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-1and* 38% *@top-10, which shows the potential of field guides for challenging real-world scenarios with many species*

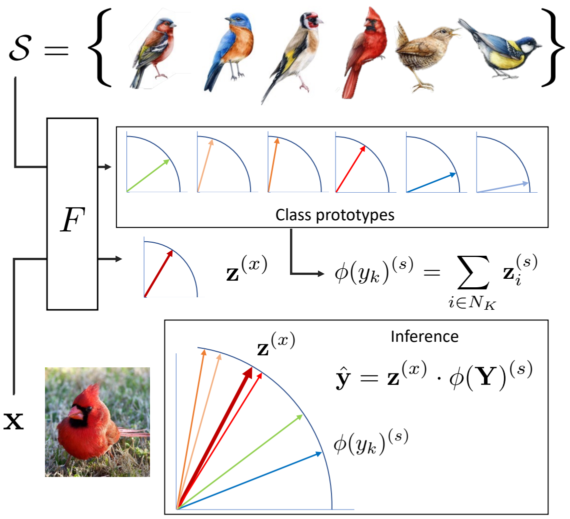
$\mathcal{S} = \{ \quad \}$
F
Class prototypes
$\mathbf{z}^{(x)}$
$\phi(y_k)^{(s)} = \sum_{i \in N_K} \mathbf{z}_i^{(s)}$
Inference
$\mathbf{z}^{(x)}$
$\hat{\mathbf{y}} = \mathbf{z}^{(x)} \cdot \phi(\mathbf{Y})^{(s)}$
$\phi(y_k)^{(s)}$
$\mathbf{x}$

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.

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 [6].

A bottleneck for automatic species recognition in the wild has long been the collection of enough observations. 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 [12] and eBirds [21] projects. The eBirds platform alone has accumulated $>$34 million images for bird species, from $\approx$800'000 contributors. Those databases make it possible to train automatic species recognition systems, which would be a valuable asset for scalable biodiversity monitoring.

When data collection is limited, one can use few-shot learning if only few labelled examples are available for certain classes [25]. 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, 8, 16, 27]. 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

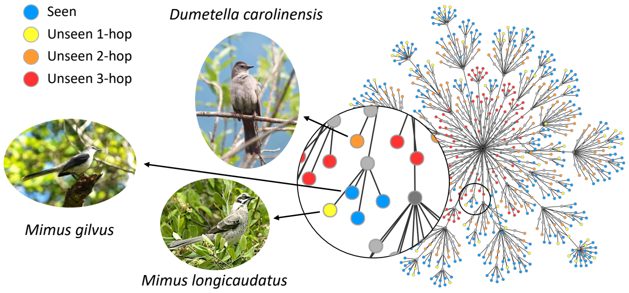

Figure 2. Hierarchical representation of the *Passeriformes* order of the iNat2021 dataset for Seen and i-hop unseen classes.

mobile apps to aid species recognition [7], 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.

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.

2. 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 [5] collected and organized by the Cornell Lab of Ornithology (CLO). 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. CLO granted us consent to use the illustrations for this research. The digital licence of the original artworks is owned by the CLO. The artworks may be accessed with a valid subscription to the *Birds of the World* project.

We use Billow with three datasets, namely Caltech-UCSD Birds-200-2011 (CUB) [26] and the bird subsets of iNaturalist 2017 [23] and iNaturalist 2021 [22]. 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 [27]. 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 [4], or to utilize genetic information [3]. We have revised these assignments, and only retain mappings for which we found a one-to-one correspondence between the common and scientific name. We obtained a match in Billow for 196 out of the 200 CUB classes.

For the iNaturalist datasets, we propose a seen/unseen split. Similar to previous ZSL work that uses ImageNet [9, 19], we construct several groups of *unseen* classes, which have different distances to *seen* classes in the label hierarchy. 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 super-class at the i-th taxonomic level). For example, the classes in the 2-hop set share the family (2$^{\text{nd}}$ 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.

3. Method

In order to utilise these illustrations for ZSL, we explore two different strategies. We first explore a two-stage strategy, where we first learn a *Contrastive eEncoding* (CE) of the illustrations that allows us to feed the codes into existing ZSL methods at a second stage. See the full version of our work for more details on this approach. We then develop a more specialized method, named *Prototype Alignment* (PA), 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.

In contrast to other types of side information for ZSL, illustrations also belong to the visual domain. We leverage this property and propose PA for ZSL with visual side information, which allows us to bypass the separate encoding step required to use illustrations with previous ZSL-methods. Inspired by [28], we explore a view of the problem through the lens of few-shot *domain adaptation*: The source domain are illustrations, the target domain are 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 $\eta(\cdot)$. 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 [15, 28], 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 in the following where possible. In every iteration, we update the memory bank in each domain $\phi(y_k)$ with the latent representation of the new samples, with momentum m with $\phi(y_k) \leftarrow \eta\big((1-m)\mathbf{z}_k + m\phi(y_k)\big)$. To promote compact and discriminative class representations, we apply a contrastive in-domain loss via a projection head h:

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

In contrast to [28] we refrain from applying a cross-domain contrastive loss to close the domain gap. Instead, we sidestep the gap by 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 class embeddings $\phi(\mathbf{Y})$ 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:

$$L_{CE}\left(\hat{\mathbf{y}}^{(s)}, \hat{\mathbf{y}}^{(x)}, \mathbf{y}\right) = L_{\text{cls}}\left(\hat{\mathbf{y}}^{(s)}, \mathbf{y}\right) + L_{\text{cls}}\left(\hat{\mathbf{y}}^{(x)}, \mathbf{y}\right). \tag{2}$$

Eq. 2 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 2 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

$$L^{(d)} = \sum_{i \in B^{(d)}} \left(\lambda_c^{(d)} L_c(\mathbf{z}_i, \phi(y_i)) + \lambda_{ce}^{(d)} L_{CE}(\hat{\mathbf{y}}_i^{(s)}, \hat{\mathbf{y}}_i^{(x)}, \mathbf{y}_i)\right), \tag{3}$$

where $B^{(d)}$ denotes indices of the samples from domain $d \in \{\mathcal{S}, \mathcal{X}\}$ in the mini-batch. Hyperparameters λ_c, λ_{ce} 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.

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

	top-1			top-10		
Model	S	U	H	S	U	H
iNat2017						
CE	**33.1** ± 0.8	2.6 ± 0.2	4.7 ± 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	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	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	**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	**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	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-10			
N-hop	1	2	3	4	1	2	3	4
iNat2017								
CE	2.3	3.4	2.9	1.6	35.1	27.8	19.0	13.8
PA	**9.1**	**9.9**	**9.3**	**7.0**	**42.3**	**35.4**	**30.5**	**25.1**
iNat2021mini								
CE	5.2	4.0	3.6	2.3	35.0	26.3	24.2	18.6
PA	**12.8**	**13.6**	**11.5**	**11.8**	**44.7**	**40.0**	**34.7**	**31.2**
iNat2021								
CE	2.6	1.9	2.1	1.6	27.3	19.6	16.4	13.3
PA	**12.3**	**13.3**	**11.4**	**10.6**	**44.8**	**39.2**	**33.6**	**30.4**

Table 1. GZSL on iNaturalist Datasets with Billow. CE: Contrastive Encoding of illustrations and TFVAEGAN. PA: Prototype Alignment. Best method is bold.

4. Results

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.

4.1. Zero-Shot Recognition, iNaturalist 2017 and 2021

We first explore using Billow illustrations for ZSL using the iNaturalist 2017 and 2021 datasets. We report experiments using CE with TFVAEGAN [18] 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-10 of 45.3% and 45.9% for iNat2021 and its iNat2021mini, respectively (see

Tab. 1a). For CE we observed a decreased performance with the larger training dataset for iNat2021 (H@top-10 30.6% and 36.1%). These results indicate that further regularization may be needed for large datasets.

Table 1b shows that the hierarchical distance to the nearest seen classes correlates strongly with performance on the unseen datasets. In line with previous results, CE showed a decrease in performance with respect to PA. This was consistent over all i-hop sets. This is aligned with what has been observed in ImageNet for ZSL [9, 14, 19]. However, it seems that ZSL on ImageNet is more challenging than in iNaturalist as the label distance on ImageNet classes might not be as meaningful as with taxonomic distances of species.

4.2. Zero-Shot Recognition, CUB

We also compare our CE and PA methods using CUB$_{196}$, which contains the 196 CUB classes also contained in Billow, divided into 148 seen and 48 unseen classes, respecting the split proposed in [27]. Class embedding vectors were generated from illustrations using CE. These embeddings were used in combination with TFVAEGAN [18], CE-GZSL [11], and LsrGAN [24] to evaluate their performance as class side information $\phi(y)$ in a ZSL setting. In Table 2a (top) we observe that the best results with CE are obtained in combination with TFVAEGAN.

In Table 2a (bottom) we present an evaluation of various supervised and unsupervised domain adaptation methods for ZSL. This was tested with DANN [10], MDD [29], MCC [13], ProtoDA [28] and CCSA [17]. Although DANN 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 2b using CUB$_{191}$, 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 [27] 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 [26], visual descriptions [20], DNA [3], and word2vec [2]. 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.

(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

$\phi(y)$	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 2. 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.

5. 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. 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.

Species recognition would benefit from further studies on how to incorporate information from a taxonomic tree into the method to improve performance, such as by explicitly modelling species similarity and patristic distances [14]; or modelling different sexes of the same bird species separately. While we have focused this work on illustrations of birds, there are many other field guides that could potentially 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. 1

[2] 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. 4

[3] 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. 2, 4

[4] 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. 2

[5] SM Billerman, BK Keeney, PG Rodewald, and TS Schulenberg. Birds of the world. cornell laboratory of ornithology, 2020. 2

[6] 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

[7] 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

[8] 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. 1

[9] 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. 2, 4

[10] 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. 4

[11] 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. 4

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

[13] Ying Jin, Ximei Wang, Mingsheng Long, and Jianmin Wang. Minimum class confusion for versatile domain adaptation. In *European Conference on Computer Vision*, 2020. 4

[14] 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. 4

[15] 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

[16] 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. 1

[17] 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. 4

[18] 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, 4

[19] 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. 2, 4

[20] 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. 4

[21] 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

[22] 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. 2

[23] 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. 2

[24] 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. 4

[25] 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. 1

[26] 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. 2, 4

[27] 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. 1, 2, 4

[28] 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, 4

[29] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, 2019. 4