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

Camera Trap Wildlife Recognition Challenges

Illumination: Animals are not always salient

Image Credit: Beery et al. 2020

Camera Trap Wildlife Recognition Challenges

Motion blur: common with poor illumination at night

Image Credit: Beery et al. 2020

Camera Trap Wildlife Recognition Challenges

Size of the region of interest (ROI): Animals can be small or far from the camera

Image Credit: Beery et al. 2020

Camera Trap Wildlife Recognition Challenges

Occlusion: e.g. by bushes or rocks

Image Credit: Beery et al. 2020

Camera Trap Wildlife Recognition Challenges





Camouflage: decreases saliency in animals' natural habitat

Image Credit: Beery et al. 2020

Camera Trap Wildlife Recognition Challenges

Perspective: Animals can be close to the camera, resulting in partial views of the body.

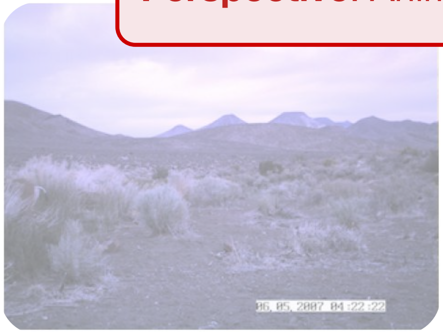


Image Credit: Beery et al. 2020

Camera Trap Wildlife Recognition Challenges

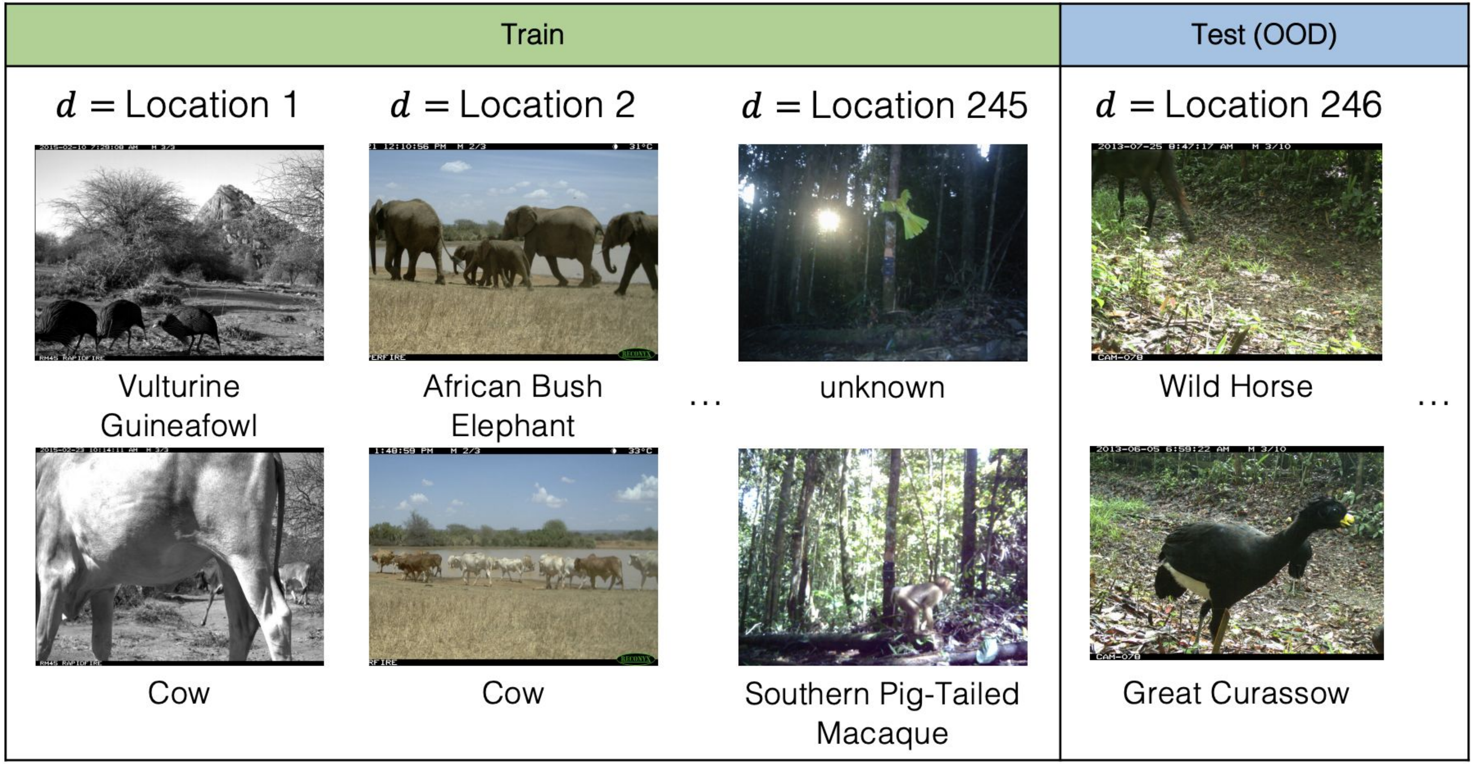
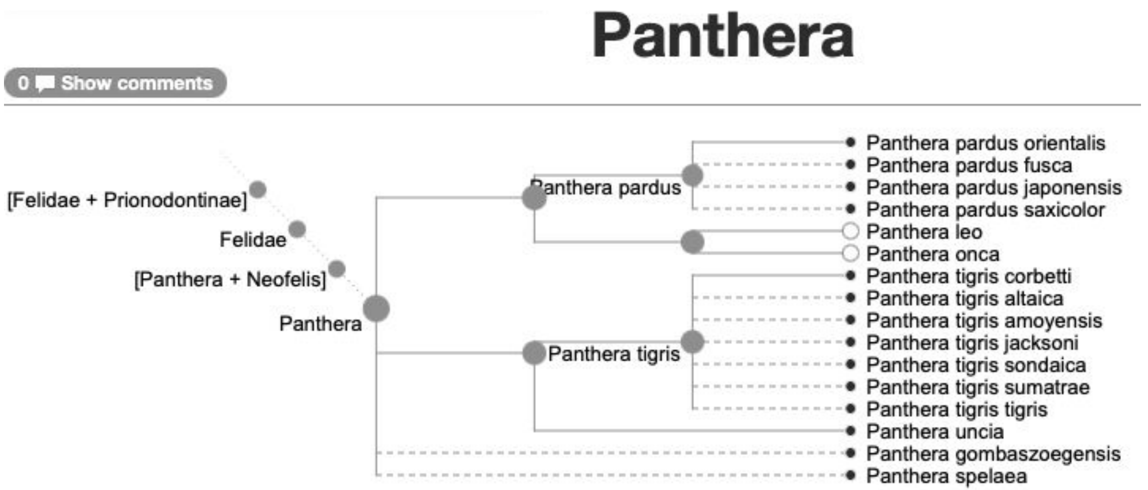


Figure Credit: Koh et al. 2021

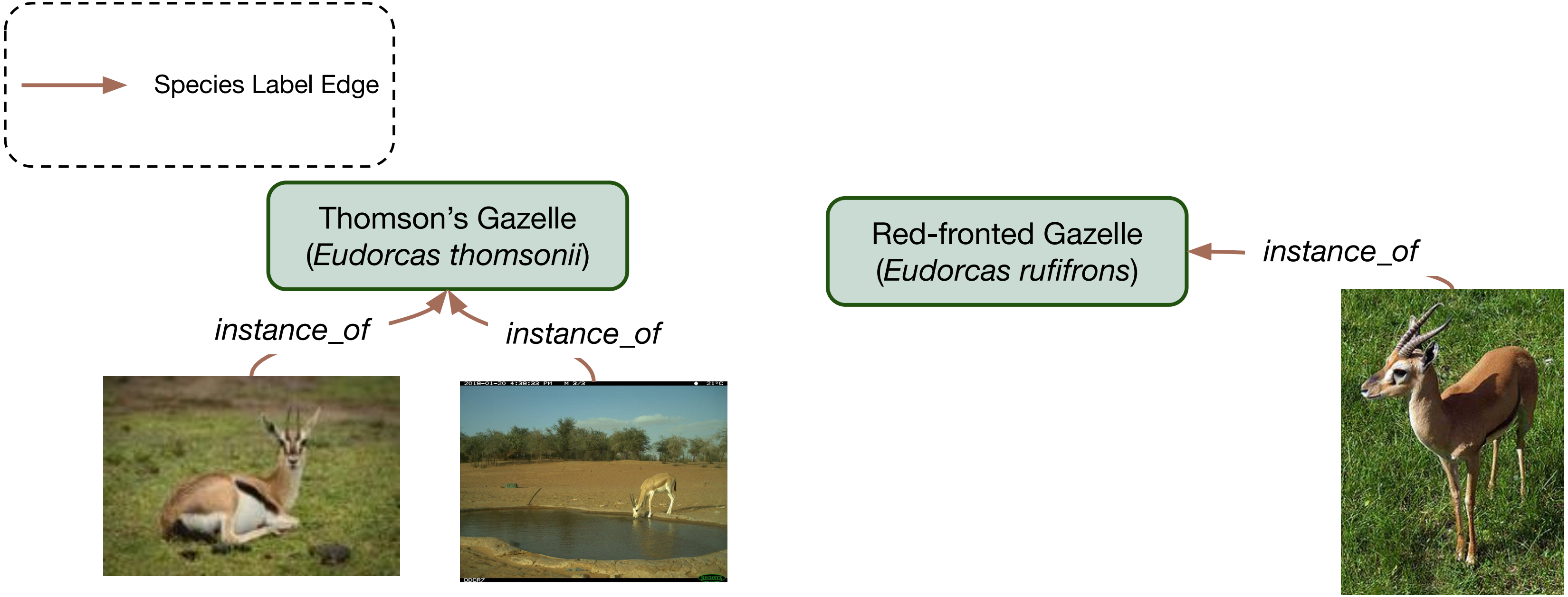
Wildlife Taxonomy

- Phylogenetic taxonomy from Open Tree of Life (OTT)



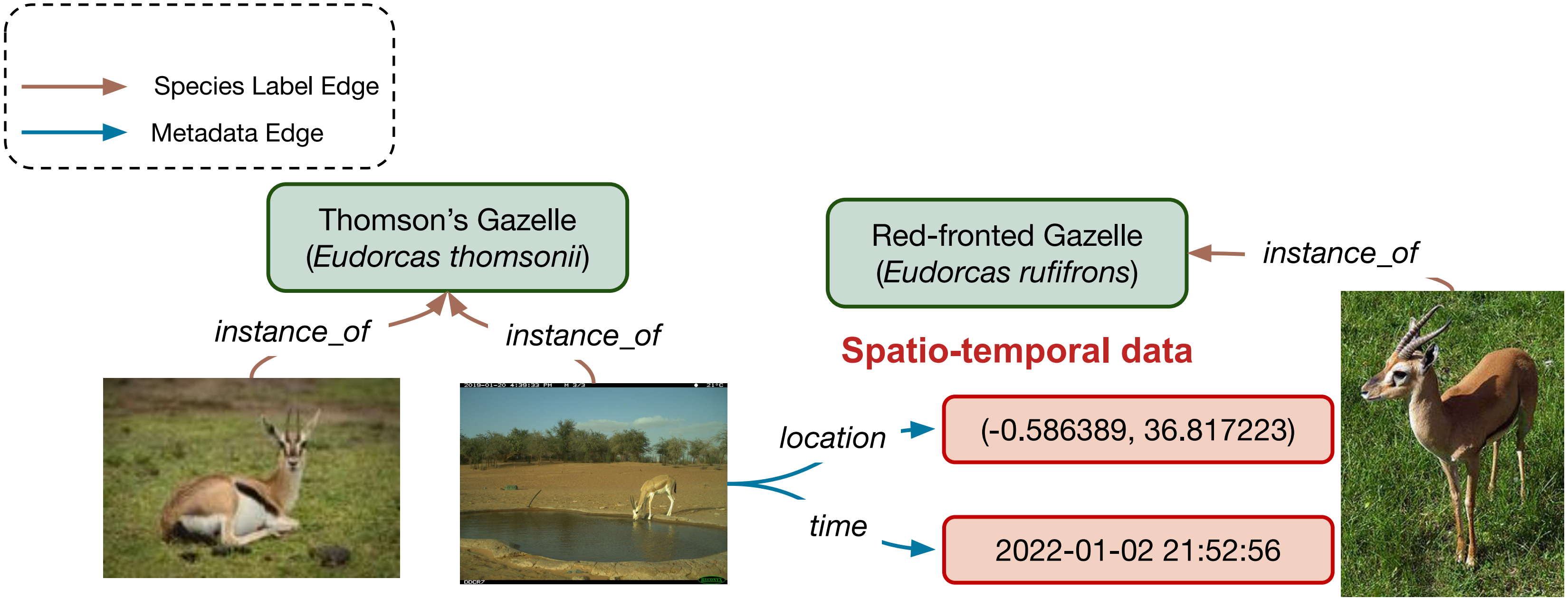
Multimodal KG

- Structured representation for semantic relationships b/w wildlife species, images, attributes and taxonomy.



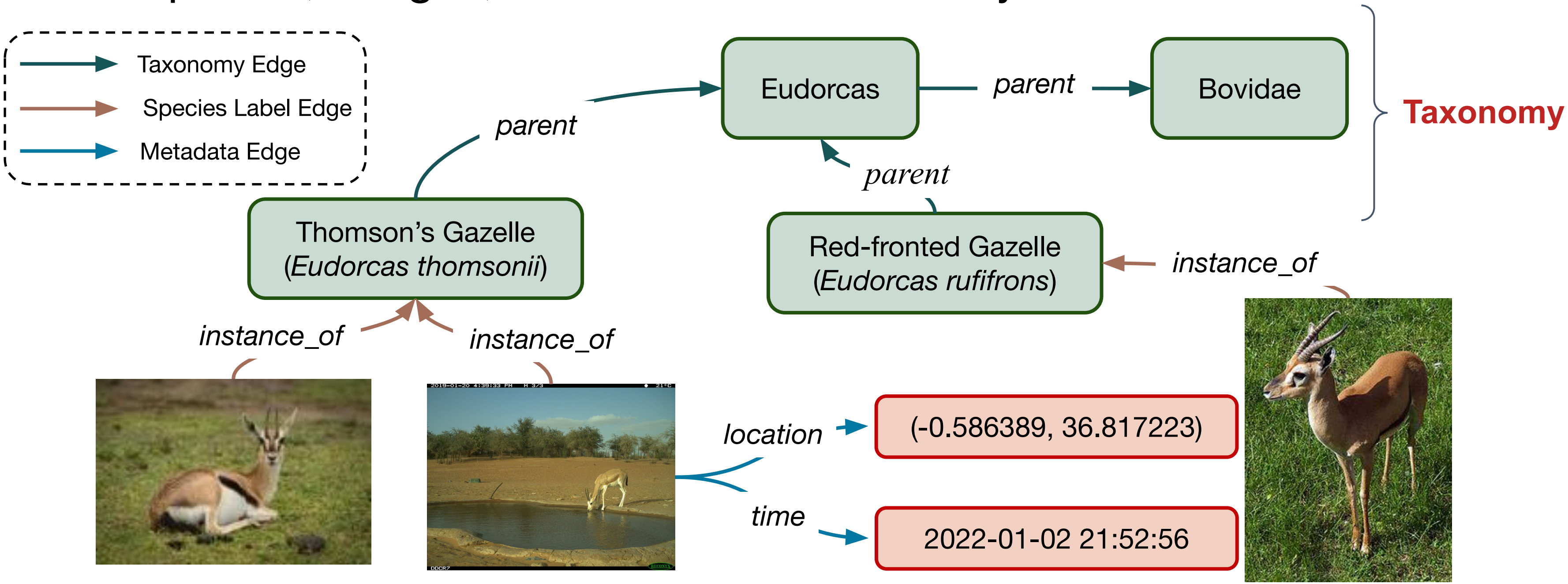
Multimodal KG

- Structured representation for semantic relationships b/w wildlife species, images, attributes and taxonomy.



Multimodal KG

- Structured representation for semantic relationships b/w wildlife species, images, attributes and taxonomy.



Multimodal KG

- Entity types:
    - Categorical (e.g. Species labels, taxons)
        - Encoder: Entity lookup table

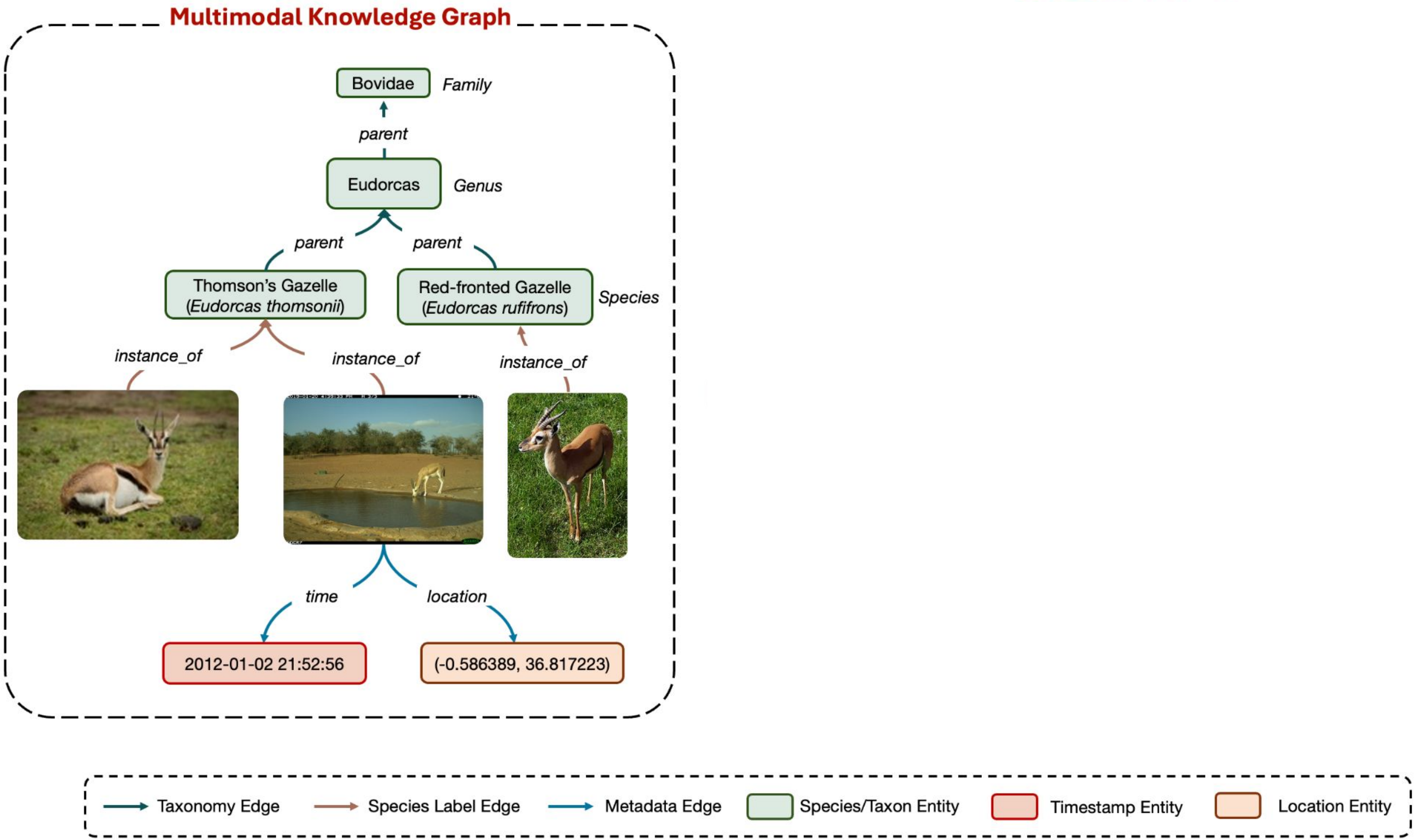
Multimodal KG

- Entity types:
    - Categorical (e.g. Species labels, taxons)
        - Encoder: Entity lookup table
    - Visual
        - Encoder: CNN/Vision Transformer

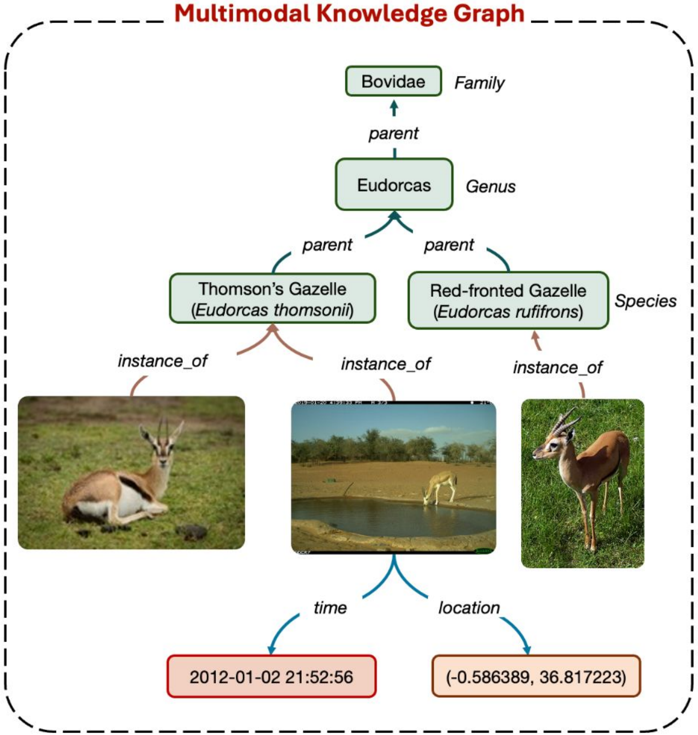
Multimodal KG

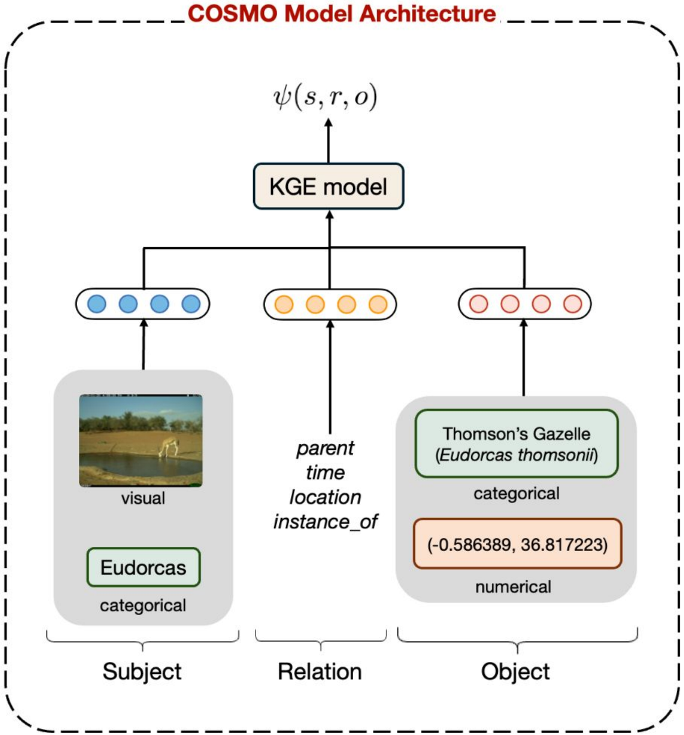
- Entity types:
  - Categorical (e.g. Species labels, taxons)
    - Encoder: Entity lookup table
  - Visual
    - Encoder: CNN/Vision Transformer
  - Numerical (e.g. location, time)
    - Encoder: MLP

Multimodal KG Embedding Model



Multimodal KG Embedding Model





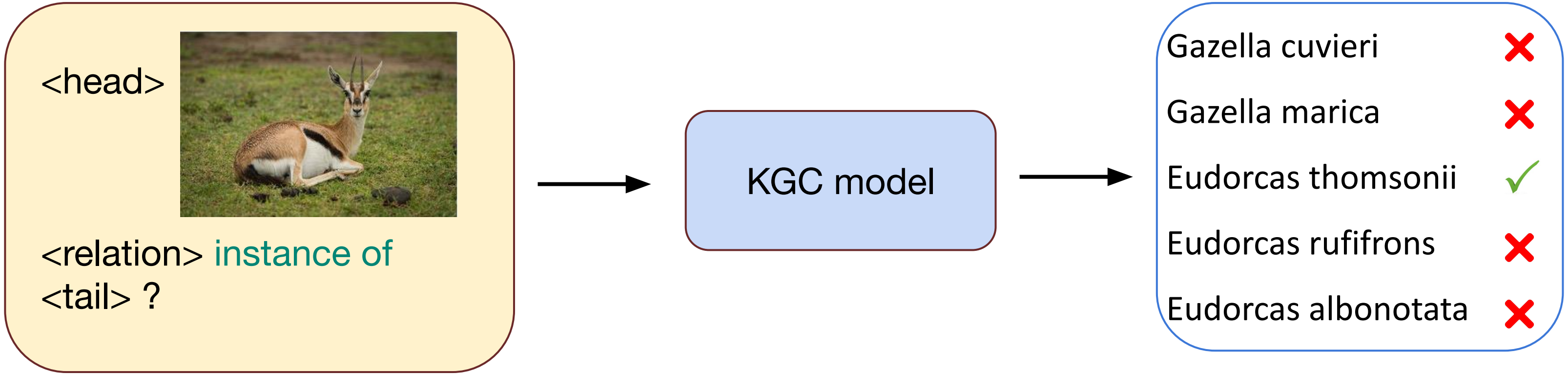
Taxonomy Edge　Species Label Edge　Metadata Edge　Species/Taxon Entity　Timestamp Entity　Location Entity

Task: Species Classification

- Task: Given a camera trap image, predict the species label.

Task: Species Classification

- Task: Given a camera trap image, predict the species label.



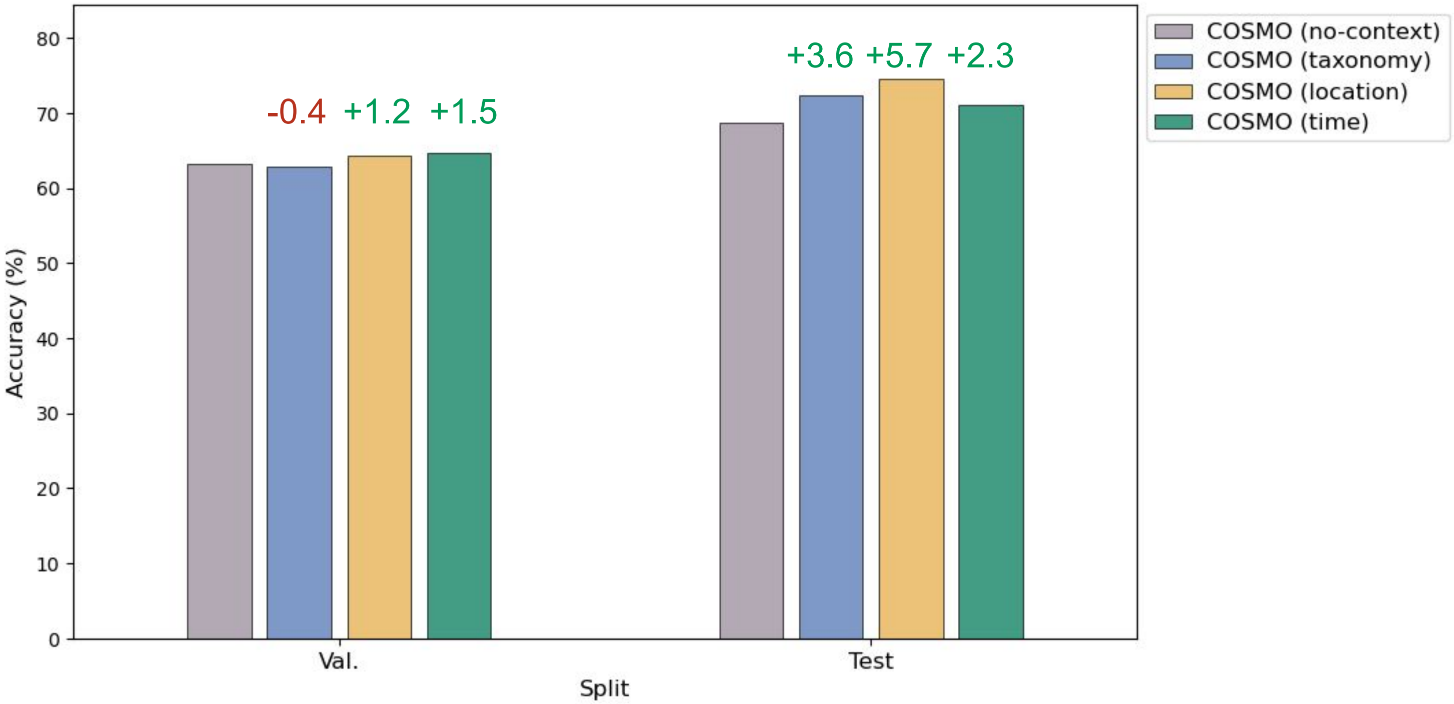
Input: Structured Query

Output: Candidate species with scores

Training

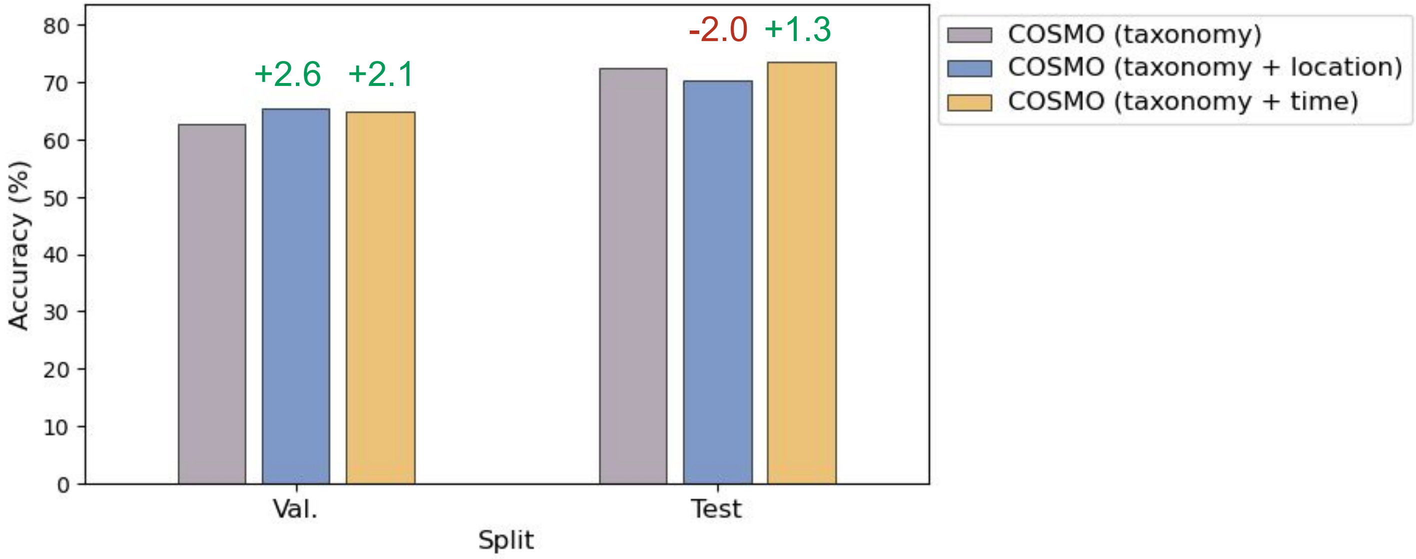
- Categorical attributes: Multi-class classification problem
- Numerical attributes: Multi-class multi-label classification problem.
- Sequentially train on each set of triples types.

Results

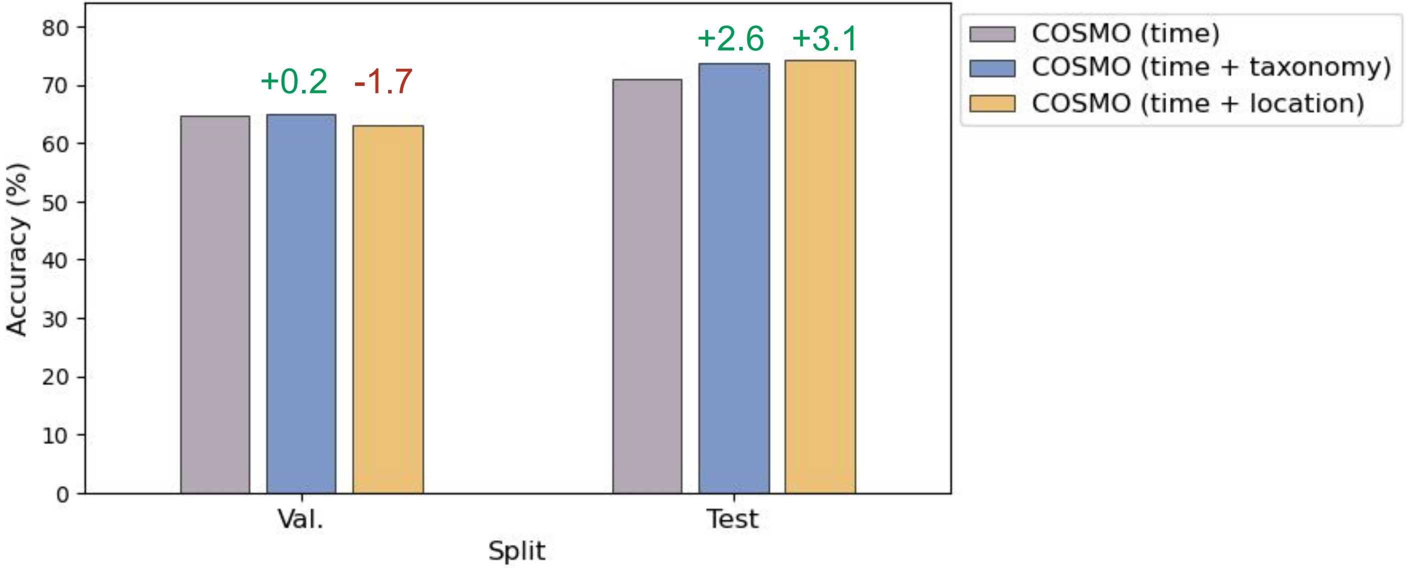


-0.4
+1.2
+1.5
+3.6
+5.7
+2.3
COSMO (no-context)
COSMO (taxonomy)
COSMO (location)
COSMO (time)
Accuracy (%)
0
10
20
30
40
50
60
70
80
Val.
Test
Split

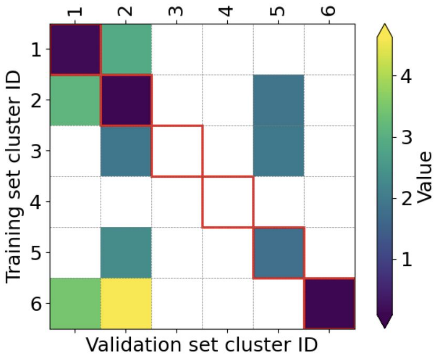
Results (Multiple contexts)



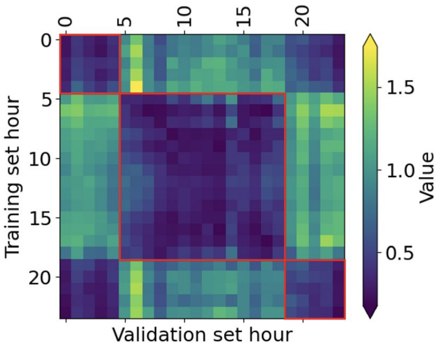
Results (Multiple contexts)



Metadata Analysis



(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).

- Similarity peaks along the diagonal for the location attribute, as well as for the day/night categorization of the time attribute.
- These metadata give a prior for species class distribution

Key Takeaways

- Unified way to leverage various forms of multimodal context for species classification in camera traps.
- Our framework achieves superior out-of-distribution generalization and competitive performance with state-of-the-art.

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

Contact: pahuja.9@osu.edu

Code: https://github.com/OSU-NLP-Group/COSMO

Thank You!

Backup

Training

- For categorical attributes, we formulate it as a multi-class classification problem and use standard cross-entropy loss to train the model.

$$\mathcal{L}(\mathcal{I}, \texttt{instance of}, s) = - \log \frac{\exp(\psi(\mathcal{I}, \texttt{instance of}, s))}{\sum_{s' \in S} \exp(\psi(\mathcal{I}, \texttt{instance of}, s'))}, \quad (2)$$

- 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.

$$\mathcal{L}(\mathcal{I}, \texttt{time}, t) = -\sum_{t'} l_{t'}^{\mathcal{I},time} \cdot \log(\sigma(\psi(\mathcal{I}, \texttt{time}, t')))+ (1 - l_{t'}^{\mathcal{I},time}) \cdot (1 - \log(\sigma(\psi(\mathcal{I}, \texttt{time}, t')))), \quad (3)$$

where $l_{t'}^{\mathcal{I},time}$ denotes the binary label indicating whether the triple $(\mathcal{I}, time, t')$ exists in the set of observed triples and

- Sequentially train on each type of triples.

Baselines

- Empirical Risk Minimization (ERM)
- Domain adaptation baselines:
  - Add a penalty to the ERM objective that encourages some form of invariance across domains.
  - **CORAL** (Sun and Saenko 2016): Encourage feature representations to have the same distribution across domains.
  - **Group DRO** (Hu et al. 2018): An algorithm that uses distributionally robust optimization to perform well on subpopulation shifts.
  - **ABSGD** (Qi et al. 2020): Assign importance weights to samples in batch.
  - **Fish** (Shi et al. 2022): Domain generalization by maximizing the inner product between gradients from different domains.

Results

	Model	Multi-modality: Taxonomy	Multi-modality: Location	Multi-modality: Time	Val. Acc. (%)	Test Acc. (%)
	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	**COSMO**	✓			62.8 (±2.2) (-0.4)	72.4 (±2.5) (+3.6)
			✓		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	**COSMO**	✓	✓		**65.4 (±0.4)** (+2.2)	70.4 (±2.1) (+1.6)
		✓		✓	<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 –.

Backup

- Bhattacharyya distance

$$D_B(P,Q) = -\ln(BC(P,Q))$$

where

$$BC(P,Q) = \sum_{x \in \mathcal{X}} \sqrt{P(x)Q(x)}$$