remote sensing

MDPI

Article

A Dual Network for Super-Resolution and Semantic Segmentation of Sentinel-2 Imagery

Saüc Abadal [1], Luis Salgueiro [1], Javier Marcello [2] and Verónica Vilaplana [1,*]

[1] Signal Theory and Communications Department, Universitat Politècnica de Catalunya (UPC), 08034 Barcelona, Spain; sauc.abadal@estudiantat.upc.edu (S.A.); luis.fernando.salgueiro@upc.edu (L.S.)
[2] Instituto de Oceanografía y Cambio Global (IOCAG), Unidad Asociada ULPGC-CSIC, 35017 Las Palmas de Gran Canaria, Spain; javier.marcello@ulpgc.es
* Correspondence: veronica.vilaplana@upc.edu

Abstract: There is a growing interest in the development of automated data processing workflows that provide reliable, high spatial resolution land cover maps. However, high-resolution remote sensing images are not always affordable. Taking into account the free availability of Sentinel-2 satellite data, in this work we propose a deep learning model to generate high-resolution segmentation maps from low-resolution inputs in a multi-task approach. Our proposal is a dual-network model with two branches: the Single Image Super-Resolution branch, that reconstructs a high-resolution version of the input image, and the Semantic Segmentation Super-Resolution branch, that predicts a high-resolution segmentation map with a scaling factor of 2. We performed several experiments to find the best architecture, training and testing on a subset of the S2GLC 2017 dataset. We based our model on the DeepLabV3+ architecture, enhancing the model and achieving an improvement of 5% on IoU and almost 10% on the recall score. Furthermore, our qualitative results demonstrate the effectiveness and usefulness of the proposed approach.

Keywords: super-resolution; semantic segmentation; deep learning; convolutional neural network; Sentinel-2

Citation: Abadal, S.; Salgueiro, L.; Marcello, J.; Vilaplana, V. A Dual Network for Super-Resolution and Semantic Segmentation of Sentinel-2 Imagery. *Remote Sens.* **2021**, *13*, 4547. https://doi.org/10.3390/rs13224547

Academic Editors: Yang-Won Lee, Jungho Im, Jaeil Cho and Chu-Yong Chung

Received: 4 October 2021
Accepted: 7 November 2021
Published: 12 November 2021

1. Introduction

Land Use and Land Cover (LULC) maps are essential tools for documenting the changes in the environment and for quantifying the human footprint on the Earth's surface [1]. Due to the increasing availability of high resolution, frequently collected remote sensing data, there is a clear need for a highly automated data processing workflow to update the land cover and land use changes [2].

In remote sensing (RS), two important concepts regarding satellite imagery are spatial resolution and spectral resolution. Spatial resolution is the ground area imaged for the instantaneous field of view of the sensor. The higher the spatial resolution, the more detail it will contain. Fine details like small buildings, cars and street lines can be seen in very high-resolution platforms (50 cm–1 m), on high-resolution (1–4 m) a tree or a bus can be distinguished, whilst medium/moderate-resolution images (4–50 m) will only show coarse features [3]. A sensor's spectral resolution specifies the number of spectral channels, and their bandwidth, in which the sensor can collect reflected radiance. The spectral reflectance signatures can be used to identify the mineral content of rocks, the moisture of soil, the health of vegetation, etc. In order to achieve high resolution in the spectral domain, images are captured using multispectral or hyperspectral sensors. Moreover, another concept that plays an important role is the revisit time of the satellite, which indicates the time needed for the space platform to collect consecutive data of a specific location of the Earth.

When analyzing among the possible sources of images, we encounter some trade-offs. On the first hand, there exist commercial satellites that provide imagery with spatial resolution of less than a meter, but these data can become expensive when needed for a

repeated experiment or to cover a large area. On the other hand, there are open-access satellites, such as the Sentinel-2, that provide lower resolution images (about 10 m/pixel), and there is also an increasing launch of smaller satellites, such as CubeSats or PocketCubes, that have cheaper sensing instruments but lower revisit times. This is why, in order to fully exploit these lower resolution images, there is a growing interest of the RS community in the development of deep learning techniques that intend to increase the spatial resolution of low-resolution images with techniques known as super-resolution.

Essentially, deep learning (DL) holds great promise to fulfill the challenging needs of remote sensing image processing since it leverages the huge computing power of modern GPUs to perform human-like reasoning and extract compact features which embody the semantics of input images. The interest of the RS community towards deep learning methods is growing fast, and many models have been proposed in recent years, often with an outstanding performance [4–7].

Taking into account the free availability of medium-resolution satellite images, like those provided by the Sentinel-2, we propose to apply deep learning techniques, specifically Convolutional Neural Networks (CNNs), to the multispectral bands from remote sensing satellites in order to automatically obtain semantic segmentation maps of high-resolution images for Land Use and Land Cover applications. More concretely, the approach taken in this work consists on applying Semantic Segmentation and Super-Resolution techniques to obtain a semantically segmented output image with higher spatial resolution than the original input image. In addition to the segmentation map, a super-resolved version of the input image is also provided in a multi-task approach.

The paper is organized as follows: Section 2 presents previous works related to deep learning methods for semantic segmentation and super-resolution. The dataset, the baseline network architecture and the proposed extensions, loss functions, training settings and evaluation metrics are detailed in Section 3. A series of experiments and results are reported in Section 4. More experiments and discussion of results are presented in Section 5 and, finally, Section 6 summarizes the main conclusions.

2. Related Work

The introduction of Deep Learning techniques in the computer vision field has led to major advances in all its different sub-domains (object detection and semantic segmentation among others). For this reason, the RS community has been recently attracted to use it in tasks like semantic segmentation or super-resolution. Particularly, CNNs [8] have been widely applied with outstanding results on different RS imaging problems [9–12]. The work related to this paper is presented in the next three sub-sections.

2.1. Semantic Segmentation

Semantic segmentation aims to assign a finite set of semantic labels, such as land cover classes, to every pixel in an image [13–15]. The network predicts a probability distribution of all classes for each pixel, and assigns the most probable class to it. Architectures based on an encoder-decoder scheme are commonly used [16–18]. In those architectures, the encoder gradually reduces the spatial dimensions of the input image in order to encode rich semantic information, whilst the decoder tries to gradually recover the spatial information so as to recover high resolution feature maps with sharp object boundaries. A very popular architecture is U-Net [17], which is broadly used due to its symmetry, achieved by maintaining skip-connections in all the levels of the encoder-decoder structure.

On the other hand, networks based on Spatial Pyramid Pooling modules [19] are able to encode rich contextual information by pooling features at different resolutions. However, detailed information related to object boundaries is missing due to the pooling or convolutions with striding operations within the network backbone. DeepLabV3 [18] employs various parallel atrous convolutions at different rates in its Atrous Spatial Pyramid Pooling (ASPP) module to capture contextual information at different rates, but lacks of a powerful decoder to recover high resolution feature maps. Atrous or dilated convolutions allow

the expansion of the receptive field without loss of resolution and avoid the max-pooling operations, so feature maps at an arbitrary resolution can be obtained. DeepLabv3+ [20] extends DeepLabv3 by adding a simple, yet effective, decoder module in order to improve the object boundaries, such as in an encoder-decoder based structure, while maintaining the rich semantic information provided by a more powerful encoder based on a Spatial Pyramid Pooling module.

In the remote sensing field, the problem of semantic segmentation has been addressed from many perspectives, ranging from statistical approaches to methods based on machine learning [21,22]. Within this group, Random Forests (RF) and Support Vector Machines (SVM) are the most widely used, as they achieve good performances and are resistant to overfitting even with small training sets. However, deep learning models are becoming the state-of-the-art technology in LULC applications and have been shown to outperform classical approaches [23–25]. As opposed to methods that perform pixel-wise classification taking into account only single pixel features, like RF or SVM, deep learning models based on CNNs use contextual information of each pixel neighborhood, which leads to the improvement of performance and the reduction of noise in the resulting segmentation maps.

Many DL based models have been recently proposed for LULC classification. In early studies, labels have been predicted pixel by pixel using patch-based CNNs, relying on a small patch around the target pixel [26–28]. This approach has been applied in problems with limited annotated data, but it is time consuming and does not guarantee the spatial continuity and integrity of labels.

Fully convolutional approaches overcome the limitations of patch-based CNNs. They use an encoder-decoder structure, where typically the encoder is one of the popular CNN architectures (like VGGNet or ResNet) pretrained on the natural-image dataset ImageNet [29], and fine-tuned on RS data. For example, this approach has been applied in [30] on Landsat 5/7 multispectral images and in [31] on WorldView-2/-3 images. Other approaches follow an object-based strategy, combining CNNs with unsupervised image segmentation (e.g., superpixels) [32,33].

2.2. Single Image Super-Resolution

Single Image Super-Resolution (SISR) aims to recover a high-resolution (HR) image from a low-resolution (LR) image [34]. These techniques seek to learn implicit redundancy that is present in the the data to recover missing HR information from a single LR instance [35], which usually implies learning local spatial correlations.

As stated in [36], there are four kinds of supervised Deep Learning-based SISR methods. One is pre-upsampling SR, which applies a conventional upsampling operation, such as a bicubic interpolation, and then refines the HR image by using a deep convolutional neural network. This approach is very computationally expensive since most of the operations are done in the high dimensional space. The second one is post-upsampling SR, which integrates learnable upsampling layers at the end of the model instead of the traditional upsampling layers, reducing the computational cost. The third one is progressive-upsampling SR; it is based on post-upsampling, but aims at gradually reconstructing high-resolution images and allows multiscale SISR. Finally, iterative up-and-down SR is based on generating intermediate images, by iteratively employing upsampling and downsampling layers, and combining them to reconstruct the final SISR image.

An alternative to the pre-upsampling method is proposed in [35], with a CNN architecture where feature maps are extracted in the low-resolution space. Moreover, an efficient sub-pixel convolution layer (known as Pixel Shuffle) is introduced, which learns an array of upsampling filters instead of using a handcrafted interpolation.

On the other hand, architectures based on Generative Adversarial Networks (GANs) [37], like SRGAN [38] or ESRGAN [39], have been proposed as they produce high resolution images with photo-realistic details. Models based on GANs have also been applied for the super-resolution of remote sensing imagery [10,40–42].

In particular, some works tackle the problem of super-resolving Sentinel-2 bands using DL approaches. Specifically, Lanaras et al. [43] propose to super-resolve the LR bands to 10 m using a CNN with skip connections (named resblocks) between feature maps, while [44] includes more resblocks and adversarial training. Other approaches, like [45], combine resblocks with self-attention mechanism and a procedure for training these models in high-performance environments. Other solutions have also been proposed in [46–49], focusing in learning difference details between the LR and HR bands.

On the other hand, to improve the spatial resolution of the Sentinel-2 10 m channels [50] uses an ESRGAN as baseline to produce SR of RGB Sentinel-2 bands with scaling factors 2 and 4, previously downsampling the dataset to form the LR-HR pairs for training. Li and Li [51] produce Sentinel-2 RGB images at 2.5 m, using GANs with the ESRGAN-style, introducing kernel estimation and noise injection to construct the pair of LR-HR from LR images. A comparison between several Sentinel-2 SR models using Wald's protocol [52] to generate the LR-HR pairs has been recently presented in [53].

2.3. Super-Resolution for Improving Semantic Segmentation

SISR can help to improve the results of semantic segmentation approaches. This idea has been explored in various works such as [54–56]. In particular, Dai et al. [54] show that applying SISR to input images of other computer vision tasks, like semantic segmentation, edge detection and object detection, improve their performance in LR imagery.

In the remote sensing field, some works apply super-resolution as a pre-processing step, using a first network for super-resolution and a second one for semantic segmentation of the super-resolved image [56,57], where both networks are separately trained.

A unified framework is proposed in [58], with a super-resolution network based on convolutional layers with residual connections and an encoder-decoder architecture for semantic segmentation, trained end-to-end. The model is trained and evaluated for the binary segmentation (object and background) of small patches with airplanes, ships and oiltanks. Another end-to-end framework is proposed in [59], using a D-DBPN for super-resolution followed by a Segnet model for semantic segmentation, and training them with a multi-task loss using images from the 2014 IEEE GRSS Data Fusion Contest dataset and the ISPRS 2D Semantic Labeling Contest [60].

Besides, a super resolution domain adaptation network was proposed in [61] to address the domain shift problem in the task of semantic segmentation of images with different resolutions (source and target domains, with low and high-resolution images, respectively). The model is trained with adversarial learning on datasets of very high resolution true orthophotos from the ISPRS 2D Semantic Labeling Contest [60].

In a recent work, Wang et al. [36] propose a two-stream model. Their model consists of three parts, a super-resolution stream, a semantic segmentation stream and a feature affinity module that helps to enhance the high-resolution features of the super-resolution stream with fine grained structural information from the super-resolution branch. The model is trained and evaluated in CityScapes and CamVid, two datasets for urban visual scene understanding. Our model adopts this dual-network approach, introducing modifications on the DeepLabV3+ architecture. Specifically, we employ more skip-connections between the encoder and both decoders, adding extra upsampling modules with a pixel-shuffle mechanism. We train our model on a subset of the Sentinel-2 Global Land Cover dataset, outperforming the baseline DeepLabV3+ trained with the same LR images, producing smooth and accurate segmentation maps and an improved version of LR input images.

3. Materials and Methods

3.1. Dataset

The S2GLC (Sentinel-2 Global Land Cover) project [2] was led by the Space Research Centre of the Polish Academy (CBK-PAN) with the support of the European Space Agency (ESA). The main goal of the project was the development of a methodology for producing high resolution global land cover maps based on Sentinel-2 imagery.

Specifically, we used the S2GLC 2017 or Land Cover Map of Europe 2017, available at [62], which is a product resulting from the Phase 2 of the S2GLC project, that restricted the methodology employed on S2GLC just to the European continent. The map was obtained by means of classifying, with a high level of automation, more than 15,000 Sentinel-2 images collected during the year 2017. The methodology for the classification of multi-temporal Sentinel-2 imagery relied on the random forest algorithm and achieved a high thematic overall accuracy, over 86% at country level. The resulting dataset legend consists of 14 land cover classes (see Figure 1). The map pixel size equals 10 m, which corresponds to the highest spatial resolution of Sentinel-2 imagery.

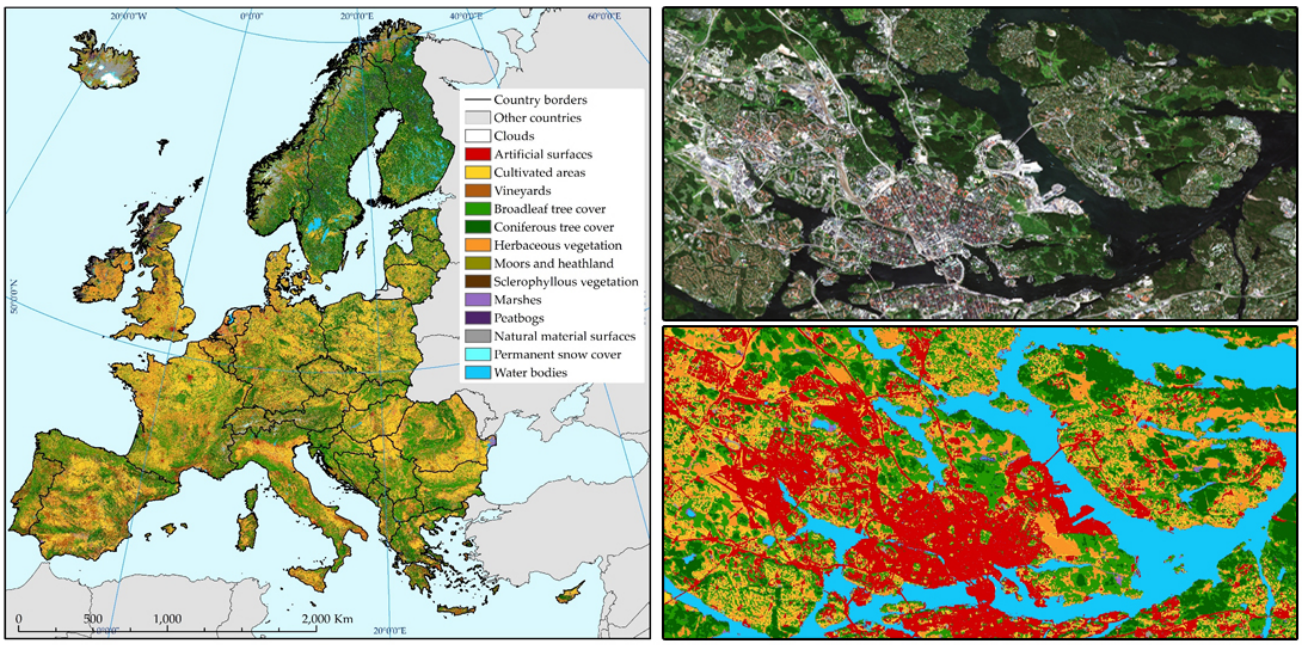


Figure 1. S2GLC 2017—Land Cover Map of Europe 2017. Source: [2].

We restricted our study area to Catalonia (Spain) and we used the S2GLC 2017 land cover map corresponding to that region as ground truth for the segmentation task. We searched for 2017 Sentinel-2 satellite images corresponding to this region (see Table 1), so as to match the date when the land cover dataset was created. We used the 10 m multispectral channels as ground truth for the super-resolution branch, composed by Bands 2, 3, 4 and 8 of Sentinel-2 images (Blue, Green, Red and Near Infrarred (NIR) channels, respectively). Then, we created our dataset (S2GLC-Cat) composed by geo-referenced pairs of the Sentinel-2 images and their corresponding land cover map from the S2GLC 2017 dataset. We cropped the S2GLC to match each Sentinel-2 image, reprojected the Sentinel-2 imagery using the coordinate system of S2GLC data and co-registered each pair. The process included locating and matching a number of ground control points in both images and then performing a geometric transformation. Automatic and manual control points were extracted to obtain a representative and well distributed set of points. Finally, a polynomial warping algorithm was applied to Sentinel-2 images.

Since both, land cover maps and Sentinel-2 multispectral images corresponding to the region of Catalunya, were too large, we formed our train and test sets by taking random patches of 512×512 from those images. It resulted in a total of 2700 images for the train set and 300 images for the test set. In order to implement the dual path approach, the input image was formed by downsampling the Sentinel-2 patches by a scale factor of 2, and we kept the full-resolution patches and labels as ground truth data.

Table 1. S2GLC-Cat Sentinel-2 images downloaded from [63].

Date	ID	Size
20170705T105031	T31TCH	9726 × 9851
20170824T105031	T31TCG	9353 × 10,134
20170705T105031	T31TBG	9679 × 9792
20170622T104021	T31TDF	2760 × 4977
20170612T104021	T31RDG	9199 × 9893
20170615T105031	T31TCF	4801 × 10,456

Moreover, we computed the histogram of both, the resulting patches and the full images, to check for any class imbalance. We concluded that the patches were representative of the full images. However, due to the non-stationarity behaviour of land cover classes, such as clouds and permanent snow surfaces, some images did not match their corresponding label. As discussed in Section 5.2, we relabeled more than 270 images in order to improve the segmentation results.

3.2. Network Architecture

Encoder-decoder networks (see Figure 2) have been successfully applied to many computer vision tasks, such as semantic segmentation, object detection and pose estimation. They are typically composed by an encoder module, that gradually reduces the spatial dimensions whilst extending the number of channels of the input image in order to encode rich semantic information; and a decoder module, which tries to gradually recover the spatial information so as to retrieve high resolution feature maps. In those architectures, it is referred as Output Stride (OS) the ratio of the input image spatial resolution to the encoder output resolution. For semantic segmentation tasks OS = 16 (or 8) is usually adopted by the feature extractor [20], meaning that the encoder output spatial resolution is 16 times smaller than the input image. From this point, the decoder gradually upsamples the feature maps generally making use of skip connections from the encoder at different levels.

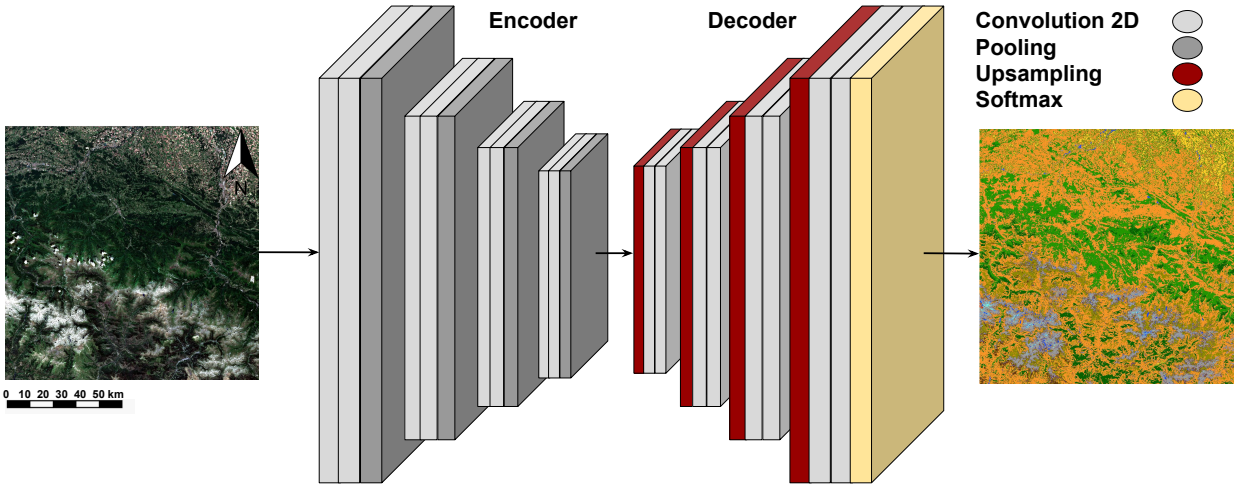


Figure 2. General overview of an encoder-decoder architecture.

The key point of our proposed architecture is a dual path network approach (DPN), which is inspired by [36]. This approach mainly consists in predicting one segmentation map and one super-resolved image, where both are twice the size of the input image. This is done simultaneously in a multitask fashion by employing two dedicated branches in the network architecture. It is worth mentioning that the model can be adjusted to work with any scaling factor by making minor changes in the decoder part.

The segmentation accuracy of the network can be related with the size of the input image (and its corresponding ground truth map): the higher the input spatial resolution, the better the performance [36]. This happens because larger input images contain finer spatial information labeled in the corresponding ground truth, so the edges of the different classes become more clear.

The motivation behind the dual-network approach (see Figure 3) is to use a low-resolution (LR) input image to predict a high resolution (HR) segmentation map, guiding the process with a HR version of the original image that is generated by a second branch. Thus, the learning paradigm consists of integrating the idea of super-resolution into an existing semantic segmentation pipeline to keep HR representations. The network, as proposed in [36], consists of a Semantic Segmentation Super-Resolution (SSSR) branch that predicts the HR segmentation map, and a Single Image Super-Resolution (SISR) branch that reconstructs a HR version of the input image, where both outputs sizes are twice the input size. Apart from those branches, there is also a Feature Affinity (FA) module that tries to enhance the HR features of the SSSR with the fine-grained structural information from the SISR by computing a loss between both outputs. More details about this FA module and the FA loss will be explained in Section 3.3.

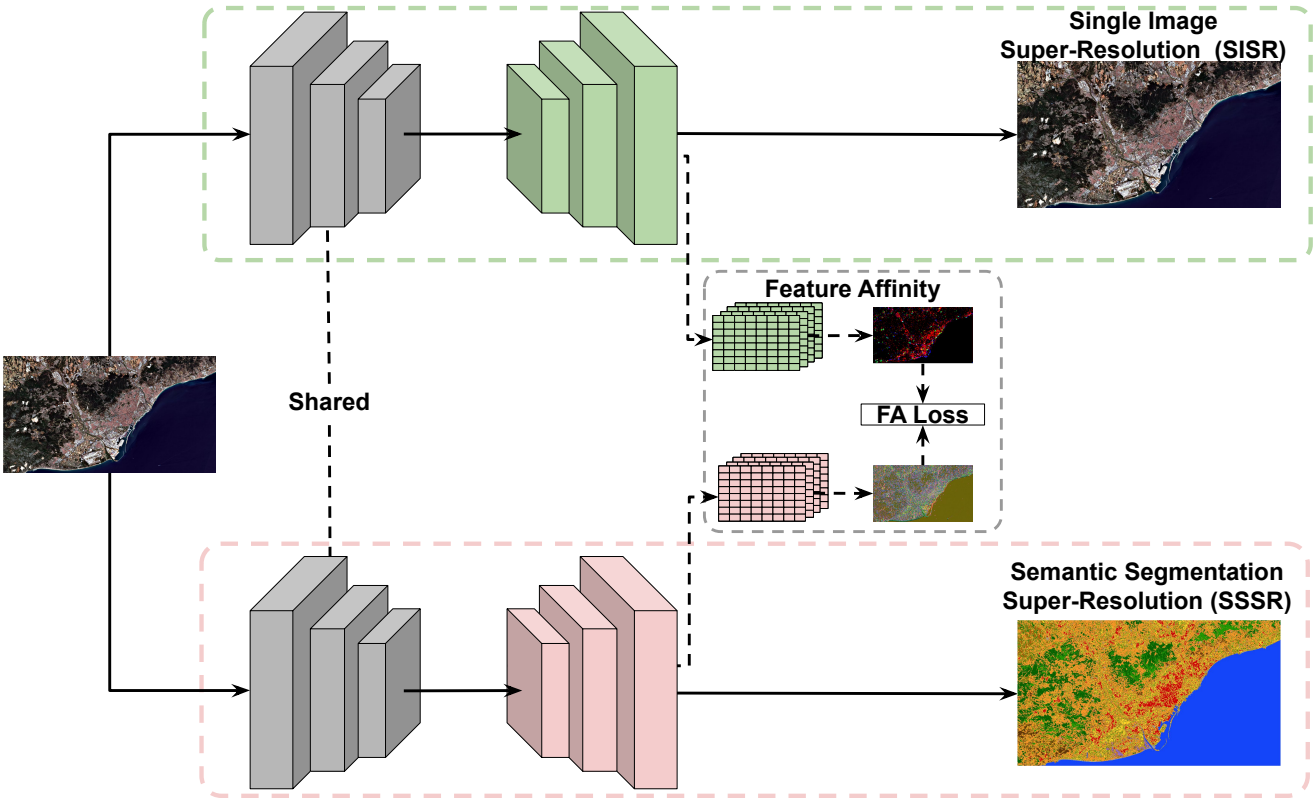


Figure 3. Overview of the Dual Super-Resolution (DSR) architecture composed by three parts: Semantic Segmentation Super-Resolution (SSSR) branch, Single Image Super-Resolution (SISR) branch, and Feature Affinity (FA) module.

The idea is that the two branches share the same encoder (feature extractor) but have their own decoder. The SSSR branch is optimized with a typical semantic segmentation loss, such as the Cross Entropy Loss, and the SISR branch is optimized with a pixel-wise loss, such as Mean Square Error. Furthermore, as commented, there is also a FA loss that tries to guide the learning of both branches. All these losses will be explained individually in Section 3.3.

In our case, we treat segmentation as the main task but we maintain the SISR output at inference time since we are also interested in predicting a HR version of the input image. Nevertheless, notice that at inference time this branch can be removed, notably reducing the computation cost, if only the segmentation map is of interest.

As stated before, the dual path network approach consists of integrating the idea of super-resolution into existing semantic segmentation architectures. We implemented this idea by appending an extra upsampling module at the end of the decoder of the DeepLabv3+ [20] network. Apart from that, we redesigned the original decoder module mainly to improve the super-resolution results, to cope with the peculiar spatial granularities of satellite imagery. We opted for considering the same design of the decoder and the extra upsampling module for both SSSR and SISR branches in order to maintain some kind of symmetry.

The DeepLabV3+ architecture (see Figure 4) extends DeepLabV3 [18] by adding a simple but effective decoder module to refine the segmentation results especially along object boundaries. The architecture is based on a powerful backbone encoder (we use ResNet101 [64]), an atrous spatial pyramid pooling module that allows encoding multi-scale contextual information, and a decoder that receives a skip-connection from the encoder low-level features to facilitate the upsampling path. DeepLabV3 is characterized by employing atrous (dilated) convolutions in the last group of layers in order to maintain the resolution of the feature maps at an arbitrary resolution. Using ResNet101 as backbone encoder, the spatial resolution of the output feature maps is 32 times smaller than the input image resolution (OS = 32).

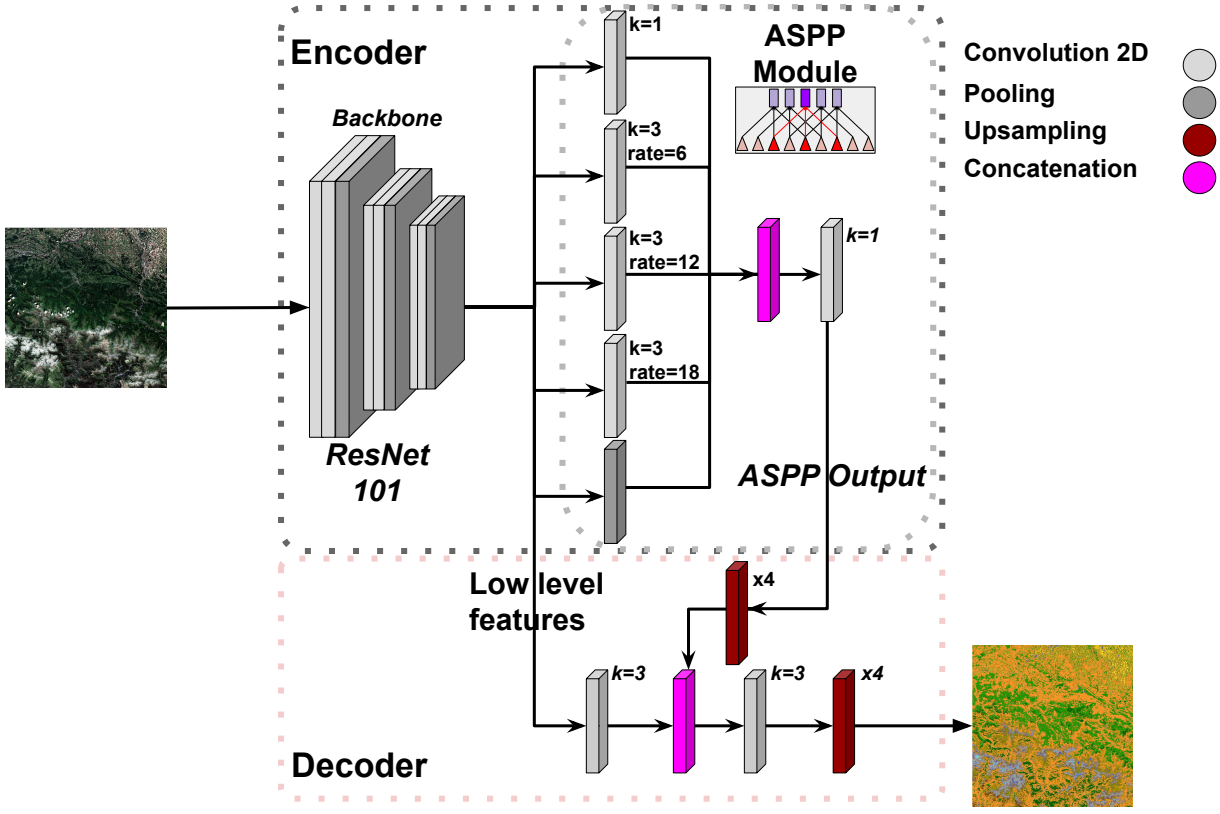


Figure 4. DeepLabV3+ architecture with Atrous Spatial Pyramid Pooling (ASPP) module.

We conducted several experiments with different versions of the decoder and the extra upsampling modules. Here we explain the models, and the results of the experiments where all of them will be presented in Section 4. We started by just adding the extra upsampling modules, consisting in a stack of 3×3 2D Convolutions, followed by a nearest neighbor upsample and another 3×3 2D Convolution, on top of each decoder.

We then proceeded to add more skip-connections to the decoders since the SISR results were not satisfactory. In a first step, we added another concatenation with a lower-level feature map (referred as model $v1$ in Section 4). Then, we added the concatenation with the bicubic interpolation of the input image before the final upsampling just in the SISR branch (model $v2$). And finally, we explored adding also the concatenation with the bicubic interpolated image in the SSSR branch (model $v3$).

We further studied the configuration of the extra upsampling module for the SSSR branch. In Figure 5 we present the final design for the model which yields the best results (model $v4$). Notice that the decoder path receives skip-connections consisting in the ResNet101 feature maps $F1$, $F2$ and the bicubic interpolation of $F0$ at the extra upsampling module (both for the SSSR and SISR branches). Here $F0$ refers to the input image, and $F1$, $F2$ refer to the ResNet101 feature maps whose spatial dimensions are respectively two and four times smaller than the input image (see Figure 5 for clarification). Moreover, for the SSSR branch, the Segmentation Head module does not convert to the desired number of classes and preserves the channel dimensions; so the extra upsampling is done directly in the feature maps and a 3×3 2D Convolution is then used to convert to the number of classes. For the SISR branch, the channel dimensionality is reduced progressively. Figure 6 shows implementation details of both decoders. Note that long skip connections from the encoder provide low-level features that help in the reconstruction of high-resolution details.

During the experiments we also explored changing the type of upsampling done in each upsample module from the DeepLabv3+ architecture. We tried setting all the upsampling modules in the architecture to (1) nearest neighbor, (2) transpose convolution, and (3) Pixel Shuffle sub-network, and we obtained the best results for case (3). For this reason, in the illustration of the architecture presented in Figure 5, the upsampling modules use pixel shuffle but are just referred as "Upsampling $\times 2$".

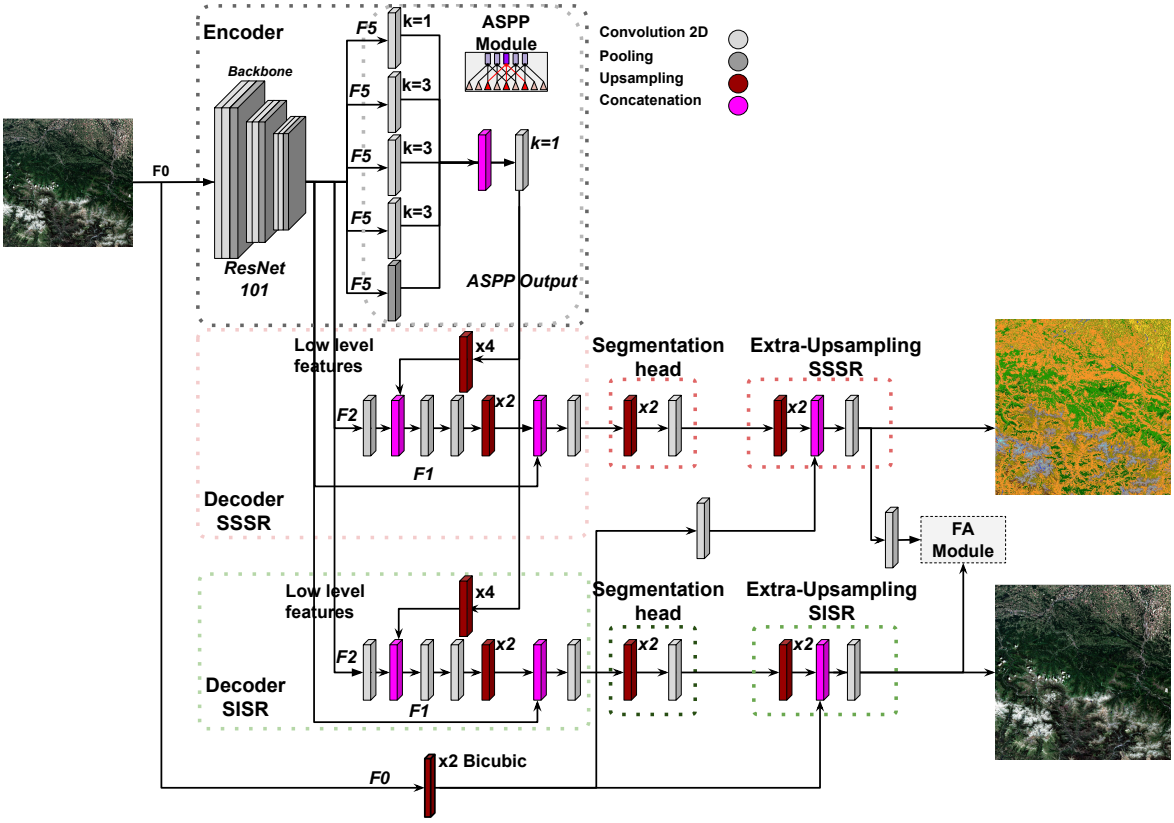


Figure 5. Dual Path Network based on DeepLabv3+. Model $v4$.

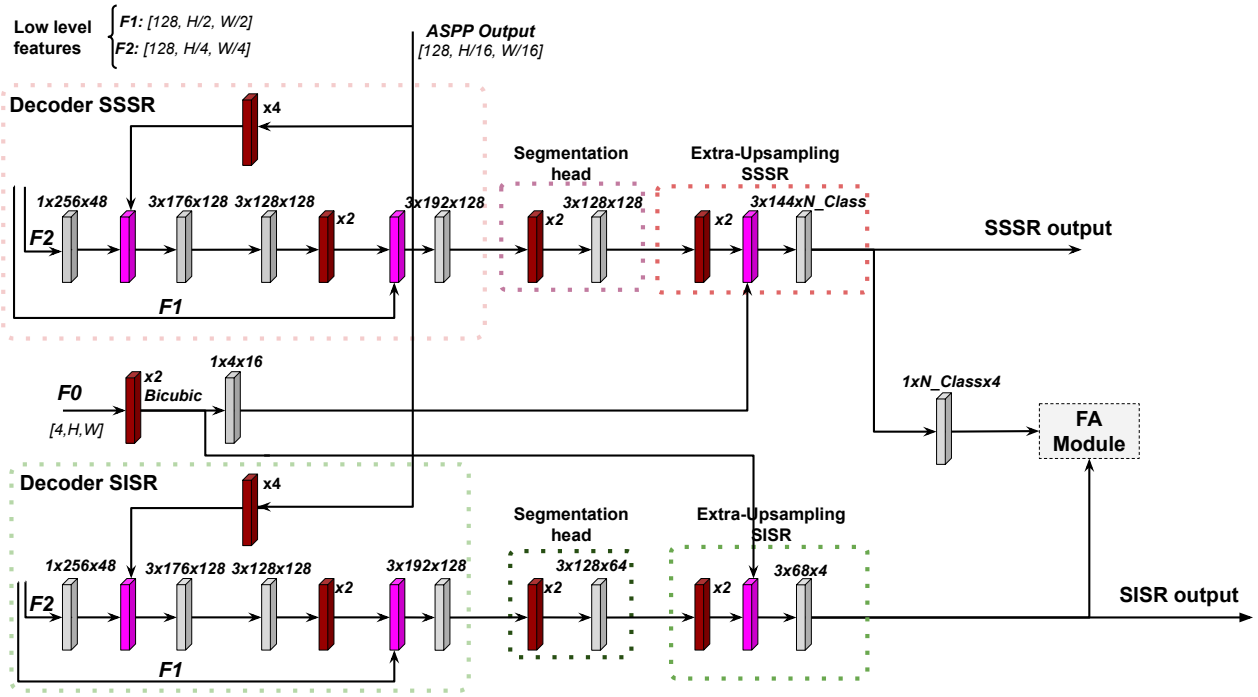


Figure 6. Implementation details of both decoders. Each convolutional layer is characterized by [kernel-size × input-channels × output-channels].

3.3. Loss Functions

In this subsection we present the different losses used for training the neural network. Since the approach consists in a multi-task model, specific losses for each task are considered. We employed Cross Entropy (CE) loss and CE with weights for semantic segmentation; two pixel-wise losses, Mean Square Error (MSE) and Mean Absolute Error (MAE), for super-resolution, and the Feature Affinity (FA) loss to guide the SSSR learning from the SISR branch.

3.3.1. Semantic Segmentation Loss (SSL)

The Cross Entropy Loss is a common loss used in multi-class segmentation problems. The expression for each training example is the following:

$$L_{CE}(\hat{y}, y) = -\sum_{k=1}^{K} y^{(k)} log(\hat{y}^{(k)}) \quad (1)$$

where $\hat{y}$ is the vector containing the predicted probabilities for each class, y is the target one-hot vector containing a "1" only at the correct class position, and K is the number of classes.

A common approach is to use the CE with weights. This variant of the CE loss is very useful when the dataset is unbalanced, i.e., there are classes that appear much less than others. CE with weights employs a rescaling weight given to each class, weighting more less frequent classes to improve the results for those classes.

3.3.2. Super-Resolution Loss (SRL)

MSE or MAE are commonly used as reconstruction loss for Single Image Super-Resolution since they compare the reconstructed image with the target one in a pixel-wise manner.

The MSE is used to minimize the error defined as the sum of all the squared differences between the true and the predicted values, where N is the number of pixel in the images:

$$L_{MSE}(\hat{Y}, Y) = \frac{1}{N}\sum_{i=1}^{N}\|\hat{Y}_i - Y_i\|^2 \quad (2)$$

Alternatively, the MAE loss aims to minimize the error which is the sum of the absolute differences between the true and the predicted values:

$$L_{MAE}(\hat{Y}, Y) = \frac{1}{N}\sum_{i=1}^{N}|\hat{Y}_i - Y_i| \quad (3)$$

3.3.3. Feature Affinity Loss (FAL)

The Feature Affinity module aims to guide the learning of the SSSR branch from the SISR branch, since the segmentation pipeline can benefit from the fact that SISR can reconstruct high-resolution fine-grained information from a low-resolution input. The idea is that the feature maps from the SSSR are enhanced by the SISR, which contains more detailed fine-grained structural information, thus obtaining a finer segmentation. Even though the structures from SISR do not directly imply semantic categories, they can be grouped by the relationship between pixel and pixel, or region and region. As proposed in [36], we modeled these details by the correlation between internal pixels.

The FA Loss aims to learn the distance between the similarity matrix of the HR features of SSSR and the HR features of SISR, where the similarity matrix describes the pairwise relationship between every pair of pixels of a given feature map (for a feature map F with dimensions $C \times W \times H$, the similarity matrix would contain $(W \times H)^2$ entries consisting in the relationship between every two pixels in the spatial dimension).

$$L_{FA}(S^{(SSSR)}, S^{(SISR)}) = \frac{1}{W^2 \cdot H^2}\sum_{i=1}^{W\cdot H}\sum_{j=1}^{W\cdot H}\|S_{ij}^{(SSSR)} - S_{ij}^{(SISR)}\|_q \quad (4)$$

where $S^{(SSSR)}$ and $S^{(SISR)}$ refer to the SSSR and SISR similarity matrix respectively, and q is the norm, set to 1 for stability. So, the loss computes the pixel-wise distance (absolute value) for all the entries in the matrices, sums them up and normalizes by the total number of entries.

Given a feature map F, the entry (i,j) of the similarity matrix of that feature map is computed by projecting the vector (dot product) taken in the channel dimension from the spatial dimension pixel number i, i.e., F_i to the vector taken from the pixel j, i.e., F_j, where the pixels are numbered in a row-wise manner from 1 to $W \times H$. This models the correlation between internal pixels. See Figure 7 for a visualization example.

$$S_{i,j} = \left(\frac{F_i}{\|F_i\|_p}\right)^T \cdot \left(\frac{F_j}{\|F_j\|_p}\right) \quad (5)$$

in this case p is the norm, set to 2 for stability.

Note that for the implementation of the computation of the whole similarity matrix, feature maps can be flattened on their spatial dimensions and the pairwise relationship between every row vector (first normalized by the 2-norm) can be computed just by multiplying the resulting matrix by its transpose (see Figure 7).

Although, as considered in [36], it is better to compute the correlation of every pair of pixels, in our implementation we subsampled the feature maps to 1/8 before computing the similarity matrix to avoid high memory overheads. Moreover, since the high-resolution feature maps of both SISR and SSSR branches have different channel distributions, the FA module (See Figure 5) also incorporates a 1×1 2D Convolution that ensures that the number of channels of SSSR matches the ones of SISR in order to reduce instabilities during the training.

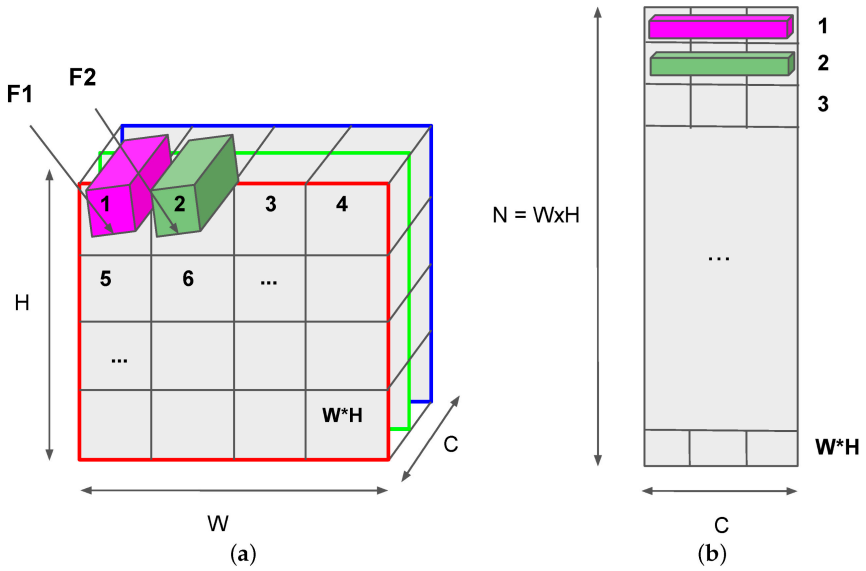


Figure 7. (**a**): Example showing how to compute the relationship, in the spatial dimensions, between pixels 1 and 2; it is the dot product of the two depicted volumes, each volume divided by the 2-norm. (**b**): Flattened version of the feature map to interpret a possible implementation of the similarity matrix computation.

3.3.4. Multi-Task Loss

Since our approach consists in a multi-task network, the whole objective function, shown in Equation (6), is composed by a linear combination of a loss for semantic segmentation (CE or weighted CE), a loss for super-resolution (MSE or MAE) and the feature affinity loss.

$$L = L_{SSL} + w_1 L_{SRL} + w_2 L_{FAL} \tag{6}$$

where w_1 and w_2 are hyper-parameters set to make the loss ranges comparable. In our case we obtained the best results weighting both w_1 and w_2 to 1.0.

3.4. Training Details

The training consists in the minimization of the multi-task loss function (Equation (6)) in an end-to-end manner.

3.4.1. Data Standardization

A common approach for speeding up the convergence of a neural network is to normalize or standardize the data prior to the training. Then, at the output of the model, the original dynamic range of the image is recovered by doing the inverse process. We obtained better results by standardizing the input image. This is done per-channel, by subtracting the mean and dividing by the standard deviation of the input image:

$$x' = \frac{(x/MAX_I) - \mu}{\sigma} \tag{7}$$

where μ and σ are the per channel computed mean and standard deviation, respectively, and MAX_I is the maximum possible pixel value of the image.

3.4.2. Weights for Unbalanced Classes

Weights can be employed in the Cross Entropy loss to try to mitigate the negative impact of class imbalance in the dataset. Those weights are inversely proportional to the frequency of occurrence of each class in the dataset, so classes that have less appearance are weighted more. We used the following expression to compute the weights, as suggested in [65]:

$$w_n = \frac{1}{ln(1.02 + \beta_n)} \tag{8}$$

where β_n corresponds to the frequency of occurrence of the class, and the term 1.02 is added for stability.

3.4.3. Optimizer

We tried different optimizers and the best results were obtained using Adam. The learning rate was initialized to 2×10^{-4}. We also explored the use of different learning rate schedulers (step decay, cosine annealing) but we did not obtain a clear improvement by using them.

3.5. Quality Assessment

In this section we will explain the metrics used to quantitatively assess the performance of the results obtained from the test set. We will differentiate between semantic segmentation metrics, used to evaluate the SSSR performance, and super-resolution metrics for the reconstruction of the SISR image.

In addition, we will present qualitative results, since the metrics are quite limited by the noisiness of the semantic segmentation ground truth.

3.5.1. Semantic Segmentation Metrics

- Intersection-Over-Union (IoU) or Jaccard Index: it is a very popular metric used in semantic segmentation. The IoU is computed as the ratio of the area of overlap between the predicted segmentation and the ground truth (intersection), and the area of union between the predicted segmentation and the ground truth. The metric ranges from 0 to 1, with 0 indicating no overlap and 1 indicating ideally overlapping segmentation. For a multi-class segmentation, the mean IoU (mIoU) is computed by averaging the per class IoU.
- Confusion matrix: it is a matrix indicating on its rows the instances of the true classes whilst in its columns indicates the instances of the predicted classes. From the confusion matrix, the per class IoU can be obtained as:

$$IoU = \frac{TP}{GT + Pred - TP} \tag{9}$$

where *TP* stands for True Positive pixels, that can be computed taking the diagonal of the confusion matrix, *GT* stands for Ground Truth pixels and are obtained by taking the sum over columns (total number of true pixels for each class), and *Pred* stands for Predictions and are obtained by taking the sum over rows (total number of predicted pixels for each class).

3.5.2. Super-Resolution Metrics

- Peak Signal-to-Noise Ratio (PSNR): it is a widely used metric to quantify the quality of reconstructed images. It is defined as follows:

$$PSNR = 10log_{10}(\frac{MAX_I^2}{MSE}) \tag{10}$$

where MAX_I is the maximum possible pixel value of the image.
- Structural Similarity Index Measure (SSIM) [66]: it is a metric used for measuring the similarity between two images. SSIM is a perception-based model that considers image degradation as perceived change in structural information. The SSIM extracts three key features from an image: luminance, contrast and structure from both the reference image (x) and the reconstructed one (y).The resulting metric ranges from -1 to 1, or is re-adjusted to be in the range [0,1]. The larger the value, the better results.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (11)$$

where μ_x is the mean of x, μ_y the mean of y, σ_x^2 the variance of x, σ_y^2 the variance of y, σ_{xy} the covariance of x and y, $c_1 = (k_1 L)^2$, $c_2 = (k_2 L)^2$ are two variables used to stabilize the division with weak denominator, L is the dynamic range of the pixel-values, $k_1 = 0.01$ and $k_2 = 0.03$ by default.

4. Experiments and Results

We conducted several experiments with modified versions of the dual network architecture. We present the results obtained for the following four architectures:

- $v1$: The basic model with the addition of an extra long skip-connection in the decoder path of both SISR and SSSR branches, consisting in the concatenation with the low level feature map of the ResNet backbone $F1$, which is two times smaller than the input image.
- $v2$: The previous $v1$ model with the addition of a skip connection consisting in the concatenation with the bicubic interpolation of the input image right before the last convolution of the extra upsampling module. The interpolated image is added only to the SISR branch.
- $v3$: The previous $v2$ but also concatenating the interpolated image to the SSSR branch, since the segmentation branch could also benefit from the structural information of the bicubic interpolated image.
- $v4$: A modification of $v3$, where the spectral information of the bicubic interpolated image is diffused by passing it through a 1×1 2D Convolution with 16 filters prior to the concatenation.

The main differences between the four models are summarized in Table 2. For each architecture, we performed several experiments varying the loss function used for semantic segmentation (CE with or without weights) and for super-resolution (MSE or MAE), the contribution of the losses in the multi-task loss, the upsampling method and the initial learning rate, and computed the super-resolution and semantic segmentation metrics on the test set. The most relevant results are presented in Table 3.

Table 2. Different versions of the dual-network architecture.

Architecture	$v1$	$v2$	$v3$	$v4$
- F1 from backbone.	x	x	x	x
- Bicubic interpolation of input image on SISR branch.		x	x	x
- Bicubic interpolation of input image on SSSR branch.			x	
- Bicubic interpolation spectrally diffused with a 1×1 Conv2d				x

From Table 3 we can conclude that the best results are achieved by model $v4$, i.e., by using the bicubic interpolated image in both branches, as it obtained better segmentation results (given by mIoU) and achieved an equal value of SSIM as well. Regarding the SR Loss, the best super-resolution metrics were obtained when using MSE and by weighting its contribution in the total loss by 1.0. Choosing either Nearest Neighbor or Pixel Shuffle in the upsampling modules lead to the best segmentation results. Even though Nearest Neighbor achieved slightly higher segmentation metrics, we opted for Pixel Shuffle since the qualitative super-resolution results were much better.

Moreover, to asses the super-resolution model branch, results obtained with nearest neighbor, bilinear and bicubic interpolation techniques ($\times 2$) on the test set are provided in Table 4.

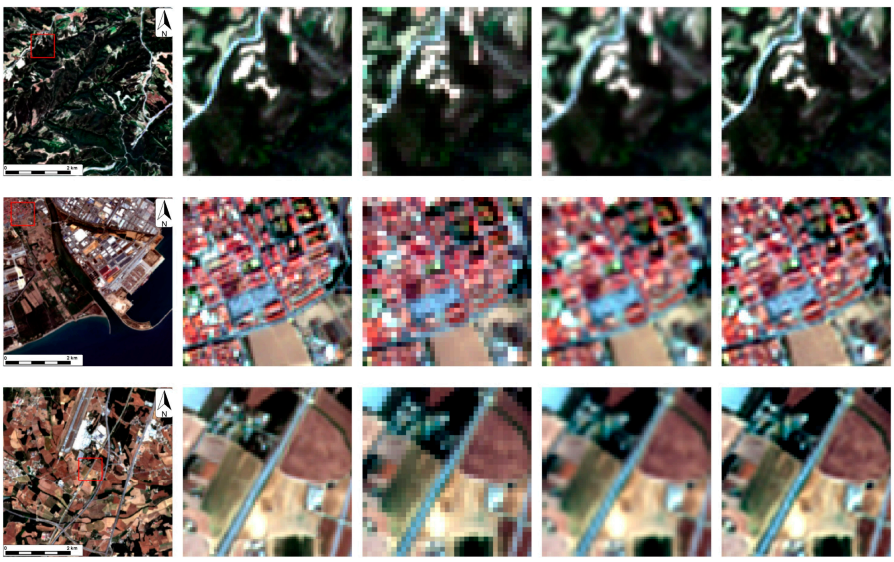
Table 3. Results obtained on the test set using the four variants of the dual network model ($v1$ to $v4$). Only the best performing configurations are shown. Upsampling techniques are NN (Nearest Neighbors) and PS (Pixel-Shuffle), lr is the learning rate, the semantic segmentation loss is weighted CE in all cases. Best values in bold.

Model	SR Loss/W	Ups.	lr	PSNR	SSIM	mIoU
$v1$	MSE, 1.0	NN	2×10^{-4}	35.318	0.770	0.450
$v2$	MSE, 0.1	NN	2×10^{-4}	35.418	**0.776**	0.480
$v2$	MSE, 1.0	NN	2×10^{-4}	35.424	**0.776**	0.475
$v2$	MAE, 0.1	NN	2×10^{-4}	35.413	0.775	0.473
$v2$	MAE, 1.0	NN	2×10^{-4}	35.423	0.775	0.475
$v3$	MSE, 1.0	NN	2×10^{-4}	**35.428**	**0.776**	0.465
$v4$	MSE, 1.0	NN	1×10^{-4}	35.419	**0.776**	**0.484**
$v4$	MSE, 1.0	PS	1×10^{-4}	35.422	**0.776**	0.482

Table 4. Results on the test set of baseline interpolation techniques and our best performing model ($v4$ with Pixel-Shuffle, SISR branch). Best values in bold.

	Interpolation Technique			
	Nearest Neighbor	**Bilinear**	**Bicubic**	**Our SISR Model**
PSNR	34.1415	34.1408	34.1574	**35.422**
SSIM	0.6729	0.6726	0.6732	**0.776**

Figures 8 and 9 show some qualitative results obtained with our best model configuration ($v4$ using Pixel-Shuffle). Images were downsampled to form the LR-HR pair. Therefore, the input LR images are at 20 m and both GT images (for SISR and SSSR branches) are at 10 m. Examples of super-resolution results using Nearest Neighbor, bicubic interpolation and our model are presented in Figure 8. Semantic segmentation examples are shown in Figure 9. It can be observed that segmentation maps are smooth and remove some of the noise that is present in the ground truth annotations (see Section 5.3).



(**a**) (**b**) (**c**) (**d**) (**e**)

Figure 8. SISR results obtained with model $v4$ using Pixel Shuffle: (**a**) full ground truth image, (**b**) selected crop downsampled to 20 m to form the LR pair, (**c**) Nearest Neighbor interpolation of LR crop, (**d**) bicubic interpolation of LR crop, (**e**) SISR result.

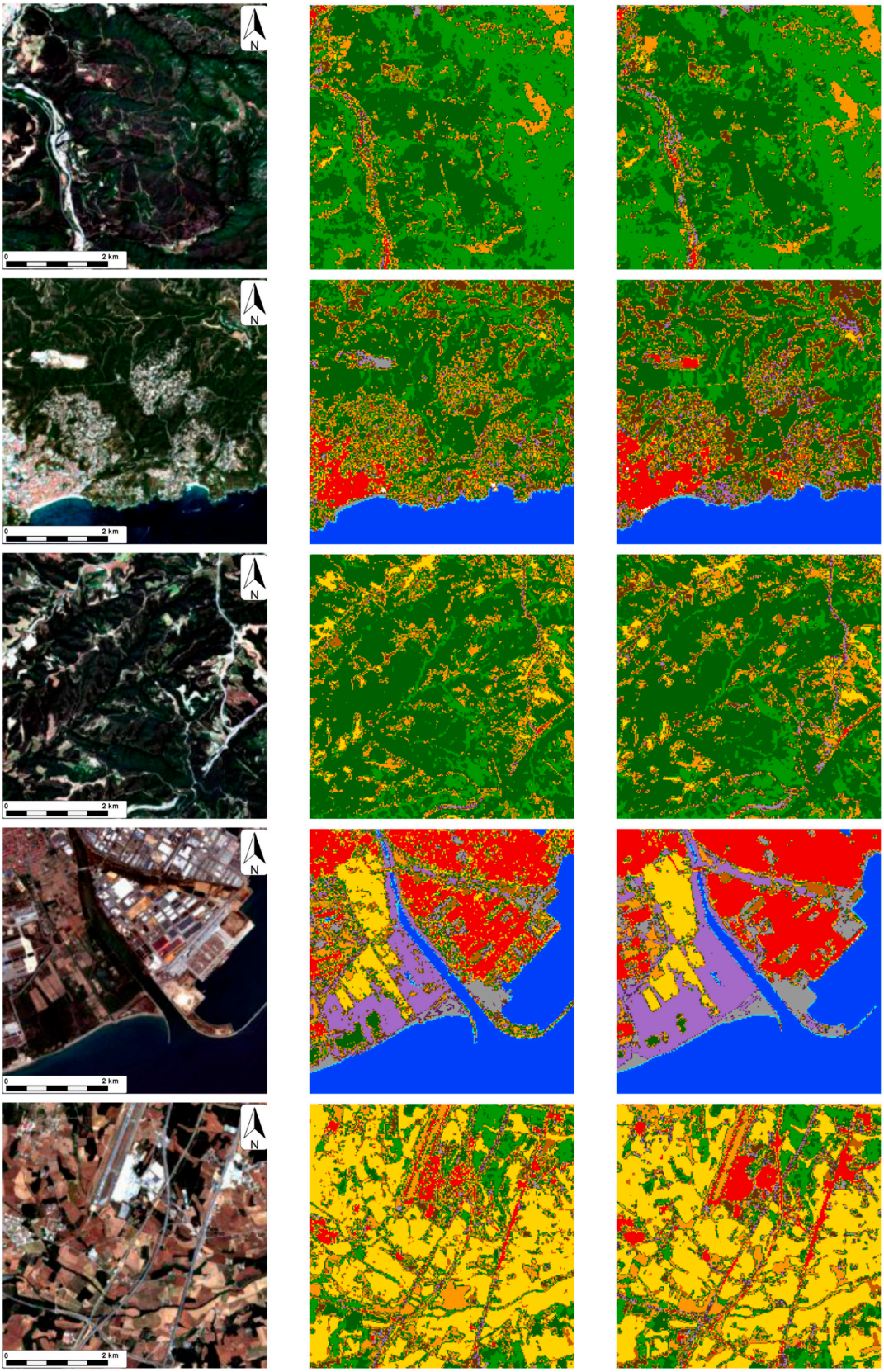


Figure 9. SSSR results obtained with model *v*4 using Pixel Shuffle: (**a**) Sentinel-2 image (SISR ground truth), (**b**) ground truth, (**c**) SSSR semantic segmentation map. The colormap is the same as depicted in Figure 1.

5. Discussion

5.1. Dual Network Architecture

We implemented a dual network approach for semantic segmentation and super-resolution based on an encoder-decoder structure, so as both tasks in the multi-task network share the same encoder as a feature extractor, but implement their own branch in the decoder path. We used as baseline the DeepLabv3+ architecture and modified it due to the particular fine grained structure of satellite images.

We showed the benefits of using a skip-connection consisting in the bicubic interpolation of the input image, and explored its concatenation either to just the SISR branch or to both SSSR and SISR branches. The best results were obtained when concatenating to both branches, but diffusing the spectral information of the interpolated image before concatenating into the SSSR branch. Regarding the type of upsampling modules employed in the whole architecture, we conclude that the best results are obtained by using the Pixel Shuffle sub-network.

5.2. Class Re-Labeling

Due to the non-stationary behaviour of land cover classes, such as clouds and permanent snow surfaces, some Sentinel-2 input images did not match their corresponding ground truth labels from the S2GLC 2017 dataset. Therefore, we decided to relabel those images in order to obtain more accurate segmentation results. We inspected the whole dataset and relabeled 270 images. After that, we trained the *v*4 *Pixel Shuffle* architecture on the relabeled dataset. Table 5 shows the confusion matrix as well as the IoU per class and mean IoU obtained with the relabeled dataset. We observe that the segmentation results on clouds and permanent snow covered surfaces increased significantly (+62.55% and +15.23%, respectively). Moreover, the global mIoU increased +5.28%, from 0.482 to 0.535.

Table 5. Confusion matrix after relabeling the dataset, normalized by rows. The IoU column shows the segmentation metric with the relabeled dataset, and the IoU* column presents the results for the original dataset. Best values are in bold.

	Class	1	2	3	4	5	6	7	8	9	10	11	12	13	14	IoU	IoU*
1	Clouds	**0.96**	0	0.01	0	0	0	0	0	0.02	0	0	0	0	0.01	**0.755**	0.129
2	Art. Surf.	0	**0.89**	0.02	0.02	0	0	0	0	0.02	0.01	0	0.04	0	0	**0.730**	**0.730**
3	Cul. Areas	0.01	0.02	**0.76**	0.06	0	0	0.07	0.01	0.02	0.01	0.02	0.02	0	0	**0.645**	**0.645**
4	Vineyards	0	0.02	0.13	**0.66**	0	0	0.09	0.01	0.05	0	0.02	0.03	0	0	**0.474**	0.465
5	Broadleaf TC	0	0	0	0	**0.78**	0.1	0.04	0.03	0.03	0.01	0	0	0	0	**0.688**	0.685
6	Coniferous TC	0	0	0	0	0.1	**0.80**	0	0.02	0.06	0.02	0	0	0	0	**0.697**	0.691
7	Herb. Veg.	0	0	0.08	0.06	0.04	0	**0.68**	0.05	0.06	0.02	0.01	0	0	0	**0.525**	0.514
8	Moors & Heathland	0	0	0.01	0.02	0.06	0.03	0.06	**0.69**	0.09	0.01	0.01	0.05	0	0	**0.470**	0.464
9	Scl. Veg.	0	0.02	0.03	0.03	0.02	0.04	0.07	0.07	**0.65**	0.04	0.01	0.03	0	0	**0.480**	0.475
10	Marshes	0	0.04	0.03	0.01	0.04	0.05	0.05	0.03	0.11	**0.63**	0.01	0.01	0	0.01	**0.308**	0.302
11	Peatbogs	0	0	0.17	0.13	0.03	0.01	0.12	0.06	0.05	0.03	**0.37**	0.03	0	0.01	0.136	**0.140**
12	Nat. Mat. Surf.	0	0.13	0.06	0.04	0	0	0.01	0.04	0.06	0.01	0.01	**0.63**	0.01	0.01	0.413	**0.430**
13	Perm. Snow	0	0	0	0	0	0	0	0.08	0.02	0	0	0.43	**0.45**	0.02	**0.365**	0.213
14	Water Bodies	0.06	0.04	0.02	0	0	0	0	0	0	0	0	0.02	0	**0.85**	0.796	**0.860**
	mIoU															0.535	0.482

Figure 10 presents some examples of relabeled images and the comparison between the predictions made by the model trained on the original dataset and trained on the relabeled one. The confusion matrix, per class IoU and mean IoU obtained with the original dataset is presented in the Apendix A (Table A1).

5.3. Noisy Annotations

The reported semantic segmentation metrics are based on comparing the model predictions with the ground truth labels provided by the S2GLC land cover maps. However, this ground truth has been generated automatically using a Random Forest classifier. A global accuracy of 86% has been reported on this dataset [2], which means that there

is an intrinsic and unavoidable inaccuracy in the ground truth that we use to train our models, which has an effect in the results obtained with them.

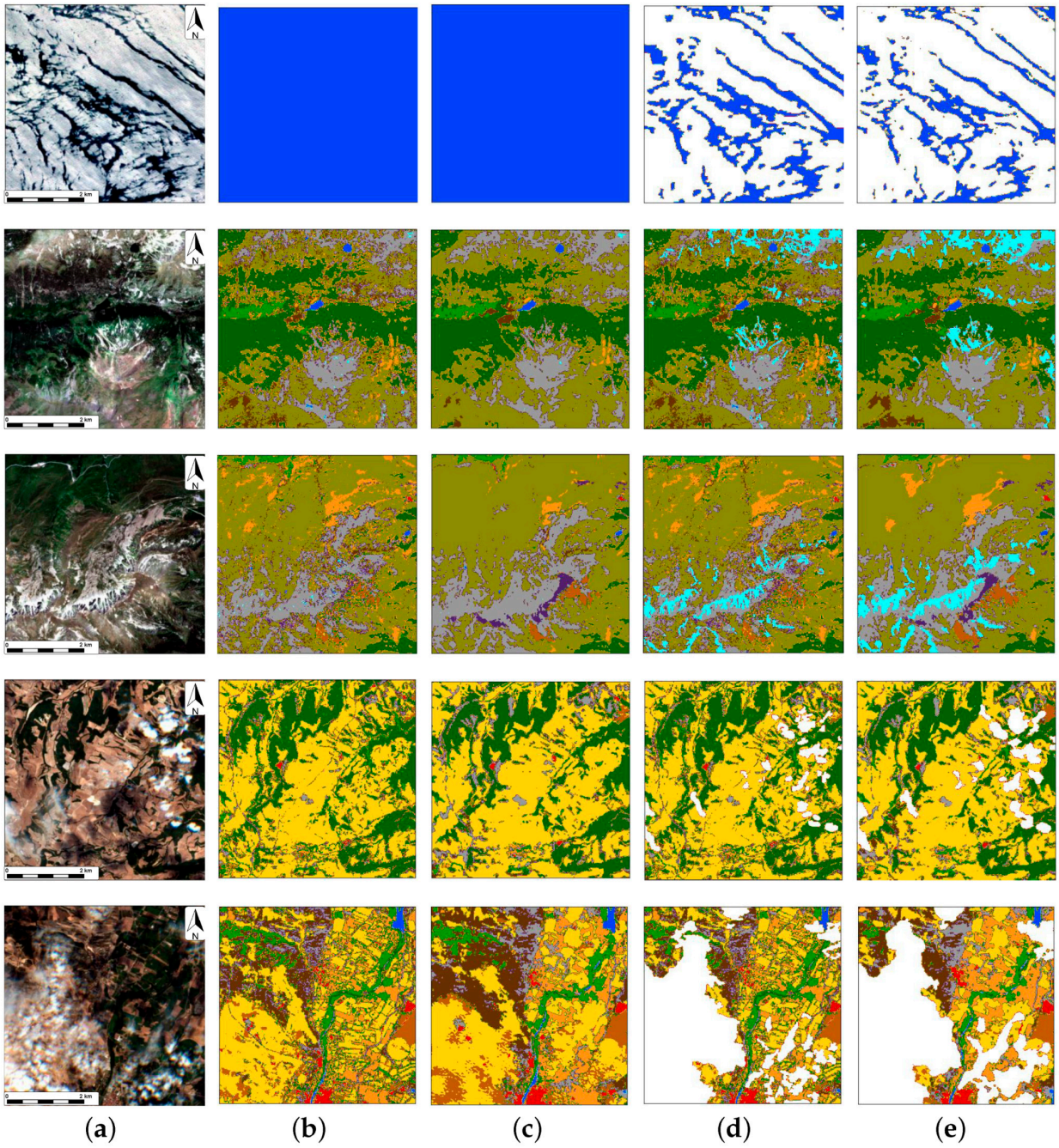


Figure 10. Samples of S2GLC 2017-Cat dataset after relabeling: (**a**) Sentinel-2 image, (**b**) original GT map, (**c**) SSSR results obtained with the original GT maps, (**d**) GT map after relabeling, (**e**) SSSR results obtained after training the model with the relabeled dataset.

In addition, the automatic procedure used to generate the ground truth land cover map is a pixel-based approach. The decision on a pixel does not take into account the pixel context, as opposed to predictions obtained by semantic segmentation models based on CNNs like our model. Therefore, ground truth annotations are noisy. Our model mitigates this noise and provides smoother segmentation maps, as can be appreciated in Figure 11. The IoU metric used for evaluating the segmentation result is not completely indicative of the performance of our model due to this high level of noise in the ground truth.

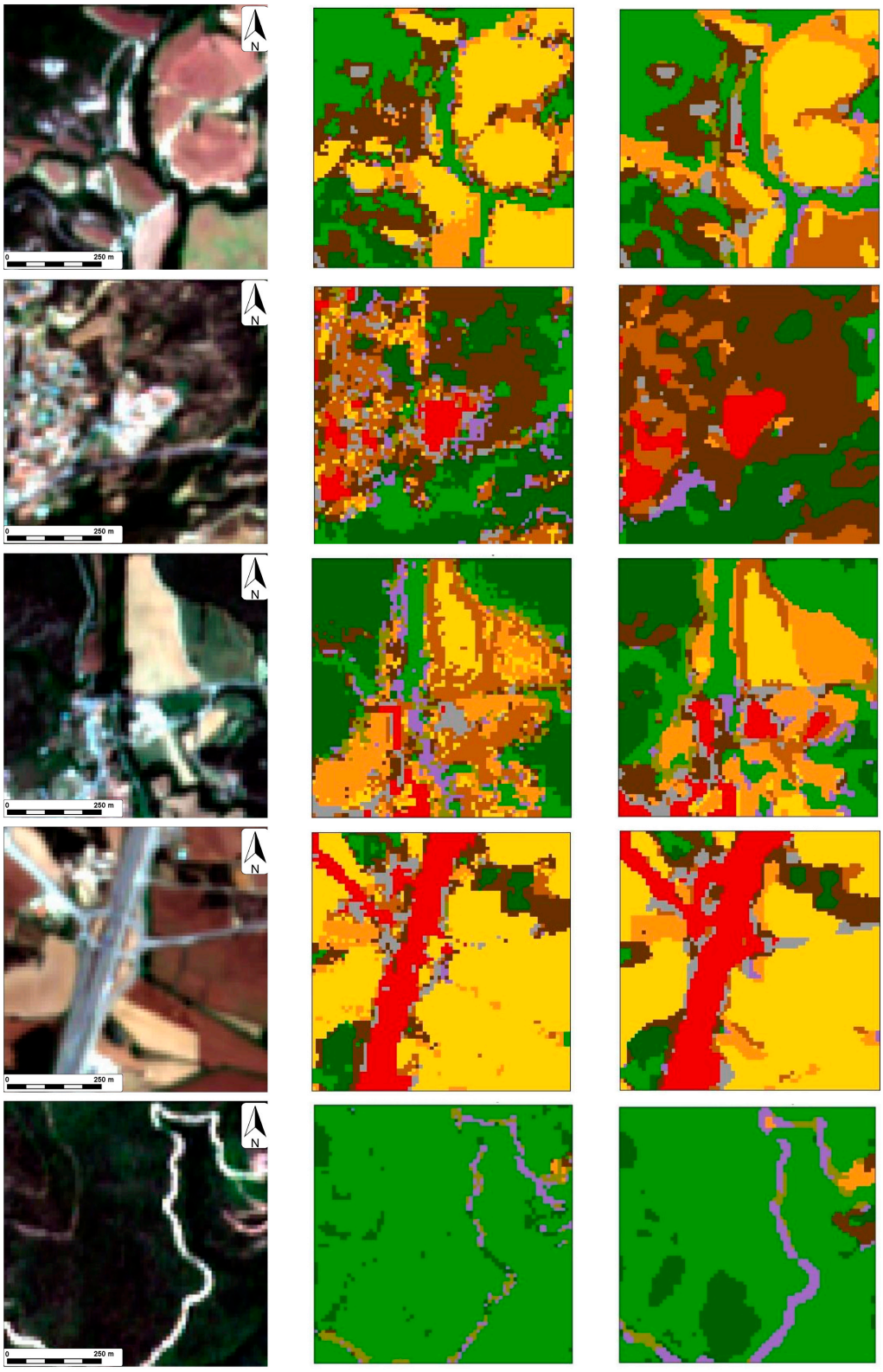


Figure 11. Illustration of the noise present in ground truth annotations on small 70 × 70 patches: (**a**) input image, (**b**) ground truth segmentation map, (**c**) semantic segmentation results (SSSR branch, model *v*4 using Pixel Shuffle).

5.4. Comparison with Low-Resolution Predictions

In order to assess the usefulness of our approach, we trained a plain DeepLabV3+ architecture just for segmentation, using again as input the downsampled version of the original images and, as ground-truth segmentation maps, the donwsampled version of the relabeled ground truth maps. That is, no super-resolution is applied in this model. The goal was to compare the segmentation results obtained with this LR model and with the SSSR model (the high resolution semantic segmentation branch).

In this experiment, training DeepLabV3+ using the same procedure explained in Section 3.4, we reached a mIoU = 0.485, while our dual network method achieved mIoU = 0.535. Table 6 presents the precision, recall and IoU scores per class for both models (best result in bold). There is an increase in IoU in most of the classes, specially for classes with low IoU scores, such as marshes, peatbogs, natural material surfaces and permanent snow covered surfaces, as well as in the mean IoU. Additionally, the precision and recall were improved in the majority of classes, achieving a mean recall increase of almost 10%. Compared with DeepLabV3+, ourl model reduces the number of false negatives, especially in less frequent classes like marshes, vineyards and peatbogs, which agrees with the gain in the IoU scores.

Some qualitative results are shown in Figure 12, demonstrating the effectiveness of our method. The high resolution segmentation maps provide more details and a better definition of contours than the LR maps.

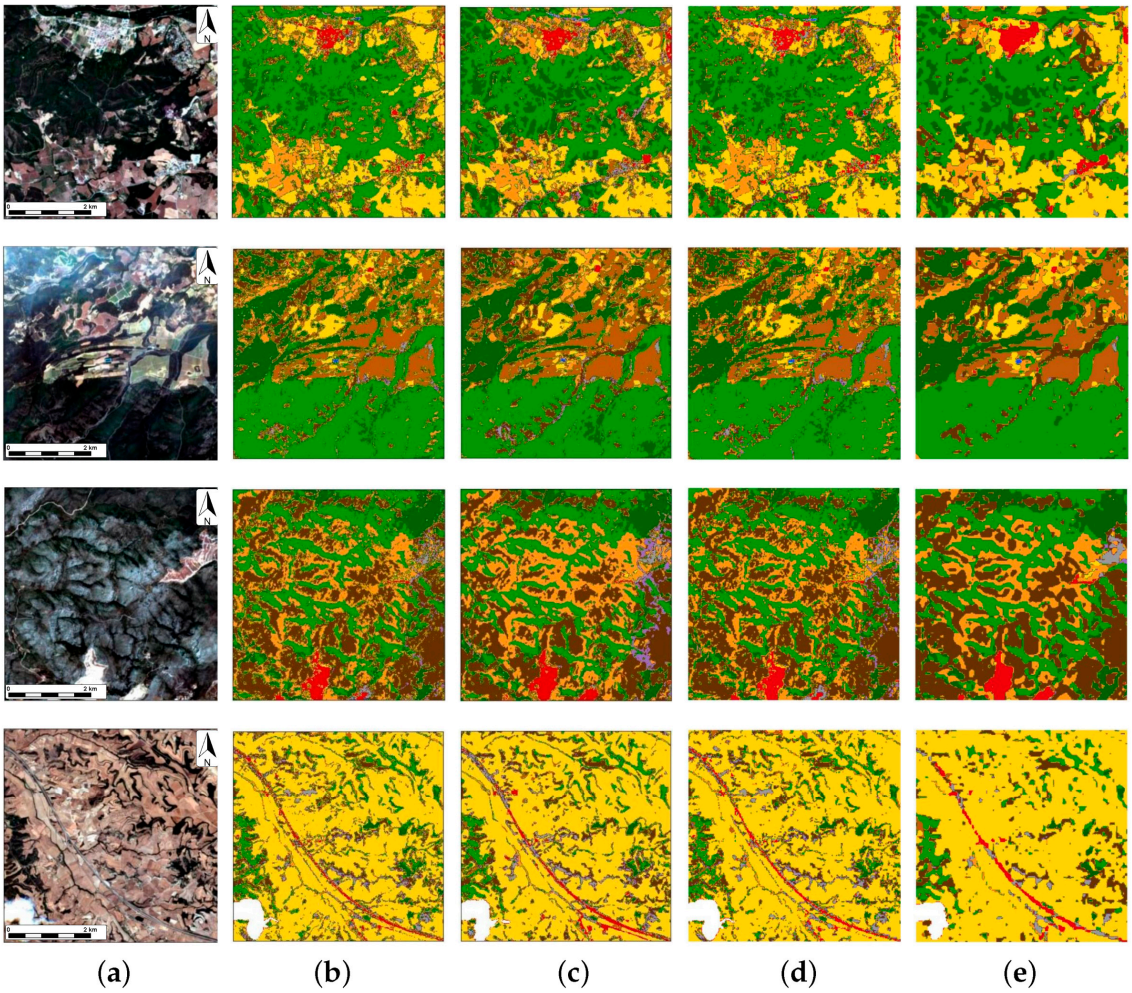


Figure 12. Comparison of segmentation results obtained with our dual approach (SSSR) and with DeepLabV3+ trained using a low-resolution dataset: (**a**) input image, (**b**) HR ground truth map, (**c**) our SSSR results, (**d**) LR ground truth map, (**e**) segmentation results on the LR image obtained with DeepLabV3+.

Table 6. Semantic segmentation results on S2GLC-Cat test set. Comparison between per-class IoU obtained by our dual approach network and the low-resolution DeepLabv3+. Best values are in bold.

	Classes	DeepLabV3+			Our SSSR Model		
		IoU	Precision	Recall	IoU	Precision	Recall
1	Clouds	**0.778**	**0.83**	0.93	0.755	0.78	**0.96**
2	Artificial surfaces and constructions	0.673	0.76	0.86	**0.730**	**0.80**	**0.89**
3	Cultivated areas	0.627	0.74	**0.80**	**0.645**	**0.82**	0.76
4	Vineyards	0.406	0.62	0.54	**0.474**	**0.63**	**0.66**
5	Broadleaf tree cover	0.634	0.76	**0.79**	**0.688**	**0.85**	0.78
6	Coniferous tree cover	0.659	0.78	**0.81**	**0.697**	0.85	0.80
7	Herbaceous vegetation	0.478	0.67	0.62	**0.525**	**0.70**	**0.68**
8	Moors and Heathland	0.445	**0.63**	0.61	**0.470**	0.60	**0.69**
9	Sclerophyllous vegetation	0.419	0.59	0.59	**0.480**	**0.65**	**0.65**
10	Marshes	0.238	**0.52**	0.30	**0.308**	0.38	**0.63**
11	Peatbogs	0.040	**0.36**	0.04	**0.136**	0.18	**0.37**
12	Natural material surfaces	0.309	**0.55**	0.41	**0.413**	0.54	**0.63**
13	Permanent snow covered surfaces	0.262	0.60	0.32	**0.365**	**0.67**	**0.45**
14	Water bodies	**0.817**	0.92	**0.88**	0.796	0.92	0.85
	Mean	0.485	0.667	0.608	**0.535**	**0.669**	**0.70**

6. Conclusions

The main objective of the work was to apply Deep Learning techniques to obtain high resolution segmentation maps from lower resolution multispectral Sentinel-2 imagery. We implemented a dual network approach based on an encoder-decoder structure where both tasks in the multi-task network share the same encoder as a feature extractor, but implement their own branch in the decoder path. The SISR branch produces a super-resolved version of the input image with a scale factor 2, and the SSSR branch that generates the semantic segmentation map also at double resolution. The model is based on the DeepLabv3+ architecture. We trained and tested the model on the S2GLC-Cat 2017 dataset.

Regarding the super-resolution metrics, we obtained a PSNR = 35.4239 and SSIM = 0.7756, which are higher than baseline interpolation methods (bicubic interpolation: PSNR= 34.1574, SSIM = 0.6732). As for the semantic segmentation metrics, we showed the increase in the mIoU due to the re-labeling task, and achieved mIoU = 0.535 on the relabeled dataset. This metric is not highly indicative due to noise produced by the method used to generate the ground truth land cover maps. Our model outperforms a DeepLabV3+ trained with the same LR images and predicts smooth, as well as accurate, segmentation maps. Quantitative and qualitative results demonstrate the effectiveness of the proposed approach.

Author Contributions: Conceptualization, S.A., L.S. and V.V.; data curation, S.A. and L.S.; methodology, S.A., L.S., J.M. and V.V.; software, S.A. and L.S.; supervision J.M. and V.V.; validation, L.S. and V.V.; formal analysis, S.A., L.S., J.M. and V.V.; resources, V.V.; writing–original draft preparation, S.A., L.S., and V.V.; writing–review and editing, S.A., L.S., J.M. and V.V.; funding acquisition, L.S., J.M. and V.V. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by the Spanish Research Agency (AEI) under project PID2020-117142GB-I00 of the call MCIN/AEI/10.13039/501100011033. L.S. would like to acknowledge the BECAL (Becas Carlos Antonio López) scholarship for the financial support.

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available at [62,63].

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

Abbreviations

The following abbreviations are used in this manuscript:

ASPP	Atrous Spatial Pyramid Pooling
CE	Cross Entropy
CNN	Convolutional Neural Network
DL	Deep Learning
DSR	Dual Super-resolution
FA	Feature Affinity
HR	High-Resolution
IoU	Intersection over Union
LR	Low-Resolution
LULC	Land Use and Land Cover
MAE	Mean Absolute Error
MSE	Mean Squared Error
NIR	Near Infra-red Band
OS	Output Stride
PSNR	Peak Signal to Noiser Ratio
RF	Random Forest
RS	Remote Sensing
SISR	Single Image Super-Resolution
S2GLC	Sentinel-2 Global Land Cover
SR	Super-Resolution
SSIM	Structural Similarity Index Measure
SSL	Semantic Segmentation Loss
SSSR	Semantic Segmentation Super-Resolution
SRL	Super-Resolution Loss
SVM	Support Vector Machine

Appendix A. Confusion Matrix Obtained for the Original Dataset

Table A1. Confusion matrix obtained for the original dataset and class IoU scores. Best values in bold.

	Classification Data	1	2	3	4	5	6	7	8	9	10	11	12	13	14	IoU
1	Clouds	0.19	0.29	0	0	0	0	0	0	0	0.05	0	0.11	0.04	**0.31**	0.129
2	Art. Surf	0	**0.87**	0.02	0.03	0	0	0	0	0.02	0.01	0	0.05	0	0	0.730
3	Cul. Areas	0	0.02	**0.76**	0.07	0	0	0.06	0.02	0.02	0.02	0.01	0.02	0	0	0.645
4	Vineyards	0	0.03	0.13	**0.68**	0	0	0.07	0.01	0.03	0	0.02	0.03	0	0	0.465
5	Broadleaf TC	0	0	0.01	0	**0.79**	0.08	0.05	0.04	0.03	0.01	0	0	0	0	0.685
6	Coniferous TC	0	0	0	0	0.10	**0.78**	0	0.04	0.05	0.02	0	0	0	0	0.691
7	Herb. Veg.	0	0	0.08	0.08	0.04	0	**0.64**	0.06	0.06	0.02	0.01	0.01	0	0	0.514
8	Moors and Heathland	0	0	0.01	0.02	0.06	0.02	0.04	**0.74**	0.07	0	0.02	0.05	0	0	0.464
9	Scl. Veg.	0	0.02	0.03	0.04	0.03	0.05	0.05	0.09	**0.63**	0.04	0	0.03	0	0	0.475
10	Marshes	0	0.03	0.02	0.02	0.04	0.04	0.04	0.03	0.11	**0.65**	0.01	0.02	0	0	0.302
11	Peatbogs	0	0	0.19	0.14	0.03	0	0.12	0.07	0.04	0.02	**0.35**	0.03	0	0.01	0.140
12	Nat. Mat. Surf.	0	0.1	0.06	0.05	0	0	0.01	0.04	0.06	0.01	0	**0.67**	0	0.02	0.430
13	Perm. Snow	0	0.01	0	0	0	0	0	0.01	0	0	0	**0.69**	0.28	0.02	0.213
14	Water Bodies	0	0.03	0.05	0	0	0	0	0	0	0.01	0	0.02	0	**0.89**	0.860
	mean															0.482

References

1. Venter, Z.S.; Sydenham, M.A.K. Continental-Scale Land Cover Mapping at 10 m Resolution Over Europe (ELC10). *Remote Sens.* **2021**, *13*, 2301. doi:10.3390/rs13122301. [CrossRef]
2. Malinowski, R.; Lewiński, S.; Rybicki, M.; Gromny, E.; Jenerowicz, M.; Krupiński, M.; Nowakowski, A.; Wojtkowski, C.; Krupiński, M.; Krätzschmar, E.; et al. Automated Production of a Land Cover/Use Map of Europe Based on Sentinel-2 Imagery. *Remote Sens.* **2020**, *12*, 3523. doi:10.3390/rs12213523. [CrossRef]
3. Alparone, L.; Aiazzi, B.; Baronti, S.; Garzelli, A. *Remote Sensing Image Fusion*; CRC Press: Boca Raton, FL, USA, 2015.

4. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 4. doi:10.1109/MGRS.2017.2762307. [CrossRef]
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. [CrossRef]
6. Tsagkatakis, G.; Aidini, A.; Fotiadou, K.; Giannopoulos, M.; Pentari, A.; Tsakalides, P. Survey of Deep-Learning Approaches for Remote Sensing Observation Enhancement. *Sensors* **2019**, *19*, 3929. doi:10.3390/s19183929. [CrossRef]
7. Ma, L.; Liu, Y.; Zhang, X.; Ye, Y.; Yin, G.; Johnson, B.A. Deep learning in remote sensing applications: A meta-analysis and review. *Isprs J. Photogramm. Remote Sens.* **2019**, *152*, 166–177. doi:10.1016/j.isprsjprs.2019.04.015. [CrossRef]
8. Goodfellow, I.; Bengio, Y.; Courville, A.; Bengio, Y. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016; Volume 1.
9. Vali, A.; Comai, S.; Matteucci, M. Deep Learning for Land Use and Land Cover Classification Based on Hyperspectral and Multispectral Earth Observation Data: A Review. *Remote Sens.* **2020**, *12*, 2495. doi:10.3390/rs12152495. [CrossRef]
10. Salgueiro Romero, L.; Marcello, J.; Vilaplana, V. Super-Resolution of Sentinel-2 Imagery Using Generative Adversarial Networks. *Remote Sens.* **2020**, *12*, 2424. doi:10.3390/rs12152424. [CrossRef]
11. Romero, L.S.; Marcello, J.; Vilaplana, V. Comparative study of upsampling methods for super-resolution in remote sensing. In Proceedings of the Twelfth International Conference on Machine Vision (ICMV 2019), Amsterdam, The Netherlands, 25–28 September 2019; pp. 417–424. doi:10.1117/12.2557357. [CrossRef]
12. Moliner, E.; Romero, L.S.; Vilaplana, V. Weakly Supervised Semantic Segmentation For Remote Sensing Hyperspectral Imaging. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 2273–2277.
13. Everingham, M.; Eslami, S.; Van Gool, L.; Williams, C.; Winn, J.; Zisserman, A. The PASCAL Visual Object Classes Challenge: A Retrospective. *Int. J. Comput. Vis.* **2015**, *111*, 98–136. doi:10.1007/s11263-014-0733-5. [CrossRef]
14. Mottaghi, R.; Chen, X.; Liu, X.; Cho, N.; Lee, S.; Fidler, S.; Urtasun, R.; Yuille, A. The role of context for object detection and semantic segmentation in the wild. In Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 891–898. doi:10.1109/CVPR.2014.119. [CrossRef]
15. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 3213–3223.
16. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 12. doi:10.1109/TPAMI.2016.2644615. [CrossRef]
17. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
18. Chen, L.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
19. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 9. doi:10.1109/TPAMI.2015.2389824. [CrossRef]
20. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
21. Khatami, R.; Mountrakis, G.; Stehman, S.V. A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research. *Remote Sens. Environ.* **2016**, *177*, 89–100. doi:10.1016/j.rse.2016.02.028. [CrossRef]
22. Talukdar, S.; Singha, P.; Mahato, S.; Shahfahad; Pal, S.; Liou, Y.A.; Rahman, A. Land-Use Land-Cover Classification by Machine Learning Classifiers for Satellite Observations—A Review. *Remote Sens.* **2020**, *12*, 1135. doi:10.3390/rs12071135. [CrossRef]
23. Kussul, N.; Lavreniuk, M.; Skakun, S.; Shelestov, A. Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 778–782. doi:10.1109/LGRS.2017.2681128. [CrossRef]
24. Ding, J.; Chen, B.; Liu, H.; Huang, M. Convolutional Neural Network With Data Augmentation for SAR Target Recognition. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 3. doi:10.1109/LGRS.2015.2513754. [CrossRef]
25. Parente, L.; Taquary, E.; Silva, A.P.; Souza, C.; Ferreira, L. Next Generation Mapping: Combining Deep Learning, Cloud Computing, and Big Remote Sensing Data. *Remote Sens.* **2019**, *11*, 2881. doi:10.3390/rs11232881. [CrossRef]
26. Li, Y.; Zhang, H.; Xue, X.; Jiang, Y.; Shen, Q. Deep learning for remote sensing image classification: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2018**, *8*, e1264. doi:10.1002/widm.1264. [CrossRef]
27. Mahdianpari, M.; Salehi, B.; Rezaee, M.; Mohammadimanesh, F.; Zhang, Y. Very Deep Convolutional Neural Networks for Complex Land Cover Mapping Using Multispectral Remote Sensing Imagery. *Remote Sens.* **2018**, *10*, 1119. doi:10.3390/rs10071119. [CrossRef]
28. Carranza-García, M.; García-Gutiérrez, J.; Riquelme, J.C. A Framework for Evaluating Land Use and Land Cover Classification Using Convolutional Neural Networks. *Remote Sens.* **2019**, *11*, 274. doi:10.3390/rs11030274. [CrossRef]
29. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Li, F.F. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.

30. Alhassan, V.; Henry, C.; Ramanna, S.; Storie, C. A deep learning framework for land-use/land-cover mapping and analysis using multispectral satellite imagery. *Neural Comput. Appl.* **2020**, *32*, 12. doi:10.1007/s00521-019-04349-9. [CrossRef]
31. Zhang, P.; Ke, Y.; Zhang, Z.; Wang, M.; Li, P.; Zhang, S. Urban Land Use and Land Cover Classification Using Novel Deep Learning Models Based on High Spatial Resolution Satellite Imagery. *Sensors* **2018**, *18*, 3717. 10.3390/s18113717. [CrossRef]
32. Zhang, C.; Sargent, I.; Pan, X.; Li, H.; Gardiner, A.; Hare, J.; Atkinson, P.M. An object-based convolutional neural network (OCNN) for urban land use classification. *Remote Sens. Environ.* **2018**, *216*, 57–70. doi:10.1016/j.rse.2018.06.034. [CrossRef]
33. Jozdani, S.E.; Johnson, B.A.; Chen, D. Comparing Deep Neural Networks, Ensemble Classifiers, and Support Vector Machine Algorithms for Object-Based Urban Land Use/Land Cover Classification. *Remote Sens.* **2019**, *11*, 1713. doi:10.3390/rs11141713. [CrossRef]
34. Yang, W.; Zhang, X.; Tian, Y.; Wang, W.; Xue, J.H.; Liao, Q. Deep Learning for Single Image Super-Resolution: A Brief Review. *IEEE Trans. Multimed.* **2019**, *21*, 3106–3121. doi:10.1109/TMM.2019.2919431. [CrossRef]
35. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1874–1883. https:/doi.org/10.1109/CVPR.2016.207.
36. Wang, L.; Li, D.; Zhu, Y.; Tian, L.; Shan, Y. Dual Super-Resolution Learning for Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 3773–3782. doi:10.1109/CVPR42600.2020.00383. [CrossRef]
37. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 8–11 December 2014; pp. 2672–2680.
38. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690.
39. Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; Change Loy, C. Esrgan: Enhanced super-resolution generative adversarial networks. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018; pp. 63–79.
40. Ma, W.; Pan, Z.; Guo, J.; Lei, B. Super-Resolution of Remote Sensing Images Based on Transferred Generative Adversarial Network. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Valencia, Spain, 22–27 July 2018; pp. 1148–1151. doi:10.1109/IGARSS.2018.8517442. [CrossRef]
41. Ma, W.; Pan, Z.; Yuan, F.; Lei, B. Super-Resolution of Remote Sensing Images via a Dense Residual Generative Adversarial Network. *Remote Sens.* **2019**, *11*, 2578. doi:10.3390/rs11212578. [CrossRef]
42. Chen, H.; Zhang, X.; Liu, Y.; Zeng, Q. Generative Adversarial Networks Capabilities for Super-Resolution Reconstruction of Weather Radar Echo Images. *Atmosphere* **2019**, *10*, 555. doi:10.3390/atmos10090555. [CrossRef]
43. Lanaras, C.; Bioucas-Dias, J.; Galliani, S.; Baltsavias, E.; Schindler, K. Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network. *ISPRS J. Photogramm. Remote Sens.* **2018**, *146*, 305–319. [CrossRef]
44. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image super-resolution using very deep residual channel attention networks. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 286–301.
45. Zhang, K.; Sumbul, G.; Demir, B. An Approach To Super-Resolution Of Sentinel-2 Images Based On Generative Adversarial Networks. In Proceedings of the Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS), Tunis, Tunisia, 9–11 March 2020; pp. 69–72.
46. Wagner, L.; Liebel, L.; Körner, M. Deep residual learning for single-image super-resolution of multi-spectral satellite imagery. In Proceedings of the Photogrammetric Image Analysis and Munich Remote Sensing Symposium, Munich, Germany, 18–20 September 2019; pp. 189–196.
47. Gargiulo, M.; Mazza, A.; Gaetano, R.; Ruello, G.; Scarpa, G. Fast super-resolution of 20 m Sentinel-2 bands using convolutional neural networks. *Remote Sens.* **2019**, *11*, 2635. doi:10.3390/rs11222635. [CrossRef]
48. Wu, J.; He, Z.; Hu, J. Sentinel-2 Sharpening via parallel residual network. *Remote Sens.* **2020**, *12*, 279. doi:10.3390/rs12020279. [CrossRef]
49. Zhu, X.; Xu, Y.; Wei, Z. Super-Resolution of Sentinel-2 Images Based on Deep Channel-Attention Residual Network. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Yokohama, Japan, 28 July–2 August 2019; pp. 628–631.
50. Gong, Y.; Liao, P.; Zhang, X.; Zhang, L.; Chen, G.; Zhu, K.; Tan, X.; Lv, Z. Enlighten-GAN for Super Resolution Reconstruction in Mid-Resolution Remote Sensing Images. *Remote Sens.* **2021**, *13*, 1104. doi:10.3390/rs13061104. [CrossRef]
51. Li, Y.; Li, B. Super-Resolution of Sentinel-2 Images at 10 m Resolution without Reference Images. *Automot. Eng.* **2021**. Available online: https://www.preprints.org/manuscript/202104.0556/v1 (accessed on 15 June 2021).
52. Wald, L.; Ranchin, T.; Mangolini, M. Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. *Photogramm. Eng. Remote Sens.* **1997**, *63*, 691–699.
53. Armannsson, S.E.; Ulfarsson, M.O.; Sigurdsson, J.; Nguyen, H.V.; Sveinsson, J.R. A Comparison of Optimized Sentinel-2 Super-Resolution Methods Using Wald's Protocol and Bayesian Optimization. *Remote Sens.* **2021**, *13*, 2192. doi:10.3390/rs13112192. [CrossRef]

54. Dai, D.; Wang, Y.; Chen, Y.; Van Gool, L. Is image super-resolution helpful for other vision tasks? In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016; pp. 1–9.
55. Haris, M.; Shakhnarovich, G.; Ukita, N. Task-Driven Super Resolution: Object Detection in Low-resolution Images. *arXiv* **2018**, arXiv:1803.11316.
56. Guo, Z.; Wu, G.; Song, X.; Yuan, W.; Chen, Q.; Zhang, H.; Shi, X.; Xu, M.; Xu, Y.; Shibasaki, R.; et al. Super-resolution integrated building semantic segmentation for multi-source remote sensing imagery. *IEEE Access* **2019**, *7*, 99381–99397. doi:10.1109/ACCESS.2019.2928646. [CrossRef]
57. Pereira, M.B.; dos Santos, J.A. How Effective Is Super-Resolution to Improve Dense Labelling of Coarse Resolution Imagery? In Proceedings of the Conference on Graphics, Patterns and Images (SIBGRAPI), Rio de Janeiro, Brazil, 28–31 October 2019; pp. 202–209. doi:10.1109/SIBGRAPI.2019.00035. [CrossRef]
58. Lei, S.; Shi, Z.; Wu, X.; Pan, B.; Xu, X.; Hao, H. Simultaneous super-resolution and segmentation for remote sensing images. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS),Yokohama, Japan, 28 July–2 August 2019; pp. 3121–3124.
59. Pereira, M.B.; Santos, J.A.d. An End-To-End Framework For Low-Resolution Remote Sensing Semantic Segmentation. In Proceedings of the IEEE Latin American GRSS ISPRS Remote Sensing Conference (LAGIRS), Santiago, Chile, 22–26 March 2020; pp. 6–11. doi:10.1109/LAGIRS48042.2020.9165642. [CrossRef]
60. ISPRS 2D Semantic Labeling Contest. Available online: https://www2.isprs.org/commissions/comm2/wg4/benchmark/semantic-labeling/ (accessed on 21 June 2021).
61. Pan, B.; Tang, Z.; Liu, E.; Xu, X.; Shi, T.; Shi, Z. SRDA-Net : Super-Resolution Domain Adaptation Networks for Semantic Segmentation. *arXiv* **2020**, arXiv:2005.06382
62. Sentinel-2 Global Land Cover Dataset. Available online: http://s2glc.cbk.waw.pl/ (accessed on 10 June 2021).
63. Copernicus Open Access Hub. European Space Agency. Available online: https://scihub.copernicus.eu/dhus/#/home (accessed on 21 March 2021).
64. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778. doi:10.1109/CVPR.2016.90. [CrossRef]
65. Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. *arXiv* **2016**, arXiv:1606.02147.
66. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]