# Multi-Branch Convolutional Descriptors for Content-based Remote Sensing Image Retrieval

Raffaele Imbriaco[a], Tunc Alkanat[a], Egor Bondarev and Peter H. N. de With
*Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven 5612AZ, The Netherlands*

Keywords: Content-based Image Retrieval, Remote Sensing, Convolutional Neural Networks, Local Feature Extraction.

Abstract: Context-based remote sensing image retrieval (CBRSIR) is an important problem in computer vision with many applications such as military, agriculture, and surveillance. In this study, inspired by recent developments in person re-identification, we design and fine-tune a multi-branch deep learning architecture that combines global and local features to obtain rich and discriminative image representations. Additionally, we propose a new evaluation strategy that fully separates the test and training sets and where new unseen data is used for querying, thereby emphasizing the generalization capability of retrieval systems. Extensive evaluations show that our method significantly outperforms the existing approaches by up to 10.7% in mean precision@20 on popular CBRSIR datasets. Regarding the new evaluation strategy, our method attains excellent retrieval performance, yielding more than 95% precision@20 score on the challenging PatternNet dataset.

## 1 INTRODUCTION

In recent years, Remote Sensing (RS) imagery has become increasingly available. RS image collections now contain a large number of pictures at high resolutions. Manual labeling and/or annotation of images is a cumbersome and expensive task. Furthermore, certain labeling or indexing methods are not suited for user-friendly retrieval (e.g. consider the case of geographic coordinates as labels). Therefore, new methods for managing RS image collections need to be developed. A technique that has achieved significant success in the RS community is Content-Based Remote Sensing Image Retrieval (CBRSIR) (Manjunath and Ma, 1996, Bai et al., 2014, Tang et al., 2018). In CBRSIR, the goal is to generate compact and robust representations of the visual content of images, to easily find similarities among them. Such systems can generally be reduced to two principal phases. First, compact image representations are generated in the feature extraction phase. Second, image similarity is computed based on special metrics, using the descriptors produced in the first phase.

Research has been commonly concentrated on the feature-extraction process (Zhou et al., 2017, Roy et al., 2018, Xiong et al., 2019). Different types of descriptors have been used for CBRSIR, which are classified according to their semantic level. Examples of low-level descriptors are SIFT (Lowe et al., 1999) and Gabor filters (Haralick et al., 1973), which describe shape, texture, color, etc. Mid-level descriptors are produced by aggregating low-level features, using methods like Bag-of-Words (Sivic and Zisserman, 2003) or Vector of Locally Aggregated Descriptors (Jégou et al., 2010). High-level descriptors encode information related to semantic concepts, such as "airplane" and "vegetation". These high-level semantic descriptors are commonly extracted from Convolutional Neural Networks (CNNs), which are trained for tasks such as classification. Several architectures and techniques are deployed in literature to improve feature extraction. Among these are, Deep Metric Learning (Roy et al., 2018, Xiong et al., 2019) and local feature extraction & aggregation (Tang et al., 2018, Imbriaco et al., 2019).

In this study, we explore the CBRSIR and present a twofold contribution. First, inspired by recent developments in the field of person re-identification, we design and deploy a part-based feature extractor and obtain state-of-the-art retrieval results. To the best of our knowledge, this is the only system to produce a single global representation, using a part-based model for CBRSIR without requiring aggregation after extraction. Second, we propose an alternative, more challenging evaluation protocol to study the generalization capabilities of CBRSIR systems. This simu-

[a]These authors contributed equally to this work.

Imbriaco, R., Alkanat, T., Bondarev, E. and de With, P.
Multi-Branch Convolutional Descriptors for Content-based Remote Sensing Image Retrieval.
DOI: 10.5220/0008895702420249
In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) - Volume 5: VISAPP*, pages 242-249
ISBN: 978-989-758-402-2; ISSN: 2184-4321

lates a more realistic scenario, where not all of the classes are available for training purposes, and where images of known classes can be acquired under different conditions or by different sensors.

# 2 RELATED WORK

In this section, a summary of the related work is presented, focusing on methods that exploit CNN features, metric learning, and local features.

## 2.1 Global Feature Extraction

As mentioned previously, one of the fundamental processing steps of any CBRSIR system is the feature extractor. Figure 1 depicts the basic architecture of an RSIR system. These are commonly classified according to the semantic complexity of the representations they generate from RS imagery. Early work (Haralick et al., 1973, Manjunath and Ma, 1996), utilized handcrafted textural features to match images with similar visual content. Richer semantic features are obtained by combining local descriptors, such as SIFT (Lowe et al., 1999) with aggregation methods like Bag-of-Words (BoW) (Sivic and Zisserman, 2003) or Vector of Locally Aggregated Descriptors (VLAD) (Jégou et al., 2010). However, most recent CBRSIR systems employ rich semantic features extracted from CNNs. Penatti *et al.* demonstrate in (Penatti et al., 2015) that CNN features are generic enough for RS imagery classification. In (Zhou et al., 2017), various methods are proposed for the extraction of descriptive representations. These include a Network-in-Network block (Lin et al., 2013) for dimensionality reduction. The above-mentioned systems produce a single, high-dimensional vector per image. These representations are commonly referred to as global descriptors, as they encode information about the whole image instead of image regions. The approaches discussed in the following subsection produce local representations and descriptors.

## 2.2 Local Feature Extraction

Alternatives to the global descriptors are presented in (Tang et al., 2018) and (Imbriaco et al., 2019). The first work presents an unsupervised framework for CBRSIR. Images are divided into patches and then fed through an auto-encoder that reconstructs the inputs. Two types of patches (uniform and superpixel) are extracted per image. The descriptors generated from each patch are aggregated using Bag-of-Words, producing a histogram representation for each image. The second work deploys attentive Deep Local Features (Noh et al., 2017) for the extraction of local descriptors at various scales. These descriptors are aggregated using VLAD (Jégou et al., 2010) and the network is trained for classification. A disadvantage of local descriptors is that the direct estimation of image similarity becomes computationally expensive. A single image may contain a large number of local descriptors, making brute-force search inefficient for large databases. Systems that exploit local descriptors generally aggregate them into a single global representation (BoW, VLAD) for efficient database search. Furthermore, all methods described above deploy networks trained for other tasks, e.g. classification. An emerging trend in CBRSIR is to train networks using metric learning for improved retrieval performance.

## 2.3 Metric Learning for RSIR

Features extracted from CNNs generalize well to tasks the network was not trained for (Penatti et al., 2015). However, networks trained using metric-learning objective functions, such as contrastive loss (Chopra et al., 2005) or triplet loss (Weinberger and Saul, 2009), show excellent performance in various retrieval tasks, like person re-identification (Hermans et al., 2017) and remote sensing image retrieval (RSIR) (Chaudhuri et al., 2019, Cao et al., 2019). Cao *et al.* present a novel method for RSIR using metric learning and study various dimensionality-reduction techniques. Their system produces global descriptors, learned with triplet loss. Dimensionality reduction and whitening of the trained descriptors is done using either Principal Component Analysis (Jolliffe, 2011), or learned using a fully-connected layer. Meanwhile, Chauduri *et al.* construct a Region Adjacency Graph, which is fed into a Graph Convolution Network to produce a global descriptor. This descriptor encodes the relationship between adjacent objects in the images. Training is done with the contrastive loss. Metric learning produces descriptors with small intra-class distances and large inter-class distances in feature space. These types of descriptors provide excellent ranking performance, even when dealing with the visual complexity occurring in RS imagery (scale, lighting and position variations). Unlike the work discussed above, our approach enables us to extract local image information without requiring neither additional post-processing of the convolutional descriptors nor computation of visual dictionaries (as in BoW and VLAD). Inspired by person re-identification, we deploy an architecture based on the Multi-Granularity Network (MGN) (Wang et al., 2018), and train it for CBRSIR. MGN's architecture enables the extrac-

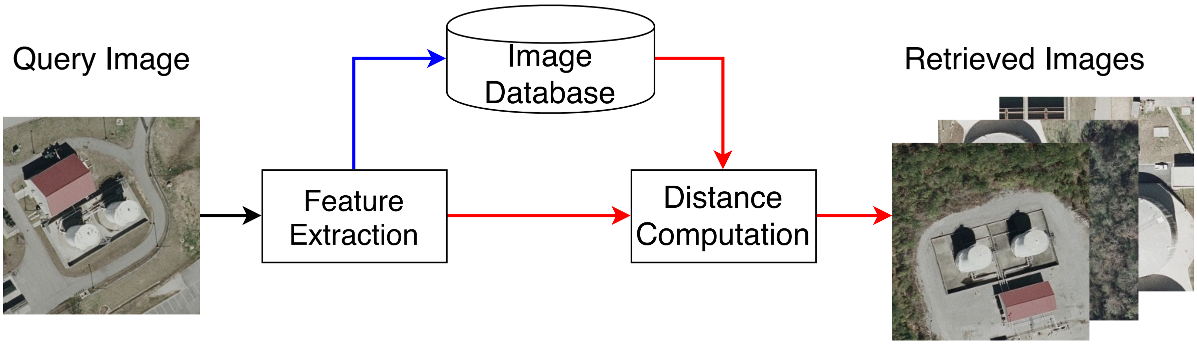

Figure 1: Diagram of the principal phases in a CBRSIR system. The blue arrow depicts the offline data flow, whereas the red arrows depict the online data flow.

tion of features from various regions and at different granularities (parts of images), thereby producing a compact and highly descriptive representation. The combined qualities of this approach enable excellent retrieval performance. Furthermore, we consider a more generic retrieval case, in which the network is trained on a dataset and retrieval is performed on different, unseen datasets. This evaluation procedure more closely resembles a real-world scenario, where the training data and retrieval database may have different domains. In conclusion, we aim at an architecture based on MGN which offers a compact and highly descriptive representation, while remaining robust for retrieval with unseen data. A more detailed description of the architecture and the design parameters are given in the following section.

## 3 METHOD

**A. Overview and Feature Extractor:** RS imagery is significantly different from street-level imaging. Images are acquired from an orthographic view and at a high altitude, thereby altering the image characteristics. Moreover, the variety of locations is large, introducing an additional demand on the generalization ability of the feature extractor. Other sources of appearance variations can be conditional on weather and lighting (e.g. illumination and occlusion), or environmental/anthropogenic (e.g. agricultural). The above-mentioned difficulties inherent to CBRSIR motivates research towards the development of better feature extraction approaches.

As discussed in Section 2.1, CNN-based feature extraction approaches are shown to be superior compared to their handcrafted counterparts. Thus, to exploit the remarkable potential of deep learning, we adopt CNNs to extract rich features from RS imagery. However, a single, universal CNN architecture that performs well on every problem does not exist. Well-performing architectures are explored for specific applications.

Recently, in other image retrieval tasks (person and vehicle re-identification), simultaneous usage of global and local features has significantly contributed to the overall retrieval performance (Chen et al., 2019, Zheng et al., 2019). Motivated by this information, we propose to jointly use global and local features, to achieve superior performance in CBRSIR. Inspired by the person re-identification approach based on MGN (Wang et al., 2018), we propose the CNN architecture shown in Figure 2.

**B. Part-based CNN Architecture:** As shown in Figure 2, our architecture is a five-branch CNN, where four of the branches extract local features and one extracts global features. We utilize four local branches of increasing granularity, to better adapt to variations in scale and position of the depicted object in the image content. Many powerful backbone architectures exist in the literature, such as ResNet (He et al., 2016), Inception (Szegedy et al., 2015) and VGG (Simonyan and Zisserman, 2014). In this study, the backbone architecture is ResNet-50, which is pre-trained on ImageNet. This decision is guided by ResNet-50's desirable computational-cost-to-performance ratio and prior success in the re-identification literature. Other backbone architectures can also be used with the proposed approach.

The shared backbone consists of ResNet-50 blocks up to and including the conv4_1 layer. Each branch uses the output of the shared backbone to concurrently extract local and global features. The features are generated with the remaining blocks of the ResNet-50 architecture up to and including the conv5_1 layer. After branching, the ResNet blocks do not share parameters. During the extraction of local branch features, we follow (Wang et al., 2018) and reduce the last stride of ResNet-50 from 2 to 1,

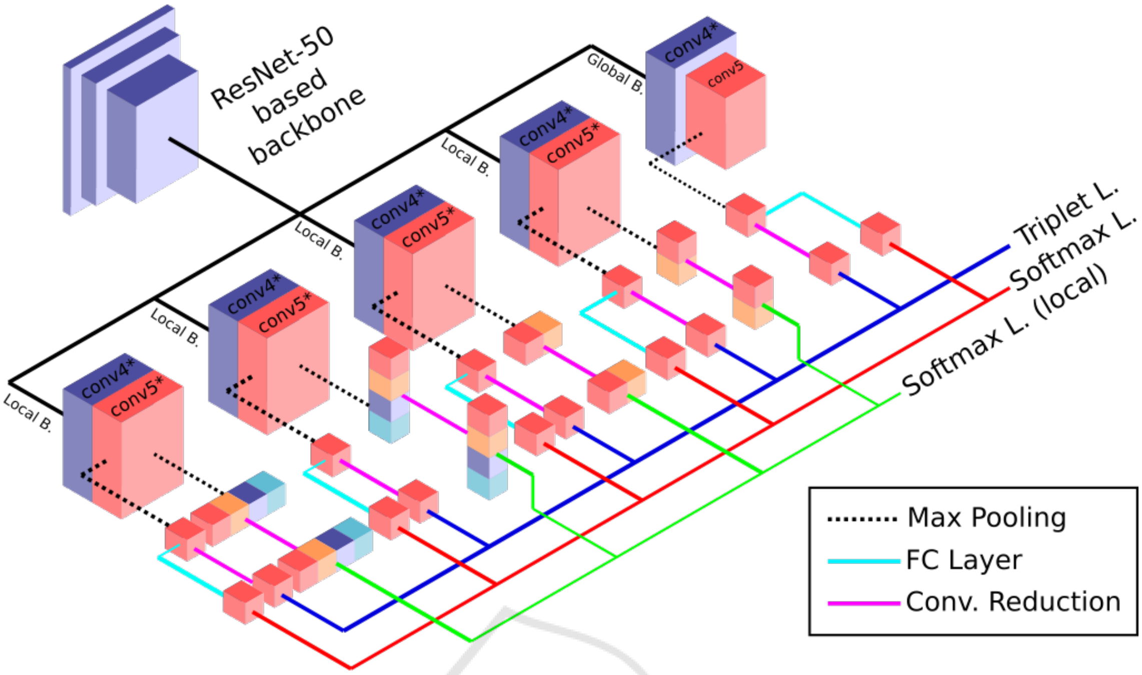

Figure 2: Overview of our proposed model. Our architecture has five branches, one being the global branch and rest are the local branches. After the shared-weight backbone, the global branch pools the feature tensor along the spatial axes without partitioning. In contrast, the local branches partition the feature tensor in different configurations prior to pooling. Then, each local feature is trained using softmax for RS classification problem. Note that, the FC layers after the convolutional reduction block of each local feature are omitted for clarity (best viewed in color).

which produces better local features. This reduction enables the local branches to extract richer features and is found to be helpful in other studies (Luo et al., 2019, Kalayeh et al., 2018). Then, as depicted in Figure 2, max-pooling is applied. Each branch has a different pooling strategy, to boost the feature extraction performance. The global branch, utilizes max-pooling over both height and width dimensions, reducing the spatial size to unity. In addition to global max-pooling, the local branches also partition the feature tensor into multiple local feature tensors and then apply individual max-pooling operations. The resulting pooled features are then trained with different approaches. We train for *global* feature extraction using the softmax cross-entropy (abbreviated as softmax) and triplet losses, while we train for *local* feature extraction using only the softmax loss.

**C. Global Features - Softmax:** Both the global and local branches learn global features with the softmax loss. To this end, global max-pooling is applied to the feature tensors after the conv5_1 layer of each branch. Afterwards, the tensor is passed through a fully-connected (FC) layer, effectively reducing the size of the pooled feature vectors to $N_c$, where $N_c$ denotes the number of classes in the training dataset. Lastly, the output of the FC layer is trained with softmax loss. Our method utilizes softmax loss in addition to triplet loss on global features, because it has shown to provide richer features.

**D. Local Features - Softmax:** The local branches divide the output of conv5_1 into horizontal or vertical partitions of increasing granularity. Horizontally, the feature tensor is split into partitions of $1 \times 2$ and $1 \times 4$ sub-tensors. The vertical partitions are generated with the same dimensions. This partitioning strategy produces high-dimensional representations that take advantage of the contextual information. Then, the max-pooling operation is applied individually to every sub-tensor. Afterwards, the resulting feature vectors are reduced in size using a convolutional reduction block. This block consists of $1 \times 1$ convolution, batch normalization, and ReLU layers, and reduces the feature size to 256, leading to relatively compact descriptors. Finally, each reduced feature is inputted to an FC layer that has $N_c$ elements and local features are learned using the softmax loss for classification.

**E. Global Features - Triplet Loss:** In addition to the global softmax loss of each branch, training is performed with the triplet loss on the reduced max-pooled feature vectors. The dimensionality-reduction strategy is identical to that of the local softmax loss training. The resulting feature vectors are trained using the triplet loss given as:

$$\mathcal{L}_{triplet} = \sum_{\substack{a,p,n \\ y_a = y_p \neq y_n}} min(0,\ m + D_{a,p} - D_{a,n}), \qquad (1)$$

where $a$, $p$, $n$ are anchor, positive and negative samples, $y_i$ is the class of sample $i$, parameter $m$ is the margin, and $D_{a,p}$ and $D_{a,n}$ are the distances between anchor-positive and anchor-negative samples, respectively. Note that we follow (Hermans et al., 2017) and apply hard-triplet mining to enhance the discrimination ability of the triplet loss. This hard-triplet mining strategy first picks $P$ classes and then $K$ images randomly from each class to construct a mini-batch. During the training, the network weights are only updated for each anchor sample once, using the hardest positive and negative images within the mini-batch. At inference time, the final feature vector used for retrieval is constructed by concatenating the triplet-loss trained features. Per branch, feature vectors are extracted from each partition, prior to their last FC layer and concatenated. This leads to a fixed-size feature vector for every image, regardless of the number of classes in a given training dataset. Then, to mathematically represent the distance between any arbitrary pair of samples, we calculate a distance metric between their descriptors. We use the $L_2$ distance to compute the similarity between two feature vectors.

# 4 EXPERIMENTS

## 4.1 Datasets

To demonstrate the efficiency of our approach, we evaluate on three widely used public datasets: UC Merced Land Use (Yang and Newsam, 2010), NWPU-RESISC45 (Cheng et al., 2017) and PatternNet (Zhou et al., 2018).

**UC Merced.** Published in 2010, this dataset includes 2,100 images equally distributed over 21 classes. The image size is 256 × 256 pixels, where pixel resolution is approximately 30 centimeters.

**NWPU-RESISC45.** This large-scale dataset includes 45 classes, each including 700 images. The image size is 256 × 256 pixels, where the pixel resolution varies between 30 and 0.2 meters. Published in 2016, this dataset is specifically challenging due to its large number of classes.

**PatternNet.** The PatternNet dataset has been published in 2017 and forms one of the most recent additions to the RSIR literature. This dataset includes 38 classes with 800 images. The 256 × 256 pixel images have been collected from Google Earth and Google Map API. As is the case for the NWPU-RESISC45 dataset, the pixel resolution is not fixed and varies between 4.693 and 0.062 meters.

## 4.2 Metrics

We adopt two popular metrics to evaluate our approach. Per experiment, the Mean Precision@k (mP@k) and the average normalized modified retrieval rank (ANMRR) (Manjunath et al., 2001, Aptoula, 2013) are computed. In the case of mP@k, we compute the ratio of correctly retrieved images in the top $k$ positions. A higher number denotes better performance. The ANMRR evaluates the retrieval performance taking both the number and rank of the retrieved results into account. In this case, a smaller number indicates better performance.

## 4.3 Splits and Evaluation Protocol

To evaluate the performance of our approach and to achieve a fair comparison with other methods, we use two evaluation protocols. In the first evaluation protocol, which is also employed in (Tang et al., 2018, Imbriaco et al., 2019,Cao et al., 2019), we train and evaluate on the same dataset. To split the datasets, we first randomly sample 20% of all images in a class-balanced manner. Then, we use the remaining part as the training set. For the evaluation, we consider each of the smaller subset images as the query and the rest of the dataset as the gallery. Retrieved matches of the same class are considered true positives. This split and evaluation protocol is commonly used for CBRSIR. However, this evaluation protocol has a significant shortcoming. In this strategy, the training and evaluation sets are not completely disjoint for the gallery images. In other words, some of the gallery images for each query are also used for training purposes.

To compensate for the shortcoming of the existing evaluation protocol and to obtain a better idea about the generalization capabilities of CBRSIR systems, we also report results on the transfer learning setting. In this evaluation approach, we train on a dataset and evaluate on another. In this way, the training and evaluation sets are completely disjoint. Moreover, this evaluation protocol reveals how successful an algorithm is on discriminating different RS structures in an image, even if it is not trained to recognize specific cues. For example, unlike dataset NWPU45, the UC Merced dataset does not contain the “swimming pool” class. A system with superior generalization ability is

Table 1: Comparison of image retrieval performances on the UC Merced, NWPU45 and PatternNet datasets.

| | UC Merced | | | NWPU45 | | | PatternNet | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | mP@10 | mP@20 | ANMRR | mP@10 | mP@20 | ANMRR | mP@10 | mP@20 | ANMRR |
| ResNet50 (Imbriaco et al., 2019) | - | 0.816 | - | - | 0.798 | - | - | - | - |
| DBOW (Tang et al., 2018) | - | 0.830 | - | - | 0.821 | - | - | - | - |
| V-DELF (MA) (Imbriaco et al., 2019) | - | 0.896 | - | - | 0.840 | - | - | - | - |
| SGCN (Chaudhuri et al., 2019) | 0.936 | - | 0.300 | - | - | - | 0.971 | - | 0.210 |
| DML (Cao et al., 2019) | 0.976 | - | 0.023 | - | - | - | **0.996** | - | **0.003** |
| Ours - Global branch only | 0.979 | 0.979 | 0.019 | 0.944 | 0.941 | **0.074** | 0.994 | 0.994 | 0.012 |
| Ours | **0.990** | **0.990** | **0.013** | **0.951** | **0.947** | 0.089 | **0.996** | **0.995** | 0.013 |

expected to be able to discriminate the images belonging to this class as a separate structure, even if it was not explicitly trained with this class.

## 4.4 Hyperparameters and Settings

Throughout our experiments, we have used an initial learning rate of 0.0002, reduced by one-tenth at epochs 60, 100 and 125. All models have been trained for a total of 150 epochs using the Adam algorithm (Kingma and Ba, 2014) and the PyTorch framework (Paszke et al., 2017). The values for weight decay and the triplet loss margin were set to 0.0005 and 1.2, respectively. Lastly, hard-triplet mining parameters $P$ and $K$ were set to $P = 4$ and $K = 5$ for PatternNet and NWPU45 datasets, where $P = 3$ and $K = 7$ were used for UC Merced dataset.

## 4.5 Results

We present the CBRSIR results organized as follows. In Table 1, we show the comparative performance evaluation of our approach against the results of existing methods. Table 2 depicts the per-class retrieval performance evaluated on the UC Merced Land Use dataset. Lastly, we present in Table 3, the retrieval performance results for transfer learning evaluation for all dataset combinations. As it can be observed from Table 1, our method outperforms the previous state-of-the-art. The multi-branch convolutional descriptors increase the retrieval performance on all three datasets, except for PatternNet, where our results are comparable to those of the state of the art. Moreover, for UC Merced and PatternNet datasets, our results exceed 99.5%, which may be interpreted as an indication of a saturated performance.

Table 2: Retrieval performances of our system on the UC Merced dataset for each class and the average.

| Class | mP@1 | mP@5 | mP@10 | mP@20 | ANMRR |
|---|---|---|---|---|---|
| agricultural | 1.000 | 1.000 | 1.000 | 1.000 | 0.008 |
| airplane | 1.000 | 1.000 | 1.000 | 1.000 | 0.008 |
| baseball d. | 1.000 | 1.000 | 1.000 | 1.000 | 0.008 |
| beach | 1.000 | 1.000 | 1.000 | 1.000 | 0.008 |
| buildings | 1.000 | 0.950 | 0.965 | 0.975 | 0.017 |
| chaparral | 1.000 | 1.000 | 1.000 | 1.000 | 0.008 |
| d. residential | 0.900 | 0.930 | 0.925 | 0.915 | 0.056 |
| forest | 1.000 | 1.000 | 1.000 | 1.000 | 0.008 |
| freeway | 1.000 | 1.000 | 1.000 | 1.000 | 0.008 |
| golf course | 1.000 | 1.000 | 1.000 | 1.000 | 0.008 |
| harbor | 1.000 | 1.000 | 1.000 | 1.000 | 0.008 |
| intersection | 1.000 | 1.000 | 1.000 | 1.000 | 0.008 |
| m. residential | 1.000 | 1.000 | 0.990 | 0.992 | 0.008 |
| m. homepark | 1.000 | 1.000 | 1.000 | 1.000 | 0.008 |
| overpass | 0.950 | 0.950 | 0.950 | 0.950 | 0.032 |
| parkinglot | 0.950 | 0.950 | 0.950 | 0.950 | 0.032 |
| river | 1.000 | 1.000 | 1.000 | 1.000 | 0.008 |
| runway | 1.000 | 1.000 | 1.000 | 1.000 | 0.008 |
| s. residential | 1.000 | 1.000 | 1.000 | 1.000 | 0.008 |
| storagetanks | 1.000 | 1.000 | 1.000 | 1.000 | 0.008 |
| tenniscourt | 1.000 | 1.000 | 1.000 | 1.000 | 0.008 |
| **Average** | **0.990** | **0.990** | **0.990** | **0.990** | **0.013** |

Table 3: Retrieval results on transfer learning setting, explained in Section 4.3. Here, the model is trained on the “Training” dataset and evaluated directly on the “Test” dataset.

| Training | Test | mP@1 | mP@5 | mP@10 | mP@20 | ANMRR |
|---|---|---|---|---|---|---|
| UCM | PNet | 0.941 | 0.916 | 0.899 | 0.880 | 0.331 |
| UCM | NWPU45 | 0.717 | 0.658 | 0.627 | 0.591 | 0.620 |
| PNet | UCM | 0.926 | 0.851 | 0.798 | 0.715 | 0.483 |
| PNet | NWPU45 | 0.769 | 0.708 | 0.670 | 0.627 | 0.633 |
| NWPU45 | UCM | 0.967 | 0.937 | 0.911 | 0.880 | 0.199 |
| NWPU45 | PNet | 0.977 | 0.970 | 0.964 | 0.956 | 0.181 |

# 5 DISCUSSION

Observing Table 2, we conclude that our performance in various classes is balanced. In UC Merced, the lowest retrieval results are associated with the “dense residential” and “overpass” classes. We conjecture that the lowered performance in those classes occurs due to the existence of visually similar categories “medium residential” and “intersection”. In PatternNet, our overall result is only slightly higher than that of UC Merced. However, considering that PatternNet is nearly 15 times larger than UC Merced, this indicates that our method produces sufficiently discriminative representations even for small datasets with limited training images. As labeling is a laborintensive task, we consider this as an advantageous property. Perhaps, the most interesting results are those of the NWPU45-RESISC dataset. This dataset is comparable in size to PatternNet and it has the highest number of classes among all three datasets. As it is the case for existing approaches, our method exhibits decreased retrieval performance for this dataset. The classes “railway”, “railway station”, “terrace” and “palace” show the lowest overall performance, scoring 0.844, 0.876, 0.868, 0.853 in mP@20, respectively. Among those, the “railway” and “railway station” classes are visually similar. Our descriptor is not capable of producing discriminative representations for classes with such small semantic differences, resulting in reduced retrieval performance for both. Overall, we conclude that, according to both the mean precision and ANMRR metrics, our method offers class-balanced retrieval performance. As explained in Section 4.3, we conjecture that the transfer learning results in Table 3 are far more informative, since they provide better insight into the generalization abilities of our system. Investigating the mean precision@20, we conclude that, the best retrieval results are obtained when NWPU45 is used as the source dataset for training. Although the PatternNet and NWPU45 datasets are of similar scale, evaluation on UC Merced reveals that training on NWPU45 offers 16.5% better results than training on PatternNet. Thus, we conclude that thanks to the higher image diversity of NWPU45, training on this dataset yields better generalization. Closer inspection of the results in Table 3 also indicate that applying transfer learning, when the training dataset is UC Merced yield only moderate results. This reinforces the idea that the training dataset should be sufficiently rich in classes and number of images.

## 6 CONCLUSIONS

In this study, we have presented our deep learning-based approach to CBRSIR. By taking advantage of combining deep global and local features, we have achieved state-of-the-art results on three publicly available and popular datasets. Moreover, our approach offers near-perfect retrieval performance for the widely-used UC Merced and PatternNet datasets, while providing balanced retrieval performance for all classes of the considered datasets.

As an additional contribution, we have also argued that the existing evaluation protocol for the CBRSIR problem has shortcomings and that it is not informative about the generalization ability of CBRSIR systems. Thus, we have proposed to utilize transfer learning evaluation to alleviate the problems of the existing evaluation approach. Furthermore, we have presented the results of our method on the transfer learning evaluation setting. We presume that this new evaluation protocol will be beneficial for the CBRSIR literature and will motivate researchers to concentrate on methods with superior generalization capability.

## ACKNOWLEDGEMENTS

The authors kindly appreciate the NVIDIA gift of a Titan Xp GPU for this research. This work was supported by the Interreg project PASSAnT and the European ITEA project PS-CRIMSON.

## REFERENCES

Aptoula, E. (2013). Remote sensing image retrieval with global morphological texture descriptors. *IEEE transactions on geoscience and remote sensing*, 52(5):3023–3034.

Bai, Y., Yu, W., Xiao, T., Xu, C., Yang, K., Ma, W.-Y., and Zhao, T. (2014). Bag-of-words based deep neural network for image retrieval. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 229–232. ACM.

Cao, R., Zhang, Q., Zhu, J., Li, Q., Li, Q., Liu, B., and Qiu, G. (2019). Enhancing remote sensing image retrieval using a triplet deep metric learning network. *International Journal of Remote Sensing*, pages 1–12.

Chaudhuri, U., Banerjee, B., and Bhattacharya, A. (2019). Siamese graph convolutional network for content based remote sensing image retrieval. *Computer Vision and Image Understanding*, 184:22–30.

Chen, H., Lagadec, B., and Bremond, F. (2019). Partition and reunion: A two-branch neural network for vehicle re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 184–192.

Cheng, G., Han, J., and Lu, X. (2017). Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883.

Chopra, S., Hadsell, R., LeCun, Y., et al. (2005). Learning a similarity metric discriminatively, with application to face verification. In *CVPR (1)*, pages 539–546.

Haralick, R. M., Shanmugam, K., et al. (1973). Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hermans, A., Beyer, L., and Leibe, B. (2017). In defense of the triplet loss for person re-identification. *ArXiv*, abs/1703.07737.

Imbriaco, R., Sebastian, C., Bondarev, E., et al. (2019). Aggregated deep local features for remote sensing image retrieval. *Remote Sensing*, 11(5):493.

Jégou, H., Douze, M., Schmid, C., and Pérez, P. (2010). Aggregating local descriptors into a compact image representation. In *CVPR 2010-23rd IEEE Conference on Computer Vision & Pattern Recognition*, pages 3304–3311. IEEE Computer Society.

Jolliffe, I. (2011). *Principal component analysis*. Springer.

Kalayeh, M. M., Basaran, E., Gökmen, M., Kamasak, M. E., and Shah, M. (2018). Human semantic parsing for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1062–1071.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lin, M., Chen, Q., and Yan, S. (2013). Network in network. *arXiv preprint arXiv:1312.4400*.

Lowe, D. G. et al. (1999). Object recognition from local scale-invariant features. In *iccv*, volume 99, pages 1150–1157.

Luo, H., Gu, Y., Liao, X., Lai, S., and Jiang, W. (2019). Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0.

Manjunath, B. S. and Ma, W.-Y. (1996). Texture features for browsing and retrieval of image data. *IEEE Transactions on pattern analysis and machine intelligence*, 18(8):837–842.

Manjunath, B. S., Ohm, J.-R., Vasudevan, V. V., and Yamada, A. (2001). Color and texture descriptors. *IEEE Transactions on circuits and systems for video technology*, 11(6):703–715.

Noh, H., Araujo, A., Sim, J., Weyand, T., and Han, B. (2017). Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3456–3465.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in pytorch. In *NIPS-W*.

Penatti, O. A. B., Nogueira, K., and dos Santos, J. A. (2015). Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Roy, S., Sangineto, E., Demir, B., and Sebe, N. (2018). Deep metric and hash-code learning for content-based retrieval of remote sensing images. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 4539–4542. IEEE.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sivic, J. and Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. In *null*, page 1470. IEEE.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Tang, X., Zhang, X., Liu, F., and Jiao, L. (2018). Unsupervised deep feature learning for remote sensing image retrieval. *Remote Sensing*, 10(8).

Wang, G., Yuan, Y., Chen, X., Li, J., and Zhou, X. (2018). Learning discriminative features with multiple granularities for person re-identification. *CoRR*, abs/1804.01438.

Weinberger, K. Q. and Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb):207–244.

Xiong, W., Lv, Y., Cui, Y., Zhang, X., and Gu, X. (2019). A discriminative feature learning approach for remote sensing image retrieval. *Remote Sensing*, 11:281.

Yang, Y. and Newsam, S. (2010). Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pages 270–279. ACM.

Zheng, F., Deng, g., Sun, X., Jiang, X., Guo, X., Yu, Z., Huang, F., and Ji, R. (2019). Pyramidal person re-identification via multi-loss dynamic training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8514–8522.

Zhou, W., Newsam, S., Li, C., and Shao, Z. (2017). Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval. *Remote Sensing*, 9(5).

Zhou, W., Newsam, S., Li, C., and Shao, Z. (2018). Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS journal of photogrammetry and remote sensing*, 145:197–209.