UNSUPERVISED CONTRASTIVE HASHING FOR CROSS-MODAL RETRIEVAL IN REMOTE SENSING

Georgii Mikriukov, Mahdyar Ravanbakhsh, Begüm Demir

Technische Universität Berlin, Berlin, Germany

ABSTRACT

The development of cross-modal retrieval systems that can search and retrieve semantically relevant data across different modalities based on a query in any modality has attracted great attention in remote sensing (RS). In this paper, we focus our attention on cross-modal text-image retrieval, where queries from one modality (e.g., text) can be matched to archive entries from another (e.g., image). Most of the existing cross-modal text-image retrieval systems in RS require a high number of labeled training samples and also do not allow fast and memory-efficient retrieval. These issues limit the applicability of the existing cross-modal retrieval systems for large-scale applications in RS. To address this problem, in this paper we introduce a novel unsupervised cross-modal contrastive hashing (DUCH) method for text-image retrieval in RS. To this end, the proposed DUCH is made up of two main modules: 1) feature extraction module, which extracts deep representations of two modalities; 2) hashing module that learns to generate cross-modal binary hash codes from the extracted representations. We introduce a novel multi-objective loss function including: i) contrastive objectives that enable similarity preservation in intra- and inter-modal similarities; ii) an adversarial objective that is enforced across two modalities for cross-modal representation consistency; and iii) binarization objectives for generating hash codes. Experimental results show that the proposed DUCH outperforms state-of-the-art methods. Our code is publicly available at https://git.tu-berlin.de/rsim/duch.

Index Terms— cross-modal retrieval, hashing, unsupervised contrastive learning, remote sensing.

1. INTRODUCTION

The increased number of recent Earth observation satellite missions has led to significant growth in remote sensing (RS) image archives. Thus, the development of scalable and accurate retrieval systems (which aim at searching for semantically similar data to a given query) for massive archives is one of the most important research topics in RS. Most existing methods in RS focus on content-based RS image retrieval (CBIR). CBIR systems take a query image and compute the similarity function between the query image and all archive images to find the most similar images to the query [1]. To achieve scalable image retrieval, deep hashing techniques have become a cutting-edge research topic for large-scale RS image retrieval [2,3]. These methods map high-dimensional image descriptors into a low-dimensional Hamming space where binary hash codes describe the image descriptors. Compared with the real-valued features, hash codes allow fast image retrieval by calculating the Hamming distances with simple bit-wise XOR operations. In addition, the binary codes can significantly reduce the amount of memory required for storing the content of images.

The abovementioned methods are defined for single-modality image retrieval problems (called uni-modal retrieval). For a given query image, uni-modal retrieval systems search for the images with semantically similar contents from the same modality image archive [4] (e.g., Sentinel-2 images). However, multi-modal data archives, including different modalities of satellite images as well as textual data, are currently available. Thus, the development of retrieval systems that return a set of semantically relevant results of different modalities given a query in any modality (e.g., using a text sentence to search for RS images) has recently attracted significant attention in RS. Due to the semantic gap among different modalities, it is challenging to search relevant multi-modal contents among heterogeneous data archives. To address this problem, cross-modal retrieval methods that aim to identify relevant data across different modalities are recently introduced in RS [5–7]. The existing cross-modal retrieval systems in RS are defined based on supervised retrieval techniques. Such techniques require the availability of labeled training samples (i.e., ground reference samples) to be used in the learning phase of the cross-modal retrieval algorithm. The amount and the quality of the available training samples are crucial for achieving accurate cross-modal retrieval. Collecting a sufficient number of reliable labeled samples is time-consuming, complex and costly in operational scenarios, significantly affecting the final accuracy of cross-modal retrieval.

Unlike RS, in the computer vision (CV) community, unsupervised and in particular self-supervised cross-modal representation learning methods (which only rely on the alignments between modalities) are widely studied [8–13]. As an example, in [9] a deep joint-semantics reconstructing hashing (DJSRH) method is introduced to learn binary codes that preserve the neighborhood structure in the original data. To this end, DJSRH learns a mapping from different modalities into a joint-semantics affinity matrix. Liu et al. [10] propose a joint-modal distribution-based similarity weighting (JDSH) method based on DJSRH, exploiting an additional objective based on cross-modal semantic similarities among samples. The existing unsupervised contrastive learning methods mainly rely on cross-modal (inter-modality) contrastive objectives to obtain consistent representations across different modalities, while the intra-modal contrastive objectives are ignored. This may lead to learning an inefficient embedding space, where the same semantic content is mapped into different points in the embedding space [14].

In this paper, we focus on the cross-modal retrieval between RS image and text sentence (i.e., caption), which consists of two sub-tasks: i) RS image retrieval with a sentence as query; and ii) sentence retrieval with an RS image as the query. In detail, we propose a novel deep unsupervised cross-modal contrastive hashing (DUCH) method that considers both inter- and intra-modality contrastive objectives for representation learning. In detail, we introduce a novel multi-objective loss function that consists of: 1) intra- and inter-modal contrastive objectives that enable similarity preser-
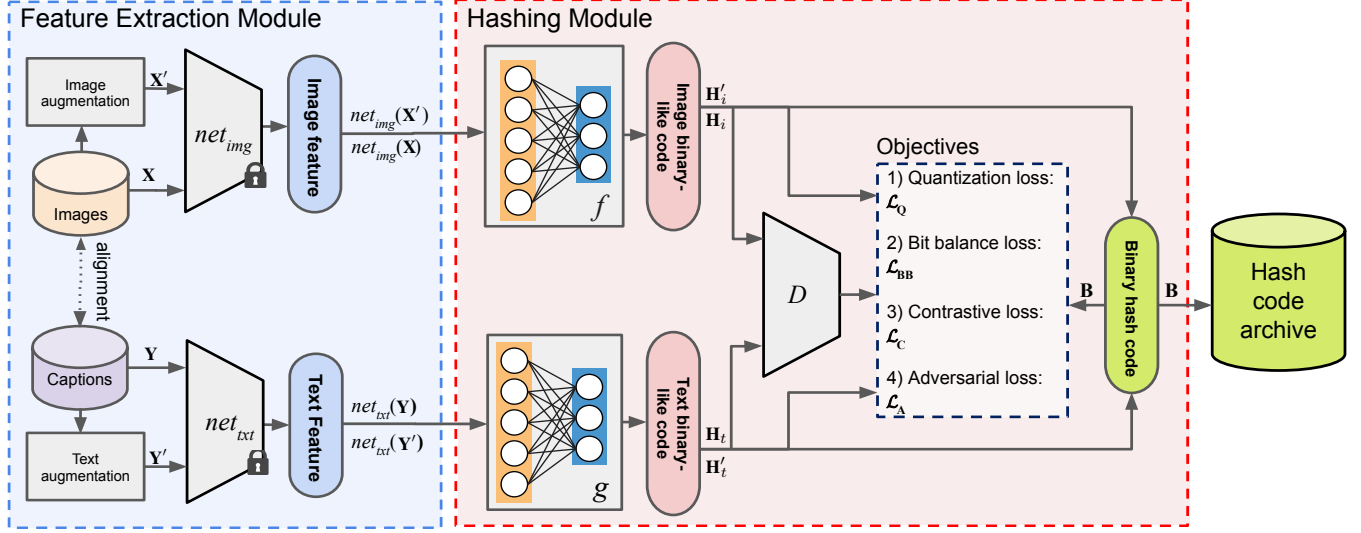
Fig. 1. Block diagram of the proposed method. In the feature extraction module, the modality-specific encoders net_{img} and net_{txt} extract deep embedding for image and text modalities, respectively. The hashing module learns two hash functions f and g from the input embedding. Discriminator D enforces the adversarial objective between two modalities for cross-modal representation consistency.

vation within and between modalities; 2) an adversarial objective enforcing cross-modal representation consistency; 3) binarization objectives to generate representative binary hash codes.

2. PROPOSED METHOD

We assume that an unlabelled multimodal training set $\mathbf{O}$ that consists of N number of image and text pairs (i.e. samples) is available, where $\mathbf{O} = \{\mathbf{X}, \mathbf{Y}\}^N$ and each image is described with only one caption. $\mathbf{X} = \{x_m\}_{m=1}^N$ and $\mathbf{Y} = \{y_m\}_{m=1}^N$ are image and text modality sets, respectively, where $x_m \in \mathbb{R}^{d_i}$ and $y_m \in \mathbb{R}^{d_t}$ are image and text feature vectors. Image and text feature vector dimensions are defined by d_i and d_t. Given the multimodal training set $\mathbf{O}$, the proposed DUCH aims at learning hash functions f and g for image and text modalities, respectively. In detail, the hash functions f and g learn to generate binary hash codes $\mathbf{B}_i = f(\mathbf{X}, \theta_i)$ and $\mathbf{B}_t = g(\mathbf{Y}, \theta_t)$, where $\mathbf{B}_i \in \{0, 1\}^{N \times B}$ and $\mathbf{B}_t \in \{0, 1\}^{N \times B}$ for image and text modalities. θ_i, θ_t are parameters of image and text hashing networks and B is the length of binary hash code. In order to learn the hash functions $f(.)$ and $g(.)$, the proposed DUCH includes two main modules: i) the feature extraction module that produces deep semantic representations for image and text modalities and input them to the next module; and ii) the hashing module that generates binary representations from the input extracted deep features. Fig. 1 shows the block diagram of the DUCH, indicating the two modules. In the following, we describe each module in detail.

2.1. Feature extraction module

The feature extraction module aims at generating deep semantic representations for both images and captions, and feed them into the next module (hashing). To this end, the feature extraction module includes two networks as modality-specific encoders: 1) an image embedding network denoted as net_{img}; and 2) a language model for encoding the textual information (i.e., captions) denoted as net_{txt}. Weights of image and text encoders are fixed during the training of

the hashing learning module. Given the training set $\mathbf{O}$, the image and text embedding are extracted by $net_{img}(\mathbf{X})$ and $net_{txt}(\mathbf{Y})$, respectively. For the sake of simplicity in the rest of this paper we refer $net_{img}(\mathbf{X})$ as $\mathbf{X}$, and $net_{txt}(\mathbf{Y})$ as $\mathbf{Y}$. For the unsupervised contrastive representation learning of DUCH, we generate a corresponding augmented set from $\mathbf{O}$, which is defined as $\mathbf{O}' = \{\mathbf{X}', \mathbf{Y}'\}^N$, where $\mathbf{X}' = \{x'_m\}_{m=1}^N$ and $\mathbf{Y}' = \{y'_m\}_{m=1}^N$ are augmented image and caption with elements $x'_m \in \mathbb{R}^{d_i}$ and $y'_m \in \mathbb{R}^{d_t}$. The embedding of augmented images and captions are extracted by net_{img} and net_{txt}, respectively. For the sake of simplicity in the rest of this paper we refer $net_{img}(\mathbf{X}')$ as $\mathbf{X}'$, and $net_{txt}(\mathbf{Y}')$ as $\mathbf{Y}'$.

2.2. Hashing module

The hashing module aims at learning hashing function to generate cross-modal binary hash codes from the image and text embeddings. To this end, we employ multiple objectives for training hash functions f and g, including contrastive loss, adversarial loss and binarization losses. The cross-modal contrastive loss is the main objective for unsupervised representation learning in the proposed DUCH. In detail, we introduce an intra-modal (modality-specific) contrastive loss to improve the learned representations. The adversarial objective ensures the modality-invariance of the binary codes. The binarization objectives are required to approximate better the generated continuous values close to the discrete binary codes.

Contrastive objectives. In order to calculate the contrastive losses we use the normalized temperature scaled cross entropy objective function adapted from [12,15]. To ensure better representation learning we consider both inter-modal and intra-modal contrastive losses. The former maps features of both modalities into a common space, while the latter performs the mapping within each modality. The inter-modal contrastive loss $\mathcal{L}_{C_{inter}}$ is defined as:

$$\mathcal{L}_{C_{inter}}(x_j, y_j) = \tag{1}$$
$$-\log \frac{S\left(f(x_j), g(y_j)\right)}{\sum_{k=1, k \neq j}^M S\left(f(x_j), f(x_k)\right) + \sum_{k=1}^M S\left(f(x_j), g(y_k)\right)},$$

where $S(\mathrm{u}, \mathrm{v}) = \exp\left(\cos\left(\mathrm{u}, \mathrm{v}\right)/\tau\right)$, and $\cos\left(\mathrm{u}, \mathrm{v}\right) = \mathrm{u}^T \mathrm{v}/\left\|\mathrm{u}\right\|\left\|\mathrm{v}\right\|$ is the cosine similarity, τ denotes a temperature, and M is a batch size. The inter-modal contrastive loss $\mathcal{L}_{C_{inter}}$ is computed between image x_j and its paired caption y_j. The intra-modal contrastive losses for image and text modalities are computed as:

$$\mathcal{L}_{C_{img}}(\mathrm{x}_j, \mathrm{x}'_j) = \tag{2}$$
$$-\log \frac{S\left(f(\mathrm{x}_j), f(\mathrm{x}'_j)\right)}{\sum_{k=1, k \neq j}^{M} S\left(f(\mathrm{x}_j), f(\mathrm{x}_k)\right) + \sum_{k=1}^{M} S\left(f(\mathrm{x}_j), f(\mathrm{x}'_k)\right)},$$

$$\mathcal{L}_{C_{txt}}(\mathrm{y}_j, \mathrm{y}'_j) = \tag{3}$$
$$-\log \frac{S\left(g(\mathrm{y}_j), g(\mathrm{y}'_j)\right)}{\sum_{k=1, k \neq j}^{M} S\left(g(\mathrm{y}_j), g(\mathrm{y}_k)\right) + \sum_{k=1}^{M} S\left(g(\mathrm{y}_j), g(\mathrm{y}'_k)\right)},$$

where $\mathcal{L}_{C_{img}}$ is the contrastive loss between image x_j and its augmentation x'_j, while $\mathcal{L}_{C_{txt}}$ is the contrastive loss between caption y_j and its augmentation y'_j. The final contrastive loss $\mathcal{L}_C$ is defined as:

$$\mathcal{L}_C = \mathcal{L}_{C_{inter}} + \lambda_1 \mathcal{L}_{C_{img}} + \lambda_2 \mathcal{L}_{C_{txt}}, \tag{4}$$

where λ_1 and λ_2 are hyperparameters for image and text intra-modal contrastive losses, respectively.

Adversarial objective. As suggested in [16], to enforce the consistency of representations across modalities, we employ an adversarial loss within DUCH. We use a discriminator network D trained in an adversarial fashion, where the text embedding is assumed to be "real" and the image embedding is considered as "fake". The adversarial objective function $\mathcal{L}_{adv}$ is defined as:

$$\mathcal{L}_{adv}\left(\mathbb{X}, \mathbb{Y}\right) = -\frac{1}{N} \sum^{N} \Big[\log\Big(D\big(g\left(\mathbb{Y}\right)\big)\Big) + \tag{5}$$
$$\log\Big(1 - D\big(f(\mathbb{X})\big)\Big)\Big],$$

where $\mathbb{X} = \{\mathbf{X}, \mathbf{X}'\}$ for the image representations and $\mathbb{Y} = \{\mathbf{Y}, \mathbf{Y}'\}$ for the text representations. $D(*, \theta_D)$ $(* = \mathbb{X}, \mathbb{Y})$ is the discriminator network D with parameters θ_D.

Binarization objectives. In order to generate representative hash codes, we consider two main objective functions for binarization: quantization loss [17] and bit balance loss [18]. The former optimizes the difference between continuous and discrete hash values, and the latter enforcing each output neuron to fire with an equal chance. The quantization loss $\mathcal{L}_Q$, and bit balance loss $\mathcal{L}_{BB}$ are calculated as:

$$\mathcal{L}_Q = \left\|\mathbf{B} - \mathbf{H}_i\right\|_F^2 + \left\|\mathbf{B} - \mathbf{H}'_i\right\|_F^2 + \left\|\mathbf{B} - \mathbf{H}_t\right\|_F^2 + \left\|\mathbf{B} - \mathbf{H}'_t\right\|_F^2, \tag{6}$$

$$\mathcal{L}_{BB} = \left\|\mathbf{H}_i \cdot \mathbf{1}\right\|_F^2 + \left\|\mathbf{H}'_i \cdot \mathbf{1}\right\|_F^2 + \left\|\mathbf{H}_t \cdot \mathbf{1}\right\|_F^2 + \left\|\mathbf{H}'_t \cdot \mathbf{1}\right\|_F^2, \tag{7}$$

where $\mathbf{H}_i = f(\mathbf{X})$, $\mathbf{H}'_i = f(\mathbf{X}')$, $\mathbf{H}_t = g(\mathbf{Y})$, $\mathbf{H}'_t = g(\mathbf{Y}')$ are binary like codes for images, augmented images, texts and augmented texts respectively, and $\mathbf{1}$ is N-dimensional vector with all values of 1. The final binary code update rule is defined as:

$$\mathbf{B} = sign\left(\frac{1}{2}\left(\frac{\mathbf{H}_i + \mathbf{H}'_i}{2} + \frac{\mathbf{H}_t + \mathbf{H}'_t}{2}\right)\right). \tag{8}$$

The overall loss function is a weighted sum of multiple objectives from (4), (5), (6), and (7):

$$\min_{\mathbf{B}, \theta_i, \theta_t, \theta_D} \mathcal{L} = \mathcal{L}_C + \alpha\mathcal{L}_{adv} + \beta\mathcal{L}_Q + \gamma\mathcal{L}_{BB}, \tag{9}$$

where α, β, γ are hyperparameters for adversarial, quantization and bit balance losses, respectively. Finally, to retrieve semantically similar captions to a query image x_q, we compute the Hamming distance between $f(net_{img}(\mathrm{x}_q))$ and hash codes in the retrieval archive. The obtained distances are ranked in ascending order and the top-K captions with the lowest distances are retrieved. Similarly, for a query caption y_q, the Hamming distance between $g(net_{txt}(\mathrm{y}_q))$ and hash codes in the retrieval archive are computed, ranked and top-K images are retrieved.
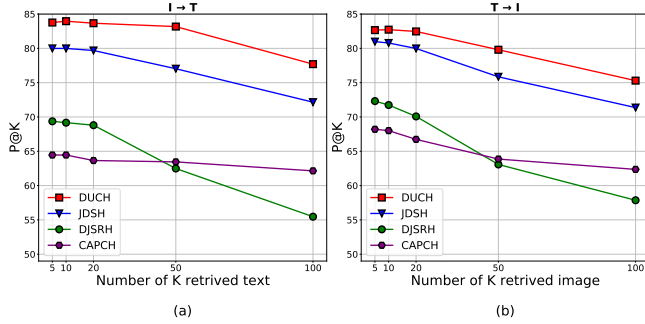
3. EXPERIMENTAL RESULTS

In the experiments we used the RSICD dataset [19] that includes 10921 images of 31 classes from aerial orthoimagery. Each image has a size of 224×224 pixels and has 5 corresponding captions. Only one randomly selected caption for each image was used during the training. The dataset was split by random selection into the train, query, and retrieval sets (50%, 10% and 40%, respectively). The image augmentation was performed by applying a Gaussian blur filter with kernel size 3×3 and $\sigma \in [1.1, 1.3]$, random rotation in the range of $[-10°, -5°]$, and 200×200 center cropping. For the text augmentation, we selected a rule-based algorithm suggested in [20], where the noun and verb tokens are replaced with semantically similar tokens. In the feature extraction module, we used a pre-trained ResNet [21] network for net_{img} the classification head of the model was removed, and the image embedding size d_i was set to 512. For the text embedding we used a pre-trained BERT [22] language model provided in [23]. The final size of the sentence embedding d_t was obtained as 768 by summing the last four hidden states of each token. The internal parameters of the image and text encoders were kept fixed during the training of the hash learning module. In the hash learning module, the networks f (image hashing) and g (text hashing) are fully connected networks with 3 layers and a batch normalization layer after the second layer. For the discriminator D we selected a simple 2 layer fully connected network. *ReLU* was used as a default activation function for all layers except for the last layers of image hashing and text hashing networks that use the *tanh* activation function. The hyperparameters α, β, γ were set to $\alpha = 0.01$, $\beta = 0.001$, $\gamma = 0.01$ based on a grid search strategy. Both intra-modal weight coefficients λ_1 and λ_2 were set to 1 (see (4)), while the batch size was set to 256 and the total number of training epochs was selected as 100. The initial learning rate was set to 0.0001 and decreased by one fifth every 50 epochs. The Adam optimizer was chosen for f, g, and D networks. We compared the performance of the proposed method with three state of the art methods: i) supervised multi-task consistency-preserving adversarial hashing (CPAH) [24] that separates the feature representations into modality-specific and modality-common features, exploiting label information to learn consistent modality-specific representations; ii) unsupervised method DJSRH [9]; iii) unsupervised method JDSH [10]. For a fair comparison, we trained all models under the same experimental setup. Results of each method are provided in terms of: i) mean average precision (mAP) and ii) precision. The mAP performance was assessed on top-20 retrieved images (denoted as mAP@20), while precision was evaluated by varying the number K of retrieved images in the range of [1-100].

In our experiments, we initially analyzed the impact of hash code lengths on retrieval performance. Table 1 shows the retrieval results obtained by the proposed DUCH and the state-of-the-art methods. From the table, one can see that in general, when the number of the hash bits increases for image-to-text and text-to-image retrieval tasks, the mAP@20 obtained by all methods increases. In addition,

Table 1. The mAP@20 results for image-to-text ($I \rightarrow T$) and text-to-image ($T \rightarrow I$) retrieval tasks.

Task	Method	B=16	B=32	B=64	B=128
$I \rightarrow T$	CPAH [24]	0.428	0.587	0.636	0.696
	DJSRH [9]	0.411	0.665	0.688	0.722
	JDSH [10]	0.385	0.720	0.796	0.815
	DUCH *(ours)*	**0.684**	**0.791**	**0.836**	**0.829**
$T \rightarrow I$	CPAH [24]	0.452	0.598	0.667	0.706
	DJSRH [9]	0.422	0.685	0.705	0.733
	JDSH [10]	0.418	0.751	0.799	0.815
	DUCH *(ours)*	**0.697**	**0.780**	**0.824**	**0.826**

Table 2. The mAP@20 results for image-to-text ($I \rightarrow T$) and text-to-image ($T \rightarrow I$) retrieval tasks under different configurations of the proposed DUCH when $B = 64$.

Method	Configuration	$I \rightarrow T$	$T \rightarrow I$
DUCH	original	0.836	0.824
DUCH-NA	excluding $\mathcal{L}_{adv}$: $\alpha = 0$	0.831	0.822
DUCH-NQ	excluding $\mathcal{L}_Q$: $\beta = 0$	0.818	0.800
DUCH-NB	excluding $\mathcal{L}_{BB}$: $\gamma = 0$	0.828	0.826
DUCH-CL	excluding $\mathcal{L}_{C_{img}}$, $\mathcal{L}_{C_{txt}}$: $\lambda_1 = 0$, $\lambda_2 = 0$	0.758	0.765
DUCH-CL-I	excluding $\mathcal{L}_{C_{txt}}$: $\lambda_2 = 0$	0.811	0.796
DUCH-CL-T	excluding $\mathcal{L}_{C_{img}}$: $\lambda_1 = 0$	0.813	0.815

Fig. 2. Precision versus number of retrieved images (P@K) obtained when B=64 for (a) image-to-text ($I \rightarrow T$) and (b) text-to-image ($T \rightarrow I$) retrieval tasks.

the proposed DUCH outperforms all the baselines, especially when hash codes with short lengths are considered. As an example, when B=16 the mAP@20 is 29.9% higher than that of JDSH (which is the second-best performing method according to our results). The mAP@20 obtained by DJSRH and CPAH are 27.4% and 25.6% smaller than DUCH, respectively. The differences in mAPs are reduced when the higher lengths of the hash codes are considered. We would like to note that DJSRH and JDSH are unsupervised methods that have several hyperparameters and are very sensitive to tweaking. We also analyzed the performance of the proposed DUCH under different values of K (the number of retrieved images) and compared it to the baseline methods. Precision versus the number K of retrieved images when B=64 are shown in Fig. 2. From the figure one can observe that the precision obtained by the DUCH (in each value of K) is higher than the baseline methods. This shows that the success of the proposed DUCH in efficiently mapping the semantic information into discriminative hash codes.

To analyze the influence of each objective, we designed different configurations by excluding individual objectives from DUCH. In detail, we compared the original DUCH (that includes all the objectives) with six different configurations of DUCH when we exclude: 1) adversarial objective (denoted as DUCH-NA); 2) quantization objective (denoted as DUCH-NQ); 3) bit balance objective (denoted as DUCH-NB); 4) image and text intra-modal contrastive losses (denoted as DUCH-CL); 5) text intra-modal contrastive loss (denoted as DUCH-CL-I); and 6) image intra-modal contrastive loss (denoted as DUCH-CL-T). The results for different configurations are shown in Table 2. The controllable objectives $\mathcal{L}_{adv}$, $\mathcal{L}_Q$ and $\mathcal{L}_{BB}$, from (9) only marginally contribute to the overall performance. The highest performance drop (2%) is observed in the case of DUCH-NQ (which does not include the quantization objective). By analyzing

the obtained mAP@20 scores from DUCH-NA and DUCH-NB, one can observe that the Adversarial and Bit balance losses have a less significant impact. This shows that the learned representations with the proposed unsupervised DUCH are consistent, and generated hash codes have independent bits. The intra-modal contrastive losses effectively boost cross-modal unsupervised training compared to the individual single intra-modal contrastive loss (see result of DUCH-CL). Without the intra-modal losses, the performance of the model significantly drops about 8%. This shows the importance of the joint use of intra-modal losses to improve the inter-modal contrastive loss. Furthermore, the performance drop in DUCH-CL-I (which does not include text intra-modal contrastive objective) is higher than DUCH-CL-T (which does not contain image intra-modal contrastive objective). This shows that the text intra-modal contrastive objective is more influential on the overall performance than the image intra-modal objective, mainly due to using a modality-specific encoder for the images pre-trained on a different domain.

4. CONCLUSION

In this paper, a novel deep unsupervised contrastive hashing (DUCH) method has been proposed for cross-modal image-text retrieval in RS. The proposed DUCH exploits a multi-objective loss function to learn a cross-modal representation in an unsupervised fashion. In detail, we have introduced a contrastive objective function that considers both inter- and intra-modal similarities. Furthermore, we have presented an adversarial objective function that assists in generating modality-invariant representations. We have demonstrated the effectiveness of the proposed inter-modality contrastive losses through the experimental results and show the superiority of the proposed DUCH over existing state-of-the-art methods. As a final remark we would like to note that through the experimental results we also observed that the intra-modal contrastive loss is less effective when the deep features are obtained from a modality-specific encoder pre-trained on a different domain. To address this problem, as a future work, we plan for end-to-end training of the feature extraction module and fine-tuning the modality-specific encoders.

5. ACKNOWLEDGMENTS

6. REFERENCES

[1] G. Sumbul, J. Kang, and B. Demir, "Deep learning for image search and retrieval in large remote sensing archives," *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences*, pp. 150–160, 2021.

[2] S. Roy, E. Sangineto, B. Demir, and N. Sebe, "Metric-learning based deep hashing network for content based retrieval of remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 2, pp. 226–230, 2021.

[3] Q. Cheng, H. Huang, L. Ye, P. Fu, D. Gan, and Y. Zhou, "A semantic-preserving deep hashing model for multi-label remote sensing image retrieval," *Remote Sensing*, vol. 13, no. 24, 2021.

[4] R. Cao, Q. Zhang, J. Zhu, Q. Li, Q. Li, B. Liu, and G. Qiu, "Enhancing remote sensing image retrieval using a triplet deep metric learning network," *International Journal of Remote Sensing*, vol. 41, no. 2, pp. 740–751, 2020.

[5] Y. Chen, X. Lu, and S. Wang, "Deep cross-modal image–voice retrieval in remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 10, pp. 7049–7061, 2020.

[6] Q. Cheng, Y. Zhou, P. Fu, Y. Xu, and L. Zhang, "A deep semantic alignment network for the cross-modal image-text retrieval in remote sensing," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 4284–4297, 2021.

[7] H. Ning, B. Zhao, and Y. Yuan, "Semantics-consistent representation learning for remote sensing image-voice retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, 2022.

[8] H. Jung, Y. Oh, S. Jeong, C. Lee, and T. Jeon, "Contrastive self-supervised learning with smoothed representation for remote sensing," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[9] S. Su, Z. Zhong, and C. Zhang, "Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3027–3035.

[10] S. Liu, S. Qian, Y. Guan, J. Zhan, and L. Ying, "Joint-modal distribution-based similarity hashing for large-scale unsupervised deep cross-modal retrieval," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 1379–1388.

[11] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[12] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the International Conference on Machine Learning*, 2020, pp. 1597–1607.

[13] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[14] M. Zolfaghari, Y. Zhu, P. Gehler, and T. Brox, "Crossclr: Cross-modal contrastive learning for multi-modal video representations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1450–1459.

[15] H. Zhang, J. Y. Koh, J. Baldridge, H. Lee, and Y. Yang, "Cross-modal contrastive learning for text-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 833–842.

[16] C. Bai, C. Zeng, Q. Ma, J. Zhang, and S. Chen, "Deep adversarial discrete hashing for cross-modal retrieval," in *Proceedings of International Conference on Multimedia Retrieval*, 2020, pp. 525–531.

[17] Q. Jiang and W. Li, "Deep cross-modal hashing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3232–3240.

[18] E. Yang, C. Deng, W. Liu, X. Liu, D. Tao, and X. Gao, "Pairwise relationship guided deep hashing for cross-modal retrieval," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017, pp. 1618–1625.

[19] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2183–2195, 2017.

[20] J. Wei and K. Zou, "Eda: Easy data augmentation techniques for boosting performance on text classification tasks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 6383–6389.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[22] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2019.

[23] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2020, pp. 38–45, Association for Computational Linguistics.

[24] D. Xie, C. Deng, C. Li, X. Liu, and D. Tao, "Multi-task consistency-preserving adversarial hashing for cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 29, pp. 3626–3637, 2020.