignment. Class prototypes (depicted here as different colored vectors) are learned using a shared feature extractor F between photographs and illustrations. At inference time the class with the largest dot-product to $\mathbf{z}^{(x)}$ is predicted.

although these typically do not attain the same performance of a model supervised with enough data. For instance, one can use few-shot learning if only few labelled examples are available for certain classes [46]. In the extreme case, **Zero-Shot Learning** (ZSL) refers to the scenario where no training samples are available at all for some target classes [1, 13, 26, 48]. This requires class-wise characteristics (side information) rather than labelled data, since labelled examples are not available for training.

Traditionally, professional as well as amateur observers rely on **field guides** to recognise animal and plant species in nature. This works remarkably well. Even if new formats of field guides arise, such as interactive maps and mobile apps to aid species recognition [12], the basic principle remains the same: the field guide provides a clear, representative visual example that emphasises the distinctive properties and visual cues needed to identify a species and to discriminate it from similar ones.

The question we explore in this paper is: Can we exploit illustrations from field guides to compensate for the lack of training data for some classes? Field guides are easily accessible, cover a broad range of species, and although they normally contain only few images of a species – sometimes only a single illustration – they allow humans to identify it in most cases. One can think of a field guide as a collection of manually created, discriminative class prototypes: the artists who create the illustrations are highly specialised professionals, and they make a conscious effort to render each species such that the illustration not only faithfully reproduces attributes like colour and shape, but optimally typifies its peculiarities and makes it distinguishable from other species. Moreover, illustrations are available also for rare, endangered and even for extinct species.

Although naturalistic illustrations resemble photographs in many ways, joint supervised training without additional regularisation leads to biases towards photographic textures, such that the classifier tends to recognise only seen classes. To tackle this problem, we propose to interpret illustrations as species-specific attribute information and leverage them in a zero-shot setting. At this point, a technical difficulty arises: Existing ZSL algorithms ingest attributes in the form of low-dimensional vectors, and we observe that the high dimensionality of illustrations, compared to conventional binary attributes (e.g., belly shape, or eye colour), challenges existing ZSL algorithms. In this work we tackle this problem and demonstrate how illustrations from birding field guides can be exploited for zero-shot learning.

We make the following contributions: (1) We introduce the *Bird Illustrations of the World* (Billow) dataset for fine-grained zero-shot classification of bird species at an unprecedented scale; (2) we propose a contrastive embedding of the illustrations that enables existing ZSL algorithms to leverage the high-dimensional side information contained in Billow; and (3) we propose a novel zero-shot learning scheme better suited for side-information in the form of illustrations. Its fundamental principle is to train a model that can process either illustrations or photographs and in both cases arrives at the same predictions and aligns the class prototypes from the illustrations with the photographs, as depicted in Figure 1.

We use Billow in conjunction with commonly used datasets with natural images commonly used in fine-grained classification. With the help of those datasets we compare our method to the state-of-the-art in ZSL as well as domain adaptation. The experiments show that Billow matches the performance of other, more structured forms of side information, confirming the hypothesis that field guides are a valuable auxiliary source of information for species recognition. We hope that our work will encourage further research into biodiversity mapping, and may serve as a first step towards unlocking the treasure trove of biological field guides, beyond Billow.

2. Related Work

Zero-Shot Learning. Early work on ZSL focused on defining class embedding spaces and visual spaces, then measuring some matching metric to predict a class [1, 13, 26]. The embedding space used for matching plays a crucial role [37, 52]. Mapping to a space closer to the class embedding can lead to a hubness problem, where a classifier is strongly biased to predict only a subset of labels.

Current state-of-the-art methods rely on generative models to map class embeddings into a visual embedding space to avoid such a problem. They use a generator to create synthetic samples that attempt to emulate real samples from unseen classes. These samples are then used to supervise the training of a machine learning algorithm along with the examples from the seen classes. One of the first studies to use a generative approach in ZSL is [49]. They use a Generative Adversarial Network (GAN) to synthesize visual examples of the unseen classes using class descriptions as conditional information. An additional classification loss ensures that the generated features have sufficient discriminative information. TFVAEGAN [29] models the embedding space using a variational formulation. The method also has a feedback network that modulates the latent representations to further improve performance. Invertible Zero-shot recognition flows [36] use invertible layers in order to learn a mapping from the class description to the visual features. Counterfactual ZSL [51] exploits "sample attributes" from the training classes to create synthetic samples with class attributes from the unseen classes. CE-GZSL [17] uses a contrastive loss that results not only in class-wise but also instance-wise supervision. LsrGAN [45] propose a novel semantic regularized loss which promotes visual features that reflect the semantic relationships between seen and unseen classes.

Side-information in ZSL. ZSL requires a sort of side-information to guide the learning and transfer knowleged from the seen classes to the unseen classes. The use of illustrations for ZSL is not new. Early work has attempted to use digital characters as side information for character recognition [27]. In [4], authors use user generated pose graphics as side information for action recognition in a ZSL setting. Sketches of objects have been used for image retrieval tasks [35].

Generative approaches work very well in cases where the side information has low dimensional embedding, and can be used as a deterministic condition by the generator to synthesize samples from unseen classes. While there are currently many types of side information used in ZSL, all of them are rather low-dimensional. In [3], authors evaluate different supervised and unsupervised embeddings for ZSL. Such types of side information include manually created binary attributes per class [2], visual descriptions [34], automatic embeddings from Wikipedia descriptions [3], and more recently learned embeddings of DNA sequences for fine-grained species classification [5]. However, it remains unclear how to use the previously discussed methods if the side information is high-dimensional, as is the case for field guide illustrations, without a low-dimensional embedding step as preprocessing.

Domain adaptation. Given that illustrations and photograph are similar in nature (as opposed to images and text embeddings or DNA sequences, for instance) we also draw from literature regarding Domain Adaptation (DA). Such studies aim to improve the performance of a model trained on the source domain in which enough data are available for supervision and applied on a target domain in which the available data are not enough for supervision.

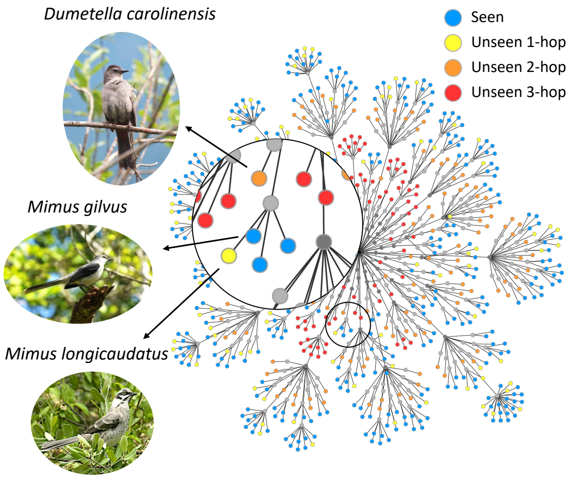

Figure 2. Hierarchical representation of the Passeriformes order of the iNat2021 dataset for Seen and i-hop unseen classes.

The most versatile form of DA is Unsupervised Domain Adaptation (UDA), in which no supervision signal is available for the target domain. Most methods aim to match the distributions of the source and target domains in a latent space, either explicitly or implicitly [21, 40, 41]. [15] propose using a gradient reversal layer which aims to make the samples from both domains statistically indistinguishable in a representation space. Since, many other methods have been proposed that use adversarial training of a discriminator to enforce alignment of the domains in a latent space [16, 42, 53]. These methods have been shown to work well in established domain adaptation benchmarks, but use cases for fine-grained classification are somewhat unexplored.

Other studies focus on the case where some supervised samples are available in the target domain. Well established methods exist to combine source and target data to improve a ML system's performance on target domain data [9]. [28] propose a unified framework for domain adaptation of deep models using a Siamese architecture to align different visual domains. Notably, [24, 50] use memory banks of latent representations of instances from both source and target domains in a few-shot learning setting, which are then used to create class prototypes for each domain.

However, applying these methods directly for ZSL is not straightforward, as seen classes tend to dominate predictions from the target domain if no additional regularization is done. Our novel method aims to close the domain gap from illustrations and photographs in a ZSL setting.

3. Bird Illustrations of the World Dataset

We introduce the Bird Illustrations of the World (Billow) dataset for Generalized Zero-Shot Learning (GZSL) in fine-grained classification. The dataset consists of illustrations from the Birds of the World project [7] collected and organized by the Cornell Lab of Ornithology. Billow includes 22’351 illustrations covering 10’631 different species, 2’279 genera, 249 families, and 41 orders.

All illustrations in the dataset share a standardized graphical style: side view in front of white background, in neutral pose. Most species have illustrations for a male and a female specimens, some also include a close-up of the bird’s head.

The original artworks may be accessed with a valid subscription to the *Birds of the World* project, and are subject to a licence of use.

For reproducibility and to support further research and comparisons, we describe here how to access all illustrations from Billow. We have used all the images available by November 1st 2021. The encoded dataset, after contrastive embedding with our method, is available along our code on Github. To download the raw illustrations we provide a python script. Note though that any further use is subject to the licensing conditions of [7].

3.1. Illustrations for Zero-Shot Learning

We use Billow with three widely used datasets in Computer Vision, namely Caltech-UCSD Birds-200-2011 (CUB) [47] and the bird subsets of iNaturalist 2017 [44] and iNaturalist 2021 [43]. The list of species included in Billow covers almost all species of the CUB dataset (196 out of 200), and also the overwhelming majority of bird species from iNaturalist 2017 (895 out of 954) and iNaturalist 2021 (1485 out of 1486). Note that the opposite is not true: even the 1485 bird classes of iNaturalist 2021 are only a small fraction of the 10’631 species present in Billow. This raises the question of whether we can leverage the rich information contained in the Billow dataset and combine it with a dataset of photographs, to advance the state-of-the-art in fine-grained (bird) species recognition.

For ZSL with CUB, there is a default split into 150 seen and 50 unseen classes [48]. CUB uses common names, not scientific names. Hence, previous work had to map the common names to scientific ones, e.g., to leverage the hierarchical label structure [6], or to utilize genetic information [5]. We have revised and merged these assignments, and only retain mappings for which we found a one-to-one correspondence between the common and scientific name. In Billow we matched 196 out of the 200 CUB classes.

Dataset	Train	Val					
		Seen	Unseen	1hop	2hop	3hop	4hop
iNat2017							
N	97,067	8,626	11,073	2,204	3,613	2,175	3,081
K	381	381	514	87	177	110	140
iNat2021							
N	211,027	7,490	7,360	1,680	2,860	1,580	1,240
K	749	749	736	168	286	158	124

Table 1. Zero-shot splits of iNaturalist bird classes. N and K denote the numbers of samples and classes, respectively, in each set

For the iNaturalist datasets, we propose a seen/unseen split. Similar to previous ZSL work that uses ImageNet [14, 31], we construct several groups of *unseen* classes, which have different distances to *seen* classes in the label hierarchy. In this way, we can assess the performance of ZSL for unseen classes that are increasingly distant from the seen ones. We first randomly select seen species, and from the remaining species we define the i-hop set as the set of all classes whose distance to the nearest seen class in the taxonomic tree is equal to i (i.e., they belong to the same superclass at the i-th taxonomic level). For example, the classes in the 2-hop set share the family (2^{nd} level) with at least one seen class, but do not share the same genus with any of them. We consider the species, genus, family and order levels to obtain 0-hop (i.e., seen classes), 1-hop, 2-hop and 3-hop sets. Classes in the 4-hop set do not have members of the same taxonomic group in any level of the seen set.

The intersection of the *Aves* super-class from iNaturalist 2017 with Billow contains 895 species. These are randomly split into 381 seen and 515 unseen classes. From the unseen ones we construct the 4 different i-hop sets for validation. We repeat the same procedure with iNaturalist 2021: the intersection of its *Birds* super-category with Billow contains 1485 species. These are split into 749 seen and 736 unseen classes. See Fig. 2 for an illustration of the validation splits, and Tab. 1 for the sizes of each split.

4. Method

In ZSL we are given a set of classes $\mathcal{Y}$ made up of two disjoint sets $\mathcal{Y}_{\text{seen}}$ and $\mathcal{Y}_{\text{unseen}}$. Side information $\mathbf{s}$ is available for every class in $\mathcal{Y}$. In most cases a single instance of $\mathbf{s}$ is available for each class, although this is not a requirement. We can think of $\mathbf{s}_y$ a (possibly incomplete) description of the class $y \in \mathcal{Y}$, e.g., a text, a list of semantic attributes, or a gene sequence. A training set of pairs $(\mathbf{x}, y)$ is available exclusively for the seen classes. In computer vision, $\mathbf{x}$ typically refers to photographic images. The goal of ZSL is to use that information to build a classifier $F(\mathbf{x})$, which can recognize samples from *unseen* classes $\mathcal{Y}_{\text{unseen}}$. Similarly, Generalized Zero-Shot Learning (GZSL) methods aim for good classification performance on test samples from both the seen *and* unseen classes $\mathcal{Y} = \mathcal{Y}_{\text{seen}} \cup \mathcal{Y}_{\text{unseen}}$.

The side information $\mathbf{s}$ provides cues about similarities (common features) between classes and serves as a bridge that enables the recognition of instances of unseen classes. Commonly the side information comes in the form of low-dimensional vectors, like for instance binary presence/absence flags for a number of attributes. On the contrary, our illustrations are images of a specific style / domain.

In order to utilise these illustration for ZSL, we explore two different strategies. We start with a two-stage strategy, where we first learn a *Contrastive Encoding* of the illustrations, such that the resulting codes can be fed into existing ZSL methods at a second stage. We then go on to develop a more advanced method, named *Prototype Alignment*, where a single end-to-end network is trained to map both illustrations and photographs to similar latent representations, in order to better leverage their similar structure.

4.1. Contrastive Encoding of Billow

Standard ZSL methods in literature assume the existence of class descriptions $\phi(y)$ in the form of low-dimensional vectors derived from some sort of side information that should help recognise seen and unseen classes. We must therefore compute $\phi(y)$ as a first step to use those methods.

Let $\mathcal{D}_{\text{side}}$ be the set of pairs $(\mathbf{s}, y)$ where $\mathbf{s} \in \mathcal{S}$ is an illustration associated with class $y \in \mathcal{Y}$. To turn illustrations into low-dimensional vectors, we use an encoding network E that produces an embedding $\mathbf{z} = E(\mathbf{s})$. These embeddings should preserve discriminative class information, so we simply add a classification head $\hat{\mathbf{y}} = C(\mathbf{z})$ and optimize E and C with a cross-entropy loss $L_{\text{cls}}(\hat{\mathbf{y}}_i, \mathbf{y}_i)$. Here, $\mathbf{y}$ represents the one-hot encoding of y.

However, the ability to discriminate classes is not enough. The embeddings should also live in a metric space where pairwise differences between them are meaningful, so as to handle the zero-shot setting. One way to achieve this is to model the overall distribution of illustrations in an embedding space. For instance, one can employ a Variational Auto-encoder (VAE), that assumes the embedding $\mathbf{z}$ to model a prior distribution in the latent space from which it is possible to draw samples and decode them to the original input space $\mathcal{X}$ [8, 19]. This approach, however, risks reducing the representation power of $\mathbf{z}$ if it is too strongly regularized by the prior distribution (see Supplementary).

An alternative is to use a contrastive loss that promotes an embedding space with a uniform distribution over the unit-sphere [23]. We apply a projection and normalization head $\tilde{\mathbf{z}} = h(\mathbf{z})$ to the embeddings before computing the contrastive loss. Following [23], our contrastive loss function is

$$L_{\text{cont}}(\tilde{\mathbf{z}}_i) = -\frac{1}{||P(i)||_1} \sum_{p \in P(i)} \log \frac{\exp\left(\frac{1}{\tau}\tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_p\right)}{\sum_{j \in B} \exp\left(\frac{1}{\tau}\tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_j\right)}, \quad (1)$$

where $P(i)$ is a set of samples in the training batch B that have the same class label as $\mathbf{x}_i$, and $\tau \in \mathbb{R}^+$ is a tunable temperature parameter. Finally we train F, C and h with both classification and contrastive losses: $L = L_{\text{cls}} + L_{\text{cont}}$. As final representation of class y, we compute

$$\phi(y) = \eta\left(\sum_{\mathbf{s} \in \mathcal{S}_y} E(\mathbf{s})\right), \quad (2)$$

where $\mathcal{S}_y$ is the set of all illustrations available for class y, and $\eta(\mathbf{z}) = \mathbf{z}/||\mathbf{z}||$ denotes L^2-normalization. The embeddings $\phi(y)$ derived from illustrations can then be used as class descriptors in different existing ZSL methods.

4.2. Prototype Alignment

In contrast to other types of side information for ZSL, illustrations also belong to the visual domain. We leverage this property and propose Prototype Alignment (PA) for ZSL with visual side information, which allows us to bypass the encoding step required by all previous ZSL-methods. Inspired by [50], we explore a view of the problem through the lens of few-shot *domain adaptation*: The source domain are illustrations, the target domain are natural, photographic images.

Let $\mathbf{s}$ and $\mathbf{x}$ be samples from the source domain $\mathcal{S}$ and the target domain $\mathcal{X}$, respectively. We have access to samples from all classes $\mathcal{Y}$ in the source domain, but only to samples of the seen classes $\mathcal{Y}_{\text{seen}}$ in the target domain. Furthermore, we also do not have unlabelled samples of unseen classes in the target domain.

We train a feature extractor network F that takes input samples from either domain and outputs a latent representation $\mathbf{z}$. The last operation in F is an L^2-normalization layer $\eta(\cdot)$, as also used in Eq. (2). During training, we keep a memory bank in each domain, with a prototype $\mathbf{z}$ of each class. For the illustrations in the source domain, that representation can be interpreted as the class embedding $\phi(y_k)^{(s)}$ that is used for ZSL. Note that, in contrast to previous approaches [24, 50], we do not keep an instance-wise memory bank, which would lead to intractable memory demands for larger datasets.

For the sake of simplicity, we omit the domain indicator from this point on where possible. In every iteration, we update the memory bank in each domain with the latent representation of the new samples, with momentum m:

$$\phi(y_k) \leftarrow \eta\left((1-m)\mathbf{z}_k + m\phi(y_k)\right). \quad (3)$$

To promote compact and discriminative class representations, we apply a contrastive in-domain loss similar to Eq. 1, via a projection head h:

$$L_c(\mathbf{z}_i, \phi(y_i)) = -\log \frac{\exp\left(\frac{1}{\tau}h(\mathbf{z}_i)h(\phi(y_i))\right)}{\sum_{k \in C} \exp\left(\frac{1}{\tau}h(\mathbf{z}_i)h(\phi(y_k))\right)}. \quad (4)$$

In contrast to [50] we refrain from applying a cross-domain contrastive loss to close the domain gap. Instead, we sidestep the gap by directly using the class prototypes from *both* domains for classification, so as to force the network F to produce class-discriminative features. To obtain class logits, we compute the dot-product between an image embedding $\mathbf{z}$ and the embeddings $\phi(\mathcal{Y})$ of the classes from both domains, $\hat{\mathbf{y}}^{(s)} = \mathbf{z}\cdot\phi(\mathbf{Y})^{(s)}$ and $\hat{\mathbf{y}}^{(x)} = \mathbf{z}\cdot\phi(\mathbf{Y}_{\text{seen}})^{(x)}$. These serve as input to a cross-entropy loss L_{cls} for supervision:

$$L_{\text{cls}}\left(\hat{\mathbf{y}}^{(s)},\hat{\mathbf{y}}^{(x)},\mathbf{y}\right) = L_{\text{cls}}\left(\hat{\mathbf{y}}^{(s)},\mathbf{y}\right) + L_{\text{cls}}\left(\hat{\mathbf{y}}^{(x)},\mathbf{y}\right). \quad (5)$$

Eq. 5 encourages sample representations that are discriminative w.r.t. prototypes from the *other* domain, which in turn aligns the two domains. Note also that the second term in Eq 5 is only computed for seen classes, as it depends on $\phi(\mathbf{Y}_{\text{seen}})^{(x)}$. The complete loss function is $L = L^{(s)} + L^{(x)}$, such that

$$L^{(d)} = \sum_{i\in B^{(d)}} \left(\lambda_{\text{c}}^{(d)} L_c(\mathbf{z}_i, \phi(y_i)) + \lambda_{\text{cls}}^{(d)} L_{\text{cls}}(\hat{\mathbf{y}}_i^{(s)}, \hat{\mathbf{y}}_i^{(x)}, \mathbf{y}_i)\right), \quad (6)$$

where $B^{(d)}$ denotes indices of the samples from domain $d \in \{\mathcal{S}, \mathcal{X}\}$ in the mini-batch. Hyperparameters $\lambda_{\text{c}}, \lambda_{\text{cls}}$ are used to balance the different losses. At test time, we can simply use the logits $\hat{\mathbf{y}} = F(\mathbf{x}) \cdot \phi(\mathbf{Y})^{(s)}$ for classification.

5. Experiments

Experimental Setup. All of our experiments are developed using PyTorch [32] and trained with Nvidia GTX 1080 GPUs. For our contrastive encoding of illustrations we use a ResNet-18 [18] pretrained on ImageNet to create the embeddings $\phi(y)$ from the illustrations. As is commonly done in ZSL literature, features from a pretrained ResNet-101 backbone without fine-tuning were used to obtain a 2048-dimensional vector representation of each image.

The PA experiments used a ResNet-101 pretrained on ImageNet data and used the Adam optimizer [25] with a base learning rate of 10^{-4}, and the convolutional layers' learning rate scaled down by 0.1. All experiments on the iNaturalist datasets ran for 40.000 iterations and experiments on CUB for 200 epochs. We set $\tau = 0.1$ in all our experiments. We retrained all baselines using their respective original implementations.

Following the convention in GZSL literature, we evaluate the performance of each algorithm using held out sets of samples of the seen classes (S) and unseen classes (U) separately. The harmonic mean of these two numbers (H) is also reported. We will make our all our code available for reproducibility.

5.1. Zero-Shot Recognition, iNaturalist 2017 and 2021

We introduce the first results with ZSL leveraging the illustrations from Billow and the images from iNaturalist datasets. We report experiments using CE with TFVAEGAN [29] in a two-stage approach, and experiments using Billow illustrations directly with PA. On all iNaturalist datasets we observed an improved performance of PA over the CE. This was consistent on all three datasets evaluated on all top-k metrics. With PA we observed a harmonic mean H@top-5 of 35.1% and 35.6% for iNat2021 and its iNat2021mini, respectively (see Tab. 2a). For CE we observed a decreased performance with the larger training dataset for iNat2021 (H@top-5 19.1% and 24.6%). These results indicate that further regularization may be needed for large datasets.

Table 2b shows that the hierarchical distance to the nearest seen classes correlates strongly with performance on the unseen datasets. Similar as previously observed, CE had a decreased performance with respect to PA. This was consistent over all i-hop sets. We also evaluated performance at different hierarchy levels and found a similar behaviour. See the Appendix for details and further analysis. This is aligned with what has been observed in ImageNet for ZSL [14, 22, 31]. However, it seems that ZSL on ImageNet is more challenging than for iNaturalist, perhaps because the label distances between ImageNet classes are not as meaningful as taxonomic distances between species.

Analysis of the number of synthetic samples. Generative approaches in GZSL generate synthetic examples for unseen classes, which is usually set to $N_{\text{syn}} = 100$ in works CUB [17, 29, 45]. The samples are added to the training set to supervise the training of a classifier. We investigate the effect that bigger values of N_{syn} could have on larger datasets with higher numbers of unseen classes. We kept all other hyperparameters constant. Results can be found in Table 3. We observe that with TFVAEGAN increasing N_{syn} results in better performance for unseen classes in all the datasets, but at the cost of lower performance for seen classes.

5.2. Zero-Shot Recognition, CUB

In addition, we compare our CE and PA proposed methods using CUB_{196}, which contains 196 CUB classes also contained in Billow, divided into 148 seen and 48 unseen classes. We always respect the proposed split by [48]. Class embedding vectors were generated from illustrations using our Contrastive Encoding. These embeddings were used in combination with TFVAEGAN [29], CE-GZSL [17], and LsrGAN [45] to evaluate their performance as class side information $\phi(y)$ in a ZSL setting. In Table 4a (top) we observe that the best results with CE are obtained in combination with TFVAEGAN.

In Table 4a (bottom) we present an evaluation of various supervised and unsupervised domain adaptation methods for ZSL. This was tested with DANN [16], MDD [53], MCC [21], ProtoDA [50] and CCSA [28]. Although DANN

(a) Seen (S), unseen (U) and harmonic mean (H) top-k accuracy. Average of 5 runs ± standard deviation.

	top-1			top-5			top-10		
Model	S	U	H	S	U	H	S	U	H
iNat2017									
CE	**33.1** ± 0.8	2.6 ± 0.2	4.7 ± 0.3	**57.5** ± 1.3	14.1 ± 0.3	22.6 ± 0.3	**66.3** ± 1.5	23.6 ± 0.2	34.8 ± 0.3
PA	23.0 ± 0.3	**8.8** ± 0.4	**12.8** ± 0.5	51.9 ± 0.4	**23.4** ± 0.8	**32.3** ± 0.8	63.8 ± 0.5	**32.9** ± 0.6	**43.5** ± 0.6
iNat2021mini									
CE	**24.2** ± 0.2	3.9 ± 0.2	6.7 ± 0.3	**46.3** ± 0.1	16.7 ± 0.4	24.6 ± 0.5	56.4 ± 0.4	26.5 ± 0.4	36.1 ± 0.3
PA	20.8 ± 0.4	**12.7** ± 0.4	**15.7** ± 0.2	46.1 ± 0.5	**29.0** ± 0.4	**35.6** ± 0.2	**56.8** ± 0.4	**38.5** ± 0.5	**45.9** ± 0.3
iNat2021									
CE	**36.6** ± 0.8	2.1 ± 0.1	3.9 ± 0.2	**61.1** ± 0.6	11.3 ± 0.4	19.1 ± 0.7	**69.7** ± 0.3	19.6 ± 0.3	30.6 ± 0.4
PA	20.9 ± 0.3	**12.2** ± 0.3	**15.4** ± 0.2	45.5 ± 0.2	**28.6** ± 0.6	**35.1** ± 0.5	56.6 ± 0.2	**37.8** ± 0.5	**45.3** ± 0.4

(b) Unseen n-hop validation sets top-k accuracy. Average of 5 runs.

	top-1				top-5				top-10			
Model	1-hop	2-hop	3-hop	4-hop	1-hop	2-hop	3-hop	4-hop	1-hop	2-hop	3-hop	4-hop
iNat2017												
CE	2.3	3.4	2.9	1.6	21.3	16.9	11.4	7.4	35.1	27.8	19.0	13.8
PA	**9.1**	**9.9**	**9.3**	**7.0**	**29.1**	**25.3**	**22.2**	**18.1**	**42.3**	**35.4**	**30.5**	**25.1**
iNat2021mini												
CE	5.2	4.0	3.6	2.3	22.9	16.6	15.1	10.7	35.0	26.3	24.2	18.6
PA	**12.8**	**13.6**	**11.5**	**11.8**	**33.5**	**30.3**	**25.9**	**23.8**	**44.7**	**40.0**	**34.7**	**31.2**
iNat2021												
CE	2.6	1.9	2.1	1.6	16.5	11.1	9.4	7.1	27.3	19.6	16.4	13.3
PA	**12.3**	**13.3**	**11.4**	**10.6**	**33.9**	**29.7**	**25.5**	**23.0**	**44.8**	**39.2**	**33.6**	**30.4**

Table 2. GZSL on iNaturalist Datasets with Billow. CE: Contrastive Encoding of illustrations and TFVAEGAN. PA: Prototype Alignment. Best method is marked in bold.

	top-1			top-5			top-10		
N_{syn}	S	U	H	S	U	H	S	U	H
iNat2017									
100	**34.0**	1.2	2.3	**57.7**	7.4	13.1	66.0	13.6	22.5
1000	33.1	2.6	4.7	57.5	14.1	22.6	**66.3**	23.6	34.8
3000	32.8	**3.1**	**5.7**	56.8	**16.1**	**25.1**	66.1	**26.0**	**37.3**
iNat2021mini									
100	**25.7**	1.7	3.2	**48.1**	9.2	15.4	**57.6**	16.7	25.9
1000	24.2	3.9	6.7	46.3	**16.7**	24.6	56.4	26.5	36.1
3000	23.1	**4.9**	**8.0**	45.3	18.6	**26.4**	55.4	**28.7**	**37.8**
iNat2021									
100	**39.2**	0.8	1.5	**62.0**	4.4	8.2	69.5	7.9	14.2
1000	36.6	2.1	3.9	61.1	11.3	19.1	69.7	19.6	30.6
3000	35.8	**3.1**	**5.8**	61.0	**14.8**	**23.8**	**69.8**	**24.1**	**35.8**

Table 3. GZSL on iNaturalist Datasets with Billow. Results with Contrastive Encoding and TFVAEGAN with different number of synthetic samples. Average of 5 runs, best method is bold. Seen, unseen and harmonic mean (H) Top-k accuracy

and ProtoDA did not completely collapse towards the seen classes, they fail to fully translate knowledge from the source domain into the target domain. Our PA approach on the other hand achieves the best performance, well above that of domain adaptation baselines and the CE approach.

Furthermore, we compared CE encodings of Billow illustrations with other types of side information in Table 4b using CUB$_{191}$, i.e., the subset of 191 CUB classes overlapping with other types of side-information and Billow, divided into 145 seen and 46 unseen classes. As in the previous experiment, the split proposed by [48] is respected and the class embedding vectors were generated from illustrations using our Contrastive Encoding. We used these embeddings in combination with TFVAEGAN, CE-GZSL and LsrGAN. We compare Billow with the following sources of side-information $\phi(y)$: binary attributes [47], visual descriptions [34], DNA [5], and word2vec [3]. These experiments show that the representation power of Billow's contrastive embedding is comparable to that of word2vec and DNA embeddings. In terms of comparison among the existing methods we can observe that TFVAEGAN achieves the best results in both scenarios.

5.3. Ablation Studies

In Table 5 we show the effect of changing different components of the PA method. First, we observed that setting the projection head h to be a small Multi Layer Perceptron decreased our performance compared to an identity function (row D). We speculate that the latent space of $\mathbf{z}$ is already too close to the label domain for it to benefit from a projection head. We computed $\hat{\mathbf{y}}$ with a learned linear classifier w_{cls} instead of using the dot product between domain embeddings and observed such modification drastically re-

(a) Experiments with Billow on CUB_{196}. Top: CE Billow (Contrastive embeddding of Billow, ours), combined with GZSL methods. Bottom: End-to-end methods to use Billow, including PA (Prototype Alignment, ours) and domain adaptation methods. † denotes UDA methods that do not use target labels

$\phi(y)$	Model	S	U	H
CE Billow (ours)	CE-GZSL	42.0 ±1.1	25.2 ±1.5	31.5 ±1.2
	LsrGAN	**69.7** ±0.3	6.4 ±0.5	11.6 ±0.9
	TFVAEGAN	45.5 ±13.1	**31.5** ±5.5	**35.8** ±1.2
Billow (end-to-end)	DANN†	24.3 ±1.8	17.5 ±2.3	20.3 ±1.6
	MDD†	1.4 ±0.4	0.7 ±0.4	0.9 ±0.4
	MCC†	6.5 ±0.5	5.8 ±0.8	6.1 ±0.4
	ProtoDA	13.8 ±0.9	13.8 ±1.8	14.4 ±2.0
	CCSA	**73.5** ±0.7	0.1 ±0.0	0.1 ±0.1
	PA (ours)	69.7 ±0.6	**36.1** ±1.5	**47.5**±1.5

(b) Experiments with Billow on CUB_{191}. Comparison with other types of side-information ($\phi(y)$) used with CUB.

$\phi(y)$	Model	S	U	H
Binary attributes	CE-GZSL	59.8 ± 1.9	48.4 ± 0.7	53.5 ± 0.7
	LsrGAN	**63.6** ± 0.2	20.4 ± 0.5	30.9 ± 0.6
	TFVAEGAN	63.4 ± 2.2	**52.8** ± 1.4	**57.6** ± 0.2
Visual descriptions	CE-GZSL	66.4 ± 0.3	65.0 ± 0.6	65.7 ± 0.4
	LsrGAN	58.7 ± 0.3	54.2 ± 0.8	56.3 ± 0.4
	TFVAEGAN	**67.8** ± 2.1	**68.4** ± 2.1	**68.1** ± 0.4
DNA	CE-GZSL	39.5 ±1.2	13.5 ±0.8	20.1 ±0.8
	LsrGAN	**69.7** ±0.1	3.9 ±0.2	7.4 ±0.4
	TFVAEGAN	30.8 ±0.4	**20.3** ±1.0	**24.5** ±0.7
word2vec	CE-GZSL	49.1 ±1.7	25.9 ±0.7	33.9 ±0.5
	LsrGAN	**62.0** ±0.5	16.5 ±0.4	26.1 ±0.5
	TFVAEGAN	45.6 ±1.0	**27.2** ±0.9	**34.1** ±0.9
CE Billow (ours)	CE-GZSL	42.7 ±1.5	27.9 ±0.8	33.8 ±1.0
	LsrGAN	**69.2** ±0.2	7.0 ±0.2	12.7 ±0.4
	TFVAEGAN	45.3 ±14.1	**31.6** ±5.4	**35.6** ±1.1

Table 4. GZSL on CUB. Seen, unseen and harmonic mean (H) Top-1 accuracy. Average of 5 runs ± standard deviation. Best method for each dataset and $\phi(y)$ is bold.

	λ_c	$h(x)$	λ_{ce}^t	w_{cls}	$L_{cls}(\hat{y}', y)$	H	S	U
A	1	Identity	0.1	learned		14.7	44.1	8.9
B	1	Identity	0.1	$\phi(Y)^{(s)}$		42.7	50.6	37.0
C	0	Identity	0.1	$\phi(Y)^{(s)}$	✓	23.5	52.4	15.2
D	1	MLP	0.1	$\phi(Y)^{(s)}$	✓	39.3	50.1	32.3
E	1	Identity	1.0	$\phi(Y)^{(s)}$	✓	46.3	69.9	34.6
F	1	Identity	0.1	$\phi(Y)^{(s)}$	✓	47.5	69.7	36.1

Table 5. Prototype Alignment experiments with different Hyperparameters on CUB dataset

duces the performance on the unseen classes (row A). The contrastive in-domain loss λ_c proved itself to be essential for achieving a good performance on the unseen classes (row C). The classification loss instead seems to be more important for the recognition of the seen classes: Removing the term completely barely changes the performance on the unseen classes while drastically reducing performance on the seen ones (row B). On the other hand having a larger $\lambda_{ce}^{(t)}$ tends to boost accuracy in seen classes with a slight reduced accuracy in unseen ones (row E). See supplementary for an ablation with different backbones.

6. Conclusion

Our experiments show that using field guides as side information for ZSL is feasible, expanding the set of fine-grained ZSL experiments to datasets with more natural distributions such as iNaturalist2017 and iNaturalist2021. Which are of a much larger scale than those commonly studied for ZSL (e.g., CUB, Animals with Attributes [48], Oxford Flowers [30]). They show that the zero-shot recognition of bird species in images is feasible with an accuracy much better than random chance. The best harmonic mean so far is obtained by the proposed PA method. Our Experiments show that a naïve implementation from domain adaptation might not yield the best results, despite a comparatively small domain gap w.r.t. photographs. iNaturalist experiments show that, while state-of-the-art ZSL combined with the contrastive encoded illustrations achieves reasonable results, our proposed PA consistently outperforms it. Still, identifying unseen birds across thousands of different species remains a challenge. The observed top-10 accuracies demonstrate that side-information provided by Billow indeed steers the classifier towards the correct (unseen) species. This is also reflected by the fact that we observe better performance for unseen classes that are closer to seen ones in the taxonomic hierarchy (Tab. 2b).

CUB has been used in combination with many types of side-information in the past. Our experiments show that leveraging illustrations in field guides can achieve comparable results to other types of side-information. Although attributes and keywords have higher accuracies on CUB than with Billow, illustrations are a valid alternative that contains several decades of knowledge that be realistically exploited. It appears more natural to describe new bird species using existing illustrations of them than comparing them to the test set of CUB to obtain visual descriptions [34], which is prone to overfit to the rather small dataset.

Species recognition would benefit from further studies on how to incorporate more side-information, such as by explicitly modelling species similarity and patristic distances [22]; or obtain a multi-source embedding using a mixture of illustrations, text descriptions or other types of side-information. More fine-grained class representation for different sexes of the same bird species could be could further improve the results. While we have focused this work on illustrations of birds, there are many other field guides that could be exploited in ZSL. We hope that our work inspires more research in this direction to assist efforts in biodiversity mapping and conservation.

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In IEEE Conference on Computer Vision and Pattern Recognition, 2013. 2

[2] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(7):1425–1438, 2015. 3

[3] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In IEEE Conference on Computer Vision and Pattern Recognition, 2015. 3, 7

[4] Stanislaw Antol, C Lawrence Zitnick, and Devi Parikh. Zero-shot learning via visual abstraction. In European Conference on Computer Vision, 2014. 3

[5] Sarkhan Badirli, Zeynep Akata, George Mohler, Christine Picard, and Mehmet Dundar. Fine-grained zero-shot learning with dna as side information. Advances in Neural Information Processing Systems, 2021. 3, 4, 7

[6] Bjorn Barz and Joachim Denzler. Deep learning on small datasets without pre-training using cosine loss. In IEEE/CVF Winter Conference on Applications of Computer Vision, 2020. 4

[7] SM Billerman, BK Keeney, PG Rodewald, and TS Schulenberg. Birds of the world. cornell laboratory of ornithology, 2020. 1, 4

[8] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE. arXiv preprint arXiv:1804.03599, 2018. 5

[9] H Daume. Frustratingly easy domain adaptation. In Annual Meeting of the Association for Computational Linguistics, 2007, 2007. 3

[10] S Díaz, J Settele, E Brondízio, H Ngo, M Guèze, J Agard, A Arneth, P Balvanera, K Brauman, S Butchart, K Chan, L Garibaldi, K Ichii, J Liu, S Subramanian, G Midgley, P Miloslavich, Z Molnár, D Obura, A Pfaff, S Polasky, A Purvis, Jona Razzaque, B Reyers, R Chowdhury, Y Shin, I Visseren-Hamakers, K Willis, and C Zayas. Summary for policymakers of the global assessment report on biodiversity and ecosystem services. Technical report, Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services, 2019. 1

[11] Seppo Fagerlund. Bird species recognition using support vector machines. EURASIP Journal on Advances in Signal Processing, 2007:1–8, 2007. 1

[12] Elizabeth J. Farnsworth, Miyoko Chu, W. John Kress, Amanda K. Neill, Jason H. Best, John Pickering, Robert D. Stevenson, Gregory W. Courtney, John K. VanDyk, and Aaron M. Ellison. Next-Generation Field Guides. BioScience, 63(11):891–899, 2013. 2

[13] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. Advances in Neural Information Processing Systems, 2013. 2

[14] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc' Aurelio Ranzato, and Tomas Mikolov. DeViSE: A deep visual-semantic embedding model. In Advances in Neural Information Processing Systems, 2013. 4, 6

[15] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In International Conference on Machine Learning, 2015. 3

[16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. Journal of Machine Learning Research, 17(59):1–35, 2016. 3, 6

[17] Zongyan Han, Zhenyong Fu, Shuo Chen, and Jian Yang. Contrastive embedding for generalized zero-shot learning. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 3, 6

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In IEEE Conference on Computer Vision and Pattern Recognition, 2016. 6

[19] Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In International Conference on Learning Representations, 2017. 5

[20] iNaturalist. https://www.inaturalist.org. California Academy of Sciences & National Geographic Society, 2011. Accessed: 26-05-2021. 1

[21] Ying Jin, Ximei Wang, Mingsheng Long, and Jianmin Wang. Minimum class confusion for versatile domain adaptation. In European Conference on Computer Vision, 2020. 3, 6

[22] Michael Kampffmeyer, Yinbo Chen, Xiaodan Liang, Hao Wang, Yujia Zhang, and Eric P Xing. Rethinking knowledge graph propagation for zero-shot learning. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 6, 8

[23] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In Advances in Neural Information Processing Systems, 2020. 5

[24] Donghyun Kim, Kuniaki Saito, Tae-Hyun Oh, Bryan A Plummer, Stan Sclaroff, and Kate Saenko. Cross-domain self-supervised learning for domain adaptation with few source labels. arXiv preprint arXiv:2003.08264, 2020. 3, 5

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014. 6

[26] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In IEEE Conference on Computer Vision and Pattern Recognition, 2009. 2

[27] Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In AAAI Conference on Artificial Intelligence, 2008. 3

[28] Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In IEEE International Conference on Computer Vision, 2017. 3, 6

[29] Sanath Narayan, Akshita Gupta, Fahad Shahbaz Khan, Cees GM Snoek, and Ling Shao. Latent embedding feedback and discriminative features for zero-shot classification. In European Conference on Computer Vision, 2020. 3, 6

[30] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, pages 722–729. IEEE, 2008. 8

[31] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. arXiv preprint arXiv:1312.5650, 2013. 4, 6

[32] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In Advances in Neural Information Processing Systems, Workshops, 2017. 6

[33] Nirosha Priyadarshani, Stephen Marsland, Julius Juodakis, Isabel Castro, and Virginia Listanti. Wavelet filters for automated recognition of birdsong in long-time field recordings. Methods in Ecology and Evolution, 11(3):403–417, 2020. 1

[34] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In IEEE Conference on Computer Vision and Pattern Recognition, 2016. 3, 7, 8

[35] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. ACM Transactions on Graphics, 35(4):1–12, 2016. 3

[36] Yuming Shen, Jie Qin, Lei Huang, Li Liu, Fan Zhu, and Ling Shao. Invertible zero-shot recognition flows. In European Conference on Computer Vision, 2020. 3

[37] Yutaro Shigeto, Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, and Yuji Matsumoto. Ridge regression, hubness, and zero-shot learning. In Machine Learning and Knowledge Discovery in Databases, 2015. 2

[38] Dan Stowell, Michael D Wood, Hanna Pamuła, Yannis Stylianou, and Hervé Glotin. Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge. Methods in Ecology and Evolution, 10(3):368–380, 2019. 1

[39] Brian L Sullivan, Christopher L Wood, Marshall J Iliff, Rick E Bonney, Daniel Fink, and Steve Kelling. eBird: A citizen-based bird observation network in the biological sciences. Biological Conservation, 142(10):2282–2292, 2009. 1

[40] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In AAAI Conference on Artificial Intelligence, 2016. 3

[41] Baochen Sun and Kate Saenko. Deep CORAL: Correlation alignment for deep domain adaptation. In European Conference on Computer Vision Workshops, 2016. 3

[42] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In IEEE Conference on Computer Vision and Pattern Recognition, 2017. 3

[43] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 1, 4

[44] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In IEEE Conference on Computer Vision and Pattern Recognition, 2018. 4

[45] Maunil R. Vyas, Hemanth Venkateswara, and Sethuraman Panchanathan. Leveraging seen and unseen semantic relationships for generative zero-shot learning. In European Conference on Computer Vision, 2020. 3, 6

[46] Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. Generalizing from a few examples: A survey on few-shot learning. ACM Computing Surveys, 53(3), 2020. 2

[47] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 4, 7

[48] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning – a comprehensive evaluation of the good, the bad and the ugly. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(9):2251–2265, 2019. 2, 4, 6, 7, 8

[49] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In IEEE Conference on Computer Vision and Pattern Recognition, 2018. 3

[50] Xiangyu Yue, Zangwei Zheng, Shanghang Zhang, Yang Gao, Trevor Darrell, Kurt Keutzer, and Alberto Sangiovanni Vincentelli. Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 3, 5, 6

[51] Zhongqi Yue, Tan Wang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Counterfactual zero-shot and open-set visual recognition. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 3

[52] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In IEEE Conference on Computer Vision and Pattern Recognition, 2017. 2

[53] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In International Conference on Machine Learning, 2019. 3, 6