COMPOSED IMAGE RETRIEVAL FOR REMOTE SENSING

Bill Psomas[1], *Ioannis Kakogeorgiou*[1], *Nikos Efthymiadis*[2], *Giorgos Tolias*[2],
Ondřej Chum[2], *Yannis Avrithis*[3], *Konstantinos Karantzalos*[1]

[1]National Technical University of Athens, [2]Czech Technical University in Prague,
[3]Institute of Advanced Research in Artificial Intelligence (IARAI)

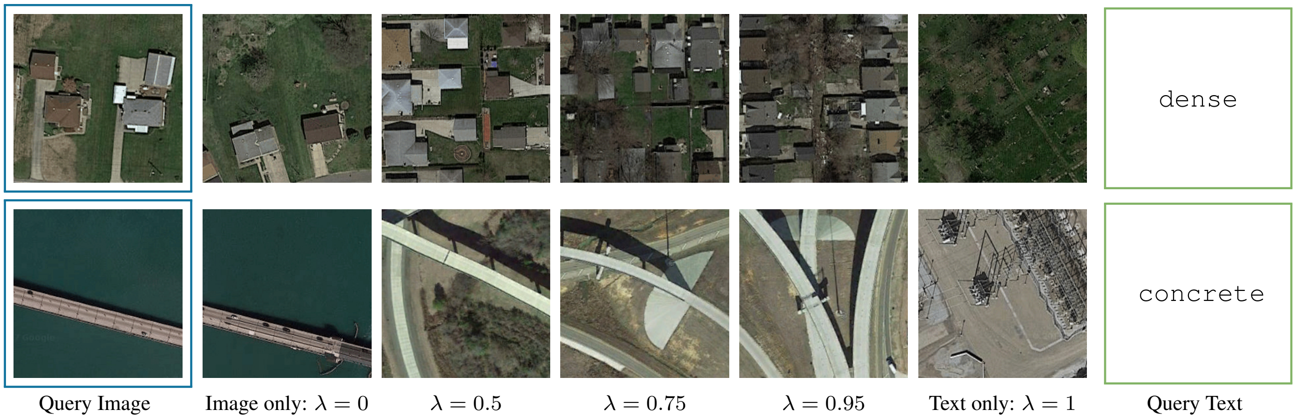


Figure 1: We introduce remote sensing composed image retrieval (RSCIR), a novel and expressive remote sensing image retrieval (RSIR) task integrating both image and text in the search query. We also introduce WeiCom, a flexible, training-free method based on vision-language models, utilizing a weighting parameter λ for more image- or text-oriented results, with $\lambda \to 0$ or $\lambda \to 1$ respectively. For each query image and query text, retrieved images shown for different λ.

ABSTRACT

This work introduces composed image retrieval to remote sensing. It allows to query a large image archive by image examples alternated by a textual description, enriching the descriptive power over unimodal queries, either visual or textual. Various attributes can be modified by the textual part, such as shape, color, or context. A novel method fusing image-to-image and text-to-image similarity is introduced. We demonstrate that a vision-language model possesses sufficient descriptive power and no further learning step or training data are necessary. We present a new evaluation benchmark focused on color, context, density, existence, quantity, and shape modifications. Our work not only sets the state-of-the-art for this task, but also serves as a foundational step in addressing a gap in the field of remote sensing image retrieval. Code at: https://github.com/billpsomas/rscir.

Index Terms— Vision-Language Models, Retrieval

1. INTRODUCTION

In recent years, earth observation (EO) through remote sensing (RS) has witnessed an enormous growth in data volume, creating a challenge in managing and extracting relevant information. The capacity to efficiently organize extensive archives and quickly *retrieve* specific images is crucial.

Remote sensing image retrieval (RSIR) [1], which aims to search and retrieve images from RS image archives, has emerged as a key solution. RSIR methods can be categorized into *unisource* and *cross-source* [2], where the categorization is based on whether the query image and the retrieved images are from the same source. In the case of unisource, there exists *single-label* [3, 4, 5, 6, 7, 8] and *multi-label* [9, 10, 11, 12, 13, 14] retrieval, depending on whether an image is associated with one or multiple labels respectively. In the case of cross-source, the term "source" is used loosely and can correspond to modality, view, etc.

In all cases, RSIR methods encounter a major limitation: the reliance on a query of single modality. This constraint often restricts users from fully expressing their specific requirements, especially given the complex and dynamic nature of Earth's surface as depicted in RS imagery. Ideally, users would benefit from a system that allows them to articulate nuanced modifications or specifications in conjunction with an image-based query. This is where composed image retrieval (CIR) [15, 16, 17, 18, 19, 20] comes into play. CIR, integrating both image and text in the search query, is designed to retrieve images that are not only visually similar to the *query image* but also relevant to the details of the accompanying *query text*. By incorporating CIR into RS, we aim to offer a more expressive and flexible search capability that aligns closely with the needs of users in this field.

In this paper, we recognize, present and qualitatively evaluate the capabilities and challenges that CIR introduces within the RS domain. We demonstrate how users can now pair a query image with a query text specifying modifications related to *color*, *context*, *density*, *existence*, *quantity*, *shape*, *size* or *texture* of one or more classes. Quantitatively, we focus on color, context, density, existence, quantity, and shape modifications, establishing a benchmark and an evaluation protocol. Our approach is training-free by using a frozen vision-language model.

In summary, we make the following contributions:

1. We are the first to introduce composed image retrieval into remote sensing, accompanied with PATTERNCOM, a benchmark dataset.
2. We introduce WEICOM, a training-free method utilizing a modality control parameter for more image- or text-oriented results according to the needs of each search, as shown in Figure 1.
3. We evaluate both qualitatively and quantitatively the performance of WEICOM, setting the state-of-the-art on remote sensing composed image retrieval.

2. RELATED WORK

Remote Sensing Image Retrieval With the aim to effectively *search* and *retrieve* information from extensive RS image archives, remote sensing image retrieval (RSIR) can be categorized into *unisource* and *cross-source* [2]. Initially, RSIR methods focus on handcrafted and low-level visual features [21, 22, 23, 24, 25, 26, 27, 28, 29, 30]. With the advent of deep learning, neural networks are utilized for unisource *single-label* retrieval: (a) as feature extractors [31, 32, 33, 34, 35, 36, 37, 38, 39, 3], (b) trained from scratch [40, 41, 42, 43, 4, 5, 44, 45], (c) integrating attention modules [6, 46, 47, 48] and (d) using metric learning [7, 49, 8, 50, 51, 52]. Neural networks are also used for unisource *multi-label* [29, 9, 10, 11, 12, 13, 14], cross-source *cross-sensors* [53, 54, 55, 56], cross-source *cross-modal* [48, 57, 58, 59, 60, 61, 62] and cross-source *cross-view* retrieval [63, 64, 65, 66, 67, 68]. Our work fills a notable gap and enhances user intent expression in RSIR by combining query image with query text.

Composed Image Retrieval Image-to-image [69, 70, 71] and text-to-image [72, 73, 74] retrieval provide ways to explore large image archives. However, the most accurate and flexible way to express the user intent is a query *composed* of both an image and a text. Composed Image Retrieval (CIR) [15, 16, 17, 18, 19, 20] aims to retrieve images not only visually similar to the query image, but also altered to align with the specifics of the query text. Traditionally, CIR methods are supervised by *triplets* of the form *query image, query text, target image* [15, 75, 76, 17, 77, 16, 19, 78]. The labor-intensive process of labeling such triplets limit early works to specific applications in fashion [79, 80, 81], physical states [82], object attributes and composition [15, 83, 84]. The emergence of vision-language models (VLMs) [85, 86, 87] led to their integration into CIR, introducing *zero-shot composed image retrieval* (ZS-CIR) [20, 88, 89]. This increases the spectrum of possible applications [88]. Methods are trained using unlabeled images [88, 20], or are not trained at all [89]. Recognizing the unexplored potential of CIR in RS, our work pioneers its introduction in this domain, particularly leveraging ZS-CIR empowered by VLMs.

Vision-Language Models The emergence of vision-language models (VLMs) [85, 86, 87, 90] revolutionizes the field of multimodal learning. Trained on large-scale datasets [91], these models map images and text into a shared embedding space. Apart from zero-shot classification, CLIP [85] can be used for detection [92], segmentation [93] and captioning [94]. CLIP can also be aligned to be used with medical data [95] or satellite data [96, 97]. In this work, we leverage CLIP and RemoteCLIP [96], a vision-language model for remote sensing, in a training-free setting.

3. METHOD

3.1. Problem formulation

In composed image retrieval, the goal is to retrieve images based on a *composed image-text query*, that is, a query that consists of a *visual* part, the query image, and a *textual* part, the query text. In this work, we introduce remote sensing composed image retrieval. To do so, we establish a benchmark and an evaluation protocol.

We denote the query image as y, its class as C_y and an attribute of the depicted class as A_y. We also denote the query text as t, which represents a modified target attribute A_t. We refer to the two queries as the composed query, $q = (y, t)$. Given an image dataset X, our goal is to retrieve images from X that share class with the query image class C_y and have the attribute A_t defined by the text query t. Retrieval aims to rank images $x \in X$ with respect to their composed similarity $s(q, x) \in R$ to the query. The task is extendable to multiple classes and multiple attributes.

To define s, we make use of pre-trained VLMs that consist of a *visual encoder* $f : \mathcal{I} \to \mathbb{R}^d$ and a *text encoder* $g : \mathcal{T} \to \mathbb{R}^d$, which map input images from image space $\mathcal{I}$ and words from the text space $\mathcal{T}$ to the same embedding space with dimension d. We extract the visual embedding $\mathbf{v}_y = f(y) \in \mathbb{R}^d$ and the text embedding $\mathbf{v}_t = g(t) \in \mathbb{R}^d$ to use as queries. Finally, the embedding of a dataset image $x \in X$ is denoted as $\mathbf{v}_x = f(x) \in \mathbb{R}^d$. All embeddings are ℓ_2-normalized.

3.2. Baselines

Unimodal baselines rely solely on a single type of query to determine similarity. We denote: *text-only* by $s_g(q, x) =$

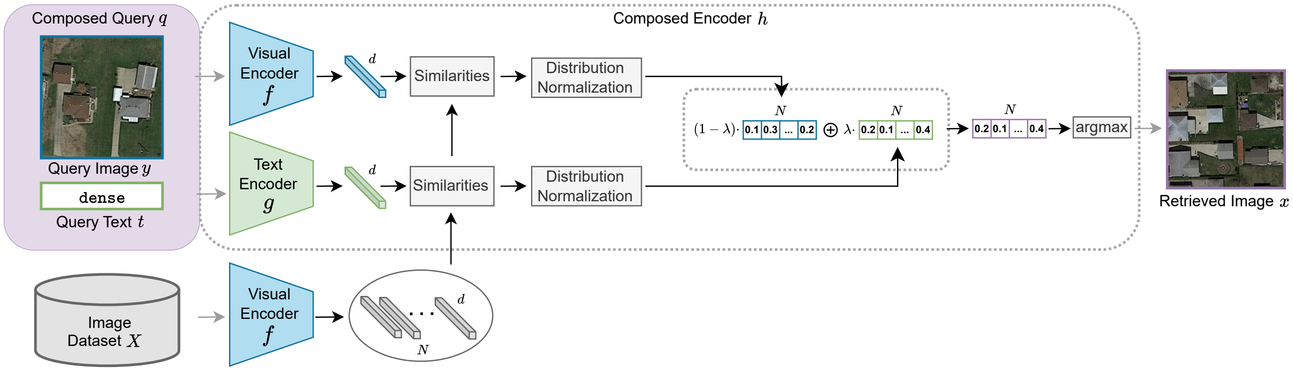


Figure 2: WEICOM*: A* WEI*ghted* COM*posed Image Retrieval Method.* It utilizes a dual-encoder approach to process both query image y and query text t. Initially, the query image is passed into a visual encoder f and the query text into a text encoder g, producing corresponding d-dimensional representations. Subsequently, similarity scores with the representations in the image dataset are calculated. These scores are then normalized and combined using a convex combination controlled by a $\lambda \in [0, 1]$. Finally, an argmax(argsort) operation identifies the most relevant retrieved image(s) x.

$g(t)^T f(x)$ and *image-only* by $s_f(q, x) = f(y)^T f(x)$. Unimodal baselines are expected to fail since the final similarity cannot embody information from both image and text.

Multimodal combines the two unimodal approaches by averaging their similarities:

$$s_a(q, x) = \frac{s_g(q, x) + s_f(q, x)}{2} \tag{1}$$

Note that this baseline is equivalent to averaging the two features $g(t), f(y)$ and then calculating the similarities once. The drawback of this approach is that the features that come from same modalities have similarities significantly greater than the cross-modal similarities, making it an approach biased in favor of the image query.

3.3. WEICOM

In our proposed method, WEICOM, we estimate the similarities of the image query $s_f(q, x)$ and the text query $s_g(q, x)$ with the database. Then we perform similarity normalization in order to have a starting point of equal contribution from both modalities and we notate $s'_f(q, x), s'_g(q, x)$. Finally, we use the weighted average of the two similarity sets using a modality control parameter λ:

$$s_{WC}(q, x) = \lambda s'_g(q, x) + (1 - \lambda) s'_f(q, x) \tag{2}$$

Similarity Normalization In order to ensure that both image and text queries contribute equally to the retrieval, we normalize their similarities with the database. We first transform the empirical distribution of similarity scores into a standard normal distribution. Subsequently, we apply the cumulative distribution function (CDF) of the standard normal distribution to the standardized data, resulting in values that range between 0 and 1. Assuming the standardized data adhere to a normal distribution, this transformation yields data that approximates a uniform distribution. Transforming data into a uniform distribution diminishes the influence of outliers and reduces skewness, smoothing any excessively peaked distributions. This approach leads to more robust similarity scores.

The modality control parameter λ After normalizing the similarities, we can control the influence of each modality using a parameter λ as a weight. Here $\lambda = 0$ refers to image-only retrieval, $\lambda = 1$ to text-only retrieval and $\lambda = 0.5$ to equal contribution of image and text, as shown in Figure 1. The full WEICOM method is summarized in Figure 2.

4. EXPERIMENTS

4.1. Datasets, networks and evaluation protocol

Datasets To evaluate quantitatively the methods, we introduce PATTERNCOM, a new benchmark based on PatternNet [98]. PatternNet is a large-scale high-resolution remote sensing image retrieval dataset. There are 38 classes and each class has 800 images of size 256×256 pixels. In PATTERNCOM, we select some classes to be depicted in query images, and add a query text that defines an attribute relevant to that class. For instance, query images of "swimming pools" are combined with text queries defining "shape" as "rectangular", "oval", and "kidney-shaped". In total, PATTERNCOM includes six attributes consisted of up to four different classes each. Each attribute can be associated with two to five values per class. The number of positives ranges from 2 to 1345 and there are more than 21k queries in total. Statistics for two out of six attributes are shown in Table 1.

Networks We use the pre-trained CLIP [85] and RemoteCLIP [96], both with a ViT-L/14 image encoder.

Evaluation Protocol We evaluate using mAP. Average Precision (AP) is the average of the precision values obtained for the set of top-k results, up to each relevant item found in the ranking. The mAP is then the mean of these AP values over all queries.

ATTRIBUTE	CLASS	VALUE	#POSITIVES	#QUERIES
color	airplane	white	672	53
		purple	53	672
	nursing home	white	85	383
		gray	383	85
	crosswalk	white	412	388
		yellow	388	412
	tennis court	blue	339	287
		brown	2	624
		gray	50	576
		green	211	415
		red	24	602
shape	swimming pool	rectangular	261	299
		oval	52	508
		kidney-shaped	247	313
	river	curved	177	623
		straight	623	177
	road	cross	800	800
		round	800	800

Table 1: *Statistics for color and shape attributes of* PATTERNCOM, the first RSCIR benchmark.

4.2. Experimental results

Qualitative results In Figure 3, we present the qualitative results of performing composed image retrieval in PATTERNCOM using WEICOM with RemoteCLIP. Each example corresponds to one of the selected attributes with the query text specifying a modification in each attribute value.

Comparison with baselines As shown in Table 2, WEICOM outperfoms both unimodal ("Text", "Image") and multimodal ("Text & Image") baselines by a large margin. In particular, it outperforms the second best by 8.95% mAP using CLIP and 15.14% mAP using RemoteCLIP on average.

(a) CLIP [85]

METHOD	COLOR	CONTEXT	DENSITY	EXISTENCE	QUANTITY	SHAPE	AVG
Text	13.47	4.83	3.58	4.38	3.31	6.22	5.97
Image	14.66	8.32	13.49	13.50	7.84	15.76	12.26
Text & Image	23.13	11.02	15.87	**13.77**	10.13	21.38	15.88
WEICOM$_{\lambda=0.5}$	46.08	17.45	16.49	9.24	18.15	23.97	21.90
WEICOM$_{\lambda=0.3}$	**46.74**	**20.97**	**22.07**	12.07	**20.96**	**26.22**	**24.83**

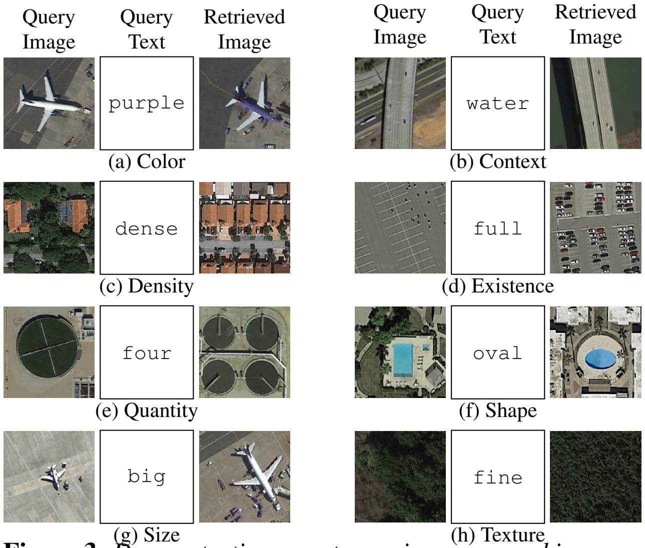
(b) RemoteCLIP [96]

METHOD	COLOR	CONTEXT	DENSITY	EXISTENCE	QUANTITY	SHAPE	AVG
Text	10.75	8.87	22.16	12.49	8.25	24.12	14.44
Image	14.40	6.62	15.11	9.29	6.99	15.18	11.27
Text & Image	23.67	10.01	18.45	10.56	7.97	19.63	15.05
WEICOM$_{\lambda=0.5}$	**43.68**	31.45	39.94	14.27	20.51	29.78	29.94
WEICOM$_{\lambda=0.6}$	41.04	**31.59**	**41.56**	**14.79**	**20.79**	**31.24**	**30.19**

Table 2: *Attribute modification mAP (%)* on PATTERNCOM using CLIP (a) and RemoteCLIP (b); comparison of WEICOM with baselines. For each attribute value of an attribute (e.g. "rectangular" of SHAPE), average mAP over all the rest attribute values (e.g. "oval" of SHAPE). AVG: average mAP over all combinations.

4.3. Ablation study

The impact of λ In Table 3 we show the impact of modality control parameter λ on WEICOM using RemoteCLIP. $\lambda = 0$ refers to image-only, $\lambda = 1$ to text-only retrieval. For $\lambda = 0.6$ we get the best average mAP, thus we set this as our method's default. The same study for CLIP gives $\lambda = 0.3$.



(a) Color (b) Context (c) Density (d) Existence (e) Quantity (f) Shape (g) Size (h) Texture

Figure 3: *Demonstrating remote sensing composed image retrieval.* Subfigures (a) to (h) depict the key attributes: color, context, density, existence, quantity, shape, size, and texture. Each one illustrates various utilizations of composed image retrieval in remote sensing. Subfigures (b), (d) are examples that extend the task to multiple classes and attributes.

λ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Color	14.5	55.3	53.0	49.6	46.4	43.7	41.0	38.2	35.0	30.4	10.8
Context	6.6	13.3	20.2	25.7	29.5	31.5	31.6	29.6	24.8	16.9	8.9
Density	15.1	23.3	29.5	34.0	37.4	39.9	41.6	42.0	40.7	35.9	22.2
Existence	9.3	10.3	11.1	12.3	13.5	14.3	14.8	15.0	14.8	14.0	12.5
Quantity	7.0	17.6	18.9	19.7	20.2	20.5	20.8	20.9	20.8	20.1	8.3
Shape	15.2	23.8	24.7	26.2	28.0	29.8	31.2	32.0	32.0	31.3	24.1
Average	11.3	23.9	26.2	27.9	29.2	29.9	30.2	29.6	28.0	24.8	14.4

Table 3: *The effect of the modality control parameter λ on* WEICOM using RemoteCLIP, measured in attribute modification mAP.

5. CONCLUSIONS

We introduce remote sensing composed image retrieval, a novel task integrating both image and text in the search query, accompanied with PATTERNCOM, a benchmark dataset. We demonstrate its versatility through use cases modifying attributes like color or shape and also introduce WEICOM, a flexible and training-free method utilizing a modality control parameter λ, setting the state-of-the-art on the task.

Acknowledgements Bill was supported by the RAMONES H2020 project (grant: 101017808). This work was also supported by the HFRI under the BiCUBES project (grant: 03943), by the Czech Technical University in Prague grant No. SGS23/173/OHK3/3T/13, the Junior Star GACR GM 21-28830M, and the CTU institutional support (Future fund). NTUA thanks NVIDIA for the donation of GPU hardware.

6. REFERENCES

[1] Peggy Agouris, James Carswell, and Anthony Stefanidis, "An environment for content-based image retrieval from large spatial databases," *ISPRS Journal of Photogrammetry and Remote Sensing*, 1999. 1

[2] Weixun Zhou, Haiyan Guan, Ziyu Li, Zhenfeng Shao, and Mahmoud R Delavar, "Remote sensing image retrieval in the past decade: Achievements, challenges, and future directions," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023. 1, 2

[3] Dongyang Hou, Zelang Miao, Huaqiao Xing, and Hao Wu, "Exploiting low dimensional features from the mobilenets for remote sensing image retrieval," *Earth Science Informatics*, vol. 13, pp. 1437–1443, 2020. 1, 2

[4] Yameng Wang, Shunping Ji, and Yongjun Zhang, "A learnable joint spatial and spectral transformation for high resolution remote sensing image retrieval," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 8100–8112, 2021. 1, 2

[5] Gencer Sumbul and Begüm Demir, "Plasticity-stability preserving multi-task learning for remote sensing image retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022. 1, 2

[6] Siyuan Wang, Dongyang Hou, and Huaqiao Xing, "A novel multi-attention fusion network with dilated convolution and label smoothing for remote sensing image retrieval," *International Journal of Remote Sensing*, vol. 43, no. 4, pp. 1306–1322, 2022. 1, 2

[7] Hongwei Zhao, Lin Yuan, Haoyu Zhao, and Zhen Wang, "Global-aware ranking deep metric learning for remote sensing image retrieval," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021. 1, 2

[8] Qimin Cheng, Deqiao Gan, Peng Fu, Haiyan Huang, and Yuzhuo Zhou, "A novel ensemble architecture of residual attention-based deep metric learning for remote sensing image retrieval," *Remote Sensing*, vol. 13, no. 17, pp. 3445, 2021. 1, 2

[9] Jian Kang, Ruben Fernandez-Beltran, Danfeng Hong, Jocelyn Chanussot, and Antonio Plaza, "Graph relation network: Modeling relations between scenes for multilabel remote-sensing image classification and retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 4355–4369, 2020. 1, 2

[10] Gencer Sumbul and Begüm Demir, "A novel graph-theoretic deep representation learning method for multilabel remote sensing image retrieval," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. IEEE, 2021, pp. 266–269. 1, 2

[11] Gencer Sumbul, Mahdyar Ravanbakhsh, and Begüm Demir, "Informative and representative triplet selection for multilabel remote sensing image retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2021. 1, 2

[12] Qimin Cheng, Haiyan Huang, Lan Ye, Peng Fu, Deqiao Gan, and Yuzhuo Zhou, "A semantic-preserving deep hashing model for multi-label remote sensing image retrieval," *Remote Sensing*, vol. 13, no. 24, pp. 4965, 2021. 1, 2

[13] Raffaele Imbriaco, Clint Sebastian, Egor Bondarev, and Peter HN de With, "Toward multilabel image retrieval for remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021. 1, 2

[14] Zhenfeng Shao, Weixun Zhou, Xueqing Deng, Maoding Zhang, and Qimin Cheng, "Multilabel remote sensing image retrieval based on fully convolutional network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 318–328, 2020. 1, 2

[15] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays, "Composing text and image for image retrieval-an empirical odyssey," in *CVPR*, 2019. 1, 2

[16] Yanbei Chen, Shaogang Gong, and Loris Bazzani, "Image search with text feedback by visiolinguistic attention learning," in *CVPR*, 2020. 1, 2

[17] Mehrdad Hosseinzadeh and Yang Wang, "Composed query image retrieval using locally bounded features," in *CVPR*, 2020. 1, 2

[18] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo, "Effective conditioned and composed image retrieval combining clip-based features," in *CVPR*, 2022. 1, 2

[19] Seungmin Lee, Dongwan Kim, and Bohyung Han, "Cosmo: Content-style modulation for image retrieval with text feedback," in *CVPR*, 2021. 1, 2

[20] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister, "Pic2word: Mapping pictures to words for zero-shot composed image retrieval," in *CVPR*, 2023. 1, 2

[21] YN Mamatha and AG Ananth, "Content based image retrieval of satellite imageries using soft query based color

composite techniques," *International Journal of Computer Applications*, vol. 7, no. 5, pp. 0975–8887, 2010. 2

[22] Caihong Ma, Qin Dai, Jianbo Liu, Shibin Liu, and Jin Yang, "An improved svm model for relevance feedback in remote sensing image retrieval," *International Journal of Digital Earth*, vol. 7, no. 9, pp. 725–745, 2014. 2

[23] Jose A Piedra-Fernandez, Gloria Ortega, James Z Wang, and Manuel Canton-Garbin, "Fuzzy content-based image retrieval for oceanic remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 9, pp. 5422–5431, 2013. 2

[24] Jiang Li and Ram M Narayanan, "Integrated spectral and spatial information mining in remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 3, pp. 673–685, 2004. 2

[25] Sitaram Bhagavathy and Bangalore S Manjunath, "Modeling and detection of geospatial objects using texture motifs," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 12, pp. 3706–3715, 2006. 2

[26] Min Wang and Tengyi Song, "Remote sensing image retrieval by scene semantic matching," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 5, pp. 2874–2886, 2012. 2

[27] ZF Shao, WX Zhou, and QM Cheng, "Remote sensing image retrieval with combined features of salient region," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 40, pp. 83–88, 2014. 2

[28] M Wang, QM Wan, LB Gu, and TY Song, "Remote-sensing image retrieval by combining image visual and semantic features," *International journal of remote sensing*, vol. 34, no. 12, pp. 4200–4223, 2013. 2

[29] Bindita Chaudhuri, Begüm Demir, Subhasis Chaudhuri, and Lorenzo Bruzzone, "Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 1144–1158, 2017. 2

[30] Osman Emre Dai, Begüm Demir, Bülent Sankur, and Lorenzo Bruzzone, "A novel system for content based retrieval of multi-label remote sensing images," in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2017, pp. 1744–1747. 2

[31] Yansheng Li, Yongjun Zhang, Chao Tao, and Hu Zhu, "Content-based high-resolution remote sensing image retrieval via unsupervised feature learning and collaborative affinity metric fusion," *Remote Sensing*, vol. 8, no. 9, pp. 709, 2016. 2

[32] Fan Hu, Xinyi Tong, Gui-Song Xia, and Liangpei Zhang, "Delving into deep representations for remote sensing image retrieval," in *2016 IEEE 13th International Conference on Signal Processing (ICSP)*. IEEE, 2016, pp. 198–203. 2

[33] Yaakoub Boualleg and Mohamed Farah, "Enhanced interactive remote sensing image retrieval with scene classification convolutional neural networks model," in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2018, pp. 4748–4751. 2

[34] Famao Ye, Hui Xiao, Xuqing Zhao, Meng Dong, Wei Luo, and Weidong Min, "Remote sensing image retrieval using convolutional neural network features and weighted distance," *IEEE geoscience and remote sensing letters*, vol. 15, no. 10, pp. 1535–1539, 2018. 2

[35] Paolo Napoletano, "Visual descriptors for content-based retrieval of remote-sensing images," *International journal of remote sensing*, vol. 39, no. 5, pp. 1343–1376, 2018. 2

[36] Yun Ge, Shunliang Jiang, Qingyong Xu, Changlong Jiang, and Famao Ye, "Exploiting representations from pre-trained convolutional neural networks for high-resolution remote sensing image retrieval," *Multimedia Tools and Applications*, vol. 77, pp. 17489–17515, 2018. 2

[37] Xu Tang, Xiangrong Zhang, Fang Liu, and Licheng Jiao, "Unsupervised deep feature learning for remote sensing image retrieval," *Remote Sensing*, vol. 10, no. 8, pp. 1243, 2018. 2

[38] Raffaele Imbriaco, Clint Sebastian, Egor Bondarev, and Peter HN de With, "Aggregated deep local features for remote sensing image retrieval," *Remote Sensing*, vol. 11, no. 5, pp. 493, 2019. 2

[39] Pouria Sadeghi-Tehran, Plamen Angelov, Nicolas Virlet, and Malcolm J Hawkesford, "Scalable database indexing and fast image retrieval based on deep learning and hierarchically nested structure applied to remote sensing and plant biology," *Journal of Imaging*, vol. 5, no. 3, pp. 33, 2019. 2

[40] Weixun Zhou, Shawn Newsam, Congmin Li, and Zhenfeng Shao, "Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval," *Remote Sensing*, vol. 9, no. 5, pp. 489, 2017. 2

[41] Maoding Zhang, Qimin Cheng, Fang Luo, and Lan Ye, "A triplet nonlocal neural network with dual-anchor triplet loss for high-resolution remote sensing image retrieval," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 2711–2723, 2021. 2

[42] Zheng Zhuo and Zhong Zhou, "Remote sensing image retrieval with gabor-ca-resnet and split-based deep feature transform network," *Remote Sensing*, vol. 13, no. 5, pp. 869, 2021. 2

[43] Yishu Liu, Yingbin Liu, Conghui Chen, and Liwang Ding, "Remote-sensing image retrieval with tree-triplet-classification networks," *Neurocomputing*, vol. 405, pp. 48–61, 2020. 2

[44] Yuebin Wang, Liqiang Zhang, Xiaohua Tong, Liang Zhang, Zhenxin Zhang, Hao Liu, Xiaoyue Xing, and P Takis Mathiopoulos, "A three-layered graph-based learning approach for remote sensing image retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 6020–6034, 2016. 2

[45] Ushasi Chaudhuri, Biplab Banerjee, and Avik Bhattacharya, "Siamese graph convolutional network for content based remote sensing image retrieval," *Computer vision and image understanding*, 2019. 2

[46] Yameng Wang, Shunping Ji, Meng Lu, and Yongjun Zhang, "Attention boosted bilinear pooling for remote sensing image retrieval," *International Journal of Remote Sensing*, vol. 41, no. 7, pp. 2704–2724, 2020. 2

[47] Wei Xiong, Yafei Lv, Yaqi Cui, Xiaohan Zhang, and Xiangqi Gu, "A discriminative feature learning approach for remote sensing image retrieval," *Remote Sensing*, vol. 11, no. 3, pp. 281, 2019. 2

[48] Ushasi Chaudhuri, Biplab Banerjee, Avik Bhattacharya, and Mihai Datcu, "Attention-driven graph convolution network for remote sensing image retrieval," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021. 2

[49] Rui Cao, Qian Zhang, Jiasong Zhu, Qing Li, Qingquan Li, Bozhi Liu, and Guoping 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. 2

[50] Yishu Liu, Zhengzhuo Han, Conghui Chen, Liwang Ding, and Yingbin Liu, "Eagle-eyed multitask cnns for aerial image retrieval and scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 9, pp. 6699–6721, 2020. 2

[51] Lili Fan, Hongwei Zhao, and Haoyu Zhao, "Global optimization: Combining local loss with result ranking loss in remote sensing image retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 8, pp. 7011–7026, 2020. 2

[52] Yishu Liu, Liwang Ding, Conghui Chen, and Yingbin Liu, "Similarity-based unsupervised deep transfer learning for remote sensing image retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 11, pp. 7872–7889, 2020. 2

[53] Yansheng Li, Yongjun Zhang, Xin Huang, and Jiayi Ma, "Learning source-invariant deep hashing convolutional neural networks for cross-source remote sensing image retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 11, pp. 6521–6536, 2018. 2

[54] Jingjing Ma, Duanpeng Shi, Xu Tang, Xiangrong Zhang, Xiao Han, and Licheng Jiao, "Cross-source image retrieval based on ensemble learning and knowledge distillation for remote sensing images," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. IEEE, 2021, pp. 2803–2806. 2

[55] Wei Xiong, Zhenyu Xiong, Yaqi Cui, and Yafei Lv, "A discriminative distillation network for cross-source remote sensing image retrieval," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 1234–1247, 2020. 2

[56] Wei Xiong, Yafei Lv, Xiaohan Zhang, and Yaqi Cui, "Learning to translate for cross-source remote sensing image retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 7, pp. 4860–4874, 2020. 2

[57] Fang Xu, Wen Yang, Tianbi Jiang, Shijie Lin, Hao Luo, and Gui-Song Xia, "Mental retrieval of remote sensing images via adversarial sketch-image feature learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 11, pp. 7801–7814, 2020. 2

[58] Yuxi Sun, Shanshan Feng, Yunming Ye, Xutao Li, Jian Kang, Zhichao Huang, and Chuyao Luo, "Multisensor fusion and explicit semantic preserving-based deep hashing for cross-modal remote sensing image retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021. 2

[59] Gencer Sumbul, Arne De Wall, Tristan Kreuziger, Filipe Marcelino, Hugo Costa, Pedro Benevides, Mario Caetano, Begüm Demir, and Volker Markl, "Bigearthnet-mm: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 3, pp. 174–180, 2021. 2

[60] Yafei Lv, Wei Xiong, Xiaohan Zhang, and Yaqi Cui, "Fusion-based correlation learning model for cross-modal remote sensing image retrieval," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021. 2

[61] Zhiqiang Yuan, Wenkai Zhang, Changyuan Tian, Xuee Rong, Zhengyuan Zhang, Hongqi Wang, Kun Fu, and Xian Sun, "Remote sensing cross-modal text-image retrieval based on global and local information," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022. 2

[62] Zhiqiang Yuan, Wenkai Zhang, Kun Fu, Xuan Li, Chubo Deng, Hongqi Wang, and Xian Sun, "Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval," *arXiv preprint arXiv:2204.09868*, 2022. 2

[63] Sixing Hu, Mengdan Feng, Rang MH Nguyen, and Gim Hee Lee, "Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7258–7267. 2

[64] Zelong Zeng, Zheng Wang, Fan Yang, and Shin'ichi Satoh, "Geo-localization via ground-to-satellite cross-view image retrieval," *IEEE Transactions on Multimedia*, 2022. 2

[65] Yuxin Tian, Xueqing Deng, Yi Zhu, and Shawn Newsam, "Cross-time and orientation-invariant overhead image geolocalization using deep local features," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 2512–2520. 2

[66] Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays, "Learning deep representations for ground-to-aerial geolocalization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5007–5015. 2

[67] Numan Khurshid, Talha Hanif, Mohbat Tharani, and Murtaza Taj, "Cross-view image retrieval-ground to aerial image retrieval through deep learning," in *International Conference on Neural Information Processing*. Springer, 2019, pp. 210–221. 2

[68] Yujiao Shi and Hongdong Li, "Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17010–17020. 2

[69] Filip Radenović, Giorgos Tolias, and Ondřej Chum, "Fine-tuning cnn image retrieval with no human annotation," *PAMI*, 2019. 2

[70] Albert Gordo, Jon Almazan, Jerome Revaud, and Diane Larlus, "Deep image retrieval: Learning global representations for image search," in *ECCV*, 2016. 2

[71] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han, "Large-scale image retrieval with attentive deep local features," in *ICCV*, 2017. 2

[72] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris, "Adversarial representation learning for text-to-image matching," in *ICCV*, 2019, pp. 5814–5824. 2

[73] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z Li, "Context-aware attention network for image-text retrieval," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3536–3545. 2

[74] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov, "Devise: A deep visual-semantic embedding model," *NeurIPS*, vol. 26, 2013. 2

[75] Yanbei Chen and Loris Bazzani, "Learning joint visual semantic matching embeddings for language-guided retrieval," in *ECCV*, 2020. 2

[76] Minghao Yin, Zhuliang Yao, Yue Cao, Xiu Li, Zheng Zhang, Stephen Lin, and Han Hu, "Disentangled non-local neural networks," in *ECCV*, 2020. 2

[77] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *NeurIPS*, 2015. 2

[78] Ginger Delmas, Rafael Sampaio de Rezende, Gabriela Csurka, and Diane Larlus, "Artemis: Attention-based retrieval with text-explicit matching and implicit similarity," 2022. 2

[79] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis, "Automatic spatially-aware fashion concept discovery," in *ICCV*, 2017. 2

[80] Tamara L Berg, Alexander C Berg, and Jonathan Shih, "Automatic attribute discovery and characterization from noisy web data," in *ECCV*. Springer, 2010. 2

[81] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris, "Fashion iq: A new dataset towards retrieving images by natural language feedback," in *CVPR*, 2021. 2

[82] Phillip Isola, Joseph J Lim, and Edward H Adelson, "Discovering states and transformations in image collections," in *CVPR*, 2015, pp. 1383–1391. 2

[83] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, “Microsoft coco: Common objects in context,” in *ECCV*, 2014. 2

[84] Andrei Neculai, Yanbei Chen, and Zeynep Akata, “Probabilistic compositional embeddings for multimodal image retrieval,” in *CVPR*, 2022. 2

[85] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., “Learning transferable visual models from natural language supervision,” in *ICML*. PMLR, 2021. 2, 3, 4

[86] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig, “Scaling up visual and vision-language representation learning with noisy text supervision,” in *ICML*, 2021. 2

[87] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi, “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in *ICML*, 2022. 2

[88] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo, “Zero-shot composed image retrieval with textual inversion,” in *ICCV*, 2023. 2

[89] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata, “Vision-by-language for training-free compositional image retrieval,” 2023, arXiv. 2

[90] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” *arXiv preprint arXiv:2301.12597*, 2023. 2

[91] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al., “Laion-5b: An open large-scale dataset for training next generation image-text models,” *NeurIPS*, 2022. 2

[92] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui, “Open-vocabulary object detection via vision and language knowledge distillation,” *arXiv preprint arXiv:2104.13921*, 2021. 2

[93] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu, “Open-vocabulary semantic segmentation with mask-adapted clip,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7061–7070. 2

[94] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf, “Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17918–17928. 2

[95] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun, “Medclip: Contrastive learning from unpaired medical images and text,” *arXiv preprint arXiv:2210.10163*, 2022. 2

[96] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, and Jun Zhou, “Remoteclip: A vision language foundation model for remote sensing,” *arXiv preprint arXiv:2306.11029*, 2023. 2, 3, 4

[97] Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm, “Satclip: Global, general-purpose location embeddings with satellite imagery,” *arXiv preprint arXiv:2311.17179*, 2023. 2

[98] Weixun Zhou, Shawn Newsam, Congmin Li, and Zhenfeng Shao, “Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval,” *ISPRS journal of photogrammetry and remote sensing*, vol. 145, pp. 197–209, 2018. 3