Deep learning-based Remote Sensing data retrieval: A cross-modal framework

Ushasi Chaudhuri
Roll No. 17431003

Supervisors:
Prof. Avik Bhattacharya, Prof. Biplab Banerjee,
Prof. Mihai Datcu (external co-supervisor)

Center of Studies in Resources Engineering
Indian Institute of Technology Bombay

31 May 2021

Motivation — Retreival

- Within the same dataset, retrieval is a relatively easy work. If we have different modalities of data, ex Pan images, multi-spectral images, SAR images, optical images, or texts, retrieval amongst cross domain becomes more challenging.
- A network learned for a particular modality may not give good performance on different data.

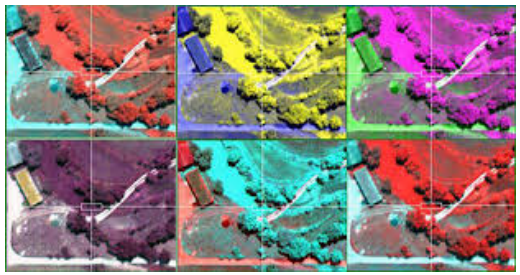
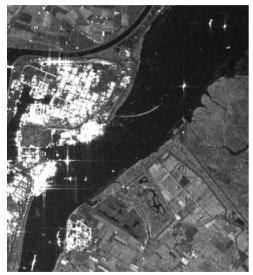
Definition (No free lunch theorem)

[1] If an algorithm performs well on a certain class of problems then it necessarily pays for that with degraded performance on the set of all remaining problems.

[1] Wolpert, D.H., Macready, W.G. (1997), "No Free Lunch Theorems for Optimization", IEEE Transactions on Evolutionary Computation 1, 67.

Motivation

- **PAN images:** High spatial resolution.
- **Multi-spectral images:** high spectral resolution.
- **SAR images:** Polarization information. No cloud clutter.
- **VHR optical image:** High spatial resolution, 3 spectral channel.

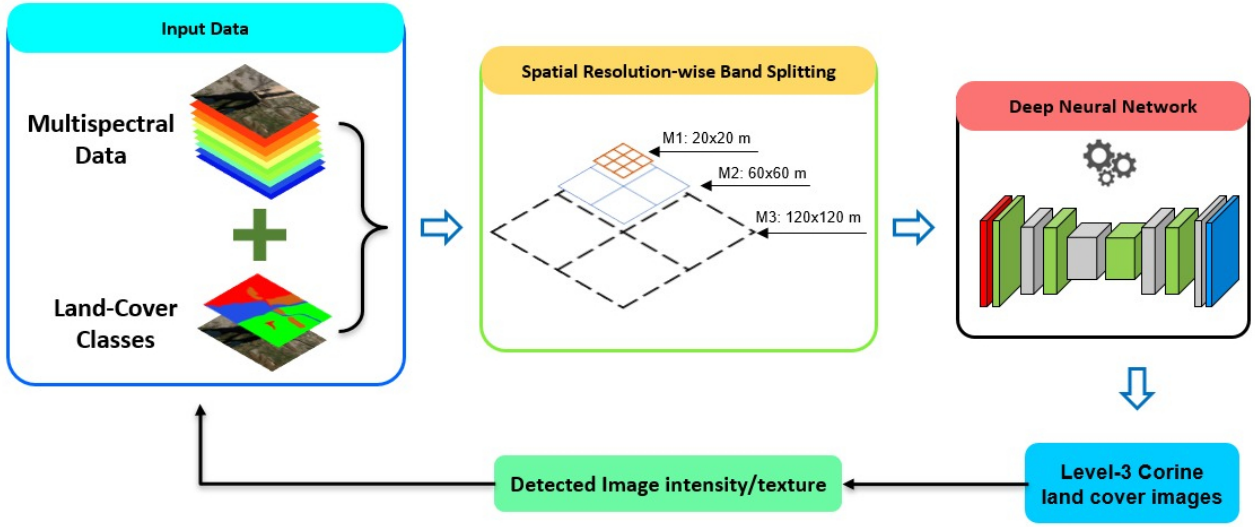

1. 1m pan IKONOS image of Valparaiso, Chile
2. source: https://www.sensorsmag.com/components/hyperspectral-and-multispectral-imaging-sensors-find-hyper-flexible-applications
3. Sentinel-1 SAR dataset with C band
4. `sparseresidentialarea` from PatternNet dataset.

2a

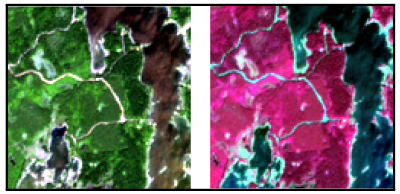
Inter-band retrieval

(a) Multi-modal classification
(b) Multi-modal retrieval

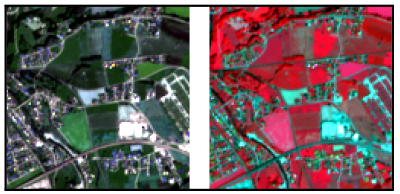
Inter-band Retreival — Captured together



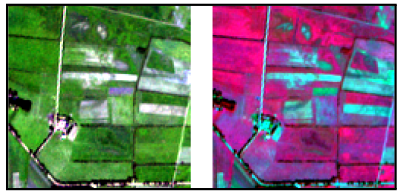
	Res.	Imagery	Utility
B1	60	Aerosol	Coastal and aerosol studies.
B2	10	Blue	Bathymetric mapping, distinguish soil/vegetation.
B3	10	Green	Emphasizes peak vegetation, helps assess plant vigor.
B4	10	Red	Discriminates vegetation slopes.
B5	20	Red Edge	Estimator of LAI, biomass content and shorelines.
B6	20	Red Edge	Estimate chlorophyll. Discriminates soil & vegetation.
B7	20	Red Edge	Applications similar to Band 5 and 6.
B8	10	NIR	Detect healthy/unhealthy veg., canopy/urban/water.
B8A	20	Red Edge	Detects bare soil and built-up areas.
B9	60	Vapour	Used for water vapour detection.
B11	20	SWIR	Monitor crops health, thermal mapping.
B12	20	SWIR	Discriminates soil/vegetation moisture, geological faults.



(a) Coniferous forest, Mixed forest, Sea and ocean



(b) Discontinuous urban fabric, Industrial or commercial units, Non-irrigated arable land



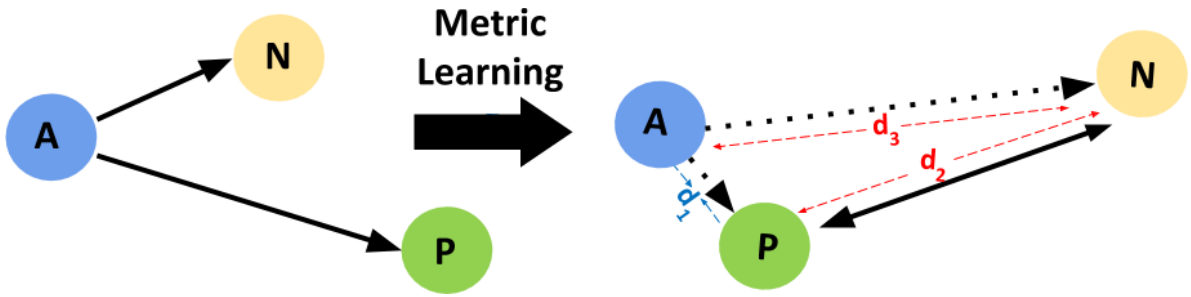
(c) Non-irrigated arable land, Pastures, Complex

Inter-band Retreival — Our contribution

- For **multi-labeled** data, we use Sigmoid activation and a threshold function to get the logits.

$$A_n = I[p_k > \alpha]$$

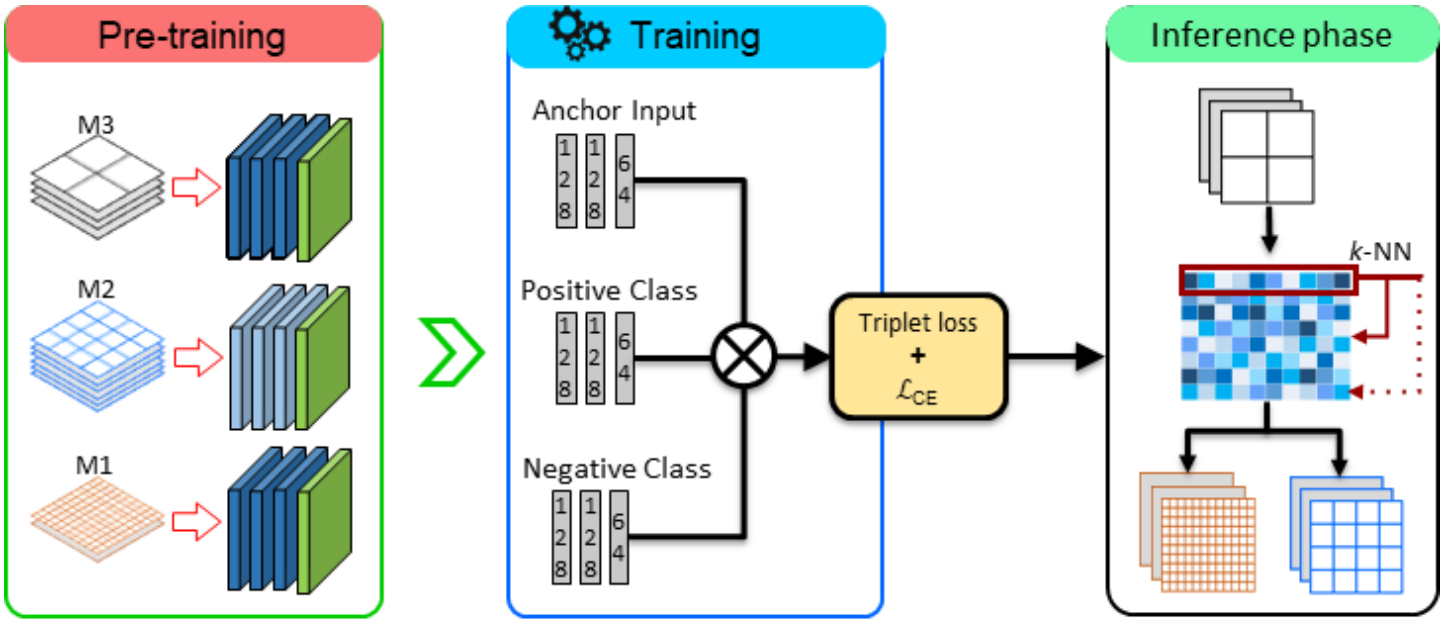
- **Modified triplet loss** for multi-labelled setup

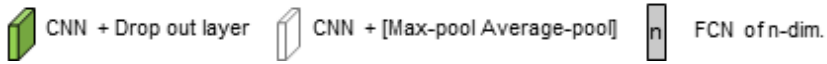


$$\mathcal{L}_{(A,P,N)} = \max(\|f(A) - f(P)\|^2 - \|f(A) - f(N)\|^2 + \alpha, 0)$$
$$+\mathbb{K}(P,N)\max(-\|f(P) - f(N)\| + \beta, 0)$$

here, $\mathbb{K}$ is an Indicator function.

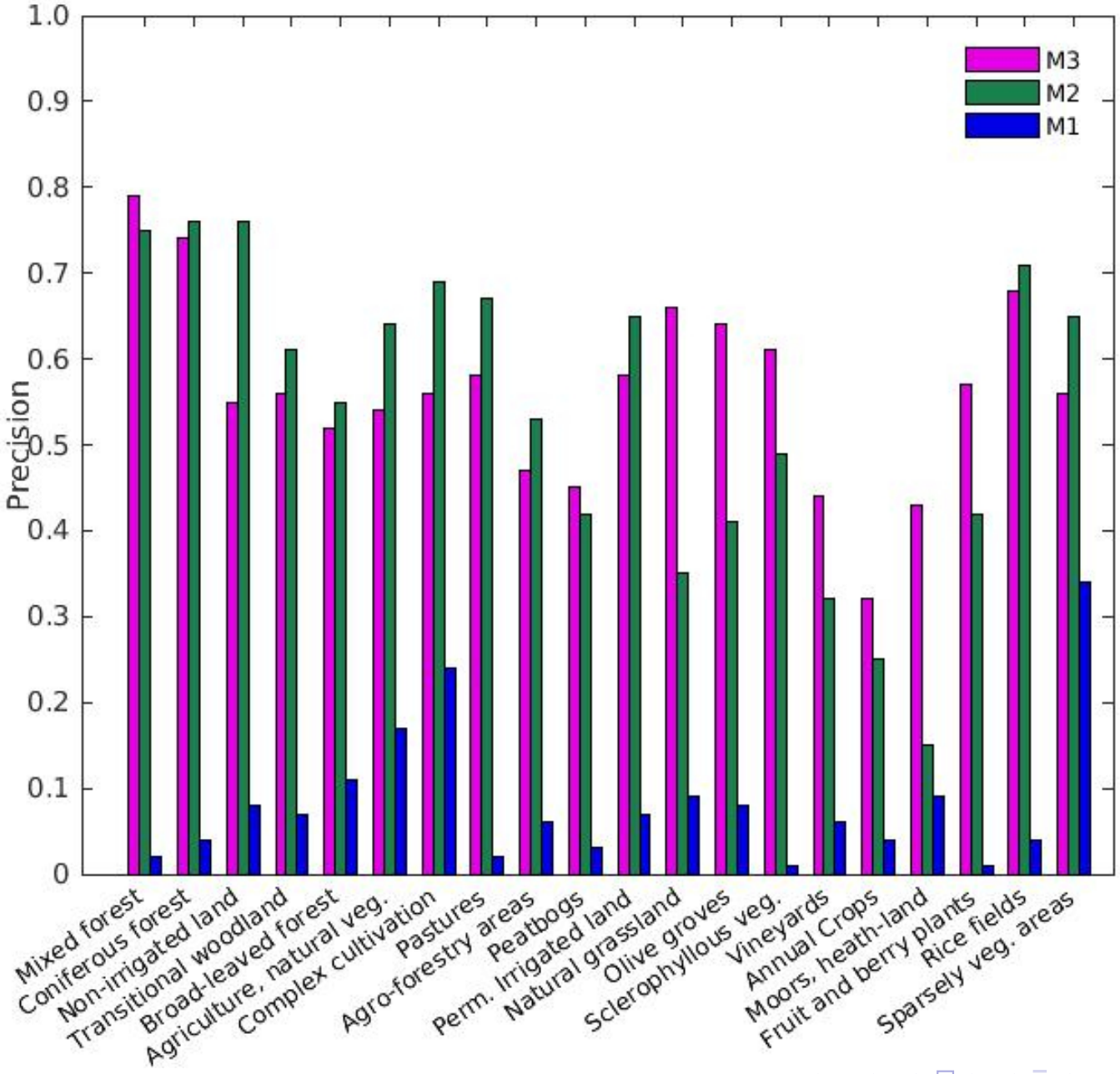
Inter-band Retreival





Method	triplet		m-triplet	
	P@10	mAP	P@10	mAP
M1–20 to M2–60	04.88	05.02	07.04	07.91
M1–20 to M3–120	08.10	07.73	09.83	14.54
M2–60 to M1–20	11.42	09.98	12.92	08.43
M2–60 to M3–120	45.98	56.87	**63.54**	**62.90**
M3–120 to M1–20	12.34	16.72	15.06	21.58
M3–120 to M2–60	59.51	63.79	**62.88**	**69.19**

Inter-band Retreival — Vegetation plot



Cross-modal retrieval

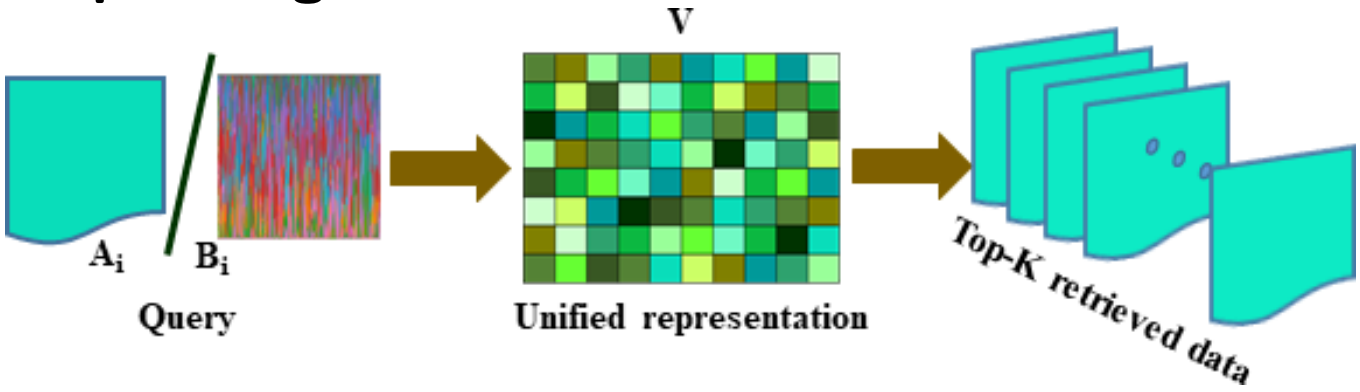
2b

(a) Cross-modal Retrieval
(b) Attention-aware cross-modal retrieval

Cross-modal Retreival

- Important problems:
  1. PAN $\Leftrightarrow$ Multispectral.
  2. Optical $\Leftrightarrow$ SAR.
  3. RGB $\Leftrightarrow$ DEM.
  4. Image $\Leftrightarrow$ text.
  5. Image $\Leftrightarrow$ sketch.
  6. LiDAR $\Leftrightarrow$ hyper-spectral.
  7. Image $\Leftrightarrow$ speech.

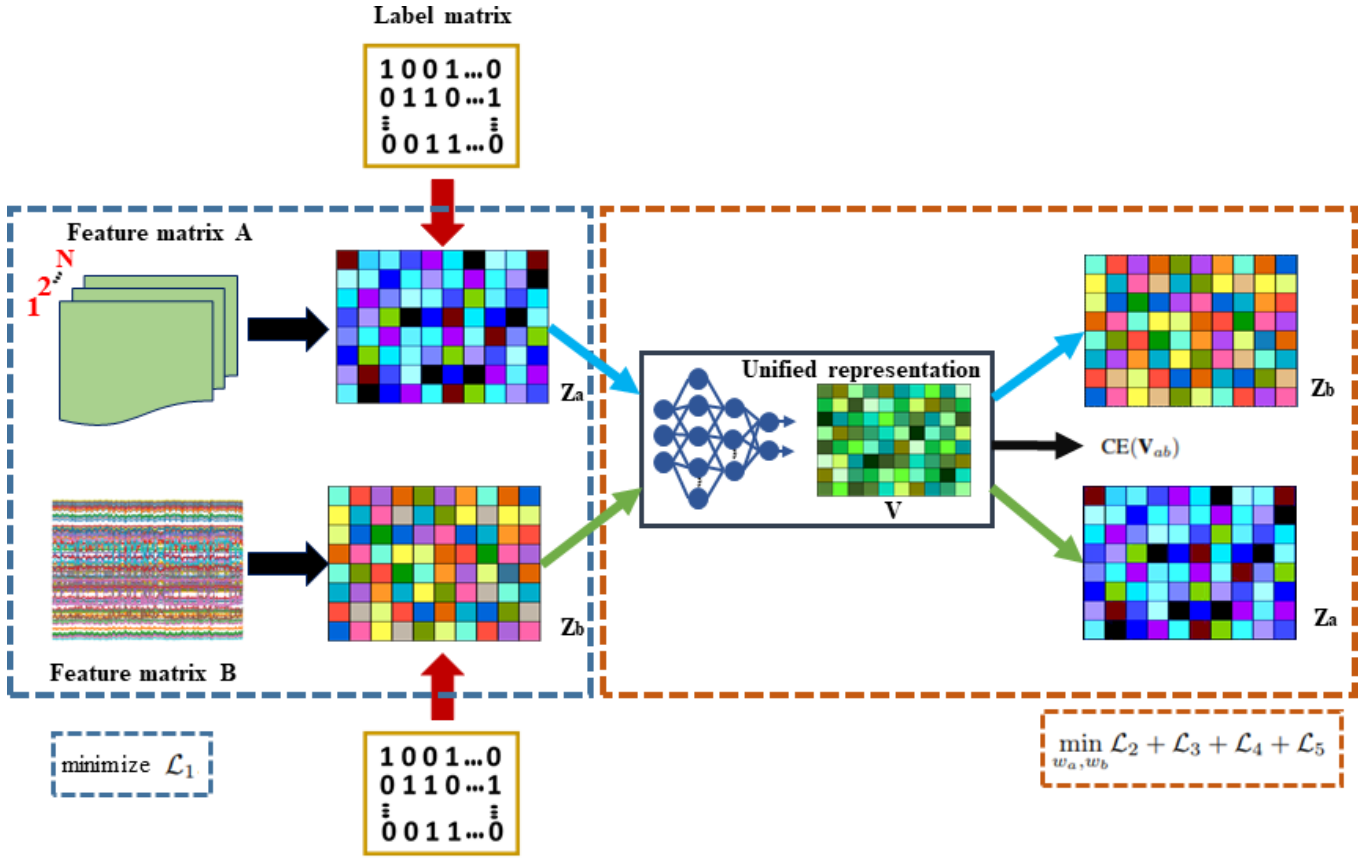
Upcoming: VideoSAR: text and audio annotation



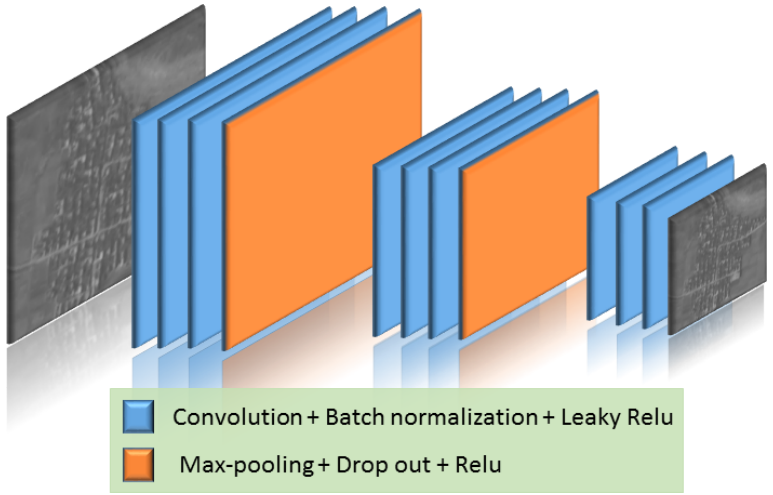
Cross-modal Retreival

Datasets used:

1. DSRSID: Panchromatic - Multi-spectral images (GF-1 satellite). 80,000 images, 8 classes. **(Single-label)**.
2. UCMerced: VHR RS images (RGB) - speech. Merced: 17 multi-labels. speech: 15 samples each. **(Multi-label)**.
3. **Earth on Canvas*:** 14 image and sketch classes.



Cross-modal Retreival



- Train 2 separate classification networks $\{(a_k, l_k)\}_{k=1}^{|\mathbf{A}|}$ and $\{(b_j, l_j)\}_{j=1}^{|\mathbf{B}|}$.
- Extracted features ($\mathbf{Z}_{a_k}$ and $\mathbf{Z}_{b_j}$) are made highly non-redundant by adding a soft orthogonality constrained.

$$\mathcal{L}_{\mathbf{A/B}} = \mathrm{CE}(\tilde{\mathbf{Z}}_{a/b}) + ||\tilde{\mathbf{Z}}_{a/b}^{T}\tilde{\mathbf{Z}}_{a/b} - \mathbf{I}||_{\mathbf{F}}^{2}$$

- $\{\mathbf{Z}_{a_i}\}$ and $\{\mathbf{Z}_{b_i}\}$ are considered as inputs for obtaining V.

Objective Function

- 1. **Difference** between each pair of corresponding i^{th} samples in $\mathbf{V}_a$ and $\mathbf{V}_b$ ($\mathcal{L}_2$):

$$\mathcal{L}_2 = ||\mathbf{V}_a - \mathbf{V}_b||_{\mathbf{F}}^2$$

- 2. **Classification** loss on $\mathbf{V}_{ab} = [\mathbf{V}_a, \mathbf{V}_b]$ ($\mathcal{L}_3$):

$$\mathcal{L}_3 = \mathrm{CE}(\mathbf{V}_{ab})$$

- 3. Separate **feature norm** loss measures on both $\mathbf{V}_a$ and $\mathbf{V}_b$, ($\mathcal{L}_4$) (Since the range of values of raw data features varies widely):

$$\mathcal{L}_4 = ||\mathbf{V}_a||_{\mathbf{F}}^2 + ||\mathbf{V}_b||_{\mathbf{F}}^2$$

- 4. **Decoder** loss which is deemed to reconstruct cross-domain samples given the latent representations: ($\mathcal{L}_5$):

Cross-modal Retreival

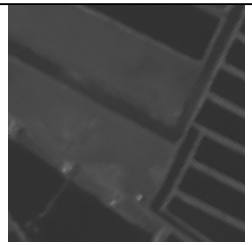
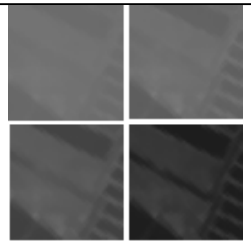
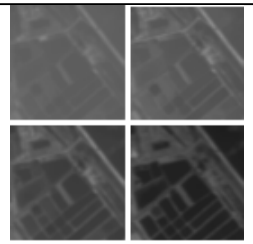
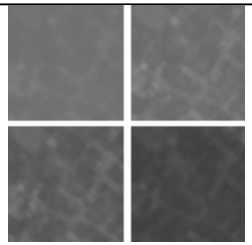
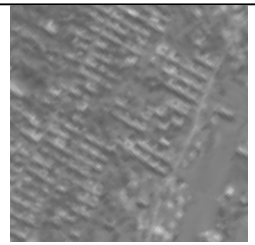
Dataset: Query	Retrived 1	Retrived 2	Retrived 3
			
			
Pavement			
	Bare Soil	Buildings	Grass

Figure: Top-3 retrieval instances from cross-modal query samples.

Cross-modal Retreival

P@50 for both PAN to MUL and MUL to PAN cases.

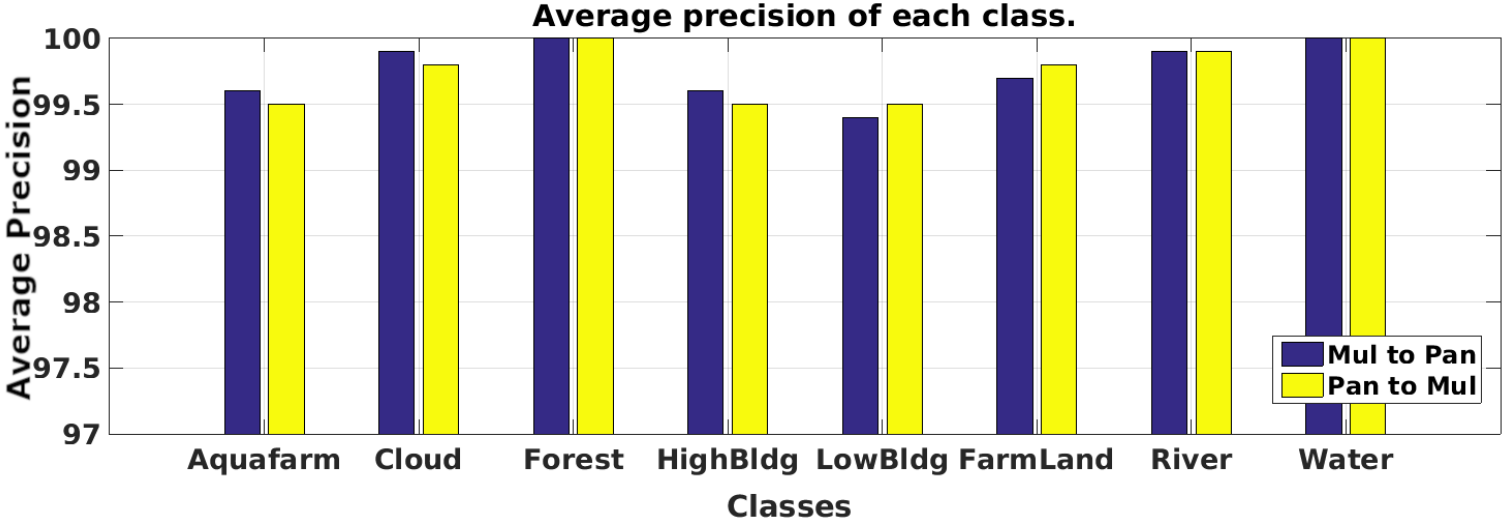


Table: Performance of the CMIR-NET framework on UC Merced-Audio dataset, with variation in embedding vector code length (d_v).

Model	d_v=32		d_v = 64		d_v = 128	
	mAP	P@10	mAP	P@10	mAP	P@10
Img→Aud	29.67	60.91	41.60	63.15	**62.11**	**64.81**
Aud→Img	21.60	40.11	42.36	51.29	**54.21**	**56.00**

Attention-aware cross-modal retrieval

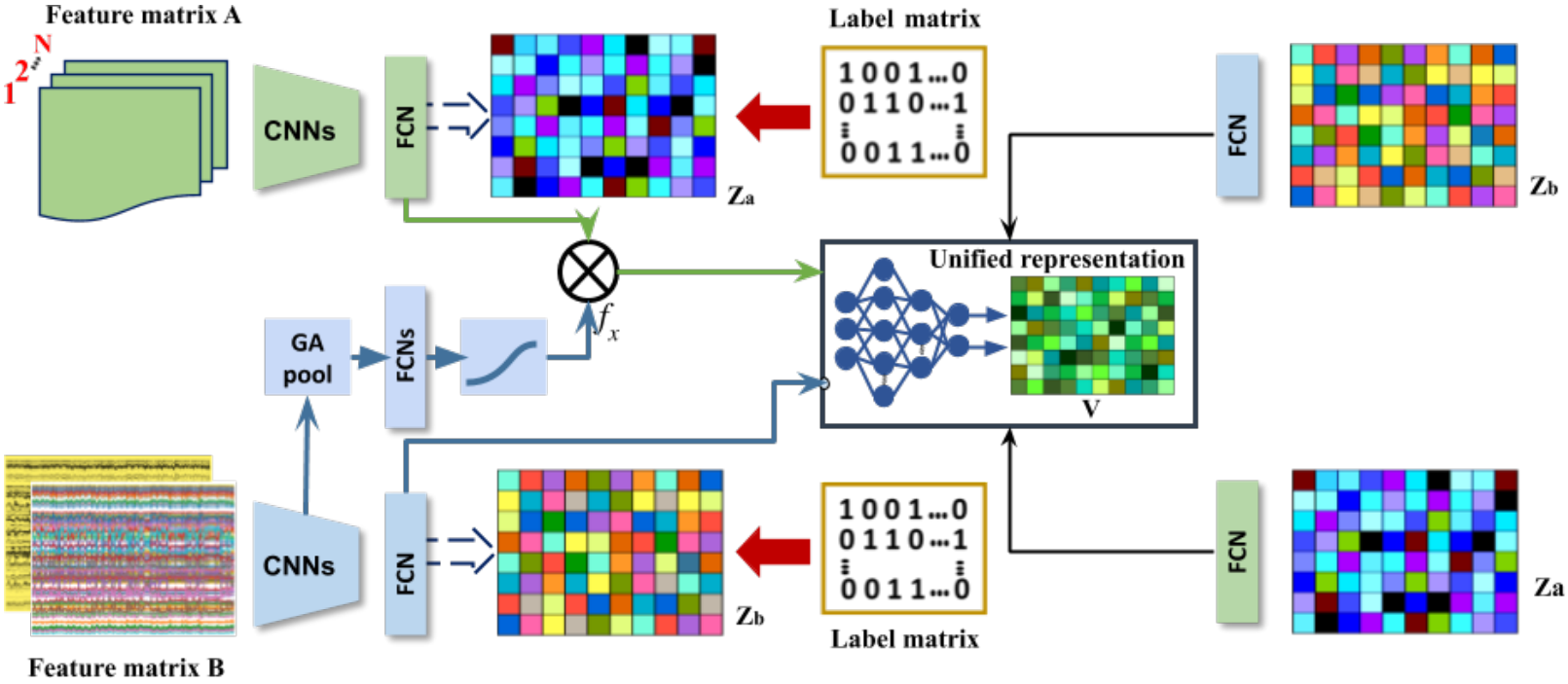
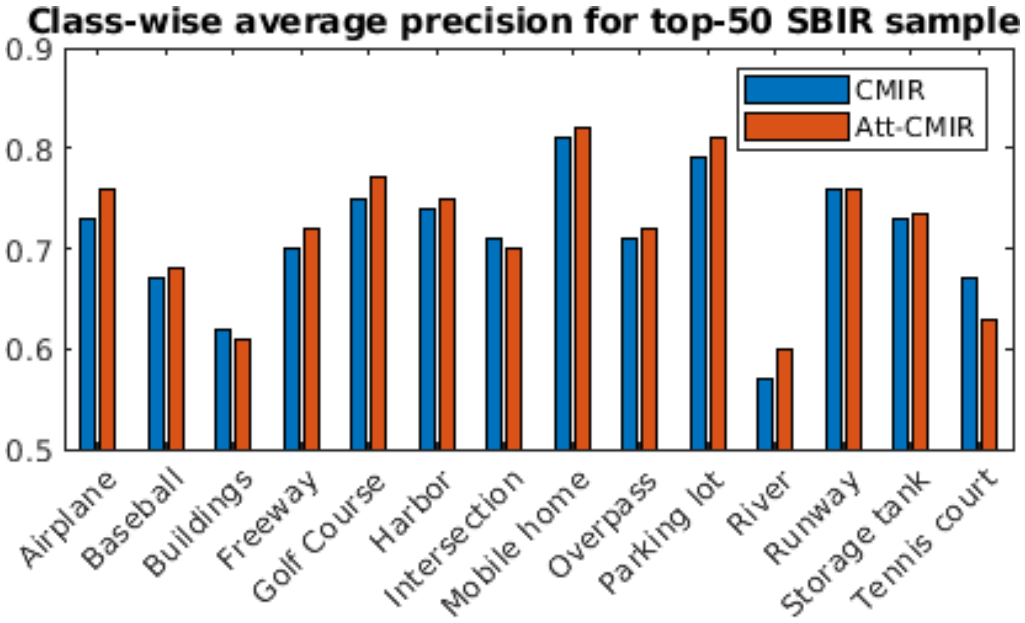
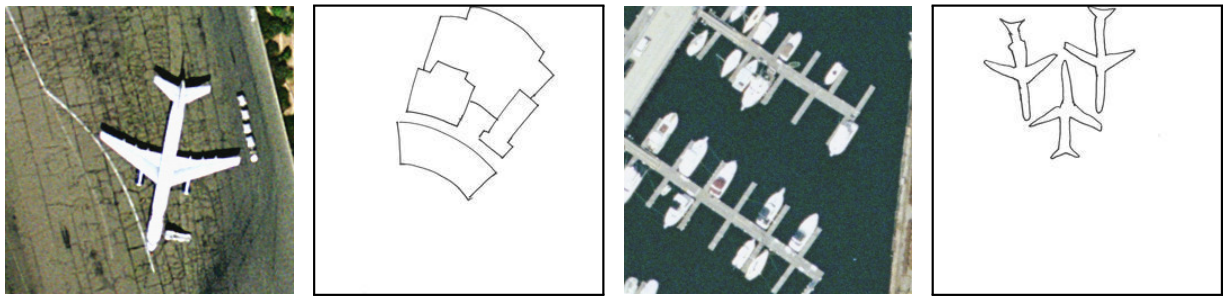


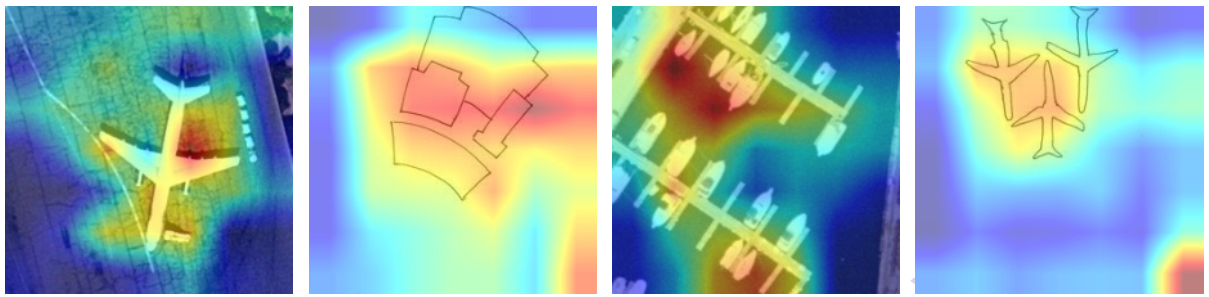
Table: Comparison with the existing literature.

Task	Model	d_v=128	
		mAP	P@10
Sketch→Photo	CMIR-Net	0.732	0.756
	Proposed	**0.753**	**0.784**
Photo→Sketch	CMIR-Net	0.696	0.708
	Proposed	**0.723**	**0.745**

Attention-aware Cross-modal Retreival







Conclusions

- Proposed a modified-triplet loss for multi-label data classification/retrieval task.
- Constructed a inter-band data retreival framework for a Sentinel-2 data archive.
- Developed an encoder-decoder based cross-modal retrieval framework, robust over various types of input data.

Related publications

1. **U. Chaudhuri**, B. Banerjee, A. Bhattacharya, M. Datcu, "Attention-driven Cross-Modal Remote Sensing Image Retrieval", In Proc. IGARSS, 2021.
2. **U. Chaudhuri**, S. Dey, M. Datcu, B. Banerjee, A. Bhattacharya, "Multi-modal Classification and Retrieval Using the Multi-labeled Sentinel-2 BigEarthNet Archive", *IEEE Journal of Selected Topics in Applied Remote Sensing* (JSTARS), 2021. (In comm.)
3. **U. Chaudhuri**, B. Banerjee, A. Bhattacharya, M. Datcu, "CMIR-NET : A deep learning based model for cross-modal retrieval in remote sensing", *Pattern Recognition Letters* (PRL), volume 131, pp 456-462, 2020.

3

Zero-shot cross-modal data retrieval

(a) Simplified ZS:SBIR

(b) **CrossAtNet:** Topology preservation + Cross-attended + Hashing ZS-SBIR

Motivation — Zero-shot

Why do we need Zero-Shot Learning?

- Necessary to determine a certain number of object classes for object recognition with high success.
- Necessary to collect as many sample images as possible for object classes.
- Exists lots of object classes that we can gather sample images.
- Also exists cases that we are not always so lucky.

This Ili pika was seen last summer in Chinas Tianshan Mountains.

3a

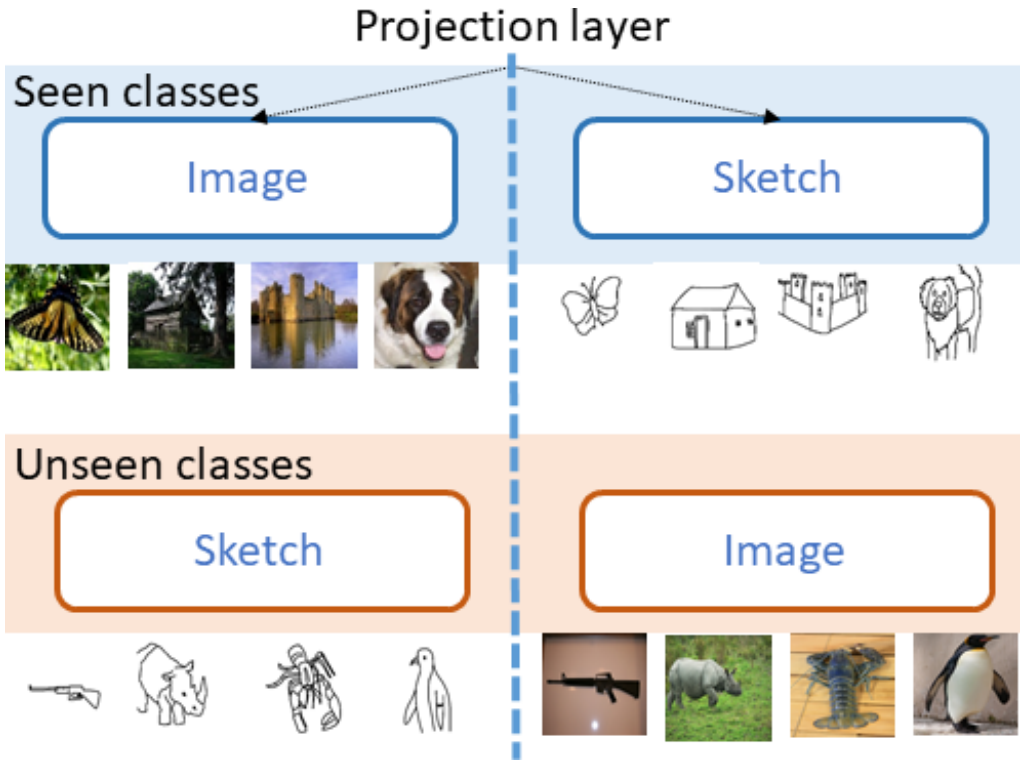
Simplified ZS-SBIR

(a) Seen-class training
(b) Unseen-class testing

Zero-shot Sketch-based Image Retrieval

Definition (Zero-shot learning)

Zero-Shot learning method aims to solve a task without receiving any example of that task at training phase.



No training samples of a few class. Only tested on these classes.

ZS-SBIR

N.b.: No samples from ZS classes to be used during training.

CNN has been seen to have **texture-bias**!! So sketches?

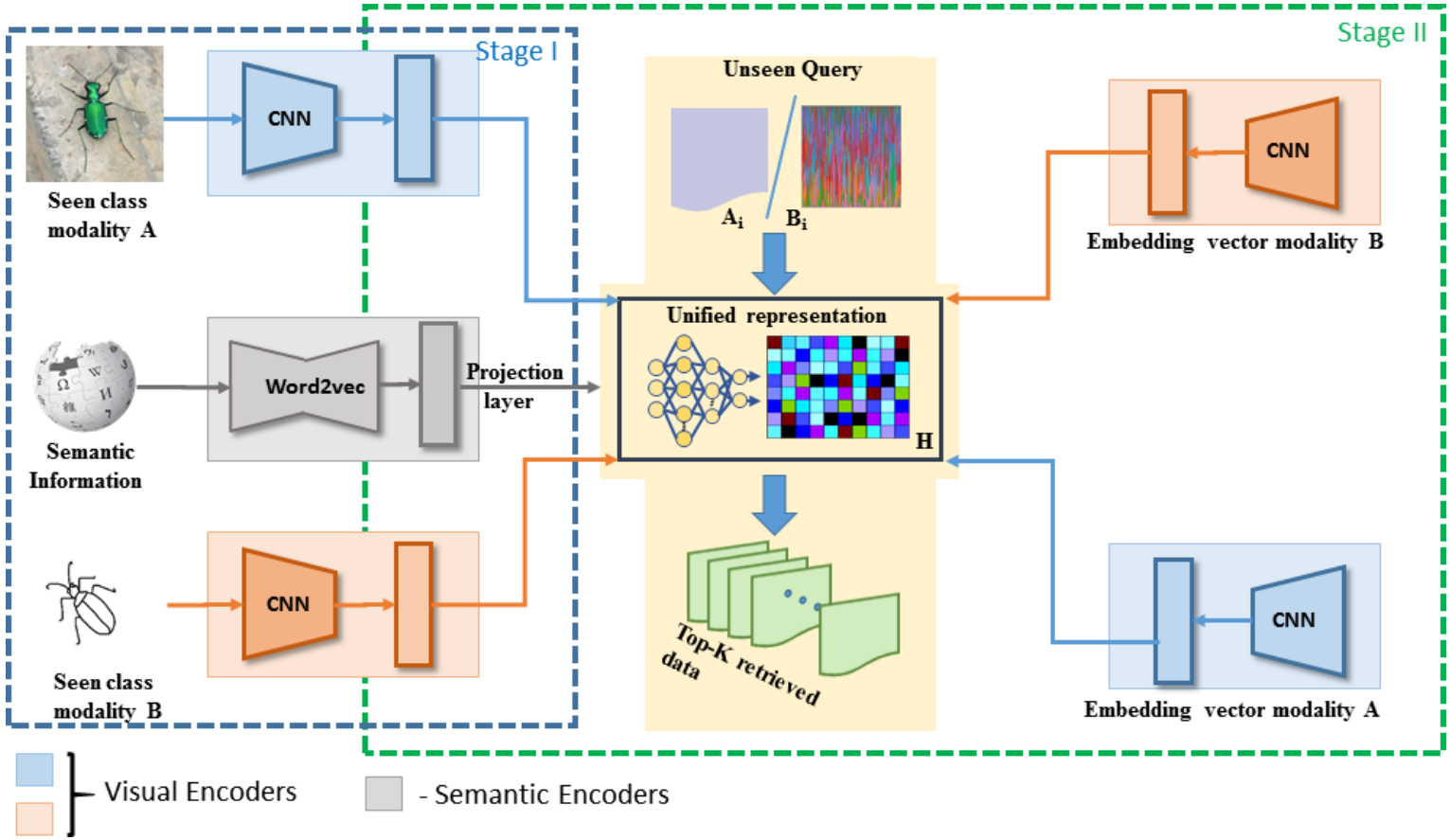
Therefore, We need 2 data representations:

1. Image
2. Class embedding as auxiliary representation

CLASSES		Image Embedding	Class Embedding
TRAINING	A	✔	✔
	B	✔	✔
	C	✔	✔
	D	✔	✔
	E	✔	✔
ZERO SHOT	F	✘	✔
	G	✘	✔
	H	✘	✔

In some experiments, class embeddings of the zero-shot classes are also not used during training.

ZS-SBIR



- Images - $\mathcal{A}$, Sketches - $\mathcal{B}$; Constraint: $\mathcal{Y}^s \cap \mathcal{Y}^u = \varnothing$
- Seen: $\mathcal{S} = \{\mathcal{A}^s, \mathcal{B}^s, \mathcal{Y}^s, \mathcal{Z}^s\}$; Unseen: $\mathcal{U} = \{\mathcal{A}^u, \mathcal{B}^u, \mathcal{Y}^u, \mathcal{Z}^u\}$.
- Modality-sp. classifiers (fine-tune Imagenet pre-trained CNN).
- Multi-stream encoder-decoder n/w for visual-semantic mapping.

ZS-SBIR

- **a) Visual Encoders:** $f_A(;,\theta_A)$ and $f_B(;,\theta_B)$ to obtain the input features corresponding to $\mathcal{A}$ and $\mathcal{B}$.
- **b) Semantic Encoders:** $f_Z(;,\theta_Z)$ for embedding the semantic information in $\mathcal{Z}$.
- **c) Cross-Modal Decoders:** $g_{AB}(;\theta_{AB})$ and $g_{BA}(;,\theta_{BA})$ which reconstructs $\mathcal{A}$ given $f_B(\mathcal{B}^s)$ and vice-versa.

★ **1. Cross-modal latent loss ($\mathcal{L}_{cmd}$):** Reduces intra-class variance. **F** represents the Frobenious norm of a matrix.

$$\mathcal{L}_{cmd} = ||f_A(\mathcal{A}_c^s) - f_Z(\mathcal{Z}_c^s)||_{\mathbf{F}}^2 + ||f_B(\mathcal{B}_c^s) - f_Z(\mathcal{Z}_c^s)||_{\mathbf{F}}^2$$

★ **2. Cross-modal triplet loss ($\mathcal{L}_{3lt}$):** Further reduce the intra-class distances, and increase the inter-class distances.

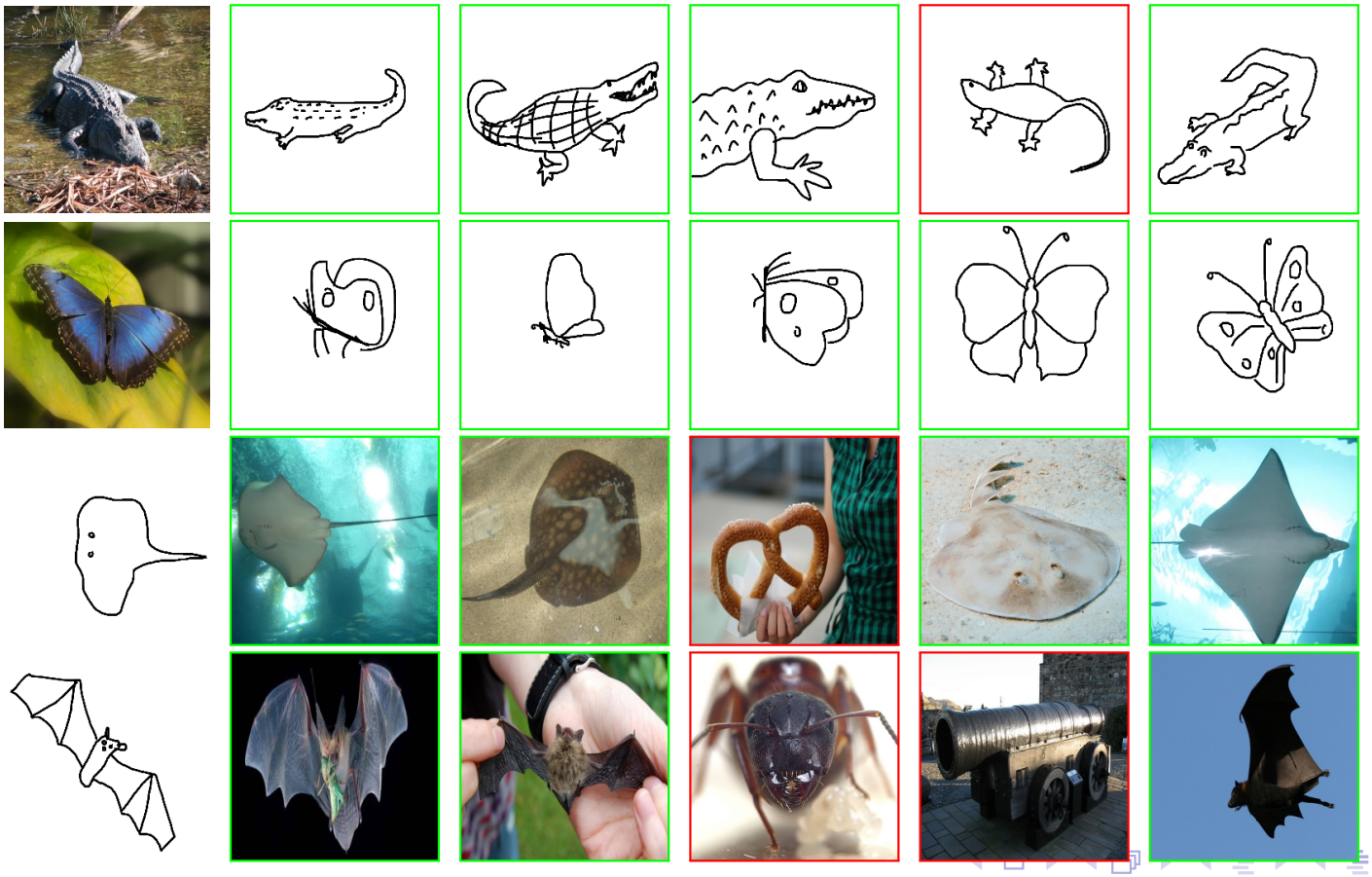
$$\mathcal{L}_{si} = \max\left(d(\mathrm{f}_A(\mathcal{A}_c^s), \mathrm{f}_B(\mathcal{B}_c^s)) - d(\mathrm{f}_A(\mathcal{A}_c^s), \mathrm{f}_B(\tilde{\mathcal{B}}_c^s)) + \alpha, 0\right)$$

★ **3. Classification loss ($\mathcal{L}_{class}$):** Class-wise discernibility.

$$\mathcal{L}_{class} = \mathrm{CE}(f_A(\mathcal{A}^s)) + \mathrm{CE}(f_B(\mathcal{B}^s))$$

Zero-shot cross-modal data retrieval

- **Sketchy:** 125 classes. Total number of photos is 12,500, while sketches is 75,471. 100:25 train:test classes split.
- **TU-Berlin:** 250 classes, with 20,000 sketches and 204,489 photos.220:30 train:test classes split.
- **Earth on Canvas*:** 14 classes $\times$ 100 samples. 10:4 train:test.



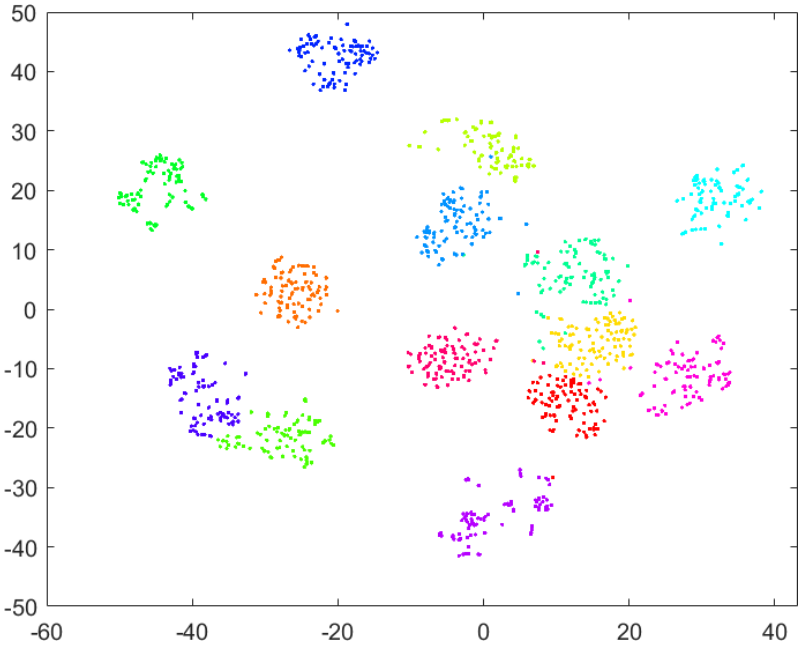
Zero-shot cross-modal data retrieval

	Task	Sketchy		TU-Berlin		size
		mAP	P@100	mAP	P@100	
SBIR	Siamese CNN	0.183	0.143	0.153	0.122	64
	SaN	0.129	0.104	0.112	0.096	512
	3D Shape	0.070	0.062	0.063	0.057	64
	DSH †	0.171	0.231	0.129	0.189	64
	GDH †	0.187	0.295	0.135	0.212	64
	GN Triplet	0.204	0.296	0.175	0.253	1024
ZSL	SSE	0.154	0.108	0.133	0.096	100
	JLSE	0.131	0.185	0.109	0.155	220
	ZSH	0.159	0.214	0.141	0.177	64
	SAE	0.216	0.293	0.167	0.221	300
ZSL:SBIR	ZS-SBIR	0.196	0.284	0.005	0.001	1024
	ZSIH †	0.258	0.342	0.223	0.294	64
	EMS	-	-	0.259	0.369	512
	CAAE	0.196	0.284	-	-	4096
	CVAE	0.225	0.333	-	-	4096
	SEM-PCYC	0.349	0.463	0.297	0.426	64
	ZSCMR	**0.467**	**0.510**	**0.362**	0.429	64

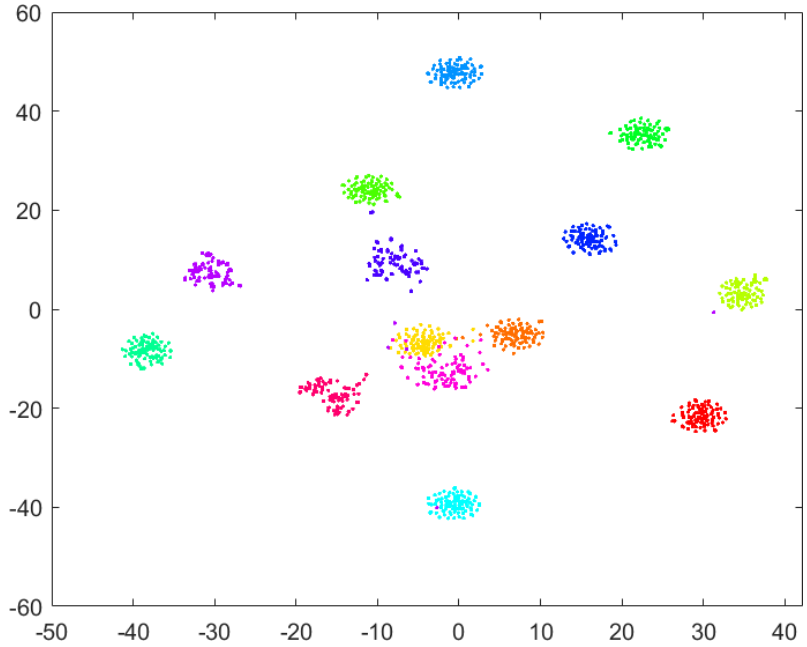
Zero-shot SBIR — Earth on Canvas

Inter-modal	EoC		Uni-modal	EoC	
	mAP	P@100		mAP	P@100
Sketch→Image	0.686	0.698	Sketch→Sketch	0.719	0.737
Image→Sketch	0.612	0.632	Image→Image	0.839	0.855

Zero-shot cross-modal data retrieval



(a) Image



(b) Sketch

Figure: Two-dimensional scatter plots of high-dimensional features generated with t-SNE of image and sketch features, in the shared latent space, trained with a fixed-semantic vector. Clusters with distinct colours denote separate classes in the dataset.

Thank You