applied sciences

MDPI

Article

Attention Neural Network for Water Image Classification under IoT Environment

Yirui Wu [1], Xuyun Zhang [2], Yao Xiao [3] and Jun Feng [1,*]

[1] College of Computer and Information, Hohai University, Nanjing 210098, China; wuyirui@hhu.edu.cn
[2] Department of Computing, Macquarie University, Sydney 2109, Australia; xuyun.zhang@mq.edu.au
[3] National Key Lab for Novel Software Technology, Nanjing University, Nanjing 210093, China; iam_xiaoyao@126.com
* Correspondence: fengjun@hhu.edu.cn

Received: 2 January 2020; Accepted: 21 January 2020; Published: 30 January 2020

Abstract: With significant development of sensors and Internet of things (IoT), researchers nowadays can easily know what happens in water ecosystem by acquiring water images. Essentially, growing data category and size greatly contribute to solving water pollution problems. In this paper, we focus on classifying water images to sub-categories of clean and polluted water, thus promoting instant feedback of a water pollution monitoring system that utilizes IoT technology to capture water image. Due to low inter-class and high intra-class differences of captured water images, water image classification is challenging. Inspired by the ability to extract highly distinguish features of Convolutional Neural Network (CNN), we aim to construct an attention neural network for IoT captured water images classification that appropriately encodes channel-wise and multi-layer properties to accomplish feature representation enhancement. During construction, we firstly propose channel-wise attention gate structure and then utilize it to construct a hierarchical attention neural network in local and global sense. We carried out comparative experiments on an image dataset about water surface with several studies, which showed the effectiveness of the proposed attention neural network for water image classification. We applied the proposed neural network as a key part of a water image based pollution monitoring system, which helps users to monitor water pollution breaks in real-time and take instant actions to deal with pollution.

Keywords: theoretical understanding of AI in the IoT; intelligent network; attention neural network; feature representation enhancement; multi-layer and channel-wise property

1. Introduction

Water ecosystems including rivers, lakes, and seas are facing great threats brought by fast development of human society. With development of sensor technology [1] and Internet of things (IoT) [2], category, volume, and quality of collected relevant data have been continuously increased and improved. With the help of collected data, researchers can develop systems to instantly monitor, control, and abate pollution, thus protecting water ecosystems. As an important research topic in water ecosystem monitoring, utilizing artificial intelligence to theoretical understand relevant data under the environment of IoT thus has been widely developed in areas of water resource management and environmental protection.

Thanks to deployment of drones, surveillance cameras, and other technologies of IoT [3,4], many relevant water data are easy to obtain. Involving big data technologies [5], water pollution monitoring system is thus greatly modified, transforming from methods of manual sampling to instant and automatic monitoring and analyses. The advantage brought by much modification is that government users can effectively know where and when pollution is happening without obvious time

delay. Core challenges of such monitoring system are thus to effectively analyze data captured by IoT technologies for detailed pollution information.

Among multiple water-relevant data types, we focus on one of the most common categories, i.e., water imagery. Furthermore, we perform image content understanding to achieve goal of water pollution monitoring. More precisely, we aim to construct a novel water pollution monitoring system, which could perform two classification tasks under the environment of IoT. First, such system can classify input water images into basic type, i.e., clean or polluted, thus knowing where pollution is happening. On the basis of general type of water images, the system should decide subcategories of input water images, which could provide sufficient information for users to take further actions. In other words, such system can not only classify clean water images into four subcategories, i.e., fountain, lake, ocean, and river, but also know what type of water pollution is happening, such as fungus, dead animals, industrial pollution, oil, and rubbish.

Based on the above goal and analysis of water image based monitoring system, the workflow for water pollution monitor can be viewed in Figure 1, in which we utilize analysis models to classify IoT captured water images into two categories and 10 subcategories. With the help of such water pollution monitoring system, users can monitor water pollution breaks in real-time and take instant actions to deal with pollution. In other words, we aim to accurately classify categories of water images to implement water pollution monitoring system. Since water images captured from environment of IoT generally suffer from artifacts, i.e., illumination variation, low contrast and complex background, it is a great challenge to accurately and efficiently perform classification tasks on water images [6]. The task of water image classification is difficult due to inherent ambiguity and low quality of IoT captured images.

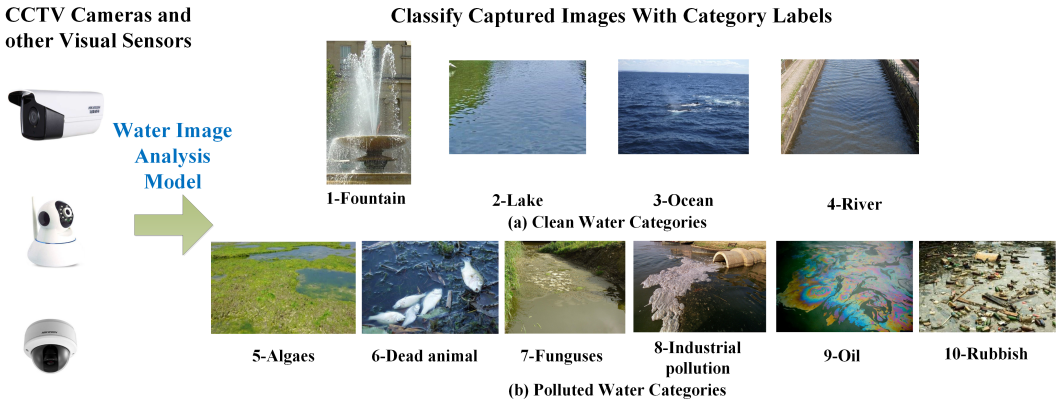

Figure 1. Workflow of water image based pollution monitoring system, where we can notice the task of the water image analysis model is to classify captured water images into two categories and 10 subcategories.

To solve this major problem of building water monitoring system, many works have been published to theoretically understand inherent meanings of water images. For example, Prasad et al. [7] fused color feature and optical flow to perform water identification, which achieves high performance in accuracy. Recently, Mettes et al. [8] built probability-based classifiers with spatiotemporal invariant descriptors to perform tasks of water detection. However, both methods need high-contrast water images with distinct boundary and simple background for reliable classification results.

Inspired by significant classification results achieved by typical Convolutional Neural Network (CNN) structure, quantity of trials on utilizing deep learning structures to comprehend water images have been made, where the core idea of such structures is to discover and apply the distinguished features for water image classification. Regarding water image as one typical category of natural scenes, Qi et al. [9] explore d deep learning architectures to extract texture related feature for scene classification. Later, Zhao et al. [10] utilized CNN and dimension reduction technology to construct

a novel spectral–spatial feature for water and other natural scene classification. However, both methods simply implement deep learning structures to solve problems of scene classification without specific modifications.

Based on the above discussion, we can notice that classification of water images is difficult with high intra-class and low inter-class properties, which can be defined as an ambiguous classification problem. Simply deploying neural networks cannot reach high accuracy in classification due to small number of training samples. Essentially, deep learning structures generally require big data to support semantic ambiguous classification. Expanding potentials on extracting representative features is proper to deal with insufficient training data. We thus propose to design a task-specific CNN for water image classification, which can achieve results with high precision and recall performance.

Essentially, the major challenge for building a task-specific CNN for water image classification is to extract highly distinguished features for processing. With years of developing and comprehending the CNN structure, researchers tend to consider a convolutional layer compute feature map with structure of variant channel filters. Inside a feature map, each slice spatially encodes visual responses under the influence of channel filter. Essentially, such filter can be regarded as pattern recognizer, where low-level visual cues such as lines and texture can be recognized by lower-layer filters, while semantic meanings such as category are recognized by higher-layer ones. To prove such supposition, researchers stack layers to view output, where features are computed to form visual abstraction with a hierarchy structure. Such phenomenon coincides with hypothesis that features extracted from CNN own important channel-wise and multi-layer characteristics.

Some features are useless for classifying water images. For example, the illumination channel can be useless for water image classification, since sensors are generally settled in the wild facing high variations in illumination. On the basis of irrelevant feature channels, water image analysis model could result in low effectiveness, since irrelevant feature channels bring noise for classification. We thus construct an attention neural network to pay attention to informative features.

In this paper, we firstly construct channel-wise attention gate structure and then build attention neural network by hierarchically involving multi-layer attention gates. With task-specific structure, the proposed network resolves informative features for water image classification. There exist three main contributions in this work:

- A context attention neural network for water image classification task is proposed, in which a task-specific attention model is proposed to encode channel-wise and multi-layer characteristics into features.
- The proposed model introduces channel-wise attention gate structure and builds hierarchical attention structure by utilizing multiple attention gates in different layers.
- Since the proposed attention neural network is simple to be implemented and deployed, we believe it can be powerful and easy to help detect inherent patterns for solving semantical problems, even facing ambiguous classification.

2. Related Work

We categorize the related methods into two groups, namely water image classification and attention mechanism, and then offer detailed descriptions.

2.1. Water Image Classification

Many studies have been applied to resolving problems of water pollution. Among them, one of the most important topics is to classify pollution types of water images. In fact, many related approaches have been developed to the benefit of accurately monitoring water information, including efforts to construct cloud-based monitoring systems [11,12].

Early, Zhang et al. [13] utilized a flip invariant shape detector for reflected contour detection. After locating reflected contours with edge features, their proposed method utilizes contour locations

to perform water image segmentation. Finally, their proposed method extracts features on sub-regions to perform detection task on water surface. Following their idea, Rankin et al. [14] detected water body by sky reflection, which computes similarity of intensity values to accurately locate water body. Later, Santana et al. [15] proposed performing water detection following the principle of dynamic texture recognition guided segmentation, which adopts entropy to model and learn texture-related property of water images. Rokni et al. [16] fused multiple technologies to detect change of water surface. Their proposed approach could produce a sharpened multi-spectral image for classification, thus providing a high accuracy result. To sum up, traditional methods to resolve water pollution problem are composed of manual feature extraction and classifier construction steps, which are often complex, costly, and time-consuming.

With the fast development of deep learning structures [17–19], researchers have applied more deep networks to perform classification tasks on water images. Inspired by CNN models to analyze close range photography, Zhao et al. [20] trained a CNN to perform classification tasks on input SAR image patches. After conducting comparative studies, they concluded that CNN is considerably better than the traditional classification methods, i.e., SVM, and has great potential to apply for SAR image interpretation. Regarding water surface as one important object category for classification, Pan et al. [21] proposed vertex component analysis network (R-VCANet) to perform multiple objects classification by inputting hyperspectral images. R-VCANet is built on the basis of spatial and spectral characteristics of HSI data, which achieves higher accuracy even with limited size of training dataset.

Following the idea of Pan et al. [21], Chen et al. [22] utilized a novel neural network to work on high-resolution remote sensing (HRRS) images for locations of urban water bodies. After segmenting inputting image into high-quality superpixels, their proposed method designs a task-specific CNN to extract semantical features of water surface, which is further applied to classify class label of superpixel: water or no-water pixel. The proposed method is similar to that of Chen et al.'s [22] in classification goal. However, we perform ambiguous classification at image level rather than binary classification at pixel level, which is more challenging than the problem they considered. Based on development of CNN-based model for water interpretation and developing of cloud-edge computing [23,24], Pan et al. [25] developed a low-cost water surveillance system with cameras as main sensors, which could automatically predict water levels via a deep CNN structure.

2.2. Attention Mechanism

Humans can easily ignore areas without salient information, while focusing on important locations for high efficiency. Such interesting phenomenon is named as visual attention mechanism, which is actually a selective strategy performed by human neurons. After observing and analyzing human attention, researchers move it to deep learning structure to focus on informative information passed through neural network. Current attention models can be divided into two kinds. Hard attention means mechanical selection on input areas, which leads the input areas to be processed as different parts with values of 0 (ignore areas) or 1 (concentrate areas). For example, He et al. [26] utilized image context to help determine salient regions at first, and then extracted related features from such regions to perform accurate text detection.

Soft attention assigns flexible weight values between 0 and 1 for selection. Therefore, it has been widely used by many deep learning applications. Spatial attention is firstly constructed to re-weight CNN generated feature map. For example, Ramanathan et al. [27] adopted soft attention to detect events in RGB videos, which focuses on the people who are responsible for the event. To utilize global attention for a higher accuracy and robustness, Liu et al. [28] iteratively optimized global attention weights for action sequence frames, which severs as an additional informative function for human action recognition. Most recently, Zhao et al. [29] involved strength of recurrent learning and attention mechanism to discover inherent context patterns among data distribution, which leads to a precise attention model and greatly promotes pedestrian recognition.

Besides spatial attention, many works have been proposed to utilize special category of channel feature for accurate recognition. For example, Chen et al. [30] proposed SCA-CNN to firstly involve both spatial- and channel-wise attention, which achieves a high accuracy for image caption task. However, their proposed CNN is specially and carefully designed, thus is too complicated to transform into other tasks.

Following the idea of extracting channel-wise features, Hu et al. [31] proposed "Squeeze-and-Excitation" (SE) block to describe dependencies among feature channels, thus extracting channel-wise context information successfully and independently. Most recently, Anderson et al. [32] proposed bottom-up and top-down attention to discover salient areas for image caption task, which has reported significant performance on accuracy.

3. Methods

We firstly introduce hierarchical architecture of attention neural network. Then, we design a task-specific channel-wise attention gate to enhance feature representation with channel-wise and multi-layer properties. Finally, we describe objective function design and training process.

3.1. Network Architecture Design

In the literature, water image classification is generally difficult, due to the ambiguous of the problem and the ineffectiveness of manual features to offer hints. Since sufficient visual information can be extracted by pre-trained deep structures on visual category classification problem, we build the proposed network on the basis of VGG-16 network. Figure 2 shows structure design of attention neural network, where we construct four layers of attention gates to involve hierarchical attention.

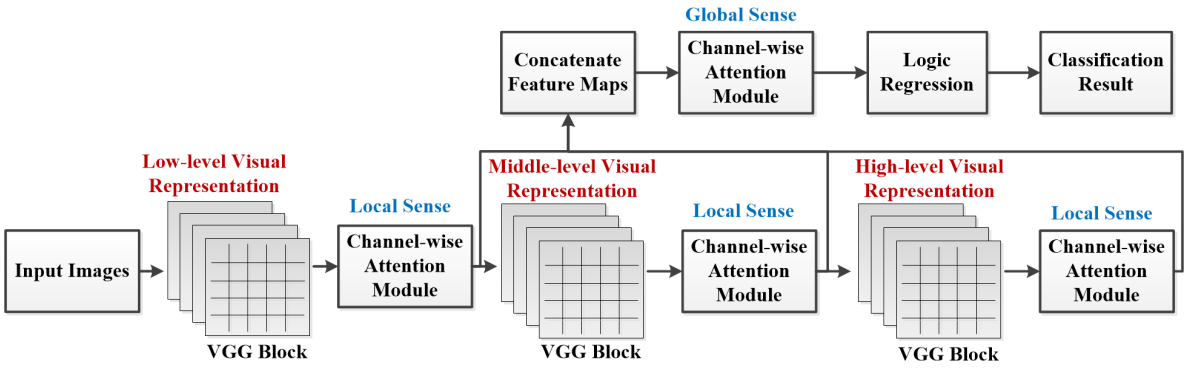

Figure 2. Structure design of attention neural network, where VGG block is designed to extract feature map, channel-wise attention module is utilized to involve channel-wise context information, and multi-layers of attention modules are used to build hierarchy structure.

We build attention neural network to emphasize informativeness of low- and middle-level features, since such features can be highly informative for ambiguous classification, i.e., water image classification problem. For example, one of the low-level features, i.e., texture, can be highly informative to classify water surface polluted by oil [15]. Without channel-wise attention on texture feature, decay of gradients and visual abstraction could result in ignoring texture feature in higher layers. This is quite true in deep neural networks that forget low- or middle-level features due to high depth of deep structures. Based on the above discussions, we build attention neural network to emphasize the importance of low- and middle-level features. To show visual cues of feature channels extracted by the proposed network, we offer samples of input images and corresponding low-, middle-, and high-level visual representations in Figure 3.

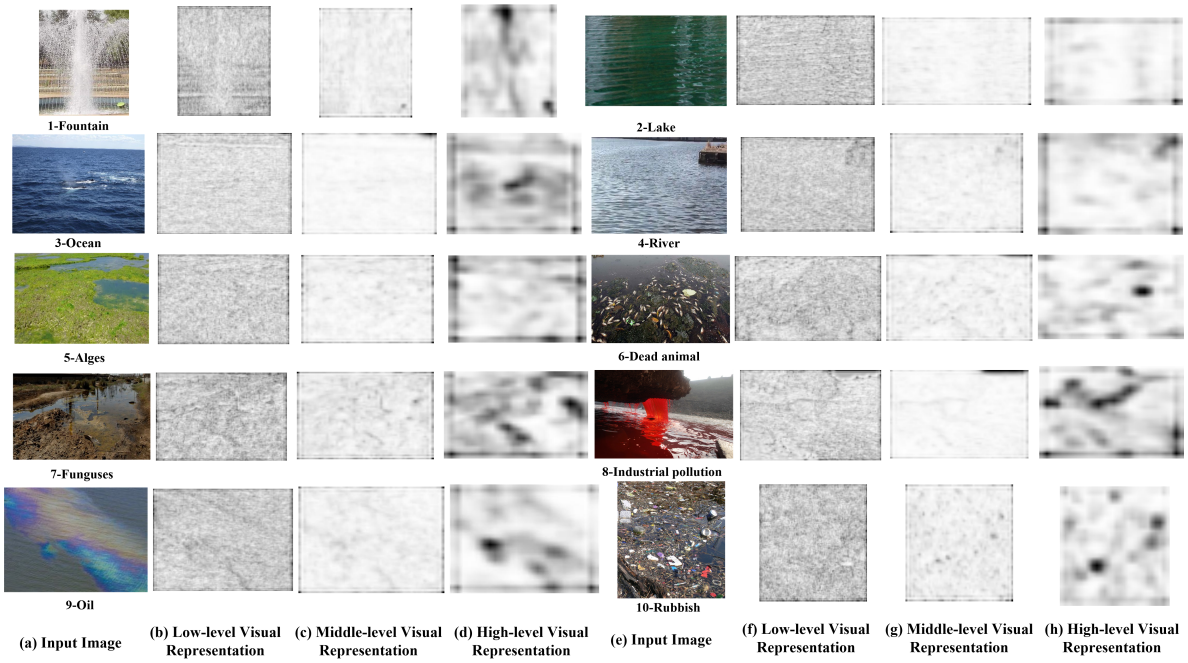

Figure 3. Input samples and corresponding intermediate results for low-, middle-, and high-level visual representations of feature maps extracted by the proposed network.

Regarding blocks of VGG16 network as the basic modules for feature extraction, we build three layers of attention gates to involve local channel-wise attention information, where such gates are located after VGG-16 blocks, as shown in Figure 2. In other words, we emphasize informative and important feature channels in different levels of visual feature representation with local attention gates. After locating feature extraction modules and attention gates, we concatenate features passing through attention gates to make up a novel representative feature, which offers assembled form of information collected from different visual levels. After concatenating, we locate another attention gate to focus on informative channels of concatenated feature, which provides selection on information in a global sense. After locally and globally emphasizing, the constructed feature is thus utilized to predict category labels with a logic regression classifier. The structure design of attention neural network is shown in Figure 2.

Specifically, we build local attention modules to pay special attention to informative features channels, which extracts context information in a local sense to improve classification. Essentially, local attention module is built as element-wise weights for different feature channels

$$\tilde{U_{k,l}} = \Phi(U_{k,l}) \cdot U_{k,l} \tag{1}$$

where l refers to the index of feature channel, k is the index of VGG16 block, $U_{k,l}$ represents one specific channel feature of generated feature map, $\cdot$ is element-wise multiplication, and function $\Phi()$ represents attention gate.

After constructing local attention modules, we concatenate weighted feature maps output by local attention modules to form global feature map. Such concatenated feature map could be regarded as a combination of multi-layer CNN feature. To utilize multi-layer characteristics, we build global channel-wise attention module to emphasize informativeness of low- and middle-level visual features.

Therefore, we further utilize attention gate to assign weight for multi-level feature map in a global sense:

$$\begin{cases} \tilde{U_k} = [\tilde{U_{k,l}}], l = 1, \ldots, c \\ \hat{U_k} = \tilde{U_k} \cdot \Phi(\tilde{U_k}) \end{cases} \tag{2}$$

where function $[\cdot]$ represents operation of matrix transforming and concatenating, which results in a single feature map. Based on the final generated feature map $\hat{U}_k$, which involves context information at both local and global levels, we apply average pooling operation to re-scale $\hat{U}_k$ and utilize a logical regression classier to classify $\hat{U}_k$ into ten sub-categories.

3.2. Design of Channel-wise Attention Gate

We firstly discuss theoretical fundament of soft attention mechanism. Afterwards, we build channel-wise attention gate based on soft attention mechanism, which essentially acts as a lightweight gating mechanism.

The proposed attention gate can describe channel-wise relationships among feature maps, which is utilized to construct attention neural network at local and global levels. Compared with traditional manually assigned weights, the proposed attention model can automatically select informative channels to perform classification tasks based on inherent characteristics of collected data, which is more flexible for different application scenarios.

Theoretical Fundament of Soft Attention Mechanism. Constructing soft attention model consists of two steps, namely computing weights according to similarity of input and trained signal, and re-scaling the feature map with weights.

To compute similarity between input and trained signal, researchers usually adopt MLP for implicit calculation, which can be defined as

$$sim(Q, W_i) = MLP(Q, W_i) \tag{3}$$

After computing similarity, Softmax function is often utilized to complete multi-tasks, i.e., normalization and assigning large values on important parts:

$$\alpha_i = softmax(sim(Q, W_i)) = \frac{e^{sim(Q,W_i)}}{\sum_{j=1}^{L} e^{sim(Q,W_i)}}; \tag{4}$$

where L is number of trained signals.

After calculating new weight for input signal, we can sum weighted parts to form new signal $Atten$:

$$Atten = \sum_{i=1}^{L} \alpha_i \cdot v_i \tag{5}$$

where v_i refers to part of input signal.

Construction of Channel-wise Attention Gate. We build the structure of the proposed attention gate shown in Figure 4, which is further utilized for description of local and global attention information. We firstly represent a feature map as $U = [u_i; i = 1, \ldots, c]$, where u_i refers to ith channel and c is the size of channels. In fact, each filter operates on pixels to compute features within a local receptive field, which fails to exploit contextual information outside its receptive field [31]. We thus collect channel-wise information to form a characteristics descriptor M by mean pooling

$$\begin{cases} M = [m_i; i = 1, \ldots, c] \\ m_i = \frac{1}{W \times H} \sum_{j=1}^{W} \sum_{k=1}^{H} u_i(j, k)) \end{cases} \tag{6}$$

where M consists of m_i, W is the width of the ith channel, and H refers to its height .

Afterwards, we construct weighting scheme to assign informativeness values to different channels. Considering that the relationship between channels is a highly nonlinear function of feature map, we construct a weighting scheme with two fully-connected layers to calculate weight vector G as

$$\begin{cases} G = [g_i; i = 1, \ldots, c] \\ g_i = Nor(sig(W_2\eta(W_1 m_i + b_1) + b_2)), \end{cases} \tag{7}$$

where function $sig()$ is sigmoid; $\eta()$ is ReLU; W_1, W_2, b_1, and b_2 are parameters to be trained; and function $Nor()$ re-scales the output to be the form of a one-hot activation.

As shown in Figure 4, G assigns weights on different channels of input feature map U to emphasize informative feature channel, which could be represented as

$$\tilde{u_i} = g_i \otimes u_i \tag{8}$$

where $\otimes$ refers to element-wise multiplication.

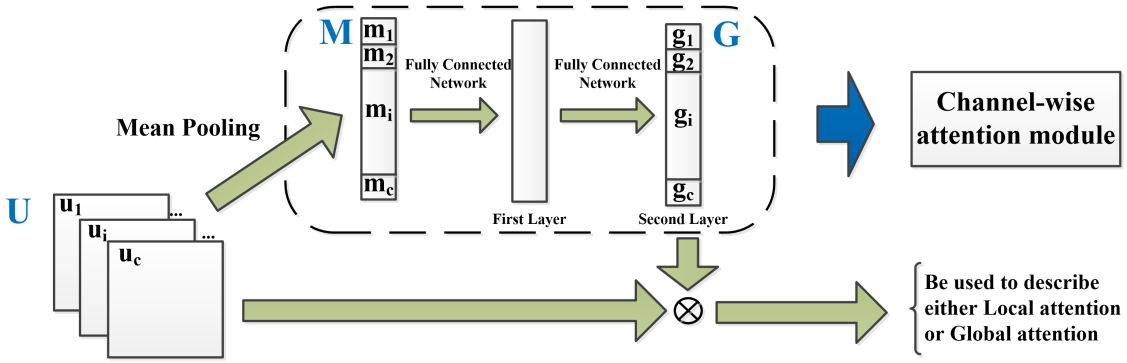

Figure 4. Architecture of the proposed attention gate for description of local and global attention.

3.3. Design of Objective Function

In this subsection, we mainly design the objective function to perform joint learning and reduce overfitting. During training, we should jointly train VGG-16 block and the hierarchical attention gates by designing a proper objective function. Under such consideration, we prefer regularized cross-entropy loss rather than mean square error, since the objective function built on cross-entropy loss encourages the proposed network to converge with fast speed, especially facing difficulty of jointly training.

We thus formulate objective function as a form with cross-entropy loss, which is defined for a single sample

$$Loss = -\sum_{i=1}^{n} y_i \log P_{y_i} + \lambda_1 \|W_N\|_2 + \lambda_2 \|B_N\|_2 \tag{9}$$

where n refers to size of ground-truth labels, P_{y_i} indicates probability to assign ith class to sample, and we define $y_i = 1$ and $y_j = 0$ for $j \neq i$, if the sample is within ith class. In Equation (9), two L2 norm items are designed to prevent overfitting. Specifically, B_N consists of b_1 and b_2. W_N denotes connection matrix by merging W_1, W_2, and other parameter matrices of VGG-16 blocks.

4. Results and Discussion

We first describe our adopted dataset and measurement. Then, we present the comparative experiments to show the performance of the proposed attention neural network. Finally, the implementation details are provided.

4.1. Dataset and Measurement

Essentially, there does not exist a proper benchmark dataset for water image classification task. We thus used a dataset obtained from Wu et al. [33] with 1000 water images. As reported in their paper, they collected water images from either standard water related videos [8] or Internet sources including Google, Bing, and Baidu. With the same criterion for sub-categories of water images, their water images are labeled with ten subclasses, namely, Fountain, Lakes, Oceans, and Rivers for clean water and Algae, Dead animals, Fungus, Industrial pollution, Oils, and Rubbish for polluted water. After analyzing the image content, the considered dataset coincides with the real situation of water pollution monitoring.

We evaluated classification with four standard measurements, namely precision, recall, f-score, and accuracy:

$$\begin{aligned} \text{Precision} &= \frac{\text{TP}}{\text{TP}+\text{FP}} \\ \text{Recall} &= \frac{\text{TP}}{\text{TP}+\text{FN}} \\ \text{F-score} &= \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision}+\text{Recall}} \\ \text{Accuracy} &= \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \end{aligned} \tag{10}$$

where TP refers to the number of classified images in positive class, TN refers to the number of classified images in negative class, and FP and FN are numbers of incorrectly classified images in positive and negative classes, respectively.

4.2. Performance Analysis

We adopted two related comparison methods in experiments: Wu et al. [33] and Mettes et al. [8]. Wu et al. [33] firstly divided input images into sub-regions by angle and spatial features of Fourier spectrum, and then extracted mean and variance features over sub-regions to form feature matrix, which represented texture properties of water images and was further utilized for classification. Mettes et al. [8] proposed spatiotemporal feature to detect water surface in inputting sequences, which does not coincide with aim of the proposed method. We utilized their proposed spatiotemporal feature to construct SVM classifier for classification goal. We implemented both methods following the instructions in their recently published papers.

Both comparative methods can be applied to perform water image classification task, since the core of their methods are feature representation of visual information of water surface images. Therefore, both methods were adopted as comparative studies to show superiority of feature representation by incorporating context information, which is the main contribution of the proposed network. In fact, simple and manually-designed features are not sufficient to deal with water image classification, due to its complexity and ambiguous characteristics. We thus offer a novel network to not only self-extract feature representation, but also pay attention to the information part after extraction. Compared with the other studies, all these features led to a better accuracy and robustness on task of water image classification.

We show quantitative results of classification experiments among the proposed network with and without channel-wise attention module and comparative studies in Table 1. Note that we report measurements on classification for clean water images (four subclasses), polluted water images (six subclasses), and total water images (ten subclasses). To better visualize performance of three classification methods, we also represent confusion matrices for clean, polluted, and total water images in Figures 5–7, respectively.

Table 1. Comparison of classification performance among the proposed method with attention module (short for WA), without attention module (short for WoA), Wu et al. [33], and Mettes et al. [8]. We performed three experiments to show the performance on classification of clean water, polluted water, and the entire dataset. Bold text indicates the best performance among the comparative studies.

Method	Clean Water				Polluted Water				Total
	Precision	Recall	F-Score	Accuracy	Precision	Recall	F-Score	Accuracy	Accuracy
The Proposed (WA)	**0.65**	**0.69**	**0.67**	**66.4%**	**0.65**	**0.67**	**0.66**	**73.6%**	**71.2%**
The Proposed (WoA)	**0.65**	**0.69**	**0.67**	63.2%	0.60	0.64	0.62	65.6%	69.2%
Wu et al. [33]	0.52	0.52	0.52	52.0%	0.44	0.45	0.44	59.2%	51.2%
Mettes et al. [8]	0.30	0.30	0.30	41.6%	0.07	0.16	0.10	35.2%	20.0%

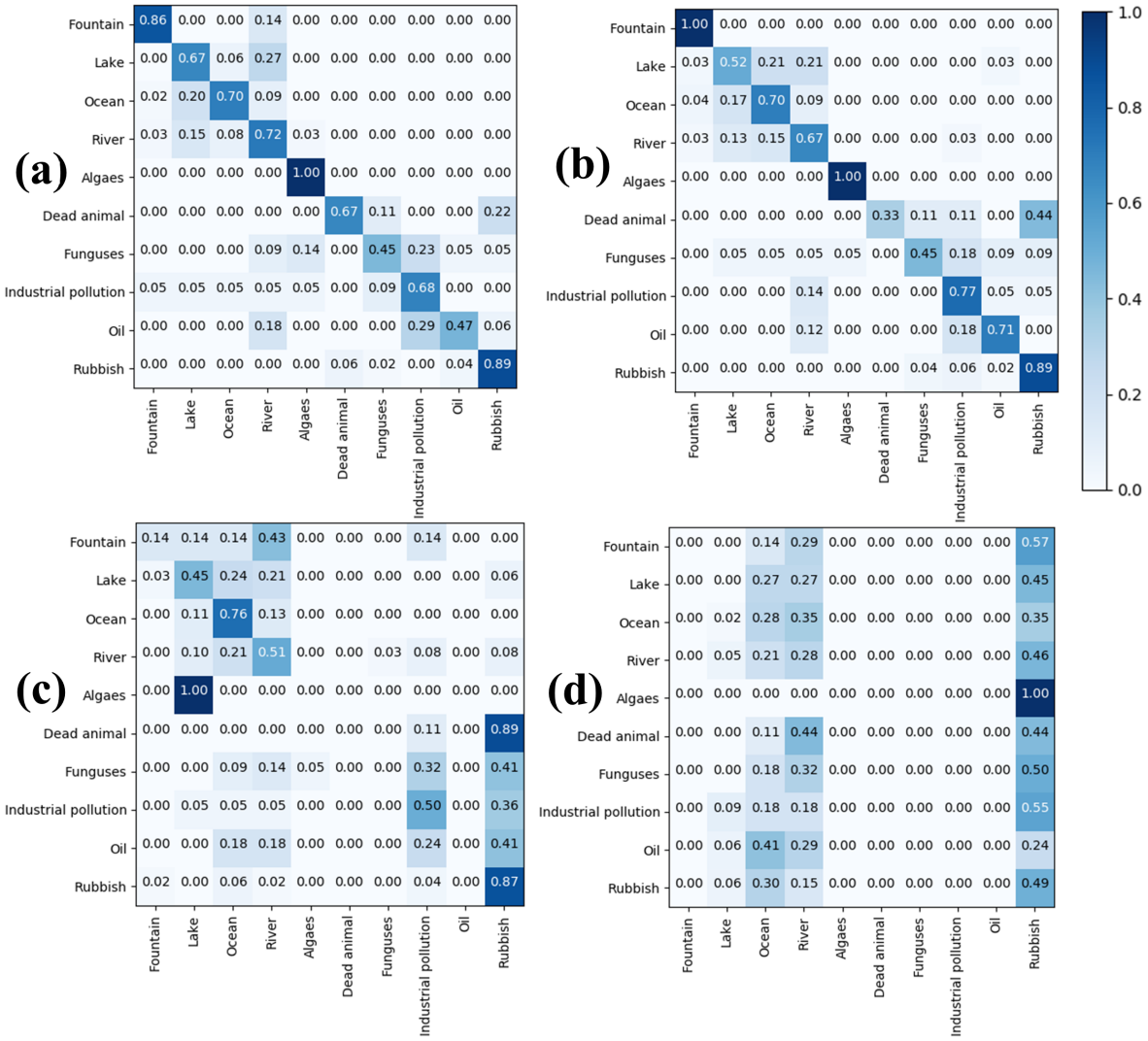

Figure 5. Classification Confusion matrices for total water images: (**a**) the matrix achieved by the proposed method (WA); (**b**) the matrix achieved by the proposed method (WoA); (**c**) the matrix achieved by Wu et al. [33]; and (**d**) the matrix achieved by Mettes et al. [8].

As shown in Table 1 and all the confusion matrix figures, we observed that performance for four clean, six polluted and ten total subclasses achieved by the proposed method (WA) is highest, which shows the proposed network successfully performs water image classification task. High performance

of the proposed method (WA) and (WoA) prove high distinguishability of deep learning structures. By comparing between the proposed WA and WoA methods, we can notice promotion on most of measurements is high due to the implementation of channel-wise attention module, which proves the effectiveness of utilizing context information for enhancement of the CNN feature map. The reason to achieve the same value in precision, recall, and F-score on clean water subclass lies in the fact that the dataset is unbalanced in numbers, where the number of clean water images is rather low and increasing measurement values with a small dataset is extremely difficult.

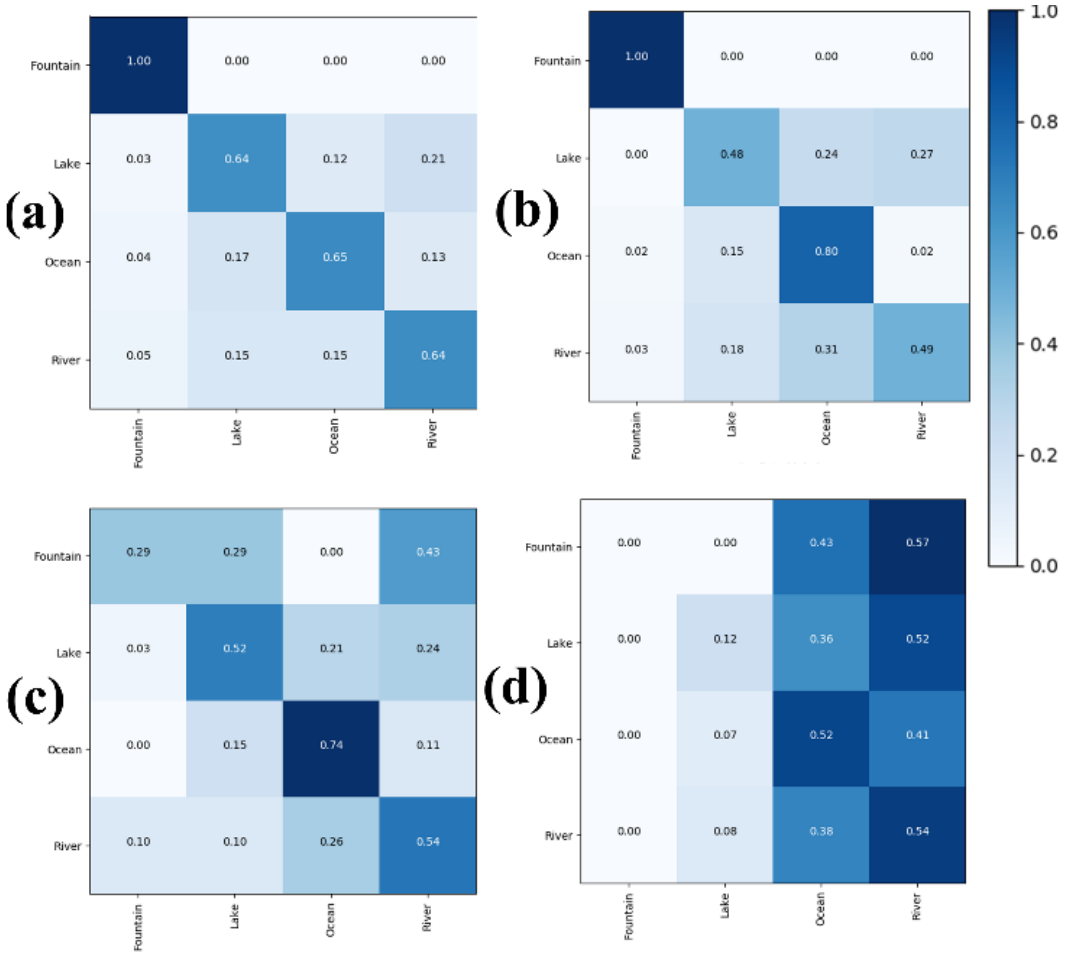

Figure 6. Classification confusion matrix for clean water images: (**a**) the matrix achieved by the proposed method (WA); (**b**) the matrix achieved by the proposed method (WoA); (**c**) the matrix achieved by Wu et al. [33]; and (**d**) the matrix achieved by Mettes et al. [8].

For comparative studies, Mettes et al. [8] achieved poor results, since their spatiotemporal descriptor is initially proposed for water surface detection. Their method has a high standard for quality of input images with high-contrast water and distinct boundary. In our adopted dataset, there exist many irregular objects, such as rubbish, dead animals, etc., in polluted images, thus leading to poor classification results. Meanwhile, Wu et al. [33] utilized HSV and Fourier spectrum to construct representative features, which overcomes the shortage of visual irregular shapes by frequency domain analyses. Therefore, their proposed method is much better than that of Mettes et al. [8] in classification performance. However, patterns of water images cannot be successfully detected by utilizing a manual feature. In Figure 5c, we can notice misclassification of dead animal as rubbish is 0.89, while correctness to classify rubbish is 0.87. After analyzing, we believe such phenomenon can be explained by the fact that the manually-designed spectrum feature proposed by Wu et al. [33] is highly effective for recognizing rubbish other than dead animals.

Essentially, low inter-class and high intra-class variations make the task of water image classification rather difficult, where one proper example is to recognize between rubbish and dead animals subclasses. Involving the ability to focus on informative visual cues, the proposed network greatly improves performance in clean water image classification, which could be proved by the confusion matrix in Figure 6. This is due to high visual resemblance of clean water images, where we should not only use high-level features, but also request assistance in classification from low-level features. However, noise could be brought by much information from low-level features to affect overall results. Therefore, a slight decrease in recall for polluted category can be viewed in Figure 7.

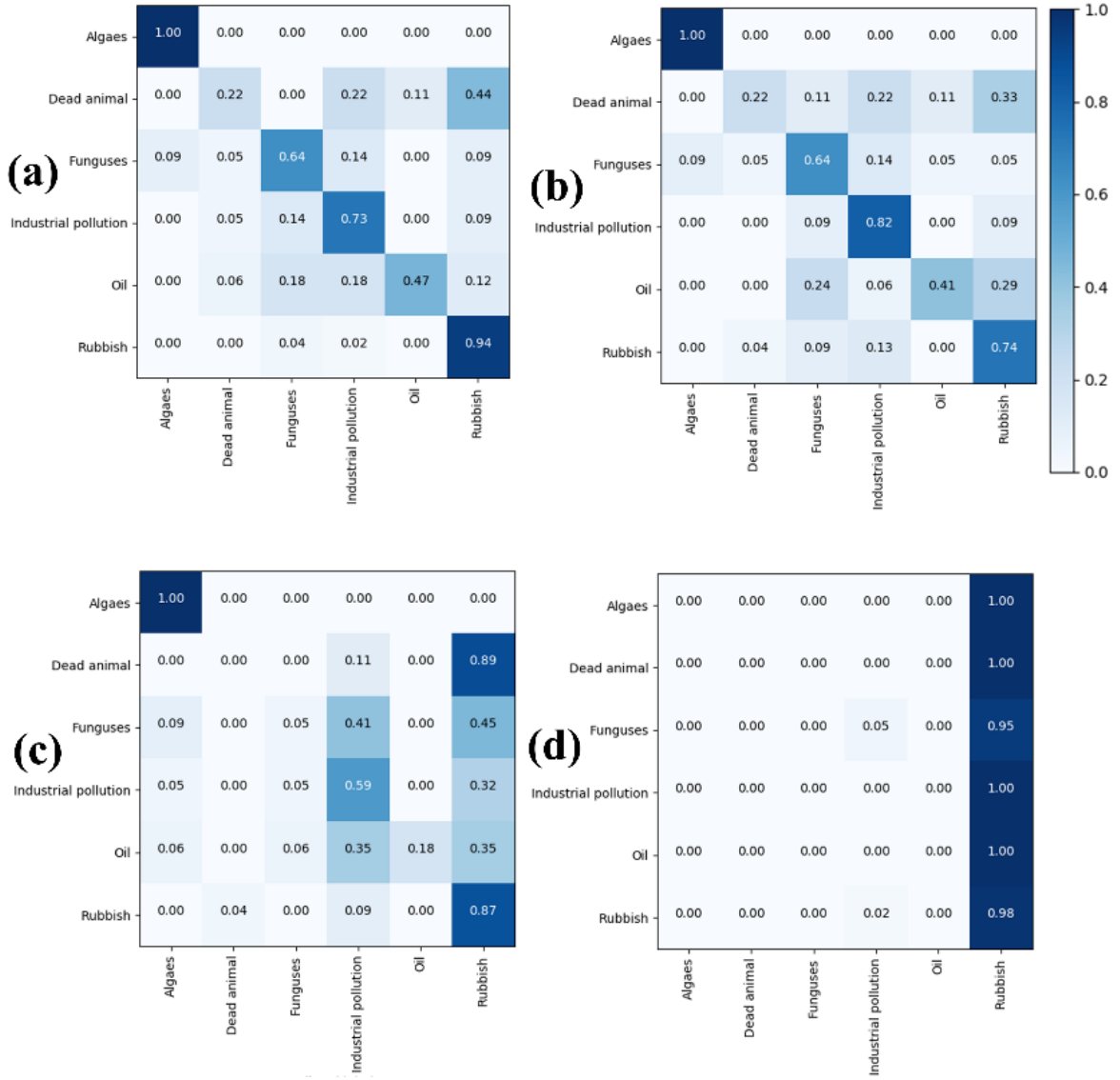

Figure 7. Classification confusion matrix for polluted water images: (**a**) the matrix achieved by the proposed method (WA); (**b**) the matrix achieved by the proposed method (WoA); (**c**) the matrix achieved by Wu et al. [33]; and (**d**) the matrix achieved by Mettes et al. [8].

Facing difficulties brought by ambiguous classification, we propose to encode context information including channel-wise and multi-layer properties for feature map enhancement, which greatly improves representation ability of generated feature map to act as qualified pattern detectors. However, misclassification still occurs in results achieved by the proposed network. For example, in Figure 5a, we can observe the 18%, 29%, and 6% of oil polluted samples are misclassified as clean river, industrial pollution, and rubbish, respectively. According to the analysis on these misclassified

images, we find small inter-class variations among samples promotes high misclassification rate. Despite coupling with attention scheme to boost performance, water image classification as an ambiguous classification is still difficult to be completely resolved. Moreover, the number of samples corresponding to different categories of pollution is not only small in size, but also unbalanced distributed among multiple categories. Therefore, an unbalanced dataset also results in a relatively high misclassification rate. Overall, the proposed attention neural network successfully boosts water image classification performance.

When applying the proposed network in real-life scenarios, it would come across mixed pollutant situations. In other words, there may exist multiple categories of pollution in one input surveillance image, such as dead animals plus rubbish, dead animals plus oil spill, etc. To solve this mixed pollutant problem, we will modify this method for higher accuracy by defining such mixed situations as new categories for classification. Since the proposed network can handle with multiple classification problem, increasing the number of categories would slightly decrease accuracy if training with sufficient samples. Moreover, mixed pollutants can be solved with the idea of fusing intermediate feature representation of samples from multiple categories, which is more efficient in representation and would be our future work to improve the proposed network.

Figure 8 offers qualitative results of correct and misclassified water images. From misclassified images, we can conclude that the major challenge of water image classification is small inter-class variation, which is easy to be found among samples from subclasses of oil and industrial pollution. Moreover, the unbalanced training dataset promotes misclassification. For example, we can find more rubbish images than images about dead animals in the dataset, which leads the proposed network to focus on emphasizing an effective feature map to correctly recognize rubbish.

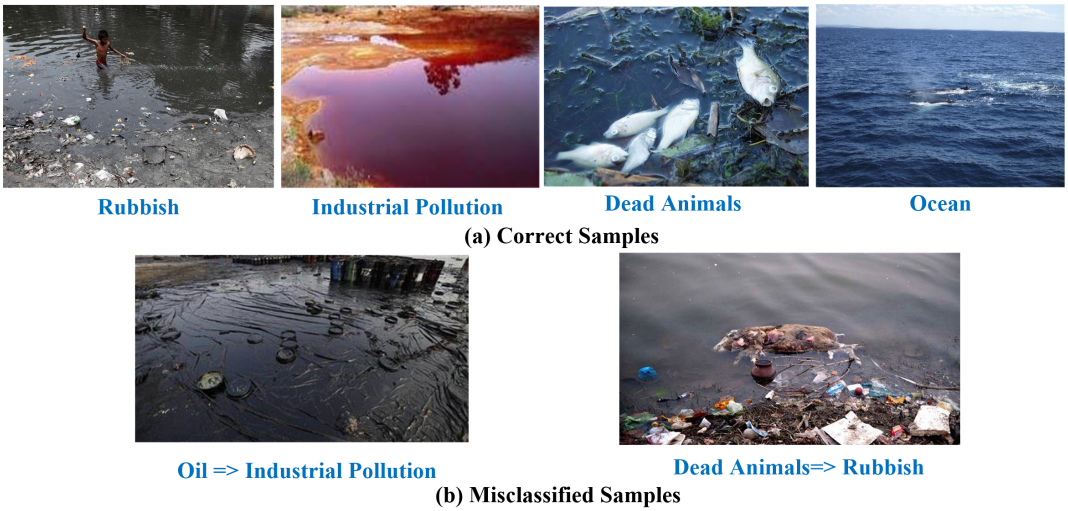

Figure 8. Samples of correct and misclassified water images achieved by the proposed attention neural network.

4.3. Implementation Details

We tested our algorithms on a PC (Intel Xeon with 6 cores @ 2.4 GHz, 60 Gb RAM, and one Nvidia GeForce GTX 1080 Ti card). During experiments, four-fold cross validation was adopted to fairly judge performance of the constructed network. Furthermore, we adopted back-propagation to minimize objective function in Equation (9) and we set corresponding weights as $\lambda_1 = e^{-4}$ and $\lambda_2 = e^{-5}$ by experiments. During training, we set batch size as 64, training iterations as 150, and learning rate for logic regression and other layers as $5e^{-3}$ and $5e^{-4}$.

5. Conclusions

We propose an attention neural network for IoT captured water images classification task, which dynamically modulates context of channel-wise and multi-layer characteristics to enhance feature map. During construction, we propose channel-wise attention gate at first and then utilize it to build hierarchical attention model. We carried out comparative experiments on an image dataset about water surface with several existing studies, which shows distinctive ability of the proposed attention neural network for water image classification.

With development of cloud computing [34] and edge computing [35], we believe an accurate and effective water image classification method is required for water pollution monitoring applications. As a key part in water pollution monitoring system, the proposed neural network is supposed to work on surveillance images with features of large volume and high quality. After instantly and efficiently analyzing on cloud-based architecture [36,37], the proposed network could compute specific category of pollution and suggest to users currently accessible solutions to deal with pollution. With such workflow, we believe the proposed network could not only largely decrease manual burden of users to keep watch on images of surveillance cameras, but also provide a total solution with features of real-time feedback and high intelligence for water pollution.

In the future, we would improve the proposed network with two novel features, i.e., ability to deal with mixed pollutants and a lightweight version to work on embedded systems such as smartphones, drones, and so on. Mixtures pollutants are common in our daily-life. Classifying mixed pollutants as a new class could solve such problems. However, such solution not only brings decrease in accuracy with many new classes, but also increased burden to collect a variety of training samples. Therefore, we would like to fuse visual feature representation to deal with mixed pollutants problem in the future. The current version of the proposed network requires large computation resource, which cannot work well on low-resource platforms. Facing difficulty that water pollution might occur in places without settled cameras, it is essential to develop a light version, which can convert all possible smartphone or drones into accessible surveillance cameras.

Author Contributions: Conceptualization, Y.W. and J.F.; methodology, Y.X.; software, Y.X.;validation, Y.W., X.Z. and Y.X.; formal analysis, Y.W.; investigation, Y.W.; resources, X.Z.; data curation, X.Z.; writing original draft preparation, Y.W.; writing review and editing, X.Z.; visualization, X.Z.; supervision, J.F.; project administration, J.F.; funding acquisition, Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by National Key R&D Program of China under Grant 2018YFC0407901, National Natural Science Foundation of China under Grant 61702160, and Natural Science Foundation of Jiangsu Province under Grant BK20170892.

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

References

1. Qi, L.; Zhang, X.; Dou, W.; Hu, C.; Yang, C.; Chen, J. A two-stage locality-sensitive hashing based approach for privacy-preserving mobile service recommendation in cross-platform edge environment. *Future Gener. Comput. Syst.* **2018**, *88*, 636–643. [CrossRef]
2. Wang, H.; Guo, C.; Cheng, S. LoC—A new financial loan management system based on smart contracts. *Future Gener. Comput. Syst.* **2019**, *100*, 648–655. [CrossRef]
3. Qi, L.; He, Q.; Chen, F.; Dou, W.; Wan, S.; Zhang, X.; Xu, X. Finding All You Need: Web APIs Recommendation in Web of Things Through Keywords Search. *IEEE Trans. Comput. Soc. Syst.* **2019**, *6*, 1063–1072. [CrossRef]
4. Xu, X.; Fu, S.; Qi, L.; Zhang, X.; Liu, Q.; He, Q.; Li, S. An IoT-oriented data placement method with privacy preservation in cloud environment. *J. Netw. Comput. Appl.* **2018**, *124*, 148–157. [CrossRef]
5. Wang, H.; Ma, S.; Dai, H.N. A rhombic dodecahedron topology for human-centric banking big data. *IEEE Trans. Comput. Soc. Syst.* **2019**, *6*, 1095–1105. [CrossRef]
6. Khan, M.; Wu, X.; Xu, X.; Dou, W. Big data challenges and opportunities in the hype of Industry 4.0. In Proceedings of the IEEE International Conference on Communications, Paris, France, 21–25 May 2017; pp. 1–6.

7. Prasad, M.G.; Chakraborty, A.; Chalasani, R.; Chandran, S. Quadcopter-based stagnant water identification. In Proceedings of the Fifth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), Patna, India, 16–19 December 2015; pp. 1–4.
8. Mettes, P.; Tan, R.T.; Veltkamp, R.C. Water detection through spatio-temporal invariant descriptors. *Comput. Vis. Image Underst.* **2017**, *154*, 182–191. [CrossRef]
9. Qi, X.; Li, C.G.; Zhao, G.; Hong, X.; Pietikäinen, M. Dynamic texture and scene classification by transferring deep image features. *Neurocomputing* **2016**, *171*, 1230–1241. [CrossRef]
10. Zhao, W.; Du, S. Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach. *IEEE Trans. Geosci. Remote. Sens.* **2016**, *54*, 4544–4554. [CrossRef]
11. Xu, X.; Xue, Y.; Qi, L.; Yuan, Y.; Zhang, X.; Umer, T.; Wan, S. An edge computing-enabled computation offloading method with privacy preservation for internet of connected vehicles. *Future Gener. Comput. Syst.* **2019**, *96*, 89–100. [CrossRef]
12. Xu, Y.; Qi, L.; Dou, W.; Yu, J. Privacy-preserving and scalable service recommendation based on simhash in a distributed cloud environment. *Complexity* **2017**, *2017*, 343785. [CrossRef]
13. Zhang, H.; Guo, X.; Cao, X. Water reflection detection using a flip invariant shape detector. In Proceedings of the IEEE International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; pp. 633–636.
14. Rankin, A.L.; Matthies, L.H.; Bellutta, P. Daytime water detection based on sky reflections. In Proceedings of the ICRA, Shanghai, China, 9–13 May 2011; pp. 5329–5336.
15. Santana, P.; Mendonça, R.; Barata, J. Water detection with segmentation guided dynamic texture recognition. In Proceedings of the IEEE International Conference on Robotics and Biomimetics (ROBIO), Guangzhou, China, 11–14 December 2012; pp. 1836–1841.
16. Rokni, K.; Ahmad, A.; Solaimani, K.; Hazini, S. A new approach for surface water change detection: Integration of pixel level image fusion and image classification techniques. *Int. J. Appl. Earth Obs. Geoinf.* **2015**, *34*, 226–234. [CrossRef]
17. Wang, H.; Wu, Y.; Min, G.; Xu, J.; Tang, P. Data-driven dynamic resource scheduling for network slicing: A Deep reinforcement learning approach. *Inf. Sci.* **2019**, *498*, 106–116. [CrossRef]
18. Zuo, Y.; Wu, Y.; Min, G.; Cui, L. Learning-based network path planning for traffic engineering. *Future Gener. Comput. Syst.* **2019**, *92*, 59–67. [CrossRef]
19. Liu, H.; Kou, H.; Yan, C.; Qi, L. Link prediction in paper citation network to construct paper correlation graph. *EURASIP J. Wirel. Commun. Netw.* **2019**, *2019*, 1–12. [CrossRef]
20. Zhao, J.; Guo, W.; Cui, S.; Zhang, Z.; Yu, W. Convolutional Neural Network for SAR image classification at patch level. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 945–948.
21. Pan, B.; Shi, Z.; Xu, X. R-VCANet: A new deep-learning-based hyperspectral image classification method. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2017**, *10*, 1975–1986. [CrossRef]
22. Chen, Y.; Fan, R.; Yang, X.; Wang, J.; Latif, A. Extraction of urban water bodies from high-resolution remote-sensing imagery using deep learning. *Water* **2018**, *10*, 585. [CrossRef]
23. Gong, W.; Qi, L.; Xu, Y. Privacy-aware multidimensional mobile service quality prediction and recommendation in distributed fog environment. *Wirel. Commun. Mob. Comput.* **2018**, *2018*, 3075849. [CrossRef]
24. Xu, X.; Li, Y.; Huang, T.; Xue, Y.; Peng, K.; Qi, L.; Dou, W. An energy-aware computation offloading method for smart edge computing in wireless metropolitan area networks. *J. Netw. Comput. Appl.* **2019**, *133*, 75–85. [CrossRef]
25. Pan, J.; Yin, Y.; Xiong, J.; Luo, W.; Gui, G.; Sari, H. Deep learning-based unmanned surveillance systems for observing water levels. *IEEE Access* **2018**, *6*, 73561–73571. [CrossRef]
26. He, T.; Huang, W.; Qiao, Y.; Yao, J. Text-Attentional Convolutional Neural Network for Scene Text Detection. *IEEE Trans. Image Process.* **2016**, *25*, 2529–2541. [CrossRef]
27. Ramanathan, V.; Tang, K.; Mori, G.; Fei-Fei, L. Learning temporal embeddings for complex video analysis. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4471–4479.
28. Liu, J.; Wang, G.; Hu, P.; Duan, L.; Kot, A.C. Global Context-Aware Attention LSTM Networks for 3D Action Recognition. In Proceedings of the IEEE Conference on Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3671–3680.

29. Zhao, X.; Sang, L.; Ding, G.; Han, J.; Di, N.; Yan, C. Recurrent Attention Model for Pedestrian Attribute Recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 December–1 January 2019; pp. 9275–9282.
30. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; Chua, T.S. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. *arXiv* **2016**, arXiv:1611.05594.
31. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. *arXiv* **2017**, *7*, arXiv:1709.01507.
32. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Lake City, UT, USA, 18–22 June 2018; pp. 6077–6086.
33. Wu, X.; Shivakumara, P.; Zhu, L.; Lu, T.; Pal, U.; Blumenstein, M. Fourier Transform based Features for Clean and Polluted Water Image Classification. In Proceedings of the International Conference on Pattern Recognition, Beijing, China, 20–24 August 2018.
34. Qi, L.; Chen, Y.; Yuan, Y.; Fu, S.; Zhang, X.; Xu, X. A QoS-aware virtual machine scheduling method for energy conservation in cloud-based cyber-physical systems. *World Wide Web* **2019**, 1–23. [CrossRef]
35. Wang, H.; Ma, S.; Dai, H.N.; Imran, M.; Wang, T. Blockchain-based data privacy management with nudge theory in open banking. *Future Gener. Comput. Syst.* **2019**. [CrossRef]
36. Xu, X.; Liu, Q.; Luo, Y.; Peng, K.; Zhang, X.; Meng, S.; Qi, L. A computation offloading method over big data for IoT-enabled cloud-edge computing. *Future Gener. Comput. Syst.* **2019**, *95*, 522–533. [CrossRef]
37. Xu, X.; Liu, Q.; Zhang, X.; Zhang, J.; Qi, L.; Dou, W. A Blockchain-Powered Crowdsourcing Method With Privacy Preservation in Mobile Environment. *IEEE Trans. Comput. Soc. Syst.* **2019**, *6*, 1407–1419. [CrossRef]