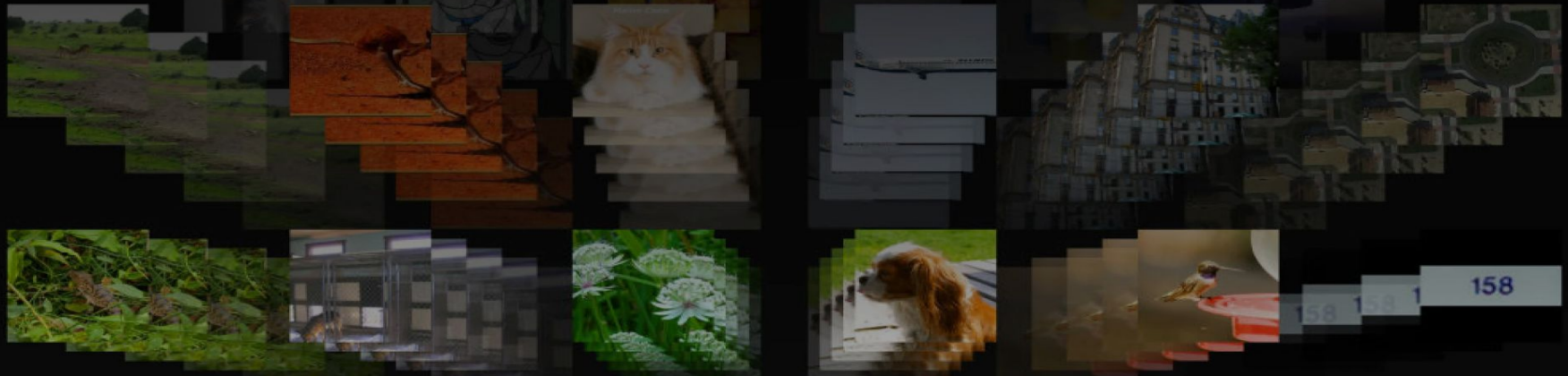


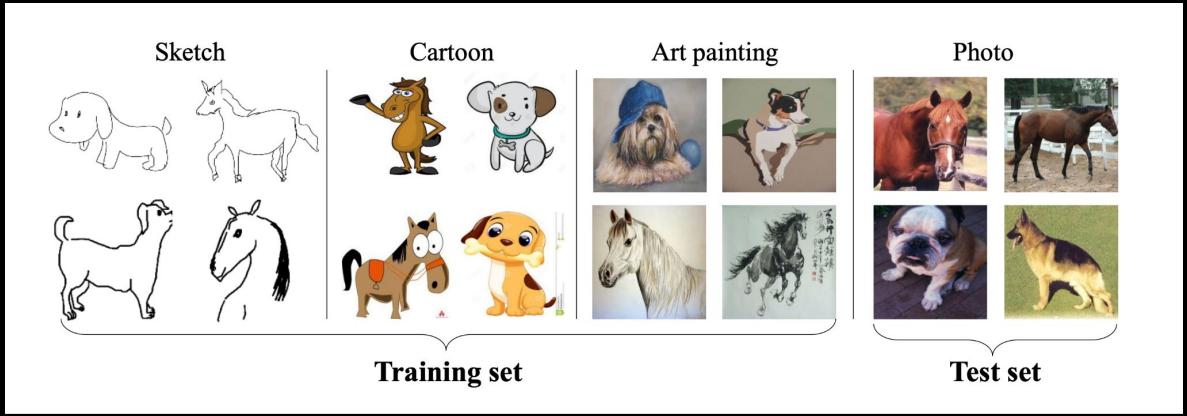
Generalizing to Unseen Domains via Text-guided Augmentation

Daiqing Qi[1], Handong Zhao[2], Aidong Zhang[1], Sheng Li[1]

University of Virginia[1], Adobe Research[2]

Background

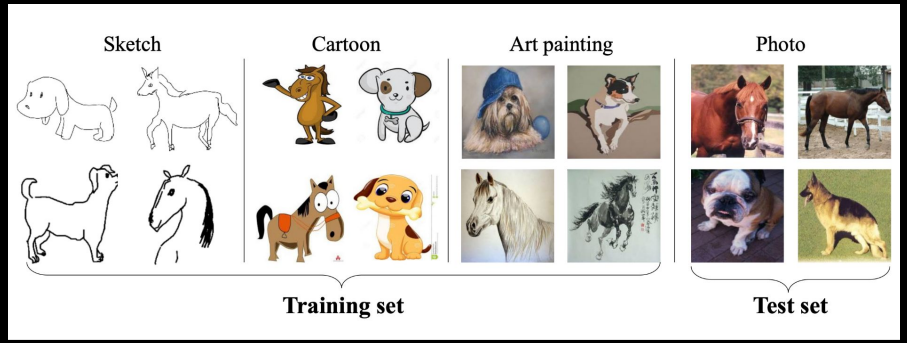
- Domain Generalization



Examples from the dataset PACS for domain generalization. The training set is composed of images belonging to domains of sketch, cartoon, and art paintings. DG aims to learn a generalized model that performs well on the unseen target domain of photos.

Background

- Domain Generalization



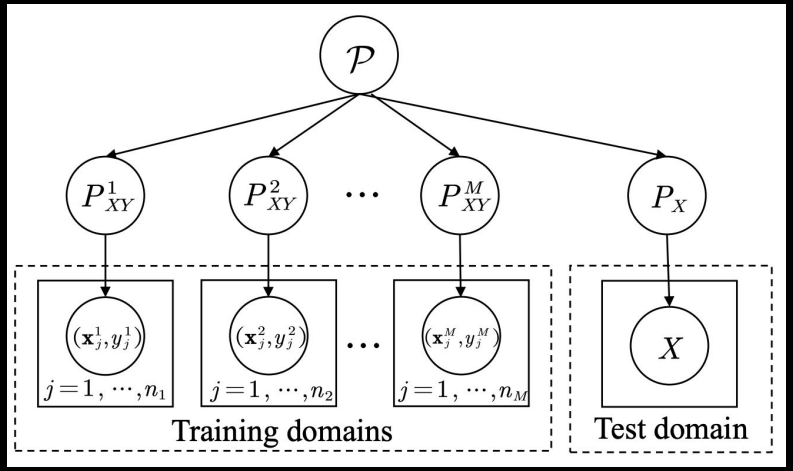
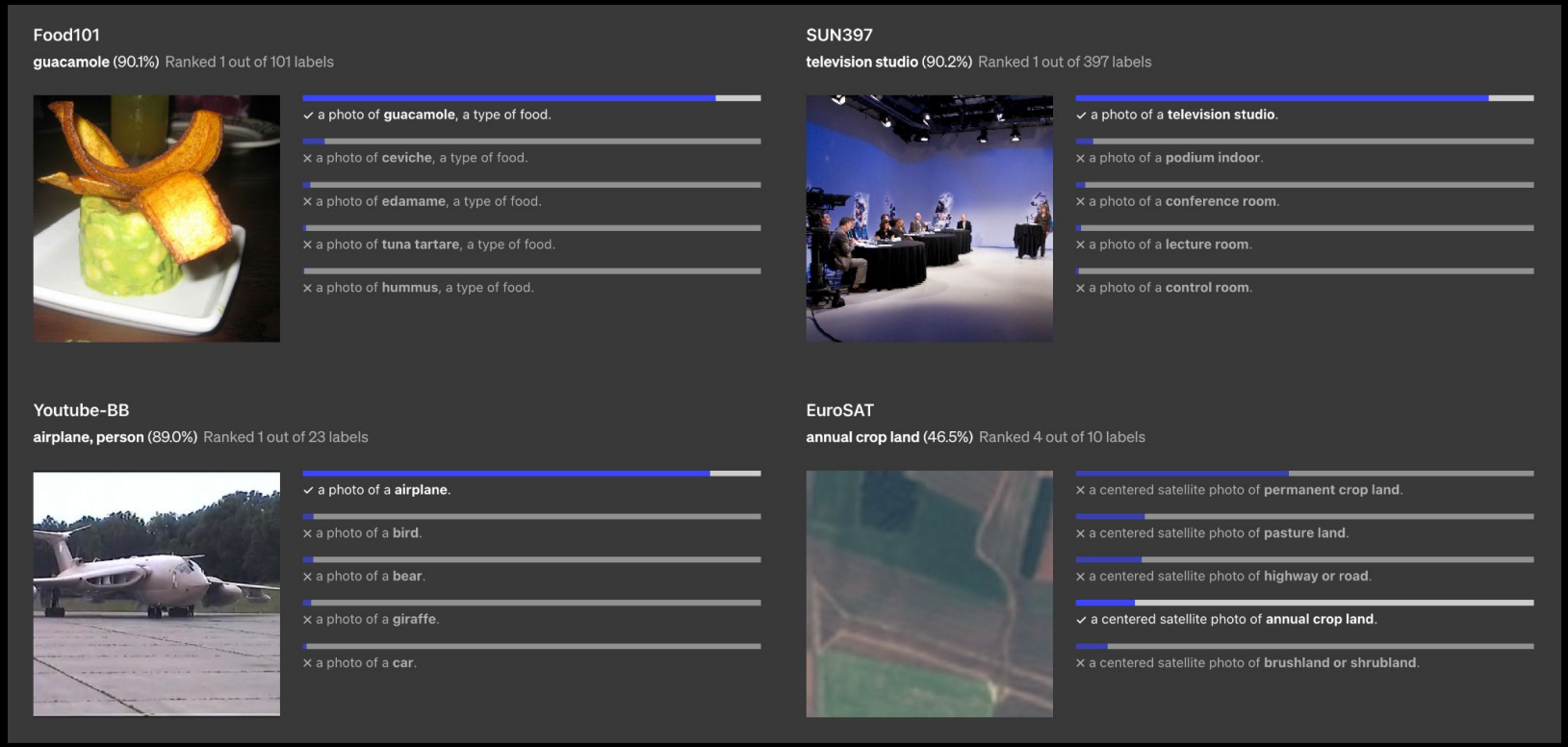


Illustration of Domain Generalization (DG)

Background

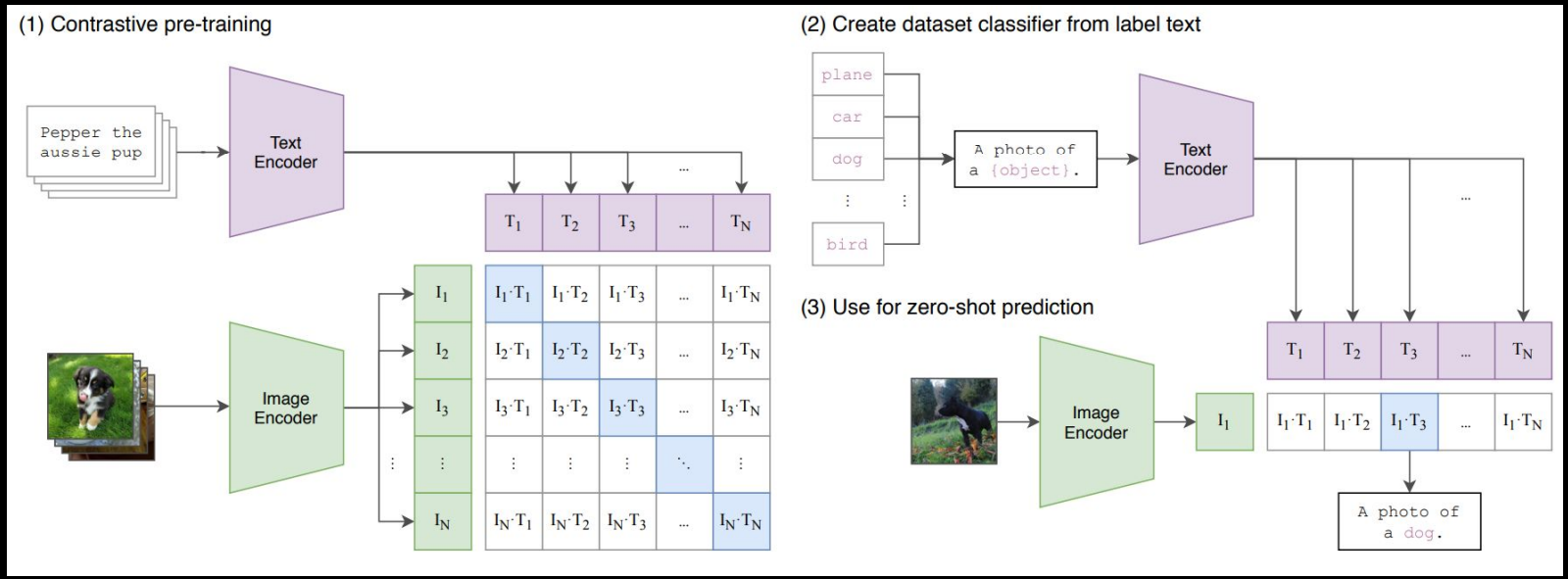
- Contrastive Image Language Pre-training (CLIP)



Learn alignment between natural language and image

Background

- Contrastive Image Language Pre-training (CLIP)



Summary of CLIP. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

Motivation

“Why is Text-guided Augmentation for Test Domains Important?”

- It is expensive to collect training data for every possible test domain
- It is easier to verbalize the test domains (e.g. “photos of birds”) and perform a text-guided augmentation during training [1]

[1] Dunlap et la., Using Language to Extend to Unseen Domains, ICLR 2023

Motivation

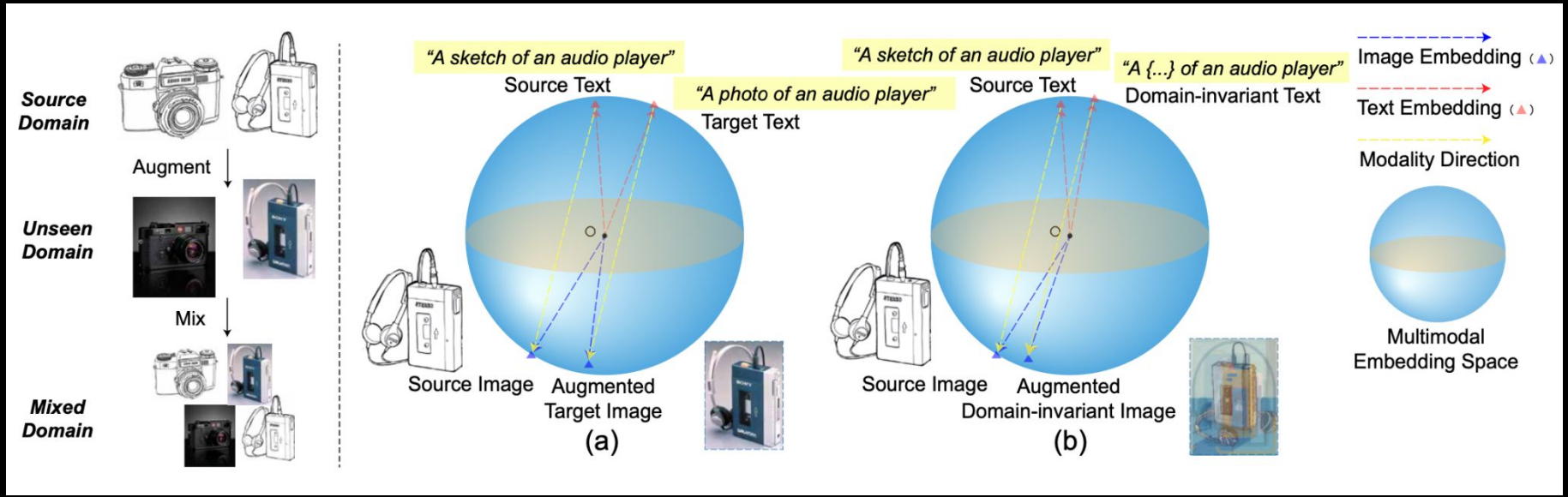
“Why is Domain-Invariant Text-guided Augmentation Important?”

- In many real-world applications, text information of test domains is not always available in advance.
- Even if we can verbalize all test domains, it is laborious for existing works [1] to train a different augmentation network for each possible unseen domain, which suffers from time and computation inefficiency.

[1] Dunlap et la., Using Language to Extend to Unseen Domains, ICLR 2023

Motivation

We benefit from the "*multimodal embedding space*" of a pre-trained vision-language model, and propose to acquire *training-free* and *domain-invariant* augmentations with text descriptions of arbitrary crafted unseen domains.



Contribution

- We explore an interesting yet under-explored problem, i.e., learning a model that extends well to test domains with only crafted text descriptions from arbitrary unseen domains (not test domains). We call it Text-driven Domain Generalization problem.

- With the multimodal embedding space of a pre-trained VL model, we pro- pose a novel training-free embedding augmentation method with theoretical guarantees, based on the geometric characteristics of the embedding distribution.

- Furthermore, combined with our training-free technique, we build a framework with our augmentation method that performs domain-invariant aug- mentations to solve the Text-driven Domain Generalization problem, which is more time-efficient while achieving better results than competing baselines.

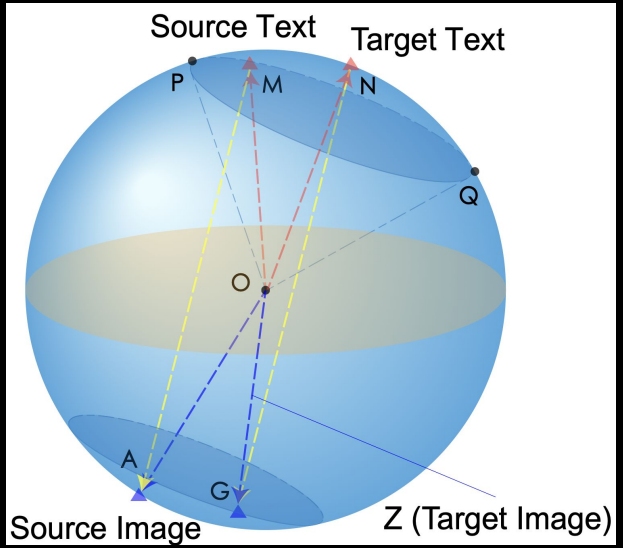
Problem Formulation

Formally, we are provided with a training dataset drawn from the source domain. Additionally, we are given a text description of the source domain and a set of text descriptions of crafted domains that are **distinct** from the test domains.

The objective is to develop a model that can generalize effectively to novel test domains, leveraging both the source and crafted domains. This involves text-driven data augmentation and training a linear probe using both the source and augmented image features.

Training-free Augmentation with Modality Direction

Modality Direction. The modality direction is defined as the difference of embeddings from one modality to another. The yellow arrows visualize the modality directions of image-text pairs.



Aligning modality direction is more appropriate than aligning global direction for training-free augmentation in following aspects:

- Better theoretical support
- Better preservation of class information
- Milder assumption for an analytical solution

Training-free Augmentation with Modality Direction

- **Modality Direction.** The modality direction is defined as the difference of embeddings from one modality to another. The yellow arrows visualize the modality directions of image-text pairs.

Training-free Augmentation with Modality Direction. Denote the output of the augmentation function $f_{\text{aug}}(\boldsymbol{x}, t_{\text{target}}; y, t_{\text{source}}; y)$ as a variable $\boldsymbol{z}$, by aligning the modality direction, the following equation should hold for each image-text pair $(\boldsymbol{x_i}, t_{\text{source}}; y_i)$:

$$\frac{\boldsymbol{z} - h^T(t_{\text{target}}\ ; y_i)}{\|\boldsymbol{z} - h^T(t_{\text{target}}\ ; y_i)\|} \cdot \frac{h^I(\boldsymbol{x_i}) - h^T(t_{\text{source}}; y_i)}{\|h^I(\boldsymbol{x_i}) - h^T(t_{\text{source}}\ ; y_i)\|} = 1 \tag{3}$$

- **Augmented Image Embedding.** The solution of (3), i.e., the value of **z** is the desired augmented image embedding.

Training-free Augmentation with Modality Direction

- **Modality Direction.** The modality direction is defined as the difference of embeddings from one modality to another. The yellow arrows visualize the modality directions of image-text pairs.

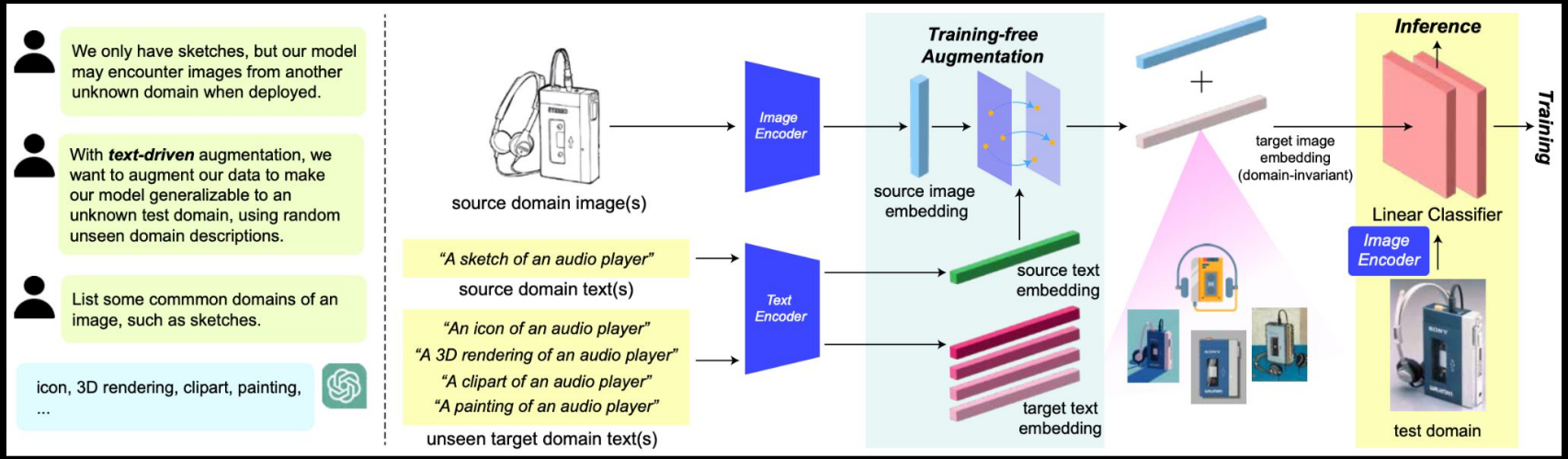
Training-free Augmentation with Modality Direction. Denote the output of the augmentation function $f_{\text{aug}}(\boldsymbol{x}, t_{\text{target}}; y, t_{\text{source}}; y)$ as a variable $\boldsymbol{z}$, by aligning the modality direction, the following equation should hold for each image-text pair $(\boldsymbol{x_i}, t_{\text{source}}; y_i)$:

$$\frac{\boldsymbol{z} - h^T(t_{\text{target}}\ ; y_i)}{\|\boldsymbol{z} - h^T(t_{\text{target}}\ ; y_i)\|} \cdot \frac{h^I(\boldsymbol{x_i}) - h^T(t_{\text{source}}; y_i)}{\|h^I(\boldsymbol{x_i}) - h^T(t_{\text{source}}\ ; y_i)\|} = 1 \tag{3}$$

- **Augmented Image Embedding.** The solution of (3), i.e., the value of **z** is the desired augmented image embedding.

- ***Private* vs. *Global* Modality Direction.** The modality direction can be given be a single image-text pair or the average of all image-text pairs.

Framework



(1) Given source domain images (sketch), we first acquire several text descriptions of unseen domains (different from the test domain).

(2) Embed all texts and images into the CLIP embedding space.

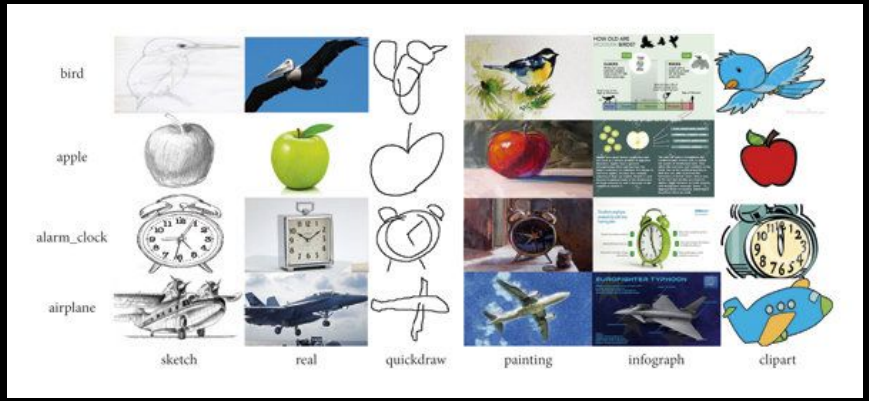
(3) A training-free augmentation is performed to obtain domain-invariant image embeddings under the guidance of crafted unseen domain descriptions.

(4) A linear classifier is trained on the mix of source and augmented embeddings.

Datasets

- **DomainNet.** Following LADS, we use a specific split of the **DomainNet** dataset which contains 40 most common classes from 4 domains: ‘sketch’, ‘real’, ‘clipart’, and ‘painting’. We train on sketches and evaluate on the three other domains.

- **CUB-Paintings.** It combines CUB-200 and CUB-200-Paintings, where there are 200 different bird species from “photo” and "painting". Following LADS, we train on phtots and evaluate on painting.



Examples from **DomainNet**

Main Results

Dataset	Method	Average	ID	OOD	Training-free (Stage-1)	Time (Stage-2)
CUB-Paintings	CLIP LP (ZS init)	75.57±0.06%	86.08±0.11%	65.05±0.05%	-	-
CUB-Paintings	WiSE-LP	73.27±0.22%	81.74±0.34%	64.80±0.10%	-	-
CUB-Paintings	LADS	74.99±0.23%	85.33±0.29%	64.85±0.26%	×	1 ×
CUB-Paintings	TEAM–*full* (Ours)	76.94±0.22%	86.61±0.21%	67.26±0.23%	✓	1 ×
CUB-Paintings	TEAM-*invar.* (Ours)	**77.16±0.18%**	**86.69±0.24%**	**67.31±0.23%**	✓	0.23 ×
DomainNet	CLIP LP (ZS init)	94.58±0.11%	95.21±0.21%	93.95±0.03%	-	-
DomainNet	WiSE-LP	94.44±0.11%	95.19±0.34%	93.68±0.12%	-	-
DomainNet	LADS	94.97±0.25%	95.29±0.33%	94.65±0.09%	×	1 ×
DomainNet	TEAM-*full* (Ours)	96.22±0.16%	**95.87±0.27%**	96.58±0.19%	✓	1 ×
DomainNet	TEAM-*invar.* (Ours)	**96.28±0.18%**	95.61±0.21%	**96.90±0.20%**	✓	0.23 ×

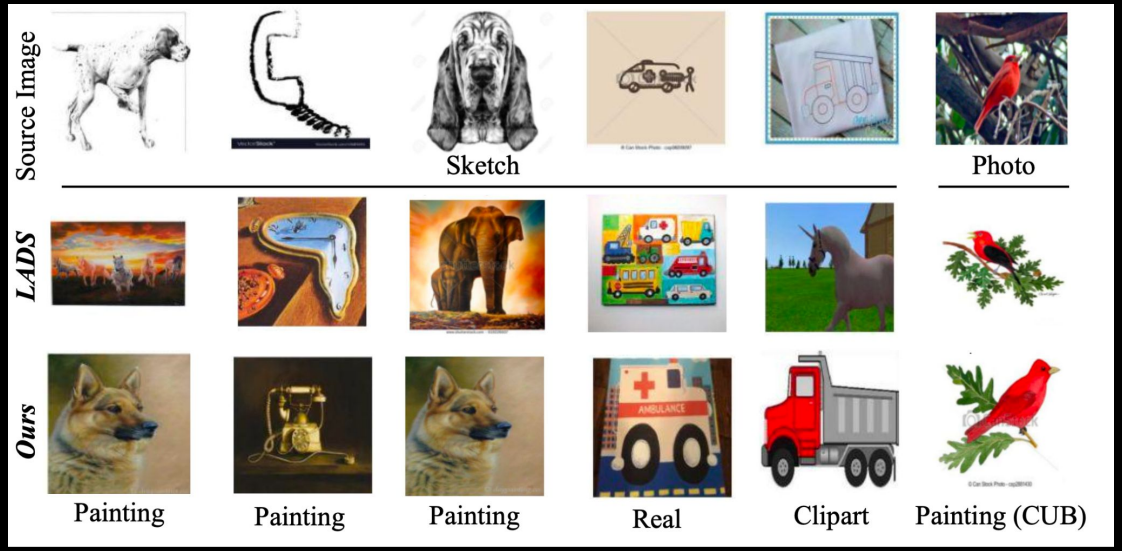
In-domain (ID), out-of-domain (OOD) and the average (of ID and OOD) accuracy on CUB-Paintings and DomainNet. Note that OOD is the major metric, where the goal is to improve OOD accuracy without eroding ID accuracy.

Ablation Study

Aug. Method	*Invar.* Mode	Average	ID	OOD	Training-free (Stage-1)	Time (Stage-2)
LADS	*None*	74.99±0.23%	85.33±0.29%	64.85±0.26%	×	1 ×
Global Dir.	Mean Pooling	74.67±0.22%	85.21±0.21%	64.12±0.21%	✓	0.23 ×
Modality Dir.	Mean Pooling	77.06±0.19%	86.61±0.22%	67.71±0.23%	✓	0.23 ×
Modality Dir.	Cosine AutoEncoder	77.18±0.21%	86.62±0.18%	67.74±0.23%	✓	0.23 ×
Modality Dir.	*None*	76.84±0.23%	86.54±0.21%	67.14±0.21%	✓	1 ×
Text only	*None*	75.28±0.19%	85.10±0.18%	65.98±0.20%	✓	1 ×

Performances of LADS and our variants on CUB-Paintings dataset. Invar-Mode refers to different methods to obtain domain-invariant representations. None means we do not use domain-invariant representations for augmentation. Text only means using text embeddings for training without being augmented to the image subspace. We report results of our TEAM (G).

Nearest Neighboring Results



Intuitive visualization of the augmented embedding quality from LADS and ours.

Thank You