7. Appendix

7.1. Encoding of Billow

Appropriate encoding of the side information can have a strong effect in the overall performance. We evaluated different ways to encode our illustrations to be used with ZSL state-of-the-art methods. From Billow, we created a separate test set and evaluated the predictive power for species classification of each encoding. This is a challenging task, since we only have a couple of illustrations (one male; one female; and , in some cases, a head detail) per species. In contrast, traditional methos usually require a large amount of images per species to train be able to perform automatic classification.

To evaluate the quality of our encoding method we measured how well can we predict the class at different hierarchical levels. Out of the 10'631 we take all the species whose genus has at least 5 species in the dataset and create a train/validation split from them. This results in 18'489 illustrations: 13362 for training, 1908 for validation and 3219 for testing. In all splits combined there are samples of 8646 species, 956 genus, 175 families and 33 orders. We evaluate top-1 and top-10 accuracy on all 4 hierarchical.

For training we explored Variational Auto-Encoder (VAE) generative models to encode our dataset. VAEs consist of two networks, an encoder E and a decoder D. A regular auto-encoder would simply use the output of the encoder and feed it to the decoder to reconstruct the input. For training the auto-encoder, the reconstruction loss is defined as $L = d(x, D(E(x)))$, where d is usually a Euclidean distance between the input and its reconstruction. VAEs assume a prior on the output of the encoder p_z and maximize the log-likelihood of the reconstruction produced by D over the entire prior distribution p_z.

Modelling such a distribution would be desirable in our case as we will use the embedding and the distances between them for ZSL. Hence we explore two VAE variants: β-VAE [8, 19] and VQ-VAE [34, 44]. Our motivation to test these 2 variations of VAE is to explore the effect of different priors on the latent distribution $p \sim z$. β-VAE uses a more constrained information bottleneck on the embedding z than vanilla-VAE (i.e. $\beta > 1$) to obtain a disentangled representation z. As a reconstruction, we slowly increase the bottleneck capacity over training as proposed by [8]. VQ-VAE on other hand imposes a discrete distribution over the embedding z, this allows to control the information bottle neck by imposing a very small dimension on the discrete distribution.

Additionally we fine-tuned a ResNet classifier pretrained on ImageNet. The classifier was supervised by $L = L_{cls} + L_{cont}$, where L_{cont} is as defined in Eq. 1. VAE experiments were trained their corresponding reconstruction loss and the supervision loss L_{cls}.

Once the model was trained we evaluated its predictive power by feeding the embedding z into a small Multi-layer Perceptron network and trained to predict the level of the label k. The results on the test set in Table 6 show that β-VAE does not perform close to the ResNet models. Along with our fine-tuned ResNet-18, we include results on pretrained modesl on Image-Net without fine-tuning the backbone. As expected larger networks achieved higher performance, but our fine-tuning on ResNet-18 improved performance in the most challenging case of fine-grained species recognition. Our ResNet-18 already achieved 100% accuracy on the training set and observed over-fitting on the validation set after fine-tuning. For this reason we decided to keep the fine-tuned ResNet18 as the default encoder for billow.

Method	top-1
VQ-VAE	0.1
β-VAE	14.2
ResNet-101*	15.5
ResNet-50*	12.0
ResNet-18*	16.5
ResNet-18 (ours)	**17.7**

Table 6. Top-1 species accuracy on a test set of Billow samples with different encoders. * indicates models without fine-tuning

Level	1-hop	2-hop	3-hop	4-hop
iNat2017				
species	9.1 ± 0.4	9.9 ± 0.5	9.3 ± 0.4	7.0 ± 0.7
genus	29.4 ± 0.5	10.2 ± 0.5	12.9 ± 0.6	11.0 ± 1.4
family	50.2 ± 0.7	36.7 ± 0.9	12.9 ± 0.6	16.8 ± 2.6
order	77.8 ± 0.7	67.7 ± 0.6	69.7 ± 1.4	16.8 ± 2.6
iNat2021mini				
species	12.8 ± 0.5	13.6 ± 0.4	11.4 ± 0.8	11.6 ± 0.4
genus	29.3 ± 0.4	13.7 ± 0.4	17.1 ± 1.1	16.2 ± 0.7
family	47.4 ± 0.8	38.3 ± 0.4	17.4 ± 1.2	20.6 ± 1.1
order	75.1 ± 1.0	69.9 ± 1.0	69.0 ± 1.3	20.6 ± 1.1
iNat2021				
species	12.3 ± 0.3	13.4 ± 0.6	11.2 ± 0.3	10.6 ± 0.4
genus	28.8 ± 0.7	13.5 ± 0.6	16.8 ± 0.6	15.7 ± 0.7
family	46.8 ± 0.5	37.8 ± 0.6	16.9 ± 0.6	20.7 ± 0.7
order	74.7 ± 0.4	70.1 ± 0.8	69.8 ± 0.8	20.7 ± 0.7

Table 7. Unseen n-hop validation sets top-1 accuracy at different label hierarchy levels. Average of 5 runs $\pm$ standard deviation.

7.2. Accuracy at different hierarchical levels

We evaluated top-1 accuracy at different label hierarchies on each n-hop validation set (which were created using the label hierarchies, see Section 3.1 for details). A prediction is correct at the family-level if the predicted species was of the same family as the target species. We present

Table 8. Prototype Alignment and Domain adaptation baseline experiments with different ResNet backbones on iNaturalist. Top-k Accuracies

Backbone	top-1			top-5			top-10		
	S	U	H	S	U	H	S	U	H
iNat2017									
ResNet-18	13.1±0.3	7.6±0.5	9.6±0.3	34.4±0.6	20.6±0.6	25.8±0.4	46.0±0.7	28.8±0.5	35.4±0.4
ResNet-50	20.9±0.3	8.2±0.3	11.8±0.4	48.8±0.6	21.8±0.7	30.1±0.7	60.9±0.6	30.9±0.5	41.0±0.5
ResNet-101	**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
iNat2021									
ResNet-18	11.1±0.2	7.9±0.2	9.3±0.1	29.2±0.2	19.8±0.3	23.6±0.2	39.7±0.4	27.4±0.4	32.4±0.3
ResNet-50	17.9±0.5	10.5±0.2	13.2±0.3	41.4±0.5	24.8±0.5	31.0±0.5	53.2±0.5	33.8±0.5	41.3±0.5
ResNet-101	**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
iNat2021mini									
ResNet-18	11.2±0.3	8.3±0.2	9.5±0.1	29.4±0.3	20.6±0.3	24.2±0.2	40.0±0.3	28.5±0.3	33.3±0.1
ResNet-50	18.3±0.4	10.8±0.3	13.6±0.3	41.9±0.5	25.5±0.5	31.7±0.5	53.5±0.3	34.3±0.7	41.8±0.5
ResNet-101	**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

Model	S	U	H
Resnet-18			
DANN [16]	13.0 ± 1.1	10.3 ± 0.5	11.5 ± 0.5
MDD [55]	1.2 ± 0.2	0.0 ± 0.1	0.0 ± 0.1
MCC [21]	4.4 ± 0.5	3.8 ± 0.4	4.0 ± 0.2
PA	48.2 ±0.4	37.2 ±0.4	42.0 ±0.2
Resnet-50			
DANN [16]	16.3 ± 0.6	14.4 ± 1.5	15.2 ± 1.1
MDD [55]	0.6 ± 0.3	0.9 ± 1.2	0.2 ± 0.5
MCC [21]	5.8 ± 0.2	6.7 ± 0.7	6.2 ± 0.4
PA	64.8 ±0.5	**37.8** ±1.0	**47.8** ±0.8
Resnet-101			
DANN [16]	24.3 ± 1.8	17.6 ± 2.4	20.3 ± 1.6
MDD [55]	1.4 ± 0.4	0.7 ± 0.5	0.9 ± 0.4
MCC [21]	6.6 ± 0.5	5.8 ± 0.6	6.1 ± 0.4
PA	**69.9** ± 0.6	34.6 ±1.5	46.3 ±1.5

Table 9. Prototype Alignment and Domain adaptation baselines experiments with different ResNet backbones on CUB Dataset. Top-1 Accuracy

these results in Table 7. The 2-hop (set where no species of the same family are part of the seen species) performance between species and genus is similar, suggesting that the Zero-shot task is equally difficult at the 2 considered hierarchy levels. Similarly for 4-hop (no overlap at any hierarchy label), the performance is equally low at any level. These results suggest that the label hierarchical distance is a meaningful strategy for evaluation. Future work exploit this label hierarchy at training time.

7.3. Ablation: Backbone Size

We explore the effect of using different backbones using end-to-end methods, including different Domain Adaptation Baselines and PA (ours). For iNaturalist datasets in Table 8, we observe better accuracies with larger networks, and slightly higher accuracies with ResNet-101. For iNat2021mini, aligned with what we observed before, we observe better performance than iNat2020 in all cases considered in these experiments. The results on CUB dataset in Table 9 show that the performance increases for all the baselines using larger ResNets, while some of the methods have slightly higher accuracies with ResNet-50 the difference are within the marging of error compared to ResNet-101.