BigEarthNet-MM: A Large-Scale, Multimodal, Multilabel Benchmark Archive for Remote Sensing Image Classification and Retrieval [Software and Data Sets]

Gencer Sumbul - Technische Universität Berlin, Berlin, Germany. gencer.suembuel@tu-berlin.de; **Arne de Wall -** Technische Universität Berlin, Berlin, Germany. a.dewall@tu-berlin.de; **Tristan Kreuziger -** Technische Universität Berlin, Berlin, Germany. tristan.kreuziger@tu-berlin.de; **Filipe Marcelino -** Direção-Geral do Território, Lisbon, Portugal. marcelino@dgterritorio.pt; **Hugo Costa -** Direção-Geral do Território, Lisbon, 1099-052, Portugal and NOVA Information Management School (NOVA IMS), Universidade Nova Lisboa, Campus de Campolide, 1070-312 Lisbon, Portugal. hugoagcosta@gmail.com; **Pedro Benevides -** Direção-Geral do Território, Lisbon, Portugal. pbenevides@dgterritorio.pt; **Mário Caetano -** Direção-Geral do Território, Lisbon, 1099-052, Portugal and NOVA Information Management School (NOVA IMS), Universidade Nova Lisboa, Campus de Campolide, 1070-312 Lisbon, Portugal. dgtcaetano@gmail.com; **Begüm Demir -** Technische Universität Berlin, Berlin, Germany. demir@tu-berlin.de; **Volker Markl -** Technische Universität Berlin, Berlin, Germany. volker.markl@tu-berlin.de

This is the accepted version of the article published in *IEEE Geoscience and Remote Sensing Magazine*

How to cite: Sumbul, G., De Wall, A., Kreuziger, T., Marcelino, F., Costa, H., Benevides, P., Caetano, M., Demir, B., & Markl, V. (2021). 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, 9(3), 174-180. https://doi.org/10.1109/MGRS.2021.3089174

Funding Information: This work is funded by the European Research Council (ERC) through the ERC-2017-STG BigEarth Project under Grant 759764 and by the German Ministry for Education and Research as BIFOLD - Berlin Institute for the Foundations of Learning and Data (01IS18025A). The authors from DGT are supported through FCT (Fundacão para a Ciência e a Tecnologia) - UIDB/04152/2020 - Centro de Investigação em Gestão de Informação (MagIC).

BigEarthNet-MM: A Large Scale Multi-Modal Multi-Label Benchmark Archive for Remote Sensing Image Classification and Retrieval

Gencer Sumbul, *Graduate Student Member, IEEE,* Arne de Wall, *Student Member, IEEE,* Tristan Kreuziger, *Student Member, IEEE,* Filipe Marcelino, Hugo Costa, Pedro Benevides, Mário Caetano, Begüm Demir, *Senior Member, IEEE,* Volker Markl

***Abstract*—This paper presents the multi-modal BigEarthNet (BigEarthNet-MM) benchmark archive made up of 590,326 pairs of Sentinel-1 and Sentinel-2 image patches to support the deep learning (DL) studies in multi-modal multi-label remote sensing (RS) image retrieval and classification. Each pair of patches in BigEarthNet-MM is annotated with multi-labels provided by the CORINE Land Cover (CLC) map of 2018 based on its thematically most detailed Level-3 class nomenclature. Our initial research demonstrates that some CLC classes are challenging to be accurately described by only considering (single-date) BigEarthNet-MM images. In this paper, we also introduce an alternative class-nomenclature as an evolution of the original CLC labels to address this problem. This is achieved by interpreting and arranging the CLC Level-3 nomenclature based on the properties of BigEarthNet-MM images in a new nomenclature of 19 classes. In our experiments, we show the potential of BigEarthNet-MM for multi-modal multi-label image retrieval and classification problems by considering several state-of-the-art DL models. We also demonstrate that the DL models trained from scratch on BigEarthNet-MM outperform those pre-trained on ImageNet, especially in relation to some complex classes, including agriculture and other vegetated and natural environments. We make all the data and the DL models publicly available at https://bigearth.net, offering an important resource to support studies on multi-modal image scene classification and retrieval problems in RS.**

***Index Terms*—Multi-modal learning, multi-label image retrieval, image classification, deep learning, remote sensing.**

I. Introduction

As a result of advancements in satellite technology, recent years have witnessed a significant increase in the volume of remote sensing (RS) image archives. Accordingly, the development of accurate scene classification and content based image retrieval (CBIR) systems in massive image archives has attracted great attention in RS. CBIR systems aim to achieve an efficient and precise retrieval of RS images from large archives that are similar to a query image [1], [2]. RS image scene classification systems aim at automatically assigning class labels to each RS image scene in a large archive [3], [4]. Deep learning (DL) based methods have recently seen a rise in popularity in the context of RS image scene classification and retrieval problems. Most DL models require a high amount of annotated images during training to optimize all parameters and reach a high performance. The availability and quality of such data determine the feasibility of many DL models. There are several benchmark archives made publicly available for different RS applications (e.g., pixel-based image classification). For a comprehensive list, we refer the reader to [5]. To the best of our knowledge, most of the existing publicly available benchmark archives for image scene classification and retrieval problems contain: 1) single-modal RS images (e.g., multispectral or SAR); and 2) single-label image annotations (i.e., each image is annotated by a single label that is associated with the most significant content of the considered image). However, multi-modal images associated with the same geographical area allow for rich characterization of RS images and thus improve image retrieval performance when jointly considered [6]. In addition, RS images usually contain areas with a high variety of semantically complex content that must be reflected by more than one class annotation through multiple class labels (multi-labels).

Thus, a benchmark archive consisting of multi-modal images annotated with multi-labels is needed. However, annotating RS images with multi-labels at a large-scale to drive DL studies is time consuming, complex, and costly in operational scenarios. To overcome this problem, a common approach is to exploit DL models with proven architectures (such as ResNet [7] or VGG [8]), which are pre-trained on publicly available general purpose datasets in the computer vision (CV) community (e.g., ImageNet [9]). The existing model is then fine-tuned on a small set of RS images annotated with multi-labels to calibrate the final layers. This strategy is also known as a transfer learning strategy. There are several versions of the above-mentioned models that have been pre-trained on large-scale datasets in CV. However, we argue that this is not a proper approach in RS, because of the differences in image characteristics in CV and RS. For example, Sentinel-2 multispectral images have 13 spectral bands associated with varying and lower spatial resolutions compared to the CV images. In addition, the semantic content present in CV and RS images is significantly different, and thus the respective semantic classes differ from each other. To address this issue, we have recently introduced BigEarthNet [10] as a large-scale single-modal benchmark archive for RS image search, retrieval

Gencer Sumbul, Arne de Wall, Tristan Kreuziger, Begüm Demir and Volker Markl are with Technische Universität Berlin, Berlin, Germany.

Filipe Marcelino, Hugo Costa, Pedro Benevides, and Mário Caetano are with Direção-Geral do Território (DGT), Lisbon, Portugal. Hugo Costa and Mário Caetano are also with NOVA Information Management School (NOVA IMS), Universidade Nova Lisboa, Campus de Campolide, 1070-312 Lisbon, Portugal.

and classification. BigEarthNet contains 590,326 Sentinel-2 image patches annotated with multi-labels provided by the CORINE Land Cover (CLC) map of 2018 (CLC 2018) [11]. The CLC nomenclature includes land cover and land use classes grouped in a three-level hierarchy, and for the BigEarthNet image patches, the most thematically detailed Level-3 class nomenclature is considered. However, there are some CLC classes that are difficult to be identified by only exploiting (single-date) Sentinel-2 images, because: i) land use concepts associated with some classes (e.g., *Dump sites*, *Sport and leisure facilities*) may not be visible from space or fully recognizable with the spatial resolution of Sentinel-2 images, and ii) RS time series, which BigEarthNet does not include, may be required to describe and discriminate some classes (e.g., *Non-irrigated arable land*, *Permanently irrigated land*). In addition, BigEarthNet is not suitable for the multi-modal learning-based algorithm development and validation purposes, since it only contains Sentinel-2 image patches.

To overcome these issues, in this paper we introduce the multi-modal BigEarthNet (BigEarthNet-MM) that contains 590,326 pairs of Sentinel-2 and Sentinel-1 image patches. We also introduce an alternative nomenclature for images in BigEarthNet-MM as an evolution of the original CLC labels. Fig. 1 shows an example of the BigEarthNet-MM image pairs and their multi-labels from the new nomenclature.

II. Description of BigEarthNet-MM

BigEarthNet-MM contains 590,326 pairs of Sentinel-1 and Sentinel-2 image patches acquired over 10 different European countries (Austria, Belgium, Finland, Ireland, Kosovo, Lithuania, Luxembourg, Portugal, Serbia, Switzerland). Sentinel-2 patches of BigEarthNet-MM are taken from the original BigEarthNet [10]. To construct these patches, 125 Sentinel-2 tiles associated with less than 1% of cloud cover and acquired between June 2017 and May 2018 were considered. All tiles were atmospherically corrected by employing Sentinel-2's Level 2A product generation and formatting tool (sen2cor) provided by the European Space Agency due to its proven success in the literature. After the atmospheric correction, the 10^{th} band of each image patch is not available anymore, as it is the cirrus band (which is omitted in the Level 2A output for its lack of surface information). Then, the tiles were divided into $590,326$ non-overlapping image patches, each of which is a section of: 1) 120×120 pixels for 10m bands; 2) 60×60 pixels for 20m bands; and 3) 20×20 pixels for 60m bands. One important goal during the tile selection process was to represent all chosen geographic locations with images acquired in different seasons. Due to the restrictions of finding tiles with a low cloud cover percentage in the relatively narrow time period, this has not been possible at each considered location. Accordingly, the following respective numbers of patches for autumn, winter, spring, and summer have been considered: 143557, 72877, 175937, and 126913. For the quality check of patches, visual inspection was also employed, which led to the identification of $70,987$ Sentinel-2 image patches that are fully covered by seasonal snow, cloud, and cloud shadow[1].

To construct the Sentinel-1 patches of BigEarthNet-MM, 325 Sentinel-1 Ground Range Detected (GRD) products acquired between June 2017 and May 2018 that jointly cover the area of all original 125 Sentinel-2 tiles with close temporal proximity were selected and processed. The selected scenes provide dual-polarized information channels (VV and VH) and are based on the interferometric wide swath (IW) mode, which is the main acquisition mode over land. All scenes were pre-processed by using the Sentinel-1 toolbox (S1TBX) and the graph processing framework (GPF) of ESA's Sentinel Application Platform (SNAP). This includes the application of precise orbit files, border and thermal noise removal, radiometric calibration, and geometric correction (i.e., Range Doppler terrain correction). Depending on the spatial extent of the scene, either the SRTM 30 (for scenes below 60° latitude) or the ASTER DEM (for scenes above 60° latitude, where no SRTM 30 exists) were employed in the geometric correction to project images from slant range to ground range. Finally, the backscatter coefficient was converted to a decibel (dB) scale. It is worth noting that, since the selection of the speckle filter is considered to be application dependent, no speckle filtering was applied in our pre-processing workflow in order to preserve the full resolution. This approach is also recommended by the Product Family Specification for SAR of the CEOS Analysis Ready Data for Land (CARD4L) framework[2]. Based on the pre-processed Sentinel-1 scenes, for each Sentinel-2 patch, a corresponding Sentinel-1 patch with a close timestamp was extracted. In addition, each Sentinel-1 patch inherited the annotations of the corresponding Sentinel-2 patch. The resulting Sentinel-1 image patches contain a spatial resolution of 10m.

A. Class-Nomenclature of BigEartNet-MM

Each pair (which is made up of Sentinel-1 and Sentinel-2 image patches acquired in the same geographical area) in BigEarthNet-MM is associated with one or more class labels (i.e. multi-labels) extracted from the CORINE land cover map of 2018. CORINE land cover (CLC) is a pioneer adventure initiated in the 80's of the last century to produce harmonized land cover land use (LCLU) maps in vector format for the member states of the European Union. According to the validation report of the CLC, the accuracy is around 85% [12]. Nowadays, CLC covers 39 countries from Europe and was produced for five reference years, 1990, 2000, 2006, 2012, and 2018. The latter was produced with data of 2017-2018, which matches the time frame of the images included in BigEarthNet. Motivations for embracing a large-scale mapping endeavor aimed at meeting the demand for spatially explicit and harmonized information on land for a variety of purposes, such as environmental management and decision making. The crude state-of-the-art of the 1980's technology and the large spectrum of potential uses of the maps led to the definition of a coarse spatial resolution and a nomenclature with some broad class definitions, mixing land cover and land use concepts. These definitions are implemented for map production by visual interpretation of RS images and additional data in most

[1]The lists are available at http://bigearth.net/#downloads.

[2]https://ceos.org/ard/

countries. Additional data may include very high spatial resolution imagery and official spatial data sets like land registers, often to infer the land use. The same technical specifications were preserved in map updating for historical consistency. Thus the produced five CLC maps have a minimum mapping unit of 25 ha and a minimum mapping width of 100 m, and provide information on land according to a hierarchical nomenclature of 44 classes at the most detailed level (Level3). The image patches in BigEarthNet-MM are representative of 43 CLC classes. In the case that CLC maps are considered as labeling sources for training the machine learning methods to automatically analyse RS images, the modified versions of the CLC nomenclature (which better fit the purpose of the considered application) are commonly preferred. One of the main reason is that RS systems directly observe the land cover rather than the land use. The CLC land-use based labels may not be fully recognizable through the RS images unless they are not associated to very high spatial resolution. As an example, in [13] CLC is used as a basis to collect training data for supervised RS image classification, but classes such as *Discontinuous urban fabric* and *Sport and leisure facilities* that depend mainly on land use were removed. A deep revision of the CLC program is actually under consideration following the concept of the EIONET Action Group on Land monitoring in Europe (EAGLE) [14].

To pay more justice to the properties of BigEarthNet-MM image pairs, we introduce a new class-nomenclature by modifying the multi-labels extracted from the CLC 2018. To this end, the CLC Level-3 nomenclature is interpreted and arranged in a new nomenclature of 19 classes[3]. Ten classes of the original CLC nomenclature are maintained in the new nomenclature, 22 classes are grouped into 9 new classes, and 11 classes are removed. The classes maintained are semantically homogeneous and largely related to land cover, such as *Broad-leaved forest* and *Beaches, dunes, sands*. Furthermore, CLC classes that are not feasible to be identified by only using single-date BigEarthNet-MM images removed, such as *Burnt areas*. Complex classes (which are often removed when undertaking image classification) are maintained, such as *Complex cultivation patterns* and *Land principally occupied by agriculture, with significant areas of natural vegetation*. The goal is to investigate the ability of DL models to learn from spatial patterns that express semantic classes. Classes are grouped when sharing similar land cover types and spectral patterns. For example, *Moors and heath land* and *Sclerophyllous vegetation* are grouped in a single class, and a new class, *Arable land*, groups similar crops that require dense time series (which not available in BigEarthNet-MM) for their discrimination (e.g. irrigated and non-irrigated crops). Classes that strongly depend on land use or need additional data for their discrimination are removed. For example, class *Airports* essentially relates to land use, and *Intertidal flats* appear in RS images either with or without water depending on the image acquisition time and hence require appropriate time series for its classification. The number of labels associated with each image pair varies between 1 and 12, while 96.80% of image pairs are not associated with more than 5 labels. Only 23 image pairs are annotated with more than 9 labels.

[3]https://bigearth.eu/BigEarthNetListofClasses.pdf

III. Experiments

A. Experimental Design

The experiments were carried out in the context of content based multi-modal multi-label RS image retrieval and classification. To achieve multi-modal learning, we stacked the VV and VH bands of Sentinel-1 image patches, and the Sentinel-2 bands associated with 10m and 20m spatial resolution into one volume for each pair in BigEarthNet-MM. To this end, we initially applied cubic interpolation to 20m bands of Sentinel-2 image patches. In the experiments, we did not use the Sentinel-2 image bands associated with 60m spatial resolution (bands 1 and 9). This is due to the fact that these bands are mainly used for cloud screening, atmospheric correction, and cirrus detection in RS applications and do not embody a significant amount of information for the characterization of semantic content of RS images. In the experiments, we considered the VGG model [8] and the ResNet model [7] at various number of layers (VGG16, VGG19, ResNet50, ResNet101, ResNet152). To fairly compare all models, we utilized the Adam optimizer [15] with an initial learning rate of 10^{-3} to decrease the sigmoid cross-entropy loss. Except the learning rate, we employed the same parameter values given in [3], [7], [8]. The batch size is set to 256 for ResNet152 and to 500 for all other models used in the experiments. We applied training from scratch for 100 epochs, while the final layers of the pre-trained models were fine-tuned separately on each modality for 10 epochs. For all the models, we added a fully connected layer that includes 19 neurons at the end of the network for the classification. For image retrieval, we extracted image features from the considered models and applied similarity matching of the features based on the χ^2-distance measure. We performed various experiments to analyze the effectiveness of: i) learning from BigEarthNet-MM directly (through training from scratch) instead of using the pre-trained models on ImageNet; and ii) state-of-the-art CNN models trained and evaluated on BigEarthNet-MM. To use the pre-trained models on ImageNet, we used the late fusion of separately fine-tuned models on Sentinel-1 and Sentinel-2 patches. In the experiments, we did not use the Sentinel-2 patches that are fully covered by seasonal snow, cloud, and cloud shadow. After the arrangements of the new class nomenclature, 57 pairs among the $590,326$ pairs are not associated with any LCLU labels. these pairs are not used in the experiments. We divided the remaining dataset into: i) the training set of $269,695$ pairs of patches, ii) validation set of $123,723$ pairs of patches, and iii) the test set of $125,866$ pairs of patches.

We performed our experiments on a cluster of 4 NVIDIA Tesla V100 GPUs. The results of multi-modal multi-label image classification were provided in terms of four performance metrics: 1) Hamming loss (HL); 2) one-error (OE); 3) recall (R); and 4) F_2-Score (F_2). For a detailed description of the considered metrics, the reader is referred to [3].

TABLE I
CLASS-BASED F_2 SCORES (%) OBTAINED WHEN: I) TRANSFER LEARNING FROM IMAGENET AND II) DIRECT LEARNING FROM BIGEARTHNET-MM ARE USED FOR MULTI-MODAL MULTI-LABEL IMAGE CLASSIFICATION.

Class	**Transfer Learning From ImageNet**	**Learning From BigEarthNet-MM**
Urban fabric	56.27	**71.99**
Industrial or commercial units	30.98	**43.21**
Arable land	80.05	**83.62**
Permanent crops	4.32	**55.52**
Pastures	50.98	**74.77**
Complex cultivation patterns	36.29	**62.03**
Land principally occupied by agriculture, with significant areas of natural vegetation	30.36	**60.63**
Agro-forestry areas	2.13	**71.87**
Broad-leaved forest	42.83	**75.39**
Coniferous forest	75.47	**86.32**
Mixed forest	72.19	**81.31**
Natural grassland and sparsely vegetated areas	14.11	**43.88**
Moors, heathland and sclerophyllous vegetation	5.29	**59.91**
Transitional woodland-shrub	41.23	**64.21**
Beaches, dunes, sands	43.67	**63.39**
Inland wetlands	8.20	**57.81**
Coastal wetlands	4.79	**42.23**
Inland waters	63.23	**82.10**
Marine waters	93.99	**97.20**
Average	39.81	**67.23**

B. Experimental Results

1) Comparison among the Strategies of Learning directly from BigEarthNet-MM and Transfer Learning from the ImageNet: In the first set of experiments, we compare the effectiveness of learning directly from BigEarthNet-MM with respect to transfer learning from ImageNet. To this end, transfer learning strategy is applied by using the pre-trained ResNet50 model trained on ImageNet, while direct learning strategy is employed by using the ResNet50 trained from scratch on BigEarthNet-MM. Table I shows the class-based F_2 classification scores (known also as macro-averaged F_2 scores [3]). By analyzing the table, one can see that learning directly from BigEarthNet-MM achieves the highest score for each class compared to the transfer learning strategy. As an example, learning directly from BigEarthNet-MM provides more than 12% and 25% higher scores for the classes *Industrial or commercial units* and *Complex cultivation patterns*, respectively, compared to the transfer learning strategy. The difference in performance between these learning strategies is more evident for more complex LULC classes. As an example, learning directly from BigEarthNet-MM improves the F_2 scores more than 54% and 69% for the classes *Moors, heathland and sclerophyllous vegetation* and *Agro-forestry areas*, respectively.

In the content of image retrieval, Fig. 1 shows an example of a query pair and the retrieved pairs of images by these strategies. By assessing the figure, one can observe that when learning is achieved directly from BigEarthNet-MM, the semantically more similar pairs of images are retrieved, containing the *Urban fabric* and *Arable land* classes present in the query. Learning directly from BigEarthNet-MM leads to retrieval of a similar pair to the query even at the 100^{th} retrieval order. However, using transfer learning strategy results in retrieval of pairs that contain *Urban fabric* and *Arable land* classes which are not present in the query pair. One can observe this behavior even at the 5^{th} retrieved pair.

TABLE II
OVERALL MULTI-LABEL CLASSIFICATION RESULTS UNDER DIFFERENT METRICS AND DL MODELS FOR BIGEARTHNET-MM.

Model	HL	$OE(\%)$	$R(\%)$	$F_2(\%)$
VGG16	0.078	7.35	76.97	76.18
VGG19	0.080	8.12	76.17	75.35
ResNet50	0.074	**5.93**	**80.05**	**78.73**
ResNet101	0.074	6.46	78.85	77.88
ResNet152	**0.073**	6.42	78.13	77.46

The main reasons of the success of directly learning from BigEarthNet-MM are due to the fact that: 1) transfer learning from ImageNet limits the accurate characterization of the spectral content of RS images; 2) fine-tuning the pre-trained model on ImageNet by using RS images can not be sufficient to eliminate the semantic gap since the category labels present in ImageNet are different from the land-cover class labels present in BigEarthNet-MM; and 3) the pre-trained model was trained for a single-label image classification scenario, and thus limits the accurate characterization of the multiple land cover classes present in BigEarthNet-MM.

2) Comparison of State-of-the-Art CNN Models: In the second set of experiments, we compare the effectiveness of the VGG and the ResNet models in the framework of multi-modal multi-label classification. Table II shows the overall classification results under different metrics (which are the sample-averaged scores [3]). By analyzing the table, one can observe that the ResNet model provides the highest scores in all metrics. As an example, ResNet50 achieves more than 2% higher recall and F_2 scores compared to VGG models. This improvement is due to the residual connections of the ResNet model and their increased depth in terms of the number of layers compared to the VGG model. Increasing the depth of the considering models does not significantly affect the performances, i.e., similar scores are obtained in all the metrics under different depth values of the same model.

IV. DISCUSSION AND CONCLUSION

In this paper, we have presented the BigEarthNet-MM benchmark archive that contains 590,326 pairs of Sentinel-1 and Sentinel-2 image patches with a new CLC-based class-nomenclature to pay more justice to the properties of the considered images. BigEarthNet-MM makes a significant advancement for the use of DL in RS, opening up promising directions to support research studies in the framework of multi-modal multi-label RS image scene classification and retrieval. BigEarthNet-MM is suitable to assess DL based methods for: i) learning from class-imbalanced multi-modal

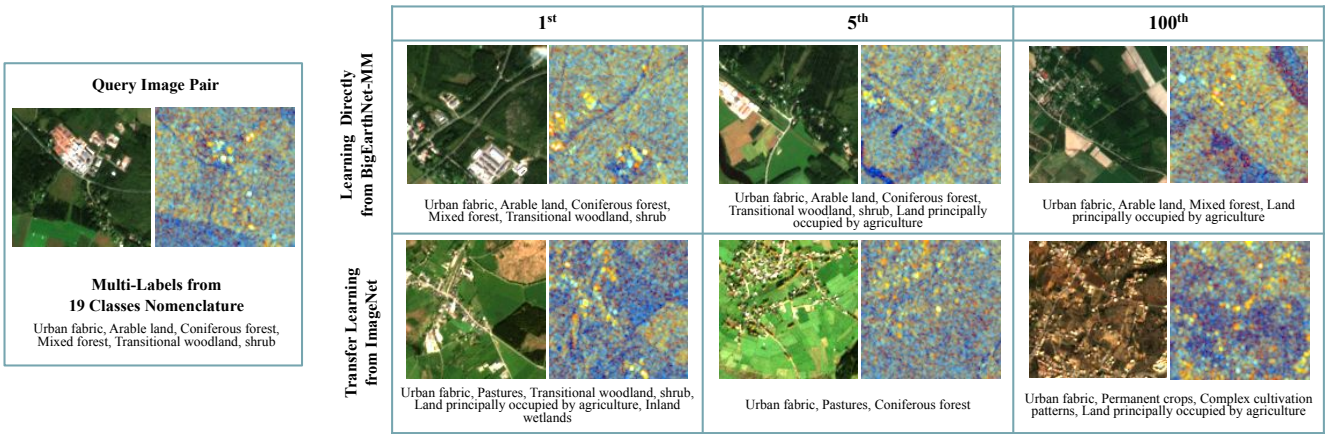

Fig. 1. An example of a query pair from the BigEarthNet-MM archive and retrieved image pairs obtained by using: 1) direct learning from BigEarthNet-MM; and 2) transfer learning from ImageNet in the framework of content-based multi-modal multi-label image retrieval.

data (since the LCLU classes are not equally represented in BigEarthNet-MM); ii) transfer learning (since BigEarthNet-MM currently contains only pairs of images from a small number of European countries); and iii) also on unsupervised, self-supervised and semi-supervised multi-modal learning for information discovery from big data archives.

It is worth noting that BigEarthNet-MM has limitations for the RS applications that require time-series data to accurately describe LCLU classes, such as *Non-irrigated arable land*, *Permanently irrigated land*. We would like to also note that some Sentinel-1 image patches can be contaminated by artefacts caused by either well-known Radio-Frequency-Interference [16] or other dataset related issues, which are independent from the pre-processing steps applied while constructing BigEarthNet-MM. As a final remark, we would like to point out that due to the use of labels from the CLC map, the BigEarthNet-MM archive can be extended to a larger scale within Europe with zero-annotation cost. As a future development of this work, we plan to enrich the BigEarthNet-MM archive by extending it to whole Europe.

ACKNOWLEDGMENT

This work is funded by the European Research Council (ERC) through the ERC-2017-STG BigEarth Project under Grant 759764 and by the German Ministry for Education and Research as BIFOLD - Berlin Institute for the Foundations of Learning and Data (01IS18025A). The authors from DGT are supported through FCT (Fundação para a Ciência e a Tecnologia) - UIDB/04152/2020 - Centro de Investigação em Gestão de Informação (MagIC).

REFERENCES

[1] S. Roy, E. Sangineto, B. Demir, and N. Sebe, "Metric-learning-based deep hashing network for content-based retrieval of remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, 2020, accepted for publication.

[2] Y. Li, Y. Zhang, X. Huang, H. Zhu, and J. Ma, "Large-scale remote sensing image retrieval by deep hashing neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 950–965, 2018.

[3] G. Sumbul and B. Demir, "A deep multi-attention driven approach for multi-label remote sensing image classification," *IEEE Access*, vol. 8, pp. 95 934–95 946, 2020.

[4] G. Cheng, X. Xie, J. Han, L. Guo, and G. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 3735–3756, 2020.

[5] Y. Long, G.-S. Xia, S. Li, W. Yang, M. Y. Yang, X. X. Zhu, and D. Zhang, L. Li, "On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 4205–4230, 2021.

[6] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, and B. Zhang, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, 2020.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 770–778.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Intl. Conf. Learn. Represent.*, 2015.

[9] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[10] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "BigEarthNet: A large-scale benchmark archive for remote sensing image understanding," in *IEEE Intl. Geosci. Remote Sens. Symp.*, 2019, pp. 5901–5904.

[11] J. Feranec, T. Soukup, G. Hazeu, and G. Jaffrain, *European Landscape Dynamics: CORINE Land Cover Data*. Boca Raton, FL, USA: CRC Press, 2016.

[12] G. Jaffrain, C. Sannier, A. Pennec, and H. Dufourmont, "Corine land cover 2012 - final validation report," European Environment Agency, Tech. Rep., 2017. [Online]. Available: https://land.copernicus.eu/user-corner/technical-library/clc-2012-validation-report-1

[13] C. Paris, L. Bruzzone, and D. Fernández-Prieto, "A novel approach to the unsupervised update of land-cover maps by classification of time series of multispectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4259–4277, July 2019.

[14] S. Arnold, B. Kosztra, G. Banko, G. Smith, G. Hazeu, M. Bock, and N. Valcarcel Sanz, "The eagle concept—a vision of a future european land monitoring framework," in *Proceedings 33th EARSeL Symposium towards Horizon*, vol. 2020. Citeseer, 2013, pp. 551–568.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Intl. Conf. Learn. Represent.*, 2014, pp. 1–41.

[16] M. Tao, J. Su, Y. Huang, and L. Wang, "Mitigation of radio frequency interference in synthetic aperture radar data: Current status and future trends," *Remote Sensing*, vol. 11, no. 20, p. 2438, 2019.