A Zero-Shot Sketch-Based Intermodal Object Retrieval Scheme for Remote Sensing Images

Ushasi Chaudhuri, *Member, IEEE*, Biplab Banerjee, *Member, IEEE*, Avik Bhattacharya, *Senior Member, IEEE*, and Mihai Datcu, *Fellow, IEEE*

***Abstract*— Domain-agnostic data retrieval has lately become essential amidst the availability of large-scale data from different types of sensors. However, the unavailability of a sufficient amount of samples of certain classes during training curtails the utility of existing retrieval models in remote sensing (RS) applications. Here, we propose a novel framework for zero-shot intermodal data retrieval of RS data. Thereupon, we design an encoder–decoder structure that ensures enhanced overlapping among the two data domains utilizing cross-triplet and cross-projection loss functions. Furthermore, we propose a sketch-based representation of the RS database *Earth on Canvas* with diverse classes. We perform a thorough benchmarking of this data set and demonstrate that the proposed framework outperforms state-of-the-art methods for zero-shot sketch-based retrieval framework for RS data.**

***Index Terms*— Cross-modal retrieval, database, earth on canvas (EoC), information retrieval, remote sensing (RS), sketches, zero-shot.**

I. Introduction

With the advancement in sensor technology, a vast amount of data are being collected from various satellites. Hence, the task of target-based data retrieval has become exceedingly challenging. Existing satellites typically scan a vast overlapping region of the Earth using different sensing techniques, such as multispectral, hyperspectral, synthetic aperture radar (SAR), video, and compressed sensing, to name a few. With increasing complexity and different sensing techniques at our disposal, it has become our primary interest to design efficient algorithms to retrieve data from multiple data modalities providing complementary information. This kind of problem is referred to as intermodal data retrieval.

In remote sensing (RS), there are primarily two important types of problems, i.e., land-cover classification and object detection. In this work, we focus on target-based object retrieval that can be categorized within the realm of object detection in RS. Object retrieval essentially requires high-resolution imagery for objects to be distinctly visible in the image. The main challenge with the conventional retrieval approach using large-scale databases is that, quite often, we do not have any query image sample of the target class at our disposal. The target of interest solely exists as a perception to the user in the form of an imprecise sketch. In such situations where a photo query is absent, it can be highly beneficial if we can quickly provide a handmade sketch of the target. Sketches are a highly symbolic and hieroglyphic representation of data. One can exploit the notion of this minimalistic representative of sketch queries for sketch-based image retrieval (SBIR) framework [1].

While dealing with satellite images, it is imperative to collect as many samples of images as possible for each class for object recognition with a high success rate. However, in general, there exist a considerable number of classes for which we seldom have any training data samples. Therefore, for such classes, we can use the zero-shot learning (ZSL) strategy. The ZSL approach aims to solve a task without receiving any example of that task during the training phase. This makes the network capable of handling an *unseen* class (i.e., a new class) sample obtained during the inference phase upon deployment of the network.

While SBIR has gained much attention in computer vision, this approach remains relatively unexplored in RS applications. A few notable efforts in this area include [2]–[4], where the authors have utilized the standard bench-marked vision data sets of TU-Berlin and Sketchy. Lately, the idea of SBIR has been recognized as hugely relevant in RS of very-high-resolution (VHR) satellite images using deep features [1], [5]. Both these tasks were accomplished using their proposed Aerial-SI data set. However, this data set remains unpublished for public usage. Furthermore, Xu *et al.* [6] exploited a related concept where they introduced the sketch-based RS image retrieval (SBRSIR) data set containing 20 classes, with 45 sketches and 200 images in each class. Even though the number of image samples is high, the number of sketch samples is comparatively low and inadequate for a conventional learning-based retrieval framework.

Moreover, since sketches lack texture properties, additional samples were required for relevant discriminative learning. Xu *et al.* [6] proposed an adversarial technique for SBIR using

Manuscript received August 11, 2020; revised December 16, 2020; accepted January 29, 2021. *(Corresponding author: Ushasi Chaudhuri.)*

Ushasi Chaudhuri, Biplab Banerjee, and Avik Bhattacharya are with the Centre of Studies in Resources Engineering (CSRE), IIT Bombay, Mumbai 400076, India (e-mail: ushasi2cool@gmail.com; getbiplab@gmail.com; avikb@csre.iitb.ac.in).

Mihai Datcu is with the German Aerospace Center (DLR), 82234 Wessling, Germany (e-mail: mihai.datcu@dlr.de).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/LGRS.2021.3056392.

Digital Object Identifier 10.1109/LGRS.2021.3056392

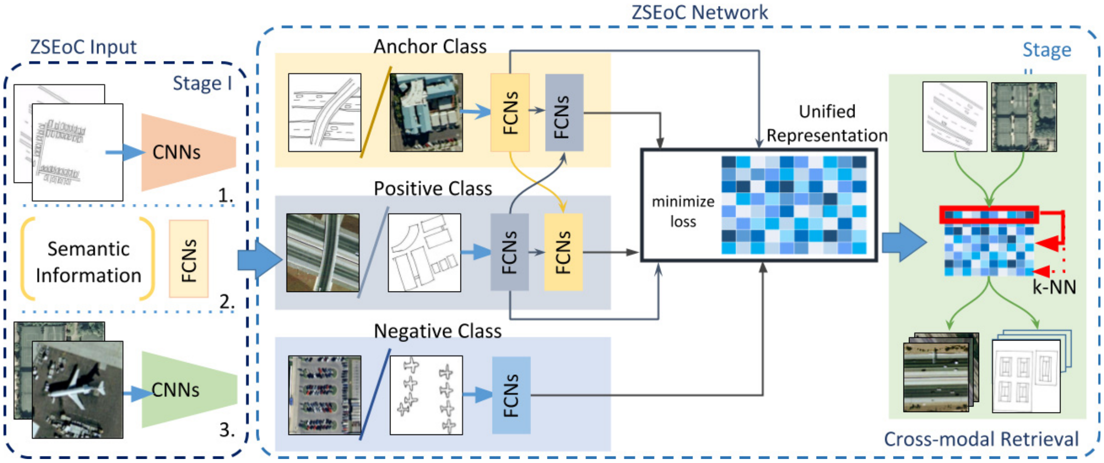

Fig. 1. Complete pipeline of the ZSEoC framework for cross-modal retrieval of image → sketch or sketch → image.

a Siamese metric learning technique. Adversarial techniques often lead to unstable training if the min–max problem is not intuitively designed. Also, they do not preserve the reverse embeddings of the image-based sketch retrieval (IBSR) aspect for making the framework intermodally retrievable. The framework becomes intermodally retrievable if we can perform both SBIR and IBSR. Although cross-modal retrieval in RS seems to be only partially explored [7], the ZSL retrieval strategy for cross-modal sketch-image data remains practically unexplored, particularly in RS applications.

In this work, we propose an efficient algorithm that performs a sketch-based intermodal retrieval from RS images (see Fig. 1). In this aspect, to the best of our knowledge, all the works in the literature only exploit the concept of sketch-based retrieval technique in RS. However, our proposed model is solely the one that introduces an intermodal retrievable aspect associated with it. Furthermore, we also extend it for a ZSL-based framework, which is a novel contribution in the field of RS. Even though we have proposed a framework for SBIR, the model is also robust to IBSR. We have carried all the experiments on an original sketch and high-resolution image bimodal data set called Earth on Canvas (EoCs), as proposed in this study. In Section II, we demonstrate the Zero-Shot retrieval architecture on the EoC data set (ZSEoC).[1]

II. Methodology

In the ZSL framework, we keep distinct classes in the training and the testing phase. The trained classes are referred to as the *seen classes*, while the classes that the network encounters only during the testing phase are referred to as the *unseen classes*. For the unseen classes, the ZSL frameworks require attribute information, which assists in recognizing the unseen classes during the testing phase. We refer to this attribute information as the *semantic information* in the remainder of the letter.

Let us denote two streams of incoming data fields for sketches and images, denoted as $\mathcal{S}$ and $\mathcal{I}$, respectively. We aim to achieve an intermodal retrieval model from $\mathcal{S}/\mathcal{I}$, given a query image from a different modality $\mathcal{I}/\mathcal{S}$. For the ZSL framework, we divide our data set into seen and unseen class images for training and testing sets, respectively. If we use L to define the set of labels and $\mathcal{W}$ as the semantic prototype, we can denote the training data as $\{\mathcal{S}^{\text{tr}}, \mathcal{I}^{\text{tr}}, L^{\text{tr}}, \mathcal{W}^{\text{tr}}\}$ and the test data as $\{\mathcal{S}^{\text{ts}}, \mathcal{I}^{\text{ts}}, L^{\text{ts}}, \mathcal{W}^{\text{ts}}\}$. Furthermore, we strictly ensure that the overlap between the seen and the unseen data is a null set ($L^{\text{tr}} \cap L^{\text{ts}} = \varnothing$). With this structure, we design a unified latent feature representation for data from both the modalities. This approach allows us to achieve a zero-shot-based intermodal retrieval framework, using the knowledge of semantic information. The workflow of our proposed architecture is shown in Fig. 1.

A. Network Construction

We organize the network using a two-stage training strategy. In the first stage, we use transfer learning from a network pretrained on the Image-Net data set [8]. In the second stage, we design an encoder–decoder-based architecture. Here, for the visual component, we use two separate encoders for the image and sketch data. In the semantic component, we encode the attribute information. We use these representations to carry out visual-to-semantic mapping for the ZSL part.

The visual encoders are a series of fully connected networks (FCNs). Using feature embeddings from the pretrained network, we stack four FCNs with 1024, 512, 256, and 128 nodes. In the semantic encoder part, we use a word-vector embedding for preserving the semantic topology. For example, runways and highways are more similar to each other than they are to parking lots, or a mobile home park is more similar to a building than it is to a baseball court. To accommodate these aspects, we use the standard `word2vec` encoding, which provides a 300-d vector embedding of the semantic class labels [9]. We use two variants of the network: 1) we use the 300-d vector directly for training and 2) we use two layers of FCNs to learn a 128-d distinct vector for semantic information.

In the encoder part of the visual streams, we design two variants of the network by taking two separate FCN networks for extracting 300-d and 128-d feature vectors in the shared latent space representation. Both these networks have similar architecture. We have used four layers of FCNs with dimensions 1024, 512, and 256, and eventually, 300 or 128 depending on the variant of the network. We have used batch-normalization

[1] More details are provided in the supplementary file.

and a leaky rectified linear unit (ReLU) function to induce nonlinearity. The network that learns the 300-d shared features is fed along with the 300-d `word2vec` embeddings directly. We refer to this model as the fixed semantic vector variant. The network that learns 128-d feature vectors from the visual encoders uses a layer of FCN to project the 300-d semantic vector onto a 128-d feature space. Since semantic information is learned from the network, we refer to this model as the latent semantic vector variant.

B. Objective Function

The overall objective function used in the proposed architecture is the sum of the following four loss functions described in the following.

1) Cross-Triplet Loss ($\mathcal{L}_{iii}$): In the first stage, we employ three branches for data input streams, which we use to create the cross-triplets. Here, we use two types of triplets [10]: 1) we chose an anchor from the image data set of class c, and we select positive and negative samples from sketch data of class c and any class other than c (essentially, c'), respectively and 2) we chose an anchor from the sketch data of class c, while we select positive and negative samples from image data of class c and c', respectively. This procedure displaces the negative class instances away from the anchor class by at least a margin α while making the same class instances of both the modalities closer. Effectively, it aids in decreasing the intermodal distance while increasing the interclass separability (2)

$$\mathcal{L}_{3a} = \max\left(d\left(w_s\mathcal{S}_c^{tr}, w_i\mathcal{I}_c^{tr}\right) - d\left(w_s\mathcal{S}_c^{tr}, w_i\mathcal{I}_{c'}^{tr}\right) + \alpha, 0\right) \quad (1)$$

$$\mathcal{L}_{3b} = \max\left(d\left(w_i\mathcal{I}_c^{tr}, w_s\mathcal{S}_c^{tr}\right) - d\left(w_i\mathcal{I}_c^{tr}, w_s\mathcal{S}_{c'}^{tr}\right) + \alpha, 0\right) \quad (2)$$

where α is a heuristically chosen margin value to push apart nonsimilar classes in the feature space. The sketch-anchored loss, $\mathcal{L}_{3a}$, and the image-anchored loss, $\mathcal{L}_{3b}$, together constitute the total cross-triplet loss, $\mathcal{L}_{iii}$. Here, $w_s\mathcal{S} = V_s$, and $w_i\mathcal{I} = V_i$, where V_s and V_i denote shared-space feature embeddings of the sketch and image instances, respectively. Even though we project the features onto a unified space, the shared features from both the modalities are different as each input is distinct. Hence, V_s and V_i denote the desired trained features of their corresponding modalities.

2) Cross-Sample Decoder Loss ($\mathcal{L}_{dl}$): The purpose of using this loss function is to make the unified latent space domain-independent. To achieve this, we bring the shared features of the two modalities closer to each other in the embedding space by performing an intermodal data instance reconstruction. This loss helps in better classwise alignment of both the modalities of data as follows:

$$\mathcal{L}_{dl} = \left\|w_i^d V_s - w_i\mathcal{I}_c^{\text{tr}}\right\|_{\text{F}}^2 + \left\|w_s^d V_i - w_s\mathcal{S}_c^{\text{tr}}\right\|_{\text{F}}^2 \quad (3)$$

where w_s^d and w_i^d are the learnable parameters for the cross-sample decoder network. Here, we want $w_i^d V_s$ to learn its corresponding class feature encoding from $\mathcal{S}$ and similarly $w_s^d V_i$ from $\mathcal{I}$ (last two FCN blocks in gray and yellow in Fig. 1).

3) Cross-Projection Loss ($\mathcal{L}_{\text{cpl}}$): To make the data projections from both the modalities closer in the feature space, we minimize the mean-square difference between these two representations and the semantic information in the embedding space. This approach offers representations from both modalities akin to the semantic projection while bringing them closer to each other as follows:

$$\mathcal{L}_{\text{cpl}} = \left\|w_i\mathcal{I}_c^{\text{tr}} - \mathcal{W}_c^{\text{tr}}\right\|_{\text{F}}^2 + \left\|w_s\mathcal{S}_c^{\text{tr}} - \mathcal{W}_c^{\text{tr}}\right\|_{\text{F}}^2. \quad (4)$$

4) Cross-Entropy Loss ($\mathcal{L}_{ce}$): An essential requirement while designing our model is to preserve the semantic class labels in the shared semantic space. For this purpose, we use the cross-entropy (CE) loss function for both the modalities to retain the semantic label information in the shared latent space as follows:

$$\mathcal{L}_{ce} = \text{CE}\left(w_s\mathcal{S}^{\text{tr}}\right) + \text{CE}\left(w_i\mathcal{I}^{\text{tr}}\right) \quad (5)$$

where w_s and w_i are the learnable parameters for creating the unified features.

5) Overall Objective Function ($\mathcal{L}$): The final objective function is an aggregate of all the losses $\mathcal{L}_{\text{total}} = \mathcal{L}_{iii} + \mathcal{L}_{dl} + \mathcal{L}_{\text{cpl}} + \mathcal{L}_{ce}$. For training this network, we perform an optimization on the final loss ($\mathcal{L}_{\text{total}}$). However, since we have a nonconvex optimization problem at hand, we perform gradient descent on each of these losses individually while holding others constant. We solve the optimization problem by alternately minimizing each loss functions individually using the minibatch gradient descent optimizer. Once we obtain the intermodal embedding of each of the data sample, we can provide a query sample from either modality $\mathcal{S}/\mathcal{I}$ and find the k-nearest neighbor (k-NN) instances from either of the modalities.

C. EoCs Data Set

We created the EoCs data set by utilizing a subset of image classes from the standard UC-Merced data set [11]. The sketches were hand-drawn on the letter by several amateur artists to avoid style-bias and were then photoscanned with a resolution of 300 dpi. To set more attention to the sketch object from the sizeable background and increase its salience in the image, we cropped along the orthogonal hull of each sketch. We pad the remaining portion with a white background to produce a size of 256 × 256 pixel dimension.

Out of the total 21 classes present in the Merced data set, we dropped four land-cover classes (namely, agricultural, beach, chaparral, and forest), as we aim to solve an object retrieval problem. Furthermore, we use a common class to represent the dense-residential, medium-residential, sparse-residential, and mobile-home park classes that describe the residential areas, as they have a similar visual appearance. Hence, the VHR-image data consists of 14 classes for which we have 100 image samples for each class (i.e., a total of 1400 optical images). We created 100 sketches for each of these 14 classes (i.e., a total of 1400 sketch images). Therefore, we have a total of 2800 images in the database containing both optical and sketch images sharing the same category labels. The classes in the data set are Airplane, Baseball diamond, Buildings, Freeway, Golf course, Harbor, Intersection, Mobile home park, Overpass, Parking lot, River, Runway, Storage tanks, and Tennis court.

TABLE I

SBIR Performance of the Proposed ZSEoC Framework on the EoC Data set in Terms of Mean Average Precision (mAP) (%) and Precision at Top-100 (P at 100) (%) Values

Task	**EoC**		
	mAP	**P@100**	**Feature dimension**
Baseline-I (VggNet-16)	0.221	0.234	4096
Baseline-II (ResNet-50)	0.236	0.254	2048
Baseline-III (ResNet-101)	0.269	0.284	2048
Baseline-IV (CNN)	0.30	0.284	128
Baseline-V (Pre-train + CNN)	0.196	0.284	128
ZS-SBIR [2]	0.395	0.421	1024
ZSIH (binary) [3]	0.452	0.487	64
ZSEoC-300 (fixed semantic vector)	**0.686**	**0.698**	300
ZSEoC-128 (latent semantic vector)	**0.674**	**0.732**	128

III. Experiments and Results

We perform all our experiments on the proposed EoC data set. In the ZSL experimental framework, we consider ten classes for training (i.e., seen classes) and four classes for testing (i.e., unseen classes). We use the last four classes, i.e., Runway, Water-tank, Tennis-court, and River classes as the unseen classes. For pretraining the network, we explore several standard models, namely, VggNet16 [12], ResNet50, and ResNet101 [13] (results reported in Table I). By using transfer learning from these pretrained networks, we perform fine-tuning on our data set to encode the class labels in the extracted feature space.

We created 14 000 triplets for each type of anchor while training the network. In each batch, we made sure that we provide an equal number of sketch-anchored triplets and image-anchored triplets to avoid any training bias. To the best of our knowledge, there does not exist any ZSL-based intermodal retrieval algorithm in the literature. Therefore, in this respect, we performed a few baselines for the sake of comparison. In Baseline-I, we used the pretrained weights from VggNet-16 and utilized a k-NN-based approach to find the top-k retrieved vales. For Baseline-II and Baseline-III, we used a similar framework but with ResNet-50 and ResNet-101, respectively. For Baseline-IV, we obtained 128-d features by using a series of 2-D convolution layers, directly from the images and sketches. We denote the convolutional layer parameters as "conv <receptive field size> − <number of channels>." We used two layers of conv3-64, followed by two layers of conv3-128. A maxpool and a batch-normalization layer follow both these pairs of convolution layers. Another conv3-256 and maxpool layer then follows this. Finally, we fed the output to a fully connected layer of dimension 128. Eventually, we combined the ResNet-101 pretrained network with three subsequent layers of 2-D convolution (last three layers from Baseline-IV) and utilized it as Baseline-V.

We compared the proposed framework with the state-of-the-art (SOTA) ZSL sketch-based image retrieval network (ZS-SBIR) [2]. Similarly, we also compared our results with the zero-shot image hashing (ZSIH) network [3]. This technique is a zero-shot sketch image retrieval model wherein the network uses a generative hashing scheme for constructing the semantic information. It is noteworthy that these are solely SBIR type networks and does not support intermodal retrieval tasks. Therefore, to maintain integrity in comparison, the SBIR method was evaluated using only the sketch-anchored triplets. It can be noted from Table I that our ZSEoC framework outperforms the current SOTA methods. Here, we also observed that the learnable semantic space precision gains a substantial boost than the fixed one. Besides, there is always a tradeoff between the accuracy and efficiency of a system. A larger feature dimension may yield more discriminative and domain-agnostic embedding features, but the retrieval efficiency will suffer due to increased computations. Table II shows the intermodal retrieval results, along with the unimodal (sketch → sketch and images → images) retrieval results.

TABLE II

Intermodal Retrieval Performance of the Proposed ZSEoC Framework on the EoC Data Set

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

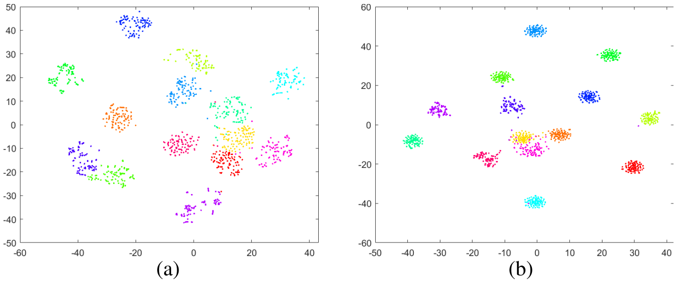

Fig. 2. 2-D scatter plots of high-dimensional features generated with t-SNE of image and sketch features, in the shared latent space, trained with a fixed-semantic vector. Clusters with distinct colors denote separate classes in the data set. (a) Image. (b) Sketch.

It is interesting to note that the unimodal retrieval results are likewise efficiently encoded in this unified feature embedding space along with the intermodal ones, thus providing better retrieval outcomes than the intermodal results. Therefore, the aforementioned indicates the efficiency of the shared embedding space. Furthermore, Fig. 2 shows 2-D scatter plots of the high-dimensional features generated with the t-distributed stochastic neighbor embedding (t-SNE) algorithm for image and sketch features, in the shared latent space that is trained with a fixed-semantic vector. Moreover, it can be noticed that the data instances for both the modalities are separated and grouped in the unified space using the ZSEoC model. Here, we also present a few intermodal retrieval results in Fig. 3, where the images with green borders indicate the correctly retrieved images. In contrast, the ones with red borders show incorrect results.

Ablation Studies: To investigate the effect of each loss function, we performed an ablation study. For this purpose, we ran our experiments on two problem sets: 1) SBIR with a latent-semantic space of 128-d and 2) IBSR. For the first set of experiments, we used the total loss function, except for the latent loss, which primarily aids in bringing the two modalities closer to each other for a consistent retrieval purpose. However, we can notice from Fig. 3 that, without

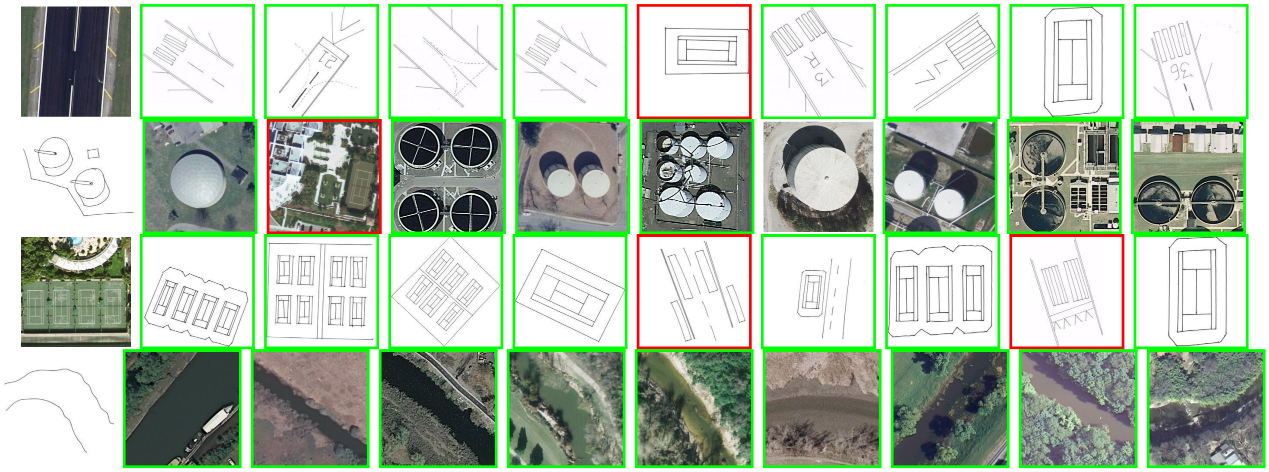

Fig. 3. Few top retrieved results from the zero-shot intermodal framework. Alternate rows represent **Sketch** → **Image** and **Image** → **Sketch** retrievals.

TABLE III

ABLATION OF LOSS FUNCTIONS AND MODEL COMPONENTS IN TERMS OF mAP VALUES. HERE, $\mathcal{L}_{\text{total}}$ DENOTES THE OBJECTIVE FUNCTION, AND "−" DEPICTS THE REMOVAL OF ONE LOSS TERM IN EACH STEP

	Experimental set up	**Sketch→Image**	**Image→Sketch**
Cross-modal	$\mathcal{L}_{total} - \mathcal{L}_{cpl}$	0.190	0.090
	$\mathcal{L}_{total} - \mathcal{L}_{ce}$	0.223	0.212
	$\mathcal{L}_{total} - \mathcal{L}_{iii}$	0.581	0.356
	$\mathcal{L}_{total} - \mathcal{L}_{dl}$	0.683	0.498
	$\mathcal{L}_{total}$	**0.686**	**0.612**

this loss function, there is a noteworthy decrease in the overall performance of the system.

In the second set of experiments, we left out the CE loss function. In doing so, the features in the shared embedding space lost their interclass distances resulting in an ineffective retrieval, which can be seen in Table III. In the third set of experiments, we excluded the cross-triplets loss function from the overall objective function. We notice a significant contrast in the performance between the SBIR and the IBSR modules from Table III. Therefore, it can be noted that the cross-triplets loss aids in boosting both the intermodal retrieval frameworks. However, we can observe that the retrieval ability of the framework significantly increases when retaining only the single cross-triplet (i.e., either sketch anchored or image anchored) while decreasing for the other. Therefore, keeping both sketch and image anchored triplets leads to an optimum tradeoff between both the performances.

In the fourth set of experiments, we excluded the decoder loss function from $\mathcal{L}$. Surprisingly, we still observed an excellent performance of the framework. However, the inclusion of this loss function provided an additional impetus in the execution of the framework, making its performance better than the SOTA. The last row in Table III displays the performance of the complete model with the total objective function.

IV. CONCLUSION

We proposed an aerial image and sketch-based intermodal ZSL framework in RS applications.[2] Our primary motivation is to project the multimodal data into a shared space for intermodal retrieval. We extended this concept to create a framework, wherein we might not have any training samples for some classes; however, there is a possibility that we might find them at any given instance. We exploited the notion of SBIR to tackle the difficulty of insufficient query image for a target within a class during the retrieval process. We proposed a novel zero-shot intermodal architecture for the RS image retrieval using the EoC data set introduced in this work. The performance of the proposed algorithm exceeds the current SOTA results in SBIR.

REFERENCES

[1] F. Xu, R. Zhang, W. Yang, and G.-S. Xia, "Mental retrieval of large-scale satellite images via learned sketch-image deep features," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 3356–3359.

[2] S. K. Yelamarthi, S. K. Reddy, A. Mishra, and A. Mittal, "A zero-shot framework for sketch based image retrieval," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 300–317.

[3] Y. Shen, L. Liu, F. Shen, and L. Shao, "Zero-shot sketch-image hashing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3598–3607.

[4] U. Chaudhuri, B. Banerjee, A. Bhattacharya, and M. Datcu, "A simplified framework for zero-shot cross-modal sketch data retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 182–183.

[5] T. Jiang, G.-S. Xia, and Q. Lu, "Sketch-based aerial image retrieval," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3690–3694.

[6] F. Xu, W. Yang, T. Jiang, S. Lin, H. Luo, and G. Xia, "Mental retrieval of remote sensing images via adversarial sketch-image feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7801–7814, Nov. 2020.

[7] U. Chaudhuri, B. Banerjee, A. Bhattacharya, and M. Datcu, "CMIR-NET: A deep learning based model for cross-modal retrieval in remote sensing," *Pattern Recognit. Lett.*, vol. 131, pp. 456–462, Mar. 2020.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[9] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[10] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, "Large scale online learning of image similarity through ranking," *J. Mach. Learn. Res.*, vol. 11, no. 3, pp. 1109–1135, 2010.

[11] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst. GIS*, 2010, pp. 270–279.

[12] G. Cheng, Z. Li, X. Yao, L. Guo, and Z. Wei, "Remote sensing image scene classification using bag of convolutional features," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1735–1739, Oct. 2017.

[13] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI*, vol. 4, 2017, p. 12.

[2] The EoCs data set and the codes developed in this work are made available at https://ushasi.github.io/Earth-on-Canvas-data set/.