Contrastive Visual Data Augmentation

Yu Zhou*, Bingxuan Li*, Mohan Tang*, Xiaomeng Jin, Te-Lin Wu, Kuan-Hao Huang, Heng Ji, Kai-Wei Chang, Nanyun Peng

Motivations

Clouded Tiger Cat

(novel specie discovered in 2024)

What is this animal?

African Leopard

GPT4o

Confusable Concept

Motivations

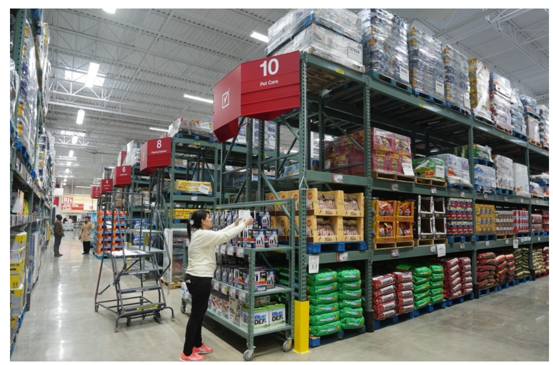
Resupply Base

(visually confusing concept)

What is this place?

Wholesale Store

LLaVA 1.6



Confusable Concept

Related Works

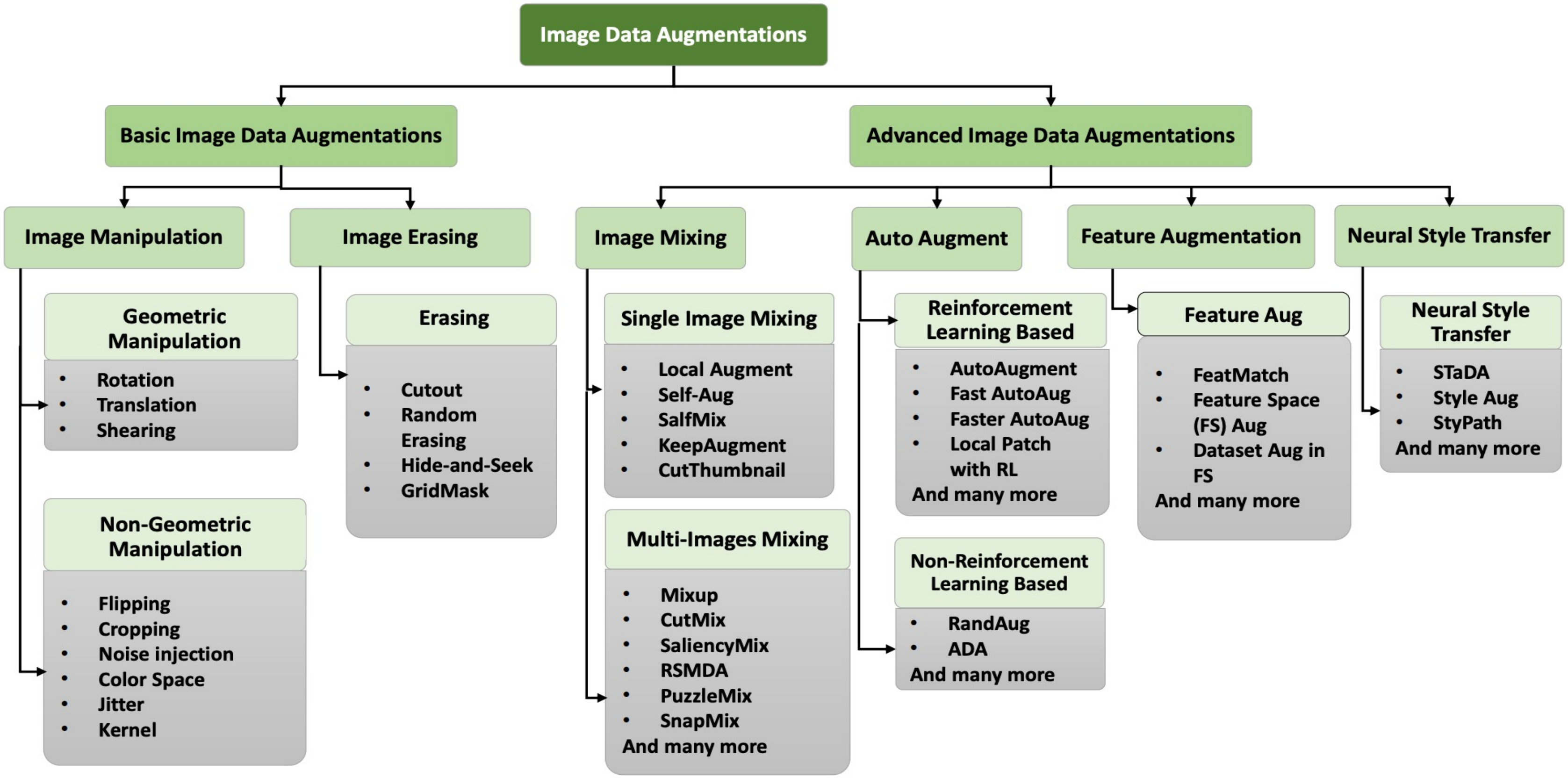
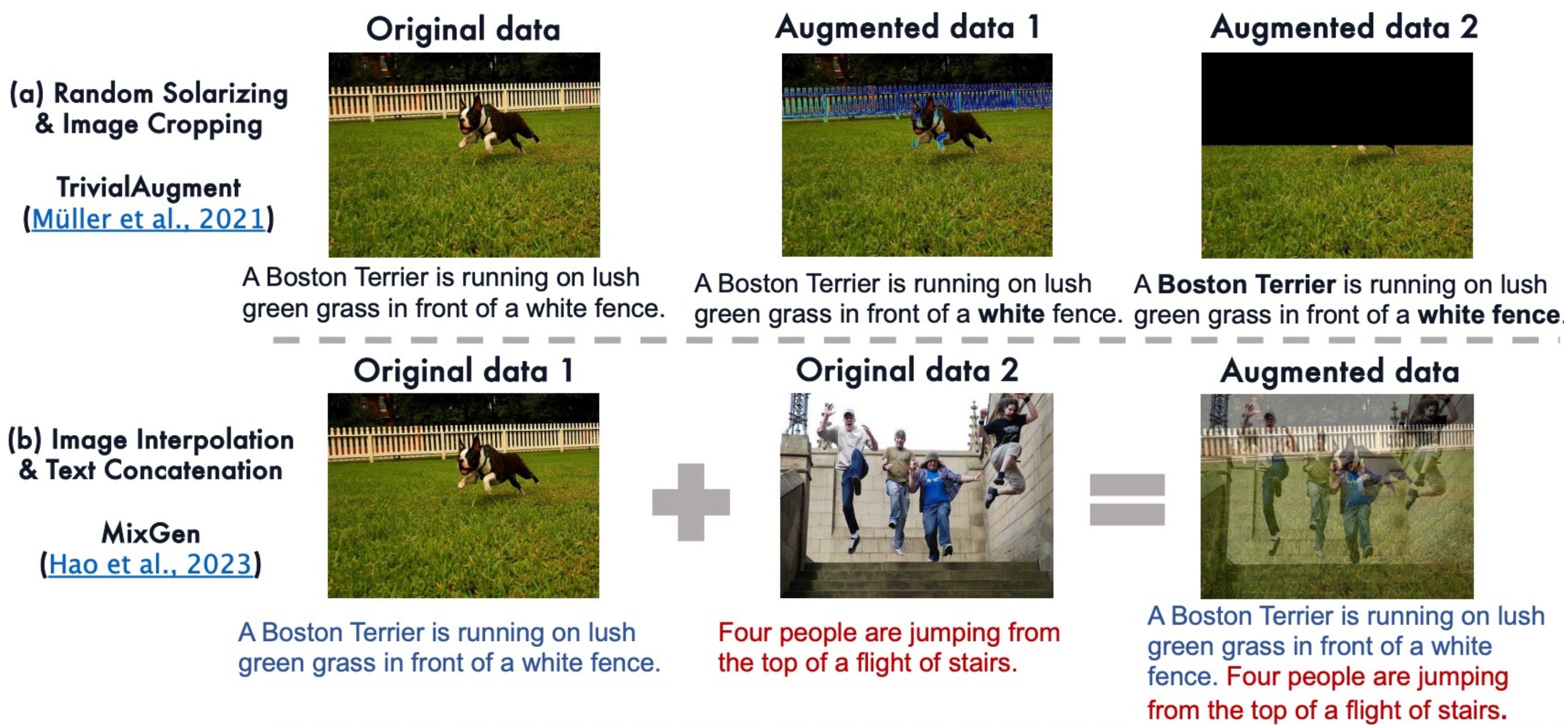


Figure from: Kumar et al. “Image Data Augmentation Approaches: A Comprehensive Survey and Future directions”. 2024

Related Works



UCLA Samueli School of Engineering

Figure from: Jin et al. “ARMADA: Attribute-Based Multimodal Data Augmentation”. 2024

Our Method

Contrastive Visual **D**ata **A**ugmentation **(CoDA)**

Target Concept

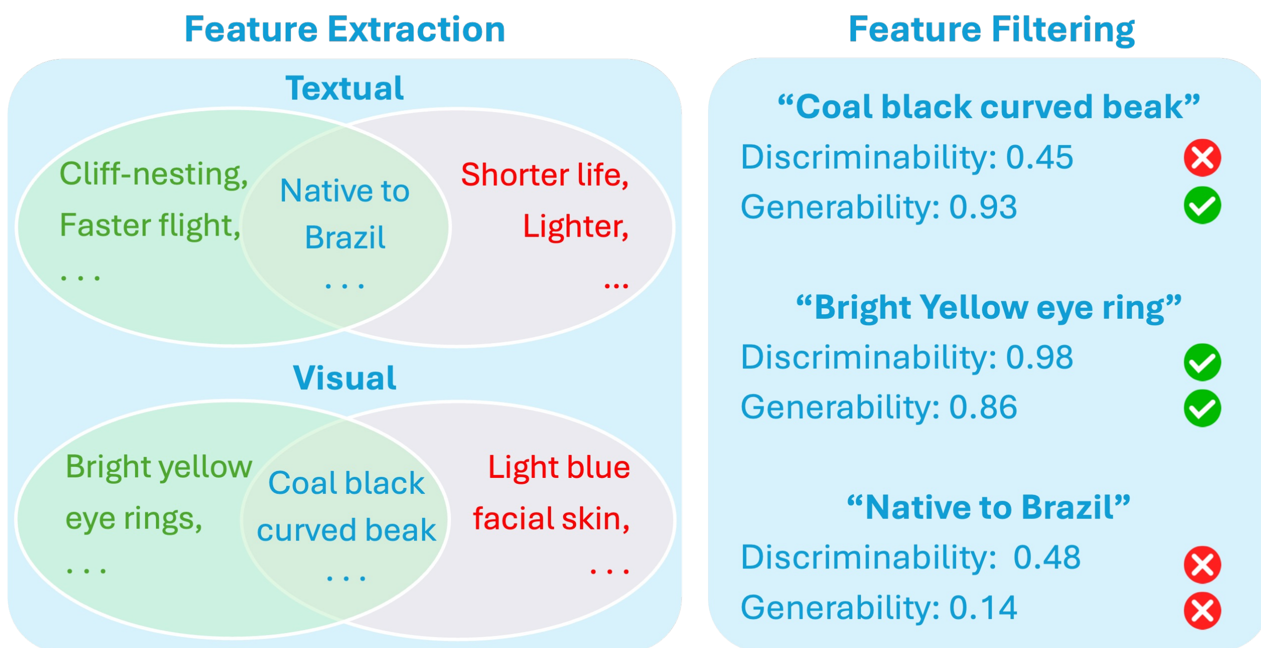
Anodorhynchus Leari
(Lear's Macaw)

Confusable Concept

Cyanopsitta Spixii
(Spix's Macaw)

UCLA Samueli School of Engineering

Our Method


Feature Extraction
Textual
Cliff-nesting,
Faster flight,
. . .
Native to
Brazil
. . .
Shorter life,
Lighter,
...
Visual
Bright yellow
eye rings,
. . .
Coal black
curved beak
. . .
Light blue
facial skin,
. . .
Feature Filtering
"Coal black curved beak"
Discriminability: 0.45
Generability: 0.93
"Bright Yellow eye ring"
Discriminability: 0.98
Generability: 0.86
"Native to Brazil"
Discriminability: 0.48
Generability: 0.14

Our Method

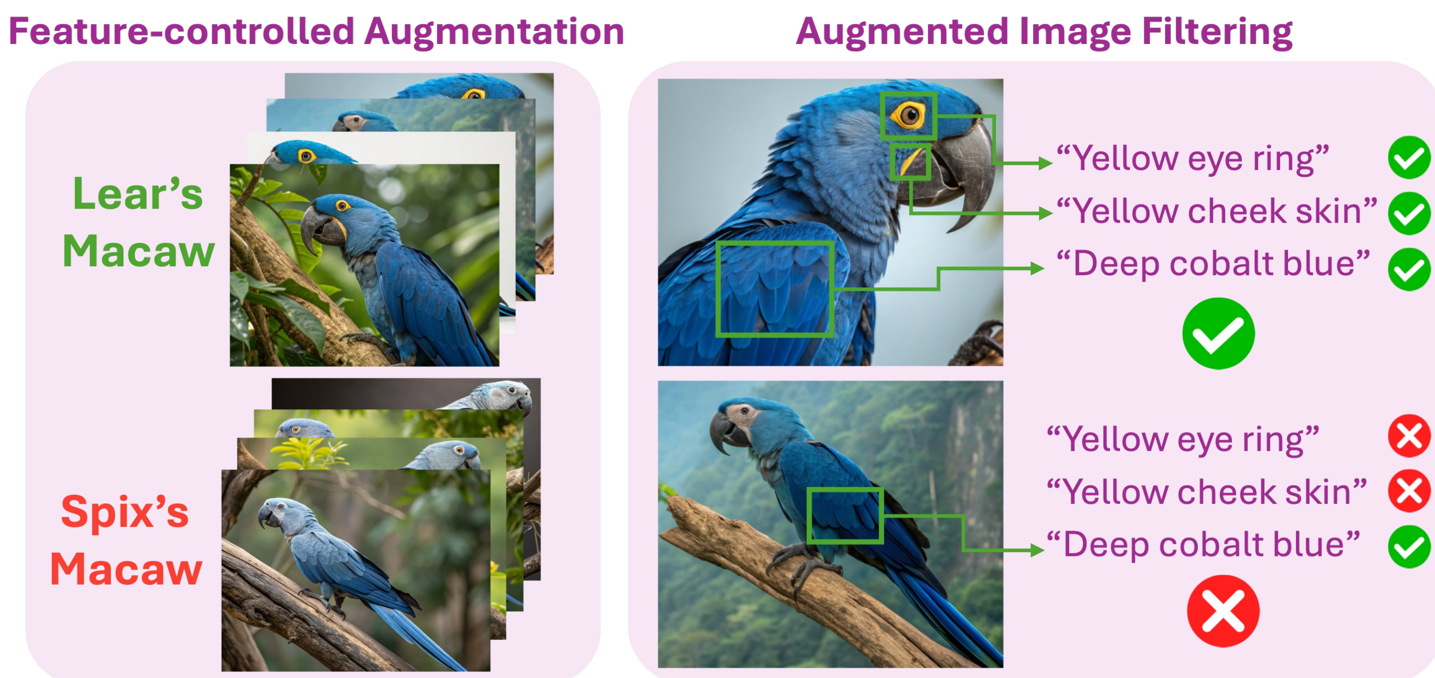
1. Discriminability (D(f, C_T , C_M)) : measures whether a feature f indeed differentiates the target class C_T from the misidentified concept C_M (check whether f is a valid feature of C_T but not C_M) .

$$D(f, \mathcal{C}_T, \mathcal{C}_M) = \sum_{i \in I} \frac{\mathrm{CLIP}(f, i^{\mathrm{real}}_{\mathcal{C}_T})}{\mathrm{CLIP}(f, i^{\mathrm{real}}_{\mathcal{C}_T}) + \mathrm{CLIP}(f, i^{\mathrm{real}}_{\mathcal{C}_M})}$$

2. Generability (G(f, C_T , C_M)) : measures whether a feature f can be properly generated by the text-toimage generative model.

$$G(f, \mathcal{C}_T, \mathcal{C}_M, g) = \sum_{i \in I} \frac{\mathrm{CLIP}(f, i^{\mathrm{synthetic}}_{\mathcal{C}_T})}{\mathrm{CLIP}(f, i^{\mathrm{synthetic}}_{\mathcal{C}_T}) + \mathrm{CLIP}(f, i^{\mathrm{real}}_{\mathcal{C}_M})}$$

Our Method



Feature-controlled Augmentation
Augmented Image Filtering
Lear's Macaw
Spix's Macaw
"Yellow eye ring"
"Yellow cheek skin"
"Deep cobalt blue"
"Yellow eye ring"
"Yellow cheek skin"
"Deep cobalt blue"

Our Method

To verify the final images contain desired target concept features, we propose a simple verification metric: Given the vanilla LMM M, a set of features F, the feature satisfaction rate S(i synthetic, F, M) for each augmented image i is:

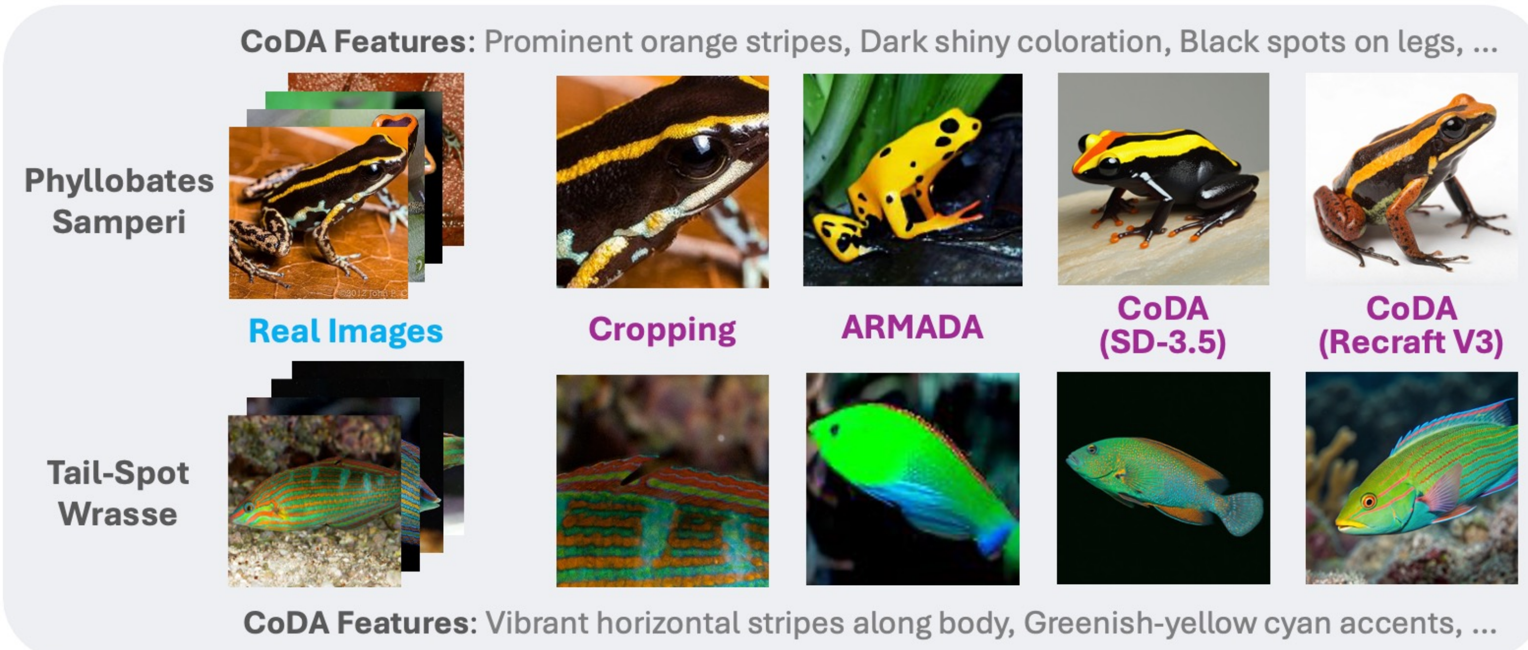
$$S(i^{\text{synthetic}}, \mathcal{F}, \mathcal{M}) = \frac{\sum_{f \in \mathcal{F}} \mathbf{1}\{\mathcal{M}(f, i^{\text{synthetic}})\}}{|\mathcal{F}|}$$

Human evaluation results on a subset of iNaturalist and the NovelSpecies dataset further verifies the reliability of our filtering pipeline:

Image Type	**Target Concept (%)**	**Misidentified Concept (%)**	**Inter-Annotator Agreement (κ)**
Real	92.51	14.32	0.87
Synthetic	83.97	-	0.82

UCLA Samueli School of Engineering

Qualitative Comparison



UCLA Samueli School of Engineering

Quantitative Experiments

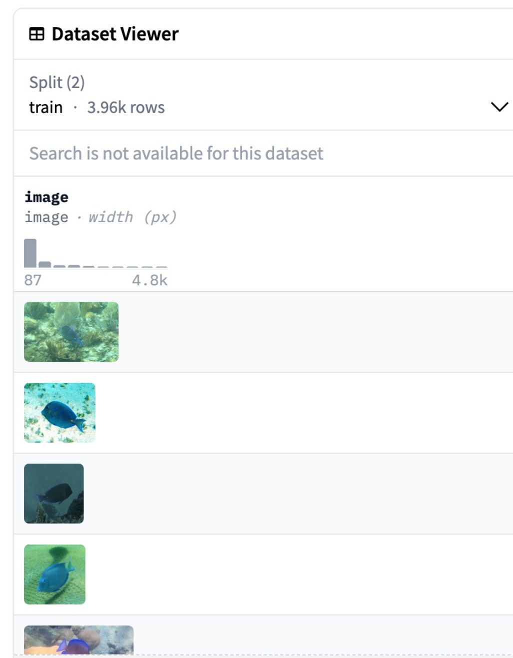
Dataset	Augmentation Method	Feature Type	Fixed Real Data (Real:Syn)				Fixed Compute (Real:Syn)		
			5:0	5:1	5:3	5:5	20:0	10:10	0:20
SUN (Xiao et al., 2010)	Baselines	All Real	**73.4**	-	-	-	**74.3**	-	-
		Cropping	-	78.3	75.8	76.3	-	77.3	76.4
		Flipping	-	75.7	78.4	74.8	-	75.2	76.1
		ARMADA	-	75.9	78.3	77.6	-	76.2	76.8
	CoDA (w/o contrastive)	Textual	-	80.6	79.7	79.4	-	81.3	80.8
		Visual	-	81.3	81.6	79.3	-	80.0	80.8
		T+V	-	82.7	80.7	80.4	-	82.8	82.1
	CoDA	Textual	-	79.2	**83.2**	82.3	-	82.8	82.1
		Visual	-	82.3	81.7	82.2	-	81.8	**83.1**
		T+V	-	**83.4**	81.7	**82.6**	-	**83.3**	82.1
iNaturalist (Van Horn et al., 2018)	Baselines	All Real	**49.2**	-	-	-	**64.3**	-	-
		Cropping	-	59.7	58.8	62.2	-	61.4	63.9
		Flipping	-	61.0	61.1	62.3	-	62.1	62.7
		ARMADA	-	60.1	60.7	61.1	-	61.6	58.5
	CoDA (w/o contrastive)	Textual	-	63.9	64.6	**66.5**	-	65.6	63.2
		Visual	-	65.0	64.7	64.3	-	65.6	63.2
		T+V	-	62.8	64.4	62.3	-	64.4	63.4
	CoDA	Textual	-	63.9	**67.8**	62.6	-	65.0	**64.9**
		Visual	-	**67.0**	66.0	65.1	-	62.5	60.9
		T+V	-	63.5	65.0	64.6	-	**67.0**	64.1

UCLA

NovelSpecies Dataset

1. Bypass the issue of evaluation data leakage with 0% risk of training data contamination.
2. Evaluate LMMs' ability to recognize novel species discovered after its knowledge cutoff.

common_name	latin_name	extinct?	sub_category	yr_discovered
Northern giant hummingbird	Patagona peruviana	No	Birds	2024
Northern silvery-cheeked antshrike	Sakesphoroides niedeguidonae	No	Birds	2024
Coapilla arboreal alligator lizard	Abronia cunemica	no	Reptiles	2024
Hussain's Eyelash-Viper	Bothriechis hussaini	no	Reptiles	2024
Khwarg's Eyelash-Pitviper	Bothriechis khwargi	no	Reptiles	2024
Peruvian Yungas pudu	Pudella carlae	no	Mammals	2024
Villa's yellow-eared bat	Vampyressa villai	no	Mammals	2024
Clouded tiger cat	Leopardus pardinoides	no	Mammals	2024



UCLA Samueli School of Engineering

Quantitative Experiments

Augmentation Method	Feature Type	LLaVA-NeXT				GPT4o-mini				ViT			
		5:0	5:1	5:3	5:5	5:0	5:1	5:3	5:5	5:0	5:1	5:3	5:5
Baselines	All Real	**61.2**	-	-	-	**84.3**	-	-	-	**75.4**	-	-	-
	Cropping	-	60.4	60.4	59.5	-	84.8	86.3	85.9	-	78.3	77.6	79.6
	Flipping	-	60.7	62.9	60.1	-	83.2	83.5	84.3	-	76.9	77.9	78.2
	ARMADA	-	60.7	60.2	61.2	-	84.1	84.3	83.9	-	76.3	76.4	78.6
CoDA (w/o contrastive)	Textual	-	74.8	75.1	74.7	-	87.6	87.2	87.0	-	82.5	84.5	84.7
	Visual	-	76.5	77.9	76.2	-	88.3	89.6	88.2	-	82.5	83.0	82.6
	T+V	-	77.6	**78.9**	78.8	-	89.5	**91.2**	87.9	-	84.3	84.9	82.5
CoDA	Textual	-	76.4	75.9	76.8	-	87.1	87.9	87.4	-	84.6	85.0	84.5
	Visual	-	77.5	78.1	77.9	-	91.3	90.8	**92.6**	-	**85.5**	84.6	85.7
	T+V	-	**78.8**	78.7	**79.2**	-	**91.6**	90.8	91.4	-	85.3	**85.8**	**86.3**

Thank you!

https://contrastive-visual-data-augmentation.github.io

https://github.com/PlusLabNLP/CoDA

https://huggingface.co/datasets/uclanlp/CoDA

https://x.com/yu_bryan_zhou/status/1917682626435113122