Bringing Back the Context: Camera Trap Species Identification as Link Prediction on Multimodal Knowledge Graphs

Vardaan Pahuja[1] Weidi Luo[1] Yu Gu[1] Cheng-Hao Tu[1] Hong-You Chen[1]
Tanya Berger-Wolf[1] Charles Stewart[2] Song Gao[3] Wei-Lun Chao[1] Yu Su[1]

[1]The Ohio State University [2]Rensselaer Polytechnic Institute
[3]University of Wisconsin-Madison

Abstract

Camera traps are valuable tools in animal ecology for biodiversity monitoring and conservation. However, challenges like poor generalization to deployment at new unseen locations limit their practical application. In this work, we leverage the structured context, such as spatiotemporal data and biological taxonomy associated with the camera trap images, to improve out-of-distribution generalization for species identification in camera traps. For example, a photo of a wild animal may be associated with information about where and when it was taken, as well as structured biology knowledge about the animal species. While typically overlooked by existing work, bringing back such context offers several potential benefits for better image understanding, such as addressing data scarcity and enhancing generalization. To effectively integrate such heterogeneous contexts into the visual domain in a unified way, we propose a novel framework that reformulates species classification as link prediction in a multimodal knowledge graph (KG). We apply this framework for out-of-distribution species classification on iWildCam2020-WILDS dataset and achieve competitive performance with state-of-the-art approaches.[1]

1. Introduction

Human activities are increasingly endangering wildlife species, resulting in a significant global decline in animal populations [2, 18, 35]. Therefore, accurately identifying and tracking wildlife species is vital for preserving ecological biodiversity. The use of camera traps [22, 42, 63] for data collection has led to the increased use of computer vision techniques for species recognition [1, 13, 28, 50, 53, 64]. Yet, a challenge has arisen: many of these models overfit to the backgrounds of their training images, diminishing their effectiveness on images from new locations [7, 37, 54]. This underscores the need for *more adaptable species classification models that perform well across diverse **contexts***.

Building on this, cognitive science research has demonstrated the profound influence of ***contextual*** information on human perception and visual recognition processes [4, 5, 43]. Particularly in wildlife monitoring, camera trap images are replete with crucial contextual data, such as where (*i.e*., camera location coordinates) and when (*i.e*., timestamps) a photo is taken. Furthermore, the structured knowledge of biology taxonomy (*e.g*., Open Tree Taxonomy [44]) can also provide valuable context for understanding the species in camera trap images. Such context provides important knowledge that can boost the recognition of visual concepts. For instance, the knowledge that a certain feline image was taken from a camera trap in Africa significantly reduces the likelihood of it representing a tiger. In addition, more robust associations might be learned with the aid of contextual information because the context provides invariable knowledge that is unbiased towards variations in the illuminations or angles of an image. This may help to compensate for domain shifts in species images resulting from such variations and potentially lead to better out-of-distribution (OOD) generalizability [6, 20].

Nevertheless, contextual information has been underexploited in the literature of image classification. Contextual information in different modalities (*e.g*., numerical values, textual descriptions, or structured taxonomies) is usually represented separately from the image in *distinct feature spaces*. The question of effectively combining features from these different spaces within a unified learning framework remains unanswered. Existing research typically treats all the features as additional input to the classifier via feature vector concatenation [6, 20, 30] or utilizes fusion to obtain aggregate representations [15, 17]. Despite their simplicity, such approaches are incapable of capturing complex structural and semantic relationships between images and various

[1]Our code is available at `https://github.com/OSU-NLP-Group/COSMO`

contextual information. Additionally, these approaches assume a uniform availability of contextual information across all images, which is often unrealistic in real-world scenarios. As a result, their flexibility is limited, especially when considering situations where certain images may lack some contextual details, such as coordinates or timestamps, like in camera trap photos.

Towards this end, we propose a new learning framework, COSMO (Classification Of Species using Multimodal cOntext), where we first organize all species images and contextual information as a *multimodal knowledge graph* (KG) and then reformulate species classification as the standard link prediction task on the KG. Specifically, we consider species images, their corresponding labels (which are available in the training data), and their associated attributes provided in the context as entities within our KG (see Figure 1 for an example). We represent the relationships between these entities as edges in our KG (see a more concrete description of our KG construction in Section 2.2). In this context, species classification can be framed as a link prediction task, where the objective is to predict the presence of an edge between an image and its corresponding species label within the KG. The learning process enables the interaction of different modalities in a *joint feature space* for robust representation learning. In addition, COSMO demonstrates greater flexibility by not assuming uniform availability of all contextual information, unlike previous methods.

The main contribution of this work is three-fold:

- We propose a novel framework, COSMO, that reformulates species classification as link prediction in a multimodal knowledge graph, which provides a unified way to incorporate heterogeneous forms of contextual information associated with images for visual recognition.
- We instantiate this framework for wildlife species classification, including the construction of a novel multimodal knowledge graph that integrates spatiotemporal information and structured biology knowledge.
- Evaluation on the iWildCam2020-WILDS dataset demonstrates that COSMO achieves competitive performance compared with standard species classification methods, especially in improving robustness and OOD generalization.

2. Methodology

2.1. Preliminaries

Multimodal KG. Given a set of KG entities with categorical values $\mathcal{E}_{\mathcal{KG}}$, multimodal entities $\mathcal{E}_{\mathcal{MM}}$, and a set of relations $\mathcal{R}$, a multimodal KG can be defined as a collection of facts $\mathcal{F} \subseteq (\mathcal{E}_{\mathcal{KG}} \cup \mathcal{E}_{\mathcal{MM}}) \times \mathcal{R} \times (\mathcal{E}_{\mathcal{KG}} \cup \mathcal{E}_{\mathcal{MM}})$ where for each fact $f = (h, r, t),\ h, t \in (\mathcal{E}_{\mathcal{KG}} \cup \mathcal{E}_{\mathcal{MM}}), r \in \mathcal{R}$.

KG Link Prediction. The task of link prediction is to infer missing facts based on known facts in a KG. Given a link prediction query $(h, r, ?)$ or $(?, r, t)$, the model ranks the target entity among the set of candidate entities.

Problem Setup. The task entails species recognition for camera trap images amidst distribution shifts. The training and test sets comprise images obtained from disjoint camera traps. During training, we use the multimodal KG to train our model, while we use just the image to make predictions for inference. The goal is to learn visual representations robust to distribution shifts by leveraging the rich structural and semantic information provided by the multimodal KG.

2.2. Building the Multimodal KG

The multimodal KG comprises entities from different modalities interconnected by heterogeneous relationships. The base KG consists of camera trap images linked with their species labels from the training set (`<image>`, `instance of`, `<species label>`). Next, we progressively augment the KG with links connecting the existing entities to contextual information. In this work, we utilize the following attributes to provide context for species classification:

- **Taxonomy**: The taxonomy forms the core of the multimodal knowledge graph, connecting distinct species to higher-order taxa. For iWildCam2020-WILDS, we obtain the phylogenetic taxonomy corresponding to the species of interest from Open Tree Taxonomy (OTT) [44] and manually link it to the species in the dataset.
- **Location**: The camera trap images are associated with the GPS coordinates of their source cameras. For iWildCam2020-WILDS, this metadata is available for a portion of the images (67%) and is obfuscated within 1 km. for privacy reasons. Animals demonstrate a preference for particular habitats; thus, the location context attribute is useful for species recognition.
- **Time**: This timestamp information proves valuable in species recognition since specific animals exhibit activity patterns tied to particular times of the day, such as feeding, hunting, or defending their territory. In our multimodal knowledge graph, we utilize the timestamp information at an hourly granularity.

Figure 1 presents a schematic representation of various contexts in a multimodal KG. For location, time, and taxonomy attributes, the corresponding RDF triples can be represented as (`<image>`, `location`, `<GPS co-ordinate>`), (`<image>`, `time`, `<timestamp>`), and (`<taxon_1>`, `parent`, `<taxon_2>`), respectively.

2.3. Model Architecture

We use DistMult [66], a strong baseline on KGE benchmarks, as our backbone KG embedding model.[2] Note that COSMO is a general framework that can leverage a variety of KG embedding models proposed in the literature. DistMult

[2] Recent work [49] showed that simple baselines like DistMult outperform more sophisticated neural network baselines when trained properly.

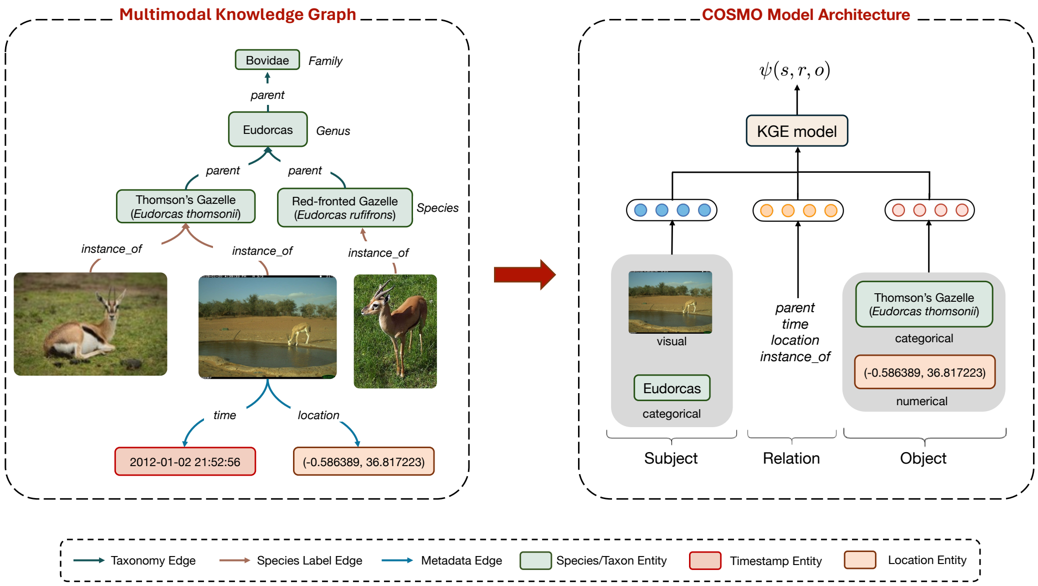


Figure 1. **Overview of our framework COSMO.** ***Left:*** Our multimodal knowledge graph for camera traps and wildlife. Photos from camera traps are jointly represented in the KG with contextual information such as time, location, and structured biology taxonomy. The taxonomy is obtained from Open Tree Taxonomy (OTT) [44]. ***Right:*** In our formulation of species classification as link prediction, the plausibility score $\psi(s, r, o)$ of each (subject, relation, object) triple is computed using a KGE model (*e.g.*, DistMult), where the subject, relation, and object are all first embedded into a vector space. Specifically, for our multimodal KG, we represent visual entities using a ResNet-50 pre-trained on ImageNet and represent numerical entities using an MLP. For categorical entities and relations, we directly represent them with embedding lookups

minimizes a bilinear scoring function between the entity embeddings of subject and object entities. For a given triplet (h, r, t), the scoring function of DistMult is defined as

$$\psi(h, r, t) = \boldsymbol{h}^T \boldsymbol{W}_r \boldsymbol{t} = \sum_{i=1}^{d} \boldsymbol{h}_i \cdot diag(\boldsymbol{W}_r)_i \cdot \boldsymbol{t}_i \quad (1)$$

Here, $\boldsymbol{h}$ and $\boldsymbol{t}$ denote the vector representations of the head entity and tail entity, respectively. The relation representation is parameterized by $\boldsymbol{W}_r \in \mathbb{R}^{d \times d}$, a diagonal matrix.

2.3.1 Multi-modality Encoders

We use an ImageNet pre-trained ResNet-50 [23] as the image encoder. The base feature of each location is represented as a 2D vector `[latitude, longitude]`. Following prior work [47], we use an MLP to project the 2D location feature to a higher dimensional space. Similarly, for temporal context, we use an MLP to project the integer value of the hour timestamp to the higher dimensional embedding space. For categorical entities such as species labels and taxa, we learn dense embeddings as representations.

2.3.2 Training

We train the model using an optimization strategy based on the modality of the tail entity. For categorical attributes, we formulate it as a multi-class classification problem and use standard cross-entropy loss to train the model. For instance, in case of a given image-species label ground truth triple $(\mathcal{I}, \texttt{instance of}, s)$, the loss is defined as $\mathcal{L}(\mathcal{I}, \texttt{io}, s) = -\log \frac{\exp(\psi(\mathcal{I}, \texttt{io}, s))}{\sum_{s' \in S} \exp(\psi(\mathcal{I}, \texttt{io}, s'))}$, where S denotes the set of all species labels, and `io` denotes the relation `instance of`.

For numerical attributes such as location and time, we formulate it as a multi-class multi-label classification problem and use a binary cross-entropy loss to optimize the parameters. This choice is motivated by the fact that images can be associated with a range of GPS coordinates and timestamps, *e.g.*, most animals are active multiple times during the day. The label space comprises all entities of ground truth modality. For instance, in the case of a given time modality ground truth triple $(\mathcal{I}, \texttt{time}, t)$, the loss is defined as:

$$\mathcal{L}(\mathcal{I}, \texttt{time}, t) = -\sum_{t'} l_{t'}^{\mathcal{I}, time} \cdot \log(\sigma(\psi(\mathcal{I}, \texttt{time}, t'))) + (1 - l_{t'}^{\mathcal{I}, time}) \cdot (1 - \log(\sigma(\psi(\mathcal{I}, \texttt{time}, t')))),$$

where $l_{t'}^{\mathcal{I}, time}$ is a binary label that indicates whether the triple $(\mathcal{I}, time, t')$ exists in the set of observed triples and $\sigma(\cdot)$ is the sigmoid activation function. We train the model by sequentially minimizing the objective on each type of context triple. Figure 1 illustrates the overall model architecture.

	Model	Multi-modality			Val. Acc. (%)	Test Acc. (%)
		Taxonomy	Location	Time		
	Empirical Risk Minimization (ERM) [28]				62.7 (±2.4)	71.6 (±2.5)
	CORAL [58]				60.3 (±2.8)	73.3 (±4.3)
	Group DRO [24]		–		60.0 (±0.7)	72.7 (±2.0)
	Fish [56]				58.0 (±0.2)	63.2 (±0.7)
	ABSGD [48]				–	72.7 (±1.8)
	MLP-concat		✓	✓	27.3 (±0.8)	39.6 (±1.0)
	COSMO (no-context)		–		63.2 (±0.4)	68.8 (±2.1)
Single context		✓			62.8 (±2.2) (-0.4)	72.4 (±2.5) (+3.6)
	COSMO		✓		64.4 (±1.0) (+1.2)	**74.5 (±3.6)** (+5.7)
				✓	64.7 (±0.4) (+1.5)	71.1 (±3.1) (+2.3)
Multiple contexts		✓	✓		**65.4 (±0.4)** (+2.2)	70.4 (±2.1) (+1.6)
	COSMO	✓		✓	<u>64.9 (±1.6)</u> (+1.7)	73.7 (±3.8) (+4.9)
			✓	✓	63.0 (±2.1) (-0.2)	<u>74.2 (±2.2)</u> (+5.4)
		✓	✓	✓	65.0 (±1.6) (+1.8)	71.5 (±2.8) (+2.7)

Table 1. Species Classification results on iWildCam2020-WILDS (OOD) dataset. The first baseline in the second section shows the no-context baseline that uses only image-species labels as KG edges. All models use a pre-trained ResNet-50 as image encoder. Parentheses show standard deviation across 3 random seeds. We highlight the best result in bold and the second best with underline. We mark the improvements over COSMO (no-context) in green. Missing values are denoted by –.

3. Experimental Setup

3.1. Datasets

We test our approach on the iWildCam2020-WILDS dataset [28], a variant of the iWildCam 2020 dataset [9]. iWildCam2020-WILDS is a benchmark dataset designed to test out-of-distribution (OOD) generalization for the task of species classification. It consists of wildlife images collected from camera traps, heat or motion-activated cameras placed in the wild [63]. Each domain corresponds to a different location of the camera trap. The training and test images belong to disjoint sets of locations in the OOD setting.

3.2. Baselines

We use the COSMO with no context that uses just the species label edges as our baseline. In addition, we compare with the following baseline algorithms for OOD generalization: Empirical Risk Minimization (ERM) [28], which trains the model to minimize average training loss, CORAL [58], a method for unsupervised domain adaptation that learns domain invariant features, Group DRO [24], an algorithm that uses distributionally robust optimization to perform well on subpopulation shifts, Fish [56] that attempts domain adaptation using gradient matching, and ABSGD [48], an optimization method for addressing data imbalance. As an alternative way of incorporating contextual information, we implement MLP-concat, a baseline which utilizes the location and temporal features at both training and inference time. It uses vanilla concatenation to fuse visual and spatiotemporal representations which are then fed into an MLP. The missing features are substituted by a mean value computed over the training dataset. All models use a pre-trained ResNet-50 as image encoder. We evaluate the models using overall accuracy as the metric.

4. Results

4.1. Performance Comparison with Addition of Multimodal Context

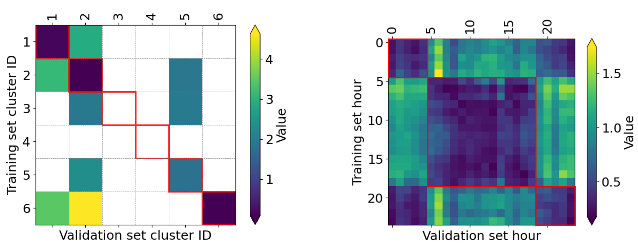
We add taxonomy, location, and temporal context information to the base KG and observe the impact on the species classification performance. Table 1 shows the results for the iWildCam2020-WILDS dataset. We make the following observations from these results:

Firstly, the addition of one or more contexts results in a performance gain over the no-context baseline in the vast majority of cases. For instance, in the case of COSMO with taxonomy, we obtain a 3.6% improvement over the no-context baseline in terms of test accuracy. Incorporating location context produces a notable 5.7% enhancement in test set accuracy, underlining the significance of auxiliary information for improved out-of-domain generalization. We further analyze the role of location in predicting the species distribution in Section 4.3. Additionally, utilizing the time attribute yields a substantial improvement over the no-context baseline, resulting in a 2.3% performance gain.

Secondly, we observe that the use of multiple contexts results in a performance boost in a majority of cases. For instance, the addition of location and time attributes improves over the taxonomy baseline by a margin of 2.6% and 2.1% respectively in terms of validation set accuracy. Similarly, the taxonomy with time baseline obtains an improvement of 1.3% and 2.6% over the taxonomy and time baselines, respectively in terms of test accuracy. Please refer to the supplementary material for additional results and analysis.

4.2. Comparison with OOD Generalization Approaches

We compare the performance of the COSMO with methods specifically designed for out-of-domain generalization. Notably, our best-performing model, which uses location as context, achieves state-of-the-art performance in terms of OOD test accuracy, outperforming the existing SOTA model (CORAL) by 1.2% on the iWildCam2020-WILDS dataset. This demonstrates the effectiveness of leveraging diverse multimodal contexts for achieving more robust OOD generalization, even in the absence of sophisticated objectives aimed at improving domain generalization, *e.g.*, CORAL [58], Group DRO [24], ABSGD [48], and Fish [56]. The MLP-concat baseline overfits to the training camera trap locations on the iWildCam2020-WILDS dataset, resulting in suboptimal performance. COSMO outperforms the MLP-concat baseline by a significant margin.



(a) Each color square shows the distance between the corresponding validation cluster centroid on x-axis and the training cluster centroid on y-axis. The correlation peaks along the diagonal (highlighted in red)[3].

(b) Each color square shows the distance between the corresponding training hour slot on x-axis and validation hour slot on y-axis. The correlation peaks for day-day and night-night hour slots (highlighted in red).

Figure 2. Correlation analysis for location and time attributes. Best viewed in color.

4.3. Correlation Analysis for Location and Time Attributes

We examined the relationship between species distribution and numerical attributes, such as location and time, to gain insights into how these contexts contribute to the task. The location coordinates can be grouped into six clusters. For each pair of cluster centroids, we compute the Bhattacharyya distance [11], a measure of similarity between probability distributions, between the training and validation set species distributions (Figure 2a). Similarly, we plot the distance between species distributions corresponding to each hour of the day (Figure 2b). We observe that the similarity (corresponds to lower distance) peaks along the diagonal for the location attribute, as well as for the day/night categorization of the time attribute. This suggests these metadata give a prior for species class distribution.

5. Discussion and Conclusion

In this work, we presented a novel framework in which the species classification task is reformulated as link prediction in a multimodal KG of species images and their diverse contextual information. This enables a unified way to leverage various forms of multimodal context, *e.g.*, numerical, categorical, and taxonomy information associated with images for species classification in camera traps. Through our experiments, we demonstrate that our framework achieves superior out-of-distribution generalization and competitive performance with state-of-the-art for species classification on the iWildCam2020-WILDS dataset.

We assume that there is a perfect linkage between these contexts and the corresponding images in the training set. However, in scenarios where such linkage is unavailable, the training procedure may introduce noise, which could lead to inferior generalization capabilities in the model. Additionally, it is important to note that the effectiveness of diverse contexts varies based on their informativeness for the given task. Interestingly, combining two or more contexts could degrade performance compared to using a single context type in some cases (Table 1). We posit that specific metadata, like location, might have a stronger regularization effect on improving generalization in species recognition tasks than other metadata. To address this, future work will involve enabling the model to assign greater importance to more informative metadata.

Furthermore, we are interested in training a foundation model for camera trap species classification across a wider spectrum of species. This model should demonstrate enhanced generalization capabilities for new camera trap setups worldwide. Additionally, we aim to integrate a broader spectrum of diverse contexts such as temperature, weather conditions, habitat, and sequence information for use with real-world camera trap deployments.

[3] The null value in row 4 is due to the absence of species overlap with respective validation clusters. The null value in columns 3 and 4 indicates the absence of these clusters in the validation set.

References

[1] Jorge A Ahumada, Eric Fegraus, Tanya Birch, Nicole Flores, Roland Kays, Timothy G O'Brien, Jonathan Palmer, Stephanie Schuttler, Jennifer Y Zhao, Walter Jetz, et al. Wildlife insights: A platform to maximize the potential of camera trap and other passive sensor wildlife data for the planet. *Environmental Conservation*, 47(1):1–6, 2020. 1

[2] Rosamund EA Almond, Monique Grooten, and T Peterson. *Living Planet Report 2020-Bending the curve of biodiversity loss*. World Wildlife Fund, 2020. 1

[3] Bilal Alsallakh, Amin Jourabloo, Mao Ye, Xiaoming Liu, and Liu Ren. Do convolutional neural networks learn class hierarchy? *IEEE Trans. Vis. Comput. Graph.*, 24(1):152–162, 2018. 11

[4] Elissa Aminoff, Nurit Gronau, and Moshe Bar. The parahippocampal cortex mediates spatial and nonspatial associations. *Cerebral cortex*, 17(7):1493–1503, 2007. 1

[5] Moshe Bar. Visual objects in context. *Nature Reviews Neuroscience*, 5(8):617–629, 2004. 1

[6] Suchet Bargoti and James Underwood. Image classification with orchard metadata. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5164–5170. IEEE, 2016. 1, 11

[7] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018. 1, 10

[8] Sara Beery, Grant Van Horn, Oisin Mac Aodha, and Pietro Perona. The iwildcam 2018 challenge dataset. *arXiv preprint arXiv:1904.05986*, 2019. 10

[9] Sara Beery, Elijah Cole, and Arvi Gjoka. The iwildcam 2020 competition dataset. *CoRR*, abs/2004.10340, 2020. 4, 9

[10] Luca Bertinetto, Romain Müller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A. Lord. Making better mistakes: Leveraging class hierarchies with deep networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 12503–12512. Computer Vision Foundation / IEEE, 2020. 10, 11

[11] Anil Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhyā: the indian journal of statistics*, pages 401–406, 1946. 5

[12] Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795, 2013. 12

[13] Ludwig Bothmann, Lisa Wimmer, Omid Charrakh, Tobias Weber, Hendrik Edelhoff, Wibke Peters, Hien Nguyen, Caryl Benjamin, and Annette Menzel. Automated wildlife image classification: An active learning tool for ecological applications. *CoRR*, abs/2303.15823, 2023. 1, 10

[14] Sanxing Chen, Xiaodong Liu, Jianfeng Gao, Jian Jiao, Ruofei Zhang, and Yangfeng Ji. HittER: Hierarchical transformers for knowledge graph embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10395–10407, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 12

[15] Grace Chu, Brian Potetz, Weijun Wang, Andrew Howard, Yang Song, Fernando Brucher, Thomas Leung, and Hartwig Adam. Geo-aware networks for fine-grained recognition. In *2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*, pages 247–254. IEEE, 2019. 1

[16] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings. In *Thirty-second AAAI conference on artificial intelligence*, 2018. 12

[17] Qishuai Diao, Yi Jiang, Bin Wen, Jia Sun, and Zehuan Yuan. Metaformer: A unified meta framework for fine-grained recognition. *arXiv preprint arXiv:2203.02751*, 2022. 1, 11

[18] Sandra Díaz, Josef Settele, Eduardo S Brondízio, Hien T Ngo, John Agard, Almut Arneth, Patricia Balvanera, Kate A Brauman, Stuart HM Butchart, Kai MA Chan, et al. Pervasive human-driven decline of life on earth points to the need for transformative change. *Science*, 366(6471):eaax3100, 2019. 1

[19] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 601–610, 2014. 12

[20] Jeffrey S Ellen, Casey A Graff, and Mark D Ohman. Improving plankton image classification using context metadata. *Limnology and Oceanography: Methods*, 17(8):439–461, 2019. 1, 11

[21] Alberto García-Durán and Mathias Niepert. Kblrn: End-to-end learning of knowledge base representations with latent, relational, and numerical features. In *Conference on Uncertainty in Artificial Intelligence*, 2018. 12

[22] Paul Glover-Kapfer, Carolina A Soto-Navarro, and Oliver R Wearn. Camera-trapping version 3.0: current constraints and future priorities for development. *Remote Sensing in Ecology and Conservation*, 5(3):209–223, 2019. 1

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 3

[24] Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 2034–2042. PMLR, 2018. 4, 5, 9, 10

[25] Mirantha Jayathilaka, Tingting Mu, and Uli Sattler. Ontology-based n-ball concept embeddings informing few-shot image classification. In *Machine Learning with Symbolic Methods and Knowledge Graphs co-located with European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2021), Virtual, September 17, 2021*. CEUR-WS.org, 2021. 11

[26] Justin Johnson, Lamberto Ballan, and Li Fei-Fei. Love thy neighbors: Image annotation by exploiting image metadata. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 4624–4632. IEEE Computer Society, 2015. 11

[27] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 9

[28] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran S. Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 5637–5664. PMLR, 2021. 1, 4, 9, 10

[29] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, 2017. 11

[30] Wen Li, Li Niu, and Dong Xu. Exploiting privileged information from web data for image categorization. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 437–452. Springer, 2014. 1, 11

[31] Xinhang Li, Xiangyu Zhao, Jiaxing Xu, Yong Zhang, and Chunxiao Xing. Imf: Interactive multimodal fusion model for link prediction. In *Proceedings of the ACM Web Conference 2023*, pages 2572–2580, 2023. 12

[32] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2181–2187. AAAI Press, 2015. 12

[33] Chengjiang Long, Roddy Collins, Eran Swears, and Anthony Hoogs. Deep neural networks in fully connected CRF for image labeling with social network metadata. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, January 7-11, 2019*, pages 1607–1615. IEEE, 2019. 11

[34] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. The more you know: Using knowledge graphs for image classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 20–28. IEEE Computer Society, 2017. 11

[35] Sean L Maxwell, Richard A Fuller, Thomas M Brooks, and James EM Watson. Biodiversity: The ravages of guns, nets and bulldozers. *Nature*, 536(7615):143–145, 2016. 1

[36] Julian J. McAuley and Jure Leskovec. Image labeling on a network: Using social-network metadata for image classification. In *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV*, pages 828–841. Springer, 2012. 11

[37] Zhongqi Miao, Kaitlyn M Gaynor, Jiayun Wang, Ziwei Liu, Oliver Muellerklein, Mohammad Sadegh Norouzzadeh, Alex McInturff, Rauri CK Bowie, Ran Nathan, Stella X Yu, et al. Insights and approaches using deep learning to classify wildlife. *Scientific reports*, 9(1):8137, 2019. 1, 10

[38] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 11

[39] Dai Quoc Nguyen, Tu Dinh Nguyen, Dat Quoc Nguyen, and Dinh Phung. A novel embedding model for knowledge base completion based on convolutional neural network. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 327–333, New Orleans, Louisiana, 2018. Association for Computational Linguistics. 12

[40] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 809–816. Omnipress, 2011. 12

[41] Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S Palmer, Craig Packer, and Jeff Clune. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115(25):E5716–E5725, 2018. 10

[42] Allan F O'Connell, James D Nichols, and K Ullas Karanth. *Camera traps in animal ecology: methods and analyses*. Springer, 2011. 1

[43] Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends in cognitive sciences*, 11(12):520–527, 2007. 1

[44] OpenTreeofLife, Karen A. Cranston, Benjamin Redelings, Luna Luisa Sanchez Reyes, Jim Allman, Emily Jane McTavish, and Mark T. Holder. Open tree of life taxonomy, 2019. 1, 2, 3, 9

[45] Vardaan Pahuja, Boshi Wang, Hugo Latapie, Jayanth Srinivasa, and Yu Su. A retrieve-and-read framework for knowledge graph link prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, page 1992–2002, New York, NY, USA, 2023. Association for Computing Machinery. 12

[46] Lain E Pardo, Sara Bombaci, Sarah E Huebner, Michael J Somers, Herve Fritz, Colleen Downs, Abby Guthmann, Robyn S Hetem, Mark Keith, Aliza le Roux, et al. Snapshot safari: A large-scale collaborative to monitor africa's remarkable biodiversity. *South African Journal of Science*, 117(1-2):1–4, 2021. 9

[47] Pouya Pezeshkpour, Liyan Chen, and Sameer Singh. Embedding multimodal relational data for knowledge base completion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3208–3218, Brussels, Belgium, 2018. Association for Computational Linguistics. 3, 12

[48] Qi Qi, Yi Xu, Wotao Yin, Rong Jin, and Tianbao Yang. Attentional-biased stochastic gradient descent. *Transactions on Machine Learning Research*, 2023. 4, 5, 9, 10

[49] Daniel Ruffinelli, Samuel Broscheit, and Rainer Gemulla. You can teach an old dog new tricks! on training knowledge graph embeddings. In *International Conference on Learning Representations*, 2020. 2

[50] Mohammad Sadegh Norouzzadeh, Dan Morris, Sara Beery, Neel Joshi, Nebojsa Jojic, and Jeff Clune. A deep active learning system for species identification and counting in camera trap images. *Methods in Ecology and Evolution*, 12 (1):150–161, 2020. 1, 10

[51] Apoorv Saxena, Adrian Kochsiek, and Rainer Gemulla. Sequence-to-sequence knowledge graph completion and question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2814–2828, Dublin, Ireland, 2022. Association for Computational Linguistics. 12

[52] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, pages 593–607. Springer, 2018. 12

[53] Stefan Schneider, Graham W. Taylor, and Stefan C. Kremer. Deep learning object detection methods for ecological camera trap data. In *15th Conference on Computer and Robot Vision, CRV 2018, Toronto, ON, Canada, May 8-10, 2018*, pages 321–328. IEEE Computer Society, 2018. 1

[54] Stefan Schneider, Saul Greenberg, Graham W Taylor, and Stefan C Kremer. Three critical factors affecting automated image species recognition performance for camera traps. *Ecology and evolution*, 10(7):3503–3517, 2020. 1

[55] Hatem Mousselly Sergieh, Teresa Botschen, Iryna Gurevych, and Stefan Roth. A multimodal translation-based approach for knowledge graph representation learning. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, pages 225–234. Association for Computational Linguistics, 2018. 12

[56] Yuge Shi, Jeffrey Seely, Philip Torr, Siddharth N, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. In *International Conference on Learning Representations*, 2022. 4, 5

[57] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. Bioclip: A vision foundation model for the tree of life. *arXiv preprint arXiv:2311.18803*, 2023. 11

[58] Baochen Sun and Kate Saenko. Deep CORAL: correlation alignment for deep domain adaptation. In *Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III*, pages 443–450, 2016. 4, 5, 9, 10

[59] Michael A Tabak, Mohammad S Norouzzadeh, David W Wolfson, Steven J Sweeney, Kurt C VerCauteren, Nathan P Snow, Joseph M Halseth, Paul A Di Salvo, Jesse S Lewis, Michael D White, et al. Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution*, 10(4):585–590, 2019. 10

[60] Michael A Tabak, Mohammad S Norouzzadeh, David W Wolfson, Erica J Newton, Raoul K Boughton, Jacob S Ivan, Eric A Odell, Eric S Newkirk, Reesa Y Conrey, Jennifer Stenglein, et al. Improving the accessibility and transferability of machine learning algorithms for identification of animals in camera trap images: Mlwic2. *Ecology and evolution*, 10 (19):10374–10383, 2020. 10

[61] Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha P. Talukdar. Composition-based multi-relational graph convolutional networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 12

[62] Jean-Christophe Vié, Craig Hilton-Taylor, Caroline Pollock, James Ragle, Jane Smart, Simon N Stuart, and Rashila Tong. The iucn red list: a key conservation tool. *Wildlife in a changing world–An analysis of the 2008 IUCN Red List of Threatened Species*, page 1, 2009. 10

[63] OR Wearn and P Glover-Kapfer. Camera-trapping for conservation: a guide to best-practices. *WWF conservation technology series*, 1(1):181, 2017. 1, 4

[64] Ben G Weinstein. A computer vision for animal ecology. *Journal of Animal Ecology*, 87(3):533–545, 2018. 1, 10

[65] Xander Wilcke, Peter Bloem, Victor de Boer, and Rein van 't Veer. End-to-end learning on multimodal knowledge graphs. *Semantic Web – Interoperability, Usability, Applicability an IOS Press Journal*, 2021. 12

[66] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 2, 12

[67] Liang Yao, Chengsheng Mao, and Yuan Luo. Kg-bert: Bert for knowledge graph completion. *arXiv preprint arXiv:1909.03193*, 2019. 12

[68] Donghan Yu, Yiming Yang, Ruohong Zhang, and Yuexin Wu. Knowledge embedding based graph convolutional network. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 1619–1628. ACM / IW3C2, 2021. 12

[69] Shu Zhang, Ran Xu, Caiming Xiong, and Chetan Ramaiah. Use all the labels: A hierarchical multi-label contrastive learning framework. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16639–16648. IEEE, 2022. 11