Noname manuscript No.
(will be inserted by the editor)

Large-scale Zero-shot Learning in the Wild: Classifying Zoological Illustrations

Lise Stork*, Andreas Weber, Jaap van den Herik, Aske Plaat, Fons Verbeek, Katherine Wolstencroft

Received: date / Accepted: date

Abstract In this paper we analyse the classification of zoological illustrations. Large archives of historical biodiversity data are stored in natural history institutions worldwide. By employing computational methods for classification, the data can be made amenable to research. The domain presents multiple challenges: the set of potential classes is large, images from only a subset of these are digitally available, and many images are unlabelled, since labelling requires domain expertise. We argue that there is a lack of research that analyses the performance of models that can cope with the aforementioned challenges for problems in the real world. Here, we explore zero-shot learning approaches to address these issues.

We first introduce a new hierarchical dataset, the *Zoological Illustration and Class Embedding* (ZICE) dataset for large-scale - covering many classes - zero-shot learning. It is the first large-scale expert dataset for zero-shot learning "in the wild", that pulls together distributed, multi-modal, auxiliary data, from a range of research institutions within the domain's community. We use the dataset to train a species embedding model using a state-of-the-art *prototypical network* for zero-shot learning. We introduce *fused prototypes* (FP) and *hierarchical prototype loss* (HPL) to optimise the model. The experiments show that the proposed approach significantly improves over the baseline. Finally, we analyse the performance of the species embedding model for use in real-world applications, exemplified by results on the ZICE dataset and an independent verification-set.

Keywords Zero-shot learning · Biodiversity · Object recognition · Fine-grained classification

*Corresponding author.
Corr. Author Address: Leiden Institute of Advanced Computer Science, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands
*E-mail:*l.stork@liacs.leidenuniv.nl

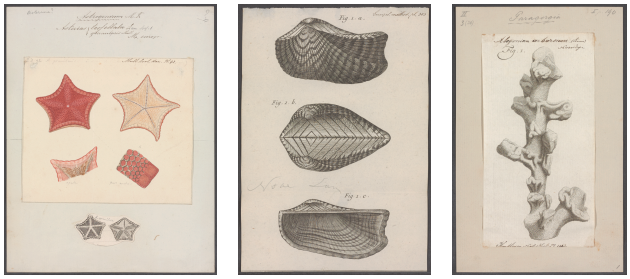


Fig. 1: Example zoological illustrations from the Iconographia Zoologica online collection[1] (best viewed in colour). Images free of known restrictions under copyright law (Public Domain Mark 1.0).

1 Introduction

Zero-shot learning (ZSL) aims to recognise objects whose instances have not yet been seen during training, based on semantic knowledge, e.g., attributes [20, 12], that are shared among seen and unseen classes. Datasets have been set up to facilitate progress in the field and demonstrate the possibilities and advantages of zero-shot learning [27, 41, 20]. We argue there is a need for research that analyses the performance of zero-shot learning models on complex real-world data, collected to fulfill a need within a certain domain. Specifically data from domains where the solution space is large and complex, and obtaining labels for training is costly or simply not feasible, e.g., [33, 39]. When algorithms are, for instance, evaluated on highly imbalanced large-scale datasets, results are poor: Xian et al. show that experiments of state-of-the-art zero-shot learning algorithms achieve only $\sim 1.3\%$ top-1 per-class accuracy on the 5,000 least populated classes in ImageNet, and only $\sim 0.4\%$ top-

[1] https://bijzonderecollecties.uva.nl/gedeelde-content/beeldbanken/iconographia.html

1 accuracy for generalised zero-shot learning (GZSL) [45], where the classifier must choose the correct class from both seen *and* unseen classes. With an increasing number of classes to choose from and less information to learn from, it becomes progressively harder for a classifier to obtain good results.

In this paper we analyse a new, sparsely populated, large-scale dataset for zero-shot learning. The dataset comes from the natural history domain, see figure 1, and is formed by consolidating data that is in use and managed by the biodiversity research community. Ultimately, automated methods can assist biodiversity experts in the formation of a global picture of historical and current biodiversity, something that is crucial given the current biodiversity crisis [17]. Our contribution is threefold:

1. We introduce the Zoological Illustration and Class Embedding dataset (ZICE) constructed from real-world data. It consists of: (i) 14,502 biological illustrations of 7973 species from the animal kingdom, with labels organised hierarchically, and (ii) class embeddings from 3 different sources - a hierarchy (a biological taxonomy), historical texts and photographs.
2. We introduce and evaluate a new zero-shot learning (ZSL) approach for fine-grained classification. We use the prototypical networks introduced by Snell et al. in [31] and introduce: *fused prototypes* (FP), and *hierarchical prototype loss* (HPL). Our approach is evaluated on the ZICE dataset.
3. We analyse the performance of our ZSL approach in a real-world scenario on an independent verification-set: a collection of 1,088 unlabelled illustrations from the animal kingdom, collected during a historical biodiversity expedition [43].

The rest of this paper is organised as follows. Section 2 describes the problem we present. In Section 3 we discuss related work on automated species classification and zero-shot learning. We discuss the data in Section 4, the methodology in Section 5, the experimental setting in Section 6 and the experiments in Section 7. We close the paper with an analysis and discussion of the results in Section 8, and our conclusions in Section 9.

2 Problem Description and Approach

Historically, the habitus illustration - a scientific illustration of a species' physical appearance - was the most important medium to convey a species' characterising traits to other scientists. In illustrations, scientists are capable of delineating and highlighting the most minuscule details, often even more so than photographs. Habitus illustrations were routinely and abundantly created and commonly served as examples for the description of newly discovered species, so-called holotypes. Additionally, they sometimes recorded the habitat or behaviour of an organism. Over the last 250 years, a large number of the earth's zoological species have been observed and documented this way, by means of expeditions to the world's most bio-diverse areas.

Research into these scientific illustrations is complicated by several challenges. First, most illustrations are stored in museum repositories and archives that are not disclosed for generic use. Digitisation projects are currently ongoing worldwide to address this challenge, but as of now, most collections remain offline [17]. Second, illustrations published as online digital collections can be used for research, but are often published online with very limited or no identifications (unique labels). To study the illustrations, the depicted organism should be annotated with a published taxonomic name. Illustrations do contain captions with handwritten *historical names*, as is demonstrated in figure 2, but these are mostly unpublished or obsolete within today's taxonomy.

In biology, taxonomy refers to the process of classifying biological organisms into taxonomic groups based on their shared characteristics. Modern names for these groups are based on a binomial nomenclature that was introduced by Carl Linnaeus in 1735. In a Linnaean taxonomy, taxon groups are hierarchically categorised into the taxonomic levels (also called *ranks*): kingdom, phylum, class, order, family, genus and species. Each level subsumes less general taxon groups, kingdom being the most general and species the most specific taxon group. The scientific name for a species serves as a unique label (once it has been published and accepted by the biodiversity community) and is formed by the binomial of the genus and species name. Over two million distinct animal species have so far been described, their names published and categorised according to a taxonomy. Consequently, the identification of an organism from a photograph or illustration, using the system of biological classification, is a complex and delicate task [3].

Automated methods can significantly reduce the time and effort required by scholars to identify and label the resources. Easy access to illustrations and their taxonomic classification facilitates research into the history of species abundance and variation, something that is paramount in biodiversity research. The current biodiversity crisis increases the importance of such historical studies as these provide a longer-term view of changes to biodiversity.

Automated species identification is not a new challenge within the field of computer vision and pattern recognition [39, 38, 6, 40, 19]. Many models have been developed for the detection and classification of species in photographs, but photographs and illustrations of species are quite distinct. In illustrations, the background (natural habitat) is often omitted and species are depicted in the form of collages

of multiple (smaller) depictions of their external and internal anatomy (e.g., bones, organs, limbs). These appear in a combination of various views (e.g., frontal, dorsal, lateral). Moreover, illustrations exist as rough pencil sketches and/or detailed colour drawings and commonly contain handwritten captions. To illustrate the differences between photographic and illustration data, three depictions and two photographs of the species *Lepas (Anatifa) anserifera Linnaeus, 1767* can be observed in figure 2 and 3.

(a) (b) (c)

Fig. 2: Scientific illustrations from the Iconographia Zoologica online collection[1] of the species *Lepas (Anatifa) anserifera Linnaeus, 1767*, with handwritten (historical) name *Anatifa laevis Bruguière, 1789* (best viewed in colour). (a) species within shell, (b) shell of species, (c) species without shell. Images free of known restrictions under copyright law (Public Domain Mark 1.0).

The large discrepancy between the two modes demands a classifier that is trained or fine-tuned on the illustrations. This is a non-trivial task due to several reasons of which we name three. First, the space of possible solutions is very large. Access to label candidates is available through many years of naming species and species systematics, but only a subset of these classes have labelled instances available for training. The Catalogue of Life (CoL)[2] estimates that *presently*, ∼2.2 million species on the planet are known to taxonomists, which, according to the current system of taxonomy, belong to ∼200,000 genera, ∼10,000 families, ∼1,600 orders, ∼400 classes, ∼100 phyla and 7 kingdoms. Second, as retrieving labelled instances is costly and difficult, few are available per species. Therefore, standard supervised classification models overfit to the training data. Third, testing the model on a test-set does not guarantee its value 'in the wild', where data often either comes from a different marginal probability distribution, or exists in a (partially) different feature space [42].

(a) (b)

Fig. 3: Photographs of the species *Lepas (Anatifa) anserifera Linnaeus, 1767 (Goose Barnacle)*, taken from iNaturalist.[3] (best viewed in colour). (a) Observation © David R.[4]. (b) Observation © mervyngreening.[5]. Images are licensed under CC BY-NC 4.0.

Here, we discuss an approach that copes with the aforementioned challenges. To address the first problem, we make use of a non-standard learning strategy called *zero-shot learning* (ZSL), with which it is possible to identify *unseen* classes: classes that are not observed by the classifier during training. As discussed in Section 2, the assumption that the Linnaean taxonomic system makes is that the members of a taxonomic group must share a cluster of similar traits [11]. As we move down the taxonomic ranks, the characteristics shared by groups of organisms become more specific. In a computer vision context, we also expect visual features (or attributes) to be shared among classes, becoming more similar and specific for classes lower down the biological taxonomy. Such assumptions are exploited by ZSL: images from a set of unseen classes can be classified through between-class feature transfer [19], for instance by embedding seen and unseen classes in a shared feature space.

We use a *prototypical network* [31], to optimise a species embedding model for zero-shot learning: we exploit auxiliary data - a hierarchy, historical texts and photographs - to obtain aforementioned class embeddings, and learn a mapping from the images to the representations. By introducing *fused prototypes* (FP), and *hierarchical prototype loss* (HPL), we aim to improve optimisation of the embedding model for classification.

To address the second problem, the difficulty of learning from few illustrations per class , we exploit image representations learned from another task - the recognition of species' photographs - to extract meaningful features for our task [26]. Moreover, by using a biological taxonomy as a label hierarchy for training (through HPL), a larger number of labelled examples are available, per group, higher up the label hierarchy.

For the third problem, we stress that a trained model should be evaluated 'in the wild' on a dataset collected under

[2] http://www.catalogueoflife.org
[3] https://www.inaturalist.org/
[4] https://www.inaturalist.org/observations/25983495
[5] https://www.inaturalist.org/observations/34793791

different conditions. Therefore, we analyse the final trained prototypical network using a second independent collection of illustrations without annotations.

3 Related Work

Below, we discuss datasets related to computer vision and biodiversity, and provide a short survey of the field of zero-shot learning.

Computer vision and biodiversity Recognising and identifying species in images is a well researched problem within the computer vision field. Most popular datasets contain classes of animals, (often birds), or plants [39,38,6,19,40,18,25]. Although some datasets contain clear, well positioned images of animals with many examples per class, newer data sets better capture the 'wildness' of real-world animal photographs. A citizen science project called iNaturalist[6], allows users to upload photographs of organism encounters in the wild. Since 2017, a new dataset has been published every year as part of the iNaturalist Competition FGVC6 for fine-grained image classification.[7] Computer vision models trained on such data sets are much better prepared for the automatic identification of species in the wild, for instance images collected from motion-triggered camera traps.[8] In addition to photographs of species, there are examples of models trained for the automated classification of plants in herbaria [5]. Herbaria are collections of dried plants on sheets of paper, which aid researchers in describing plant species.

While a great deal of work is spent capturing often unclear images of species in the wild, a wealth of detailed zoological illustrations are under-utilised. A reason could be that samples are small, many classes are under-represented, and numerous institutions have yet to start with the digitisation of their collections [10].

Zero-shot learning While standard supervised image classification methods learn to recognise classes from examples of those classes, *zero-shot learning* (ZSL) aims to recognise unseen classes from examples of other classes, using between-class feature transfer. To share knowledge between classes, all classes $y \in Y$ are embedded in a feature space, $\varphi(y) : Y \rightarrow \tilde{Y}$, called *class embeddings*. They are either (i) created manually, through class annotations or attributes [20,12], or (ii) learned from auxiliary information such as taxonomies [4,36] or text [23,28,16]. Attribute embeddings encode if a certain attribute - from a set of predefined attributes - is present for a specific class. Attribute embeddings can be either binary or continuous, e.g., $\{$*wing: 0.1, red: 0.4, tail: 0.7*$\}$ and fall within the interval [0,1]. Learned embeddings are continuous and represent similarities between classes more abstractly. Class embeddings from various sources can be used to complement one another; combining them often results in a higher accuracy [33,2,1]. Combining class embeddings can be done in different ways, for instance by concatenating the class embeddings or combining compatibility scores [2]. We refer to [2] for an extensive evaluation of class embeddings. Similarly to classes, images from classes Y_{tr} are embedded in a different feature space, $\theta(x) : X \rightarrow \tilde{X}$, resulting in *image embeddings*. Most commonly, θ is a Convolutional Neural Network (CNN), that learns filters that allow the extraction of important features from an image. After training the CNN, the top of the network - often just the softmax layer - is removed and an embedding function remains. Finally, a function is learned that maps image embeddings to class embeddings: $f : \tilde{X} \rightarrow \tilde{Y}$. By mapping an image from an unseen class into class embedding space, the image can be classified by assigning to it, for instance, the label of the nearest class embedding vector.

Most common ZSL methods learn either a linear [1,13,2,29] or a non-linear [44,32] compatibility function. Prototypical networks [31] belong to the latter group. They learn linear deep visual-semantic models, such as DeViSe [13] and Cross-modal transfer (CMT) [32], in which the visual object recognition network is trained to predict the class embedding vector learned from auxiliary data. However, where DeViSe aims to maximise hinge rank loss and CMT aims to minimise a distance function, prototypical networks produce a distribution over distances to class embedding vectors and minimise the negative log-probability.

While all methods achieve impressive results on small- and medium-scale datasets, the more realistic variant *generalised zero-shot learning* (GZSL), that aims to classify both seen and unseen classes, performs poorly for unseen classes [32]. The function learned by the model overfits to the seen classes and will therefore, during testing, favour seen over unseen classes. Therefore, zero-shot learning models embedded in real world applications should include novelty detection. For an extensive comparison of state-of-the-art of zero-shot learning and generalised zero-shot learning methods, we point to the work of Xian et al. [45]. In our work we use prototypical networks for zero-shot learning because they are state-of-the-art models within the few- and zero-shot learning domain [31].

4 The Data

In this section, we discuss the Zoological Illustration and Class Embedding (ZICE) dataset (see Subsection 4.1), used for training, validating and testing our zero-shot learning ap-

[6] https://www.inaturalist.org/
[7] https://www.kaggle.com/c/inaturalist-2019-fgvc6
[8] https://github.com/microsoft/CameraTraps

proach, and an independent verification-set (in Subsection 4.2) to verify the zero-shot learning results.

4.1 The ZICE dataset

The Zoological Illustration and Class Embedding (ZICE) dataset contains illustrations, from the Iconographia Zoologica online collection,[1] and class embeddings obtained from online data sources within the biodiversity domain.

The Iconographia Zoologica is a 19th century collection of biological illustrations from the Artis Library of the University of Amsterdam. The collection was formed by three collectors: the well-known collector and naturalist Th. G. van Lidth de Jeude, the zoologist R.T. Maitland and the curator of the shell collection at the Amsterdam Zoo, Abraham Oltman, together with the Amsterdam society *Natura Artis Magistra*. In the 21st century, the collection was digitised and labelled with either complete binomial species names (genus and specific epithet) or corresponding genera. The full online collection contains over 26,500 pages of zoological illustrations.

Table 1: A biological illustration of the canonical name 'Achatina columnaris' (Achatina columnaris - - Print - Iconographia Zoologica - Special Collections University of Amsterdam - UBAINV0274 088 12 0011.tif) and its GBIF classification meta-data.

Achatina columnaris - - Print - UBAINV0274 088 12 0011

GBIF taxonID	*2295965*
Kingdom	*Animalia*
Phylum	*Mollusca*
Class	*Gastropoda*
Order	*Stylommatophora*
Family	*Achatinidae*
Genus	*Achatina*
Specific epithet	*columnaris*

We have cross-referenced the illustration labels with the June 2018 backbone taxonomy [30] of the Global Biodiversity Information Facility (GBIF),[9] a central repository for biodiversity occurrence data. For 14,502 illustrations, labels could be cross-referenced directly with GBIF without extra domain expert curation. Matches were only accepted when the names had the status "accepted", as propagating labels with the status "unaccepted" or "synonym" could prove problematic for the performance of a network trained with these labels. As a result of the automated matching process, all illustrations in the ZICE dataset are organised according to a taxonomy with six ranks: kingdom, phylum, class, order, family, genus, genus + epithet (species). In the rest of this paper, we refer to this taxonomy with the term *label hierarchy*. The various ranks of the hierarchy we call *levels*. Below we illustrate our work as follows. Table 1 contains an image of an illustration, its label and the matched GBIF classification. Figure 4 shows twelve example illustrations.

Fig. 4: Cropped example illustrations from the ZICE trainset (best viewed in colour). Image (f), for example, depicts the skull of a *Rhinosceros unicornis* and image (j) the tail of a *Squilla hoevenii*. Images free of known restrictions under copyright law (Public Domain Mark 1.0)

For our approach, we have generated class embeddings, whose classes match those from the illustrations, from three different sources: (i) hierarchical embeddings φ^h based on the GBIF backbone taxonomy [30], (ii) text embeddings φ^t based on literature from the Biodiversity Heritage Library (BHL) [15] and (iii) photograph embeddings φ^p based on images from the iNaturalist 2018 challenge dataset [39]. Information on how these embeddings were produced is given in Section 5.

4.2 Data from the Committee for Natural History

The Committee for Natural History of the Netherlands Indies (1820-1850) was founded by King William I of the United Kingdom of the Netherlands. Their primary task was the collection of information on natural resources in the Dutch Indies. In addition, they were deployed to observe and describe the local flora and fauna. As a result, many specimens, biological illustrations and observation descrip-

[9] https://www.gbif.org/

tions were brought back to the Netherlands for closer investigation, with the aim to publish results on the natural diversity of the Dutch Indies [43]. Currently, the physical collection is stored at the *Naturalis Biodiversity Center* in Leiden. In 2008 the archival part of the collection was digitised (scanned), but due to a lack of annotation, it still remained inaccessible to biodiversity researchers. Currently, the collection serves as a use-case for the Making Sense of Illustrated Handwritten Archives Project[10] of which this work is part. We use 1,088 illustrations from the collection to evaluate the model in a realistic setting. Example illustrations are presented in figure 5.

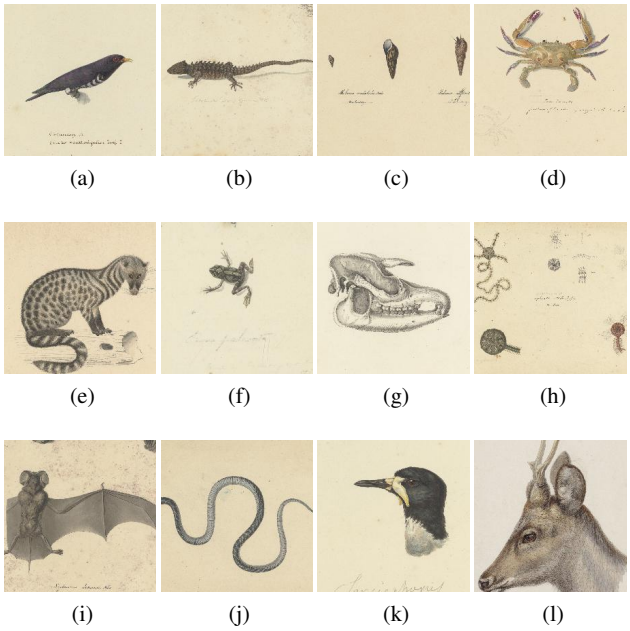


Fig. 5: Cropped example illustrations from the verification-set: data from the Committee for Natural History (best viewed in colour). Labels are unknown. Images free of known restrictions under copyright law (Public Domain Mark 1.0)

5 Methodology

In this section, we describe the mathematical formulation of the zero-shot learning model (ZSL) (in Subsection 5.1), image embeddings (in Subsection 5.2), class embeddings (in Subsection 5.3), a new method for (i) combining class embeddings: *fused prototypes* (FP) (in Subsection 5.4), and (ii) for calculating *hierarchical prototype loss* (HPL) based on the label hierarchy (in Subsection 5.5).

[10] www.makingsenseproject.org

5.1 Zero-shot Learning Model

Prototypical networks for few-shot learning, as described in [31], compute for each class an m-dimensional representation $\mathbf{c}_k \in \mathbb{R}^m$ or class *prototype*. They do so by embedding *support points* $\in S$, e.g., images belonging to a subset of classes k, $\mathbf{x}_k \in \mathbb{R}^n$, with an embedding function $f_\phi : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and taking the per-class average of the resulting embedded support points, see equation 1. These k points then serve as prototypes for those k classes. We further refer to the space $\mathbb{R}^m$ by the term *prototype space*.

$$\mathbf{c}_k = \frac{1}{|S_k|} \sum_{(\mathbf{x}_i, y_i) \in S_k} f\phi(\mathbf{x}_i) \tag{1}$$

To train the network, prototypical network loss (PNL), is calculated by embedding *query points*: images belonging to the same k classes that are embedded in prototype space using the same function f_ϕ. Distances from the query points to the k prototypes are computed so that, based on a softmax over these distances, a distribution over classes is obtained. Parameters ϕ are learned by minimising the negative log-probability of the true class k via Stochastic Gradient Descent. The network is trained with mini-batches, consisting of k classes, q query points and s support points is called an *episode*.

For zero-shot learning, Snell et al. [31] mention that rather than embedding support points in prototype space, prototypes can be constructed by embedding auxiliary information, e.g., class embeddings in the form of attribute annotations, in prototype space. In their paper they use binary attribute vectors from the CUB-200-2011 dataset [41]. They extract features from different crops of the images using GoogleNet [34] and map them to prototype space using a one-layer linear model. Similarly, they use a one-layer linear model to map the attributes to prototype space and prototypical training proceeds as in the few-shot setting. Rather than relying on one source (such as attributes), we rely on a combination of class embeddings from three distinct sources.

5.2 Image embeddings

We extract deep features from zoological illustrations using a deep Convolutional Neural Network (CNN). Training a deep CNN from scratch from small samples results in features that overfit the dataset. When all images from a class are depicted by a specific illustrator, the model learns features that are specific to the illustrator, such as a mark or a label. Therefore, we transfer image representations learned from photographs (the *source* dataset) to illustrations (the *target* dataset) to bootstrap the learning of the image embedding space [26]. We use the inception V3 model [35],

and import weights learned on the iNaturalist 2018 competition dataset.[11] For zero-shot learning, image embeddings are often generated using CNNs pre-trained on another task (e.g., the ImageNet task [8]). The choice of model is crucial as the quality of the image embeddings has a big impact on the performance of the ZSL model. Therefore, we have chosen to use a model that was trained on a task more similar to ours. Xian et al. [45] mention that class overlap between the classes from the source dataset and classes represented in the target dataset leads to an unwanted positively biased result. However, our goal is not to compare between various state-of-the-art zero-shot learning methods, but rather to provide insights for training a model that is able to generalise to new data within the target domain.

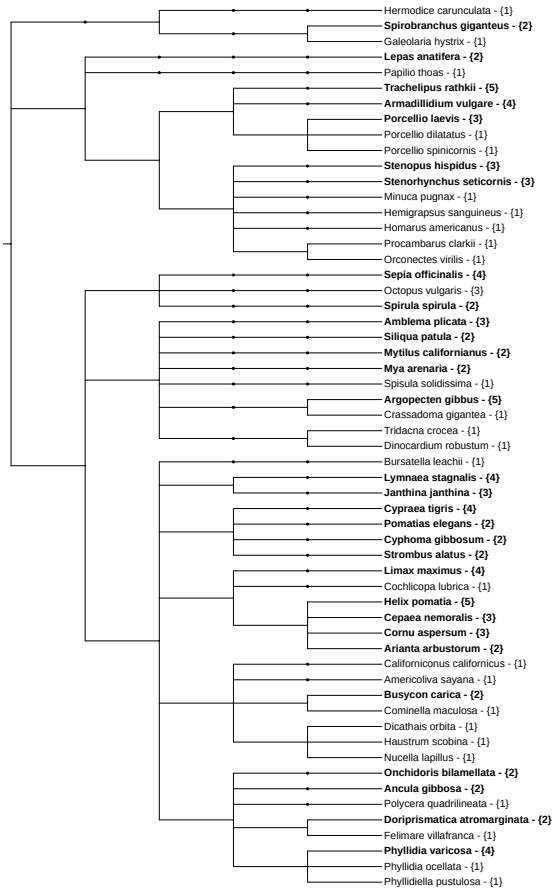


Fig. 6: A subset of classes from the ZICE dataset and the corresponding hierarchy. Leafnodes present species names. Nodes indicate taxon groups from kingdom, (left) to species (right), becoming increasingly specific. It contains species from the three animal phyla (top to bottom): *Annelida*, *Mollusca* and *Anthropoda*. Bold names indicate classes used for training, and numbers indicate number of instances within that class.

[11] https://github.com/macaodha/inat_comp_2018

5.3 Class embeddings

Below we describe details concerning the embedding functions ϕ used to create the class embeddings φ from the three different sources: (i) φ^h : the GBIF hierarchy, (ii) φ^t : texts from the Biodiversity Heritage Library (BHL) and (iii) φ^p : photographs from the 2018 iNaturalist competition dataset. As each embedding comes from a different domain, all embeddings are l_2-normalised.

φ^h: After all 7973 classes from the ZICE dataset were matched with GBIF names, we had access to the ground truth list of higher taxon labels for nearly all of them: 7920 in total. For the missing 53 classes, no (or an incomplete) higher classification was available in the GBIF taxonomy backbone. Using the deterministic algorithm from Barz et al. [4], we projected all 7920 classes onto a unit sphere of dimensionality n - where n is the number of classes. The negated dot product between classes on the sphere represents their semantic similarity. This similarity is based on the ratio of overlap between their ground truth list of higher taxon labels - nodes in the hierarchy. Part of the hierarchy of classes is given in figure 6.

φ^t: To facilitate semantic search over large textual biodiversity archives, Nguyen et al. have constructed an inventory of name variants and synonyms from a large textual biodiversity corpus [24]. For this task, they have computed the semantic similarity between all single and multi-word terms - ”chipping sparrows” becomes ”chipping_sparrows” - mentioned in the corpus. The semantic similarity is the output of a similarity function computed from word embeddings. Furthermore, they compared multiple methods to compute word embeddings: *continuous-bag-of-words* (CBOW) [22], *count-based* [37] and *Global Vectors* (GloVe) [28]. From these three, we rely on the 300 dimensional multi-word GloVe embeddings.

φ^p: Features in photographs are quite distinct from those in illustrations, but their features are well able to capture the semantic similarity of the different classes they represent. To this end, we have extracted 2048 dimensional features from the iNaturalist 2018 dataset photographs, using the inception V3 model trained on the corresponding dataset (previously mentioned in Section 5.2).

5.4 Combining class embeddings

Below we describe two methods for generating singular class prototypes from three distinct embeddings, each with a different dimensionality.

Concatenated embeddings (CE) One method that is often employed to combine the different embeddings is concatenation: the dimensions of each class embedding (from the

three distinct sources) are concatenated together. This results in one sparse matrix with a large dimensionality. Similarly to Snell et al. [31], we use a one-layer linear model to map the concatenated embeddings to prototype space.

Fused prototypes (FP) Our new idea is to implement *fused prototypes*, which we derive from the prototypical few-shot learning approach. Instead of using images as support points, class embeddings from distinct sources are mapped into prototype space, each source with a one-layer linear model. In prototype space, the resulting prototypes are fused together, similarly to the way support points are fused for few-shot learning, see formula 2. In that formula, E denotes the set of class embedding vectors from the various sources i. $\mathbf{e_k}$ refers to the class embedding vector $\in E_i$ that is labelled with class k. f_i refers to the linear model that embeds the class embedding vectors from source E_i to a prototype space.

$$\mathbf{c}_k = \sum_{(\mathbf{e_k}, y_k) \in E_i} f_i \phi(\mathbf{e}_k) \tag{2}$$

We hypothesise that fused prototypes will perform better than concatenated embeddings, as the latter introduce one large sparse input space whereas fused embeddings learn from multiple dense input spaces.

5.5 Hierarchical prototype loss

Hierarchical prototype loss (HPL) extends prototypical network loss (PNL), and is defined as the sum of the losses for each level of the label hierarchy. One other example of a type of hierarchical loss is hierarchical triplet loss (HTL) [14], which extends triplet loss (TL) used in triple-based models.

The loss for a specific level l is calculated by first generating new prototypes for that level: all class embeddings that belong to a class $k \in l$ are averaged to form new prototypes. As described in Section 5.1, distances of the query points to the k new prototypes are then computed and the loss is calculated over these distances. The hierarchical prototype loss accumulates the losses of each level in the hierarchy L for which we have labels (in our case there are six), see equation 3.

$$\text{HPL} = argmin_w \sum_{i=0}^{L} -log(p(\mathbf{y}^i|\mathbf{x}, \mathbf{w}) \tag{3}$$

By implementing HPL, we take a multi-objective (or multi-granularity) approach: we enforce a clearer separation of classes not only for the finest grain, but also for coarser taxonomic groups. As more labels are available for each level higher up in the label hierarchy, this intuitively supports the discovery of more robust features for the classification of coarser classes.

6 Experimental setting

In this section we discuss details regarding the settings of the experiment: the dataset splits (in Subsection 6.1), data augmentation (in Subsection 6.2), evaluation criteria (in Subsection 6.3) and experimental tasks (in Subsection 6.4).

6.1 Dataset splits

As recommended by [45], we split the classes for training and evaluation based on the number of instances each of them contain. Since our dataset contains so few instances per class, ($n_k \in [1, 283]$, μ: 1.79, σ: 3.93). We have used all classes with n $\geq$ 2 per class for the training set Y_{tr}. Two examples per class is not sufficient to learn a good class representation, but the features of these illustrations are useful for between super-class feature sharing. Moreover, we exploit them for learning representations of classes on a higher taxonomic level, since a larger number of instances are available higher up the label hierarchy. All remaining classes with n = 1 were split and used for the validation set Y_v, and the test set Y_{ts}. Table 2 shows the number of classes and images within each super-class, split, and embedding. Since not all of the classes were represented in each source (GBIF, BHL and iNaturalist), each embedding (φ^h, φ^t, and $Y\varphi^p$ respectively) represents a subset of Y. However, together they span the totality of classes $y \in Y$. The super-class *Animalia* is used for classes that are not assigned to a phylum.

6.2 Data augmentation

For training, we have used image embeddings extracted from augmented versions of all images, in order to increase the ability of the classifier to generalise the classification with respect to the data. Before cropping all images, the largest side of each image was first resized to 300. During resizing, we keep the aspect ratio identical to the original image. Since the images are rather large and contain a large amount of white space, it is quite clear that not resizing the images first would result in many empty crops. 2048-dimensional features are extracted by applying the pre-trained Inception V3 model to crops (middle, upper left, upper right, lower left and lower right) of each resized original illustration and its horizontally flipped version. Crops containing only white space or text were manually discarded.

Table 2: Dataset statistics: number of classes and instances per super-class (phylum), number of classes per split, number of instances per split, and number of classes per embedding

Super-class (phylum)	Y_{tot}	n_{tot}	Y_{tr}	Y_v	Y_{ts}	n_{tr}	n_v	n_{ts}	Y_{φ^h}	Y_{φ^t}	Y_{φ^p}
Arthropoda	2977	3740	620	1106	1251	1383	1112	1245	2977	218	14
Chordata	2903	7358	1281	870	752	5736	878	744	2901	2050	475
Mollusca	1423	2384	488	464	471	1449	464	471	1385	475	40
Cnidaria	179	299	58	47	74	178	48	73	179	88	5
Echinodermata	111	180	36	33	42	105	33	42	111	62	10
Annelida	109	171	32	44	33	94	44	33	106	61	3
Porifera	59	79	17	17	25	37	17	25	59	11	-
Platyhelminthes	56	75	9	38	9	28	38	9	55	9	-
Bryozoa	45	67	10	12	23	32	12	23	45	23	-
Brachiopoda	37	38	1	23	13	2	23	13	37	2	-
Nematoda	18	24	4	9	5	10	9	5	18	8	-
Rotifera	17	20	2	7	8	5	7	8	17	12	-
Ctenophora	14	33	5	2	7	24	3	6	14	6	-
Nemertea	6	8	2	3	1	4	3	1	4	4	-
Sipuncula	5	6	1	3	1	2	3	1	5	-	-
Acanthocephala	4	5	1	1	2	2	1	2	4	2	-
Nematomorpha	2	6	1	1	-	5	1	-	2	1	-
Onychophora	2	2	-	2	-	-	2	-	0	2	-
Cephalorhyncha	1	1	-	1	-	-	1	-	0	1	-
Chaetognatha	1	2	1	-	-	2	-	-	1	1	-
Entoprocta	1	1	-	1	-	-	1	-	0	1	-
Animalia	3	3	-	2	1	-	2	1	0	3	-
Total	7973	14502	2569	2684	2717	9098	2702	2702	7920	3040	547

6.3 Evaluation criteria

In our experimental ZSL results (Subsection 7.2) we report two accuracy metrics: top-k accuracy and hierarchical accuracy@k. Below we will note down our motivations for doing so.

Top-k accuracy When measuring the performance of a classifier on a large-scale hierarchical dataset, we should be alert on the fact that a flat top-1 accuracy does not accurately portray the classifier's capabilities. Assuming the solution space is large, it is valuable for domain experts to obtain top-k predictions as the correct label might be among them, as exemplified later in figure 8. We therefore report top-k accuracy, $k \in \{1, 2, 5, 10\}$. This metric is computed by the percentage of images for which the correct label is among the top k predictions.

Hierarchical accuracy@k Classifying an illustration of a *Boiga nigriceps* as a *Boiga dendrophila* - both tree snakes - is less problematic than classifying it as a *Procyon lotor*, a common raccoon. In the former case, the classifier has learnt important coarser features that allowed it to classify the image as a tree snake, providing researchers with a partially correct classification. To this end, we would like to shed light on the quality of the entire predicted classification for each illustration. Therefore, we additionally report *hierarchical accuracy*. Hierarchical @k precision is sometimes used as a metric for hierarchical datasets [13]. We report a new metric that we deem more informative in our context: *average per-level accuracy*, or *hierarchical accuracy*. It is computed by calculating the accuracy for each level in the label hierarchy and averaging over these, see formula 4. In formula 4, L refers to the set levels for which we have labels and l to a specific level $l \in L$.

$$\textit{Hierarchical acc} = \sum_{l=1}^{|L|} \frac{n \text{ correct preds in } l}{n \text{ samples in } l} \tag{4}$$

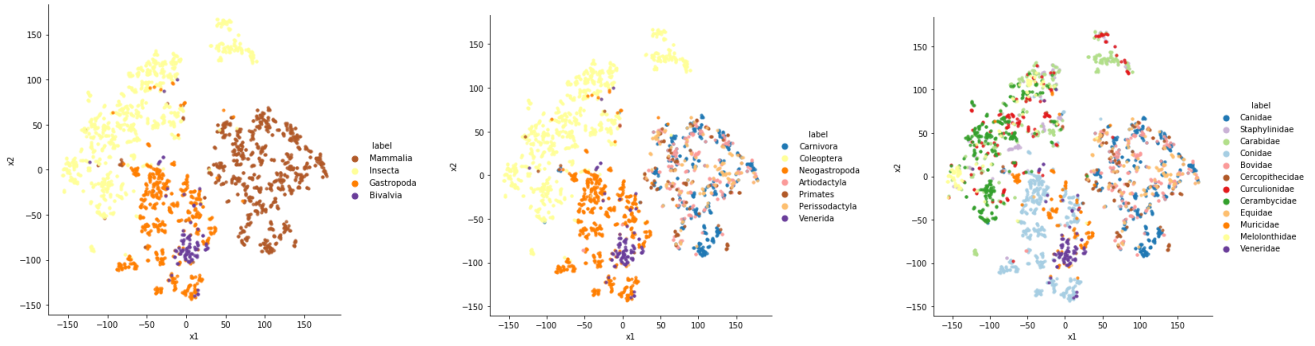
Additionally, we report accuracies for labels k levels up the label hierarchy, where $k \in \{1, 2, 3\}$, thus referring to the accuracy for labels one, two and three levels up the label hierarchy respectively.

6.4 Experimental tasks

In this section we describe the setup of our experiments, dividing them up in separate tasks, each task evaluating one of the components of our proposed zero-shot learning approach (discussed in Section 5).

In a common supervised classification setting we evaluate (in Subsection 7.1):

Task 1 the image embeddings. Specifically, we train a Support Vector Machine (SVM) on the image embeddings.



(a) Class labels (b) Order labels (c) Family labels

Fig. 7: T-SNE plots showing image embeddings of images from the ZICE dataset (should be viewed in colour). From left to right, label levels become more specific. Family labels come from a selection of 12 families of which the binomial name was not present in the iNaturalist 2018 dataset. The t-SNE algorithm was run for 5,000 iterations with perplexity 100.

In a fine-grained zero-shot learning setting we evaluate (in Subsection 7.2):

Task 2 the class embeddings. Specifically, we train multiple prototypical networks using the image embeddings for every possible combination of class embeddings.

Task 3 how class embeddings are combined. We compare the results of a network that uses *fused prototypes* (FP) to those of a network trained using *concatenated embeddings* (CE).

Task 4 the hierarchical prototype loss (HPL): loss accumulated for all levels of the hierarchy.

Task 5 the final results by an in-depth analysis of the network on the main test-set.

7 Experimental Results

Before we present zero-shot learning results, we start by evaluating the quality of the image embeddings (**Task 1**) in a separate supervised classification setting below.

7.1 Supervised classification and visualisation

We show classification results from a Support Vector Machine (SVM) trained on the image embeddings, in table 3, and a t-Distributed Stochastic Neighbor Embedding (t-SNE) [21] visualisation of the image embeddings, figure 7. For this supervised classification task, we have selected image embeddings from the set of species that is disjoint from the set of species represented in the iNaturalist 2018 dataset, so as to obtain a deeper insight into the generic quality of the embeddings. Since both the SVM and t-SNE do not perform well on small samples, we select twelve most populated taxon groups from two levels higher up the label hierarchy: the *family* level. We additionally present higher taxon labels for both the SVM results as the t-SNE visualisation. Table 3 shows per-class, micro, macro and weighted average precision and recall results for the Support Vector Machine (SVM) trained on top of the image embeddings. The weighted average alters the macro metric to account for label imbalance. The support column indicates the number of actual occurrences of that class in the given dataset. The SVM was trained using a stratified 80% 20% split for the train and test-set respectively.

Looking at figure 7, we see that same-class image embeddings are visibly clustered, but that image embeddings within the order *Coleoptera* (in yellow) and within class *Mammalia* (in brown) overlap. This effect is reflected in table 3: the image embeddings from only one of four families

Table 3: Classification precision and recall results in % (rounded off to whole integers) for a Support Vector Machine (SVM) trained on the image embeddings belonging to 12 families (also visualised in figure 7). The top-1 per-class average accuracy is 43.58%.

Class	Family	prec.	rec.	f1	support
Mammalia	Bovidae	0	0	0	19
Mammalia	Canidae	48	100	65	33
Insecta	Carabidae	44	74	56	27
Insecta	Cerambycidae	56	85	68	26
Mammalia	Cercopithecidae	0	0	0	9
Gastropoda	Conidae	87	98	92	41
Insecta	Curculionidae	0	0	0	14
Mammalia	Equidae	0	0	0	12
Insecta	Melolonthinae	100	22	36	9
Gastropoda	Muricidae	67	55	60	11
Insecta	Staphylinidae	0	0	0	10
Bivalvia	Veneridae	82	90	86	10
	micro avg	60	60	60	221
	macro avg	40	44	38	221
	weighted avg	46	60	50	221

Table 4: Zero-shot learning (ZSL) classification results in % for **Task 2, 3 and 4**: each combination of embeddings, concatenated embeddings (CE) versus fused prototypes (FP), fused prototypes plus hierarchical prototype loss (FP + HPL), and final results for the best configuration. The 50-way classification accuracy for the final model was 35.53%, calculated by averaging results over 6,000 randomly drawn episodes.

				top-k acc Y_{ts}				**Hierarchical acc@k** Y_{ts}			
Method	φ^h	φ^t	φ^p	*1*	*2*	*5*	*10*	*1*	*2*	*3*	*avg*
N/A	✓	✗	✗	**2.29**	**4.12**	**8.9**	**15.34**	**5.93**	13.23	43.74	36.38
	✗	✓	✗	0.41	0.66	1.14	1.72	0.72	1.22	7.33	12.53
	✗	✗	✓	0.55	0.85	1.47	2.15	1.03	2.81	15.29	18.26
FP	✓	✓	✗	2.13	3.89	**8.79**	**15.11**	5.51	13.56	43.21	35.96
	✓	✗	✓	**2.50**	**4.26**	**8.91**	**15.26**	**6.05**	**14.24**	**45.69**	**36.85**
	✗	✓	✓	0.53	0.84	1.45	2.06	1.04	2.02	9.41	13.50
	✓	✓	✓	**2.42**	**4.29**	**9.10**	**15.37**	**5.98**	**14.22**	**45.09**	**36.70**
CE (baseline)	✓	✓	✓	2.09	**4.05**	**8.96**	**15.54**	5.45	13.42	**44.76**	36.41
FP	✓	✓	✓	**2.42**	**4.29**	**9.10**	**15.37**	**5.98**	**14.23**	**45.09**	**36.70**
FP	✓	✓	✓	**2.42**	**4.29**	**9.10**	**15.37**	**5.98**	14.23	45.09	36.70
FP + HPL	✓	✓	✓	2.12	3.88	**8.88**	**15.03**	**6.23**	**15.71**	**51.10**	**39.35**
Final model	✓	✗	✓	2.77	4.74	9.64	16.02	6.94	16.65	50.71	39.67
Majority guess	-	-	-	0.04	0.07	0.19	0.37	2.85	3.26	21.87	18.66

subsumed under the class *Mammalia* can be classified correctly (*Canidae*, with 100% recall). From the precision value (48%) we find that many other image embeddings are also classified as Canidae.

The results shows us that the features learned from the iNaturalist 2018 task are not specific enough to properly classify all fine-grained classes in our task well. Further improving the image features would therefore improve zero-shot learning results, but other approaches should be devised that can do so with the aid of small data samples. Visualisation of the features after dimensionality reduction can give an indication up to which grain the features within specific taxon groups are informative enough for proper classification. How well classes can be distinguished within a certain taxonomic group also depends upon the inter-class variation of that group, which can be quite small. Some species within the order *Coleoptera* (beatles), for instance, can only be accurately identified after a close inspection of their genitalia [7].

7.2 Fine-grained zero-shot learning

In this section we provide general training details for our zero-shot learning experiments, present results for the evaluation of our approach (**Task 2 and 3**), and show final results for a model that incorporates the results from **Task 2 and 3** (**Task 4**).

Training details All prototypical networks were trained using Stochastic Gradient Descent (SGD) with Adam. Episodes for training were comprised of $k = 50$, $q = 1$ and $s = 0$, similar to a balanced mini-batch of size 50. The validation loss was monitored during training and if, for 10 iterations, the loss did not decrease, the learning rate was decreased with a factor of 0.5. We tuned hyper-parameters using hyper-parameter optimisation - tree-structured parzen estimators - and ended up with a learning rate of 10^{-4} and a weight decay of 10^{-5}. Early stopping on the validation loss was used to determine the optimal number of epochs for training. For each model, five different networks were trained. As a statistical test for comparing classifiers we used the McNemar test [9] for each classifier pair for all predictions of 5 runs accumulated. It is a test that works well for testing statistical significance when dealing with paired nominal data for comparing classifiers trained, validated and tested on the same splits of a dataset. Bold numbers indicate statistical superiority over other values within that column and cell. Multiple bold numbers in one row are not statistically different. A final model was trained, again 5 times, with the configuration that we found to work best. The last row of table 4 indicates accuracy values for the majority guess, where the model simply always predicts the majority class.

Evaluation Table 4 shows results for networks trained, validated and tested with each k-combination of the set of embeddings E, with $k \in (1, 2, 3)$. In order for the results to be comparable between all combinations, we used all classes to train, validate and test the networks, despite the fact that each embeddings spans a subset of all classes. In case a class was not represented in an embedding, those dimensions were set to zero. In this context, these results say more about the contribution of each embedding to the performance of the networks, than about their quality.

The results in table 4 show us that φ^h is the most informative embedding, mainly because it embeds almost all classes, as is visible in table 2. A network trained with φ^t ap-

Table 5: Zero-shot learning (ZSL) classification results in % for **Task 5** on the test-set per super-class (phylum).

			top-k acc Y_{ts}				**Hierarchical acc@k** Y_{ts}			
Super-class (phylum)	avg n_{tr}	avg n_{ts}	*1*	*2*	*5*	*10*	*1*	*2*	*3*	*avg*
Chordata	5736	744	4.7	7.39	14.65	24.06	14.65	53.36	81.05	50.22
Mollusca	1449	471	3.4	6.16	11.89	20.59	29.3	47.56	73.25	47.77
Arthropoda	1383	1245	1.61	2.97	6.59	10.6	15.74	60.88	80.0	50.1
Cnidaria	178	73	8.22	9.59	16.44	30.14	19.18	31.51	41.1	29.86
Echinodermata	105	42	4.76	7.14	9.52	21.43	9.52	11.9	33.33	19.05
Annelida	94	33	0.0	0.0	0.0	0.0	0.0	0.0	3.03	1.21
Porifera	37	25	0.0	8.0	8.0	16.0	4.0	8.0	44.0	20.0
Bryozoa	32	23	0.0	0.0	0.0	8.7	0.0	4.35	4.35	3.48
Platyhelminthes	28	9	0.0	0.0	0.0	0.0	0.0	0.0	11.11	4.44
Ctenophora	24	6	0.0	0.0	33.33	33.33	0.0	0.0	0.0	3.33
Nematoda	10	5	20.0	20.0	40.0	40.0	20.0	40.0	40.0	32.0
Rotifera	5	8	0.0	0.0	0.0	12.5	0.0	0.0	0.0	0.0
Nemertea	4	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Sipuncula	2	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Brachiopoda	2	13	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Acanthocephala	2	2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Animalia	0	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Per super-class average	534.76	158.94	2.51	3.6	8.26	12.79	6.61	15.15	24.19	15.38
Per sub-class average (species)	9098	2702	2.96	4.96	9.96	16.65	7.11	17.10	52.26	40.05

pears to perform better than the majority guess for the top-k acc metric. However, even though φ^t spans almost half of the classes; 3040 in total, it harms the learning ability of the network when used in combination with other embeddings. This could be due to a myriad of factors: it could be that the embedding is better suited for finding synonyms between taxon terms - as similar species are described similarly. It could also be that some names in the Biodiversity Heritage Library [15] are ambiguous: referring to one species in the historical texts, while they refer to another in modern taxonomy. Particularly, any historical unpublished name could have been published today as a different species. Matching them with sources from a modern taxonomy could therefore be problematic. φ^p as an addition to φ^h improves the accuracy of the model, specifically the hierarchical acc@2 (13.23% to 14.24%) and @3 (43.74% to 45.69%). Using more instances and more fine-grained classes to generate φ^p would therefore further improve results.

Table 4 further presents results for the comparison of fused prototypes (FP) with concatenated embeddings (CE), and the effects of hierarchical prototype loss (HPL). CE represents the baseline model: it is comparable to the method used by Snell et al. [31] for zero-shot learning. Results show that by using our fused prototypes (FP) formulation, we can increase the top-1 accuracy from 2.09% to 2.42%. Such an increase is non-trivial. Since the test-set contains an instance per class, with 2702 classes (on the finest grain), an increase of 0.33% of the top-1 accuracy equals the capability of the classifier to correctly classify illustrations from an *additional* 9 classes from different parts of the biological taxonomy. Fused embeddings also induce a higher hierarchical accuracy compared to concatenated embeddings: from 5.45% to 5.98% and 13.42% to 14.23% for hierarchical acc@1 and @2 respectively. We anticipate that when class embeddings from additional (informative) sources are used, this effect which we discuss in Section 5.3 will become more evident: the value of using fused prototypes over concatenated embeddings will increase.

As expected, table 4 shows that adding hierarchical prototype loss (HPL) as a loss function improves the average hierarchical accuracy significantly - from 36.70% to 39.35%. However, a decrease is measured for the top-1 and top-2 accuracy: from 2.42% to 2.12% and 4.29% to 3.88% respectively, which demonstrates the inter super-class variation of taxon groups. It is unsurprising that the models that were trained with neither the hierarchical prototype loss nor the hierarchical embeddings have low hierarchical accuracy values compared to the majority class guess.

Final results A final model was trained 5 times using the best configuration - $\{\varphi^t, \varphi^p\}$, FP and HPL. Although implementing HPL decreases the top-1 and top-2 accuracy, a substantial increase of the average hierarchical accuracy was measured. Table 4 shows per-network averaged top-k and hierarchical acc@k accuracies for the final model on the test-set, and table 5 provides results for the same metrics, calculated from predictions of the final model's best network, and detailed per super-class. Table 5 serves to provide a deeper insight into the trained network. Evidently, illustrations from some of the super-classes were not recognised at all due to their limited contribution to the training of the network - visible from the column avg n_{ts} - and per definition, most feature sharing occurs within super-classes.

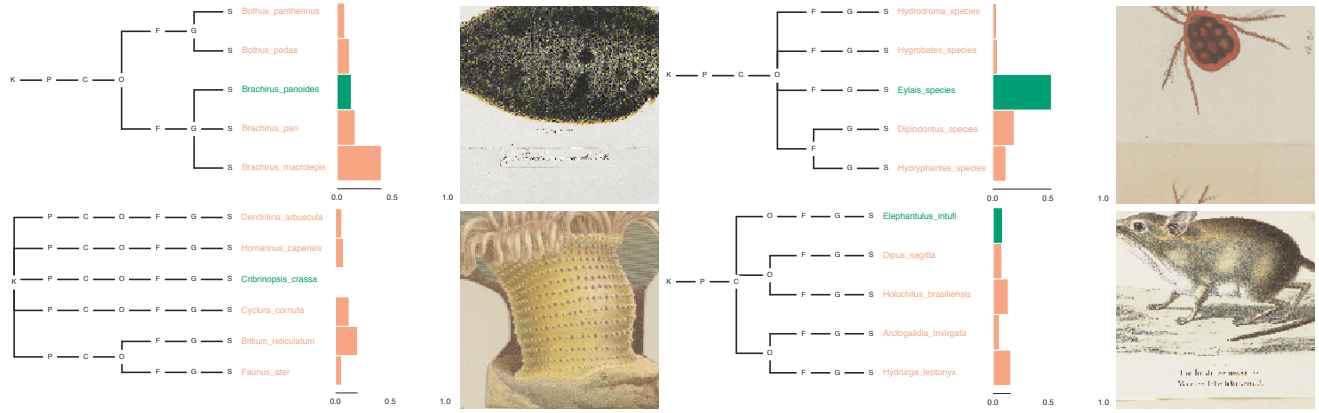


Fig. 8: Example of 4 test images and the confidence values of their top 5 predictions (best viewed in colour). Confidence values, per illustration, have a sum of one. Labels are organised hierarchically (K: kingdom to S: species) to show the diversity of predictions and how close - in the label hierarchy - the classifier is to the real label. For the *Cribrinopsis crassa*, the correct label was not among the top 5 predictions. A dark green label denotes the correct label.

Furthermore, table 6 details results for GZSL. Here, the network could select label candidates from *all* classes rather than just the *unseen* ones during classification. The top-k accuracies for GZSL are poor: during classification, a network trained for ZSL tends to favor seen classes over unseen classes [32]. This however, logically, does not affect the average hierarchical accuracy as much; good predictions are generally from the same family, order or another label level higher up the label hierarchy (see figure 8), and seen and unseen classes share common super-classes.

Figure 8 shows images from the test-set and the confidence values of their top 5 predictions. The first (1) and second (2) image (from top to bottom) are examples of good predictions. Although the prediction of the first image (1) is incorrect (the classifier is most confident that the correct label is *Brachirus macrolepis*), it is a species from the correct genus *Brachirus*. Moreover, the top 3 predictions are all from the same (correct) genus, and the remaining two predictions are from a different genus within the same family. The predictions for the second image (2) are very good: the confidence value for the correct label is high and the remaining predictions are from the same order. Visually, there are very few distinctions between species from these families, so mistakes reflect the inability of the classifier to learn very specific fine-grained features from the available data. The third image (3) is a very poor prediction as (i) the correct label is not among the top 5 predictions and (ii) almost all predictions are from a different phylum. Interestingly, however, the most confident predictions do have something in common: they share the illustration's most salient feature - a dotted pattern. Lastly, the fourth image (4) contains an incorrect prediction, but is correct up to the class level. The correct label belongs to the order *Macroscelidea* (Elephant shrew), and incorrect predictions belong to the *Rodentia* (Rodents) and *Carnivora* (Carnivores). Predictions from the *Rodentia* are from two different mouse families. Generally, elephant shrew visually resemble mice or gerbils; both rodents. The most salient feature that would allow a classifier to make the distinction between a mice or gerbil and an elephant shrew, interestingly, is cut off from the illustration: its long trunk-like nose resembling the trunk of an elephant. It is therefore good to consider that cropping the image at its center in a standardised way can cause the loss of information that is vital for proper classification.

Table 6: Generalised zero-shot learning (GZSL) classification results in % for final model

	top-k acc Y_{ts}				**Hier. acc@k** Y_{ts}		
Method	*1*	*2*	*5*	*10*	*1*	*2*	*avg*
GZSL	0.04	0.21	1.24	3.25	4.47	16.03	38.19
M. guess	0.01	0.03	0.06	0.13	2.85	3.26	18.66

8 Analysis and discussion

The results presented in this work show us, first of all, that normal supervised classification can not cope with the full scope of the problem presented in this paper. Training a CNN from scratch would overfit filters to our dataset. Using filters from a pre-trained model would partially circumvent this issue, but a network trained on these features would show similar issues. Additionally, it could only be employed to classify previously seen classes (classes with many training examples), whereas in cases like this, it is difficult to retrieve labelled digitised example illustrations for all classes.

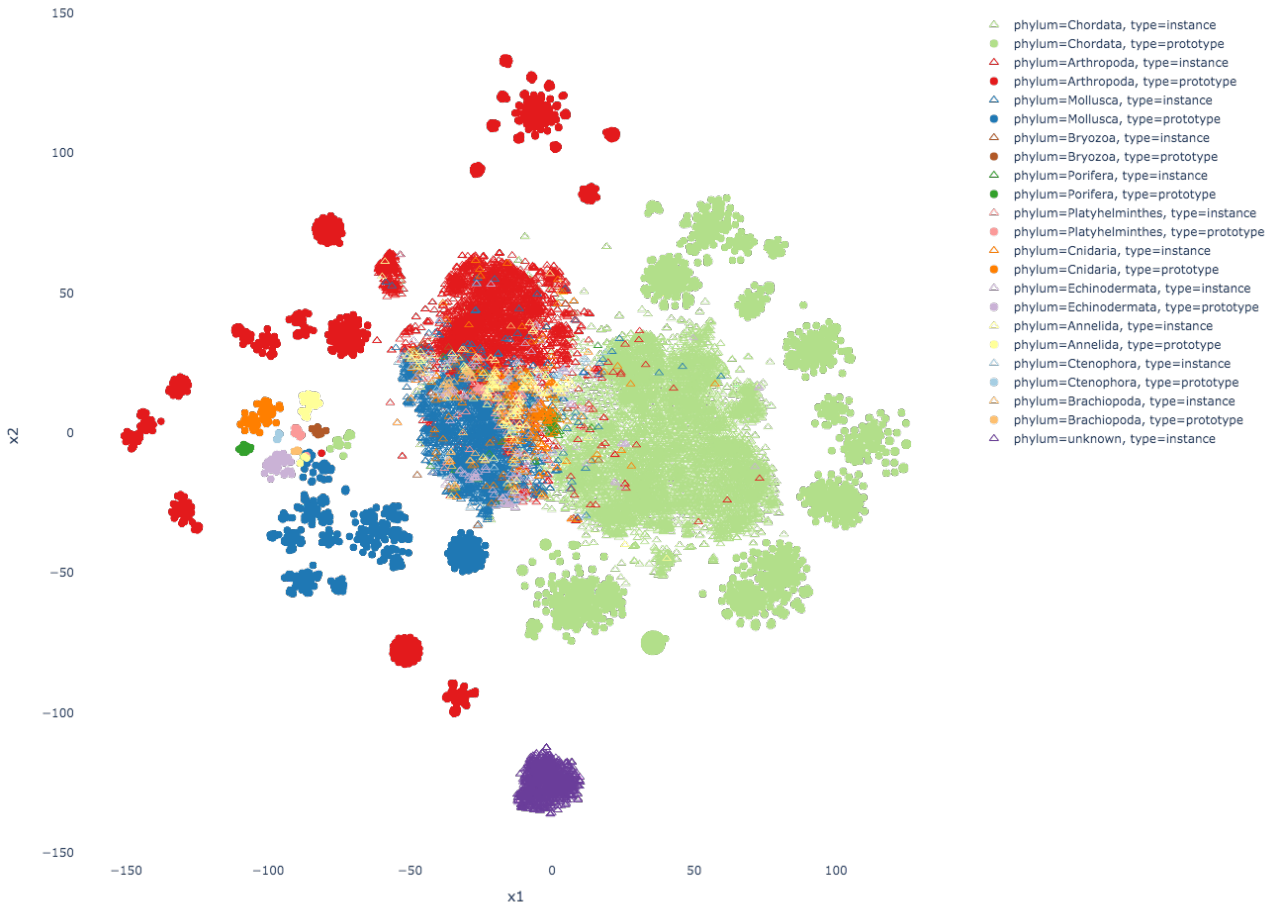


Fig. 9: A t-SNE plot showing all prototypes (closed circles) and instances (open triangles) from the 12 most populated phyla (indicated by different colours), embedded by the final prototypical network (should be viewed in colour). The plot includes prototypes and instances from training, validation and test classes. Instances from the verification set (bottom cluster) are indicated by the label 'unknown'. Note that t-SNE does not accurately preserve distances between clusters.

Table 5, 6 and figure 8 show us that with the aid of transfer learning, background knowledge, auxiliary data, and a label hierarchy, we can learn to capture coarse to finer-grained features that are shared among taxon groups. Practically, however, for many of these problem domains, it is challenging to transfer obtained results to real-world scenarios. First of all, table 5 shows us that novelty detection is a requirement for GZSL, as seen classes are favored over unseen classes. Second, using a trained network in real-world applications can prove problematic.

The verification set that we have presented in Section 4 serves as an example real-world dataset. Illustrations from this dataset, shown in figure 4, depict the same task as illustrations from the ZICE dataset, shown in figure 5. However, both datasets appear to come from a distinct marginal probability distribution. In order to demonstrate this effect, figure 9 shows a t-SNE visualisation of images and prototypes from the ZICE dataset, embedded by the final network. In addition, it shows the images from the verification-set (depicted as purple triangles), similarly embedded by the final network. It is clear from the t-SNE visualisation that the embedded images from the verification-set lie on a different manifold than those from the ZICE dataset. Through training, the prototypes have enveloped the manifold of embedded images from the ZICE dataset. The classifier therefore performs poorly on the verification-set, selecting only classes from the phylum *Anthropoda*, as its prototypes are closest. The domain shift most likely resulted from the use of different types of paper, sketching techniques and materials. We argue that it is vital that a computer vision system is tested on an independent dataset that represents the same

task. Models can then be developed to, for instance, align domain marginal probability distributions [42].

9 Conclusions

In this paper we have analysed the problem of classifying species in zoological illustrations. Fur this purpose, we have introduced a dataset representative of the problem, and had to deal with very limited access to labelled examples to learn good, robust data representations. For most classes, only a few example instances were available, whereas for other classes, no examples were available for training. Zero-shot learning and transfer learning allowed us to re-use and share features of images through the deployment of auxiliary data sources, and hereby to push the boundaries of automated recognition for this specific problem: from the classes that contained zero example instances for training, illustrations from 80 classes could be classified correctly. We have introduced fused prototypes (FP) and hierarchical prototype loss (HPL) to improve the results further, and conclude that these alterations improve over the baseline substantially. FP most notably allowed us to classify examples from an additional 9 unseen fine-grained classes, and learning with HPL has increased the average hierarchical accuracy substantially (from 36.41% to 39.35%). Finally, a verification-set has shed light on the robustness of the network to differences in marginal probability distributions between datasets.

We have demonstrated how intrinsically complex it can be to develop computer vision models for real-world applications, but our results have shown that computational methods can be used as decision support for researchers. Our model can help biodiversity researchers classify their historical and present-day scientific illustrations, which reside underutilised in natural history museums globally. Online datasets that store domain knowledge and auxiliary data of species can and should be exploited to develop embedding models for classification. They can guide and improve the recognition of illustrations from previously seen *and* unseen species of living organisms on various levels of the biological taxonomy.

References

1. Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C.: Label-embedding for image classification. IEEE transactions on pattern analysis and machine intelligence **38**(7), 1425–1438 (2015). DOI 10.1109/TPAMI.2015.2487986
2. Akata, Z., Reed, S., Walter, D., Lee, H., Schiele, B.: Evaluation of output embeddings for fine-grained image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2927–2936. IEEE (2015). DOI 10.1109/CVPR.2015.7298911
3. Austen, G.E., Bindemann, M., Griffiths, R.A., Roberts, D.L.: Species identification by experts and non-experts: comparing images from field guides. Scientific Reports **6**, 33634 (2016)
4. Barz, B., Denzler, J.: Hierarchy-based image embeddings for semantic image retrieval. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 638–647. IEEE (2019). DOI 10.1109/WACV.2019.00073
5. Belhumeur, P.N., Chen, D., Feiner, S., Jacobs, D.W., Kress, W.J., Ling, H., Lopez, I., Ramamoorthi, R., Sheorey, S., White, S., Zhang, L.: Searching the world's herbaria: A system for visual identification of plant species. In: Proceedings of the European Conference on Computer Vision, pp. 116–129. Springer (2008). DOI 10.1007/978-3-540-88693-8_9
6. Berg, T., Liu, J., Woo Lee, S., Alexander, M.L., Jacobs, D.W., Belhumeur, P.N.: Birdsnap: Large-scale fine-grained visual categorization of birds. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2011–2018. IEEE (2014). DOI 10.1109/CVPR.2014.259
7. Choate, P.M.: Introduction to the identification of beetles (coleoptera). Dichotomous keys to some Families of Florida Coleoptera pp. 23–33 (1999)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009). DOI 10.1109/CVPR.2009.5206848
9. Dietterich, T.G.: Approximate statistical tests for comparing supervised classification learning algorithms. Neural computation **10**(7), 1895–1923 (1998). DOI 10.1162/089976698300017197
10. Drew, J.A., Moreau, C.S., Stiassny, M.L.: Digitization of museum collections holds the potential to enhance researcher diversity. Nature ecology & evolution **1**(12), 1789 (2017)
11. Ereshefsky, M.: The poverty of the Linnaean hierarchy: A philosophical study of biological taxonomy. Cambridge University Press (2000). DOI 10.1017/CBO9780511498459
12. Ferrari, V., Zisserman, A.: Learning visual attributes. In: Advances in Neural Information Processing Systems, pp. 433–440 (2007)
13. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: Devise: A deep visual-semantic embedding model. In: Advances in Neural Information Processing Systems, pp. 2121–2129 (2013)
14. Ge, W., Dong, D., Scott, M.R.: Deep metric learning with hierarchical triplet loss. In: Proceedings of the European Conference on Computer Vision, pp. 272–288. Springer (2018). DOI 10.1007/978-3-030-01231-1_17
15. Gwinn, N.E., Rinaldo, C.: The biodiversity heritage library: sharing biodiversity literature with the world. IFLA journal **35**(1), 25–34 (2009). DOI 10.1177/0340035208102032
16. Harris, Z.S.: Distributional structure. Word **10**(2-3), 146–162 (1954). DOI 10.1080/00437956.1954.11659520
17. Hedrick, B.P., Heberling, J.M., Meineke, E.K., Turner, K.G., Grassa, C.J., Park, D.S., Kennedy, J., Clarke, J.A., Cook, J.A., Blackburn, D.C., Edwards, S.V., Davis, C.C.: Digitization and the future of natural history collections. BioScience **70**(3), 243–251 (2020). DOI 10.1093/biosci/biz163
18. Kumar, N., Belhumeur, P.N., Biswas, A., Jacobs, D.W., Kress, W.J., Lopez, I.C., Soares, J.V.B.: Leafsnap: A computer vision system for automatic plant species identification. In: Proceedings of the European Conference on Computer Vision, pp. 502–516. Springer (2012). DOI 10.1007/978-3-642-33709-3_36
19. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 951–958. IEEE (2009). DOI 10.1109/CVPR.2009.5206594
20. Lampert, C.H., Nickisch, H., Harmeling, S.: Attribute-based classification for zero-shot visual object categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(3), 453–465 (2013). DOI 10.1109/TPAMI.2013.140

21. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov), 2579–2605 (2008)
22. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Workshop proceedings of the International Conference on Learning Representations. arXiv preprint arXiv:1301.3781 (2013)
23. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, vol. 2, pp. 3111–3119 (2013)
24. Nguyen, N.T.H., Soto, A.J., Kontonatsios, G., Batista-Navarro, R., Ananiadou, S.: Constructing a biodiversity terminological inventory. PLoS ONE **12**(4), e0175277 (2017). DOI 10.1371/journal.pone.0175277
25. Nilsback, M.E., Zisserman, A.: A visual vocabulary for flower classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 1447–1454. IEEE (2006). DOI 10.1109/CVPR.2006.42
26. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1717–1724. IEEE (2014). DOI 10.1109/CVPR.2014.222
27. Patterson, G., Hays, J.: Sun attribute database: Discovering, annotating, and recognizing scene attributes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2751–2758. IEEE (2012). DOI 10.1109/CVPR.2012.6247998
28. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1532–1543. Association for Computational Linguistics (2014). DOI 10.3115/v1/D14-1162
29. Romera-Paredes, B., Torr, P.H.S.: An embarrassingly simple approach to zero-shot learning. In: Visual Attributes, pp. 11–30. Springer (2017). DOI 10.1007/978-3-319-50077-5_2
30. Secretariat, G.: Gbif backbone taxonomy. Checklist Dataset https://doi. org/10.15468/39omei (2017)
31. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems, pp. 4077–4087 (2017)
32. Socher, R., Ganjoo, M., Manning, C.D., Ng, A.: Zero-shot learning through cross-modal transfer. In: Advances in Neural Information Processing Systems, pp. 935–943 (2013)
33. Sumbul, G., Cinbis, R.G., Aksoy, S.: Fine-grained object recognition and zero-shot learning in remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing **56**(2), 770–779 (2018). DOI 10.1109/TGRS.2017.2754648
34. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015). DOI 10.1109/CVPR.2015.7298594
35. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826. IEEE (2016). DOI 10.1109/CVPR.2016.308
36. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. Journal of Machine Learning Research **6**(Sep), 1453–1484 (2005)
37. Turney, P.D., Pantel, P.: From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research **37**(1), 141–188 (2010)
38. Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., Perona, P., Belongie, S.: Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 595–604. IEEE (2015). DOI 10.1109/CVPR.2015.7298658
39. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8769–8778. IEEE (2018). DOI 10.1109/CVPR.2018.00914
40. Wah, C., Branson, S., Perona, P., Belongie, S.: Multiclass recognition and part localization with humans in the loop. In: Computer Vision (ICCV), 2011 IEEE International Conference on, pp. 2524–2531. IEEE (2011)
41. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
42. Wang, M., Deng, W.: Deep visual domain adaptation: A survey. Neurocomputing **312**, 135–153 (2018). DOI 10.1016/j.neucom.2018.05.083
43. Weber, A.: Collecting colonial nature: European naturalists and the netherlands indies in the early nineteenth century. BMGN-Low Countries Historical Review **134**(3) (2019). DOI 10.18352/bmgn-lchr.10741
44. Xian, Y., Akata, Z., Sharma, G., Nguyen, Q., Hein, M., Schiele, B.: Latent embeddings for zero-shot classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 69–77 (2016). DOI 10.1109/CVPR.2016.15
45. Xian, Y., Lampert, C.H., Schiele, B., Akata, Z.: 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). DOI 10.1109/TPAMI.2018.2857768