Article

A Multi-Scale Deep Neural Network for Water Detection from SAR Images in the Mountainous Areas

Lifu Chen ^{1,2} , Peng Zhang ^{1,2}, Jin Xing ^{3,*} , Zhenhong Li ^{3,4} , Xuemin Xing ^{2,5} and Zhihui Yuan ^{1,2}

¹ School of Electrical and Information Engineering, Changsha University of Science and Technology, Changsha 410114, China; lifu_chen@csust.edu.cn (L.C.); 17105020588@stu.csust.edu.cn (P.Z.); yuanzhihui@csust.edu.cn (Z.Y.)

² Laboratory of Radar Remote Sensing Applications, Changsha University of Science and Technology, Changsha 410114, China; xuemin.xing@csust.edu.cn

³ School of Engineering, Newcastle University, Newcastle NE17RU, UK; zhenhong.li@newcastle.ac.uk

⁴ College of Geological Engineering and Geomatics, Chang'an University, Xi'an 710054, China

⁵ School of Traffic and Transportation Engineering, Changsha University of Science and Technology, Changsha 410114, China

* Correspondence: jin.xing@newcastle.ac.uk; Tel.: +44-191-208-6080

Received: 2 August 2020; Accepted: 28 September 2020; Published: 1 October 2020

Abstract: Water detection from Synthetic Aperture Radar (SAR) images has been widely utilized in various applications. However, it remains an open challenge due to the high similarity between water and shadow in SAR images. To address this challenge, a new end-to-end framework based on deep learning has been proposed to automatically classify water and shadow areas in SAR images. This end-to-end framework is mainly composed of three parts, namely, Multi-scale Spatial Feature (MSF) extraction, Multi-Level Selective Attention Network (MLSAN) and the Improvement Strategy (IS). Firstly, the dataset is input to MSF for multi-scale low-level feature extraction via three different methods. Then, these low-level features are fed into the MLSAN network, which contains the Encoder and Decoder. The Encoder aims to generate different levels of features using residual network of 101 layers. The Decoder extracts geospatial contextual information and fuses the multi-level features to generate high-level features that are further optimized by the IS. Finally, the classification is implemented with the Softmax function. We name the proposed framework as MSF-MLSAN, which is trained and tested using millimeter wave SAR datasets. The classification accuracy reaches 0.8382 and 0.9278 for water and shadow, respectively; while the overall Intersection over Union (IoU) is 0.9076. MSF-MLSAN demonstrates the success of integrating SAR domain knowledge and state-of-the-art deep learning techniques.

Keywords: SAR; deep learning; convolutional neural network; attention mechanism; water and shadow classification

1. Introduction

Detecting water bodies from Synthetic Aperture Radar (SAR) images has been a very active research field [1]. Detection results are widely applied to reduce errors in SAR phase unwrapping, monitor flooding and assess damage, track water storage changes over various periods, investigate the area increase and decrease of wetlands and delineate shoreline movement [2–4]. Due to the speckle noise in SAR images and confusing characteristics of other objects (e.g., shadows) in SAR images, automatic water body detection from SAR images with high precision remains an open challenge. Karvonen et al. [5]

have proposed a semi-automated classification algorithm of water areas using Radarsat-1 SAR imagery, in which the average accuracy was merely around 70%. Moreover, Martinis et al. [6] compared advantages and disadvantages of four semi-automatic/automatic water detection algorithms and these algorithms achieved similar overall accuracy (i.e., ~85%) in the evaluation experiments. Their study has emphasized the importance of auxiliary datasets (e.g., digital elevation models) and parameter tuning in reaching decent accuracy. Other semi-automatic/automatic water detection approaches have also been explored but their performance is unsatisfying [7–9].

Shadow is one of the most important challenges in the process of water detection, especially in mountainous regions [10]. Both water and shadow appear as dark areas in SAR images (i.e., low gray values), which often results in false detection of water bodies. To achieve higher accuracy in automatic water detection, simultaneous SAR image classification of water and shadow is employed in this paper. SAR image classification is to identify different targets from SAR images, which is being ameliorated considerably by deep learning techniques [11]. Deep learning mimics the workflow of human brain using artificial neural networks that combine low-level features to form higher-level features for object detection, language processing and decision-making [12]. Krizhevsky et al. [13] adopted a convolutional neural network (CNN) to achieve 10% accuracy improvement in the ImageNet contest. This was the first time the deep learning outperformed traditional handcrafted features with shallow models in object detection. From then on, the blooming of deep learning has begun in computer vision research, such as object detection [14] and semantic segmentation [15]. The success of deep learning has inspired us to design and develop a special neural network to address the challenge of simultaneous water and shadow classification in this paper.

The study of SAR image classification using deep learning techniques is developing rapidly. Zhu et al. [16] summarized the progress of deep learning in remote sensing and highlighted research challenges of employing deep learning in the analytics and interpretation of SAR images. Hänsch et al. [17] proposed the first application of Complex-valued CNN (C-CNN) for object recognition in Polarization SAR (Pol-SAR) data. Though the architecture was only one single convolutional layer, the results have surpassed those from standard CNN. Song et al. [18] proposed an efficient and accurate water area classification method, which required precise Digital Elevation Model (DEM) and Digital Surface Model (DSM) datasets. Huang et al. [9] investigated a tree-based automatic classification approach to analyze surface water using Sentinel-1 SAR data. However, this method was problematic in distinguishing shadow and water areas. Therefore, how to employ deep learning techniques without auxiliary datasets for water detection becomes a major challenge to solve.

We seek solutions from SAR domain knowledge and geospatial analytics. Although there is great advancement in deep learning to perform image classification, domain knowledge in SAR image analytics has not been integrated in the design and development of deep neural networks. Because most deep learning techniques (e.g., CNN) are designed for the classification of daily objects (e.g., cats and dogs), image features extracted by neural networks cannot provide satisfying performance in water and shadow classification due to their high difference [11]. Meanwhile, the high computation cost of deep neural network training has also been noticed [16]. To address these problems, we embed prior SAR image feature extraction algorithms into the architecture of deep neural network to achieve higher classification accuracy and lower training costs.

Meanwhile, geospatial contextual information should also be considered in bridging deep learning and SAR image analytics. Notti et al. [7] assessed the potential and limitations for the flood mapping using various satellite imagery datasets, including MODIS, Proba-V, Landsat, Sentinel-2 and Sentinel-1. They emphasized that geospatial contextual information in the automatic detection of inundated areas was essential but had not been given enough attention. Contextual image features usually encode spatial structures of geographic objects, which play pivotal roles in classifying similar objects [19]. Although multi-level contextual image features could be extracted by the spatial pyramid pooling technique within a image tile [20], there is no guarantee that the given image tile could cover the right geospatial extent of target objects. This further echoes the necessity of merging geospatial analytics within deep learning.

In this paper, we propose a new end-to-end framework called Multi-scale Spatial Feature Multi-Level Selective Attention Network (MSF-MLSAN), which is built on state-of-the-art deep learning techniques, SAR domain knowledge and geospatial context analytics, to implement water and shadow classification. MSF-MLSAN contains three parts. The first part is the Multi-scale Spatial Features (MSF) to extract multi-scale and low-level SAR image features. The second part is the Multi-Level Selective Attention Network (MLSAN), which extracts and refines middle-level and the high-level SAR image features. It contains Encoder module and Decoder module. The third part is named Improvement Strategy (IS), including the score map weighting and the splicing strategy to optimize simultaneous classification results. The contribution of this paper is two-fold. On the one hand, we propose MSF-MLSAN, which integrates existing domain knowledge in SAR image analytics with deep learning techniques, as a successful approach to achieve simultaneous classification of water and shadow areas. On the other hand, the pivotal role of geospatial context has been noticed in the design of SAR-specific neural networks to implement the automatic detection of geographic objects with high similarities in SAR images.

The rest of this paper is organized as follows—the methodology of our proposed neural network, MSF-MLSAN, for simultaneous classification of water and shadow is presented in Section 2. Section 3 introduces the datasets, performance assessment and result interpretation. The improvement and extension of MSF-MLSAN are discussed in Section 4. Finally, the conclusion is given in Section 5.

2. Methodology

2.1. Foundation

Great progress has been achieved to classify daily objects with deep learning techniques, for example, RefineNet, which combines image features from different resolutions using multipath refinement, has improved the intersection-over-union (IoU) score of PASCAL VOC 2012 dataset to 83.4 [21]. However, the lack of geospatial contextual information remains an open challenge in processing remote sensing images, since geographic objects present much larger scale variance than daily objects that routinely requires the support of contextual information across a number of image tiles to detect [22]. In this paper, we employ the attention mechanism and spatial pyramid pooling to handle geospatial contextual information.

(1) Attention mechanism

Attention mechanism has developed from human visual system [23]. It scans the whole image to detect important regions, to which more attention needs to be assigned. More attention means more weights of features from these regions need be considered for classification. In recent years, attention mechanism has been successfully applied in many fields, such as natural language processing [24] and image classification [25].

Based on this mechanism, Squeeze-and-Excitation Network (SENet) was proposed and won the first place in ILSVRC 2017 classification competition, achieving a ~25% relative improvement over the winner of 2016 [26]. It first performed a Squeeze operation on the feature map obtained by the convolution to extract channel-level features, then employed an Excitation operation on these features to learn the relationship among channel as assigned weights, which were used to calculate final features for classification.

(2) Pyramid pooling

While SENet can extract the channel-wise geospatial contextual information from feature maps, the pyramid pooling [20] techniques have also been successful in extracting contextual information from the spatial structure of feature maps. Generally, the size of a receptive field determines how much contextual information could be extracted using pyramid pooling. Since geographic objects usually cover larger spatial extents than the size of receptive fields, we combine the pyramid pooling with the Global Attention Pooling (GAP) [27] to capture contextual information with high scale variance.

2.2. The Framework of MSF-MLSAN

MSF-MLSAN, the innovative framework for water classification from SAR images, is shown in Figure 1. MSF-MLSAN mainly contains three parts: Multi-scale Spatial Features (MSF) extraction, Multi-Level and Selective Attention Network (MLSAN) and Improvement Strategy (IS).

Figure 1. The framework of the Multi-scale Spatial Feature Multi-Level Selective Attention Network (MSF-MLSAN) algorithm.

First, SAR imagery datasets are input to the MSF module. MSF extracts three prior SAR image features from the input images, namely, the Gabor feature, Gray-Level Gradient Co-occurrence Matrix (GLGCM) feature and Multi-scale Omnidirectional Gaussian Derivative Filter (MSOGDF) feature. These three types of low-level features could be further fused to forge a group of three-channel feature maps. Therefore, we can get three groups of feature maps from these three of low-level features. In addition, the fourth group of three-channel feature map will be generated by SAR images. Then the four groups of three-channel feature maps will be input to MLSAN to extract corresponding high-level features, which produces four Score Maps (SMs). SMs are processed by the Improvement Strategy to generate final maps for classification. Details of each part of MSF-MLSAN are delineated in the following discussion.

2.3. Multi-Scale Spatial Features (MSF)

MSF aims to acquire multi-scale spatial features from SAR images by integrating prior SAR image feature extraction algorithms with deep neural networks. Different features of objects could be enhanced by transformations of original SAR images, which are utilized to achieve better classification accuracy. In this paper, we employ prior SAR features extraction methods, namely GLGCM, Gabor transformation and MSOGDF to extract low-level features and then we fuse them as higher-level features.

2.3.1. GLGCM Extraction

GLGCM can effectively exploit the grayscale and the gradient of adjacent pixels in SAR images and generate various statistical characteristics, including 15 commonly used ones. We select grayscale mean, gradient mean and grayscale mean square error in this paper according to Reference [11].

2.3.2. Gabor Transformation

Gabor transformation is a kind of Windowed Fourier transformation. It can extract relevant features at different scales and directions in the frequency domain. The Gabor transformation resembles the biological function of human eyes, since it is often used to recognize the texture of targets. There are usually four characteristics from four directions and we select the two characteristics from 45° and 135° following our previous work [11].

2.3.3. MSOGDF

The spatial structure features contain important information about the visual characteristics of images. To acquire these features, a Multi-Scale Omnidirectional Gauss Derivative Filter (MSOGDF) is

introduced, which can generate features with different directions from the image. We heuristically select the features with two directions on the specific scale, which are 45° and 135° [11].

2.3.4. Multi-Scale Space Statistical Features Fusion

Features extracted by GLGCM, Gabor and MSOGDF are concatenated as the input of MLSAN. In this paper, four groups of three-channel fusion feature maps are generated. The first group is produced by concatenating the gray-mean, the gradient mean and the gray mean square error characteristics of SAR images from GLGCM. The second group is generated by concatenating the 45° feature and 135° feature from Gabor transformation and the SAR image. The third group is generated by merging the 45° feature and 135° feature from MSOGDF and SAR image, which uses $\sigma = 1$ and the $k = 1$. The fourth group is the original SAR images. All four groups of fusion feature maps are ground-referenced.

2.4. MLSAN

There are several classical and widely used networks for image classification proposed in recent years, such as FCN [28], SegNet [15], DeepLab [29], RefineNet [21] and PSPNet [20]. These networks have been successfully applied in image classification. However, there are still problems in these networks, such as the poor classification results due to information lose in the pooling operation and inter-class overlapping caused by the lack of multi-scale contextual information [16]. MLSAN network is proposed to address the geospatial context challenge in SAR image analysis, which is quite different from the classification of daily objects.

MLSAN contains two parts, the Encoder part and the Decoder part, as shown in Figure 1.

2.4.1. The Encoder

The Encoder part is based on residual networks [30] to extract the middle-level and high-level features of the input images. In this paper, resnet_v2_101 is adopted. The forward operation of the whole network is a process of continuously solving the residuals, accompanied by the resolution decreasing of feature maps and the dimension increasing. Supposing the size of the input image is $512 \times 512 \times 3$, then the feature map F0 with the size of $128 \times 128 \times 64$ is first generated by a convolution and a pooling operation. Second, the feature map F1 with the size of $128 \times 128 \times 256$ is generated from Res-1 including three residual units [30]. Third, the feature map F2 with the size of $64 \times 64 \times 512$ is outputted from Res-2 using four residual units. Fourth, we generate the feature map F3 with $32 \times 32 \times 1024$ from Res-3 via 23 residual units. Finally, the feature map F4 of size $16 \times 16 \times 2048$ will be outputted from Res-4 through three residual units.

The residual network is built by the stacking of residual units. The residual unit is based on the idea of identity mapping to guarantee the propagation of gradient information to the lower layer, which enables the parameters training through the whole network. Resnet_v2_101 used in this paper makes use of the model pre-trained with ImageNet. We follow [31] to use feature migration instead of the random initialization of parameters.

2.4.2. The Decoder

The high-level features are usually used to distinguish different categories, while low-level features often offer too many details of the objects. Therefore, we carefully design the decoder to merge and enhance high-level and low-level features.

There are four feature maps generated from the Encoder network, namely, the high-level feature F4 ($16 \times 16 \times 2048$) from Res-4, the middle-level features F3 ($32 \times 32 \times 1024$), F2 ($64 \times 64 \times 512$) and F1 ($128 \times 128 \times 256$) from Res-3, Res-2 and Res-1, respectively. To reduce the computational complexity of the Decoder part and the number of redundant features, convolution is incurred between the encoding and decoding operations to perform the dimension reduction. The final dimensions of feature maps inputted to the Decoding network are $16 \times 16 \times 512$, $32 \times 32 \times 256$, $64 \times 64 \times 256$ and $128 \times 128 \times 256$.

The Decoding network is the opposite of the encoding network. It uses the bottom-up approach, in which, high-level and sub-high-level features outputted from the encoder network is stepwise merged into new feature maps. Meanwhile, resolutions of feature maps increase while dimensions decrease.

All the four features maps (F_1 , F_2 , F_3 and F_4) are inputted to decoder modules, M_1 , M_2 , M_3 and M_4 , respectively (Figure 1). The three modules (M_1 , M_3 and M_4) are designed with the same architecture, namely, the Feature Map Adaption and Fusion (FMAF) and the Residual and Attention Pooling (RAP), as shown in Figures 2 and 3. The M_2 module contains two parts, namely, the FMAF and the Pyramid and Attention Pooling (PAP), which is depicted in Figure 4.

Figure 2. The Feature Map Adaption and Fusion (FMAF) architecture.

Figure 3. The Residual and Attention Pooling.

Figure 4. The pyramid and attention pooling.

- FMAF module

The FMAF is designed for adjusting the size of feature maps and fusing features at different levels. The architecture is shown in Figure 2. The inputs are the feature maps at different levels. Then they will be processed by two residual convolutional units (RCU), one convolution and one up-sampling for scaling. Finally, they will be summed to produce the fusion map.

- RAP module

The Residual and Attention Pooling (RAP) aims to capture geospatial contextual information from a large area and pool the input features by multi-pooling layers. The architecture of RAP is shown in Figure 3. It is composed of a series of pooling modules and each module contains one convolution layer and one Maximum pooling layer. The output of the previous pooling layer is used as the input to the latter pooling layer. Therefore, the subsequent pooling layer can handle image features over larger areas with small pooling windows. After fusing features with pooling layers, the attention module is used to weight the fused features, which can highlight useful features and decline redundant features. Finally, weighted features are fused with initial input features and results will be inputted to the RCU. In the RAP, four pooling modules are employed. The window size of each pooling layer is 5×5 and the stride is 1×1 .

- PAP module

The design of Pyramid and Attention Pooling (PAP) module is shown in Figure 4. Processed by the pooling layer, the size of the input feature map (i.e., $64 \times 64 \times 256$) will become $1 \times 1 \times 256$, $2 \times 2 \times 256$, $3 \times 3 \times 256$ and $6 \times 6 \times 256$. Then, their dimension will be reduced by the convolution layer to $1 \times 1 \times 64$, $2 \times 2 \times 64$, $3 \times 3 \times 64$ and $6 \times 6 \times 64$. Furthermore, they will all be up-sampled to $64 \times 64 \times 64$ by the bilinear interpolation and concatenated to form one feature map with the initial dimension ($64 \times 64 \times 256$). It will also be weighted by the attention module. Then, it will be concatenated with the input feature map to generate a feature map ($64 \times 64 \times 512$) and recovered to the initial dimension ($64 \times 64 \times 256$) by a convolution layer. Finally, the feature map will be processed by an RCU to generate features for further processing. The pyramid pooling fuses four different scales of features, including global pooling and local features of different sub-regions, in which, the attention mechanism is employed to enhance the extraction of the contextual information.

In this paper, the combination of RAP and PAP enhances the network's ability to extract contextual information significantly and also increases the variety of extracted information. By this means, the context of geographic objects has been taken into account in the architecture of deep neural networks. Meanwhile, all modules of the network are built on top of the residual connection, which ensures the efficiency of the network training.

The detailed decoding process is depicted as follows:

According to Figure 1, the feature map **F4** outputted from **Res-4** enters the **M4** module. In which, **F4** will be processed by the RAP module to extract contextual information and the new feature map **F4_2** will be inputted to the module **M3**. Therefore, there are two inputs for **M3** module, namely, **F4_2** and the middle-level feature map **F3** outputted from **Res-3**. First, these two different-level feature maps will be processed by the FMAF module for feature fusion. Then, the fused feature map will be processed by the RAP module to extract contextual information with different scales. Finally, the new feature map **F3_2** is generated. For **M2** module and **M1** module, there are similar operations as **M3** module. The difference lies in that PAP is used to extract contextual information in **M2** module, while RAP is used in the other three modules (**M1**, **M3** and **M4**). Once the **F1_2** feature map is generated, it will be processed by two RCUs to increase the non-linearity. The size of the feature map is $128 \times 128 \times 256$ at this stage. Later, one up-sampling layer and one convolution layer are used to recover the dimension of the feature map to $512 \times 512 \times N_{class}$ (N_{class} is the number of the classes). Finally, a dense score map will be generated by the Softmax layer.

2.5. Improvement Strategy (IS)

Although MSF-MLSAN could be employed to classify large-scale images, the challenge of context still exists—it is possible that the given image tile contains very limited contextual information or even covers the target object partially. To better handle the geographic context of classification and improve the overall accuracy, two strategies are invented, namely, splicing and weighting.

Inspired by Reference [32], we splice the two adjacent image slices with the method shown in Figure 5. While testing, we will use a sliding window of size 512×512 to cut the large-scale images and the step for the sliding is 256, such as the sliding windows $S_{11}, S_{12}, S_{21}, \dots, S_{mn}$ shown in Figure 5. To solve the discontinuity of the classification results at the junction of two adjacent windows by direct splicing, the sliding step of the next cutting window is only 256, which is half of the two adjacent windows are overlapping (such as S_{11} and S_{12} or S_{11} and S_{21}). After both two windows are tested, the final classification result of the overlapping area will be generated by averaging the results of the two adjacent windows. By this splicing, we can get the score map tested with minimum splitting border errors [22].

Figure 5. The sketch map of the splicing strategy.

In this paper, four trained models are generated eventually. And we can obtain four score maps via the splicing method. Then, the four score maps will be weighted to generate the final score map according to their respective classification performance (shown in Figure 1).

The whole network is organized by long-range residual connections between the blocks in encoder and decoder. Hence, it consists of many stacked “Residual Units.”

One Residual Unit performs the following computation:

$$y_l = h(X_l) + F(X_l, W_l) \quad (1)$$

$$X_{l+1} = f(y_l), \quad (2)$$

where X_l and W_l denote the input feature and a set of weights (and biases) associated with the l -th residual unit. F is the residual function. The function h and f are set as an identity mapping: $h(X_l) = X_l$, $f(X_l) = X_l$.

By applying the recursion several times, the output of the residual unit admits a representation of the form

$$X_L = X_l + \sum_{i=l}^{L-1} F(X_i, W_i). \quad (3)$$

The implementation detail of training for the whole network is summarized in the following Algorithm 1.

Algorithm 1: Training the whole network

Input: datasets, including SAR slice images, corresponding coherence maps, phase maps and ground truth.

- 1: Multi-scale features extraction and features fusion to generate four groups of fusion maps.
- 2: Initialization: The encoder part loads the model pre-trained from ImageNet. The decoder partis initialized by truncated normal distribution.
- 3: Training the network by BP algorithm.
- 4: Fine-tuning the network:

Calculate the loss function

$$loss = -\sum \left(f_{ijc} - \log \sum_{d=1}^D e^{f_{ijd}} \right),$$

where f is the feature extracted from the last layer in the network. c denotes the ground truth class and d denotes the dimension of the feature map.

- 5: **for** epoch = 1 to i **do**

6: **Forward:**

$$X_1 = X_0 + F(X_0, W_0),$$

where X_0 denotes the input image.

... ..

$$X_L = X_0 + \sum_{i=0}^{L-1} F(X_i, W_i).$$

7: **Backward:**

$$\frac{\partial loss}{\partial X_i} = \frac{\partial loss}{\partial X_L} \cdot \frac{\partial X_L}{\partial X_i} = \frac{\partial loss}{\partial X_i} \left(1 + \frac{\partial}{\partial X_i} \sum_{i=1}^{L-1} F(X_i, W_i) \right)$$

$$\frac{\partial loss}{\partial X_i} = \frac{\partial loss}{\partial X_i} \cdot \frac{\partial X_i}{\partial W_i} = \frac{\partial loss}{\partial W_i} \left(\frac{\partial loss}{\partial X_L} + \frac{\partial loss}{\partial X_L} \frac{\partial}{\partial X_i} \sum_{i=1}^{L-1} F(X_i, W_i) \right),$$

which means the information can directly propagated to any shallower unit and solves the problem of gradient disappearance.

$$W'_i = W_i - \alpha \frac{\partial loss}{\partial W_i},$$

where α denotes the learning rate.

8: **end for**

Output: the model of the network

3. Experiments

3.1. Datasets

The experiments use imagery data from the millimeter SAR system to evaluate MSF-MLSAN. The central frequency of the system is 35 GHz and the resolutions are 0.13 m in the range and 0.14 m in the azimuth. In the SAR system, there are considerable water and shadow areas in the mountainous areas because of the relative low flight height (2000 m to 4000 m). There are nine large-scale SAR images used in experiments and the size of each is 10240×13050 pixels. Eight of these images are used for training and validation, while the remaining one is used for model testing. We choose three labels, namely, water, shadow and background. The ground truth labels are generated by manual marking and confirmed by SAR experts from Beijing Institute of Radio Measurement (the data provider). The 8 large-scale images are cut to lots of SAR tiles with a size of 720×720 for training and validation purposes. There are 1288 SAR image tiles in the datasets generated from the eight images. We select 80% of the samples randomly as the training datasets and the rest are kept as the validation datasets. Figure 6 shows examples of the datasets, in which, (a) and (b) denote the SAR image and the ground truth respectively.

Figure 6. An example of Synthetic Aperture Radar (SAR) datasets. (a) SAR image. (b) Ground truth.

Our MSF-MLSAN has been developed using Tensorflow 1.10, CUDA 9.0 and Python 3.6 on the server with Titan xp 12 GB, CPU of i7-6800k, 16 GB memory and 2 TB hard disk. The detailed testing steps can be found in Appendix A (Algorithm A1). In the experiment, 100 epochs are trained, one of which is to train all the pictures in the training set once. The total training time of MSF-MLSAN for the dataset is about 20 h.

3.2. Performance Indices

To assess the classification accuracy, two important indices are used in the paper, namely, Overall Accuracy (OA) and Intersection over Union (IoU).

- Overall Accuracy

OA is a vital index to evaluate the classification performance of the given algorithm, which is computed by the following equation supposing for type A.

$$OA_A = \frac{GT_A \cap D_A}{GT_A}, \quad (4)$$

where GT_A is all the pixels of ground truth for type A and D_A is all the pixels detected as type A. $GT_A \cap D_A$ denotes all the pixels of intersection area of GT_A and D_A .

As shown in Figure 7, the red rectangles contain all the pixels of ground truth for type A (B or C) and the blue rectangles contain all pixels detected as type A (B or C). While the overlapping areas denote successfully detected results. Therefore, OA is the ratio of correctly classified pixels (the total overlapping pixels) to pixels of ground truth. The larger the OA value is, the higher the classification accuracy we would achieve.

Figure 7. The schematic diagram of computing overall accuracy. (A–C) are the overlapping areas of detected results and ground truth for type A, type B and type C.

- Intersection over Union

IoU is an effective metric for evaluating classification accuracy, which is an important supplement to OA. According to Figure 7, the pixel number of ground truth for type A is the same as type C but the number of pixels detected as type C (D_C) is much larger than the number of pixels detected as type A (D_A). So, the value of OA for type C is higher than that of type A. But there are more pixels detected as type C, which belong to other types. Since OA cannot fully delineate the classification performance, IoU has been incurred in our experiment.

IoU is computed by Equation (2) for type A

$$IoU = \frac{GT_A \cap D_A}{GT_A \cup D_A}. \quad (5)$$

Therefore, IoU is the ratio of the intersection of GT_A and D_A to the union of GT_A and D_A . Then, for Figure 7A,C, IoU for type C is lower than that of type A because many pixels have been falsely detected as type C.

Hence, for a classification experiment, the performance is better with higher values of both OA and IoU.

3.3. MSF Results

The MSF integrates prior SAR image features with deep learning techniques for multi-level feature extraction. Figure 8 presents the Gabor features of the SAR image. Figure 8a shows the SAR image, which covers some water and shadow areas. Figure 8b,c indicate the Gabor feature with the direction of 45-degree and 135-degree respectively. From the two figures we can find out that they effectively embody the local texture features in different directions. Figure 8d gives the fused feature map concatenated from Figure 8a–c, which clearly reflects the texture of local areas.

Figure 8. Gabor transformation results. (a) SAR image. (b) Gabor transformation of 45-degree. (c) Gabor transformation of 135-degree. (d) the concatenation results of Gabor transformation.

Figure 9 depicts the GLGCM feature of the SAR image. Figure 9a illustrates the gray mean, which is useful for reducing the speckle noise. Figure 9b presents the gradient mean map, which highlights the water and shadows to improve the classification accuracy. Figure 9c demonstrates the gray mean square error which blurs other objects but water and shadow areas. Figure 9d is the fusion map of concatenating Figure 9a–c, from which the water and shadow edges are enhanced. Therefore, GLGCM offers strong low-level features for water and shadow classification.

Figure 9. The Gray-Level Gradient Co-occurrence Matrix (GLGCM) results. (a) the gray-mean result. (b) the gradient mean result. (c) the gray mean square error. (d) the concatenation results of GLGCM.

Figure 10 delineates MSOGDF features from SAR images. Figure 10a shows the original SAR image. Figure 10b,c are the MSOGDF features with the direction of 45-degree and 135-degree respectively. Figure 10d is the fusion result of Figure 10a–c, as an enhanced feature for simultaneous classification of water and shadow.

Figure 10. The results of Multi-scale Omnidirectional Gaussian Derivative Filter (MSOGDF). (a) SAR image. (b) MSOGDF characteristic of 45-degree. (c) MSOGDF characteristic of 135-degree. (d) the concatenation results of MSOGDF.

3.4. Classification Results and Analysis

In order to test MSF-MLSAN, we compare it with six widely applied classification methods: RefineNet [21], DeepLabV3 [33], MLSAN, GLGCM-MLSAN, Gabor-MLSAN and MSOGDF-MLSAN. MLSAN means the datasets are processed by the proposed framework (Figure 1) without the MSF part. GLGCM-MLSAN means we only utilize GLGCM feature in the MSF part. Gabor-MLSAN and MSOGDF-MLSAN are implemented with Gabor feature and MSOGDF feature respectively in the proposed network.

In the experiment, two SAR images are tested: one covers more water but fewer shadow areas and the other contains fewer water areas but more shadow areas. These two experiments are designed to test if MSF-MLSAN could simultaneously classify water and shadow, given their coverage is heterogenous.

3.4.1. The Experiment with More Water but Fewer Shadow Areas

Figure 11 shows the data of the first experiment, which includes the SAR image and the ground truth. The size of the SAR image in the first experiment is 4608×4096 pixels, covering considerable water and shadow areas.

Figure 11. The first experiment data. (a) SAR image. (b) Ground truth.

In this experiment, we mainly test the performance of MSF-MLSAN in distinguishing water and shadow. Figure 12 shows the fusion map with SAR image and the classification results of different methods. Figure 12a depicts the ground truth of the SAR image, which is generated by SAR experts. The two colors of green and blue denote the shadow and water, respectively.

Figure 12. The fusion map of classification results by different methods. (a) Ground truth. (b) RefineNet. (c) DeepLabV3. (d) Multi-Level Selective Attention Network (MLSAN). (e) GLGCM-MLSAN. (f) Gabor-MLSAN. (g) MSOGDF-MLSAN. (h) MSF-MLSAN.

Figure 12b gives the classification result generated by RefineNet. Compared with the ground truth, we find there are several obvious errors, such as the areas of the red rectangles which are incorrectly classified between water and shadow areas, especially many shadow areas are classified as water. Figure 13a shows the confusion matrix of RefineNet, of which, W, S and B denote the types of water, shadow and background respectively. The overall accuracy for water, shadow and background are 0.865, 0.891 and 0.8512 and the IoU for them are 0.7166, 0.7524 and 0.8174 respectively.

Figure 12c illustrates the classification result of DeepLabV3. Compared with RefineNet, its performance of distinguishing water and shadow areas is a little worse than RefineNet. We can see many shadow areas are classified as water areas in the red boxes and many shadow areas are not detected in the pink boxes. Therefore, the accuracy and IoU for water and shadow are both reduced a lot according to Figure 13b, because of the false classification and miss detection.

Figure 12d presents the classification result of MLSAN. Compared with RefineNet, many false alarm areas for water have disappeared, which displays the satisfying performance of MLSAN. Figure 13c shows the confusion matrix of MLSAN. There is a small increase in water IoU due to the reduction of false alarms, despite the small decrease in water OA. In addition, the OA and IoU for the shadow and the background are both improved.

Figure 12e indicates the classification results of GLGCM-MLSAN. Compared with Figure 12d, some missed water areas are detected, which results in the increase in OA and IoU of the water showing in Figure 13d. However, there is a slightly decline in OA and IoU of the shadow and background due to misclassification.

Figure 12f gives the classification result of Gabor-MLSAN. Compared with Figure 12e, Figure 12f reduces a few false alarms for water though it still has some miss detections. Therefore, Gabor-MLSAN can produce better IoU though its OA is a little bit lower than GLGCM-MLSAN (shown in Figure 13e). In addition, Gabor-MLSAN brings higher OA and IoU for shadows than GLGCM-MLSAN, due to the reduction of misclassification of shadow areas. Furthermore, they basically have similar performance in background classification.

Figure 12g presents the classification results for MSOGDF-MLSAN. Compared with GLGCM-MLSAN and Gabor-MLSAN, it gains the worst performance in water classification, as the lowest OA and IoU shown in Figure 13f. It has a little worse performance for shadow classification than Gabor-MLSAN but a little better than GLGCM-MLSAN. However, it has the best classification performance for background than the other five methods.

Figure 13. The classification accuracy and IoU of different methods. (a) RefineNet. (b) DeepLabV3. (c) MLSAN. (d) GLGCM-MLSAN. (e) Gabor-MLSAN. (f) MSOGDF-MLSAN. (g) MSF-MLSAN. (W, S and B denote Water, Shadow and Background, respectively).

Through analyzing the classification performance of methods above, we find GLGCM-MLSAN and Gabor-MLSAN could achieve better classification results for water and shadow, while MLSAN and MSOGDF-MLSAN slightly outperform in the classification of background. Therefore, the weights of 0.1, 0.35, 0.45 and 0.1 are selected heuristically for the score maps of MLSAN, GLGCM-MLSAN, Gabor-MLSAN and MSOGDF-MLSAN respectively to generate the MSF-MLSAN score map. The classification results of MSF-MLSAN is given in Figure 12h. It generates better classification results than MSOGDF-MLSAN or Gabor-MLSAN. Figure 13g shows the OA and IoU of MSF-MLSAN, from which we can see the IoU for water is the highest. As shadow classification is concerned, the OA is slightly lower than Gabor-MLSAN but the IoU is the highest. For the background, it reaches the highest values of OA and IoU among all the seven methods. This evaluation proves the prominence performance of MSF-MLSAN for simultaneous classification of water and shadow.

To further investigate the classification accuracy, Figure 14 shows three maps for the three types of targets classified by different methods. Figure 14a is the classification accuracy for water. The DeepLabV3 achieves the lowest accuracy in OA and IoU and MSF-MLSAN generates the best classification result. For shadow classification illustrated in Figure 14b, DeepLabV3 also achieves relatively poor classification accuracy and MSF-MLSAN offers the highest accuracy in OA and IoU. Figure 14c indicates that RefineNet gets the lowest accuracy in background classification and MSF-MLSAN achieves the best performance.

Figure 14. Classification accuracy for the three types of targets by different methods for the first experiment. (a) Water. (b) Shadow. (c) Background.

3.4.2. The Experiment with Fewer Water but More Shadow Areas

Figure 15 shows the data of our second experiment, which includes the SAR image and the ground truth. The size of SAR image is 4608×3548 pixels, containing few water areas and considerable shadow areas. The mountain areas in this test scenario incur many shadow areas in the SAR image, which proves a good test to MSF-MLSAN in classifying water and shadow.

Figure 15. The second experiment data. (a) SAR image. (b) Ground truth.

The results of this experiment are shown in Figure 16. Figure 16a is the ground truth of this scene, generated by SAR experts. Figure 16b presents the classification results of RefineNet, which contain many false alarms for water detection highlighted in the red rectangles. That's the reason why the IoU of water classification is low (as shown in Figure 17a), though its OA is high. Its classification performance for shadow and background are good with IoU of 0.7935 and 0.783 respectively. Figure 16c illustrates the classification results of DeepLabV3. Compared with RefineNet, the classification performance for water and shadow are even worse. We can see many shadow areas are classified as water in red boxes and also many shadow areas are not detected in pink boxes. which causes the classification accuracy of water and shadow areas are reduced, as depicted in Figure 17b.

Figure 16. The fusion map of classification results by different methods. (a) Ground truth. (b) RefineNet. (c) DeepLabV3. (d) MLSAN. (e) GLGCM-MLSAN. (f) Gabor-MLSAN. (g) MSOGDF-MLSAN. (h) MSF-MLSAN.

Figure 17. The classification accuracy and IoU of different methods for the second experiment. (a) RefineNet. (b) DeepLabV3. (c) MLSAN. (d) GLGCM-MLSAN. (e) Gabor-MLSAN. (f) MSOGDF-MLSAN. (g) MSF-MLSAN. (W, S and B denote Water, Shadow and Background, respectively).

Figure 16d gives the classification results of MLSAN. Compared with Figure 19b, the considerable false alarms for water have been resolved except some red box areas. Hence, the water detection

performance has been significantly improved in MLSAN depicted as Figure 17c, achieving 0.6383 in IoU. In the classification of shadow, there are still false alarm areas such as the left red box region and some missed detection areas. However, the accuracy for shadow classification has also been improved to 0.925 in OA and 0.8121 in IoU. Because of the reduction in false alarms of background, the IoU is improved to 0.8181.

Figure 16e delineates the classification result of GLGCM-MLSAN. Compared with Figure 16d, it reduces the false alarm areas for water classification, though it still has small areas shown in the two little red boxes. The OA and IoU for water classification have been greatly improved to 0.8762 and 0.7869, respectively, as shown in Figure 17d. For shadow classification, the OA and IoU have also been increased to 0.9485 and 0.8189, though there are still some false detected areas highlighted by the large red box on the left and missed detected areas in the pink box. However, the OA and IoU for background classification have declined a little bit by 1.84% and 0.55% compared with Figure 16d.

Figure 16f shows the classification result of Gabor-MLSAN. Compared with GLGCM-MLSAN, there are slightly more false alarms for water, such as the red boxes, so the OA and IoU for water are reduced by 0.94% and 4.61% (refer to Figure 17e). While for shadow classification, the accuracy increases a little in IoU though there is a slightly reduction in OA compared with GLGCM-MLSAN. However, the classification performance for the background is worse than GLGCM-MLSAN, with a reduction of 1.57% in OA and 1.32% in IoU.

Figure 16g demonstrates the classification result of MSOGDF-MLSAN, in which we can see there are one large area where water detection fails as highlighted by the long red box and there are also several false alarms for water in the red boxes. The OA and IoU of MSOGDF-MLSAN is reduced to 0.6498 and 0.6246 shown in Figure 17f, slightly better than RefineNet. In addition, there are some false detected areas for shadow depicted in red boxes on the left and some missed detections in the pink box. So, the OA and IoU for shadow is slightly lower than MLSAN, GLGCM-MLSAN and Gabor-MLSAN. As background classification is concerned, it only gets better performance than Gabor-MLSAN and RefineNet.

Finally, Figure 16h is the classification result of the proposed method of MSF-MLSAN. It utilizes the same weights as the first experiment for the four score maps of MLSAN, GLGCM-MLSAN, Gabor-MLSAN and MSOGDF-MLSAN, which are 0.1, 0.35, 0.45 and 0.1, respectively. According to Figure 17g, we can see its classification results are better than the former five results in water and shadow with the IoU of 0.8085 and 0.8206 (as shown in Figure 17g). For the background, it is 0.8575 in OA and 0.8154 in IoU, which are slightly lower than MLSAN in OA and IoU but higher than the other methods.

Figure 18 illustrates the classification accuracy for the three types of targets produced by the seven methods, showing the OA and IoU of each target from Figure 18a–c. In Figure 18, we can see MSF-MLSAN can achieve the best classification performance for the three types of targets.

Figure 18. Classification accuracy for the four types of targets by different methods for the second experiment. (a) Water. (b) Shadow. (c) Background.

Figure 19 shows the Loss curve and IOU diagram of MSF-MLSAN network when training this dataset. In this project, 100 epochs are trained and the corresponding Loss and IOU values are obtained

after being verified every five epochs. From these two graphs, it can be found that Loss converges steadily and the verification accuracy tends to be stable during the training process.

Figure 19. The Loss curve and the IoU curve with epochs.

According to the classification results of the two experiments, the mean OA and IoU for each type of target as well as the overall IoU of the proposed framework (MSF-MLSAN) are computed and shown in Table 1.

Table 1. Mean Classification results of the two experiments.

	Types	Ground Truth			
		Water	Shadow	BG	IoU
Detected results	Water	0.8382	0.0018	0.0050	0.7727
	Shadow	0.0811	0.9278	0.0784	0.8006
	BG	0.0807	0.0704	0.9032	0.8588
Overall Accuracy		0.9076			

From Table 1, we can see the mean OAs for water, shadow and background are 0.8382, 0.9278 and 0.9032; while the IoUs are 0.7757, 0.8006 and 0.8588. The overall accuracy for the two experiments is 0.9076. The high OA and IoU of MSF-MLSAN have proven its success in the simultaneous classification of water and shadow. The complexity of context has been addressed by the combination of decoder (attention mechanism and pyramid pooling operation) and splicing.

4. Discussion

MSF-MLSAN, a new end-to-end framework based on deep learning is presented in the paper and prominent results are obtained for simultaneous water and shadow classification in SAR images. However, there are two extension to accommodate broader applications in the future for MSF-MLSAN: the weights optimization for feature fusion and the generalization of MSF-MLSAN.

4.1. Weights Optimization for Feature Fusion

In MSF-MLSAN, the weights for the four score maps are selected to build the final fused map, from which classification results are generated. The weights are determined according to their separate classification performance heuristically, following [11]. But this approach requires high-level expertise and is difficult to transfer to other fields. Therefore, an automatic method to determine optimal weight combination needs to be developed in the future. And we plan to employ the Ho-Kashyap algorithm depicted in Reference [34] to realize it.

4.2. Generalization

In this paper, the dataset comes from the millimeter wave SAR system. The center frequency is 35 GHz and the resolutions are 0.13 m in the range and 0.14 m in the azimuth. With the following two extensions, MSF-MLSAN can be applied to analyze SAR imagery datasets with different resolutions or different bands.

4.2.1. Resolution

Resolution is a key parameter in SAR systems. It is a measure of the ability of an imaging system to distinguish level of details and also a key indicator for measuring the performance of SAR imaging system. The characteristics of the same target could be different in SAR images with different resolutions. The higher the resolution, the richer the texture and detailed information the image would offer. The lower the resolution, the more homogeneity of targets will be. In addition, the spatial extents of the same target may vary significantly under different image resolutions, which will have a big impact on the classification. For example, if we test SAR images from TerraSAR (3 m resolution) directly with the model trained by the millimeter SAR dataset used in this paper, the classification results may not be as good as it is reported in this paper, due to the large difference of their resolutions.

To solve this problem, two methods could be employed. One is to sample a small portion of SAR images with the different resolutions, then re-train the sample dataset within MSF-MLSAN to generate a new model. Finally, SAR images of different resolutions can be analyzed with these new models respectively. The accuracy depends on the quality of the samples and the re-training process is time-consuming.

Another way is to use the interpolation or down-sampling techniques [35]. Then the accuracy of MSF-MLSAN is subject to the quality of the re-scaling process and we also notice the risk of incurring additional errors in the interpolation or potential information loss due to down-sampling operations [36].

4.2.2. Frequency Band

It is also an important problem to use the model trained by the millimeter wave system to test SAR images acquired with different bands (e.g., Sentinel-1 and TerraSAR). Different bands of SAR systems have been employed for different applications. If the wavelength of the microwave is longer (i.e., the frequency band is smaller), the penetration will become stronger, which is usually used to detect underneath targets (e.g., forest soil conditions). On the other hand, more detailed information of the targets can be acquired if the wavelength is shorter.

Transfer learning aims to apply knowledge or patterns learned in one field or task to different but related domains, which could serve as an appealing approach to address the band heterogeneity [37]. In summary, transfer learning allows MSF-MLSAN trained on a given band to be applied to SAR data with different frequency bands. Transfer learning within MSF-MLSAN will be an important part of our future work.

5. Conclusions

In the paper, a new end-to-end framework with deep learning to classify water and shadow in SAR imagery synthetically has been proposed, which demonstrates the success of integrating SAR domain knowledge and geospatial contextual information in deep learning for SAR image analytics. MSF-MLSAN could automatically classify these two highly similar objects simultaneously, to avoid the error-prone coordination of their respective classification results. MSF-MLSAN mainly contains three parts, namely the MSF, the MLSAN network and the IS. The MSF is to extract the low-level features of SAR images with GLGCM, Gabor and MSOGDF, as the integration of SAR domain knowledge with deep neural networks. The MLSAN network contains two parts, namely, the encoder and the decoder. Finally, the score maps are weighted to get the fusion score map to generate the final classification result within the IS.

In this paper, two experiments from millimeter wave SAR images are performed. One is the scene with more water areas and the other is the scene with more shadow areas. The classification results are compared with other six methods, including RefineNet, DeepLabV3, MLSAN, GLGCM-MLSAN, Gabor-MLSAN and MSOGDF-MLSAN. The proposed MSF-MLSAN method has the best classification results for water and shadow in mean OA of the two experiments, as 0.8382 and 0.9278; and the

mean IoUs of them are 0.7727 and 0.806. The mean overall IoU of MSF-MLSAN is 0.9076. Therefore, the proposed framework MSF-MLSAN achieves outstanding accuracy for simultaneous water and shadow classification in millimeter wave SAR images.

The success of MSF-MLSAN has proven the necessity of integrating SAR domain knowledge and geospatial analytics into the design of deep neural networks. Without such integration, MSF cannot capture appropriate SAR image features. On the other hand, geospatial contextual information of geographic objects has been handled via the selective attention pooling (the RAP module and the PAP module) and splicing in IS. The former encodes multiscale spatial structures within the given image tile, while the latter guarantees the right spatial coverage of given image tiles. In our future work, we will explore new deep learning techniques to solve scale gaps between SAR images analytics and daily object classification.

The proposed framework can also be applied to spaceborne SAR systems and InSAR systems for classification and object detection studies, such as the TanDEM-X system and the TerranSAR system. The weighting optimization and generalization strategies will play pivotal roles in extending MSF-MLSAN for these data analytics, as a focus of our future research. MSF-MLSAN paves the way for further integration of SAR domain knowledge and state-of-the-art deep learning techniques. We hope this paper will inspire more innovative works in this direction.

Author Contributions: L.C., J.X. and P.Z. proposed the new network, designed the experiments and produced the results; P.Z. made the SAR dataset; P.Z., L.C., Z.L., X.X. and Z.Y. contributed to the discussion of the results. L.C. and P.Z. drafted the manuscript. All authors contributed to the study, reviewed and approved the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been funded by National Natural Science Foundation of China (Refs: 41201468, 41701536, 61701047, 42074033 and 41941019), by the Fundamental Research Funds for the Central Universities, CHD (Ref: 300102260301/087 and 300102260404/087) and by Scientific Research Fund of Hunan Provincial Education Department (Refs: 16B004 and 18A148).

Acknowledgments: We would like to thank Beijing Institute of Radio Measurement, who provided the materials used for experiments.

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

Appendix A

Implementation Details of the Testing.

Algorithm A1: Testing process

Input: the input images are cut from a large-scale SAR image using a sliding window and the size of the sliding window is 512×512 with a step of 256. Then the different features with different scales are extracted by Gabor, GLGCM and MSOGDF. The inputs of the testing network include SAR images, corresponding Gabor features, GLGCM features and MSOGDF features.

Initialization: load the trained model and initialize the network.

Forward:

for iteration = 1:n (images) do

$$y = F_{net}(w, x)$$

where w, x, F and y denote the weight of the network, the input image, the output score map and the forward calculation of the entire network.

end for

Output: $y_{ori}, y_{Gabor}, y_{GLGCM}, y_{MSOGDF}$.

Postprocessing:

1. Weight and fuse the score maps.

$$y_{fuse} = w_1 \cdot y_{ori} + w_2 \cdot y_{Gabor} + w_3 \cdot y_{GLGCM} + w_4 \cdot y_{MSOGDF}$$

where w_1, w_2, w_3 and w_4 are the weights for the score maps and they satisfy:

$$w_1 + w_2 + w_3 + w_4 = 1.$$

2. merge all classification results of the small cut images to generate the final classification results of the large-scale image.

Remove 50 columns or 50 rows of pixels from the right side or the bottom of the current score map and remove 50 columns or 50 rows of pixels from the left side or top of the next neighboring small score map. Then weight and merge the overlapping 156 columns or 50 rows of pixels and the weights are both 0.5 for them.

3. Input the final score map matrix to the Softmax function to get the belief map.

$$b = \text{Softmax}(y_{fuse})$$

4. Output the maximum probability index of each pixel of the belief map, then combine it with the Colormap to generate the final classification results.

References

1. Ferretti, A.; MontiGuarnieri, A.; Prati, C.; Rocca, F.; Massonet, D. InSAR Principles-Guidelines for SAR Interferometry Processing and Interpretation. *J. Financ. Stabil.* **2007**, *19*, 156–162. [\[CrossRef\]](#)
2. Liao, H.; Wen, T. Extracting urban water bodies from high-resolution radar images: Measuring the urban surface morphology to control for radar's double-bounce effect. *Int. J. Appl. Earth Obs. Geoinf.* **2020**, *85*, 102003. [\[CrossRef\]](#)
3. D'Addabbo, A.; Refice, A.; Pasquariello, G.; Lovergine, F.P.; Capolongo, D.; Manfreda, S. A Bayesian Network for Flood Detection Combining SAR Imagery and Ancillary Data. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 3612–3615. [\[CrossRef\]](#)
4. Ding, X.; Li, X. Shoreline movement monitoring based on SAR images in Shanghai, China. *Int. J. Remote Sens.* **2014**, *35*, 3994–4008. [\[CrossRef\]](#)
5. Karvonen, J.; Simila, M.; Makynen, M. Open water detection from Baltic Sea ice Radarsat-1 SAR imagery. *IEEE Geosci. Remote Sens. Lett.* **2005**, *2*, 275–279. [\[CrossRef\]](#)
6. Martinis, S.; Kuenzer, C.; Wendleder, A.; Huth, J.; Twele, A.; Roth, A.; Dech, S. Comparing four operational SAR-based water and flood detection approaches. *Int. J. Remote Sens.* **2015**, *36*, 3519–3543. [\[CrossRef\]](#)
7. Notti, D.; Giordan, D.; Calo, F.; Pepe, A.; Zucca, F.; Galve, J.P. Potential and Limitations of Open Satellite Data for Flood Mapping. *Remote Sens.* **2018**, *10*, 1673. [\[CrossRef\]](#)
8. Behnamian, A.; Banks, S.; White, L.; Brisco, B.; Millard, K.; Pasher, J.; Chen, Z.; Duffe, J.; Bourgeau-Chavez, L.; Battaglia, M. Semi-Automated Surface Water Detection with Synthetic Aperture Radar Data: A Wetland Case Study. *Remote Sens.* **2017**, *9*, 1209. [\[CrossRef\]](#)
9. Huang, W.; DeVries, B.; Huang, C.W.; Lang, M.W.; Jones, J.F.; Creed, I.L.; Carroll, M. Automated Extraction of Surface Water Extent from Sentinel-1 Data. *IGARSS 2018*, *10*, 797. [\[CrossRef\]](#)
10. Zhang, P.; Chen, L.; Li, Z.; Xing, J.; Xing, X.; Yuan, Z. Automatic Extraction of Water and Shadow from SAR Images Based on a Multi-Resolution Dense Encoder and Decoder Network. *Sensors* **2019**, *19*, 3576. [\[CrossRef\]](#)
11. Chen, L.; Cui, X.; Li, Z.; Yuan, Z.; Xing, J.; Xing, X.; Jia, Z. A New Deep Learning Algorithm for SAR Scene Classification Based on Spatial Statistical Modeling and Features Re-Calibration. *Sensors* **2019**, *19*, 2479. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Hinton, G.E. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504–507. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [\[CrossRef\]](#)
14. Lenz, I.; Lee, H.; Saxena, A. Deep learning for detecting robotic grasps. *arxiv* **2013**, arXiv:1301.3592.
15. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [\[CrossRef\]](#)
16. Zhu, X.; Tuia, D.; Mou, L.; Xia, G.; 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*, 8–36. [\[CrossRef\]](#)
17. Hänsch, R.; Hellwich, O. Complex-valued convolutional neural networks for object detection in PolSAR data. In Proceedings of the 8th European Conference on Synthetic Aperture Radar, Aachen, Germany, 7–10 June 2010; pp. 1–4.
18. Song, Y.S.; Sohn, H.G.; Park, C.H. Efficient water area classification using Radarsat-1 SAR imagery in a high relief mountainous environment. *Photogramm. Eng. Remote Sens.* **2017**, *73*, 285–296. [\[CrossRef\]](#)
19. Ma, X.; Geng, J.; Wang, H. Hyperspectral image classification via contextual deep learning. *EURASIP J. Image Vide* **2015**, *1*, 2153–7003. [\[CrossRef\]](#)
20. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. *arxiv* **2016**, arXiv:1612.01105.
21. Lin, G.; Milan, A.; Shen, C.; Reid, I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. *arxiv* **2016**, arXiv:1611.06612.
22. Xing, J.; Sieber, R.; Kalacska, M. The challenges of image segmentation in big remotely sensed imagery data. *Ann. GIS* **2014**, *20*, 233–244. [\[CrossRef\]](#)
23. Chaudhari, S.; Polatkan, G.; Ramanath, R.; Mithal, V. An attentive survey of attention models. *arxiv* **2019**, arXiv:1904.02874.
24. Luong, M.T.; Pham, H.; Manning, C.D. Effective approaches to attention-based neural machine translation. *arxiv* **2015**, arXiv:1508.04025.

25. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; Tang, X. Residual attention network for image classification. *arxiv* **2017**, arXiv:1704.06904.
26. Hu, J.; Shen, L.; Sun, G.; Wu, E. Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [[CrossRef](#)]
27. Li, H.; Xiong, P.; An, J.; Wang, L. Pyramid attention network for semantic segmentation. *arxiv* **2018**, arXiv:1805.10180.
28. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [[CrossRef](#)]
29. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *40*, 834–848. [[CrossRef](#)]
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. *arxiv* **2015**, arXiv:1512.03385.
31. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks. *arxiv* **2014**, arXiv:1411.1792.
32. Xing, J.; Sieber, R. Sampling based image splitting in large scale distributed computing of earth observation data. *IGARSS* **2014**, 1409–1412. [[CrossRef](#)]
33. Chen, L.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arxiv* **2017**, arXiv:1706.05587.
34. Serpico, S.; Moser, G. Weight parameter optimization by the Ho–Kashyap algorithm in MRF models for supervised image classification. *IEEE Trans. Geosci Remote Sens* **2006**, *44*, 3695–3705. [[CrossRef](#)]
35. Stich, T.; Linz, C.; Albuquerque, G.; Magnor, M. View and Time Interpolation in Image Space. *Comput. Graph. Forum.* **2008**, *27*, 1781–1787. [[CrossRef](#)]
36. Xing, J.; Sieber, R.; Caelli, T. A scale-invariant change detection method for land use/cover change research. *Isprs. Photogramm.* **2018**, *141*, 252–264. [[CrossRef](#)]
37. Lu, J.; Vahid, B.; Hao, P.; Zuo, H.; Xue, S.; Zhang, G. Transfer learning using computational intelligence: A survey. *Knowl. Based Syst.* **2015**, *80*, 14–23. [[CrossRef](#)]

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).