Letter

Multi-Attention Fusion and Fine-Grained Alignment for Bidirectional Image-Sentence Retrieval in Remote Sensing

Qimin Cheng, Yuzhuo Zhou, Haiyan Huang, and Zhongyuan Wang

Dear editor,

Cross-modal retrieval in remote sensing (RS) data has inspired increasing enthusiasm due to its merit in flexible input and efficient query. In this letter, we address to establish semantic relationship between RS images and their description sentences. Specially, we propose a multi-attention fusion and fine-grained alignment network, termed MAFA-Net, for bidirectional cross-modal image-sentence retrieval in RS. While multiple attention mechanisms are fused to enhance the discriminative ability of visual features for RS images with complex scenes, fine-grained alignment strategy is introduced to study the hidden connection between RS observations and sentences. To validate the capability of MAFA-Net, we leverage four captioning benchmark datasets with paired RS images and descriptions, i.e., UCM-Captions, Sydney-Captions, RSICD and NWPU-Captions. Experimental results on the four datasets demonstrate that MAFA-Net can yield better performance than the current state-of-the-art approaches.

Related work: The accelerated advancement in earth observation technology witnesses an explosive growth of multi-modal and multi-source remote sensing data. Cross-modal retrieval in RS facilitates flexible and efficient query, which has attracted extensive interest in recent years and can be applied to natural disaster early-warning and military intelligence generation, etc.

Significant efforts have been devoted to cross-modal retrieval for natural images. To probe fine-grained relationships among images and sentences, Chen *et al*. [1] proposed a cross-modal retrieval model (IMRAM) based on a recurrent attention technique. Lee *et al*. [2] proposed a stacked attention mechanism-based graphic retrieval model (SCAN) to learn more discriminative textual and visual feature representations. Wang *et al*. [3] proposed a multi-modal tensor fusion network (MTFN) to directly measure the similarity between different modalities through rank-based tensor fusion. Wang *et al*. [4] proposed a position focused attention network (PFAN) to improve cross-modal matching performance. Besides, to satisfy industrial requirement, Wu *et al*. [5] proposed a hashing approach to achieve large-scale cross-modal retrieval via learning a unified hash representation and deep hashing functions for different modalities in a self-supervised way. Although these achievements gained inspiring results for retrieval tasks in natural images, their robustness and generalization ability need to be verified when transfer to RS fields due to the intrinsic and extrinsic properties of RS data.

Motivated by the burgeoning demands for multi-modal requests in RS like military intelligence generation, researchers have paid more attention to RS cross-modal retrieval during the recent several years. To explore semantic correlation between visual features and textual description of RS data, Abdullah *et al*. [6] proposed a novel deep bidirectional ternary network (DBTN) for Text-to-Image (T2I) matching task through features fusion strategy. With regard to Image-to-Text (I2T) retrieval for RS data, Cheng *et al*. [7] proposed to use a cross-attention mechanism and a gating mechanism to enhance the association between RS images and descriptions, which is the first attempt to prove the possibility of bidirectional T2I and I2T retrieval in RS. Afterwards, Lv *et al*. [8] proposed a fusion-based correlation learning model (FCLM) to capture multi-modal complementary information and fusion features and to further supervise the learning of the feature extraction network. Yuan *et al*. [9] proposed an asymmetric multimodal feature matching network (AMFMN) to extract the salient visual features of RS images through a multi-scale visual self-attention technique, and exploited it to guide textual feature extraction. Moreover, they further designed a concise and efficient version of their cross-modal retrieval model, namely LW-MCR [10] on the basis of knowledge distillation. For fast and efficient retrieval on large-scale RS data, Mikriukov *et al*. [11] introduced a novel deep unsupervised cross-modal contrastive hashing model. Except for image-sentence retrieval, there has been some work on visual-audio retrieval [12], image-sketch retrieval [13], cross-source panchromatic-multispectral image retrieval [14], [15] and zero-shot image-word matching [16].

It is no doubt that all the above work partly advances the cross-modal retrieval in RS from different aspects including visual feature representation and description optimization strategy, etc. However, current work on bidirectional image-sentence retrieval in RS is deficient in 1) Achievements on bidirectional image-sentence retrieval for RS data are very limited and comprehensive analysis is still lacking. Current work [6]−[11] conducts comparative experiments with the baseline for natural images unexceptionally; 2) The generalization of existing approaches on much larger and more challenging RS captioning datasets needs to be verified. The size of the datasets applied by existing approaches [6], [8]−[11] is limited (with the maximum of 24 333 original captions in RSICD [17] and 23 715 granular captions in RSITMD [9]); 3) Semantic ambiguity of complex scenes of RS data remains unsolved.

To address these limitations, we propose a novel cross-modal network for bidirectional T2I and I2T retrieval in RS. The contribution of our work lies in: 1) We aim to differentiate visual features for complex scene representation through fusing multiple attention mechanisms and reinforce the intra-modality semantic association through fine-grained alignment strategy. 2) We evaluate the effectiveness and robustness of our proposed network on a much larger dataset, NWPU-Captions with 157 500 captions in total, along with the several popular benchmark datasets.

MAFA-Net: The motivation of MAFA-Net includes two aspects. The first one is to depict RS images, especially those complex scenes, with more abstract and discriminative feature representation. The second one is to address semantic ambiguity existed in different modality of RS data through establishing fine-grained relevance between RS image region and visual words.

To this end, MAFA-Net consists of two main parts: a multi-attention fusion module and a fine-grained alignment module. The multi-attention fusion module aims to weaken interference from background noise in RS images and enhance the salient objects, thereby to improve the discriminative ability of the visual features. The fine-grained alignment module exploits sentence features as context information to further optimize and update the visual features of RS images. The overall architecture of MAFA-Net is shown in Fig. 1.

Multi-attention fusion module: It is designed to filter out redundant information, refine salient features, and capture contextual correlation from the extracted visual features of RSIs.

For RS images, we firstly use $conv_2$, $conv_3$ and $conv_4$ in the ResNet152 network to obtain their corresponding convolutional features (F_{c2}, F_{c3} and F_{c4}), respectively. Then, we up-sample the

Corresponding author: Qimin Cheng.

Citation: Q. M. Cheng, Y. Z. Zhou, H. Y. Huang, and Z. Y. Wang, "Multi-attention fusion and fine-grained alignment for bidirectional image-sentence retrieval in remote sensing," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 8, pp. 1532–1535, Aug. 2022.

Q. M. Cheng and Y. Z. Zhou are with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: chengqm@hust.edu.cn; zhouyuzhuo@hust.edu.cn).

H. Y. Huang is with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: eduhuanghaiyan@163.com).

Z. Y. Wang is with the School of Computer Science, Wuhan University, Wuhan 430079, China (e-mail: wzy_hope@163.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2022.105773

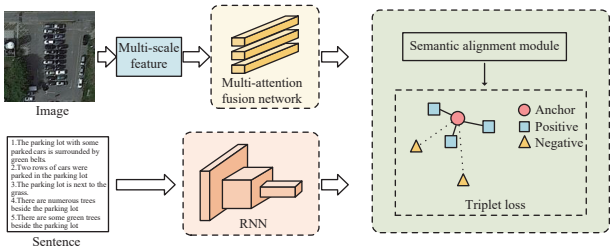

Fig. 1. The overall architecture of MAFA-Net.

F_{c3} and F_{c4} separately to make them the same size as F_{c2}, and let F_{c2} pass through an inception module to expand the receptive field. Finally, the three feature maps are added and fused to obtain the multi-scale visual feature F_{MS}

$$F_{MS} = Inception(F_{c2}) + upsample(F_{c3}) + upsample(F_{c4}). \quad (1)$$

Fig. 2 presents the architecture of multi-attention fusion module. The channel attention branch suppresses background noise and redundant information while enhancing salient features. The spatial attention branch enables the modal to focus on semantically rich regions through aggregating pixels with higher relevance together. The position attention branch encodes contextual information into local features. Finally, the three attention values are integrated adaptively through weighted fusion with learned weight parameters.

$$\alpha_C = soft\max(W_i \tanh((W_C \oplus F_{MS} + b_C) \oplus W_{hC} h_{t-1}) + b_i) \quad (2)$$

$$\alpha_S = soft\max(W_i^{'} \tanh((W_S F_{MS} + b_S) \oplus W_{hS} h_{t-1}) + b_i^{'}) \quad (3)$$

$$\alpha_P = soft\max(W_i^{''} \tanh((W_P \otimes F_{MS} + b_P) \oplus W_{hP} h_{t-1}) + b_i^{''}) \quad (4)$$

$$F_{AF} = \alpha_C + \beta\alpha_S + \lambda\alpha_P. \quad (5)$$

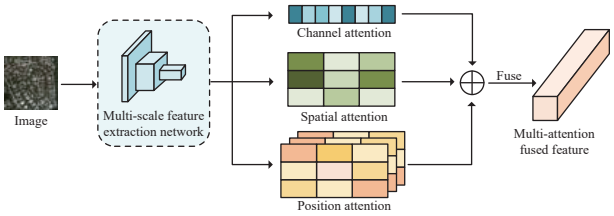

Fig. 2. The architecture of multi-attention fusion module.

In which α_C, α_S and α_P mean channel, spatial and position attention feature, W_C, W_S and W_P are the to-be-learned weight matrices, and b_C, b_S and b_P are bias vectors. The parameter β and λ are the weight coefficient to be learned in training process. F_{AF} is the fused attention feature.

Fine-grained alignment module: We reinforce the inter-modality semantic consistency with the help of the fine-grained semantic alignment module, in which the text features are used as contextual information to guide the gate mechanism for visual adaption.

Before semantic alignment, the feature representation of sentences needs to be extracted. In this letter, bidirectional GRU is utilized to extract sentence features

$$y_i = W_y w_i, \quad i \in \{1, 2, \ldots, N\} \quad (6)$$

$$\vec{h}_i = \overrightarrow{GRU}(\vec{h}_{i-1}, y_i) \quad (7)$$

$$\overleftarrow{h}_i = \overleftarrow{GRU}(\overleftarrow{h}_{i+1}, y_i) \quad (8)$$

$$t_i = (\vec{h}_i + \overleftarrow{h}_i)/2 \quad (9)$$

where y_i means word-level vector of word w_i, in dimension of 300. W_y represents the mapping matrix, and t_i is the word feature extracted by Bi-GRU. The dimension of text feature is 2048.

We define visual features to consist of K regional features $F_{AF} = \{f_1, f_2, \ldots, f_K\}$, and sentence features to consist of N word features $T_S = \{t_1, t_2, \ldots, t_N\}$. Then, these two sets of cross-model features are semantically associated via an attention mechanism. A high attention score typically indicates tight coupling between an image region and a sentence, and vice versa. Specifically, for N words together with K image regions, the fine-grained alignment module first performs $K \times N$ pair-wise cosine similarity calculation by

$$s_{ij} = \frac{f_i^T t_j}{\|f_i\| \cdot \|t_j\|} \quad (10)$$

$$[s_{ij}]_+ = \max\{s_{ij}, 0\} \quad (11)$$

$$\bar{s}_{ij} = \frac{[s_{ij}]_+}{\sqrt{\sum_{i=1}^{K} [s_{ij}]_+^2}}. \quad (12)$$

A row-wise softmax$(\cdot)$ operation is performed on the similarity scores, followed by weighting each word feature with its corresponding attention score to acquire the sentence-level textual feature of each image region:

$$e_i = \sum_{j=1}^{N} t_j \times soft\max(\bar{s}_{ij}). \quad (13)$$

Two gate functions involving an update one g_i and a new memory one c_i are designed to enhance the discrimination of the final feature

$$l_i(W, b) = concat(f_i, e_i) \times W + b \quad (14)$$

$$g_i = sigmoid[l_i(W_1, b_1)] \quad (15)$$

$$c_i = sigmoid[l_i(W_2, b_2)] \quad (16)$$

$$\tilde{f}_i = (1 - g_i) \times f_i + g_i \times c_i \quad (17)$$

where W_1, b_1, W_2 and b_2 are learnable parameters, with g_i filtering out insignificant information and c_i preserving discriminative information. Eventually, we regard $\tilde{f}$ as the remodeled image region feature.

Loss functions: The consistency between a RS image and a sentence can be measured by $S(V, T)$

$$S(V, T) = \frac{1}{K} \sum_{i=1}^{K} \frac{\tilde{f}_i^T e_i}{\|\tilde{f}_i\| \cdot \|e_i\|}. \quad (18)$$

We adopt the triplet loss function to train our proposed model, which is defined as

$$L = \sum_{M} \{[\delta - S(V_i, T_i^p) + S(V_i, T_i^n)]_+ + [\delta - S(V_i^p, T_i) + S(V_i^n, T_i)]_+\} \quad (19)$$

where $S(V_i, T_i^p)$ and $S(V_i^p, T_i)$ are positive image-sentence pairs, and the $S(V_i, T_i^n)$ and $S(V_i^n, T_i)$ are the hard negative image-sentence pairs. δ is the margin threshold for triplet loss. The training data is divided into M mini-batches to mitigate the computational burden during the training process. In this way, the hardest negative example is only searched in its corresponding mini-batch.

Dataset and metrics: Four RS datasets are selected to evaluate the performance of different approaches in the cross-modal image-sentence retrieval task.

1) UCM-Captions: This dataset is released by [18] based on the UCMerced dataset. The size of each image is 256×256, and the pixel resolution is 0.3048 m. Each image is described with five different sentences and hence contains 10 500 descriptions in total.

2) Sydney-Captions: This dataset is released by [18] based on the Sydney dataset and includes 3065 descriptions for 613 cropped images. The original images in it are with size of 18 000×14 000 and pixel resolution of 0.5 m. Each cropped image is described by five varied sentences.

3) RSICD: There are totally 10 921 RS images and 24 333 original descriptions in this dataset [17], the scale of which is larger than the aforementioned two datasets. Images in it are resized to 224×224 pixels, meanwhile 54 605 sentences are utilized by randomly duplicating existing descriptions.

4) NWPU-Captions: NWPU-Captions is provided by Wuhan University and Huazhong University of Science and Technology based on the NWPU-RESISC45 dataset. It incorporates 45 different labels with each one including 700 instances. Each image is described by five sentences according to certain annotated rules and the total number of descriptions is 157 500. This dataset is challenging due to its large scale and big variations.

We use the criteria R@K (K = 1, 5, 10) to evaluate the performance

of different approaches. Larger R@K indicates better performance.

Experimental settings: In the training process, we set the batch size to 16 and the learning rate to 0.0005 which decreases by 0.7 after every 20 epochs. Totally, 120 epochs are conducted. The margin threshold δ in the loss function is set to 0.2. The visual feature of image region is of 2048-dimensional while the word feature is of 300-dimensional. The hidden dimension of Bi-GRU is 2048. During training, word features are initialized randomly and fed to Bi-GRU.

Results and analysis: We conduct experiments on the four benchmark datasets and Tables 1−4 report the experimental results of various methods including representative cross-modal models for natural images like IMRAM [1], SCAN [2], MTFN [3], PFAN [4] and latest models for RS data like FCLM [8], AMFMN [9] and LW-MCR [10].

It can be seen from Tables 1−4 that generally MAFA-Net achieves better retrieval performance than other models on four datasets. Although, on the first three datasets, MAFA-Net occasionally slightly underperforms others on some metrics. This might be related to the relatively small amount of data in the UCM-Captions dataset and the Sydney-Captions dataset, and the unbalanced distribution of data categories in the Sydney-Captions dataset itself. However, on the much larger and challenging NWPU-Captions dataset, MAFA-Net achieves best on all evaluation metrics. The results of MAFA-Net on four different datasets also demonstrate its robustness.

Table 1. Comparative Experimental Results on UCM-Captions

Method	Image retrieval (T2I)			Sentence retrieval (I2T)		
	R@1	R@5	R@10	R@1	R@5	R@10
IMRAM [1]	11.6	36.2	60.5	12.2	36.2	65.2
SCAN [2]	12.8	45.2	69.5	12.4	46.8	91.9
MTFN [3]	**14.1**	52.3	78.9	10.4	47.4	64.2
PFAN [4]	10.1	28.6	53.8	11.5	38.1	70.0
AMFMN [9]	12.86	**53.24**	79.43	16.67	45.71	68.57
LW-MCR [10]	13.14	50.38	79.52	**18.10**	47.14	63.81
MAFA-Net	10.3	48.2	**80.1**	14.5	**56.1**	**95.7**

Table 2. Comparative Experimental Results on Sydney-Captions

Method	Image retrieval (T2I)			Sentence retrieval (I2T)		
	R@1	R@5	R@10	R@1	R@5	R@10
IMRAM [1]	9.8	45.1	56.8	10.9	50.2	66.1
SCAN [2]	6.2	33.5	51.0	20.6	53.4	67.2
MTFN [3]	13.7	55.5	77.5	20.6	51.7	68.9
PFAN [4]	14.0	51.3	61.9	21.8	49.6	68.5
AMFMN [9]	14.83	56.55	77.89	**24.14**	51.72	75.86
LW-MCR [10]	**15.52**	58.28	80.34	20.69	60.34	**77.59**
MAFA-Net	13.1	**61.4**	**81.9**	22.3	**60.5**	76.4

We also conduct ablation experiments to evaluate the contribution of multi-attention fusion module (MA) and fine-grained alignment module (FA) to MAFA-Net. Table 5 reports the results on NWPU-Captions, in which _nMA_nFA means the basic network without the two modules, _nMA means the network without MA module and _nFA means the network without FA module. It can be seen that the two modules can significantly improve the retrieval performance of the MAFA-Net separately, while their contributions are relatively close. Table 5 also tabulates the training and testing time for executing different models on NWPU-Captions.

We further show the visualization results of our MAFA-Net in Figs. 3−6.

It can be seen that most of the retrieval results match the input, which indicates that the MAFA-Net proposed in this letter can maintain a good semantic correspondence between RS images and sentences. It is worth mentioning that even for the challenging high-density scenes with a great of small and clustered objects, MAFA-Net still performs well (see Fig. 6).

Conclusion: In this letter, we propose a multi-attention fusion and fine-grained alignment network (MAFA-Net) to conduct the cross-modal image-sentence retrieval task in the remote sensing domain. MAFA-Net aims at addressing the properties of multiscale properties and the problem of semantic ambiguity existed in cross-modal retrieval of RS data. Specifically, we design a multi-attention fusion module to improve the feature representation ability. Meanwhile, a fine-grained alignment module is designed to make the information between two different modalities (e.g., visual and textural) interact. Besides the three public available benchmark datasets, a much larger captioning dataset, NWPU-Captions, is utilized to evaluate the performance of MAFA-Net. Experimental results prove that MAFA-Net outperforms current approaches and even for challenging high-density scenes, MAFA-Net can get satisfying results. In the future, we would like to consider more modalities like LiDAR or multi-spectral images and domain adaption [19] for RS visual applications.

Table 3. Comparative Experimental Results on RSICD

Method	Image retrieval (T2I)			Sentence retrieval (I2T)		
	R@1	R@5	R@10	R@1	R@5	R@10
SCAN [2]	7.6	25.0	41.4	6.2	24.8	46.8
IMRAM [1]	7.1	9.5	26.7	4.8	17.1	35.2
MTFN [3]	2.4	7.1	33.8	3.8	21.4	39.0
PFAN [4]	4.8	14.3	29.0	4.3	18.1	37.1
AMFMN [9]	4.99	18.28	31.44	5.39	15.08	23.40
LW-MCR [10]	4.30	18.85	32.34	4.39	13.35	20.29
FCLM [8]	9.11	31.61	**50.00**	11.27	**37.94**	54.41
MAFA-Net	**12.9**	**32.4**	47.6	**12.3**	35.7	**54.8**

Table 4. Comparative Experimental Results on NWPU-Captions

Method	Image retrieval (T2I)			Sentence retrieval (I2T)		
	R@1	R@5	R@10	R@1	R@5	R@10
SCAN [2]	9.5	35.7	63.2	12.6	40.1	81.6
IMRAM [1]	7.1	25.3	51.1	4.3	31.0	65.3
MTFN [3]	5.2	21.8	56.1	11.9	37.0	74.2
PFAN [4]	7.6	22.9	47.5	11.2	32.6	71.0
MAFA-Net	**13.5**	**39.4**	**67.1**	**15.2**	**46.0**	**83.5**

Table 5. Ablation Experimental Results on NWPU-Captions

Method	Image retrieval			Sentence retrieval			Training time (m)	Testing time (s)
	R@1	R@5	R@10	R@1	R@5	R@10		
_nMA _nFA	8.2	29.7	57.2	9.6	31.5	61.3	260	27.62
_nFA	10.4	35.3	64.8	12.7	40.6	75.6	320	49.25
_nMA	12.5	37.6	64.1	13.2	42.9	76.2	340	54.68
MAFA -Net	**13.5**	**39.4**	**67.1**	**15.2**	**46.0**	**83.5**	410	67.43

Acknowledgments: This work was supported by the National Natural Science Foundation of China (42090012), Special Research and 5G Project of Jiangxi Province in China (20212ABC03A09), Guangdong-Macao Joint Innovation Project (2021A0505080008), Key R & D Project of Sichuan Science and Technology Plan (2022YFN0031), and Zhuhai Industry University Research Cooperation Project of China (ZH22017001210098PWC).

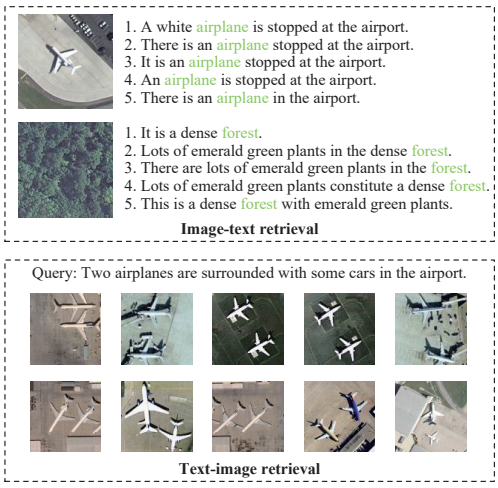

Fig. 3. Visualization results of MAFA-Net on UCM-Captions.

Fig. 4. Visualization results of MAFA-Net on Sydney-Captions.

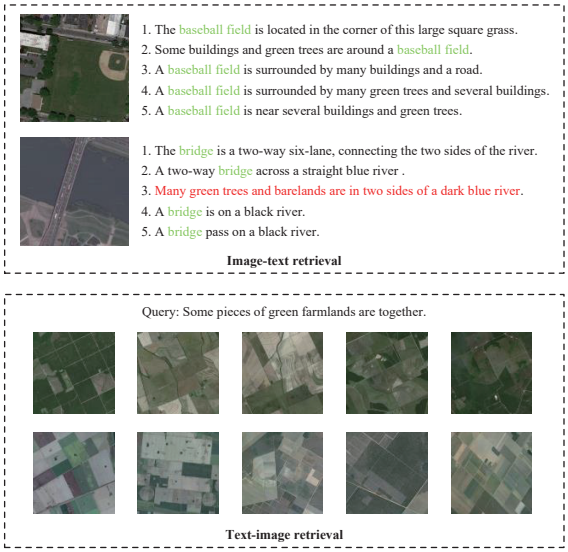

Fig. 5. Visualization results of MAFA-Net on RSICD.

Fig. 6. Visualization results of MAFA-Net on NWPU-Captions.

References

[1] H. Chen, G. Ding, X. Liu, Z. Lin, J. Liu, and J. Han, "IMRAM: Iterative matching with recurrent attention memory for cross-modal image-text retrieval," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Jun. 2020, pp. 12652−12660.

[2] K. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proc. 15th European Conf. Computer Vision*, Sep. 2018, pp. 201−216.

[3] T. Wang, X. Xu, Y. Yang, A. Hanjalic, H. Shen, and J. Song, "Matching images and text with multi-modal tensor fusion and re-ranking," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 12-20.

[4] Y. Wang, H. Yang, X. Qian, L. Ma, and X. Fan, " Position focused attention network for image-text matching," in *Proc. 28th Int. Joint Conf. Artificial Intelligence*, Aug. 2019, pp. 3792−3798.

[5] G. Wu, J. Han, Z. Lin, G. Ding, B. Zhang, and Q. Ni, "Joint image-text hashing for fast large-scale cross-media retrieval using self-supervised deep learning," *IEEE Trans. Industrial Electronics* , vol. 66, no. 12, pp. 9868–9877, Dec. 2019.

[6] Abdullah, Ba zi, Rahhal A, *et al*, " TextRS: Deep bidirectional triplet network for matching text to remote sensing images," *Remote Sensing*, vol. 12, no. 3, pp. 405–423, Jan. 2020.

[7] 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 J. Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 4284–4297, Apr. 2021.

[8] Y. Lv, W. Xiong, X. Zhang, and Y. Cui, " Fusion-based correlation learning model for cross-modal remote sensing image retrieval," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, Jan. 2022.

[9] Z. Yuan, W. Zhang, K. Fu, X. Li, C. Deng, H. Wang, and X. Sun, "Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval," *IEEE Trans. Geoscience and Remote Sensing*, vol. 60, p. 4404119, May. 2022.

[10] Z. Yuan, W. Zhang, X. Rong, X. Li, J. Chen, H. Wang, K. Fu, and X. Sun, " A lightweight multi-scale cross-modal text-image retrieval method in remote sensing," *IEEE Trans. Geoscience and Remote Sensing*, vol. 60, p. 5612819, Apr. 2022.

[11] G. Mikriukov, M. Ravanbakhsh, and B. Demir. " Deep unsupervised contrastive hashing for large-scale cross-modal text-image retrieval in remote sensing," arXiv preprint arXiv: 2201.08125v1, Jan. 2022.

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

[13] U. Chaudhuri, B. Banerjee, A. Bhattacharya, and M. Datcu, "Attention-driven cross-modal remote sensing image retrieval", in *Proc. IEEE Int. Geoscience and Remote Sensing Symposium*, 2021, pp. 4783−4786.

[14] U. Chaudhuri, B. Banerjee, A. Bhattacharya, and M. Datcu, "CMIR-NET: A deep learning based model for cross-modal retrieval in remote sensing," *Pattern Recognition Letters*, vol. 131, no. 2, pp. 456–462, 2020.

[15] Y. Li, Y. Zhang, X. Huang, and J. Ma, "Learning source-invariant deep hashing convolutional neural networks for cross-source remote sensing image retrieval," *IEEE Trans. Geoscience and Remote Sensing* , vol. 56, no. 11, pp. 6521–6536, Nov. 2018.

[16] Y. Li, D. Kong, Y. Zhang, Y. Tan, and L. Chen, " Robust deep alignment network with remote sensing knowledge graph for zero-shot and generalized zero-shot remote sensing image scene classification," *ISPRS J. Photogrammetry and Remote Sensing*, vol. 179, pp. 145–158, 2021.

[17] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Trans. Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2183–2195, Apr. 2018.

[18] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *Proc. Int. Conf. Computer, Inform. and Telecomm. Syst.*, pp. 124−128, Jul. 2016.

[19] H. J. Hu, H. S. Wang, Z. Liu, and W. D. Chen, "Domain-invariant similarity activation map contrastive learning for retrieval-based long-term visual localization," *IEEE/CAA J. Autom. Sinica* , vol. 9, no. 2, pp. 313–328, Feb. 2022.