Article

SatImNet: Structured and Harmonised Training Data for Enhanced Satellite Imagery Classification

Vasileios Syrris [1,*], Ondrej Pesek [2] and Pierre Soille [1]

[1] European Commission, Joint Research Centre (JRC), 21027 Ispra, Italy; pierre.soille@ec.europa.eu
[2] Faculty of Civil Engineering, Czech Technical University in Prague, 16636 Prague, Czechia; ondrej.pesek@ext.ec.europa.eu
* Correspondence: vasileios.syrris@ec.europa.eu; Tel.: +39-0332-789525

Abstract: Automatic supervised classification with complex modelling such as deep neural networks requires the availability of representative training data sets. While there exists a plethora of data sets that can be used for this purpose, they are usually very heterogeneous and not interoperable. In this context, the present work has a twofold objective: i) to describe procedures of open-source training data management, integration, and data retrieval, and ii) to demonstrate the practical use of varying source training data for remote sensing image classification. For the former, we propose SatImNet, a collection of open training data, structured and harmonized according to specific rules. For the latter, two modelling approaches based on convolutional neural networks have been designed and configured to deal with satellite image classification and segmentation.

Keywords: image classification; semantic segmentation; remote sensing; convolutional neural network; data management; data fusion; open data sets

1. Introduction

Data-driven modelling requires sufficient and representative samples that capture and convey significant information about the phenomenon under study. Especially in the case of deep and convolutional neural networks (DNN–CNN), the usage of big training sets is a requisite to estimate adequately the high number of model weights (i.e., the strength of the connection between neural nodes) and avoid over-fitting. The lack of sizeable and labelled training data may be addressed by data augmentation [1], deep modelling techniques such as the generative adversarial networks [2], transfer learning [3] and domain adaptation [4]. Regarding transfer learning, for instance, there exist large collections of pre-trained models dealing with image classification [5,6]. However, these models have been trained on true colour images showing humans, animals, or landscapes, and it remains an open question whether transfer learning, without meticulous domain adaptation and fine-tunings, improves the generic purpose classification or segmentation of satellite images with various spectral, spatial, and temporal resolutions.

The collection of good quality training sets for supervised learning is an expensive, error-prone [7], and time-consuming procedure. It involves manual or semi-automatic label annotation, verification, and deployment of a suitable sampling strategy like systematic, stratified, reservoir, cluster, snowball, time-location, and many other sampling techniques [8]. In addition, consideration of sampling and non-sampling errors and biases need to be taken into account and corrected for. In satellite image classification, factors such as the spectral, spatial, radiometric, and temporal resolution, type of sensor (active or passive [9]), and data processing level (radiometric and geometric calibration, geo-referencing, atmospheric correction), to name a few, synthesise a manifold of concepts and features that need to be accounted for.

On the other hand, deep supervised learning as well as complex and multi-parametric modelling require big training sets. A potential solution to this issue is the collection of different training data sets and their exploitation in such a way that they complement each other and act in a synergistic manner. However, the joint use of two or more existing training data sets require careful examination of their individual features and organisation. To ease this process, we propose a methodology to organise open and freely

available training data sets designed for satellite image classification in view of fusing them with other Earth observation (EO)-based products. The proposed methodology is then applied to a series of seven free and open data sets and leads to the SatImNet collection. The process of information retrieval has been optimized over a distributed disk storage system. We also demonstrate the blending of different information layers by exploiting deep neural network modelling approaches with the objective to solve concurrently the tasks of image classification and semantic segmentation. This work ties in with the framework of the European strategy for open science [10] which strives to make research more open, global, collaborative, reproducible, and verifiable. Accordingly, all the data, models, and programming code presented herein are provided under the FAIR [11] (findable, accessible, interoperable, and reusable) conditions.

The paper is structured as follows: Section 2 discusses the significant features that make the varying source training sets interoperable for satellite image classification, and Section 3 introduces the SatImNet collection which organises existing training sets in an optimised and structural way. Section 4 demonstrates CNN models that have been trained on blended training sets and solve satisfactorily a satellite image classification and semantic segmentation task. Section 5 describes the computing platform over which the experimental process has been performed. Section 6 underlines the contribution of the present work and outlines the way forward.

2. Major Features of an Interoperable Training Set

In the following, we define the minimal and essential attributes that should characterise a training set when examined under the prism of interoperability and completeness. For best clarity, we have grouped the attributes according to three categories enumerated with their associated attributes hereafter.

1. *Attributes related to the scope of the training data:*

    - Classification problem: denotes the family/genre of the supervised learning problem that the specific data set can serve in a more effective way;
    - Intended application: explains the primary purpose that the specific data set serves according to the data set designers;
    - Definition of classes: signifies how the class annotations are originally provided by the data set designer;
    - Annotation: provides information about the class label annotation, whether derived from a manual or automated procedure, or it is based on expert opinion, volunteering effort, or organised crowd-sourcing; it serves as a qualitative indicator about the reliability of the provided data;
    - Verification: linked with the former feature, it refers as well to the belief level and reliability of the information transmitted by the data;
    - Licence: the existence of terms, agreements, and restrictions as were explicitly stated by the data publishers;
    - URL: the original web link to the data.

2. *Attributes related to the usage and sustainability of the training data:*

    - Geographic coverage: reveals the terrain characteristics, the morphological features of the objects that shape the surface, and the landscape variability, as well as the potential irregularities covered by the candidate data set;
    - Timestamp: the image sensing time or the time window to which the image refers is crucial information for change detection and seasonality-based applications. This piece of information is closely related to the concept of temporal resolution but cannot be used interchangeably;
    - Data volume: helps to determine disk storage and memory requirements;
    - Data lineage: necessary information for precise reproducibility of the data processing workflow, including pre-processing transformations such as standardisation, normalisation, clipping of values, quantisation, projection, resampling, correction (atmospheric, terrain, cirrus), etc;

- Name of classes: the semantic name that describes the category to which a single pixel, a group of pixels, or a rectangular window (patch) belongs to;
- Number of classes: shows the plurality and the exhaustiveness of the targets to be detected or identified;
- Naming convention: whether the file name conveys additional information such as sensing time, class name, location and so on;
- Quality of documentation: a qualitative annotation assigned by the users (in this case, by us) about the existence of sufficient explanatory material;
- Continuous development: a qualitative indicator about data sustainability, error correction and quality improvement that is based on the information provided by the data designers.

3. *Intrinsic image attributes:*

- File format: indicates file compression, the file reader and encoding/decoding type, availability of meta-information (e.g., GeoTIFF, PNG, etc.), number of channels/bands;
- Image dimensions: a quick reference (image height and width) to estimate the batch size during the training phase, expressed as image rows $\times$ columns;
- Number of bands: number of channels packed into a single file or number of separate single-band files belonging to a dedicated subfolder associated with the image name;
- Name of bands: standard naming of the image channels like RGB (red/green/blue) or specific naming that follows the product convention such as the naming of Sentinel-2 products;
- Data type per band: essential information about the range of band values that differentiates the data distributions and impacts the data normalisation/standardisation operations;
- No data value: specifies the presence of invalid values that affect processes such as data normalisation and masking;
- Spatial resolution: it determines what types of targets can be detected and indirectly points at the physical size of the training samples. Spatial resolution is often expressed in meters;
- Spectral resolution: it refers to the capacity of a satellite sensor to measure specific wavelengths of the electromagnetic spectrum. In data fusion context, spectral resolution helps to compare and match image bands delivered by different sensors;
- Temporal resolution: the amount of time that elapses before a satellite revisits a particular point on the Earth's surface. Although temporal resolution is an important attribute, there are no training sets that currently cover this aspect in detail;
- Type of original imagery: a piece of information with reference to the sensor type and source, the availability of masks, the existence of geo-reference, and other auxiliary details;
- Orientation: information referring mainly to image or target rotation/positioning; in the case of non-explicit statement, this feature contains basic photogrammetry information such as rectification and geometric correction;
- Metadata: extra information that accompanies an image. It concerns mostly the geo-referenced and time-stamped images.

Additional informative features could be the different/alternative usage of a data set with respect to other research works, number of citations, impact gauging, and the physical location (file path) at the local or remote storage system. In some cases, data publishers provide the date of the data set release, but this should not be confused with the formerly mentioned, pivotal attribute of timestamp.

For clarification purposes, we also detail the image classification applications that a training set with the above mentioned attributes could serve. Image classification is a generic term to describe the process of labelling an entire image or part of it. It differs from image segmentation wherein the image is partitioned in two or more connected sets of pixels; in this context, no semantics is needed since the segmentation can be based on pixel value similarities/differences and connectivity rules [12]. Depending on

the granularity of classification (pixel, group of pixels, block/window/patch), which is strongly impacted by the spatial resolution of the input imagery (it can be equally extended to the temporal resolution as well), we distinguish the following applications that can be considered as a semantic segmentation (a label is assigned to every segment/part of the image) at different levels of detail:

- Block/window/patch-based classification: In this case, the input image is divided in several overlapping or non-overlapping units (blocks/windows/patches) and all or part of these units receive a label by the classifier. The assumption under this configuration is that a considerable amount of pixels that compose the unit belong to one class, the one that has been assigned to the unit by the classifier. Labelling a unit signifies that the specific unit contains the target which is subject to detection. Target localisation is when drawing a rectangle which actually frames the image area that incorporates the target.
- Group of pixels classification: A label is assigned to spatially connected pixels, forming patterns with concrete dimensions and structure that resemble pre-defined models of physical or artificial objects. This type of application is known as instance segmentation or object delineation.
- Pixel-wise classification: The target is every single pixel that can form a distinct class. The pixel labelling is performed by the classifier without considering the classification of the adjacent pixels. This type of operation occurs i) when the spatial resolution is lower than the real dimensions of the target (for instance, the finer resolution of the Sentinel-2 radiometric bands is 10 m, permissible for a building detection but not admissible for the detection of cars), or ii) for land cover classification or big areas identification such as the road or train network.

The term target conveys the abstract concept of a pattern that is detectable and sometimes identifiable. It is subjected to the satellite sensor capacity according to the supported spatial, spectral, temporal, and radiometric resolution. Pixel targets are considered mostly for classification of land cover types such as forest, urban, water, crops, bare soil, etc., i.e., classes that may have irregular morphological characteristics such as shape or compactness. Group of pixels are considered for object detection or recognition. In this context, the pixel formation visually resembles real-world objects such as trees, houses, cars, boats, etc., usually targets with concrete structure, distinct boundaries, cohesion, and other characteristics that make them separable from their surrounding area. Patches (rectangular windows) can be employed in both of the aforementioned applications. In object detection, the image patch encloses completely the target, i.e., comprises the group of pixels that form the object. In land cover classification, the rectangular window surrounds an ample number of pixels that represent a single land cover class or a mixture of land cover types.

3. SatImNet Collection

In this section, we describe the initial edition of the SatImNet (Satellite Image Net) collection, a compilation of seven open and free training sets targeting various EO applications. Then, we elaborate on the rationale behind our choices for the data structuring. Lastly, we tabulate the training sets under consideration according to the defined attributes of Section 2.

3.1. Description of the Training Sets

The initial edition of SatImNet consists of seven diverse training sets:

1. DOTA: A large-scale Dataset for Object deTection in Aerial images, used to develop and evaluate object detectors in aerial images [13];
2. xView: contains proprietary images (DigitalGlobe's WorldView-3) from complex scenes around the world, annotated using bounding boxes [14];
3. Airbus-ship: combines Airbus proprietary data with highly trained analysts to support the maritime industry and monitoring services [15];

4. Clouds-s2-taiwan: contains Sentinel-2 True Colour Images (TCI) and corresponding cloud masks, covering the area of Taiwan [16];
5. Inria Aerial Image Labeling: comprises aerial ortho-rectified colour imagery with a spatial resolution of 0.3 m and ground truth data for two semantic classes (building and no building) [17];
6. BigEarthNet-v1.0: a large-scale Sentinel-2 benchmark archive consisting of Level-2A (L2A: Bottom Of Atmosphere reflectance images derived from the associated Level-1C products) Sentinel-2 image patches, annotated by the multiple land-cover classes that were provided from the CORINE Land Cover database of the year 2018 [18];
7. EuroSAT: consists of numerous Level-1C (L1C: top-of-atmosphere reflectances in cartographic geometry) Sentinel-2 patches provided in two editions, one with 13 spectral bands and another one with the basic RGB bands; all the image patches refer to 10 classes and are used to address the challenge of land use and land cover classification [19].

With regard to satellite imagery, one of the unique features is the spatial resolution which determines substantially the type of application. The high-resolution imagery provided by DOTA, xView, Airbus-ship, and Inria Aerial Image Labeling is suitable for object detection, localisation, and semantic segmentation. The remaining three data sets are fitting mostly applications relevant to both image patch and pixel-wise classification.

As last note, we underline the fact that although the images of xView and Airbus-ship data sets have an ownership, they are provided for free under the licenses described in the subsequent Table 1.

3.2. SatImNet Data Model

This section explains the rationale behind the chosen data model. The SatImNet collection has been structured in a modular fashion, preserving the unique characteristics of each constituent data set while providing a meta-layer that acts in a similar way as an ontology does, recording links and modelling relations among concepts and entities from the different data sets. The first two abstract layers of this meta-layer consist of the following keys: (1) built-up: residential, industrial, facilities, infrastructure, construction, areas; (2) transport means: vehicle, flying, vessel; (3) object: man-made; (4) natural areas: air, land, water. The third layer is composed by the leaf nodes (terminal nodes) representing all the classes of the seven data sets. While the intuitive choice was to consolidate the names of the classes, we decided to leave the original class names in order to retain a backward compatibility from SatImNet to the seven data sets. Accordingly, we preserve class names such as "residential building" although there is another leaf node "bulding" and a parent node "residential". Another consequence is that leaf nodes have duplicates, as it is the case of "damaged building" that belongs to more than one parent nodes such as the "residential" and "industrial" nodes.

Apart from the meta-layer that has formed as a lattice, the other structures that compose the data model are represented by nested short tree hierarchies which have been proven to be quite efficient in information retrieval tasks [20]. Accordingly, a *json* file *content_public.json* has been created for each data set that contains mostly the intrinsic attributes of the images of the specific data set in a hierarchical, tree-based format. These attributes represent information such as the physical path of the file, its size in bytes, the file type (genre), the image acquisition time or the time the image refers to, the class label (if any), the meta-information like projection, number of bands and so on. Figure 1 displays a subset of the hierarchy and shows the typical route a query follows across the central semantic meta-layer and each information module which condenses the essential information that characterises every file.

Since the entire or part of the collection is accessible over the network, we selected a database-free solution for the tree hierarchies based on *json* files. This portable layout grants a standalone character to the modules of the collection, independent of specialised software and transparent to the non-expert end-user. The lack of indexing which impacts critically the query speediness is tackled (whenever is feasible) by keeping the depth and breadth of every single tree in moderate sizes. A consequence of this is the creation of multiple *json* files for each data set (e.g., 13 files in the case of BigEarthNet-v1.0).

At this point, we underline the fact that the baseline system upon which we optimise all the processes is the EOS open-source distributed disk storage system developed at CERN [21], having as front-end a multi-core processing platform [22,23]. This configuration allows multi-tasking and is suitable for distributed information retrieval out of many files. One bounding condition set by the baseline system is the prevention of generating many small-sized files, given that EOS guarantees minimal file access latencies via the operation of in-memory metadata servers which replicate the entire file structure of the distributed storage system. For this reason, the files of the training sets have been zipped into larger archives, the size of which has been optimised in a way as to allow admissible information retrieval whilst sustaining efficient data transfer across the network. Reading individual files from *zip* files can be achieved through various interfaces and drivers. In our case, we employ the open source Geospatial Data Abstraction Library (GDAL) [24] which provides drivers for all the standard formats of raster and vector geospatial data. Jupyter notebooks demonstrating the execution of queries as well as the respective information retrieval from the *json* files and subsequently by the *zip* archives are referred to in Section 5. Although the decision to zip the files was based on the technical characteristics of the multi-petabyte EOS open storage solution, the files can be of course unzipped if more suitable in other environments. To harmonise the class annotations provided in different file formats (json, geojson, text, and csv), binary or labelled image masks were created for every single training sample. Although the aforementioned data model has been optimised upon a distributed storage system, it turns out to be a general-purpose model built on the principles of simplicity, speediness, extensibility, customisation, and portability.

3.3. Characterisation of the Data Sets

The three attribute categories presented in Section 2 have been used as a basis to characterise the seven training sets of SatImNet. Tables 1–3 respectively show the results of this analysis. In Table 1, the feature *Conversion of class representation* has been added, indicating whether the original type of class annotations has been converted into an image mask upon creation of SatImNet. The dash "-" has been used to denote either non-existent information or non-clear definition.

Table 1. Attributes related to the scope of the training data.

	DOTA	xView	Airbus-Ship	Clouds-s2- Taiwan	Inria Aerial Image Labeling	BigEarthNet- v1.0	EuroSAT
Classification problem	object detection	object detection	object detection	pixel-based detection	pixel-based & object detection	patch-based land cover classification	patch-based land cover classification
Intended application	object detection in aerial images	(1)	locate ships in images	clouds classification	building classification	land cover classification	land use and land cover classification
Definition of classes	bounding boxes in txt	bounding boxes in geojson	boxes encoding in csv	GeoTIFF images	GeoTIFF images	tags in json	name of the files: RGB jpg; 13-band GeoTIFF
Conversion of class representation	txt to png	geojson to GeoTIFF	csv to png	no conversion	no conversion	no conversion	no conversion
Annotation	manual; experts	-	-	manual	(2)	based on CLC 2018	manual
Verification	visual	-	-	-	visual	visual	visual
Licence	For academic purposes	CC BY-NC-SA 4.0	Non-commercial purposes	-	Non-commercial purposes	CDLA- Permissive-1.0	-
URL	https://captain-whu.github.io/DOTA/dataset.html	http://xviewdataset.org	https://www.kaggle.com/c/airbus-ship-detection/overview	https://www.mdpi.com/2072-4292/11/2/119/s1	https://project.inria.fr/aerialimagelabeling	http://bigearth.net	https://github.com/phelber/eurosat

[1] Enables discovery of more object classes; improves detection of fine-grained classes. [2] Combines public domain imagery with public domain official building footprints.

Table 2. Attributes related to the usage and sustainability of the training data.

	DOTA	xView	Airbus-Ship	Clouds-s2- Taiwan	Inria Aerial Image Labeling	BigEarthNet- v1.0	EuroSAT
Geographic coverage	variable	1415 km^2 (1)	variable	Taiwan	810 km^2 (2)	10 European countries (3)	34 European cities
Timestamp	-	-	-	May 2018	-	June 2017–May 2018	variable
Data volume	19.9 GB	36.4 GB	29.5 GB	123 MB	25.3 GB	106 GB	2.88 GB
Data lineage	-	-	-	-	-	sen2cor	L1C
Name of classes	(4)	(5)	ship/no ship	cloud/no cloud	building/no building	CLC nomenclature (6)	(7)
Number of classes	15	60	2	2	2	12 (8)	10
Naming convention	no	no	no	yes	yes	yes	at class level
Quality of documentation	good	moderate	not detailed	good	good	very good (9)	good
Continuous development	-	-	no	no	no	yes	-

[1] part of cities around the world. [2] towns around the world: Austin, Chicago, Kitsap County, Western Tyrol, Vienna. [3] Austria, Belgium, Finland, Ireland, Kosovo, Lithuania, Luxembourg, Portugal, Serbia, Switzerland. [4] plane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, large vehicle, small vehicle, helicopter, roundabout, soccer ball field, swimming pool (and one additional: container-crane). [5] various aircraft types, vehicles, boats/vessels, buildings, and man-made objects such as containers, pylons, and towers. [6] level-3 CORINE Land Cover class labels. [7] industrial buildings, residential buildings, annual crop, permanent crop, river, sea & lake, herbaceous vegetation, highway, pasture, forest. [8] multiple classes per patch. [9] http://bigearth.net/static/documents/BigEarthNetManual.pdf.

Table 3. Intrinsic image attributes.

	DOTA	**xView**	**Airbus-Ship**	**Clouds-s2- Taiwan**	**Inria Aerial Image Labeling**	**BigEarthNet- v1.0**	**EuroSAT**
File format	png	GeoTIFF	jpg	GeoTIFF	GeoTIFF	GeoTIFF	RGB: jpg; 13-band: GeoTIFF
Image dimensions (rows × cols)	from 800 × 800 to about 4000 × 4000	various small patches	768 × 768	224 × 224	5000 × 5000	120 × 120, 60 × 60, or 20 × 20	64 × 64
Number of bands	3	3	3	10	3	12	13-band & RGB
Name of bands	RGB	RGB and 8-band	RGB	Sentinel-2 bands	RGB	Sentinel-2 L2A bands	class name
Data type per band	8	8	8	16	8	16	RGB: 8; 13 band: 16
No data value	-	0	-	-	0 (2)	-	-
Spatial resolution	variable	0.3 m	-	20 m	0.3 m	10 m; 20 m; 60 m	10 m; 20 m; 60 m
Spectral resolution	-	-	-	(1)	-	(1)	(1)
Type of original imagery	multiple sensors; non geo-referenced	WV-3; geo-referenced	non geo-referenced	geo-referenced	geo-referenced	Sentinel-2 patches; geo-referenced	Sentinel-2 patches; geo-referenced
Orientation	variable	ortho-images	ortho-images	S2 L1C ortho-images	ortho-images	S2 L2A ortho-images	S2 L1C ortho-images
Metadata	no	yes	no	yes	yes	yes	yes

[1] https://sentinel.esa.int/web/sentinel/missions/sentinel-2/instrument-payload/resolution-and-swath. [2] explicitly defined in some cases only.

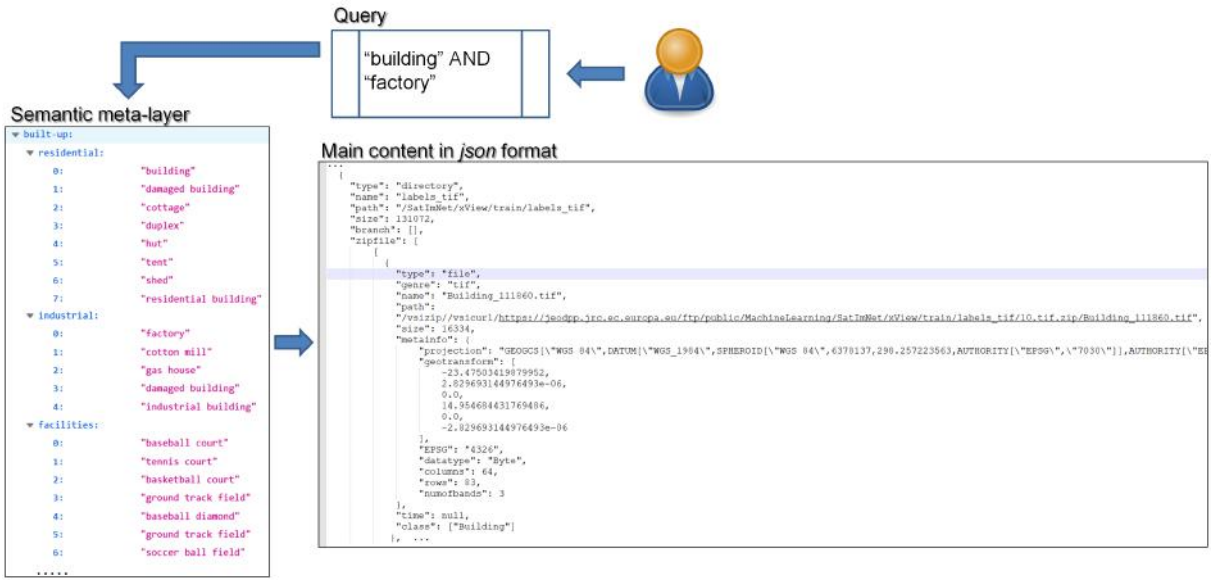

Figure 1. Schematic representation of the information retrieval task: the user forms a query by selecting one or more keywords, e.g., *building* and *factory*. Then, the query passes first through the semantic meta-layer which seeks for conceptual similarities (e.g., *built-up*, *residential*, *industrial*), and then performs search across the *json* files that retain information about the intrinsic attributes of every individual image.

4. Fusion of Heterogeneous Training Sets: Experimental Results

This section shows the added value of SatImNet for enhanced information extraction from Earth Observation products. The fusion of training sets from multiple sources is challenging for the data are created with different technical specifications: source imagery, pre-processing workflow, assumptions, manual/automatic refinement, different spatial and temporal resolution, etc. Within SatImNet context, we investigate data fusion by conducting experiments demonstrating a satellite image classification and segmentation application based on CNN models. Working within the open and free data framework, we decided to employ satellite imagery provided by the Copernicus Sentinel-2 (S2) [25] mission as input to the models since the S2 products are delivered for free and the experiments can be reproduced by anyone. The highest spatial resolution of S2 products is 10 m and fits better with applications such as land cover classification. This condition guided us to select the appropriate training data and exclude data sets that match better to object detection in very high resolution imagery. We note here that the objective of the SatImNet collection is not to provide a single big data set for any type of image classification or segmentation application but rather to offer varying source data that are homogenised to the greatest extent possible, thereby facilitating the user to select and combine different data sets that fit her purpose.

Two examples of applications are illustrated in this section. The first one deals with land cover classification formulated as a patch-based classification problem. From the SatImNet collection, we chose two sets: (i) the EuroSAT, and (ii) the BigEarthNet-v1.0 (see Tables 1–3 for the technical characteristics). We underline that the EuroSAT L1C data set has been used as a backbone set (main corpus). In some of the experiments, this set has been enriched with training samples selected from the BigEarthNet-v1.0 L2A data set: (i) we added 1000 samples from the BigEarthNet-v1.0 *Annual crops associated with permanent crops* class to the EuroSAT class *annual crop*, and (ii) we augmented the size of the EuroSAT class *forest* with 1000 samples taken randomly from the BigEarthNet-v1.0 classes *broad-leaved forest*, *coniferous forest*, and *mixed forest*.

The second application is a semantic segmentation problem. In addition to the two above mentioned data sets, we chose the following two products:

1. In connection with the water class, we used the Global Surface Water (GSW) [26], a collection of global and spatio-temporal water maps created from individual full-resolution 185 km^2 global reference

system II scenes (images) acquired by the Landsat 5, 7, and 8 satellites and distributed by the United States Geological Survey (USGS). Two out of the ten EuroSAT classes refer to water variants (*river* and *sea lake*), depicting areas that are partially or totally covered by water. From the BigEarthNet-v1.0, we randomly selected image patches referring to the classes *coastal lagoons* and *sea and ocean* (1000 samples from each category).

2. With regard to the EuroSAT classes *industrial* and *residential*, the image segmentation was based on the European Settlement Map [27] (ESM 2015, R2019 dataset), a spatial raster data set that is mapping human settlements in Europe based on Copernicus Very High Resolution (VHR) optical coverage, having 2015 as the reference year. From the BigEarthNet-v1.0, we randomly selected image patches referring to the classes *continuous urban fabric* and *discontinuous urban fabric* (500 samples from each category).

The specific data fusion approach is just one of the many combinations and associations someone could follow, and can be deemed as a representative scheme rather than the optimal strategy.

The selected BigEarthNet-v1.0 image patches have been resized from their original size of 120 × 120 to 64 × 64 images by applying a Gaussian filter to smooth the data and then by using bi-linear interpolation. On the basis of both EuroSAT and BigEarthNet-v1.0 geo-referenced image patches, we warped and clipped the GSW 2018 yearly classification layer, producing in that way the necessary water masks. Similarly, we clipped the 10 m up-scaled ESM and considered all the pixels pointing at residential and non-residential built-up areas. The non-residential built-up areas refer to detected industrial buildings, big commercial stores, and facilities. All the produced masks were resampled to 10 m spatial resolution using nearest neighbour interpolation. These masks compose an auxiliary data set that has been added to the SatImNet collection under the name BDA.

4.1. Convolutional Neural Network Modelling

Although CNNs have been experimentally proved to be more adequate for object detection [28,29] and semantic segmentation [30,31] in very high spatial resolution ($\leq$5 m), there is lately a considerable number of works [32–35] demonstrating promising results at coarser spatial resolutions such as those of the S2 imagery. Nevertheless, the majority of these works focus on image patch and not on pixel-wise classification, which is a more complex problem.

The experimental configuration in this study has been chosen to explore the challenging problem of image segmentation together with image classification in the spatial resolution of 10 m of the S2 products. In order to assess better the added value of the data fusion, rather than using pre-trained models, we have designed and tested two customised lightweight CNN approaches instead, trained from scratch for 100 epochs. The input–output schema for the CNN models is depicted in Figure 2.

The first CNN architecture (named CNN-dual) is a two-branch dual output CNN architecture that segments the image according to two classification schemas: *left-branch output* is the classification result of assigning to each pixel one of the 10 EuroSAT classes (annual crop, forest, herbaceous vegetation, highway, industrial, pasture, permanent crop, residential, river, and sea lake), and *right-branch output* is the pixel-wise classification result with reference to the aggregation classes water, built-up and other, as instructed by the GSW and ESM layers. The input to such a model is a 5 rows × 5 columns × N bands image (Figure 2). Table 4 summarises the basic parametrisation of the model. The same classification task could be formulated as a 12-class problem modelled by a single branch CNN; nevertheless, experimental results showed that CNN-dual provides consistently better results. There are three layer-couplings which intertwine the intermediate outputs across the two branches and sufficiently high structural capacity. This neural network architecture should not be confused with the twin neural network topology (Siamese) that uses the same weights while pairing two different inputs to compute comparable outputs. Figure 3 displays the CNN-dual model for the segmentation of S2 image patches.

The second CNN approach comprises two independent networks. The left network (CNN-class) as it appears in Figure 4 takes a 64 rows × 64 columns × N bands image and performs patch-based classification,

i.e., it assigns one label from the 10 EuroSAT classes to all the 64×64 image pixels (Figure 2). The right network (CNN-segm) has been designed for image segmentation according to the three aggregated classes (water, built-up, and other). CNN-segm is applied solely on the $64 \times 64 \times N$ image patches classified by CNN-class as *water* or *built-up*. In this case, the $64 \times 64 \times N$ array disintegrates in blocks of size $5 \times 5 \times N$ following a sliding-window approach. Table 4 shows the choices for the basic parameters of the two models.

Table 4. Parametrisation of the customised CNN models.

Model	# Trainable Parameters	Activation Function	Dropout Rate	Random Weights Initial- Isation	Batch Normal- Isation	Loss Function	Optimiser	Output
CNN-dual	1,259,277	relu (last layer: softmax)	0.1	He uniform variance scaling [36]	✓[37]	categorical cross- entropy	stochastic gradient descent (0.01 learning rate)	10 classes and 3 classes
CNN-class	2,507,018	relu (last layer: softmax)	0.09	Xavier normal initializer [38]	✓	categorical cross- entropy	Adam [39] (0.01 learning rate)	10 classes
CNN-segm	860,163	tanh (last layer: softmax)	0.1	Xavier normal initializer	✓	categorical cross- entropy	Adam (0.001 learning rate)	3 classes

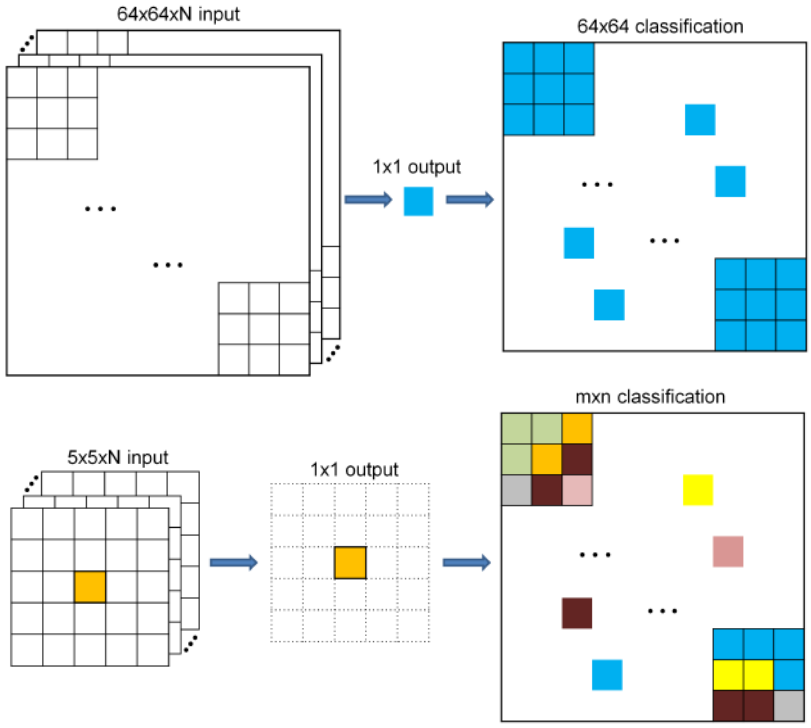

Figure 2. Input and output with respect to the CNN-class model (**top**) and both the CNN-segm and CNN-dual models (**bottom**). The variable N signifies the number of bands. The variables m and n denote the number of rows and columns, respectively, of the output image.

All the above mentioned parametrisations as well as the proposed CNN topologies derived from an extensive repetitive experimental process (threefold grid search). We note here that the purpose of this case study is not to conduct a comparison analysis of widely accepted CNN-based classification or segmentation models against the proposed ones. The presented CNN topologies are lightweight modelling approaches useful for evaluating the impact an enriched training set brings on the classification performance. Table 5 depicts the classification accuracy in terms of two metrics, the F1-score and the Kappa score, while considering the four S2 bands with 10 m spatial resolution. Regarding the CNN-dual results, there are two values for each metric corresponding to 10 (left) and 3 (right) classes, respectively. Figure 5 displays some indicative screenshots.

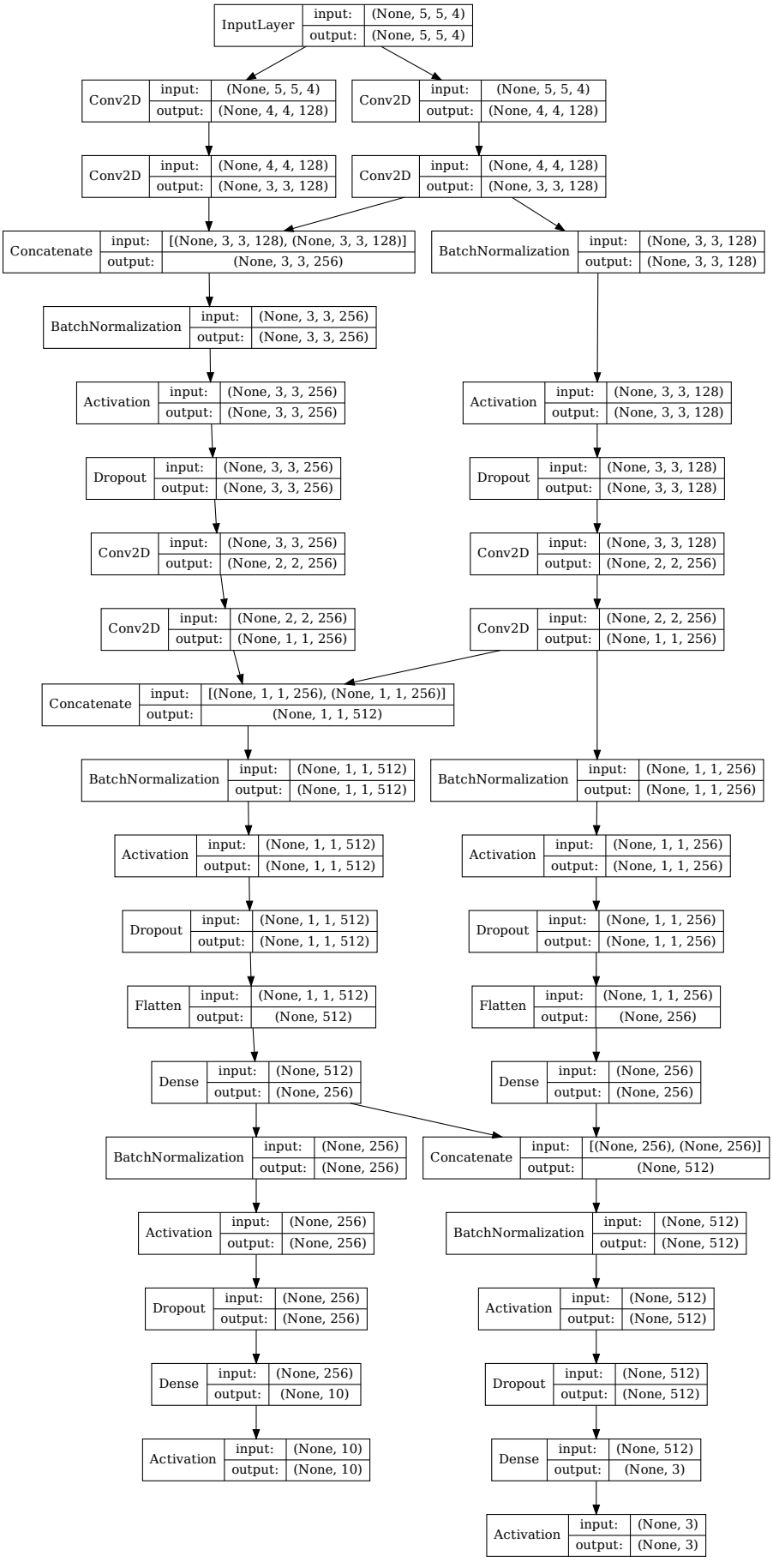

Figure 3. A two-branch dual output CNN topology for image segmentation.

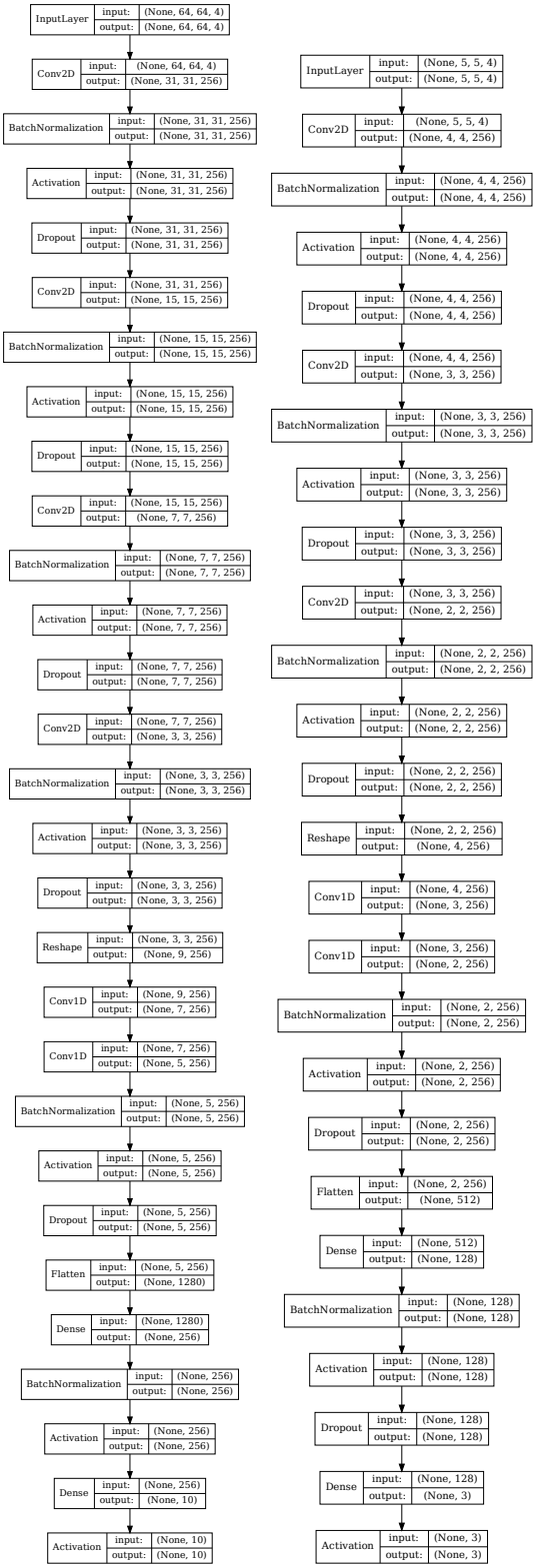

Figure 4. Two independent CNN topologies: the left one performs a patch-based classification and the right one a pixel-wise classification.

We concluded also that the four S2 10 m bands, blue (B02), green (B03), red (B04), and NIR (B08) give for the most part of the performed experiments the most consistent results. Actually, we found out that there is an intense dissimilarity between the data distributions of the EuroSAT-provided bands B09 and

B10 (the data set creators claim that all the S2 images have been downloaded via Amazon S3) and the respective bands of the MSI Sentinel-2 products we downloaded from the Copernicus Open Access Hub (https://scihub.copernicus.eu/).

Table 5. Accuracy performance by the proposed CNN modelling approaches, computed via tenfold cross-validation. The training, validation, and testing sets have been partitioned according to the rule 80/10/10. The subheadings 10cl and 3cl refer to 10-class and 3-class result, respectively.

Model	Training Set	Testing Set	F1-Score (%)		Kappa Score (%)	
Patch-based classification			10cl		10cl	
CNN-class	EuroSAT	EuroSAT	96.65		96.28	
CNN-class	EuroSAT	EuroSAT & BigEarthNet-v1.0	87.23		84.84	
CNN-class	EuroSAT & BigEarthNet-v1.0	EuroSAT	98.76		98.62	
CNN-class	EuroSAT & BigEarthNet-v1.0	EuroSAT & BigEarthNet-v1.0	94.44		93.77	
Image segmentation			10cl	3cl	10cl	3cl
CNN-dual	EuroSAT	EuroSAT	72.13	83.70	70.71	66.51
CNN-dual	EuroSAT	EuroSAT & BigEarthNet-v1.0	67.33	77.17	66.12	61.71
CNN-dual	EuroSAT & BigEarthNet-v1.0	EuroSAT	77.29	88.01	75.52	70.99
CNN-dual	EuroSAT & BigEarthNet-v1.0	EuroSAT & BigEarthNet-v1.0	74.89	86.89	72.15	67.98

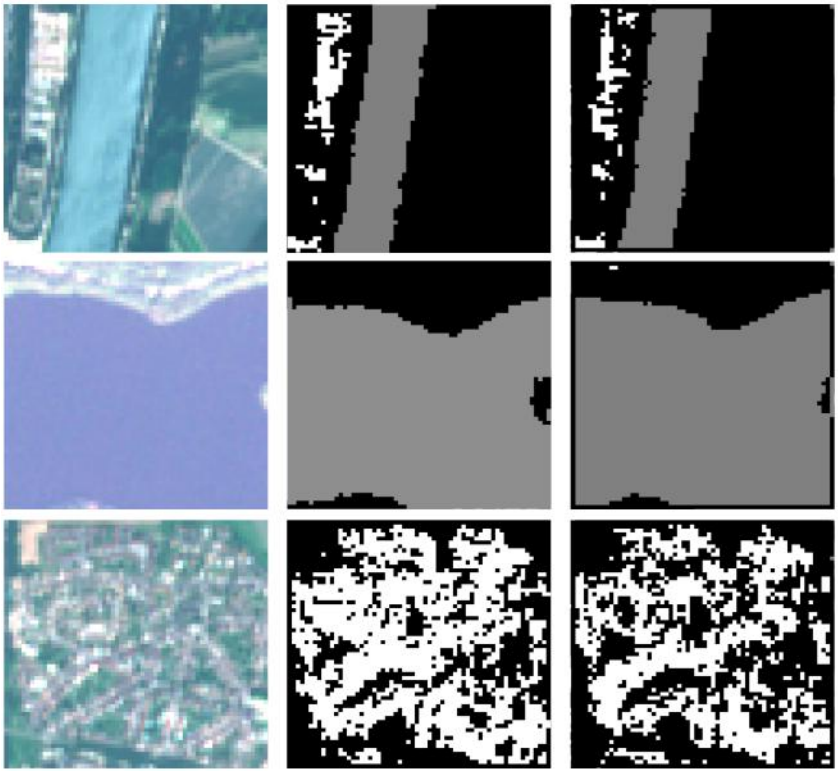

Figure 5. Image patches (64 × 64) of the EuroSAT data set kept out for testing. The first column displays the RGB composition, the middle column shows the output of the CNN-segm, and the last column displays the segmentation result of the CNN-dual model. The CNN-class classifies correctly the three image patches to *river*, *sea lake*, and *residential*, respectively.

Visual inspection of the classification results over (i) additional S2 MSI images (i.e., images not included in the collected data sets), (ii) geographic areas outside of the European continent where the training patches have been extracted, such as in China and USA (Figure 6), and (iii) Level-2A (surface reflectance in cartographic geometry) S2 products (Figure 7), while the majority of the training samples have been derived by L1C images, is in accordance with the results obtained from the L1C tests. The visual

inspection confirmed an agreement of more than 60% with respect to the segmentation results and around 80% for the patch-based classification. There is a consistent confusion among the classes *residential*, *industrial*, and *highway*, as well as among the classes *forest*, *herbaceous vegetation*, and *pasture*. The latter confusion is attributed mostly on the seasonality whereas the former happens as a result of the similar radiometric signature of the build-up structures. We observed also a known problem with the misclassification of the shadow cast by mountains or forest trees to one of the water classes. This can be potentially solved by incorporating training samples which represent this pattern. To quantitatively affirm our findings derived from visual observation, we have included an indicative confusion matrix (Figure 8) that displays the agreement between the classification result of the model CNN-dual when applied over an S2 product covering an area in USA (Brawley—south region of Salton sea) and the global land cover product FROM-GLC [40]. We employ the term agreement and not accuracy since the class nomenclature differs significantly between the two layers. To bridge the two layers, we applied the re-classification schema of Table 6. The FROM-GLC classes *tundra*, *bareland*, and *snow/ice* have not been considered. The comparison has been performed at the spatial resolution of FROM-GLC (30 meters) by up-scaling the segmentation output through the statistical *mode* operator.

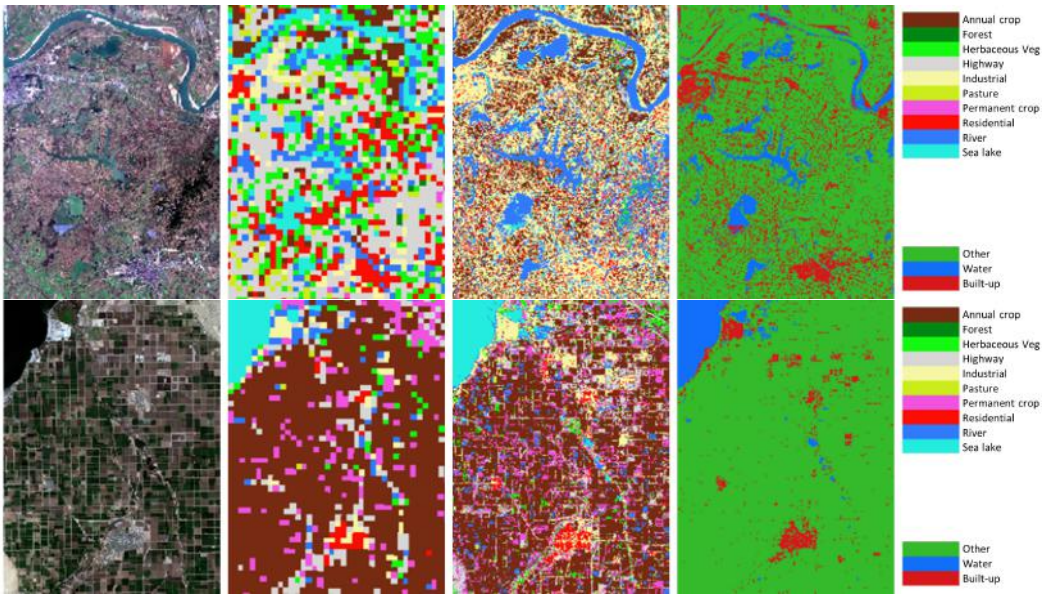

Figure 6. Transfer learning on geographic locations different than the location from which the training samples have been selected. Two exemplar areas from China and USA in the first and second row, respectively; all the training samples have been selected from Europe. The columns from left to right: (i) RGB, (ii) CNN-class, (iii) 10-class CNN-dual, and (iv) 3-class CNN-dual output.

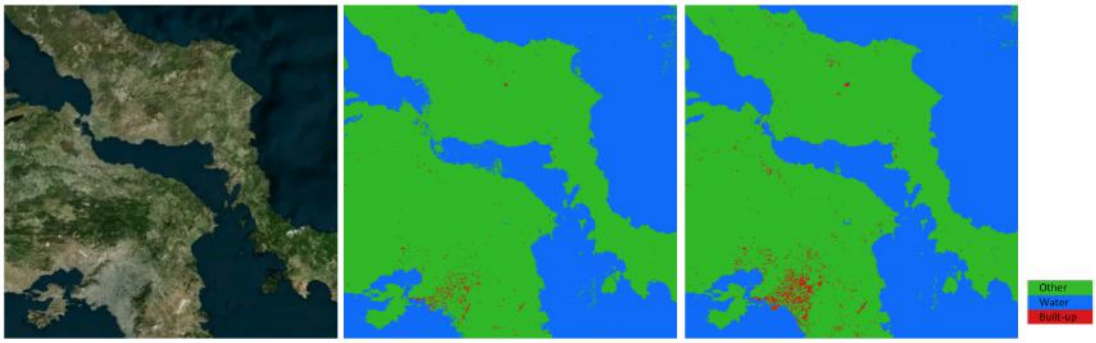

Figure 7. Testing model robustness on L2A product (geographic area in Europe). The columns from left to right: (i) RGB, (ii) 3-class CNN-segm, and (iii) 3-class CNN-dual output.

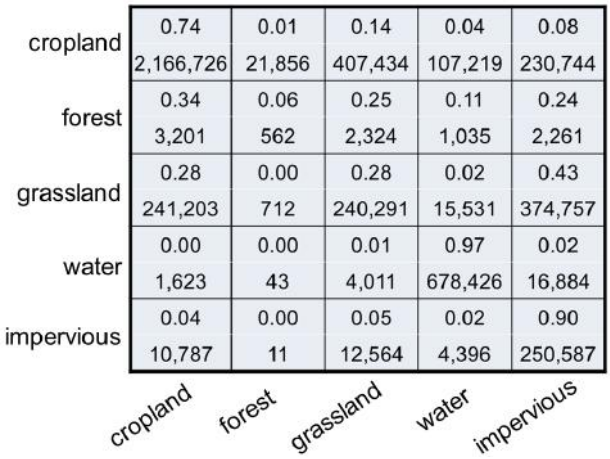

Figure 8. Confusion matrix between the CNN-dual segmentation output and the FROM-GLC land cover map at 30 m spatial resolution. The top row in each cell represents the agreement percentage and the bottom row the actual number of classified pixels.

Table 6. Re-classification of CNN-dual segmentation output according to FROM-GLC nomenclature.

EuroSAT Class		FROM-GLC Class
annual crop, permanent crop	→	cropland
forest	→	forest
herbaceous vegetation, pasture	→	grassland, shrubland
highway, industrial, residential	→	impervious surface
river, sea lake	→	water, wetland

4.2. Transfer Model

To have a more complete picture and with regards to patch-based classification, we show results (Table 7 and Figure 9) of a widely used CNN model, the *ResNet50* [41] with the following two adaptations:

- M1: We kept frozen the main topology of the model and we added one 2D convolutional layer at the beginning of the model that merges the four-band input tensor to a standard RGB tensor, and two dense layers (512 and 10 nodes, respectively) at the end of the model's architecture, resulting in 23,587,712 non-trainable parameters and 1,054,233 trainable parameters. The non-trainable parameters have been tuned based on the *ImageNet* [42] data set.
- M2: We adjusted the *ResNet50* topology in such a way to accept a four-band input tensor and, as in M1, we added two dense layers (512 and 10 nodes, respectively) at the end of the model's architecture, resulting in a new model with 53,120 non-trainable and 24,591,946 trainable parameters. The parameters of the model obtain as initial value the optimal values based on the *ImageNet* training.

Table 7. Classification accuracy of the adapted *ResNet50* models M1 and M2 (100 epochs of training).

Model	Non-Trainable Parameters	Trainable Parameters	Training	Testing	F1-Score (%)	Kappa Score (%)
M1	23,587,712	1,054,233	EuroSAT	EuroSAT	5.74	3.72
M1	23,587,712	1,054,233	EuroSAT	BigEarthNet-v1.0	6.64	4.10
M2	53,120	24,591,946	EuroSAT	EuroSAT	96.42	96.02
M2	53,120	24,591,946	EuroSAT	BigEarthNet-v1.0	88.47	87.08

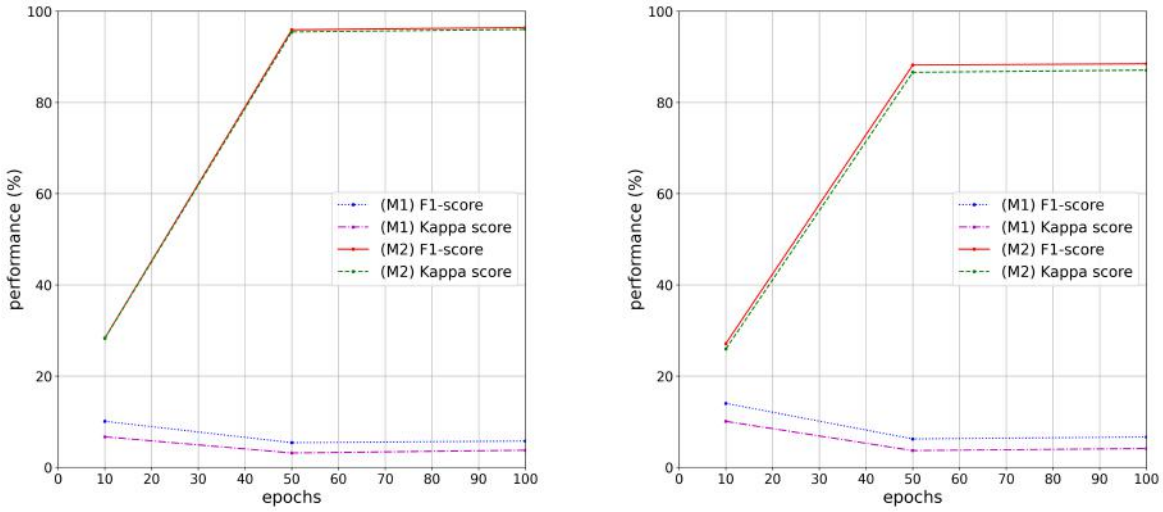

Figure 9. Accuracy performance of the adapted *ResNet50* model. M1: 23,587,712 non-trainable and 1,054,233 trainable parameters, and M2: 53,120 non-trainable and 24,591,946 trainable parameters. (**a**) Both training and testing is based on the EuroSAT set; (**b**) training with the EuroSAT set and testing on the BigEarthNet-v1.0.

4.3. Data Augmentation

So far, the data augmentation has been derived by the blending of two different training sets. In this section, we investigate whether there is a gain in terms of accuracy by employing the more conventional data augmentation by means of standard image transformations. Since the image patches of size 5×5 are already quite small and their amount reaches very high numbers, we applied the image transformations over the 64×64 image patches solely. More specifically, we applied step-wise rotation of step 90°, and flipping in the left-right and up-down direction, resulting in 162,000 samples from the initial 27,000 of EuroSat only, and in 192,300 samples from the initial 32,050 of the mixed EuroSat & BigEarthNet-v1.0 data set. The results are summarised in Table 8.

Table 8. Accuracy performance of CNN-class and adapted ResNet50 (model M2) on the 10-class patch-based classification problem, computed via tenfold cross-validation, and using data augmentation based on image transformations. The training, validation, and testing sets have been partitioned according to the rule 80/10/10.

Model	Training Set	Testing Set	F1-Score (%)	Kappa Score (%)
CNN-class	EuroSAT	EuroSAT	99.34	99.27
CNN-class	EuroSAT	EuroSAT & BigEarthNet-v1.0	89.62	88.13
CNN-class	EuroSAT & BigEarthNet-v1.0	EuroSAT	99.87	99.86
CNN-class	EuroSAT & BigEarthNet-v1.0	EuroSAT & BigEarthNet-v1.0	97.08	96.73
M2	EuroSAT	EuroSAT	98.07	97.85
M2	EuroSAT	EuroSAT & BigEarthNet-v1.0	88.32	86.22
M2	EuroSAT & BigEarthNet-v1.0	EuroSAT	99.01	98.90
M2	EuroSAT & BigEarthNet-v1.0	EuroSAT & BigEarthNet-v1.0	96.20	95.74

4.4. Experimental Findings

One of the objectives of the experimental study was to validate whether the presence of bigger amount of representative training samples impact the classification performance comparably or even better than the utilisation of pre-trained and adapted sophisticated models. To verify this hypothesis, we did not use a transfer learning approach in terms of pre-trained and pre-configured NN layers, but instead, we designed relatively light CNN topologies, keeping the number of model weights as low as possible while retaining

the structural capacity of the model in adequate levels. The maximum tenfold cross-validation overall accuracy reaches 99.37% for CNN-class and 95.41% for the 3-class output of the CNN-dual. One remark here is that the same image patches shifted for several pixels may produce different results but always in line with the most dominant class. We do not report the accuracy results of CNN-segm because they are slightly worse than the results given by CNN-dual. Nevertheless, the role of CNN-segm is to segment the patches already classified by the CNN-class model in order to refine further the classification (3-class). CNN-dual has more parameters than CNN-segm and has been designed to provide concurrently a 10-class and 3-class segmentation of the image. Both CNN-dual and CNN-segm have been trained on 15,836,588 samples and validated on 3,959,147 samples.

The summarised results of Table 5 justify comparatively the gain from the mixing of the training sets in terms of accuracy and robustness. The two measures, F1 and Kappa score, improve significantly when the training is based on both data sets than employing a single one (the EuroSAT in this case).

The results shown in Figure 8 corroborate the confusion between the "green" vegetation classes that is due to the seasonality effect. The fact that several *cropland* and *grassland* pixels have been classified to the *impervious* class can be explained by the fact that pixels reflecting rural roads or small settlements pull their surrounding pixels as well in many cases, resulting in thicker objects. This phenomenon is strongly dependent on the detection capacity of the sensor. Nevertheless, the agreement between the two layers is noteworthy if we consider that the selection of the 5×5 image patches have not undergone a refinement in terms of class labelling. More precisely, as the original 64×64 image patches have been divided into 5×5 blocks, likewise, the class label of each 64×64 image patch has been assigned to all its constituent 5×5 blocks, causing several false class assignments since there are more than one classes in most of the 64×64 image patches. Currently, we are working in this front, trying to find an automatic way for the precise assignment of class labels to the smaller blocks, something that will improve a lot the classification results.

The very low classification performance shown in Table 7 with respect to model (M1) states that this type of transfer learning is not adequate at all since the low number of trainable parameters is not sufficient to help the model adapt into the new domain. In regard to the model (M2), the same Figures show that the classification performance is similar to the results of the patch-based classification (CNN-class) as presented in Table 5, especially in the case of 100 epochs. Nevertheless, the lower number of parameters of CNN-class, adjusted from scratch upon the combined EuroSAT and BigEarthNet-v1.0 training set seems more adequate option in terms of computational complexity and adaptability than using a more complex model trained in one of the two sets.

The positive effect of data augmentation through standard image transformations is apparent as shown by the results of Table 8. However, if we compare the pair *training: BigEarthNet-v1.0 & EuroSAT* and *testing: EuroSAT* of Table 5 with the pair *training: EuroSAT* and *testing: BigEarthNet-v1.0 & EuroSAT* of Table 8, we acknowledge the fact that data augmentation via image transformations alone, when applied on EuroSAT, does not improve the classification result so much compared to the data augmentation that has been derived by the blending of the two training sets BigEarthNet-v1.0 & EuroSAT.

We do not provide comparable results of transfer learning for image segmentation because, as we have already mentioned, the original purpose of both EuroSAT and BigEarthNet-v1.0 data sets was the patch-based classification, and the only way to convert the problem into a semantic segmentation problem is to deal with image blocks of size much smaller than the 64×64 patch size supported by the data sets. Although some of the standard pre-trained segmentation models can operate with the original 64×64 image blocks for the 3-class image segmentation, they cannot answer on the 10-class image segmentation problem due to the lack of detailed reference masks. This challenging problem has been tackled by our customised CNN model (CNN-dual) which operates at pixel level and can be trained fast from scratch upon the fused data set. It supports as well our hypothesis that the existence of big enough, representative, and enriched data sets combined with customised modelling approaches provides added-value to the classification outcome compared to transfer learning.

5. Open Access to Data and Workflows

Training deep neural networks requires hardware and software libraries to be fine-tuned for array-based intensive computations. Multi-layered networks rely heavily on matrix math operations and demand immense amounts of computing capacity (mostly floating-point). For some years now, the state of the art in such type of computing and especially for image processing is shaped by powerful machinery such as the graphics processing units (GPUs) and their optimised architectures.

In this regard, the JRC (Joint Research Centre) Big Data Analytics project, having as a major objective to provide services for large-scale processing and data analysis to the JRC scientific community and the collaborative partners, is constantly increasing the fleet of GPU-based processing nodes, including NVIDIA Tesla K80, GeForce GTX 1080 Ti, Quadro RTX 6000, and Tesla V100-PCIE cards. Dedicated Docker images with CUDA [43] parallel model, back-end, and deep learning frameworks such as TensorFlow, Apache MXNet and PyTorch [44] and adequate application programming interfaces have been configured to facilitate and streamline the prototyping and large-scale testing of working models.

We mention these technical details in order to underline the fact that although operations such as transfer learning, domain adaptation, model customization, and hyper-parameter fine-tuning are much lighter than training and optimising a deep neural network from scratch, they also require dedicated hardware and software for the exploration of various scenarios and the shortening of the experimentation process.

The entire experimental setting presented here has been performed onto the JRC's high-throughput computing platform, the so-called JEODPP [23]. Jupyter notebooks and Docker images are open and accessible upon request. This decision is in conformity with the FAIR Guiding Principles for scientific data management and stewardship, and promotes open science. Complete or sectional download of the SatImNet collection can be done via *ftp* (https://jeodpp.jrc.ec.europa.eu/ftp/public/MachineLearning/SatImNet) service. In addition, individual files are directly accessible through the support of *vsizip* and *vsicurl* drivers. The open repository (https://github.com/syrriva/SatImNet) contains the Python scripts (in the form of Jupyter notebooks) to access and query the SatImNet collection via *ftp*.

6. Conclusions

The availability and plurality of well-organised, complete, and representative data sets is critical for the efficient training of machine learning models (specifically of deep neural networks) in order to solve satellite image classification tasks in a robust and operative fashion. Working under the framework of open science and very closely to policy support which invites for transparent and reproducible processing pipelines at which data and software are integrated, are open and freely available through ready-to-use working environments, the contribution of this paper is aligned with three goals: (i) to define the functional characteristics of a sustainable collection of training sets, aiming at covering the various specificities that delineate the landscape of automated satellite image classification; (ii) according to the defined attributes, to structure and compile a number of heterogeneous data sets, and (iii) to demonstrate a potential fusion of training sets by using deep neural network modelling and solving concurrently an image classification and segmentation problem.

Future work involves systematic harvesting of training sets across Internet, automation of the quality control of the discovered data sets, and continuous integration of the distinct modules of the working pipeline. Apart from the accumulation of data sets which have been designed and provided by the research community, another scheduled activity concerns the methodical building of in-house training sets, targeting wide scope Earth observation applications such as crop and surface water monitoring, deforestation, and crisis management.

Author Contributions: V.S. and P.S. planned the formulation of the SatImNet data set; V.S. and O.P. selected the open data and O.P. downloaded the data sets in their original version; V.S., O.P., and P.S. defined the attributes on which the data sets have been categorised; V.S. designed and implemented the experimental study, and structured the data; V.S. created all the Jupyter notebooks and O.P. tested them; V.S. wrote the paper and all the authors reviewed and edited it. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The authors would like to thank Tomáš Kliment, Panagiotis Mavrogiorgos, Pier Valerio Tognoli, and Paul Hasenohr for their contribution in data management, ftp service set up, and Docker images configuration and maintenance.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shorten, C.; Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 1–48.
2. Howe, J.; Pula, K.; Reite, A.A. Conditional generative adversarial networks for data augmentation and adaptation in remotely sensed imagery. In *Applications of Machine Learning*; Zelinski, M.E., Taha, T.M., Howe, J., Awwal, A.A.S., Iftekharuddin, K.M., Eds.; International Society for Optics and Photonics, SPIE: San Diego, California, United States, 2019; Volume 11139, pp. 119–131. doi:10.1117/12.2529586.
3. Pan, S.J.; Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. doi:10.1109/TKDE.2009.191.
4. Wang, M.; Deng, W. Deep visual domain adaptation: A survey. *Neurocomputing* **2018**, *312*, 135–153. doi:10.1016/j.neucom.2018.05.083.
5. Hoeser, T.; Kuenzer, C. Object Detection and Image Segmentation with Deep Learning on Earth Observation Data: A Review-Part I: Evolution and Recent Trends. *Remote Sens.* **2020**, *12*, 1667. doi:10.3390/rs12101667.
6. Bianco, S.; Cadène, R.; Celona, L.; Napoletano, P. Benchmark Analysis of Representative Deep Neural Network Architectures. *IEEE Access* **2018**, *6*, 64270–64277. doi:10.1109/ACCESS.2018.2877890.
7. Ball, J.E.; Anderson, D.T.; Chan, C.S. Comprehensive survey of deep learning in remote sensing: Theories, tools, and challenges for the community. *J. Appl. Remote Sens.* **2017**, *11*, 042609.
8. Thompson, S.K. *Sampling*, 3rd ed.; John Wiley & Sons, Inc., USA: 2012. doi:10.1002/9781118162934.
9. Schott, J.R. *Remote Sensing: The Image Chain Approach*, 2nd ed.; Oxford University Press, USA: 1996.
10. European Commission. A European Strategy for Data. 2020. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1582551099377&uri=CELEX:52020DC0066 (accessed on 28 February 2020).
11. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. doi:10.1038/sdata.2016.18.
12. Soille, P. Constrained connectivity for hierarchical image partitioning and simplification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 1132–1145.
13. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
14. Lam, D.; Kuzma, R.; McGee, K.; Dooley, S.; Laielli, M.; Klaric, M.; Bulatov, Y.; McCord, B. xView: Objects in Context in Overhead Imagery. *arXiv* **2018**, arXiv:cs.CV/1802.07856.
15. Airbus-Kaggle. Airbus Ship Detection Challenge. 2018. Available online: https://www.kaggle.com/c/airbus-ship-detection (accessed on 28 February 2020).
16. Liu, C.C.; Zhang, Y.C.; Chen, P.Y.; Lai, C.C.; Chen, Y.H.; Cheng, J.H.; Ko, M.H. Clouds Classification from Sentinel-2 Imagery with Deep Residual Learning and Semantic Image Segmentation. *Remote Sens.* **2019**, *11*, 119. doi:10.3390/rs11020119.
17. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Can Semantic Labeling Methods Generalize to Any City? The Inria Aerial Image Labeling Benchmark. In Proceedings of the IEEE International Symposium on Geoscience and Remote Sensing (IGARSS), Fort Worth, TX, USA, 23–28 July 2017.

18. Sumbul, G.; Charfuelan, M.; Demir, B.; Markl, V. Bigearthnet: A Large-Scale Benchmark Archive for Remote Sensing Image Understanding. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 5901–5904. doi:10.1109/IGARSS.2019.8900532.
19. Helber, P.; Bischke, B.; Dengel, A.; Borth, D. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 2217–2226.
20. Baeza-Yates, R.; Ribeiro-Neto, B. *Modern Information Retrieval the Concepts and Technology Behind Search*, 2nd ed.; Addison-Wesley Publishing Company, USA: 2011.
21. Peters, A.; Sindrilaru, E.; Adde, G. EOS as the present and future solution for data storage at CERN. *J. Phys. Conf. Ser.* **2015**, *664*, 042042. doi:10.1088/1742-6596/664/4/042042.
22. Soille, P.; Burger, A.; Marchi, D.D.; Hasenohr, P.; Kempeneers, P.; Rodriguez, D.; Syrris, V.; Vasilev, V. The JRC Earth Observation Data and Processing Platform. In Proceedings of the Conference on Big Data from Space (BiDS'17), Toulouse, France, 28–30 November 2017; pp. 271–274. doi:10.5281/zenodo.3239211.
23. Soille, P.; Burger, A.; De Marchi, D.; Kempeneers, P.; Rodriguez, D.; Syrris, V.; Vasilev, V. A Versatile Data-Intensive Computing Platform for Information Retrieval from Big Geospatial Data. *Future Gener. Comput. Syst.* **2018**, *81*, 30–40. doi:10.1016/j.future.2017.11.007.
24. GDAL/OGR contributors. *GDAL/OGR Geospatial Data Abstraction software Library*; Open Source Geospatial Foundation: 2020.
25. ESA. Sentinel-2 Products Specification Document. 2018. Available online: https://sentinel.esa.int/web/sentinel/document-library/content/-/article/sentinel-2-level-1-to-level-1c-product-specifications (accessed on 28 February 2020).
26. Pekel, J.F.; Cottam, A.; Gorelick, N.; Belward, A. High-resolution mapping of global surface water and its long-term changes. *Nature* **2016**, *540*, 418–422. doi:10.1038/nature20584.
27. Corbane, C.; Sabo, F.; Syrris, V.; Kemper, T.; Politis, P.; Pesaresi, M.; Soille, P.; Osé, K. Application of the Symbolic Machine Learning to Copernicus VHR Imagery: The European Settlement Map. *IEEE Geosci. Remote Sens. Lett.* **2019**, 17, 1153–1157. doi:10.1109/LGRS.2019.2942131.
28. Cheng, G.; Han, J. A survey on object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. doi:10.1016/j.isprsjprs.2016.03.014.
29. Li, Y.; Zhang, H.; Xue, X.; Jiang, Y.; Shen, Q. Deep learning for remote sensing image classification: A survey. *WIREs Data Min. Knowl. Discov.* **2018**, *8*, e1264. doi:10.1002/widm.1264.
30. Witharana, C.; Civco, D.L.; Meyer, T.H. Evaluation of data fusion and image segmentation in earth observation based rapid mapping workflows. *ISPRS J. Photogramm. Remote Sens.* **2014**, *87*, 1–18. doi:10.1016/j.isprsjprs.2013.10.005.
31. Audebert, N.; Saux, B.; Lefèvre, S. *Semantic Segmentation of Earth Observation Data Using Multimodal and Multi-Scale Deep Networks*; In Computer Vision–ACCV 2016. Springer International Publishing, Switzerland: 2017; pp. 180–196. doi:10.1007/978-3-319-54181-5_12.
32. Shendryk, Y.; Rist, Y.; Ticehurst, C.; Thorburn, P. Deep learning for multi-modal classification of cloud, shadow and land cover scenes in PlanetScope and Sentinel-2 imagery. *ISPRS J. Photogramm. Remote Sens.* **2019**, *157*, 124–136. doi:10.1016/j.isprsjprs.2019.08.018.
33. Sharma, A.; Liu, X.; Yang, X.; Shi, D. A patch-based convolutional neural network for remote sensing image classification. *Neural Netw.* **2017**, *95*, 19–28. doi:10.1016/j.neunet.2017.07.017.
34. Syrris, V.; Hasenohr, P.; Delipetrev, B.; Kotsev, A.; Kempeneers, P.; Soille, P. Evaluation of the Potential of Convolutional Neural Networks and Random Forests for Multi-Class Segmentation of Sentinel-2 Imagery. *Remote Sens.* **2019**, *11*, 907. doi:10.3390/rs11080907.
35. Liu, S.; Qi, Z.; Li, X.; Yeh, A.G.O. Integration of Convolutional Neural Networks and Object-Based Post-Classification Refinement for Land Use and Land Cover Mapping with Optical and SAR Data. *Remote Sens.* **2019**, *11*, 690. doi:10.3390/rs11060690.
36. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)Santiago, Chile, 7–13 December 2015; pp. 1026–1034. doi:10.1109/ICCV.2015.123.

37. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 7–9 July 2015; Bach, F., Blei, D., Eds.; PMLR: Lille, France, 2015; Volume 37, pp. 448–456.
38. Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS'10), Sardinia, Italy, 13–15 May 2010. Society for Artificial Intelligence and Statistics : 2010.
39. Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations, Banff, AB, Canada, 14–16 April 2014.
40. Peng Gong, E.A. Stable classification with limited sample: transferring a 30-m resolution sample set collected in 2015 to mapping 10-m resolution global land cover in 2017. *Sci. Bull.* **2019**, *64*, 370–373. doi:10.1016/j.scib.2019.03.002.
41. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arxiv:1512.03385.
42. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009.
43. Whitehead, N.; Fit-florea, A. Precision & Performance: Floating Point and IEEE 754 Compliance for NVIDIA GPUs. 2018. Available online: https://developer.download.nvidia.com/assets/cuda/files/NVIDIA-CUDA-Floating-Point.pdf (accessed on 16 February 2020).
44. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Vancouver, BC, Canada, 8–14 December 2019; Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R., Eds.; Curran Associates, Inc., USA: 2019; pp. 8024–8035.