remote sensing

MDPI

Article

Finer Resolution Mapping of Marine Aquaculture Areas Using WorldView-2 Imagery and a Hierarchical Cascade Convolutional Neural Network

Yongyong Fu [1], Ziran Ye [1], Jinsong Deng [1,*], Xinyu Zheng [2], Yibo Huang [1], Wu Yang [1], Yaohua Wang [3] and Ke Wang [1]

1 College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China
2 College of Information Engineering, Zhejiang A&F University, Hangzhou 311300, China
3 Zhejiang Key Laboratory of Exploitation and Preservation of Coastal Bio-resource, Zhejiang Mariculture Research Institute, Wenzhou 325005, China
* Correspondence: jsong_deng@zju.edu.cn; Tel.: +86-571-8898-2623

Received: 13 June 2019; Accepted: 12 July 2019; Published: 15 July 2019

Abstract: Marine aquaculture plays an important role in seafood supplement, economic development, and coastal ecosystem service provision. The precise delineation of marine aquaculture areas from high spatial resolution (HSR) imagery is vital for the sustainable development and management of coastal marine resources. However, various sizes and detailed structures of marine objects make it difficult for accurate mapping from HSR images by using conventional methods. Therefore, this study attempts to extract marine aquaculture areas by using an automatic labeling method based on the convolutional neural network (CNN), i.e., an end-to-end hierarchical cascade network (HCNet). Specifically, for marine objects of various sizes, we propose to improve the classification performance by utilizing multi-scale contextual information. Technically, based on the output of a CNN encoder, we employ atrous convolutions to capture multi-scale contextual information and aggregate them in a hierarchical cascade way. Meanwhile, for marine objects with detailed structures, we propose to refine the detailed information gradually by using a series of long-span connections with fine resolution features from the shallow layers. In addition, to decrease the semantic gaps between features in different levels, we propose to refine the feature space (i.e., channel and spatial dimensions) using an attention-based module. Experimental results show that our proposed HCNet can effectively identify and distinguish different kinds of marine aquaculture, with 98% of overall accuracy. It also achieves better classification performance compared with object-based support vector machine and state-of-the-art CNN-based methods, such as FCN-32s, U-Net, and DeeplabV2. Our developed method lays a solid foundation for the intelligent monitoring and management of coastal marine resources.

Keywords: marine aquaculture areas; WorldView-2 imagery; fully convolutional network (FCN); land-use and land-cover (LULC) mapping

1. Introduction

Marine aquaculture, the farming of aquatic organisms such as marine fish, shellfish, aquatic plants in the marine environment, provides great potential for the increasing demand of seafood production and the economic development in coastal areas [1–3]. Globally, the production of marine aquaculture has increased to 28.7 million tons in 2016, which doubles the almost 14.2 million tons in 2000 [4,5]. The rapid growth faces limitation in the availability of suitable land space and the environmental carrying capacity of land-based sites. Therefore, marine aquaculture, especially the widely used raft culture and cage culture areas that are mainly cultivated with marine plants and fish, respectively,

has been rapidly developed in inshore areas. However, extensive and disordered marine aquaculture might cause serious environmental problems and socio-economic losses [6–8]. Although the Chinese government has formulated a series of laws and regulations at local and national levels, such as the Marine Environmental Protection Law, overall marine functional zonation, and nature reserve schemes, it is still a big challenge for comprehensive coastal management in China. Thus, accurate mapping and monitoring of marine aquaculture are imperative for the management and sustainable development of coastal marine resources.

In facing of various spatial and temporal scales in a complex marine environment, remote sensing technology has substantially improved our ability to observe remote and vast areas at a fraction of the cost of traditional surveys [9]. To extract the marine aquaculture areas from remotely sensed images, previous studies have tried various methods including visual interpretation, spatial structure enhanced analyses [10,11], object-based image analysis (OBIA) [12–14], and deep convolutional neural networks (CNNs) [15]. Visual interpretation is used less because it is labor-intensive and time-consuming. Spatial structures enhancement analysis (such as texture and neighborhood characteristics analyses) is frequently used in pixel-based classification methods. OBIA has been widely used in the past few decades. It firstly segments the image and then performs classification based on these segments [16]. Thus, it can achieve a good classification performance by utilizing abundant features based on the representative segments.

In recent years, deep CNNs consisting of multiple trainable layers that can automatically learn representative and discriminative features [17,18] have achieved great success in the computer vision field [19,20]. In the remote sensing domain, deep CNNs have also been actively studied and shown obvious improvements on object detection [21] and scene classification [22]. Recent studies have further explored the ability of deep CNNs for dense prediction on the remotely sensed images. A straightforward method is to directly label a pixel by performing classification with its adjacent areas in a sliding-window way [23–25]. However, such methods have limited classification performance due to their fixed receptive field and huge time consumption [26]. Although some studies attempt to solve these problems by using the segment-based patches as basic classification union [27–29], they can be largely influenced by the segmentation accuracy. Besides, most of them are not trained end-to-end. To solve these problems, most recent studies have tried to perform pixel-wise classification exploiting fully convolutional networks (FCNs) [30], which replace fully connected layers with convolutional layers in classical CNN schemes. The main advantage of FCNs is that they allow pixel-wise labeling while the whole image as input.

However, there are some critical limitations for the FCNs to label the marine aquaculture areas in high spatial resolution (HSR) images accurately. The first challenge is the coexistence of confusing objects of various sizes, such as the large and continuous island areas versus a high diversity of small aquaculture areas in the sea areas. To tackle such problems, many researchers have concentrated on the use of multi-scale features, where objects at different scales can be prominent accordingly. One of the commonly used methods is to use multi-scale images as input to the deep CNNs [31–33]. However, such methods usually take more time because of the repetitive computing for multi-scale versions of the input images. On the other hand, some studies also try to aggregate multi-scale features, which are created by atrous convolution [15,34] or pooling operations [35,36] at multiple scales, or multi-kernel convolution [37]. However, as pooling with larger pooling sizes or convolution operation with larger atrous rates becomes less effective (i.e., more pooling or convolution operations would be applied to the padded zeros instead of the valid filter weights), such methods are limited to certain ranges of reception fields, resulting in a limitation for achieving better classification performance.

Meanwhile, due to the consecutive down-sampling processes in FCNs, the final feature maps are much smaller than the original image, leading to coarser prediction results and a decrease in classification accuracies. Therefore, it is a tough problem to perform accurate semantic labeling with such coarse feature maps, especially for marine objects with detailed structures in HSR images. To solve this problem, researchers have tried to restore the detailed spatial information by combining fine

resolution features from shallow layers, such as multi-level feature fusing [38–40], up-pooling or deconvolution with recorded pooling indices [41,42]. However, most current methods directly stack these multi-level feature maps, ignoring the adverse noises from the shallow layers. Meanwhile, some studies also attempted to refine the classification results by combining them with boundary detection results [43,44]. However, this requires extra modules and supervision for boundary detection.

In summary, although current FCN-based approaches have achieved great success in dense prediction, it is difficult to perform a fine mapping of the marine aquaculture areas fully exploiting the information in HSR images. First, most current approaches are less effective at acquiring multi-scale contextual information, making it difficult to detect various objects in the marine environment. Second, most of the existing strategies are less effective for the utilization of finer feature maps from shallow layers, making it difficult to restore the detailed structures of marine objects in HSR images.

To solve these problems, it is necessary to combine much effective multi-scale contextual information and fine resolution features from shallow layers. Inspired by this idea, we propose a novel model called the hierarchical cascade convolutional neural network (HCNet) to address the problems of fine mapping of marine aquaculture areas from HSR images. In addition, we also employ several attention-based modules throughout the network to refine the feature space. Finally, we compare our proposed HCNet with the conventional OBIA method and several state-of-the-art FCN-based methods. Both of them have been widely used and achieved great success in the classification of HSR images or nature images.

2. Study Area

A typical marine aquaculture area of 110 km^2 around Sandu Island was selected as our study area, which is located at Ningde City, Fujian Province, China (Figure 1). It is located in the subtropical monsoon climate zoon with the annual average precipitation of 1631 mm and a mean temperature of 14.7–19.8°C. As located in the semi-closed natural harbor, which helps weaken typhoons and accumulate nutrients in seawater, it has developed extensive marine aquaculture areas of various sizes that mainly include cage culture areas (CCA, see a1 and a2 in Figure 1) and raft culture areas (RCA, see b1 and b2 in Figure 1).

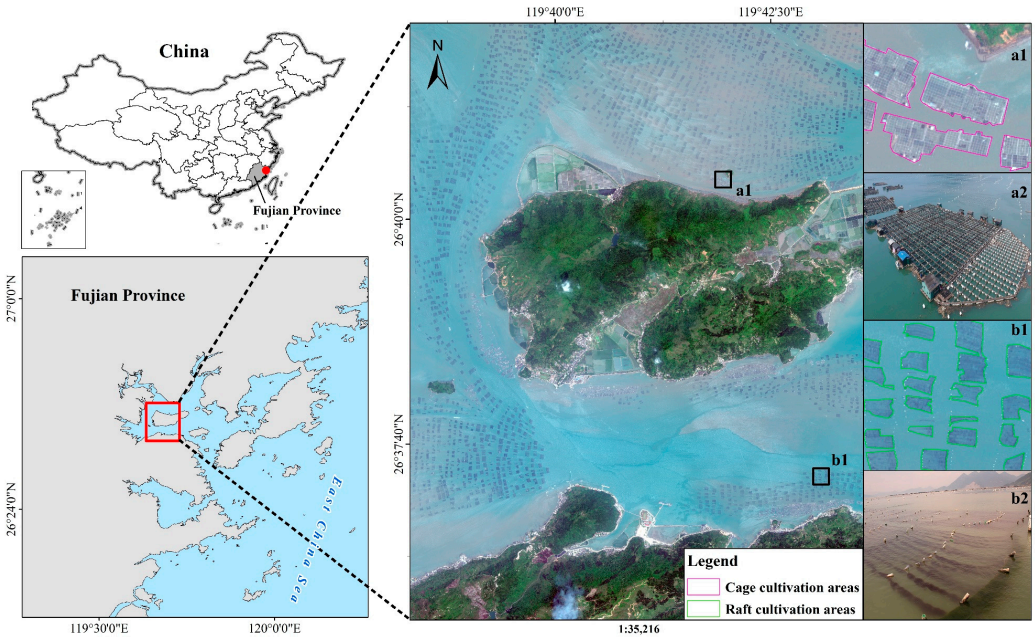

Figure 1. The study area Sandu Island is a typical marine aquaculture area in Ningde City, Fujian Province, China. The image here shows a Worldview-2 image of the study area in true color with image examples for cage culture areas (CCA) and raft culture areas (RCA) on the satellite (**a**1 and **b**1, respectively) and ground (**a**2 and **b**2, respectively).

The CCA are composed of accommodations and a large number of fish cages, which are constructed of plastic foam float and woodblocks. Since most of them are not standard productions from the factory, each of them has a complex and unique structure, making their extraction from HSR images difficult.

The RCA are generally cultivated with kelp or agar, which are widely distributed in the study area. The cultivated plants are usually twined on the belt that are linked to the fixed styrofoam floats. Therefore, the cultivated areas are mainly influenced by the density of plants and cultivated belts, making the RCA largely different from each other in HSR images. Meanwhile, as the cultivated belts are submerged in seawater, the features of RCA in HSR images may also be influenced by the unstable environment, such as waves or turbid seawater.

3. Materials and Methods

3.1. Data and Preprocessing

We selected a WorldView-2 (WV-2) image acquired on 20 May 2011 as the data source. The WV-2 image was selected in this study because of its high spatial resolution compared with other frequently used HSR imagery (e.g., IKNOS, SPOT-5, QuickBird, GaoFen-2). The satellite provides eight multispectral bands (MSS) with a spatial resolution of 2 m: coastal (400–450 nm), blue (450–510 nm), green (510–580 nm), yellow (585–625 nm), red (630–690 nm), red edge (705–745 nm), near infrared-1 (PAN, 770–895 nm), and near infrared-2 (860–1040 nm). It also provides a panchromatic band (450–800 nm) in sub-meter spatial resolution of 0.5 m [45].

As there is no cloud or haze in the whole aquaculture areas, we did not perform atmospheric correction in the preprocessing steps [46]. The MSS images and PAN image were firstly orthorectified into the Universal Transverse Mercator (UTM) projection system, and fused using Gram–Schmidt pan-sharpening method in ENVI (v5.3.1, Exelis Visual Information Solutions, Boulder, CO, USA, 2014). Eventually, we used the fused imagery consisting of eight bands with a spatial resolution of 0.5 m in the following classification process.

3.2. Hierarchical Convolutional Neural Network

As illustrated in Figure 2, the general workflow of the proposed HCNet mainly consists of three steps. Specifically, we first used a conventional CNN as an encoder to extract the high-dimensional feature maps based on the input imagery. Based on the output feature maps from the encoder, we used a hierarchical cascade structure to extract and aggregate the semantic information from local to global scale gradually. With the extracted multi-scale contextual information, we applied a coarser-to-fine strategy to restore the detailed information of marine objects in HSR imagery. In the following section, we will describe four important parts of the proposed HCNet, including (1) encoder based on VGG-Net [47]; (2) hierarchical contextual information aggregation; (3) detailed structure refinement; and (4) feature space refinement.

3.2.1. Encoder Based on VGG-Net

As illustrated in Figure 2, we first used the encoder network to transform the input imagery to high-dimensional abstract feature maps. To this aim, we employed the widely used VGG-16 network as the backbone of our proposed HCNet for its high performance. The VGG-16 network is structured with five blocks of convolutional layers followed by three fully connected layers. Detailed information about the model architecture can be found in [47]. To avoid the loss of spatial information and accelerate the training process, following similar encoder architecture to [38,41,42], we directly removed all the fully connected layers of the original model, which contain approximately 89% of the total 138 million parameters. As high-resolution feature maps are instrumental in the following process of our multi-scale context feature extractor, we avoided down-sampling after the last two max-pooling layers by setting these pooling layers with both stride and padding of one. As a result,

our encoder can obtain high-resolution feature maps, which are 1/8 of the input size instead of 1/32 in the original VGG-16 network.

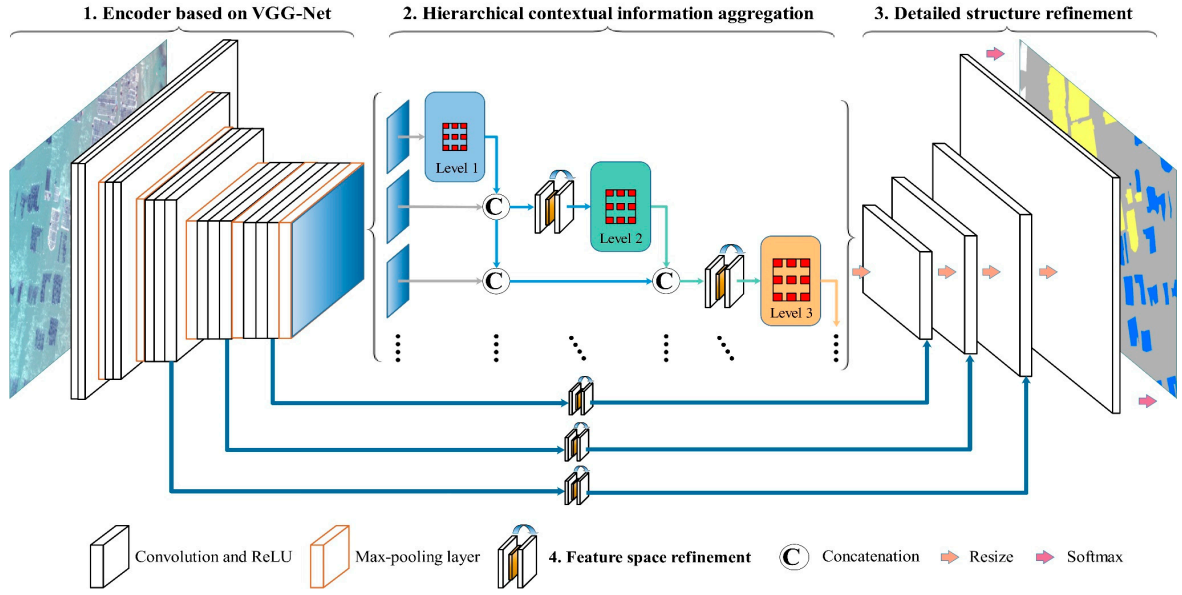

Figure 2. Overview of the proposed hierarchical cascade convolutional neural network (HCNet).

3.2.2. Hierarchical Contextual Information Aggregation

In CNN, extensive and powerful semantic information can be obtained by increasing the depth of the architecture, gaining from larger reception field and more non-linear operations [48]. However, semantic information captured from a single scale may lose the hierarchical dependence of the objects with their surrounding environment, leading to a decrease in the ability to recognize confusing objects of various sizes. Therefore, multi-scale semantic information, which can capture the relationships between the target objects and their surrounding environment, is important for the identification of confusing marine objects.

To obtain multi-scale feature maps at different reception fields, we applied atrous convolution [49] in this study. As shown in Figure 3a, an atrous kernel can increase its reception field without increasing additional parameters by dilating the kernel with zeros [34]. However, since the number of valid weights of feature maps decrease as the atrous rate increases, it is still difficult to obtain a larger reception field using a larger atrous rate with the current fusing strategy (e.g., direct concatenation, as shown in Figure 3b). For example, when applying a 3×3 kernel with an atrous rate that is close to the size of feature maps, only the central weight of the kernel is valid, which functions the same as the kernel with a size of one. To solve this problem, we developed a novel hierarchical cascade architecture, as illustrated in the middle part of Figure 2. By using a hierarchical cascade architecture, it was expected to enlarge the reception fields and increase the sampling rate while acquiring multi-scale contextual information. For instance, as shown in Figure 3a, the reception field of the original convolutional layer with an atrous rate of 4 is 9, with contributions from only three pixels. In a hierarchical cascade architecture, as the layer at a higher level calculates features based on feature maps from lower levels, the reception field at a higher level has increased to 13, as shown in Figure 3c. Meanwhile, the final calculation contributes from the information of seven pixels instead of the original three.

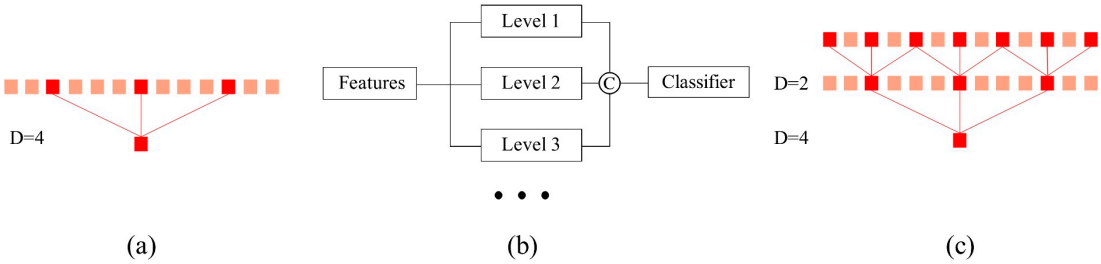

Figure 3. (**a**) One dimensional atrous convolution with an atrous rate of 4. 'D' represents the atrous rate. (**b**) An illustration of the commonly used direct concatenation strategy. (**c**) One dimensional atrous convolution with an atrous rate of 4 in the hierarchical cascade way.

Specifically, we acquired a series of feature maps with local to global contextual information by organizing the atrous convolutional layers in a hierarchical cascade fashion, where the atrous rates increase layer by layer (2, 4, 6, and 8 in our experiment). A layer with a smaller atrous rate was put in the upper part, while a layer with a larger atrous rate was put in the lower part. The outputs of each atrous convolutional layer were concatenated with the input feature maps and all the outputs from previous atrous convolutional layers. The concatenated feature maps were then fed into the following atrous convolutional layer. In this way, we can obtain increasingly larger reception fields in the following atrous convolutional layers. Meanwhile, each intermediate concatenated feature maps contain semantic information from different scales. Each atrous convolutional layer in the hierarchical structure can be formulated as:

$$F_1 = C_{k,D_1}[F_o], \tag{1}$$

$$F_l = C_{k,D_l}[\mathcal{L}(F_o C F_1 C F_2 C F_3 C \ldots C F_{l-1})],\ l > 1, \tag{2}$$

$$D_1 < D_2 < D_3 < \ldots < D_l, \tag{3}$$

where F_o represent feature maps from the output of our encoder network. $C_{k,D_l}[\cdot]$ represents an atrous convolution operation with kernel size k and atrous rate D at l-level. F_l ($l = 1, \ldots, n$) represent the feature maps at l-level in the hierarchical cascade structure. 'C' represents the concatenation operation. '$\mathcal{L}(\cdot)$' is the feature space refinement process, which is used to refine the fused multi-scale features and will be described in Section 3.2.4. D_l represents the atrous rate for capturing the corresponding feature maps at l-level.

3.2.3. Detailed Structure Refinement

Apart from the confusing marine objects of various sizes, the objects with fine structures in HSR images also increase the difficulty for accurate mapping. In fact, with increased down-sampling (i.e., "striding") and pooling operations, CNN causes a decrease in the size of the feature map. Taking the widely used VGG-Net [47] as an example, the last feature maps have only a size of 1/32 of the original image size. Thus, it is difficult to restore the detailed information in the original resolution, especially for the objects with detailed structures.

In CNN, it has been found that fine resolution feature maps from shallow layers can help restore the fine structures [38,50]. Based on such findings, we proposed to combine the low-level feature maps from the encoder for detailed structure refinement with a coarse-to-fine strategy. However, due to the existence of inherent semantic gaps between different-level feature maps, which presented as adverse noises from the shallow layers, directly stacking these feature maps might not be the best way to proceed. To solve this problem, we gradually concatenated the refined feature maps from shallow layers and the up-sampling feature maps from previous layers by using long-span connections.

After that, we fused them by using a convolution operation (which was 512, 512, and 256 kernels with a size of 3 × 3 for each operation in our experiments), as illustrated in Figure 2. It can be formulated as:

$$F_{\mathrm{r}} = C_{k,m}[\mathcal{L}(F_i\prime)\mathrm{C}\Upsilon(F_i)], \tag{4}$$

where F_i represent feature maps produced from the previous layers. F_i' represent the reutilized feature maps from corresponding shallow layers in the encoder network. $\prime\mathcal{L}(\cdot)\prime$ is the feature space refine process, which will be described in Section 3.2.4. $\prime\Upsilon(\cdot)\prime$ is the bilinear interpolation process. 'C' represents the concatenation operation. $C_{k,m}[\cdot]$ represents a convolution operation with m kernels and a size of k. F_{r} represent the generated feature maps.

3.2.4. Feature Space Refinement

To increase the feature representation and decrease the semantic gaps between different-level feature maps, we proposed to refine the feature space by using the attention mechanism: focusing on the important parts and suppressing adverse noise or unnecessary parts of the feature maps.

As shown in Figure 4, the proposed strategy for feature space refinement includes two aspects: channel and spatial refinement by using simple yet effective attention-based structures. Each of the single refining processes can be formulated as:

$$F\prime_{\mathrm{c}} = F \otimes \Phi_{\mathrm{c}}(F), \tag{5}$$

$$F\prime_{\mathrm{c\_s}} = F'_{\mathrm{c}} \otimes \Phi_{\mathrm{s}}(F'_{\mathrm{c}}), \tag{6}$$

where F represent the feature maps to be utilized from a shallow or previous layer. Φ_{c} is the channel attention module. $\prime\otimes\prime$ represents the element-wise multiplication. $F\prime_{\mathrm{c}}$ represent the channel refined feature maps. Φ_{s} is the spatial attention module. $F\prime_{\mathrm{c\_s}}$ represent the final channel and spatial refined feature maps. The following paragraphs describe the details of each attention module.

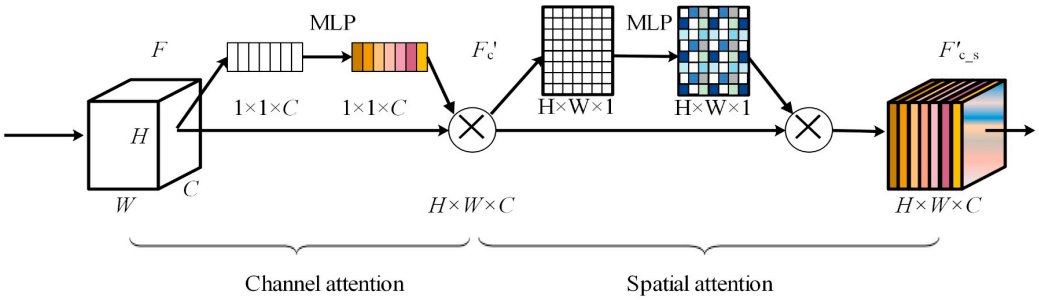

Figure 4. Overview of the channel and spatial attention mechanism used in this study. F represent the original feature maps. H, W, C represent the height, width and channel number of the feature maps, respectively. MLP represents the multi-layer perception with one hidden layer. F_c' represent the channel refined feature maps. $\prime\otimes\prime$ represents the element-wise multiplication. F'_{c_s} represent the final channel and spatial refined feature maps.

We produced the attention maps by exploiting the inter-channel or inter-spatial relationships of feature maps. To produce the channel attention map, we firstly aggregated global spatial information of a feature map by employing the global average pooling operation, generating a global spatial context descriptor. After that, the descriptor was fed into a multi-layer perceptron (MLP) with one hidden layer to produce the channel attention map. To control the capacity and computational cost, we reduced the size of the hidden layer to 1/r, where r is the reduction ratio (16 in our experiments). The process of acquiring channel attention can be formulated as:

$$\Phi_c(F) = \sigma(\mathrm{MLP}(\mathrm{C\_AvgPool}(F))) = \sigma(W_2^{1\times1\times C} \times \delta(W_1^{1\times1\times\frac{C}{r}} \times \mathrm{C\_AvgPool}(F))), \tag{7}$$

where C represents the channel number of the feature maps. W_1 and W_2 represent the weights of MLP layers with a size of $1 \times 1 \times \frac{C}{r}$ and $1 \times 1 \times C$, respectively. C_AvgPool$(\cdot)$ represents global average pooling operation on each channel of the feature maps. δ is the ReLU activation function. σ is the sigmoid function.

Differently from the channel attention map, the spatial attention map is expected to find the most informative region of the feature maps. To compute the spatial attention map, we first applied the global average pooling operation along the channel axis, generating a global channel context descriptor. After that, similar to the process for acquiring channel attention map, we fed the flattened descriptor to the MLP with one hidden layer at a reduced ratio of r (16 in our experiments). Finally, we reshaped the output to the two-dimensional spatial attention maps. The process of acquiring our spatial attention map can be formulated as:

$$\Phi_s(F) = \sigma(\mathrm{MLP}(\mathrm{S\_AvgPool}(F))) = \sigma(W_2^{H \times W \times 1} \times \delta(W_1^{\frac{H \times W}{r} \times 1} \times \mathrm{S\_AvgPool}(F))), \quad (8)$$

where H and W represent the height and width of the feature maps, respectively. W_1 and W_2 represent the weights of MLP layers with a size of $\frac{H \times W}{r} \times 1$ and $H \times W \times 1$, respectively. S_AvgPool$(\cdot)$ represents global average pooling operation along the channel axis of the feature maps.

3.3. Implementation Details

As shown in Figure 2, we employed the encoder, which is a variant of VGG-Net with 16-layers, to produce high-dimensional abstract features from input imagery. Based on the output of the encoder network, we captured the hierarchical contextual information by using a group of atrous convolution operations with the atrous rates of 2, 4, 6, and 8. Meanwhile, to avoid growing too wide and controlling the model's size, we used kernels with a size of 1×1 after each concatenation in the hierarchical cascade structure, making all channels of the concatenated feature maps reduce to 512, which is same as the output of the encoder. We also set 512 as the kernel numbers for all the atrous convolution layers to make weights for contextual information of all levels equal.

As for the detailed structure refinement, we only chose three layers in the encoder part for refinement as illustrated in Figure 2. The reasons are as follows: (1) although shallow layers carry much detailed information, those layers contain much noise that is adverse for restoring the detailed structures; (2) it is also hard to train the CNN well with more complex structures and parameters, especially with a typical small dataset in remote sensing. Besides, we chose the last convolution layer in each block before the pooling layers for refinement, because they contain much detailed information in these layers. We then used a 1×1 kernel after each concatenation to control the model's size, reducing the feature maps to a specific number (i.e., 512, 512, and 256, respectively), which is the same as the corresponding convolution layers in the encoder. Finally, a convolution layer with four 1×1 kernels was employed to predict the label maps, which were further up-sampled by a factor of eight and passed through softmax activation, where the categorical cross entropy is employed to measure the error between the predicted and actual values.

In the training process, 6141 patches with a size of 256×256 cropped from the pre-processed imagery were utilized as inputs of our proposed HCNet. The ground truth map of each patch was obtained by visual interpretation and corrected by ground survey (released at https://github.com/yyong-fu/HCNet). Among them, we randomly selected 70% of the dataset for training and the remaining 30% for testing.

In the experiments, we implemented the HCNet using the high-level application programming interface Keras (version 2.2.4) with tensorflow (version 1.8.0) as the computation backbone. All the algorithms were programmed using python 3.5.2. We trained the HCNet for 20 epochs using a batch size of four, and Adam optimization with a learning rate of 0.0001, β_1 of 0.9, and β_2 of 0.999. We carried all the experiments on a computer with a 4.20-GHz Intel(R) Core i7-7700K CPU, 16 GB of memory, and an NVIDIA GeForce GTX 1070 graphics processing unit (GPU).

3.4. Comparison Methods

3.4.1. Object-Based Support Vector Machine (SVM) Classification

To provide a comparison with our proposed approach, we compared HCNet with the widely used OBIA approach. Over the past decades, OBIA has achieved great success in the classification of remote sensing, especially for HSR images. Meanwhile, a wide range of remote sensing applications proved that the support vector machine (SVM) is an effective and reliable classifier [51]. Thus, as a typical and reliable method for classification of remote sensing images, the object-based SVM is a suitable method for our classification and comparison purposes.

The first important process is to obtain image objects via segmentation, which are the basic classification units. Here, we employed the widely used multi-resolution segmentation (MRS) algorithm, which is implemented in the eCognition software (version 9.0), to produce semantically meaningful image objects. Each of them was expected to be seen as a proper representation of an instance of some type of geo-object. Three key parameters control the segmentation process: scale parameter (SP), shape, and compactness. Instead of using the "trial and error" method, we employed the Estimation of Scale Parameter (ESP) 2 tool [52] for the selection of optimal SP. The ESP 2 tool iteratively segments the image with SPs increasing in a fixed step size, and calculates the local variance value, which is the mean standard deviation of the objects, for every step. Figure 5 shows the local variance values that are plotted against the corresponding scale parameters. Based on this figure, the local maximum points of the curve indicate the candidates of optimal SP. The graph shows that the scale of 112 represents the first sharp break after a continue decreasing. Thus, we set 112 as the optimal SP. We gave the weight of shape parameter less importance by assigning a value of 0.1, as the various shapes of CCA and RCA exist in the study area. We then assigned the weight of compactness value of 0.5 to treat them equally.

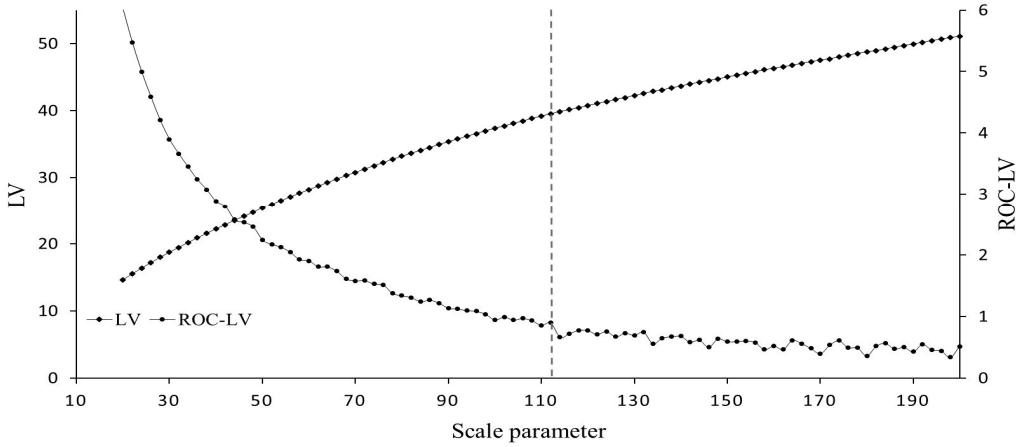

Figure 5. Scatter diagrams produced by the Estimation of Scale Parameter (ESP) 2 tool. The local variance (LV) and the rate of change of LV (ROC-LV) values are plotted against corresponding scale parameters. The gray vertical dotted line shows our selected optimal SP.

Once the semantically meaningful image objects were obtained, we constructed the initial feature space with 45 commonly used features, which consist of the typical spectral, geometric, and textural aspects of the segments (Table 1). Detailed information about these features can be found in [53].

Table 1. Object features used for image analysis with the object-based SVM method. GLCM: gray-level co-occurrence matrix. GLDV: gray-level difference vector.

Feature Type	Features
Spectral features	Brightness; Maximum difference; Mean layer i (i = 1, 2, 3, 4, 5, 6, 7, 8); Standard deviation layer i (i = 1, 2, 3, 4, 5, 6, 7, 8)
Geometry features	Area; Asymmetry; Border index; Border length; Compactness; Density; Elliptic fit; Length/width; Length; Main direction; Rectangular fit; Roundness; Shape index; Width; Volume
Textural features	GLCM ang.2nd moment; GLCM contrast; GLCM dissimilarity; GLCM entropy; GLCM homogeneity; GLCM mean; GLDV ang.2nd moment; GLCM correlation; GLDV contrast; GLDV entropy; GLDV mean; GLCM standard deviation

To select the most representative features from the initial feature space, we utilized a wrapper method which is implemented in the Weka software (v3.8, University of Waikato, New Zealand, 2016). The wrapper method evaluates attribute subsets by using a learning scheme. Cross-validation was employed to estimate the accuracy values for every subset of the attributes. Eventually, we selected 18 features for classification: spectral features (mean (bands 4, 5, 7, and 8), standard deviation (band 3, and 6), Max.diff.), geometrical features (border length, width, border index, roundness), and textural features with all directions (homogeneity, contrast, dissimilarity, entropy, mean, correlation calculated from gray-level co-occurrence matrix, and entropy calculated from gray-level difference vector). For the configuration of SVM classifier, we employed Radial Basis Function as the kernel function. We then used a simple grid search method to determine the optimal penalty factor and the gamma parameter based on LibSVM [54]. Finally, the optimal penalty factor and gamma parameter value were 1.6 and 0.14, respectively.

3.4.2. FCN-Based Methods

Because of the high performance in recent remote sensing applications, we also selected several state-of-the-art FCN-based methods for comparison. For the FCN-based models, we directly selected the FCN-32s [30], U-Net [38], and DeeplabV2 [34] for comparison. We selected these models because all these models are either VGG-16 Net or similar architectures-based networks, with long-span connections or multi-scale contextual aggregation strategies, which are very suitable to compare with our proposed structures. The FCN-32s is the first proposed FCN-based method, which does not use the multi-scale contextual information or any long-span connections. Therefore, it represents a baseline for all the FCN-based methods. The U-Net has a U-shaped structure containing an encoder on the left side and a decoder on the right side. The up-sampled features in the decoder are combined with symmetric high-resolution features from the encoder to enable precise localization and high classification performance. Unlike U-Net, the DeeplabV2 proposed to use atrous spatial pyramid pooling to capture objects and image context at multi-scales and then used a fully connected conditional random field to improve the localization and classification performance. Detailed information about these model architectures can be found in [30,34,38].

We selected the same patches employed in our proposed method for training or testing these deep models. In the training phase, we modified the number of outputs to four for all these models. After that, we trained all these models from scratch. The training parameters and strategies adopted for these models are the same as ours.

3.5. Accuracy Assessment and Comparison

In this study, we compared our proposed HCNet with the widely used object-based SVM method and several FCN-based models. We conducted accuracy assessments on the final classification results of the testing dataset, with totally 30% of the whole study area. To construct the error matrix, we confirmed whether these pixels were correctly labeled by visual interpretation. Finally, we calculated the accuracy statistics based on the error matrix, including producer accuracy (PA), user accuracy (UA), overall accuracy (OA), and kappa coefficient.

To quantitatively assess the classification performance of our proposed method and other methods, two commonly used overall accuracy metrics, including F1 score (F1) and intersection over union (IoU), were calculated. F1 is calculated as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{9}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{10}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

where TP, FP, and FN refer to true positives, false positives, and false negatives, respectively.

IoU is calculated as:

$$\text{IoU} = \frac{\left|A_p \cap A_{GT}\right|}{\left|A_p \cup A_{GT}\right|} \tag{12}$$

where A_p is the set of predicted pixels. A_{GT} is the set of ground truth pixels. '$\cup$'and '$\cap$' represent the union and intersection operation, respectively. $|\cdot|$ represents the number of pixels in the set.

To evaluate the classification performance of RCA and CCA using different methods, we calculated these accuracy metrics for each category. In addition, we used the mean overall accuracy metrics of RCA and CCA to evaluate the average performance of different methods.

4. Results and Comparison

4.1. Classification Results and Accuracy Assessment

The final classification results using the proposed HCNet are shown in Figure 6. After a visual inspection on the final classification results, most of the RCA and CCA were identified successfully. We also noticed that some ponds in the land area are misclassified as sea area. This is because all of them are complete seawater within an image patch for its limited size.

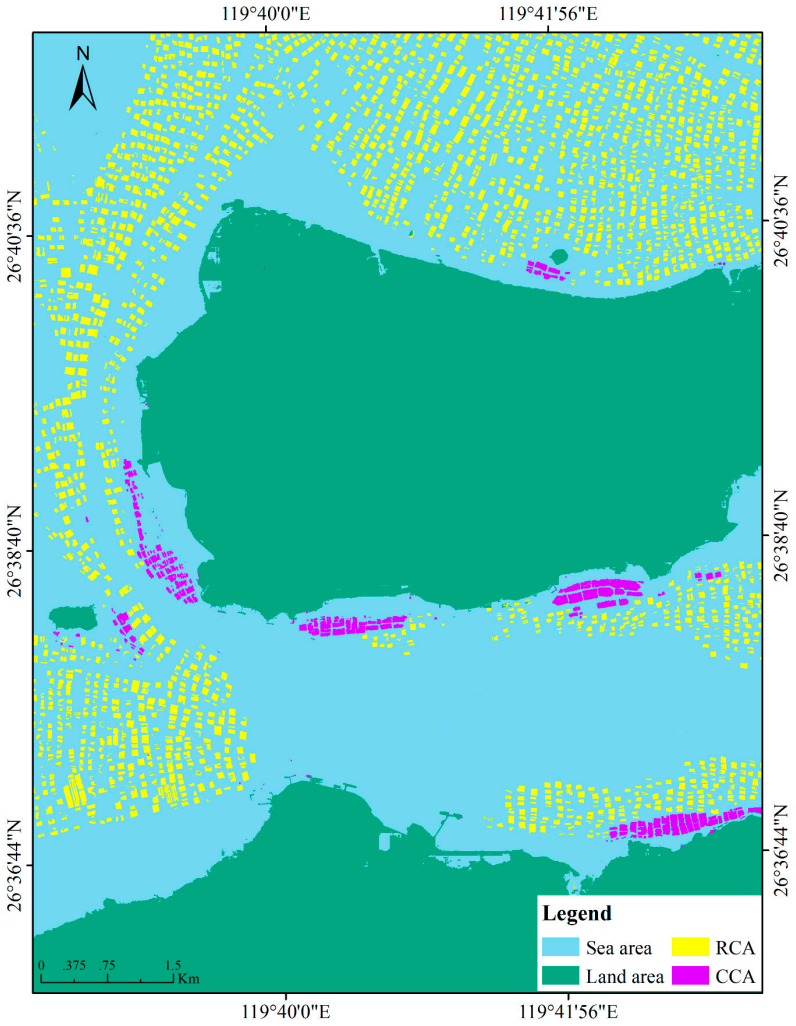

Figure 6. Classification results of CCA and RCA using our proposed method.

To quantitatively assess the classification performance, we used a testing dataset with over 1200 randomly selected patches for accuracy assessment, which accounts for 30% of the whole area. Table 2 shows the confusion matrix of the classification results. We find that the sea area and land area have the best classification performance, with over 98% of PA and UA values. The RCA and CCA have relatively UA values of 95.1% and 96.4%, respectively, indicating that over 95% of the classified CCA and RCA are indeed CCA and RCA, respectively. The CCA also have a relatively high PA value of 96.5%, indicating that over 95% of the CCA in the imagery are correctly labeled. Thus, the CCA and RCA are classified successfully, with OA greater than 95%, and a high kappa coefficient value of 0.97.

Table 2. Confusion matrix for the final classification results.

Predicted Class	Ground Truth					
	Sea Area	**Land Area**	**RCA**	**CCA**	**Sum**	**UA:**
Sea area	38394007	350996	355428	34576	39135007	98.1%
Land area	240583	34755462	9996	2922	35008963	99.3%
RCA	256058	3131	5009423	0	5268612	95.1%
CCA	33990	3706	335	1027595	1065626	96.4%
Sum	38924638	35113295	5375182	1065093		
PA:	98.6%	99.0%	93.2%	96.5%		
Overall accuracy:	98.4%					
Kappa coefficient:	0.97					

4.2. Accuracy Comparison

In this study, we compared our proposed approach with the object-based SVM and several state-of-the-art FCN-based methods. Table 3 shows the experiments setup and time complexity of different classification schemes. The time complexity was obtained by averaging the time to perform classification on the testing dataset, which contains over 1200 images with a size of 256 × 256 pixels. As is shown in Table 3, the OB-SVM spends the longest time for inference compared with other methods. With the acceleration of GPU, our proposed method and other FCN-based methods take less time. Furthermore, our proposed HCNet takes the least time for inference. This is mainly because we reduced the trainable parameters in our model: (1) we removed the last three fully connected layers in the original VGG-16 architecture in our encoder; (2) we used the 1 × 1 convolution operations to control the model size; (3) we used the bilinear interpolation instead of deconvolution for up-sampling operations.

Table 3. Experiments setup and computational complexity using different classification schemes. OB-SVM: object-based SVM classification method. Ours-HCNet: our proposed HCNet method. GPU: NVIDIA GeForce GTX 1070 graphics processing units. CPU: inter i7 7700k with 16 Gb memory.

Experimental Details	OB-SVM	FCN-32s	U-Net	DeeplabV2	Ours-HCNet
learning rate	-	0.0001	0.0001	0.0001	0.0001
batch size	-	4	4	4	4
platform	CPU	GPU	GPU	GPU	GPU
time (ms)	184.4	80.0	30.1	35.8	28.5

To provide a quantitative assessment for the performance of different methods, several commonly used accuracy metrics, including F1 and IoU were calculated on the testing dataset for the CCA and RCA, respectively (Table 4). The mean F1 and IoU values of CCA and RCA were also calculated to assess the global performance. As shown in Table 4, approaches using U-Net and DeeplabV2 achieve a similar accuracy level, with a mean IoU value of approximately 88%. The object-based SVM achieves the lowest accuracy values, with only a mean IoU value of approximately 80%. Our proposed method achieves the best performance, with the highest mean F1 value of 95.29%, and the highest mean IoU value of 91.03%.

Table 4. Quantitative comparison between our method and other methods at the pixel level, where the best values in bold and the second-best values are underlined.

Methods	RCA		CCA		Mean	
	F1	IoU	F1	IoU	Mean F1	Mean IoU
OB-SVM	87.43%	77.67%	90.70%	82.98%	89.07%	80.33%
FCN-32s	89.76%	81.42%	92.15%	85.44%	90.95%	83.43%
U-Net	92.12%	85.39%	95.49%	91.38%	93.81%	88.38%
DeeplabV2	92.75%	86.84%	94.87%	90.24%	93.91%	88.54%
Ours-HCNet	**94.13%**	**88.91%**	**96.46%**	**93.15%**	**95.29%**	**91.03%**

5. Discussion

5.1. OBIA vs. Our Approach

In this study, we first compared our proposed HCNet with the object-based SVM method. The object-based methods have been widely used for classification in the past few years, especially for the HSR images. Differently from the traditional pixel-based methods, object-based methods use segments of an imagery as basic units for classification. Classification based on segments has a lot of benefits, including a decrease on spectral variability and an increase on spatial and contextual information such as geometrical features [55]. Thus, uniform spectral character and abundant features of the image objects increase the classification accuracy and eliminate salt-and-pepper noise. However, when the spectral and geometric features are similar, it is hard for the segmentation algorithms to obtain high-quality image objects. Besides, it also takes extra time and computation. In the classification phase, it is also hard to design and choose discriminative features as input of the classifier. Thus, both of the uncertainties in the necessary procedures limit the classification performance.

To overcome such limitations, our proposed methods mainly contributes in two aspects. First, differently from the standard procedure of "segmentation and then classification" in OBIA, our method is implemented in an "end-to-end" way, which can avoid the segmentation error and is more efficient for large-scale HSR image classification. Second, the two methods are different in the feature design process. The feature space employed in the OBIA generally consists of handcrafted features, which are designed based on statistical analyses of a local area in the HSR imagery. There remains an inherent tradeoff for these handcrafted features between high discrimination performance and robustness [56]. In contrast, our proposed HCNet can automatically learn multi-level semantic information from local to scene scales. Thus, our proposed method can achieve a better classification performance, with a nearly 10% improvement in terms of mean IoU at the pixel level.

5.2. Conventional FCN-Based Methods vs. Our Approach

FCN, which is a fully convolutional version of CNN, has become the most state-of-the-art dense classification method in recent years [26,57,58]. However, there are mainly two problems for conventional FCN-based methods to precisely identify the boundaries of marine aquaculture areas. First, it is difficult to capture semantic information of confusing marine objects of various sizes through a single and fixed reception field. In addition, consecutive pooling operations largely reduce the resolution of the final feature maps. Thus, it is difficult for the predicted results to restore the original resolution as input imagery by learning. As shown in Figure 7, the predicted boundaries of CCA and RCA from the FCN-32s model have been largely smoothed. Meanwhile, some small objects are also misclassified or neglected by the FCN-32s model.

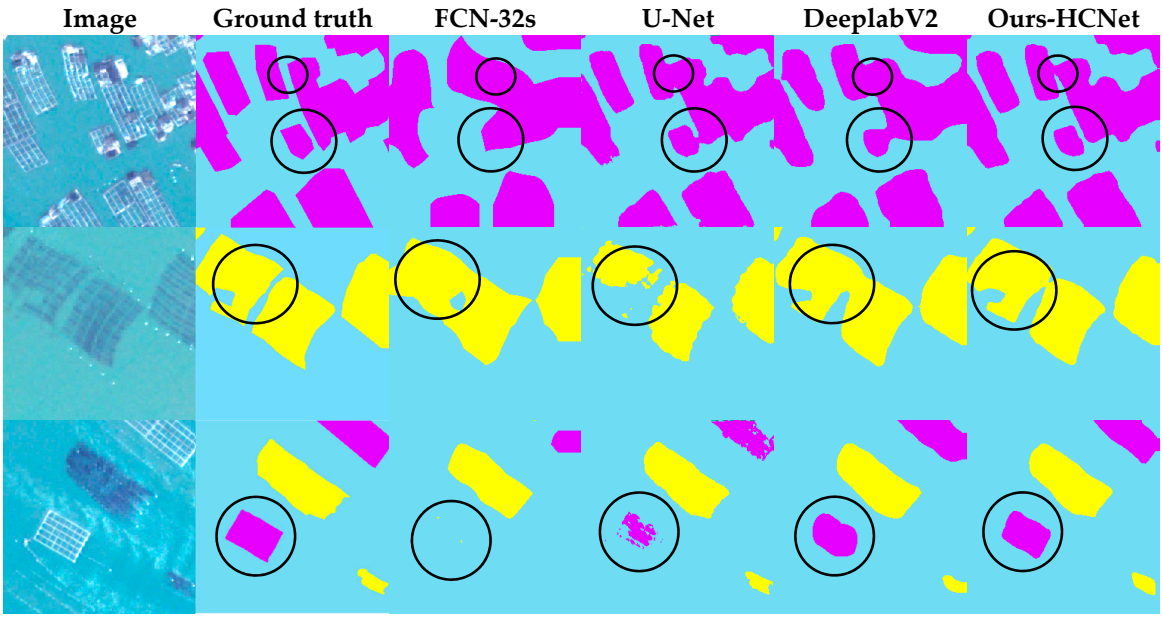

Figure 7. The classification results of CCA and RCA by using our proposed method and other comparison methods. The black circles indicate that our proposed HCNet obtains the best performance.

Although some studies, i.e., U-Net, DeeplabV2, try to improve the classification results by incorporating information from shallow layers or feature pyramid, it is still hard for them to identify objects at different scales while retaining the detailed information. In our study, we fully combined the information from shallow layers and feature pyramid to improve the classification results. In addition, we also enlarged the reception field for feature maps from the feature pyramid and increased the representation of feature maps from shallow layers, which is helpful for the prediction. Thus, as shown in Figure 7, our methods significantly improved the classification performance, with an improvement in the mean IoU value of nearly three percentage points.

5.3. Ablation Analysis

To explore the benefits brought by different proposed structures, we conducted ablation experiments on our proposed HCNet. We used the simplest encoder based model, which is mainly composed of the encoder (see Figure 2) followed by an up-sampling rate of eight and the classification layer, as the baseline method. Table 5 shows the accuracy assessment results for variants of HCNet by adding different structures gradually. As can be seen, the classification performance of each category improves by adding our proposed structures. As shown in Table 5, multi-scale contextual information fused in a parallel stack way can only improve the classification performance slightly. In contrast, our proposed hierarchical cascade structure can substantially improve classification performance, with an improvement in the mean IoU value of nearly 2.4 percentage points. Moreover, when applying with the detailed structure refinement and feature space refinement strategies, the classification performance improves even further.

Table 5. Quantitative comparison for the ablation experiments on our proposed HCNet. 'Mul' represents aggregating the multi-scale information in the commonly used direct stacking way. 'Mul+HCI' represents aggregating the multi-scale information in our proposed hierarchical cascade way. 'Mul+HCI+DIR' represents aggregating the multi-scale information by using our proposed hierarchical cascade method and adding the detailed structure refinement strategy. 'Mul+HCI+DIR+FSR' represents aggregating the multi-scale information by using our proposed hierarchical cascade method and adding the detailed structure refinement and feature space refinement strategies, as shown in Figure 2.

Methods	RCA		CCA		Mean	
	F1	IoU	F1	IoU	Mean F1	Mean IoU
Baseline	91.67%	84.62%	94.01%	88.70%	92.84%	86.66%
+Mul	93.11%	87.10%	93.10%	87.10%	93.11%	87.10%
+Mul+HCI	93.01%	86.93%	95.33%	91.08%	94.17%	89.00%
+Mul+HCI+DSR	93.61%	87.98%	96.45%	93.14%	95.03%	90.56%
+Mul+HCI+DSR+FSR	94.13%	88.91%	96.46%	93.15%	95.29%	91.03%

5.4. Potential Applications and Limitations

There are four carefully designed structures in our proposed HCNet, which mainly include the encoder, hierarchical cascade architecture, long-span connections, and the attention-based module. The encoder can be employed in most present CNN architectures for its fast convergence, and reduced consumption of memory. The combination of the hierarchical cascade architecture and long-span connections is helpful in capturing a large contextual information while maintaining the detailed information. Thus, it would be helpful for the classification of objects or geographical landscapes with complex components from HSR imagery, such as buildings [59,60], urban function zones [61], and fashions of rural settlements [62]. In addition, some analyses of natural imagery may also benefit from such structures, such as the identification of urban street scenes [63,64], agricultural trees [65], cells [66,67], and bacteria [68,69]. In addition, as the attention-based module is very helpful for the neural networks to find the most representative parts from abundant features, it can also be helpful for the feature space refinement in high spectral resolution image applications [70].

Meanwhile, there are several limitations of our proposed HCNet. First, although the HCNet can successfully identify marine aquaculture areas from HSR imagery with a spatial resolution of 0.5 m, it is relatively time consuming for the HCNet to perform classification on all the split patches, because some patches may not cover the targets. Thus, further research on the detection of the existence of targets before classification may also be helpful. Second, future studies may try to accelerate the training and inference process by using a series of model compressing methods, such as parameter pruning and sharing [71], low-rank factorization [72], network quantization [73], and knowledge distillation [74]. Third, our proposed method only applies to marine aquaculture areas covering the water surface. However, there are still a few submersible cages in some aquaculture areas, such as Shandong Province in the northeast of China.

6. Conclusions

In this study, we proposed a novel end-to-end hierarchical cascade neural network to identify and discriminate different types of marine aquaculture areas from HSR imagery. Our proposed HCNet achieves a high classification performance by focusing on three aspects: (1) a hierarchical cascade structure has been employed to capture multi-scale contextual information by enlarging the reception field, which is helpful to identify confusing objects of various sizes; (2) a coarse-to-fine refinement strategy is proposed to refine the target objects gradually, which is helpful for restoring the detailed information for marine objects with detailed structures; and (3) an attention-based module is proposed to refine the feature space, including both the channel and spatial dimensions.

Experimental results show that our proposed HCNet successfully identified the CCA and RCA, with OA greater than 95%, and a high kappa coefficient value of 0.97. Compared with the

conventional OBIA and the state-of-the-art FCN-based methods, our proposed HCNet achieves significant improvements on both visual and quantitative performances. In addition, our proposed method also has less time complexity than comparable methods.

Future studies may focus on testing our method on discriminating other types of confusing land cover and land use with detailed structures. Meanwhile, to speed up the process of mapping aquaculture areas from HSR images and enhance its applicability, researchers may focus on finding an approach to apply image segmentation preprocessing for interesting areas and accelerate the deep model. Additionally, as the training process of deep models needs a lot of precious manually labeled ground truth, it is necessary to investigate a method to train the models in less supervised way.

Author Contributions: Funding acquisition, J.D. and W.Y.; Methodology, Y.F.; Supervision, K.W.; Validation, Y.W.; Visualization, X.Z. and K.W.; Writing—original draft, Y.F.; Writing—review and editing, Z.Y., J.D., X.Z., Y.H. and W.Y.

Funding: Funding for this work was provided by Zhejiang Provincial Natural Science Foundation (LY18G030006), Ministry of Science and Technology of China (2016YFC0503404), and Programs of Science and Technology Department of Zhejiang Province (2018F10016).

Acknowledgments: We would like to express our appreciation to the editors and anonymous reviewers for their valuable comments, which greatly improved the paper's quality.

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

References

1. Gentry, R.R.; Froehlich, H.E.; Grimm, D.; Kareiva, P.; Parke, M.; Rust, M.; Gaines, S.D.; Halpern, B.S. Mapping the global potential for marine aquaculture. *Nat. Ecol. Evol.* **2017**, *1*, 1317–1324. [CrossRef] [PubMed]
2. Campbell, B.; Pauly, D. Mariculture: A global analysis of production trends since 1950. *Mar. Policy* **2013**, *39*, 94–100. [CrossRef]
3. Burbridge, P.; Hendrick, V.; Roth, E.; Rosenthal, H. Rosenthal Social and economic policy issues relevant to marine aquaculture. *J. Appl. Ichthyol.* **2001**, *17*, 194–206. [CrossRef]
4. FAO. *The State of World Fisheries and Aquaculture*; FAO: Rome, Italy, 2004; ISBN 9251051771.
5. FAO. *The State of World Fisheries and Aquaculture*; FAO: Rome, Italy, 2018; ISBN 9789251305621.
6. Grigorakis, K.; Rigos, G. Aquaculture effects on environmental and public welfare—The case of Mediterranean mariculture. *Chemosphere* **2011**, *85*, 899–919. [CrossRef] [PubMed]
7. Cao, L.; Wang, W.; Yang, Y.; Yang, C.; Yuan, Z.; Xiong, S.; Diana, J. Environmental impact of aquaculture and countermeasures to aquaculture pollution in China. *Environ. Sci. Pollut. Res.* **2007**, *14*, 452–462.
8. Tovar, A.; Moreno, C.; Mánuel-Vez, M.P.; García-Vargas, M. Environmental impacts of intensive aquaculture in marine waters. *Water Res.* **2000**, *34*, 334–342. [CrossRef]
9. Lillesand, T.; Kiefer, R.W.; Chipman, J. *Remote Sensing and Image Interpretation*, 5th ed.; John Wiley & Sons: Hobokan, NJ, USA, 2004; ISBN 0471152277.
10. Fan, J.; Chu, J.; Geng, J.; Zhang, F. Floating raft aquaculture information automatic extraction based on high resolution SAR images. In Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015; pp. 3898–3901.
11. Lu, Y.; Li, Q.; Du, X.; Wang, H.; Liu, J. A Method of Coastal Aquaculture Area Automatic Extraction with High Spatial Resolution Images. *Remote Sens. Technol. Appl.* **2015**, *30*, 486–494. [CrossRef]
12. Zheng, Y.; Wu, J.; Wang, A.; Chen, J. Object-and pixel-based classifications of macroalgae farming area with high spatial resolution imagery. *Geocarto Int.* **2017**, *33*, 1048–1063. [CrossRef]
13. Fu, Y.; Deng, J.; Ye, Z.; Gan, M.; Wang, K.; Wu, J.; Yang, W.; Xiao, G. Coastal aquaculture mapping from very high spatial resolution imagery by combining object-based neighbor features. *Sustainability* **2019**, *11*, 637. [CrossRef]
14. Wang, M.; Cui, Q.; Wang, J.; Ming, D.; Lv, G. Raft cultivation area extraction from high resolution remote sensing imagery by fusing multi-scale region-line primitive association features. *ISPRS J. Photogramm. Remote Sens.* **2017**, *123*, 104–113. [CrossRef]
15. Shi, T.; Xu, Q.; Zou, Z.; Shi, Z. Automatic Raft Labeling for Remote Sensing Images via Dual-Scale Homogeneous Convolutional Neural Network. *Remote Sens.* **2018**, *10*, 1130. [CrossRef]

16. Blaschke, T.; Hay, G.J.; Kelly, M.; Lang, S.; Hofmann, P.; Addink, E.; Queiroz Feitosa, R.; van der Meer, F.; van der Werff, H.; van Coillie, F.; et al. Geographic Object-Based Image Analysis—Towards a new paradigm. *ISPRS J. Photogramm. Remote Sens.* **2014**, *87*, 180–191. [CrossRef] [PubMed]
17. Farabet, C.; Couprie, C.; Najman, L.; Lecun, Y. Learning hierarchical features for scene labeling. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1915–1929. [CrossRef] [PubMed]
18. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
19. Arel, I.; Rose, D.; Karnowski, T. Deep machine learning-A new frontier in artificial intelligence research. *IEEE Comput. Intell. Mag.* **2010**, *5*, 13–18. [CrossRef]
20. Schmidhuber, J. Deep Learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef]
21. Dong, Z.; Wu, Y.; Pei, M.; Jia, Y. Vehicle Type Classification Using a Semisupervised Convolutional Neural Network. *IEEE Trans. Intell. Transp. Syst.* **2015**, *16*, 2247–2256. [CrossRef]
22. Hu, F.; Xia, G.S.; Hu, J.; Zhang, L. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. *Remote Sens.* **2015**, *7*, 14680–14707. [CrossRef]
23. Sharma, A.; Liu, X.; Yang, X.; Shi, D. A patch-based convolutional neural network for remote sensing image classification. *Neural Netw.* **2017**, *95*, 19–28. [CrossRef]
24. Santara, A.; Mani, K.; Hatwar, P.; Singh, A.; Garg, A.; Padia, K.; Mitra, P. Bass net: Band-adaptive spectral-spatial feature learning neural network for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 5293–5301. [CrossRef]
25. Lagrange, A.; Le Saux, B.; Beaupere, A.; Boulch, A.; Chan-Hon-Tong, A.; Herbin, S.; Randrianarivo, H.; Ferecatu, M. Benchmarking classification of earth-observation data: From learning explicit features to convolutional networks. In Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015; pp. 4173–4176.
26. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Convolutional Neural Networks for Large-Scale Remote-Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 645–657. [CrossRef]
27. Audebert, N.; Le Saux, B.; Lefèvre, S. How Useful is Region-based Classification of Remote Sensing Images in a Deep Learning Framework? In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 5091–5094.
28. 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. [CrossRef]
29. Fu, Y.; Liu, K.; Shen, Z.; Deng, J.; Gan, M.; Liu, X.; Lu, D.; Wang, K. Mapping Impervious Surfaces in Town-Rural Transition Belts Using China's GF-2 Imagery and Object-Based Deep CNNs. *Remote Sens.* **2019**, *11*, 280. [CrossRef]
30. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
31. Eigen, D.; Fergus, R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 2650–2658.
32. Zhao, W.; Du, S. Learning multiscale and deep representations for classifying remotely sensed imagery. *ISPRS J. Photogramm. Remote Sens.* **2016**, *113*, 155–165. [CrossRef]
33. Liu, Y.; Zhong, Y.; Fei, F.; Zhang, L. Scene semantic classification based on random-scale stretched convolutional neural network for high-spatial resolution remote sensing imagery. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 763–766.
34. 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.* **2018**, *40*, 834–848. [CrossRef] [PubMed]
35. 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*, 1904–1916. [CrossRef] [PubMed]
36. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6230–6239.

37. Audebert, N.; Le Saux, B.; Lefèvre, S. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. In Proceedings of the Asian Conference on Computer Vision (ACCV16), Taipei, Taiwan, 20–24 November 2016; pp. 180–196.
38. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015), Munich, Germany, 5–9 October 2015; pp. 234–241.
39. Hariharan, B.; Arbeláez, P.; Girshick, R.; Malik, J. Hypercolumns for object segmentation and fine-grained localization. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 447–456.
40. Pinheiro, P.O.; Lin, T.Y.; Collobert, R.; Dollár, P. Learning to refine object segments. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 75–91.
41. Noh, H.; Hong, S.; Han, B. Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1520–1528.
42. 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] [PubMed]
43. Bertasius, G.; Shi, J.; Torresani, L. Semantic Segmentation with Boundary Neural Fields. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 3602–3610.
44. Marmanis, D.; Schindler, K.; Wegner, J.D.; Galliani, S.; Datcu, M.; Stilla, U. Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS J. Photogramm. Remote Sens.* **2018**, *135*, 158–172. [CrossRef]
45. Wolf, A. Using WorldView 2 Vis-NIR MSI Imagery to Support Land Mapping and Feature Extraction Using Normalized Difference Index Ratios. In *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery*; SPIE: Baltimore, MD, USA, 2012; Volume 8390, p. 83900N.
46. Lin, C.; Wu, C.C.; Tsogt, K.; Ouyang, Y.C.; Chang, C.I. Effects of atmospheric correction and pansharpening on LULC classification accuracy using WorldView-2 imagery. *Inf. Process. Agric.* **2015**, *2*, 25–36. [CrossRef]
47. Karen, S.; Andrew, Z. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
48. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. In *Computer Vision—ECCV 2014*; Springer: Zurich, Switzerland, 2014; pp. 818–833.
49. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
50. Zhang, Y.; Qiu, Z.; Yao, T.; Liu, D.; Mei, T. Fully Convolutional Adaptation Networks for Semantic Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6810–6818.
51. Mountrakis, G.; Im, J.; Ogole, C. Support vector machines in remote sensing: A review. *ISPRS J. Photogramm. Remote Sens.* **2011**, *66*, 247–259. [CrossRef]
52. Drăguţ, L.; Csillik, O.; Eisank, C.; Tiede, D. Automated parameterisation for multi-scale image segmentation on multiple layers. *ISPRS J. Photogramm. Remote Sens.* **2014**, *88*, 119–127. [CrossRef] [PubMed]
53. eCognition Developer. *Trimble eCognition Developer 9.0 Reference Book*; Trimble Germany GmbH: Munich, Germany, 2014.
54. Fan, R.; Chen, P.; Lin, C. Working Set Selection Using Second Order Information for Training Support Vector Machines. *J. Mach. Learn. Res.* **2005**, *6*, 1889–1918.
55. Blaschke, T. Object based image analysis for remote sensing. *ISPRS J. Photogramm. Remote Sens.* **2010**, *65*, 2–16. [CrossRef]
56. Zhang, L.; Zhang, L.; Du, B. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 22–40. [CrossRef]
57. 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*, 8–36. [CrossRef]
58. Fu, G.; Liu, C.; Zhou, R.; Sun, T.; Zhang, Q. Classification for high resolution remote sensing imagery using a fully convolutional network. *Remote Sens.* **2017**, *9*, 498. [CrossRef]

59. Lu, Z.; Im, J.; Rhee, J.; Hodgson, M. Building type classification using spatial and landscape attributes derived from LiDAR remote sensing data. *Landsc. Urban Plan.* **2014**, *130*, 134–148. [CrossRef]
60. Fauvel, M.; Benediktsson, J.A.; Chanussot, J.; Sveinsson, J.R. Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 3804–3814. [CrossRef]
61. Song, J.; Lin, T.; Li, X.; Prishchepov, A.V. Mapping Urban Functional Zones by Integrating Very High Spatial Resolution Remote Sensing Imagery and Points of Interest: A Case Study of Xiamen, China. *Remote Sens.* **2018**, *10*, 1737. [CrossRef]
62. Zheng, X.; Wu, B.; Weston, M.V.; Zhang, J.; Gan, M.; Zhu, J.; Deng, J.; Wang, K.; Teng, L. Rural settlement subdivision by using landscape metrics as spatial contextual information. *Remote Sens.* **2017**, *9*, 486. [CrossRef]
63. 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 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 3213–3223.
64. Ha, Q.; Watanabe, K.; Karasawa, T.; Ushiku, Y.; Harada, T. MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems, Vancouver, BC, Canada, 24–28 September 2017.
65. Torres-Sánchez, J.; López-Granados, F.; Serrano, N.; Arquero, O.; Peña, J.M. High-throughput 3-D monitoring of agricultural-tree plantations with Unmanned Aerial Vehicle (UAV) technology. *PLoS ONE* **2015**, *10*, e0130479. [CrossRef] [PubMed]
66. Nguyen, K.; Bredno, J.; Knowles, D.A. Using contextual information to classify nuclei in histology images. In Proceedings of the 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), New York, NY, USA, 16–19 April 2015; pp. 995–998.
67. Wei, X.; Li, W.; Zhang, M.; Li, Q. Medical Hyperspectral Image Classification Based on End-to-End Fusion Deep Neural Network. *IEEE Trans. Instrum. Meas.* **2019**, 1–12. [CrossRef]
68. Sousa, A.M.; Machado, I.; Nicolau, A.; Pereira, M.O. Improvements on colony morphology identification towards bacterial profiling. *J. Microbiol. Methods* **2013**, *95*, 327–335. [CrossRef] [PubMed]
69. Turra, G.; Conti, N.; Signoroni, A. Hyperspectral image acquisition and analysis of cultured bacteria for the discrimination of urinary tract infections. In Proceedings of the 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, Italy, 25–29 August 2015; pp. 759–762.
70. Signoroni, A.; Savardi, M.; Baronio, A.; Benini, S. Deep Learning Meets Hyperspectral Image Analysis: A Multidisciplinary Review. *J. Imaging* **2019**, *5*, 52. [CrossRef]
71. Li, H.; Kadav, A.; Durdanovic, I.; Samet, H.; Graf, H.P. Pruning Filters for Efficient ConvNets. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017; pp. 1–13.
72. Zhang, X.; Zou, J.; He, K.; Sun, J. Accelerating Very Deep Convolutional Networks for Classification and Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1943–1955. [CrossRef] [PubMed]
73. Venkatesh, G.; Nurvitadhi, E.; Marr, D. Accelerating Deep Convolutional Networks using low-precision and sparsity. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 2861–2865.
74. Yim, J.; Joo, D.; Bae, J.; Kim, J. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1063–6919.

© 2019 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/).