Deep Learning Techniques for Remote Sensing Image Retrieval

Ushasi Chaudhuri

Under the guidance of
Avik Bhattacharya, Biplab Banerjee &
Mihai Datcu(external)

Center of Studies in Resources Engineering (CSRE)
IIT Bombay

February 2020

Content Based Image Retrieval:

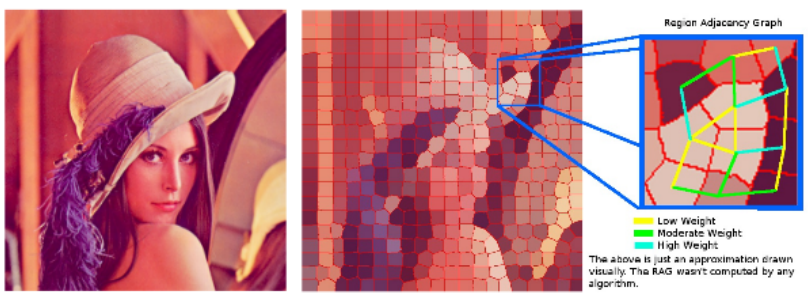
- Due to availability of wide range of satellite sensors, accumulation of an unprecedented volume of remote sensing images.
- Availability of large volumes of data, arises the necessity of sophisticated information extraction strategies.
- Image retrieval is one such popular mechanism where the aim is to retrieve a number of visually coherent images from a predefined database given a query.
- CBIR performance depends on 2 factors:
  1. **representation capability** of extracted features
  2. efficiency of **similarity measure**.
- Problem of exhaustive high dimensional feature matching:
  1. improve the **search criteria**
  2. **dimensionality reduction** in the feature space.

Highlights of first APS:

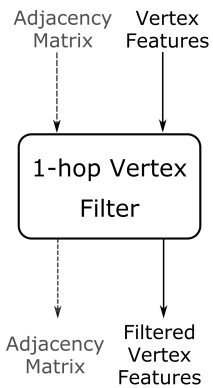
Graph convolutional network for content based remote sensing image retrieval

We plan to focus our attention on two important problems:

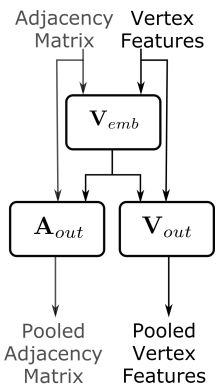
1. **global image representation**, highlighting the interactions among the **local scene** constructs
2. discriminative learning of an embedding space from an **irregular spatial distribution** of regions for the purpose of designing a CBIR system for VHR RS data.



input image regions (here superpixels) RAG



Convolution

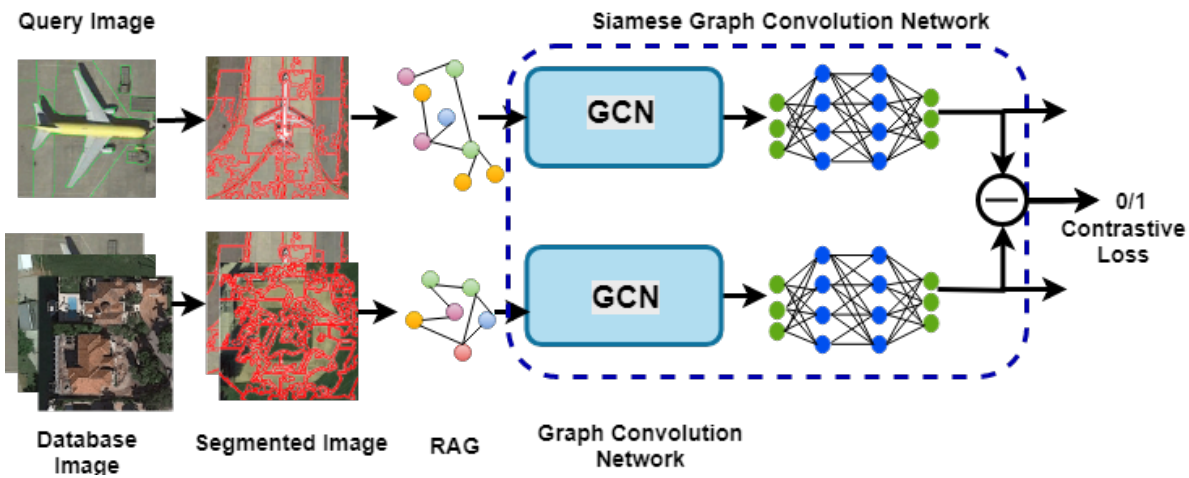


Pooling

Highlights of first APS:

1. **Siamese graph convolutional network for remote sensing CBIR**

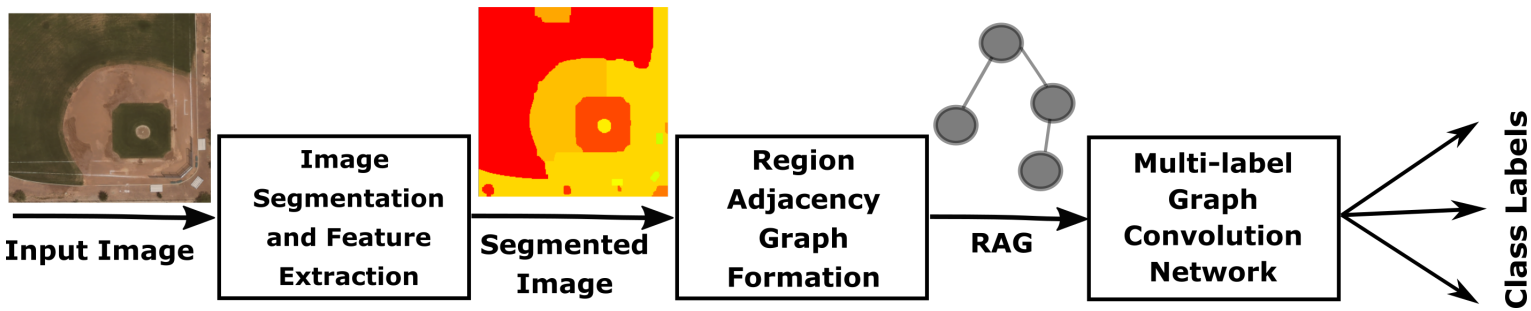
- Very high resolution (VHR) remote sensing image retrieval.
- Novel Siamese graph Convolutional network architecture which is end to end trainable.
- Improved retrieval performance using the embeddings from the Siamese graph convnet.
- Extensive experiments on the benchmark UC-Merced and large-scale PatternNet datasets.



Highlights of first APS:

2. **Graph convolutional network for multi-label VHR remote sensing scene recognition**

- Addressed the problem of multi-label scene classification.
- Novel GCN based architecture, which is end to end trainable.
- Improved classification performance using the embeddings from the GCN.
- Extensive experiments on the benchmark UC-Merced dataset.



Cross-modal retrieval- Introduction:

- 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)

[a] 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.

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

Literature Survey - TPAMI 2019

2466 IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 41, NO. 10, OCTOBER 2019

Label Consistent Matrix Factorization Hashing for Large-Scale Cross-Modal Similarity Search

Di Wang, Xinbo Gao, *Senior Member, IEEE*, Xiumei Wang, and Lihuo He, *Member, IEEE*

Abstract—Multimodal hashing has attracted much interest for cross-modal similarity search on large-scale multimedia data sets because of its efficiency and effectiveness. Recently, supervised multimodal hashing, which tries to preserve the semantic information obtained from the labels of training data, has received considerable attention for its higher search accuracy compared with unsupervised multimodal hashing. Although these algorithms are promising, they are mainly designed to preserve pairwise similarities. When semantic labels of training data are given, the algorithms often transform the labels into pairwise similarities, which gives rise to the following problems: (1) constructing pairwise similarity matrix requires enormous storage space and a large amount of calculation, making these methods unscalable to large-scale data sets; (2) transforming labels into pairwise similarities loses the category information of the training data. Therefore, these methods do not enable the hash codes to preserve the discriminative information reflected by labels and, hence, the retrieval accuracies of these methods are affected. To address these challenges, this paper introduces a simple yet effective supervised multimodal hashing method, called label consistent matrix factorization hashing (LCMFH), which focuses on directly utilizing semantic labels to guide the hashing learning procedure. Considering that relevant data from different modalities have semantic correlations, LCMFH transforms heterogeneous data into latent semantic spaces in which multimodal data from the same category share the same representation. Therefore, hash codes quantified by the obtained representations are consistent with the semantic labels of the original data and, thus, can have more discriminative power for cross-modal similarity search tasks. Thorough experiments on standard databases show that the proposed algorithm outperforms several state-of-the-art methods.

Index Terms—Hashing, multimodal, supervised, similarity search, cross-modal

1 Introduction

With the explosive growth of multimedia data, performing efficient and accurate similarity searches on large-scale data sets has become a challenging issue [1], [2]. To address large-scale similarity search problems, various hashing-based methods that, map similar data points in the images. Therefore, it is desirable to design hashing methods over multimodal data sets.

To facilitate cross-modal retrieval, many multimodal hashing methods have been proposed in recent years [15], [16]. According to whether supervised information is used,

Literature Survey

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. X, NO. X, MONTH YEAR 1

Recipe1M+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images

Javier Marín[1]*, Aritro Biswas[1]*, Ferda Ofli[2], Nicholas Hynes[1], Amaia Salvador[3], Yusuf Aytar[1], Ingmar Weber[2], Antonio Torralba[1]
[1]Massachusetts Institute of Technology [2]Qatar Computing Research Institute, HBKU
[3]Universitat Politècnica de Catalunya
{abiswas,nhynes}@mit.edu, {jmarin,yusuf,torralba}@csail.mit.edu, amaia.salvador@upc.edu, {fofli,iweber}@hbku.edu.qa

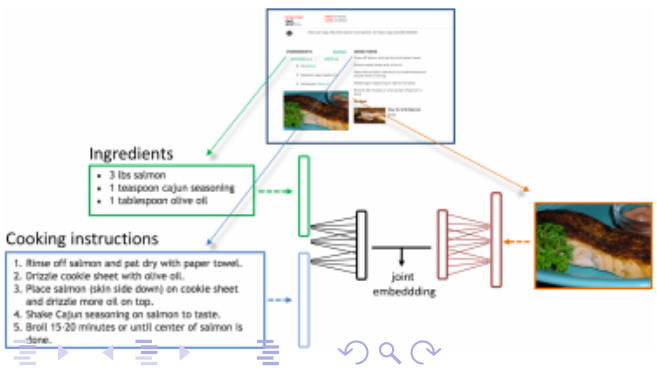
Abstract—In this paper, we introduce Recipe1M+, a new large-scale, structured corpus of over one million cooking recipes and 13 million food images. As the largest publicly available collection of recipe data, Recipe1M+ affords the ability to train high-capacity models on aligned, multimodal data. Using these data, we train a neural network to learn a joint embedding of recipes and images that yields impressive results on an image-recipe retrieval task. Moreover, we demonstrate that regularization via the addition of a high-level classification objective both improves retrieval performance to rival that of humans and enables semantic vector arithmetic. We postulate that these embeddings will provide a basis for further exploration of the Recipe1M+ dataset and food and cooking in general. Code, data and models are publicly available.

Index Terms—Cross-modal, deep learning, cooking recipes, food images

1 Introduction

THERE are few things so fundamental to the human experience as food. Its consumption is intricately linked to our health, our feelings and our culture. Even migrants starting a new life in a foreign country often hold on to their ethnic food longer than to their native language. Vital as it is to our lives, food also offers new perspectives on topical challenges in computer vision like finding representations that are robust to occlusion and deformation (as occur during ingredient processing).

The profusion of online recipe collections with user-submitted photos presents the possibility of training machines to automatically understand food preparation by jointly analyzing ingredient lists, cooking instructions and food images. Far beyond applications



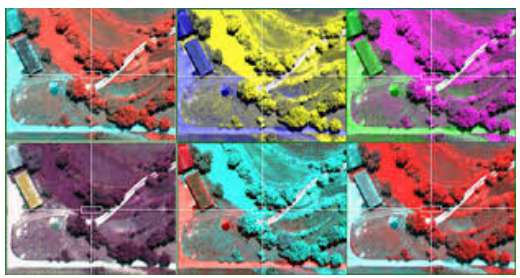
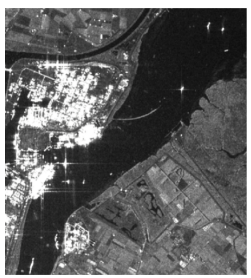
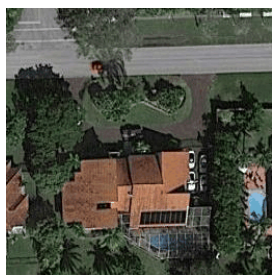
Motivation - TPAMI 2020

Why limit CBIR to just one data stream?

Advantages of each modality:

- **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 channels.

Can be an ill-posed problem.

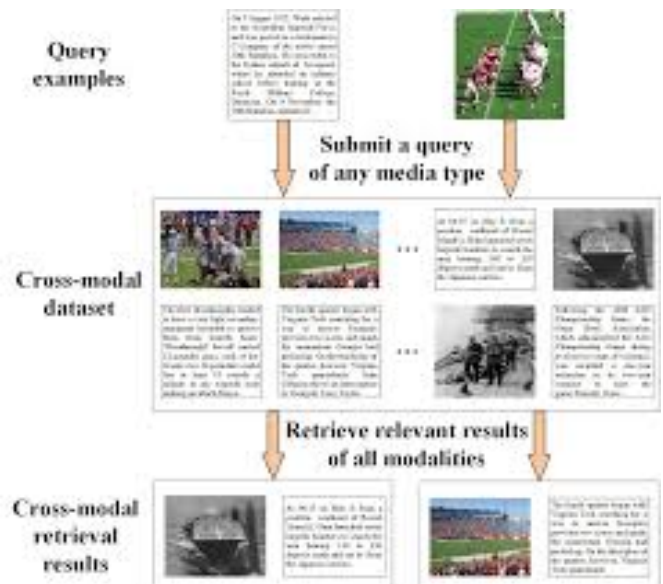
1. 1m pan IKONOS image of Valparaiso, Chile, 2. multispectral-imaging-sensors
3. Sentinel-1 SAR dataset with C band, 4. `sparseresidentialarea` from PatternNet dataset.

Motivation

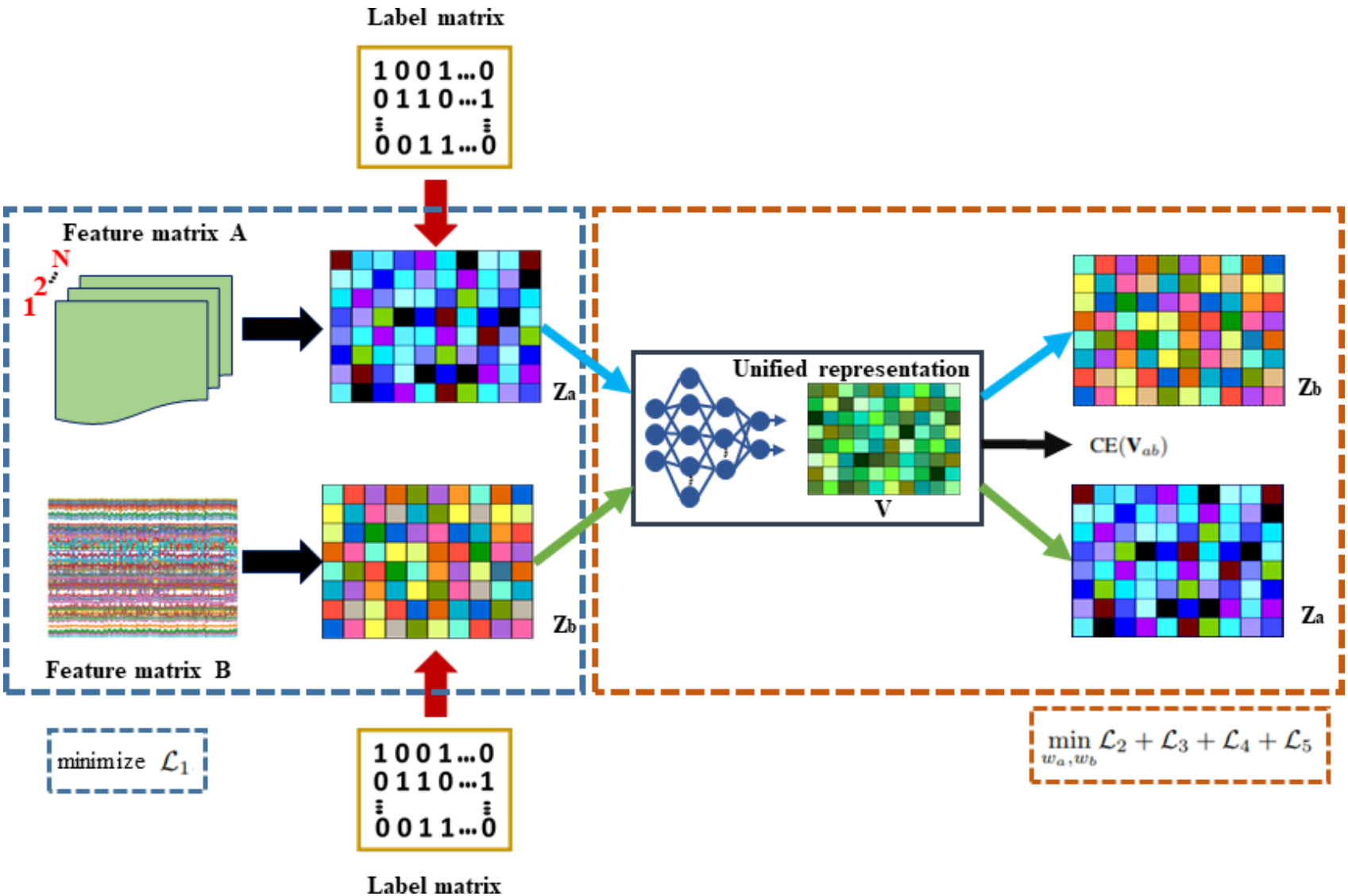
Why not try this in RS with so many available sensors?

- Important problems:
  1. Panchromatic ⇔ Multispectral.
  2. Optical ⇔ Synthetic Aperture Radar (SAR).
  3. RGB ⇔ Digital elevation maps/LiDAR point clouds.
  4. Image ⇔ text.
  5. Image ⇔ audio.

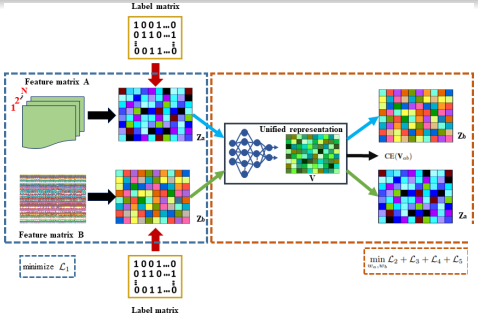
Upcoming: VideoSAR: text and audio annotation

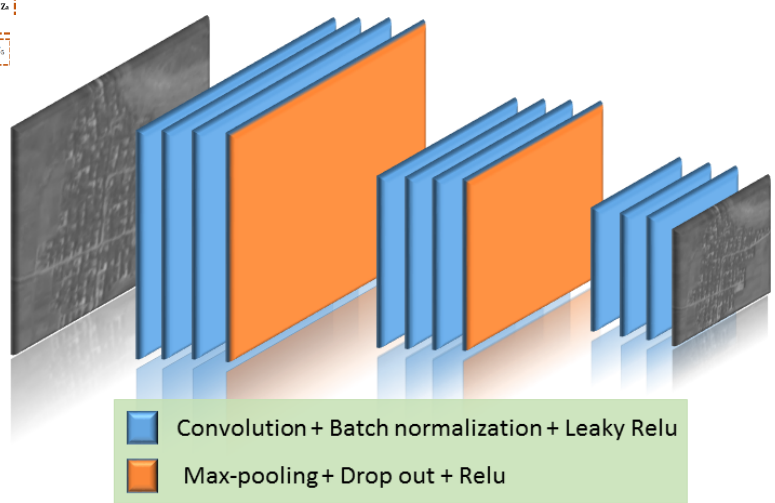


Overall Block Diagram



Pre-training



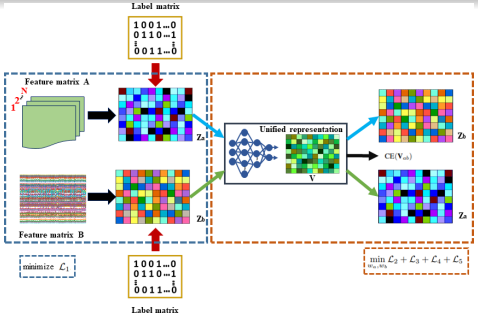


- Train 2 separate classification networks $\{(a_k, l_k)\}_{k=1}^{|\mathbf{A}|}$ and $\{(b_j, l_j)\}_{j=1}^{|\mathbf{B}|}$.
- Extracted features ($\mathbf{Z}_{a_k}$ and $\mathbf{Z}_{b_j}$) are made highly non-redundant by adding a soft orthogonality constrained.

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

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

Construction of V



To construct $\{\mathbf{V}_{a_i}\}$ and $\{\mathbf{V}_{b_i}\}$ from $\{\mathbf{Z}_{a_i}\}$ and $\{\mathbf{Z}_{b_i}\}$, we use a neural network based discriminative encoder-decoder architecture which minimizes:

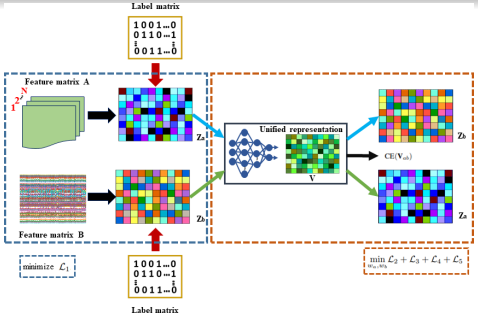
- 1. **Difference** between each pair of corresponding i^{th} samples in $\mathbf{V}_a$ and $\mathbf{V}_b$ ($\mathcal{L}_2$):

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

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

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

Construction of V



- 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: $(\mathcal{L}_5)$:

$$\mathcal{L}_5 = ||w_{ab}\mathbf{V}_a - \mathbf{Z}_b||_{\mathbf{F}}^2 + ||w_{ba}\mathbf{V}_b - \mathbf{Z}_a||_{\mathbf{F}}^2$$

where $\mathbf{Z}_{a/b} = \{Z_{a_i/b_i}\}$.

Loss Functions

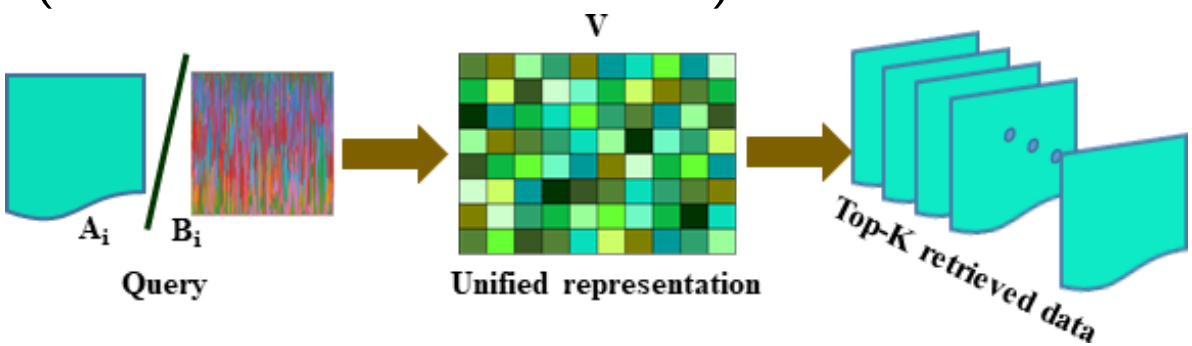
Overall equation:

$$\mathcal{L} = \lambda_1 \mathcal{L}_2 + \lambda_2 \mathcal{L}_3 + \lambda_3 \mathcal{L}_4 + \lambda_4 \mathcal{L}_5 + \lambda_5 \mathcal{R}$$

Where, given the non-negative weights λ_{1-5} and $\mathcal{R}$ defines the standard ℓ_2 regularizer on w_a and w_b.

$$\mathcal{R} = ||w_a - \alpha||_{\mathbf{F}}^2 + ||w_b - \alpha||_{\mathbf{F}}^2$$

for $\alpha \geq 0$. (**To avoid:** Trivial solution)



Algorithm

Algorithm 1 The proposed training and inference stage

Input: $\{(a_k, l_k)\}$, $\{(b_j, l_j)\}$, and $\mathbf{X}$

Output: Unified representations $\mathbf{V}_{a/b}$ ($w_a\mathbf{Z}_a$ and $w_b\mathbf{Z}_b$).

1: Normalize $\mathbf{A}$ and $\mathbf{B}$.

2: Generate intermediate representations $\{\mathbf{Z}_{a_k}\}$ and $\{\mathbf{Z}_{b_j}\}$ by minimizing $\mathcal{L}_{\mathbf{A/B}}$.

3: Train the network to obtain V by optimizing $\mathcal{L}$. The optimization follows the following stages:

4: **do**

5:

$$\min_{w_a, w_b} \lambda_1\mathcal{L}_2 + \lambda_2\mathcal{L}_3 + \lambda_3\mathcal{L}_4 + \lambda_4\mathcal{L}_5 + \lambda_5\mathcal{R} \tag{8}$$

6: **while** until convergence

7: **return** w_a and w_b (for projecting data onto V)

Input: $a \in \mathbf{A}$ or $b \in \mathbf{B}$

Output: Top-K retrieved data.

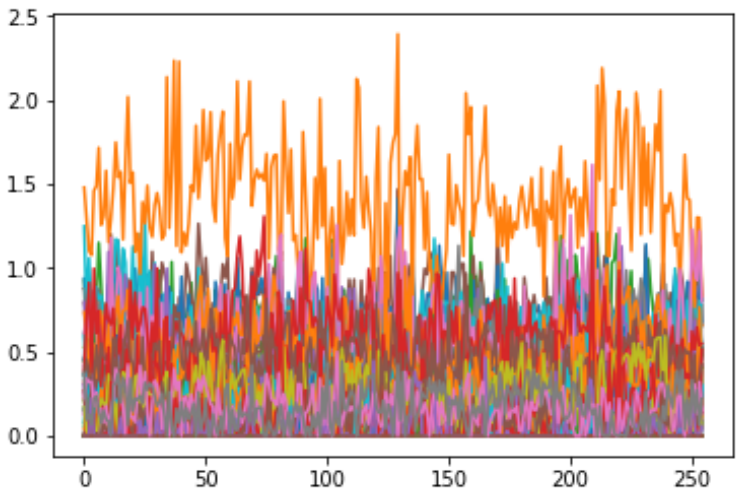
8: Uni-modal retrieval using K-NN from $w_a\mathbf{Z}_a$ **or** $w_b\mathbf{Z}_b$.

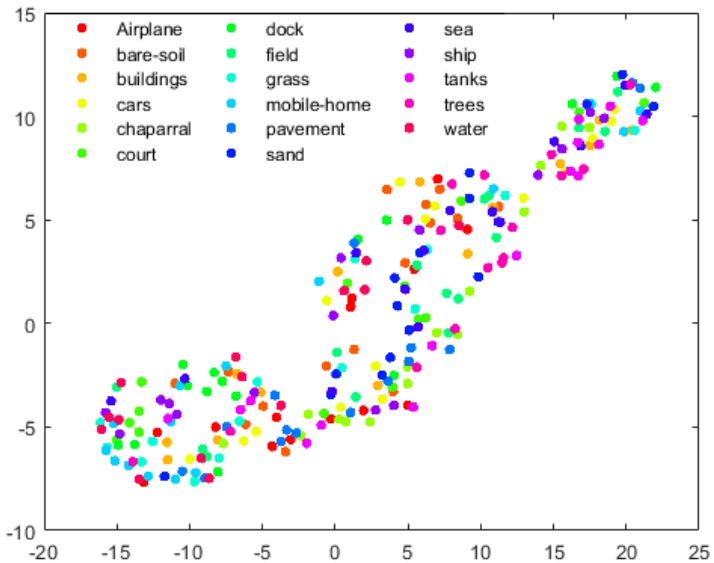
9: Cross-modal retrieval using K-NN from $w_a\mathbf{Z}_a$ **and** $w_b\mathbf{Z}_b$.

Cross-modal data

Datasets used:

1. DSRSID: Panchromatic - Multi-spectral (GF-1 satellite). **(Single-label, paired)**.
2. UCMerced: VHR RS images (RGB) - speech. **(Multi-label, un-paired)**.





Results

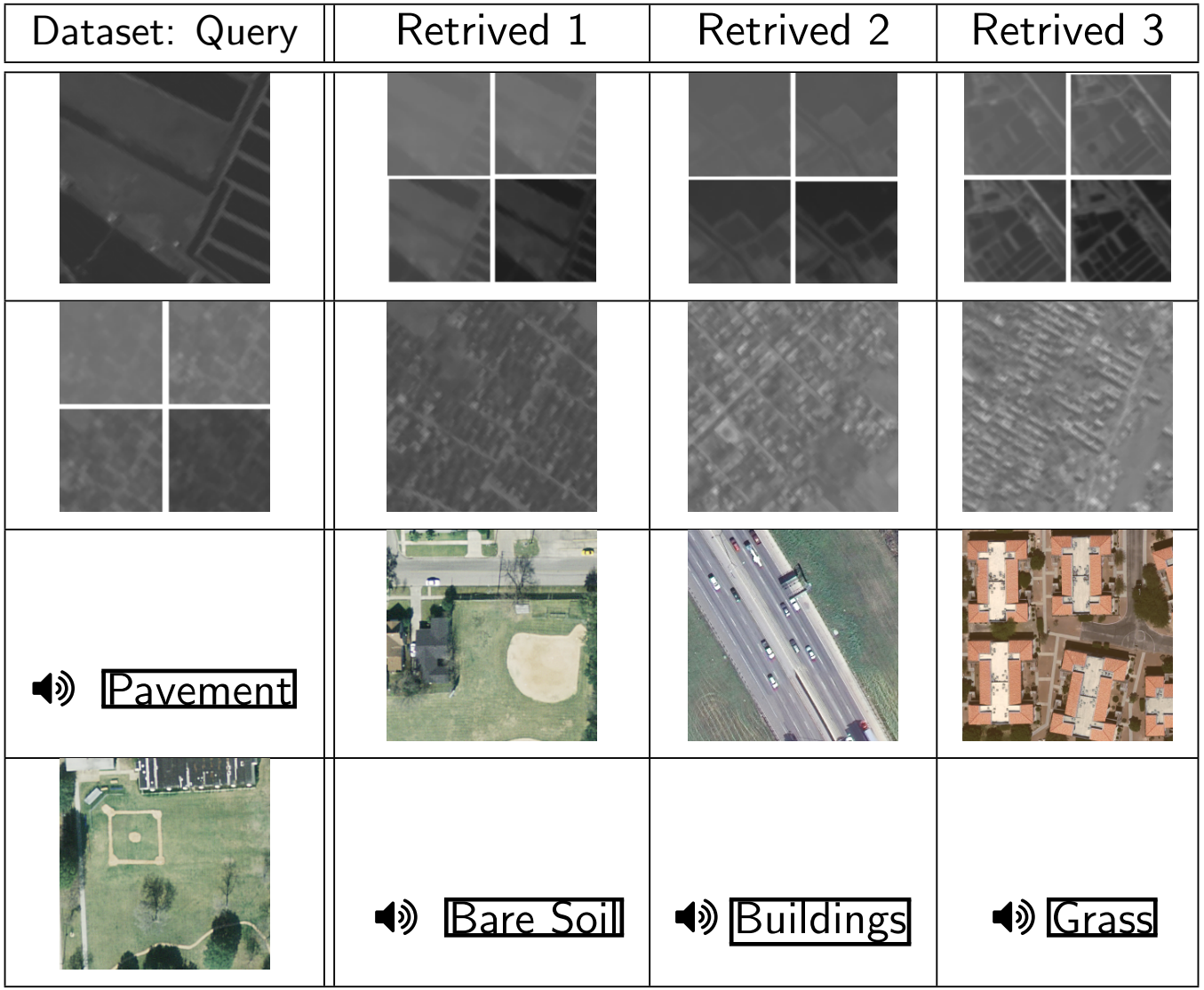


Figure: Top-3 retrieval instances from cross-modal query samples.

1. Results on DSRSID dataset

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

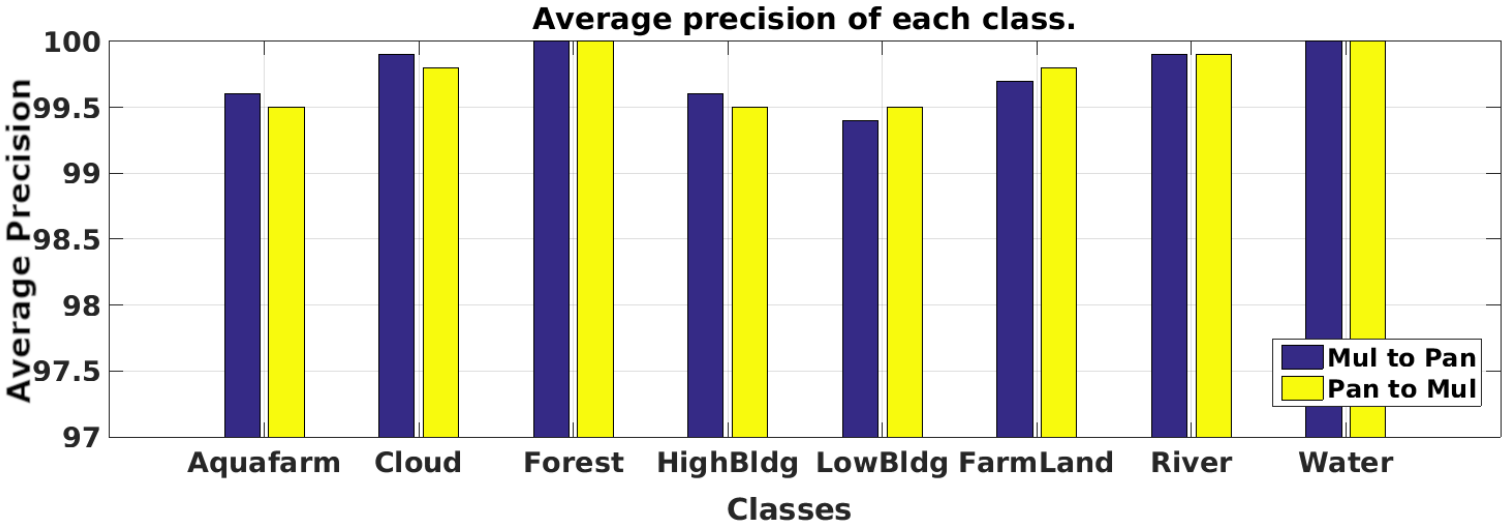


Table: Performance on the DSRSID dataset, under different embedding vector code lengths (d_v).

Task	d_v=16		d_v = 32		d_v = 64	
	mAP	P@10	mAP	P@10	mAP	P@10
Pan→Mul	95.52	97.10	98.96	98.99	99.05	99.40
Mul→Pan	98.77	99.00	97.95	97.99	98.93	98.60
Pan→Pan	99.41	99.82	98.11	98.40	98.69	99.40
Mul→Mul	99.55	99.69	98.18	98.60	98.25	98.40

1. Results on DSRSID dataset

Table: Comparison of mAP values of our proposed framework and other comparative algorithms based on the code length of the embedding space.

Task	Model	d_v=16	d_v = 32
		mAP(%)	mAP(%)
Pan→Mul	SCM [**zhang2014large**]	34.72	37.67
	DCMH [**jiang2017deep**]	80.76	85.09
	SIDHCNN [**dsrisd**]	**95.53**	96.43
	Proposed CMIR-NET	95.52	**98.96**
Mul→Pan	SCM [**zhang2014large**]	36.71	38.71
	DCMH [**jiang2017deep**]	80.23	84.45
	SIDHCNN [**dsrisd**]	97.25	97.89
	Proposed CMIR-NET	**98.77**	**97.95**

$$\text{Precision} = \frac{t_p}{t_p + f_p} \qquad \text{Precision} = \frac{\text{relevant documents} \cap \text{retrieved documents}}{\text{retrieved documnets}}$$

2. Results on UC Merced-Speech dataset

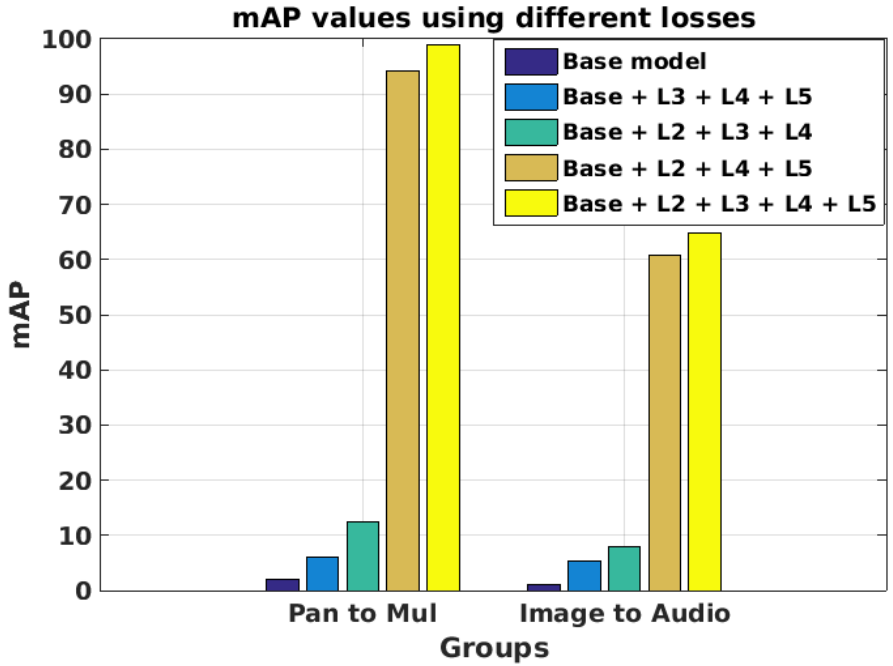
Table: Performance of the CMIR-NET framework on UC Merced-Audio dataset in terms of mAP and P@10 values, with variation in embedding vector code length (d_v).

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**

Table: Performance of the CMIR-NET framework in terms of mAP values, with variations in hyper-parameters (with $d_v = 128$).

Model	λ_1		$\lambda_3 = \lambda_4$		λ_5	
	value	mAP	value	mAP	value	mAP
Aud→Img	0.000001	49.02	0.001	53.29	0.01	23.57
	0.00001	**54.21**	0.01	**54.21**	0.1	51.84
	0.0001	52.76	0.1	53.87	1	**54.21**

Results - Ablation studies



$\mathcal{L}_2$ (difference) $\mathcal{L}_3$ (classification) $\mathcal{L}_4$ (norm) $\mathcal{L}_5$ (decoder)

$$\mathcal{R} = ||w_a - \alpha||_{\mathbf{F}}^2 + ||w_b - \alpha||_{\mathbf{F}}^2$$

Table: Sensitivity to critical parameter α for the UC-Merced↔ Audio.

Model	$\alpha = 0$		$\alpha = 1$		$\alpha = 2$	
	mAP	P@10	mAP	P@10	mAP	P@10
Img→Aud	0	0	**62.11**	**64.81**	32.09	52.77
Aud→Img	0	0	**54.21**	**56.00**	33.64	45.01

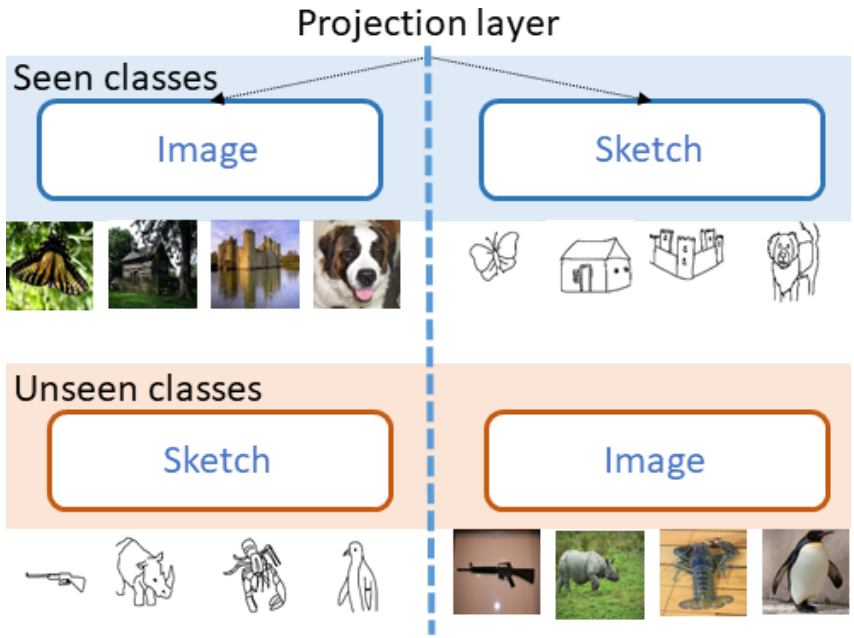
Conclusions

- We propose a novel framework for cross-modal information retrieval.
- Beats the state-of-the-art results for uni-modal retrieval, while also allowing cross-modal retrieval.
- Framework focuses on learning a unified and discriminative embedding space from different input modalities.
- Generic enough to handle both uni-modal and cross-modal retrieval.
- **Future work:** Can we try our hand on self-supervised cross-modal learning or zero-shot cross-modal IR?

2. Extension to Zeros-Shot Cross-modal 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.

Motivation

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 China's Tianshan Mountains.

Literature Survey (IGARSS 2019)

SKETCH-BASED AERIAL IMAGE RETRIEVAL

Tianbi Jiang[1], *Gui-Song Xia*[1], *Qikai Lu*[2]

[1]*State Key Lab. LIESMARS, Wuhan University, Wuhan 430079, China*
[2]*EIS, Wuhan University, Wuhan 430079, China*
{jiangtianbi, guisong.xia, qikai_lu}@whu.edu.cn,

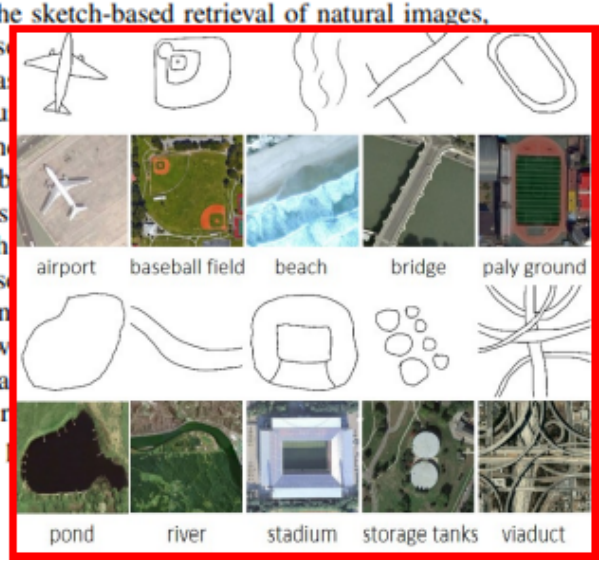
ABSTRACT

Notwithstanding aerial image retrieval is an important and obligatory task, existing retrieval systems lose their efficiency when there is no available aerial image used as the exemplar query. In this paper, we take free-hand sketches into consideration and address the problem of sketch-based aerial image retrieval. This is an extremely challenging task due to the complex surface structures and huge variations of resolutions of aerial images, and few works have been devoted to it. For the first time to our knowledge, we propose a framework to bridge the gap between sketches and aerial images. Specifically, an aerial sketch-image dataset is first collected. Sketches and aerial images are augmented to varied levels of details and used to train a multi-scale deep hierarchical model. The fully-connected layers of the deep model are used as cross-domain features, and the similarity between aerial images and sketches is measured by the Euclidean distance. Experiments on several public aerial image datasets demonstrate the efficiency and superiority of the proposed method.

Index Terms— Sketch, aerial image retrieval, multi-scale deep model

ploy modified *histogram-of-gradient* (HOG) descriptors combined with mid-level encoding to retrieve natural images. Recently, convolutional neural networks (CNNs) have also been used to accomplish sketch-based image retrieval and shown superior performance [10–12]. In a word, it has been demonstrated that cross-domain image-sketch comparison can overcome the problem of image retrieval without available query image.

Inspired by the sketch-based retrieval of natural images, this article propos
image retrieval ta
is the complex su
age resolutions an
more difficult to b
natural images us
problem of sketch
tic sketches are us
We first collect an
SI. Specifically, w
sketch-image data
and images in Aer
ing scales which



Literature Survey (CVPR 2019)

Semantically Tied Paired Cycle Consistency for Zero-Shot Sketch-based Image Retrieval

Anjan Dutta
Computer Vision Center
Autonomous University of Barcelona
adutta@cvc.uab.es

Zeynep Akata
Amsterdam Machine Learning Lab
University of Amsterdam
z.akata@uva.nl

Abstract

Zero-shot sketch-based image retrieval (SBIR) is an emerging task in computer vision, allowing to retrieve natural images relevant to sketch queries that might not been seen in the training phase. Existing works either require aligned sketch-image pairs or inefficient memory fusion layer for mapping the visual information to a semantic space. In this work, we propose a semantically aligned paired cycle-consistent generative (SEM-PCYC) model for zero-shot SBIR, where each branch maps the visual information to a common semantic space via an adversarial training. Each of these branches maintains a cycle consistency that only requires supervision at category levels, and avoids the need of highly-priced aligned sketch-image pairs. A classification criteria on the generators' outputs ensures the visual to semantic space mapping to be discriminating. Furthermore, we propose to combine textual and hierarchical side information via a feature selection auto-encoder that selects discriminating side information within a same end-to-end model. Our results demonstrate a significant boost in zero-shot SBIR performance over the state-of-the-art on the challenging Sketchy and TU-Berlin datasets.

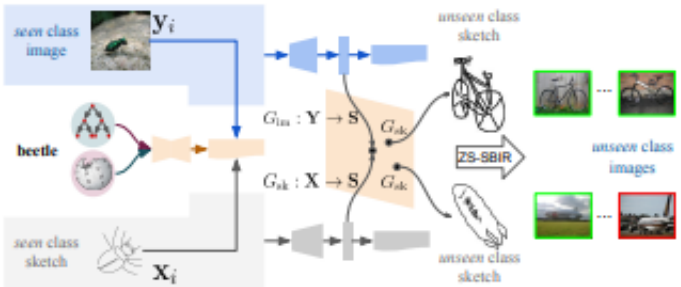


Figure 1. The proposed SEM-PCYC model learns to map visual information from sketch and image to a semantic space through an adversarial training based on the *seen* classes. During the testing phase the learned mappings are used for generating embeddings on the *unseen* classes for zero-shot SBIR.

As the visual information from all the classes gets explored by the system during training, with overlapping training and test classes, existing SBIR methods perform well [63]. Since in practice there is no guarantee that the training data would include all possible queries, a more realistic setting is *zero-shot* SBIR [43, 23] which combines zero-shot learning (ZSL) [25, 54] and SBIR as a single task, where the aim is an accurate class prediction and a competent retrieval performance. However, zero-shot SBIR is ex-

Methodology

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

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

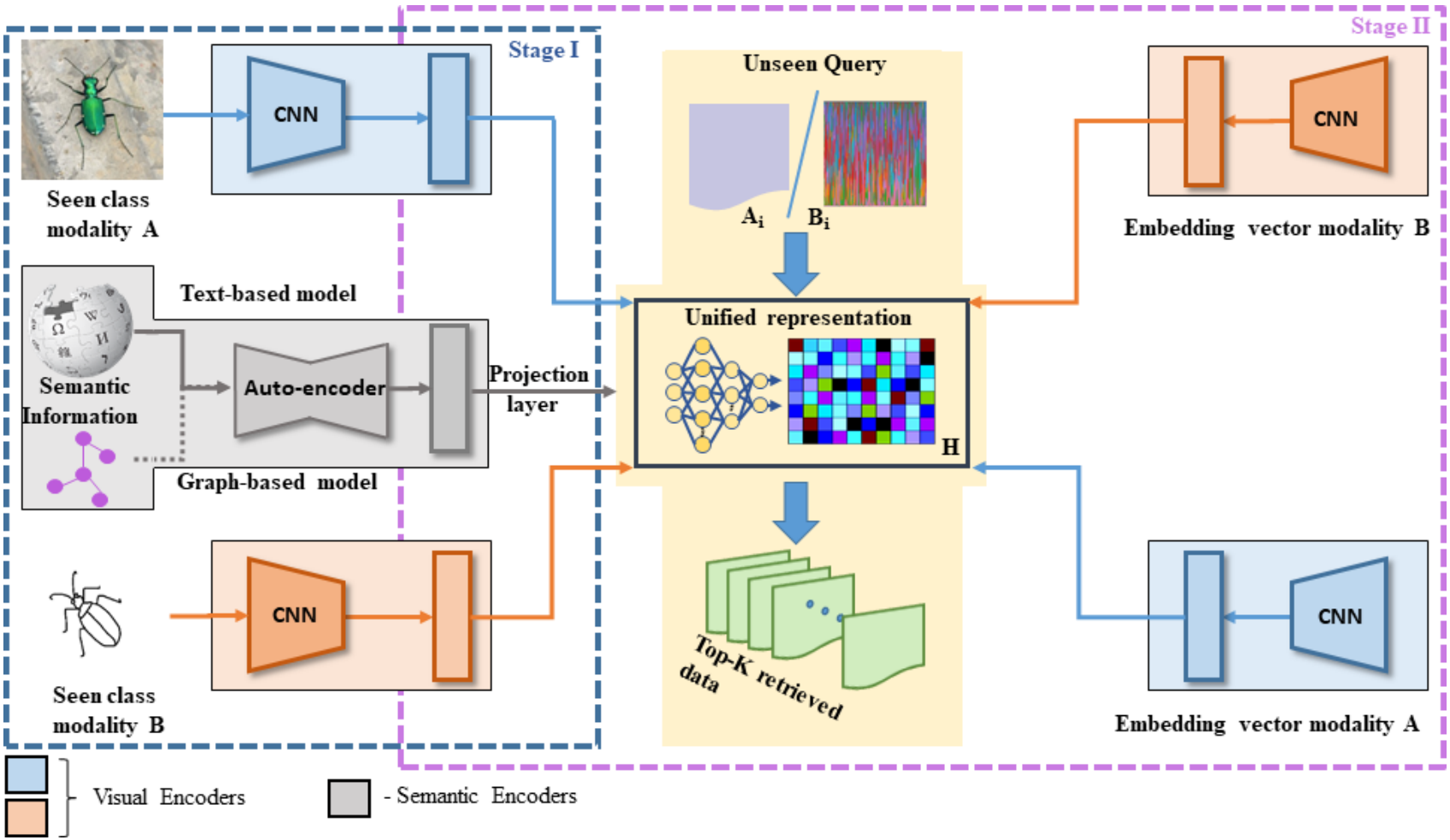
$\therefore$ We need 2 data representations:

1. Image embedding
2. Class embedding — as auxiliary representation

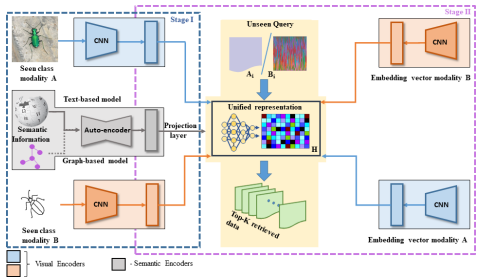
CLASSES		Image Embedding	Class Embedding
TRAINING	A	✔	✔
	B	✔	✔
	C	✔	✔
	D	✔	✔
	E	✔	✔
ZERO SHOT	F	✘	✔
	G	✘	✔
	H	✘	✔

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

Methodology



Methodology



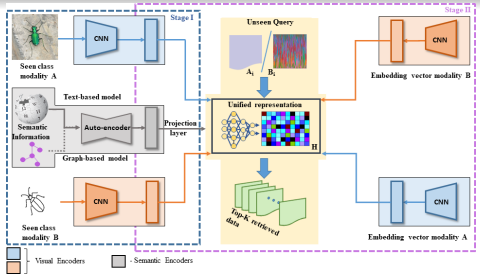
Preliminaries

- Images - $\mathcal{A}$, Sketches - $\mathcal{B}$
- Seen: $\mathcal{S} = \{\mathcal{A}^s, \mathcal{B}^s, \mathcal{Y}^s, \mathcal{Z}^s\}$
- Unseen: $\mathcal{U} = \{\mathcal{A}^u, \mathcal{B}^u, \mathcal{Y}^u, \mathcal{Z}^u\}$
- Constraint: $\mathcal{Y}^s \cap \mathcal{Y}^u = \varnothing$

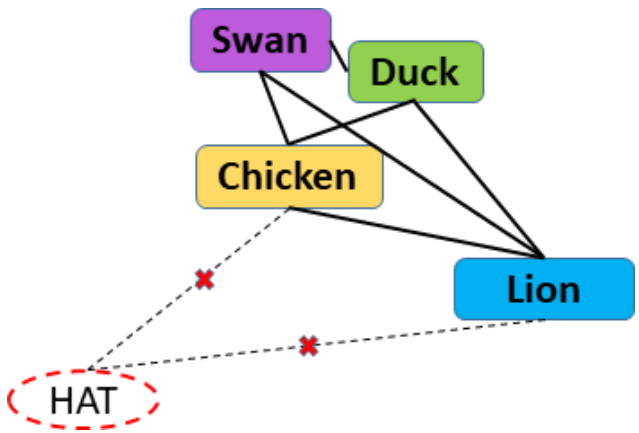
Overall ZSCMR Architecture:

1. Model the two modality-specific classifiers (fine-tune an Imagenet pre-trained CNN).
2. A multi-stream encoder-decoder based neural network for visual-semantic mapping task.

Methodology

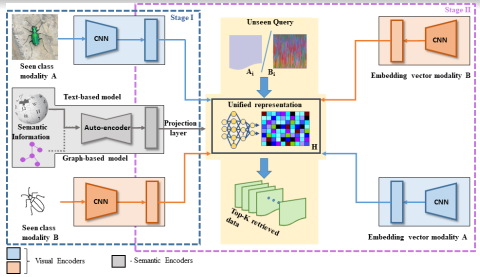


- **a) Visual Encoders:** Encoders $f_A(;,\theta_A)$ and $f_B(;,\theta_B)$ to obtain the input features corresponding to $\mathcal{A}$ and $\mathcal{B}$.
- **b) Semantic Encoders:** semantic encoder $f_Z(;,\theta_Z)$ with learnable parameter set θ_Z for embedding the semantic information in $\mathcal{Z}$.



- **c) Cross-Modal Decoders:** decoder branches $g_{AB}(;\theta_{AB})$ and $g_{BA}(;,\theta_{BA})$ which reconstructs the instances of $\mathcal{A}$ given $f_B(\mathcal{B}^s)$ and vice-versa.

Loss Functions



- **1. Cross-modal latent loss ($\mathcal{L}_{cmd}$):** Bring closer $f_A(a_i^s)$ and $f_B(b_i^s)$, by reducing their distances from z^s. This, reduces the cross-modal intra-class variance substantially.

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

F represents the Frobenious norm of a matrix.

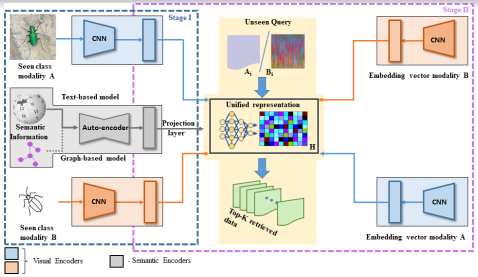
- **2. Cross-modal triplet loss ($\mathcal{L}_{3lt}$):** Further reduce the intra-class distances, and increase the inter-class distances by pushing the samples far apart in the embedding space.

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

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

$$\mathcal{L}_{3lt} = \mathcal{L}_{is} + \mathcal{L}_{si}$$

Loss Functions



- **3. Decoder loss ($\mathcal{L}_{rcs}$):**
  To enforce the domain-invariance,
  we use cross-domain sample reconstruction, which reduces distributions-gap between $\mathcal{A}$ and $\mathcal{B}$.

$$\mathcal{L}_{rcs} = ||g_{AB}(f_A(\mathcal{A}_c^s)) - f_B(\mathcal{B}_c^s)||_{\mathbf{F}}^2 + ||g_{BA}(f_B(\mathcal{B}_c^s)) - f_A(\mathcal{A}_c^s)||_{\mathbf{F}}^2$$

- **4. Classification loss ($\mathcal{L}_{class}$):** To preserve the class information in the shared space

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

Retrieval

- **Overall Objective function ($\mathcal{L}$):** linear combination of sum of the above-mentioned losses

$$\mathcal{L} = \mathcal{L}_{cmd} + \mathcal{L}_{rcs} + \mathcal{L}_{3lt} + \mathcal{L}_{class}$$

- **Cross-modal Retrieval:** Save the trained network and save the weights for unseen classes. Using their feature embeddings, find the k-nearest neighbor distance of query q from remaining samples.

$$\mathbf{D} = ||\mathbf{H}_{\mathcal{A}/\mathcal{B}^{\mathbf{u}}} - \mathbf{H}_{\mathcal{A}/\mathcal{B}^{\mathbf{u}}(\mathbf{q})}||_{\mathbf{2}}^{\mathbf{2}}$$

Idea: design such a robust shared embedding space that even a simple l_2-norm is enough for retrieval.

Algorithm:

Algorithm 1 The proposed training and inference stage

Input: $\mathcal{S} = \{\mathcal{A}^s, \mathcal{B}^s, \mathcal{Y}^s, \mathcal{Z}^s\}$

Output: Unified representations **H**.

1: **Stage 1:** Normalize and pre-train $\mathcal{A}^s$ and $\mathcal{B}^s$.

2: **Stage 2:** Find the Word2Vec embeddings of $\mathcal{Y}^s$.

3: Construct a graph from the Word2Vec embeddings.

4: Train the network to obtain **H** by optimizing $\mathcal{L}$.

5: **do**

6:

$$\min_{\theta_A,\theta_B,\theta_{AB},\theta_{BA},\theta_C} \mathcal{L}_{cmd} + \mathcal{L}_{3lt} + \mathcal{L}_{class} + \mathcal{L}_{rcs} \qquad (11)$$

7: Train Z

8: **while** until convergence

9: **return** $\theta_A, \theta_B, \theta_{AB}, \theta_{BA}, \theta_C$ (for projecting data onto **H**)

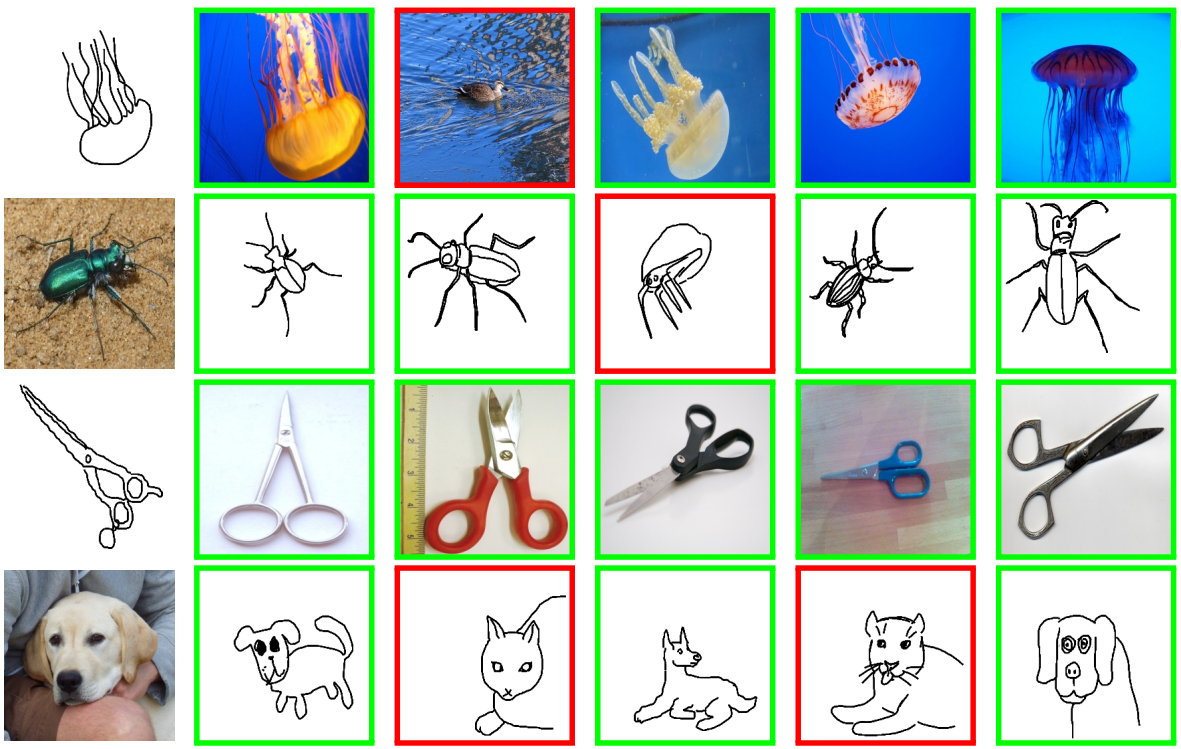
Input: $a_i^u \in \mathcal{A}^u$ or $b_i^u \in \mathcal{B}^u$ and $\mathcal{Z}^u$

Output: Top-K retrieved data.

10: Cross-modal zero-shot retrieval using k-NN.

Datasets used:

- **Sketchy dataset:** 125 classes. Total number of photos is 12,500, while sketches is 75,471. 25 classes for the testing, 100 for training.
- **TU-Berlin:** 250 classes, with 20,000 sketches and 204,489 photos. We train the network on 220 classes and test on 30 classes.



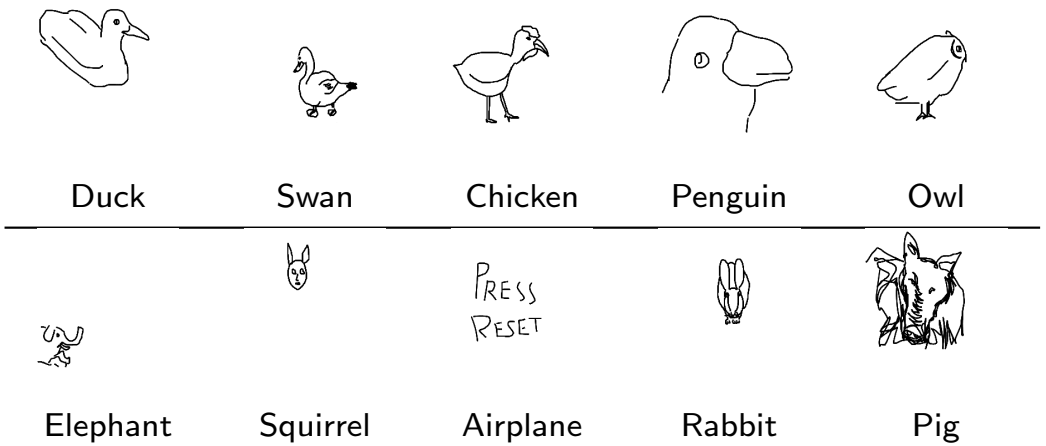
Results:

	Task	Sketchy			TU-Berlin		
		mAP	P@100	size	MAP	P@100	size
SBIR	DSH (binary) [**liu2017deep**]	0.171	0.231	64	0.129	0.189	64
	GDH (binary) [**zhang2018generative**]	0.187	0.295	64	0.135	0.212	64
	GN Triplet [**sangkloy2016sketchy**]	0.204	0.296	1024	0.175	0.253	1024
ZSL	JLSE [**zhang2016zero**]	0.131	0.185	100	0.109	0.155	220
	ZSH [**yang2016revisiting**]	0.159	0.214	64	0.141	0.177	64
	SAE [**kodirov2017semantic**]	0.216	0.293	300	0.167	0.221	300
ZSL:SBIR	ZS-SBIR [**kiran2018zero**]	0.196	0.284	1024	0.005	0.001	1024
	ZSIH (binary) [**shen2018zero**]	0.258	0.342	64	0.223	0.294	64
	SEM-PCYC [**dutta2019semantically**]	0.349	0.463	64	0.297	0.426	64
	SEM-PCYC (binary) [**dutta2019semantically**]	0.344	0.399	64	0.293	0.392	64
	ZSCMR (without Graph, unseen)	**0.421**	**0.458**	64	0.285	0.326	64
	ZSCMR (with Graph, unseen)	**0.510**	**0.535**	300	**0.334**	**0.353**	300
	ZSCMR (Image→Sketch)	**0.415**	**0.417**	300	**0.264**	**0.281**	300

Task	Sketchy		TU-Berlin	
	mAP	P@100	mAP	P@100
Sketch→Image	0.547	0.561	0.373	0.399
Image→Sketch	0.415	0.417	0.264	0.281
Sketch→Sketch	0.429	0.457	0.238	0.257
Image→Image	0.675	0.713	0.454	0.498

Results:

Row 1: Confusion classes. **Row 2:** Idiosyncratic sketches.



Ablation study with different experimental setups.

Task	Sketchy	TU-Berlin	dim
With Graph	0.547	0.373	64
Without Graph	0.467	0.352	300
With decoder loss $\mathcal{L}_{rcs}$	0.467	0.352	300
Without decoder loss $\mathcal{L}_{rcs}$	0.215	0.178	300
With triplet loss $\mathcal{L}_{3lt}$	0.467	0.352	300
Without triplet loss $\mathcal{L}_{3lt}$	0.387	0.280	300
Fixed semantic space	0.431	0.378	300
Latent semantic space	0.467	0.352	300
Pretraining with VGG-16	0.538	0.435	64
Pretraining with ResNet-50	0.560	0.441	64
Pretraining with ResNet-101	0.542	0.448	64
Seen+Unseen class graph	0.547	0.373	300
Unseen class graph	0.510	0.334	300

Conclusions:

- We propose a cross-modal zeros-shot retrieval framework and evaluate the results on sketch-based data.
- Main motive: to project different domain data onto a common embedding space, wherein discrimination between different classes within different data can be effectively possible.
- Beats the current state-of-the-art results in unseen classes SBIR.
- Also, encodes the IBSR and two uni-modal retrieval in the unseen classes.
- **Future work:** Its extension to aerial sketches, try it on multi-spectral satellites (different bands, based on spatial resolution).

Publications:

Journal:

1. U. Chaudhuri, B. Banerjee, A. Bhattacharya, **Siamese graph convolutional network for content based remote sensing image retrieval**, In *Computer Vision and Image Understanding* 184 (2019): 22-30.
2. N. Khan, U. Chaudhuri, B. Banerjee, Subhasis Chaudhuri, **Graph Convolutional Network for Multi-label VHR Remote Sensing Scene Recognition**, In *Neurocomputing* 357 (2019): 36-46.
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* 131 (2020): 456-462.
4. U. Chaudhuri, B. Banerjee, A. Bhattacharya, M. Datcu, **ZSCMR - A Novel Deep Representation Learning Technique for Zero-Shot Cross-Modal Sketch Based Image Retrieval**, *Computer Vision and Image Understanding* (in preparation).

Conference:

1. S. Rakshit, U. Chaudhuri, B. Banerjee, S. Chaudhuri, **Class Consistency Driven Unsupervised Deep Adversarial Domain Adaptation**, In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0-0. 2019.


THANK
YOU!