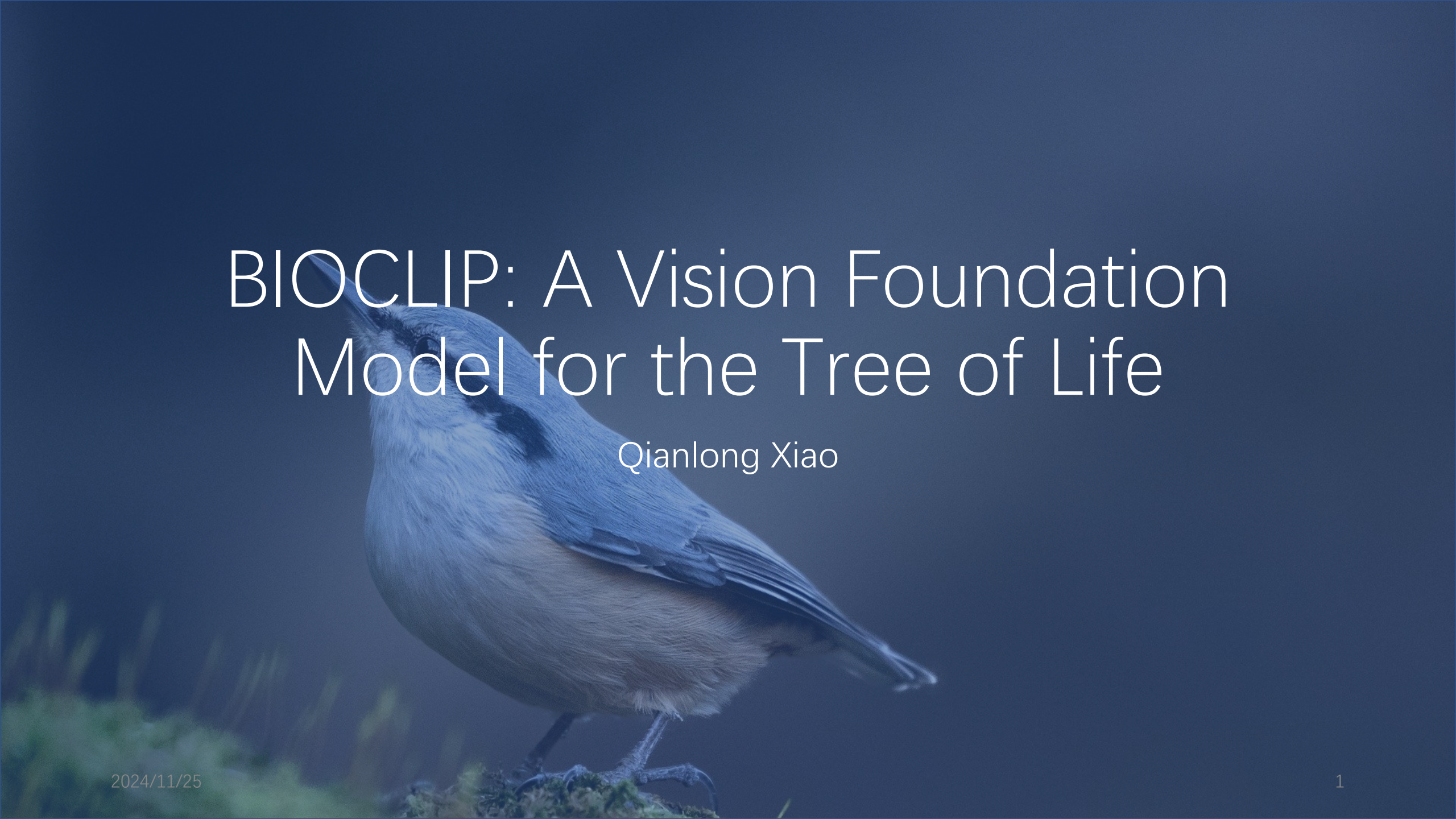


BIOCLIP: A Vision Foundation Model for the Tree of Life

Qianlong Xiao

Agenda

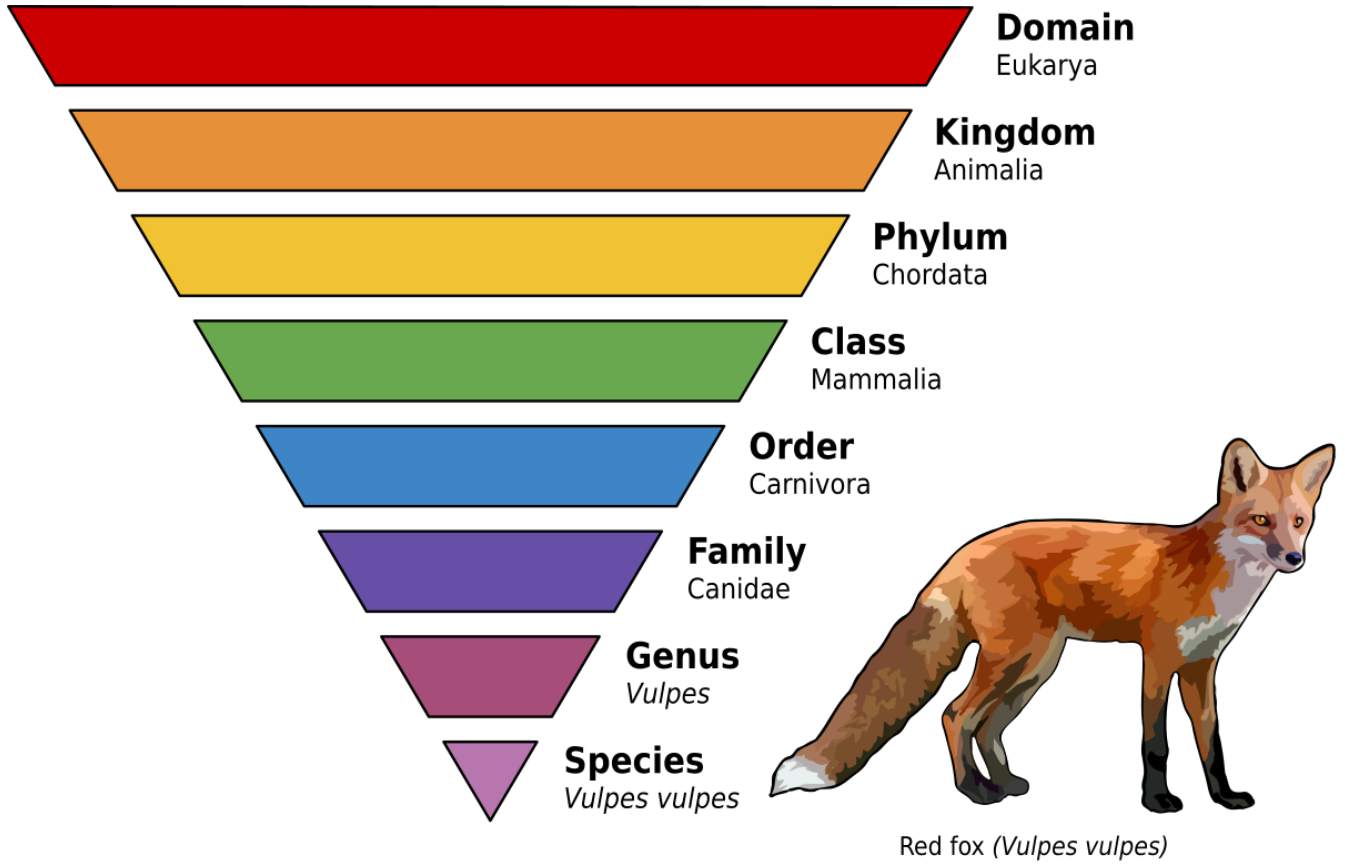
1. Motivation
2. OpenAI CLIP model
3. BioCLIP model
    1. Dataset
    2. Training Process
4. Discussion
5. Conclusion and Future Outlook

Terminology

- standard taxonomic hierarchy

Most General

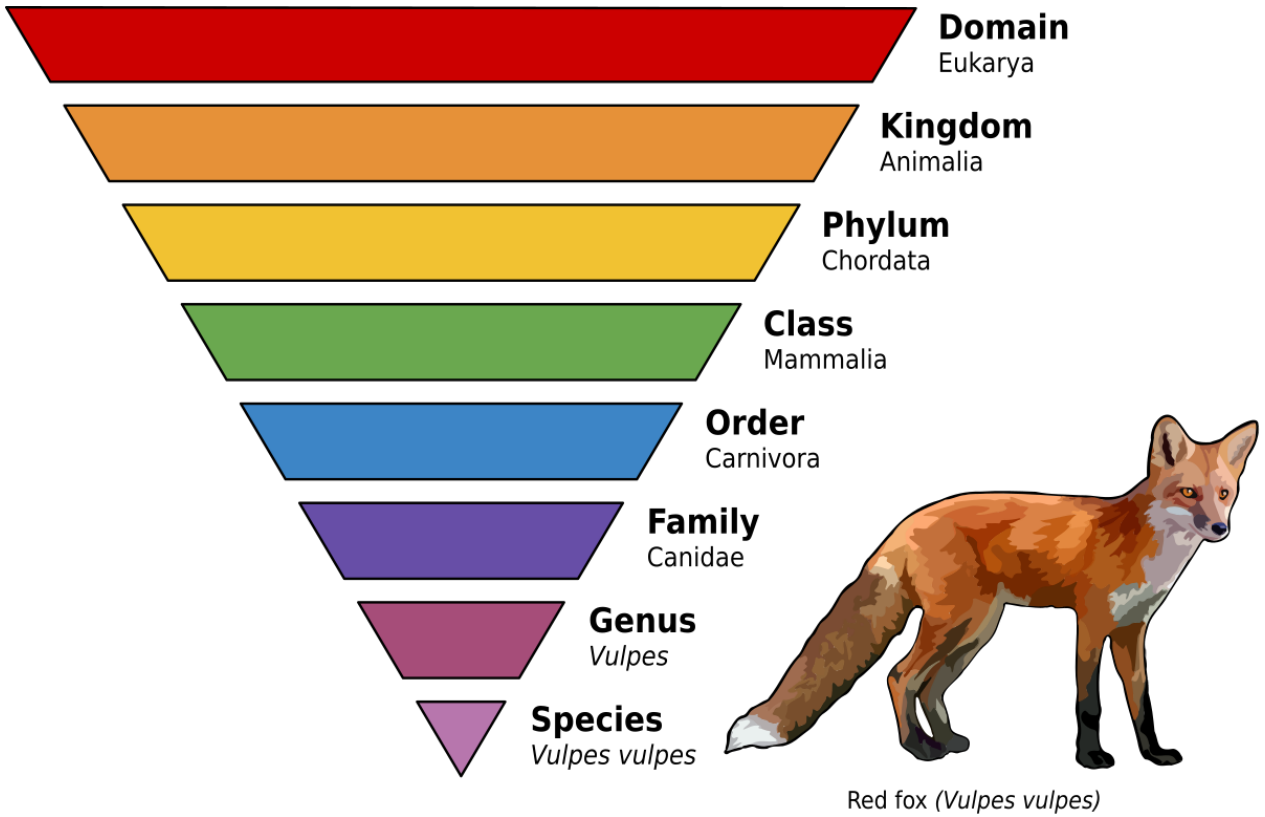
Most Specific



Ref: https://en.wikipedia.org/wiki/Taxonomic_rank

Terminology

- standard taxonomic hierarchy



Taxon: A taxon is a **unit or group** in biological classification that represents a category of organisms **at any rank** in the taxonomic hierarchy.
Taxa: The **plural form** of taxon, referring to **multiple groups or units** in the taxonomic hierarchy.

Ref: https://en.wikipedia.org/wiki/Taxonomic_rank

Motivation

- Firstly, let look up these two pictures:

Onoclea sensibilis

Onoclea hintonii

They are so similar. Could any model distinguish them?

Motivation

- Fine-grained Classification Limitations
  - Existing models struggle to differentiate closely related species, such as Onoclea sensibilis and Onoclea hintonii.
- Limitations of Current Datasets
  - Most biological image datasets are either too small or lack sufficient coverage and fine-grained labels, hindering their applicability.
- Need for Specialized Pretraining Strategies
  - Leveraging the hierarchical structure of biological taxonomy can enhance pretraining strategies, improving performance in zero-shot and few-shot learning scenarios.

BioCLIP overview

- Development of TREEOFLIFE-10M
  - the **largest and most diverse** ML-ready dataset of biology images
- Development of BIOCLIP, a **foundation model** for the tree of life
- Achieved significant improvements (**16%-17% absolute increase**) over baseline models
- BIOCLIP has learned a **hierarchical representation** conforming to the tree of life

OpenAI CLIP model

- **CLIP** (Contrastive Language–Image Pre-training) is a **multimodal** AI model
- Multimodal **Contrastive** Learning
- Multimodal Learning
  - Vision Encoder (ResNet or Vision Transformer) and a Text Encoder (Transformer-based).
- Dataset
  - Trained on 400M image-text pairs from the internet
- Enable Zero-shot Learning

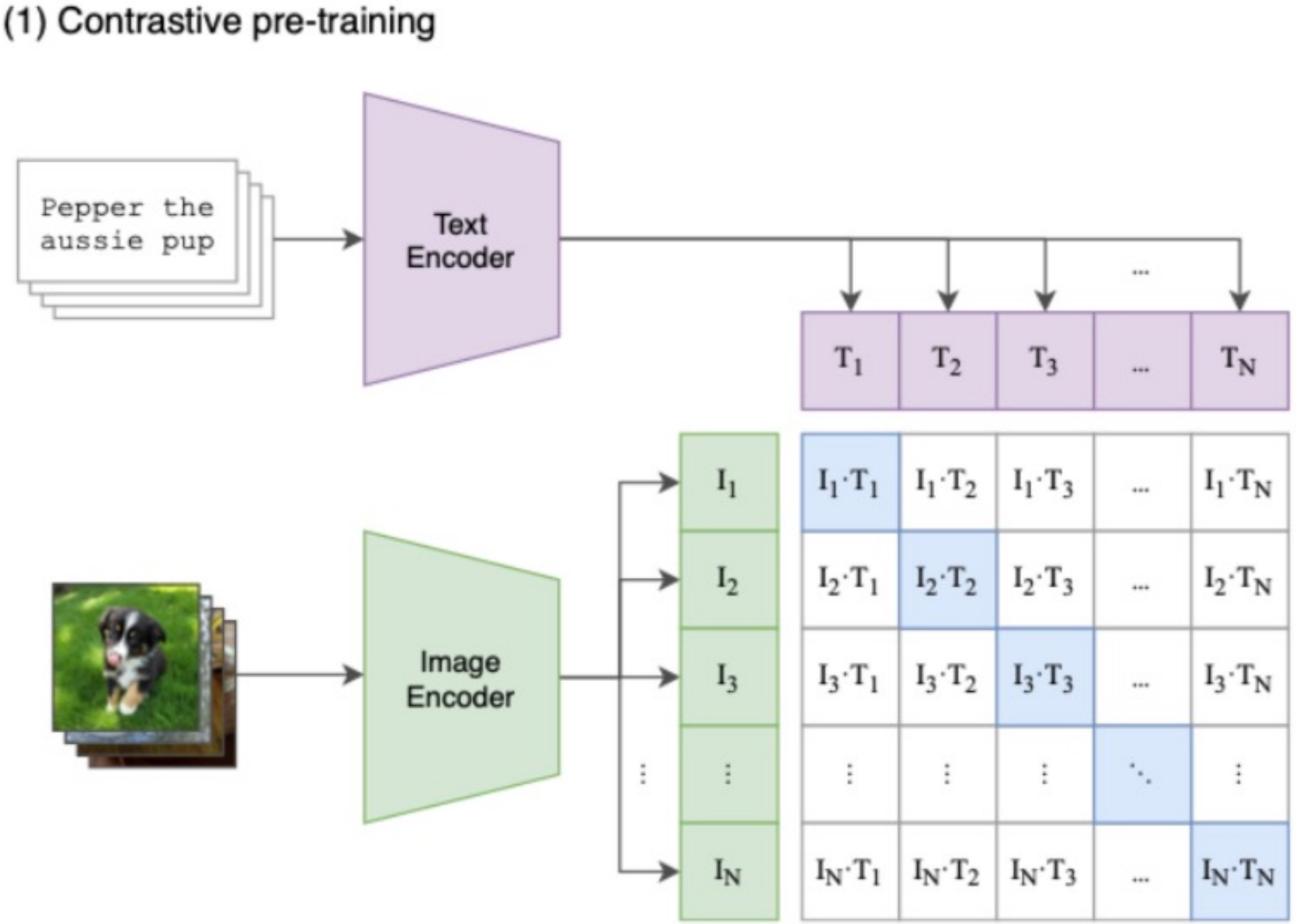
Limitations of Existing Datasets

- MS-COCO (Lin et al., 2014), Visual Genome (Krishna et al., 2017)
  - High-quality crowd-labeled datasets.
  - Small by modern standards (~100,000 training photos each).
- YFCC100M:
  - Large scale (~100 million photos).
  - Metadata is sparse and inconsistent (e.g., auto-generated filenames).
  - Filtering for natural language descriptions reduces the dataset to ~15 million photos, similar to ImageNet.

Constructing the WIT Dataset

- Scale:
  - 400 million (image, text) pairs collected from public sources on the internet.
- Query-Driven Collection:
  - Used 500,000 diverse text queries to find pairs.
  - Capped at 20,000 pairs per query to ensure approximate class balance.
- Broad Visual Concept Coverage:
  - Designed to cover a wide range of visual concepts.
- Comparable Word Count:
  - Similar to the WebText dataset used for GPT-2 training.

How the CLIP model works



Similarity Calculation

$$\text{cosine\_similarity}(u, v) = \frac{u \cdot v}{\|u\|\|v\|}$$

U: Embedding of the image from the image encoder.
V: Embedding of the text from the text encoder.

	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

Loss Function

- InfoNCE (Noise Contrastive Estimation) loss
- **maximizes** the similarity of matched pairs while
- **minimizing** the similarity of mismatched pairs

$$L_{\text{image-to-text}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(S_{ii}/\tau)}{\sum_{j=1}^{N}\exp(S_{ij}/\tau)}$$

$$L_{\text{text-to-image}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(S_{ii}/\tau)}{\sum_{j=1}^{N}\exp(S_{ji}/\tau)}$$

$$L = \frac{1}{2}(L_{\text{image-to-text}} + L_{\text{text-to-image}})$$

Image Encoder

- ResNet50(He et al., 2016a)
- Optimizations
  - ResNet-D improvements
  - rect-2 blur pooling
  - attention pooling
- RN50x4 (4x wider), RN50x1 (16x wider), and RN50x64 (wider).
- for initial and medium-scale experiments.

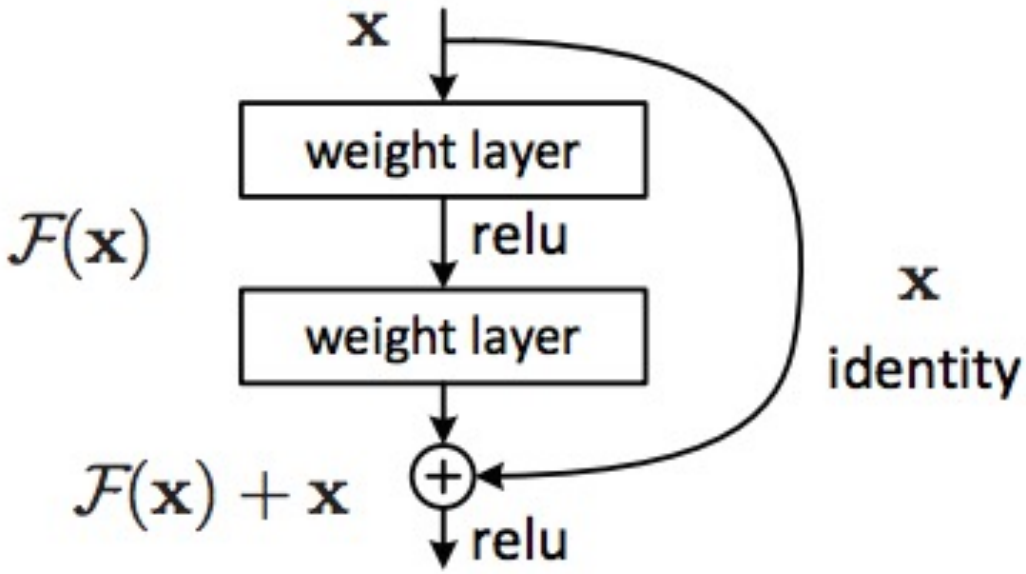
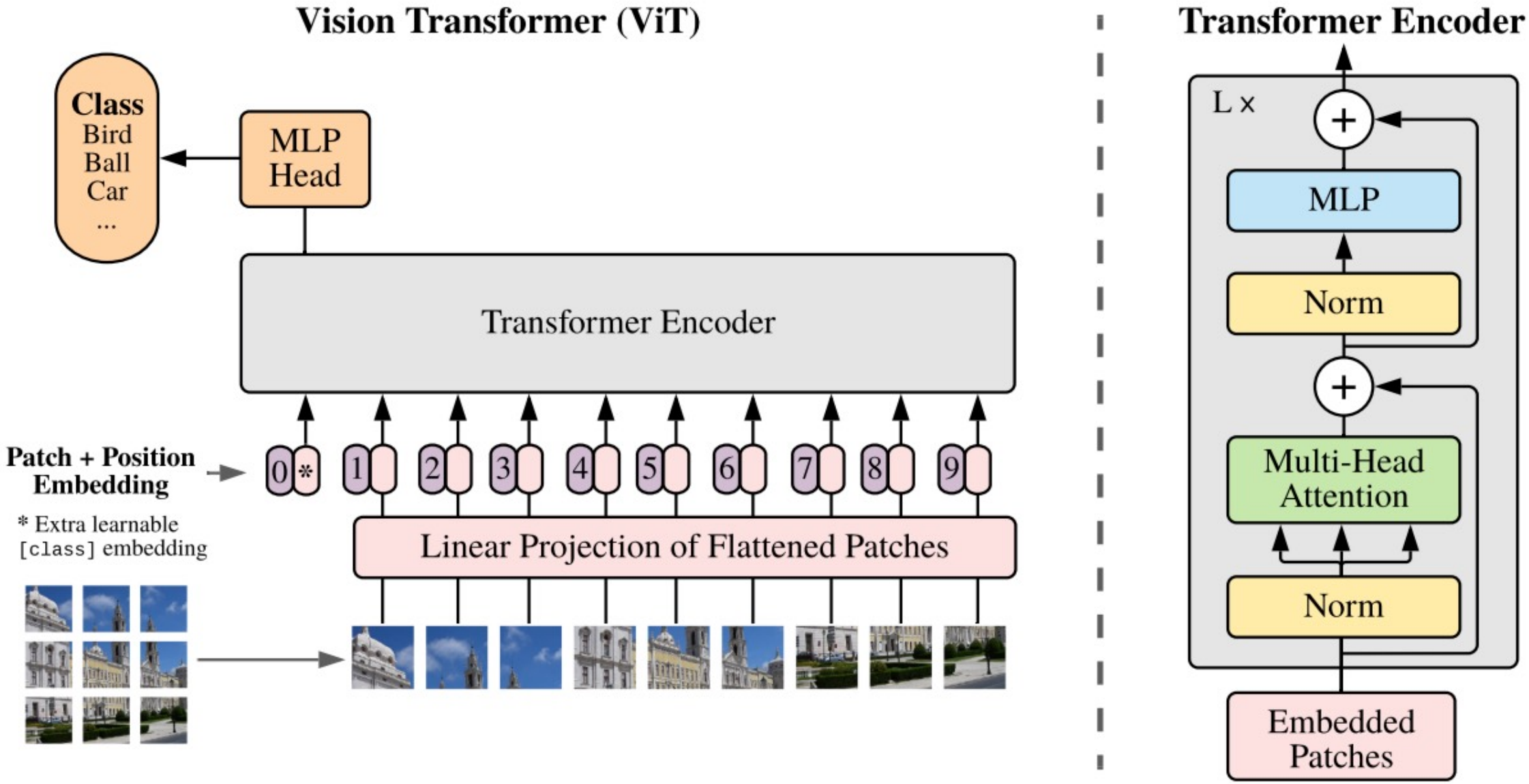


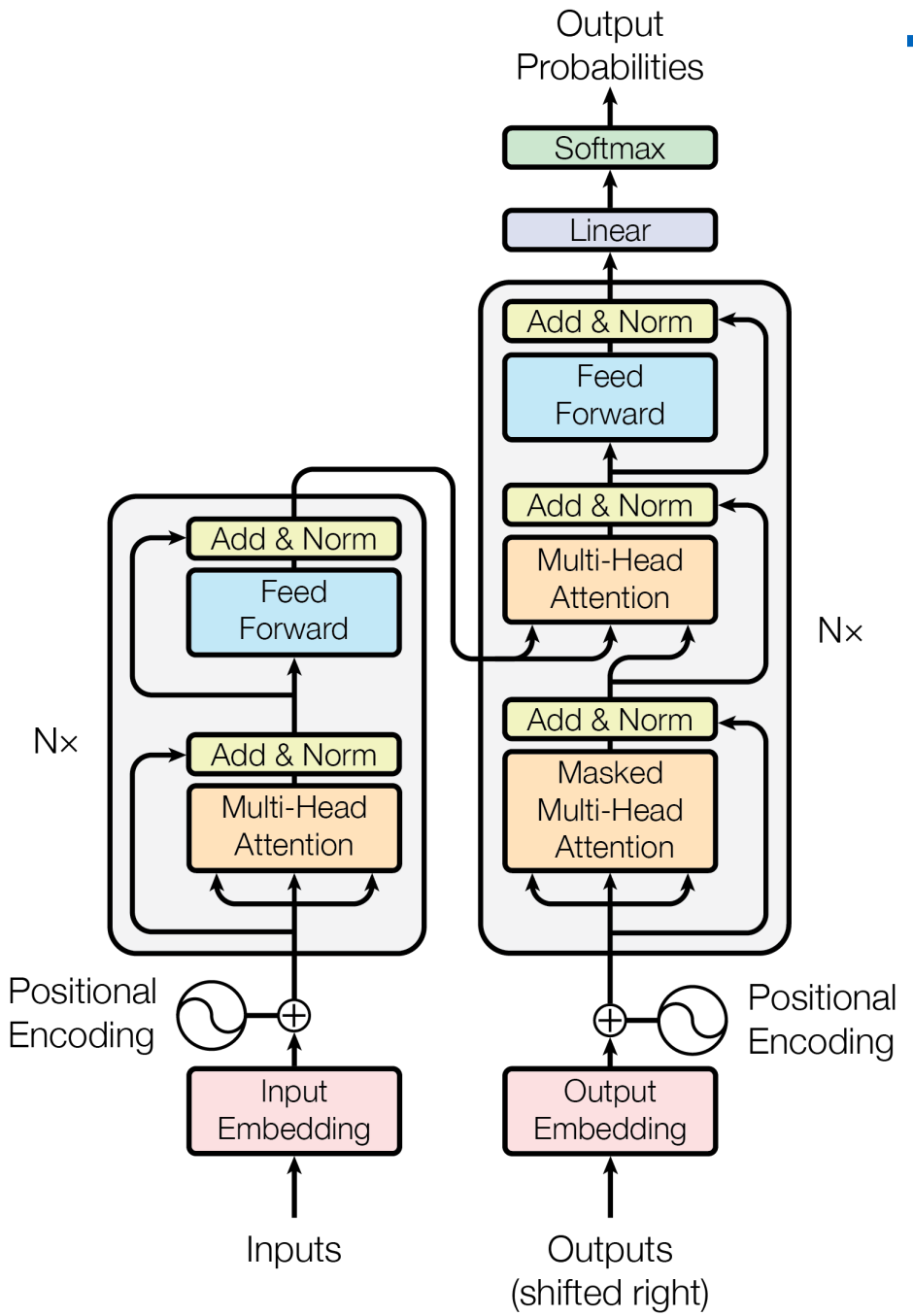
Image Encoder

- Vision Transformer (ViT) (Dosovitskiy et al., 2020)
- Large-scale Model Stage
- Variants Used:
  - ViT-B/32
  - ViT-B/16
  - ViT-L/14, ViT-L/14@336px
- Global Context Modeling
- High-Resolution Efficiency


Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0 * 1 2 3 4 5 6 7 8 9
Linear Projection of Flattened Patches
Transformer Encoder
L x
MLP
Norm
Multi-Head Attention
Norm
Embedded Patches

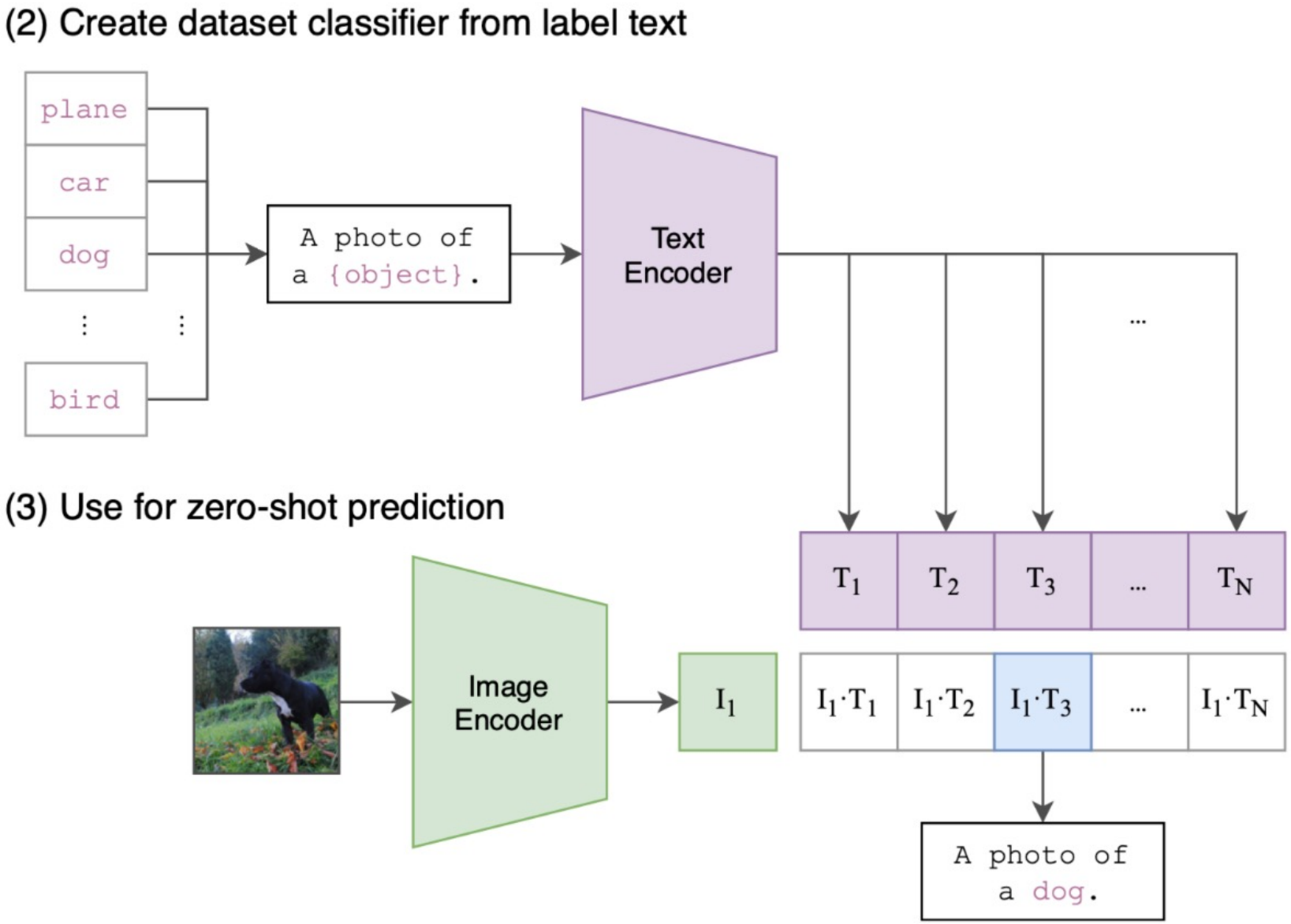
Text Encoder

- Transformer (Vaswani et al., 2017)
- Tokenized text is processed into embeddings using self-attention layers.
- Outputs are mapped to a shared multimodal embedding space.



CLIP:

Zero-shot



Zero- shot

- Embedding Text Labels
- Embedding Input Images
- Matching via Cosine Similarity

- Dynamic Classifier Creation
- Generalization to Unseen Data

BIOCLIP Overview

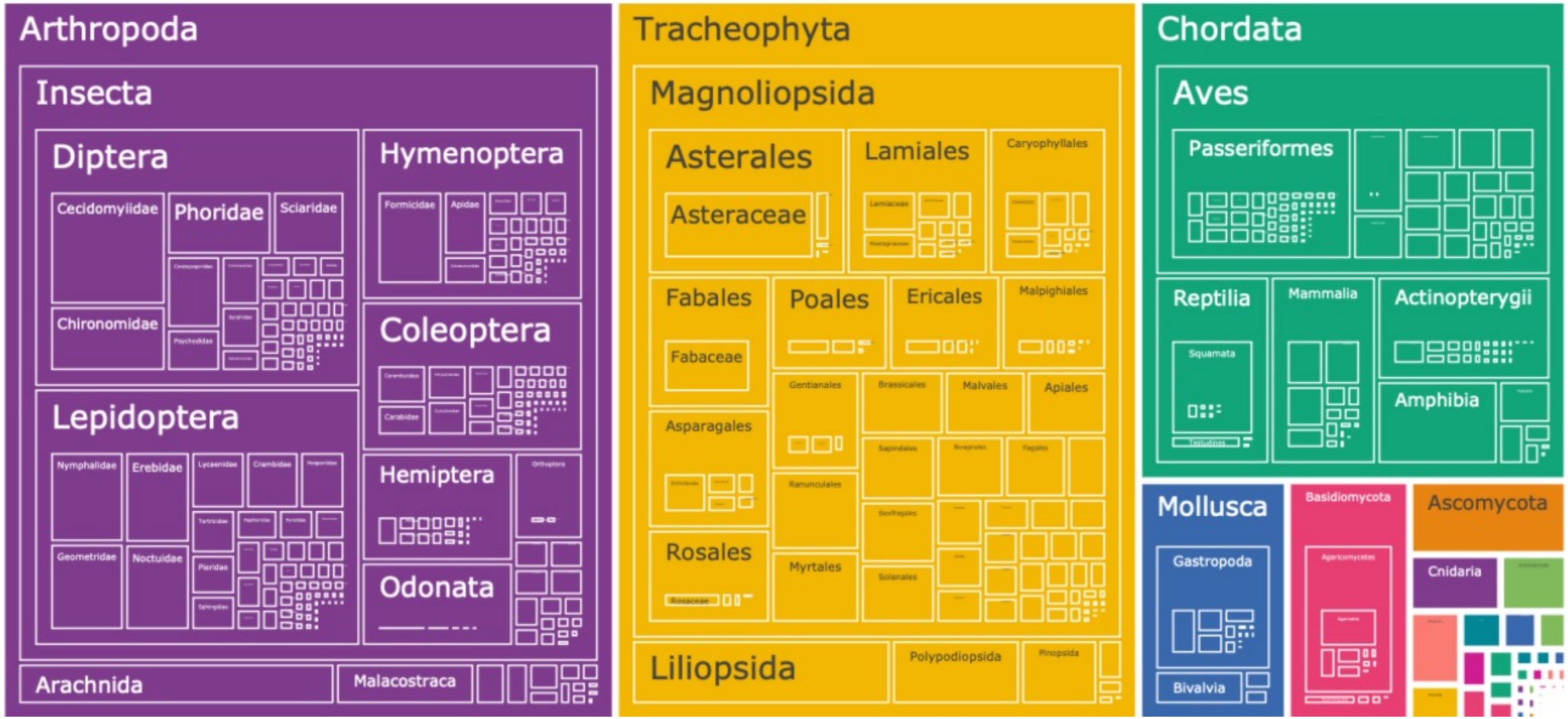
- initialized from OpenAI’ s public CLIP checkpoint
- continually pre-trained on TREEOFLIFE-10M
- multimodal contrastive learning objective

BIOCLIP Dataset

Dataset	Description	Images	Unique Classes
iNat21	Citizen scientist labeled image dataset from iNaturalist for fine-grained classification.	2.7M	10,000
BIOSCAN-1M	Expert labeled image dataset of insects for classification.	1.1M	7,831
EOL	A new dataset with citizen scientist images sourced from Encyclopedia of Life and taxonomic labels standardized by us.	6.6M	448,910
TREEOFLIFE-10M	**Largest-to-date ML-ready dataset of biology images with taxonomic labels.**	**10.4M**	**454,103**

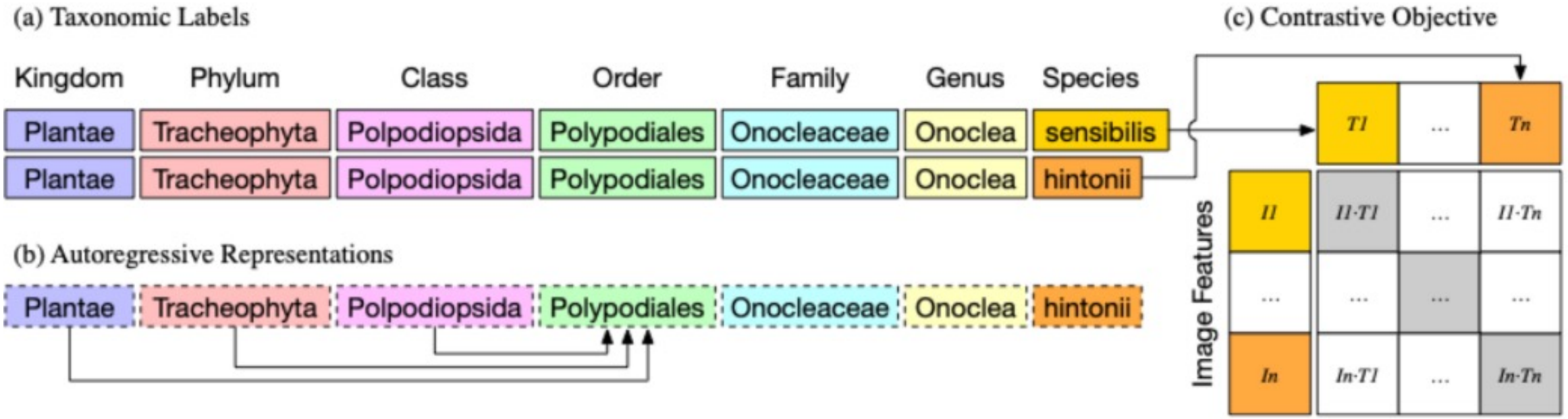
most diverse large-scale public ML-ready dataset for computer vision models in biology

Dataset Overview



Phyla
Classes
Orders
families

Architecture



maximize feature similarity between **positive (image, text) pairs** and **minimize** feature similarity between **negative (image, text) pairs**

Model Initialization

- Image Encoder: a ViT-B/16 vision transformer with OpenAI’ s CLIP weights
- Text Encoder: 77-token causal autoregressive transformer
- 100 epochs on TREEOFLIFE-10M
- Learning Rate: Cosine learning rate schedule.

Feature	**Baseline Model**	**BIOCLIP Model**
Hardware	4 NVIDIA A100 GPUs (1 Node)	8 NVIDIA A100 GPUs (2 Nodes)
Global Batch Size	16,384	32,768
Dataset	iNat21	TREEOFLIFE-10M

Model hyperparameter

Hyperparameter	Value
Architecture	ViT-B/16
Max learning rate	1×10^{-4}
Warm-up steps	1,000
Weight Decay	0.2
Input Res.	224×224

Table D1. Common hyperparameters among all models we train.

Model hyperparameter 2

Dataset	Text Type	Batch Size	Epoch
TREEOFLIFE-10M	Mixture	32K	100
iNat21 Only	Mixture	16K	65
TREEOFLIFE-1M	Common	16K	86
	Scientific		87
	Taxonomy		87
	Sci+Com		87
	Tax+Com		86
	Mixture		91

Table D2. Hyperparameters that differ between the various models we train. We use a smaller batch size and only 1M examples for our text-type ablation because of limited compute.

Datasets for evaluation

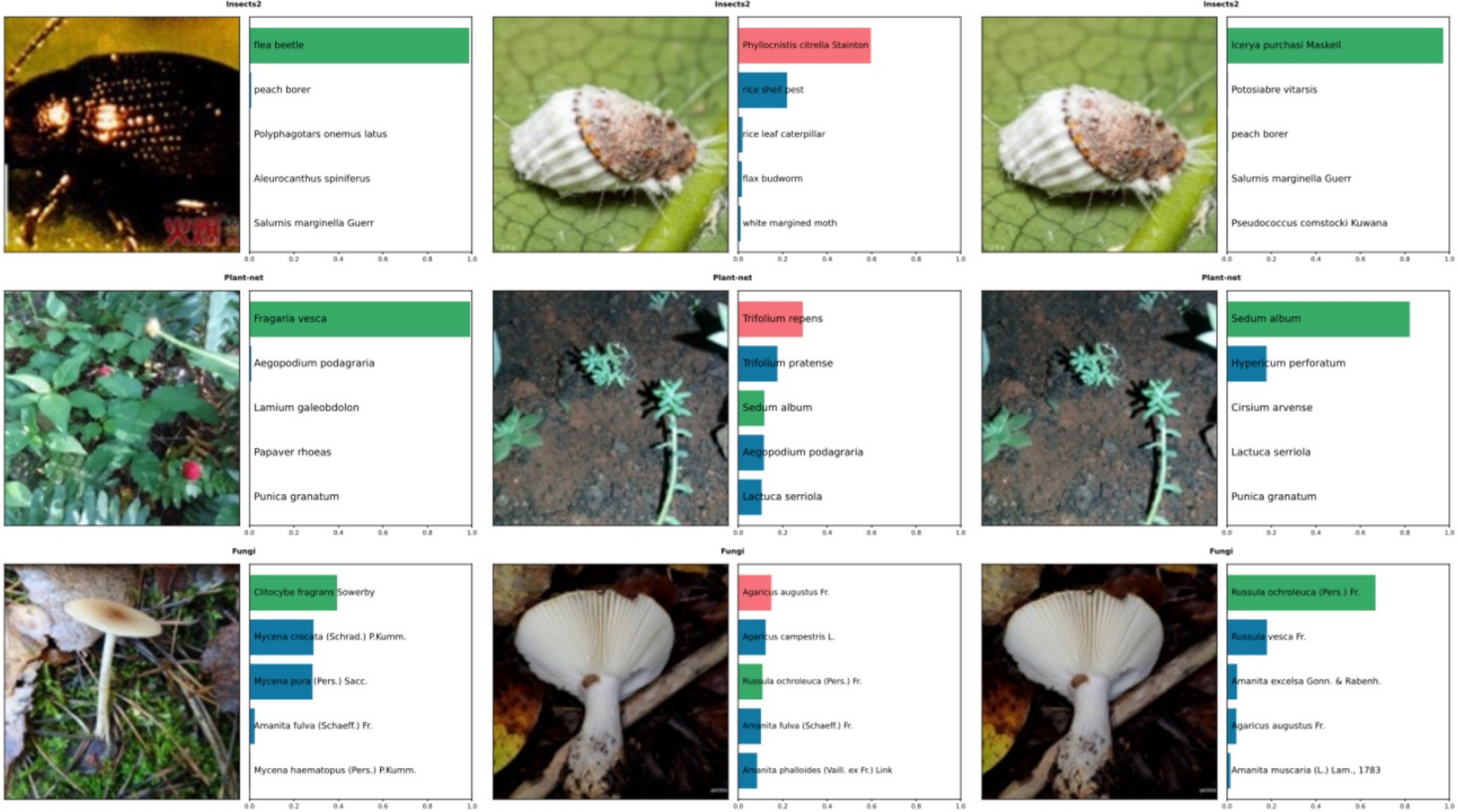
	Name	Description	Examples	Classes	Labels
Animals	Birds 525	Scraped dataset of bird images from web search. [68]	89,885	525	Taxonomic
	Plankton	Expert-labeled in situ images of plankton [35].	4,080	102	Mixed
	Insects	Expert and volunteer-labeled in-the-wild citizen science images of insects [74].	4,680	117	Scientific
	Insects 2	Mixed common and scientific name classification for insect pests [91].	4,080	102	Mixed
Plants & Fungi	PlantNet	Citizen science species-labeled plant images, some drawings [27].	1,000	25	Scientific
	Fungi	Expert-labeled images of Danish fungi [66].	1,000	25	Scientific
	PlantVillage	Museum-style leaf specimens labeled with common names [25].	1,520	38	Common
	Medicinal Leaf	Species classification of leaves from mature, healthy medicinal plants [71].	1,040	26	Scientific
	PlantDoc	17 diseases for 13 plant species [76].	1,080	27	Common
	RARE SPECIES	Subset of species in the IUCN Red List categories: Near Threatened through Extinct in the Wild (iucnredlist.org).	12,000	400	Taxonomic

Can BIOCLIP Generalize to Unseen Taxa?

- BIOCLIP substantially outperforms both baseline CLIP models
- Especially on **unseen taxa(Rare Species)** and **zero-shot classification**

Model	Animals				Plants & Fungi							
	Birds 525	Plankton	Insects	Insects 2	PlantNet	Fungi	PlantVillage	Med. Leaf	PlantDoc	Rare Species	Mean (Δ)	
Random Guessing	0.2	1.2	1.0	1.0	4.0	4.0	2.6	4.0	3.7	0.3	2.2	
Zero-Shot Classification												
CLIP	49.9	3.2	9.1	9.8	58.5	10.2	5.4	15.9	26.1	31.8	21.9	–
OpenCLIP	54.7	2.2	6.5	9.6	50.2	5.7	8.0	12.4	25.8	29.8	20.4	−1.5
BIOCLIP	**72.1**	**6.1**	**34.8**	**20.4**	**91.4**	40.7	**24.4**	**38.6**	**28.4**	**38.0**	**39.4**	+17.5
– iNat21 Only	56.1	2.6	30.7	11.5	88.2	**43.0**	18.4	25.6	20.5	21.3	31.7	+9.8
One-Shot Classification												
CLIP	43.7	25.1	21.6	13.7	42.1	17.2	49.7	70.1	24.8	28.5	33.6	–
OpenCLIP	53.7	32.3	23.2	14.3	45.1	18.4	53.6	71.2	26.8	29.2	36.7	+3.1
Supervised-IN21K	60.2	22.9	14.7	14.4	46.7	16.9	**62.3**	58.6	27.7	28.0	35.2	+1.6
DINO	40.5	**37.0**	23.5	16.4	30.7	20.0	60.0	79.2	23.7	31.0	36.2	+2.6
BIOCLIP	71.8	30.6	**57.4**	**20.4**	64.5	**40.3**	58.8	**84.3**	**30.7**	**44.9**	**50.3**	+16.7
– iNat21 Only	**74.8**	29.6	53.9	19.7	**67.4**	35.5	55.2	75.1	27.8	36.9	47.5	+13.9
Five-Shot Classification												
CLIP	73.5	41.2	39.9	24.6	65.2	27.9	71.8	89.7	35.2	46.0	51.5	–
OpenCLIP	81.9	52.5	42.6	25.0	68.0	30.6	77.8	91.3	42.0	47.4	55.9	+4.4
Supervised-IN21K	83.9	39.2	32.0	25.4	70.9	30.9	**82.4**	82.3	44.7	47.3	53.9	+2.4
DINO	70.8	**56.9**	46.3	28.6	50.3	34.1	82.1	94.9	40.3	50.1	55.4	+3.9
BIOCLIP	90.0	49.3	**77.8**	**33.6**	**85.6**	**62.3**	80.9	**95.9**	**47.5**	**65.7**	**68.8**	+17.3
– iNat21 Only	**90.1**	48.2	73.7	32.1	84.7	55.6	77.2	93.5	41.0	55.6	65.1	+13.6

Zero-, one- and five-shot classification top-1 accuracy for different models



Correct BIOCLIP predictions; Images that CLIP incorrectly labels but BIOCLIP correctly labels

zero-shot predictions results

Text types

Text Type	Example
Common	black-billed magpie
Scientific	*Pica hudsonia*
Taxonomic	*Animalia Chordata Aves Passeriformes Corvidae Pica hudsonia*
Scientific + Common	*Pica hudsonia* with common name black-billed magpie
Taxonomic + Common	*Animalia Chordata Aves Passeriformes Corvidae Pica hudsonia* with common name black-billed magpie

- mixed text type training strategy
  - retains the **generalization benefits**
  - providing more **flexibility**

Discussion: How Text Types Affect Generalization ?

- **Taxonomic + Common** Names Boost Generalization
- **Mixed Text Types** Enhance Flexibility
- Data Diversity Matters(10M>2M)

Dataset	Train↓Test→	Com	Sci	Tax	Sci+Com	Tax+Com
ToL-1M	Com	24.9	9.5	10.8	22.3	21.0
	Sci	11.0	22.3	4.5	21.5	8.0
	Tax	11.8	10.1	26.6	16.0	24.8
	Sci+Com	24.5	12.9	12.6	28.0	24.9
	Tax+Com	20.5	8.0	19.7	24.0	30.4
	Mixture	**26.1**	**24.9**	**26.7**	**29.5**	**30.9**
iNat21-2.7M	Mixture	20.4	14.7	15.6	20.9	21.3
ToL-10M	Mixture	31.6	30.1	34.1	37.0	38.0

Zero-shot accuracy on species not seen during training (RARE SPECIES task)

Discussion: Is the CLIP Objective Necessary?

Comparison:

- Standard Classification: CEL
- Hierarchical Classification: summing CEL across all levels.

Objective	Mean 1-Shot	Mean 5-shot
Cross-entropy	16.5	26.2
Hier. cross-entropy	19.3	30.5
CLIP	**44.7**	**63.8**

Table 6: One- and five-shot classification top-1 accuracy for different pre-training objectives on TREEOFLIFE-1M(ViT-B/16 models)

Results:

- Hierarchical classification outperforms simple classification
- The CLIP massively outperforms both baselines.

Discussion: Can BIOCLIP Classify More Than Species?

- **Plant Disease Diagnosis**
- **BIOCLIP outperforms baselines** in classification
- a **9.1% higher mean accuracy** compared to zero-shot

	Animals				Plants & Fungi							
Model	Birds 525	Plankton	Insects	Insects 2	PlantNet	Fungi	PlantVillage	Med. Leaf	PlantDoc	Rare Species	Mean (Δ)	
Random Guessing	0.2	1.2	1.0	1.0	4.0	4.0	2.6	4.0	3.7	0.3	2.2	
Zero-Shot Classification												
CLIP	49.9	3.2	9.1	9.8	58.5	10.2	5.4	15.9	26.1	31.8	21.9	–
OpenCLIP	54.7	2.2	6.5	9.6	50.2	5.7	8.0	12.4	25.8	29.8	20.4	−1.5
BioCLIP	**72.1**	**6.1**	**34.8**	**20.4**	**91.4**	40.7	**24.4**	**38.6**	**28.4**	**38.0**	**39.4**	+17.5
– iNat21 Only	56.1	2.6	30.7	11.5	88.2	**43.0**	18.4	25.6	20.5	21.3	31.7	+9.8
One-Shot Classification												
CLIP	43.7	25.1	21.6	13.7	42.1	17.2	49.7	70.1	24.8	28.5	33.6	–
OpenCLIP	53.7	32.3	23.2	14.3	45.1	18.4	53.6	71.2	26.8	29.2	36.7	+3.1
Supervised-IN21K	60.2	22.9	14.7	14.4	46.7	16.9	**62.3**	58.6	27.7	28.0	35.2	+1.6
DINO	40.5	**37.0**	23.5	16.4	30.7	20.0	60.0	79.2	23.7	31.0	36.2	+2.6
BioCLIP	71.8	30.6	**57.4**	**20.4**	64.5	**40.3**	58.8	**84.3**	**30.7**	**44.9**	**50.3**	+16.7
– iNat21 Only	**74.8**	29.6	53.9	19.7	**67.4**	35.5	55.2	75.1	27.8	36.9	47.5	+13.9
Five-Shot Classification												
CLIP	73.5	41.2	39.9	24.6	65.2	27.9	71.8	89.7	35.2	46.0	51.5	–
OpenCLIP	81.9	52.5	42.6	25.0	68.0	30.6	77.8	91.3	42.0	47.4	55.9	+4.4
Supervised-IN21K	83.9	39.2	32.0	25.4	70.9	30.9	**82.4**	82.3	44.7	47.3	53.9	+2.4
DINO	70.8	**56.9**	46.3	28.6	50.3	34.1	82.1	94.9	40.3	50.1	55.4	+3.9
BioCLIP	90.0	49.3	**77.8**	**33.6**	**85.6**	**62.3**	80.9	**95.9**	**47.5**	**65.7**	**68.8**	+17.3
– iNat21 Only	**90.1**	48.2	73.7	32.1	84.7	55.6	77.2	93.5	41.0	55.6	65.1	+13.6

Discussion: Does BIOCLIP Learn the Hierarchy?

- Intrinsic Evaluation
- Method: t-SNE Visualization:
- At higher ranks
  - both CLIP and BIOCLIP show good separation.
- At lower ranks
  - BIOCLIP demonstrates richer clustering and produces more separable features.
  - CLIP’ s features are cluttered and lack clear structure.
- **Fine-grained** representation in BIOCLIP enables **superior performance**, especially for challenging tasks with **limited data**.

BIOCLIP:

OPENAI CLIP:

BIOCLIP:

OPENAI CLIP:

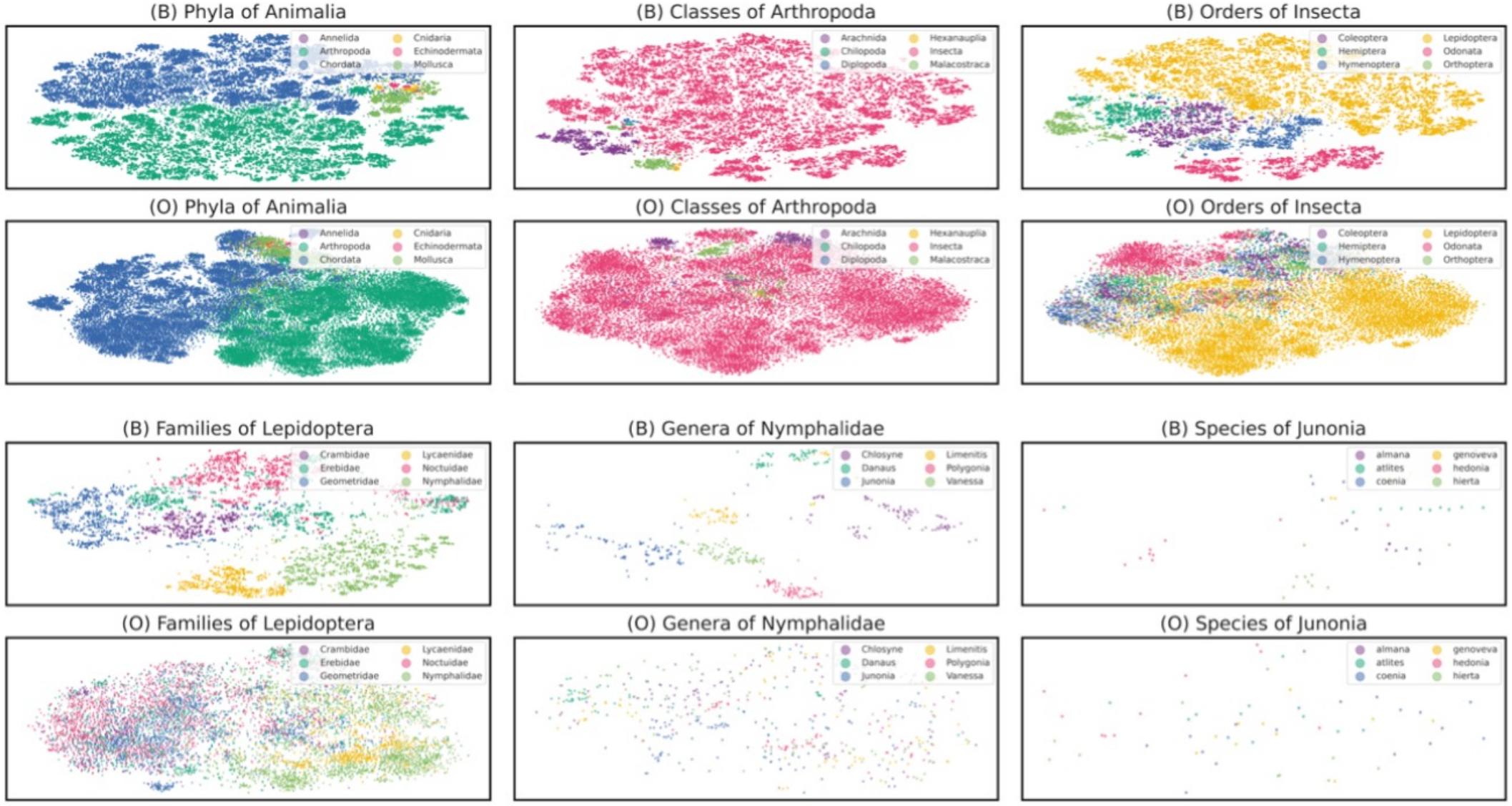


Figure 3. T-SNE visualization of image features, colored by taxonomic labels. BIOCLIP (B) is visualized in the first and third row and OpenAI's CLIP (O) is visualized in the second and fourth rows. BIOCLIP's features better preserve the hierarchical structure: while both BIOCLIP and CLIP cleanly separate the phyla in the Animalia Kingdom (top left), only BIOCLIP successfully separates the orders in the Insecta Class (top right) and the families in the Lepidoptera Order (bottom left).

Conclusion and Future Outlook

- **TREEOFLIFE-10M**: A large-scale and diverse biological image dataset.
- **BIOCLIP**: A foundation model designed for the tree of life, capable of fine-grained biological classification.
  - Strong classification capabilities in both **zero-shot** and **few-shot** settings.
  - **the entire taxonomic name** leads to better generalization
  - Visualization shows that BIOCLIP’ s **image embeddings align** well with the **taxonomic hierarchy**
  - Model learns visual representations for over **450K taxa**

Future Directions

- **Scaling Data**: Expanding to over **100M** research-grade images from iNaturalist.
- **Richer Text Descriptions**: Collecting detailed species descriptions to enable BIOCLIP to capture fine-grained **trait-level** representations.

Any questions?

Thanks for your attention