



Article

ICENETv2: A Fine-Grained River Ice Semantic Segmentation Network Based on UAV Images

Xiuwei Zhang ^{1,2,*} , Yang Zhou ^{1,2} , Jiaojiao Jin ^{1,2}, Yafei Wang ³, Minhao Fan ⁴, Ning Wang ⁵ and Yanning Zhang ^{1,2}

- ¹ School of Computer Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China; zy2019@mail.nwpu.edu.cn (Y.Z.); jinjiaojiao@mail.nwpu.edu.cn (J.J.); ynzhang@nwpu.edu.cn (Y.Z.)
- ² National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, Xi'an 710072, China
- ³ Ningxia–Inner Mongolia Hydrology and Water Resource Bureau, Baotou 014030, China; wangyafei1@swj.yrcc.gov.cn
- ⁴ Hydrology Bureau of the Yellow River Conservancy Commission, Zhengzhou 450004, China; fanminhao@swj.yrcc.gov.cn
- ⁵ School of Microelectronics, Xi'an Jiaotong University, Xi'an 710072, China; wangning07@stu.xjtu.edu.cn
- * Correspondence: xwzhang@nwpu.edu.cn

Abstract: Accurate ice segmentation is one of the most crucial techniques for intelligent ice monitoring. Compared with ice segmentation, it can provide more information for ice situation analysis, change trend prediction, and so on. Therefore, the study of ice segmentation has important practical significance. In this study, we focused on fine-grained river ice segmentation using unmanned aerial vehicle (UAV) images. This has the following difficulties: (1) The scale of river ice varies greatly in different images and even in the same image; (2) the same kind of river ice differs greatly in color, shape, texture, size, and so on; and (3) the appearances of different kinds of river ice sometimes appear similar due to the complex formation and change procedure. Therefore, to perform this study, the NWPU_YRCC2 dataset was built, in which all UAV images were collected in the Ningxia–Inner Mongolia reach of the Yellow River. Then, a novel semantic segmentation method based on deep convolution neural network, named ICENETv2, is proposed. To achieve multiscale accurate prediction, we design a multilevel features fusion framework, in which multi-scale high-level semantic features and lower-level finer features are effectively fused. Additionally, a dual attention module is adopted to highlight distinguishable characteristics, and a learnable up-sampling strategy is further used to improve the segmentation accuracy of the details. Experiments show that ICENETv2 achieves the state-of-the-art on the NWPU_YRCC2 dataset. Finally, our ICENETv2 is also applied to solve a realistic problem, calculating drift ice cover density, which is one of the most important factors to predict the freeze-up data of the river. The results demonstrate that the performance of ICENETv2 meets the actual application demand.

Keywords: fine-grained river ice; position attention; channel attention; drift ice cover density; semantic segmentation



Citation: Zhang, X.; Zhou, Y.; Jin, J.; Wang, Y.; Fan, M.; Wang, N.; Zhang, Y. ICENETv2: A Fine-Grained River Ice Semantic Segmentation Network Based on UAV Images. *Remote Sens.* **2021**, *13*, 633. <https://doi.org/10.3390/rs13040633>

Academic Editor: Hyungtae Lee
Received: 2 November 2020
Accepted: 5 February 2021
Published: 10 February 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 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 (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Ice caps, ice plugs, or ice dams are often formed in high latitude rivers in winter, which could change the hydraulic, thermal, and geometric boundary conditions of water flow and form a unique ice phenomenon in winter [1]. Ice plugs or ice dams, i.e., drift ice in the river channel blocking the cross section of water flow, may cause water level rise, inundate farmland houses, damage the coastal hydraulic structures, cause shipping interruption, or cause hydraulic power loss [2,3]. Therefore, river ice monitoring is necessary in preparing for potential hazards. Accurate fine-grained ice semantic segmentation is a key technology in the study of river ice monitoring, which can provide more information for ice situation analysis, change trend prediction, and so on.

Many efforts have been spent on studying river ice segmentation based on imaging monitoring. In terms of monitoring, there are typically three kinds, including satellite-based monitoring [4–7], shore-based terrestrial monitoring [8,9], and UAV (unmanned aerial vehicle)-based aerial monitoring [10–12]. The advantage of satellite-based monitoring is that its observation range is very large and it is not limited by national boundaries and geographical conditions. While, it is often hard or costly to realize real-time observation. The revisit period of general satellites like Landsat is often several days. Some satellites with short revisit periods are often expensive or have coarse resolution. Shore-based terrestrial monitoring can observe river ice at any time. But, installing surveillance equipment on the shore is often limited by geographical conditions, especially in mountainous regions. UAV-based aerial monitoring has the advantages of having a wide monitoring range, a fast response speed, and is uneasily disturbed by terrain. It has served as an important supplementary way to monitor river ice. Therefore, this paper focuses on the monitoring of river ice using UAV images.

In terms of ice segmentation methods, they are divided into three categories: Traditional threshold methods, traditional machine learning-based methods, and neural network-based methods. Note that, here we also consider the methods of sea ice segmentation, since they are similar to river ice segmentation to a certain extent.

Traditional threshold methods. The threshold-based segmentation methods use the difference in gray scale of different objects to be extracted from the image, and divide the pixels into several categories by setting appropriate threshold values to achieve the segmentation of different objects. Engram et al. [6] utilized a threshold approach to distinguish floating ice and bedfast ice using SAR images across Arctic Alaska. Beaton et al. [7] presented a river coverage segmentation method, in which a threshold technique was adopted to reduce the effectiveness of cloud obstruction and maximize river coverage.

Traditional machine learning-based methods. With the development of machine learning, many methods have been utilized to segment an ice region from images. They can be summarized into two categories, unsupervised and supervised methods. On the part of unsupervised methods, Ren et al. [13] presented a multi-stage method using k-means clustering for sea ice SAR image segmentation. Dang et al. [14] presented two methods for SAR sea ice image segmentation: The k-means clustering method and threshold-based segmentation method. The result showed that the k-means clustering method outperformed the threshold-based segmentation method. The former can gain a clear segmentation boundary and complete segmentation region. Chu and Lindenschmidt [4] classified the river covers into four categories, including smooth rubble ice, intact sheet ice, rough rubble ice, and open water, based on fuzzy k-means clustering with Moderate Resolution Imaging Spectroradiometer (MODIS) and RADARSAT-2 images. On the part of supervised methods, Zhang et al. [15] presented a CART decision tree method to retrieve sea ice from MODIS images in the Bohai Sea. Romanov [5] adopted decision tree to detect ice using AVHRR images.

Deep neural network-based methods. In recent years, we have witnessed important advances in image semantic segmentation based on deep neural networks. A new generation of algorithms based on FCN [16] keeps improving state-of-the-art performance on different benchmarks. Singh et al. [11] adopted several deep neural network-based semantic segmentation models to segment river ice images into anchor ice, frazil ice, and water, and achieved great outcomes. These models are mainly UNet [17], SegNet [18], and DeepLab [19]. In 2020, we [12] also designed a semantic segmentation deep convolution neural network, named ICENET, for river ice semantic segmentation. It revealed that applying deep convolutional neural networks into ice detection and fine-grained ice segmentation is promising.

In conclusion, the traditional threshold method directly utilizes the gray-scale differential characteristics of the images and has the advantages of simple calculation and high efficiency. However, an appropriate threshold is hard to be determined and is sensitive to many factors such as noise and brightness. Therefore, this kind of method usually has a

poor generalization ability. The traditional machine learning-based methods can achieve good results for tasks in simple scenes, but with images of complex scenes, they are still relatively simple and require manual intervention, which cannot guarantee the segmentation effect. With the popularity of deep learning, semantic segmentation has made great progress. Convolutional neural network-based semantic segmentation methods far exceed traditional methods, under its strong nonlinear fitting ability and learning ability. Hence, the deep neural network-based method is adopted in this paper.

In this paper, we study fine-grained river ice segmentation based on the deep neural network technique. This study is different from the methods mentioned above. It distinguishes shore ice, drift ice, water, and bank, which has an important application significant in river ice monitoring. Therefore, to study it, a UAV image dataset was built. By analysis, this fine-grained river ice segmentation has these characteristics: (1) The scale of river ice varies greatly, ranging from several pixels to thousands of pixels; (2) the appearances of river ice are diverse, even for the same kind; and (3) sometimes, drift ice and shore ice look similar, since they could become each other in some conditions. These characteristics will be analyzed in Section 2.2. Aiming towards these characteristics, we designed a novel semantic segmentation network structure, which effectively exploits the multilevel features fusion, dual attention module, and new up-sampling strategy to generate high-resolution predictions. Our main contributions are as follows:

- A UAV image dataset named NWPU_YRCC2 was built for fine-grained river ice semantic segmentation. All UAV images were collected in the Ningxia–Inner Mongolia reach of the Yellow River, since the ice phenomenon in this reach is very typical and diverse. The dataset consists of 1525 precisely labeled images covering typical river ice images with different characteristics;
- A novel network is proposed for fine-grained river ice segmentation, named ICENETv2. In this network, multiscale low-resolution semantic features and high-resolution finer features are effectively fused to generate different scale predictions, since the scale of river ice changes greatly even in the same image. Additionally, we adopt a dual attention module to highlight distinguishable features and use a learnable up-sampling strategy to improve the details of the segmentation and increase the semantic segmentation accuracy of fine-grained river ice;
- Compared with DeepLabV3 [20], PSPNet [21], RefineNet [22], and BiSeNet [23], our ICENETv2 has the state-of-the-art performance on the NWPU_YRCC2 dataset. Besides, our ICENETv2 is applied to solve a practical problem, i.e., calculating drift ice cover density, which is one of the most intuitive information for predicting the freeze-up date of a river. By using the predicted fine-grained river ice semantic segmentation map, the drift ice cover density error is only 5.6%. The results show that its performance meets the actual application demand.

2. Study Area and Materials

2.1. Study Area

We selected the Ningxia–Inner Mongolia reach of the Yellow River as our study area, which is indicated by the red ellipse in Figure 1. The Yellow River is one of the world's longest rivers, with a total length of 5464 km and a drainage area of 752,443 km². Since the middle section of the river runs through the Loess Plateau, it carries a large amount of sediment [24]. It spans 23 longitudes from east to west and 10 latitudes from north to south, hence there are considerable differences in elevation between the east and west, and in landforms among different regions. Since the basin is located in the middle latitude zone, the influence of atmospheric circulation and monsoon circulation is relatively complex [25]. Due to the complexity of climate and landform, the ice phenomenon is very typical and diverse in spring and winter, especially in the Ningxia–Inner Mongolia reach [26]. Therefore, this reach is selected so as to study fine-grained river ice segmentation.

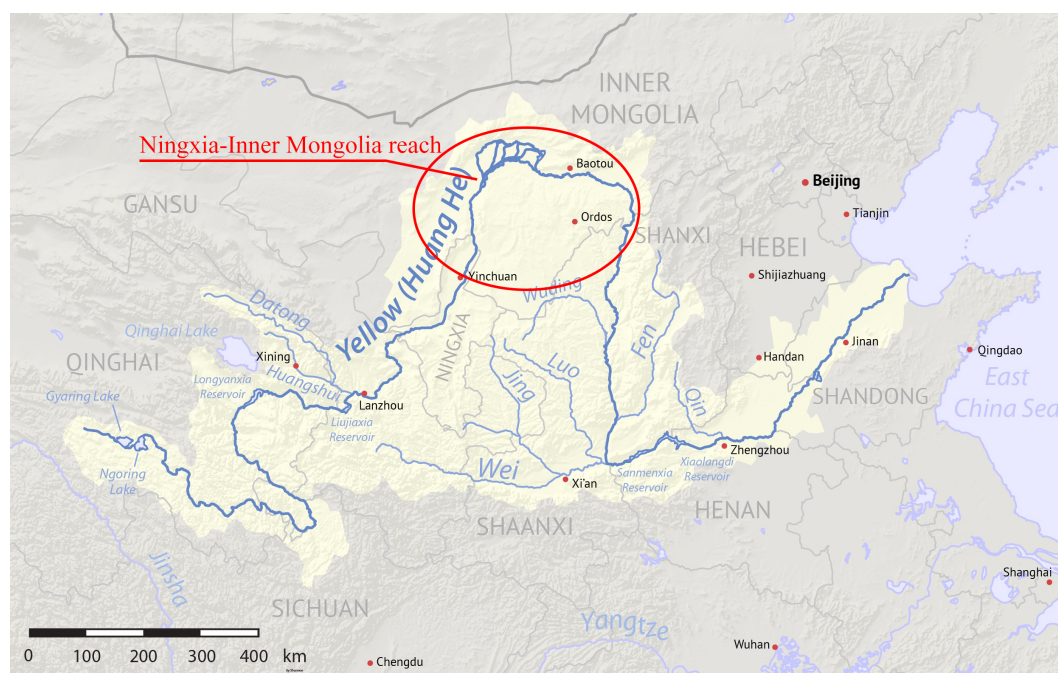


Figure 1. Our study area is indicated by the red ellipse. The base map is from the Yellow River entry in Wikipedia [27].

2.2. Dataset and Analysis

There are no suitable UAV image datasets for the fine-grained river ice segmentation of the Yellow River. Therefore, based on our previous NWPU_YRCC dataset [12], we further built the NWPU_YRCC2 dataset. The NWPU_YRCC2 dataset contains four categories: Shore ice, drift ice, water, and others. It is necessary to distinguish shore ice and drift ice, since the calculation of drift ice cover density, which is an important factor in the actual freeze-up date forecast, just considers the ratio of drift ice and excludes shore ice.

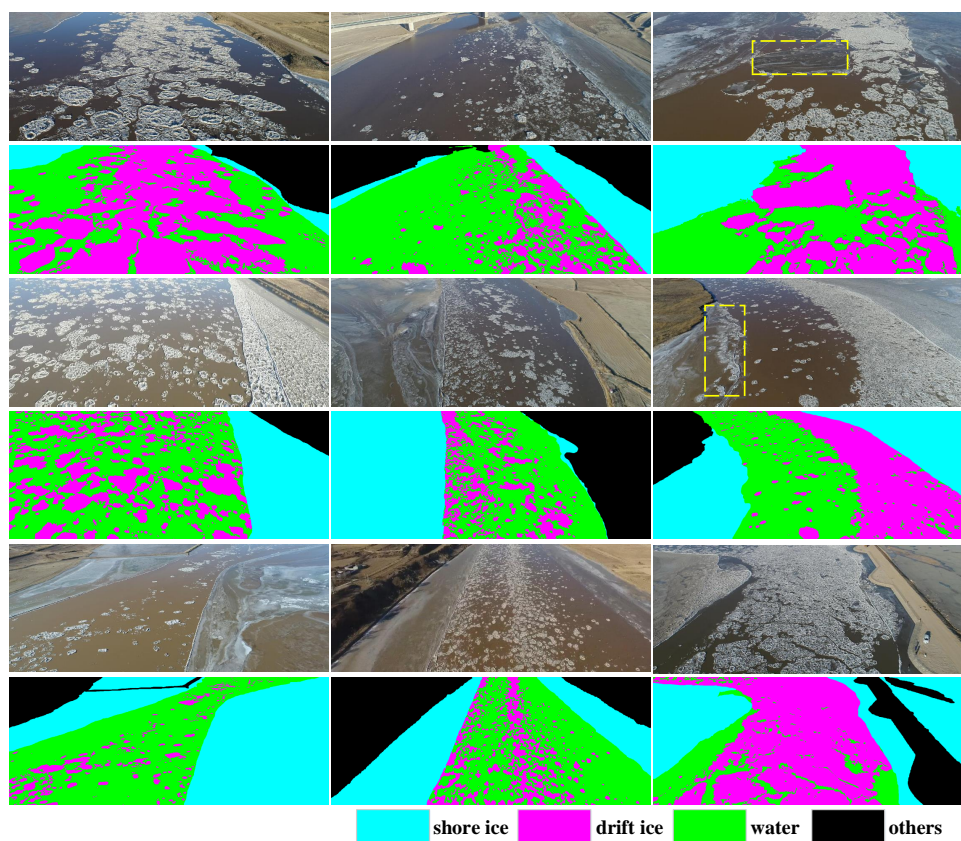
The dataset building process is basically similar to that of NWPU_YRCC [12]. The aerial images were taken annually from 2015 to 2019 at the Ningxia–Inner Mongolia reach of the Yellow River from November to March. During data collection, an ASN216 fixed-wing drone with a visible light camera Canon 5DS and a DJI Inspire 1 were used to capture images and videos, shown in Figure 2. The details of UAVs are shown in Table 1. The UAV images are taken in nadir view or oblique view during data capture. However, when calculating the drift ice cover density, the UAV images are required to be captured in nadir view. The flying height of the drones ranges from 30 m to 600 m. Finally, 200 videos were captured, ranging in length from 10 min to 50 min. The maximum image resolution and maximum video resolution of the Canon 5DS camera on ASN216 are 8688×5792 and 1920×1080 , respectively, while the maximum image resolution and maximum video resolution of the camera on DJI Inspire 1 are 4000×3000 and 4096×2160 , respectively. To keep the resolution of the input image of the model consistent, we resized the images to 1600×640 . From the videos and images obtained from aerial photography, 305 typical images containing four categories of targets are carefully selected. These images are mainly collected during the freeze-up period. We use Photoshop software to label each pixel of the images into four categories: Shore ice, drift ice, water, and others. The reason why we do not use other annotation tools such as Labelme and Image Labeler is that they are difficult to use for marking the boundary between these four categories. It is worth mentioning that this annotation work is very time-consuming. We divided 305 images into training set, validation set, and test set at a ratio of 6:2:2. Then, they are expanded to 1525 images to form the NWPU_YRCC2 dataset by data augmentation operations, including the brightness adjustment, flipping, and clipping. The brightness adjustment includes two ways, increasing brightness and decreasing brightness. For flipping, we used both horizontal and vertical flipping.

Table 1. Details of fixed-wing UAV (unmanned aerial vehicle) ASN216 and DJI Inspire 1.

Parameters	ASN216	DJI Inspire 1
Max take-off weight	30 kg	3.4 kg
Max speed	120 km/h	22 m/s
Sensor type	CMOS	Exmor R CMOS
Max image resolution	8688 × 5792	4000 × 3000
Max video resolution	1920 × 1080	4096 × 2160
Effective pixels	50.6 million pixels	12.4 million pixels

**Figure 2.** Photographs of fixed-wing UAV ASN216 and DJI Inspire 1.

Several typical images and their annotations are shown in Figure 3. Drift ice is ice that flows with water on or in water. The appearance of drift ice varies greatly in different images in regards to scale, color and texture, as shown in the first two rows of Figure 3. Sometimes, aggregated drift ice could crash shore ice and make part of them break and divorce from the original shore ice to form new drift ice, as shown in the right image of the first row.

**Figure 3.** Some typical river ice images and their corresponding annotation maps.

Shore ice is a zone of ice frozen along a river bank. According to its formation time and condition, it can be divided into different kinds, such as newborn shore ice, fixed shore ice, alluvial shore ice, and regenerative shore ice, and so on. In this paper, we do not distinguish these fine-grained shore ice, since it is not really necessary for actual applications. However the appearances of different kinds of shore ice differ greatly. The middle two rows present some typical images of shore ice and their annotation maps. From these images, shore ice exhibits a different appearance even in the same image. It leads to the difficulty of accurate fine-grained river ice semantic segmentation. It is especially to distinguish drift ice and shore ice when drift ice stops at the edge of shore ice and freezes up to become a new part of shore ice, shown in the yellow box of the right image.

The last two rows show some typical images with different drift ice cover density and their annotation maps. Therefore, the scale of drift ice changes very much, ranging from several pixels to thousands of pixels and even in the same image.

In brief, accurate fine-grained river ice segmentation is a hard task because it has the following difficulties: (1) The scale of river ice varies greatly in different images and even in the same image; (2) the same kind of river ice differs greatly in color, shape, texture, size, and so on; and (3) the appearances of different kinds of river ice appear similar sometimes, due to the complex formation and change procedure.

3. Proposed Method

We propose a novel semantic segmentation network, named ICENETv2, to deal with the challenges of fine-grained river ice segmentation. ICENETv2 is developed based on our previous ICENET and it differs from ICENET in four main respects. Aiming to extract more multi-scale high-level semantic information, the ratios of the feature maps of Res1, Res2, Res3, and Res4 to the input image are changed from (1/8, 1/16, 1/16, 1/16) to (1/4, 1/8, 1/16, 1/32), respectively. Both position attention and channel attention are adopted to highlight the distinguishable semantic features between drift ice and shore ice. A learnable up-sampling strategy is used to further reconstruct the finer information, since the appearance of drift ice is diverse and sometimes its scale is prone to be small. Finally, a joint loss function is utilized to sufficiently train the network.

In this section, firstly, the network architecture of our ICENETv2 is illustrated, as shown in Figure 4. Secondly, the four principle sub-modules of ICENETv2, namely attention module, fusion module, sub-pixel upsample module [28], and the loss function are presented in Sections 3.2–3.5 respectively.

3.1. Network Architecture

In fine-grained river ice semantic segmentation, the scales of the drift ice and shore ice vary diversely, ranging from several pixels to hundreds of pixels, even thousands of pixels. The effective fusion of multiscale features can significantly enhance the segmentation precision of multiscale targets. Therefore, inspired by BiSeNet, the proposed ICENETv2 also adopts a two-branch architecture, as shown in Figure 4. The deep branch aims to extract high-level semantic context features, in which a parallel dual attention mechanism combining channel attentive features and positional attentive features is adopted to extract more comprehensive and sufficient semantic features. The shallow branch extracts low-level features that can encode high-resolution finer spatial details. Finally, deep features and shallow features are sufficiently fused to generate prediction. Besides, instead to the commonly used bilinear interpolation, a learnable up-sampling strategy, i.e., sub-pixel method [28], is utilized to further refine the segmentation detail.

To be specific, the input image is fed into a convolution block, of which stride is 2 and kernel size is 7. Then, in order to acquire semantic context features and detailed finer features, the output features are respectively input into two branches. The deep branch is based on ResNet-101 and consists of four residual blocks, namely Res1, Res2, Res3, and Res4. The four residual blocks are original residual blocks in ResNet-101. The stride of Res1 is 1 and that of the other three blocks is 2. Res3 and Res4 are separately followed

by a parallel dual attention module, in which channel attention and positional attention are parallel imposed on the output of the residual block, then the two attentive features are combined by element-wise sum as the output. A global average pooling is performed on the output of Res4 to produce a global contextual vector. Then, the output features of dual attention module after Res4 are weighted by the global contextual vector and the weighted output is up-sampled twice by the sub-pixel method [28]. Finally, we concatenate the weighted features after up-sampling and output features of the dual attention module after Res3, then up-sample them twofold by sub-pixel to generate multiscale semantic features as the output of the deep branch.

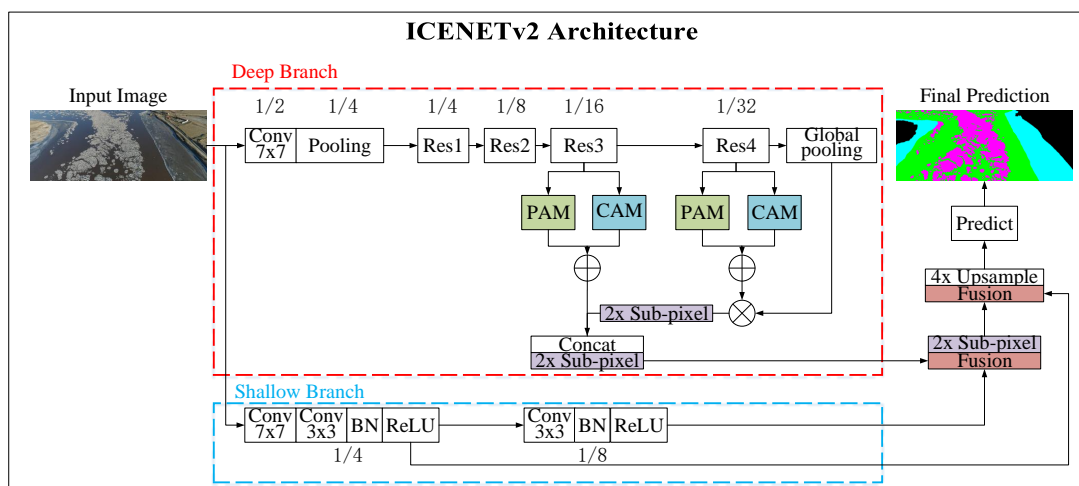


Figure 4. The architecture of ICENETv2. It mainly contains two branches: The deep branch in the red dashed box and the shallow branch in the blue dashed box. Res1, Res2, Res3, and Res4 are four original residual blocks from residual blocks in ResNet-101. ‘CAM’ and ‘PAM’ is the channel attention module and the position attention module, respectively. The notation ‘Sub-pixel’ means sub-pixel up-sampling module. ‘Fusion’ is the fusion module. The fractions in the two branches are the ratios of the feature map size to the resolution of the input.

The shallow branch only contains two convolution blocks, of which stride is 2 and kernel size is 3. Therefore, there are two outputs corresponding to the two convolution blocks. The size of the two output feature maps is one-quarter and one-eighth of the input image size, respectively. Since there are many small drift ice blocks in the problem of fine-grained river ice semantic segmentation, the shallow branch is designed to gain two different scales of high-resolution spatial information.

Finally, the fusion module shown in Figure 5 is adopted to integrate the two branch outputs to generate ultimate prediction. Firstly, the output feature maps of the deep branch and shallow one-eighth resolution output feature maps are fused and up-sampled twice by sub-pixel. Then, the former fusion result and the one-quarter resolution output feature maps of the shallow branch are fused again and 4X up-sampled by the bilinear interpolation, to achieve the final semantic segmentation prediction.

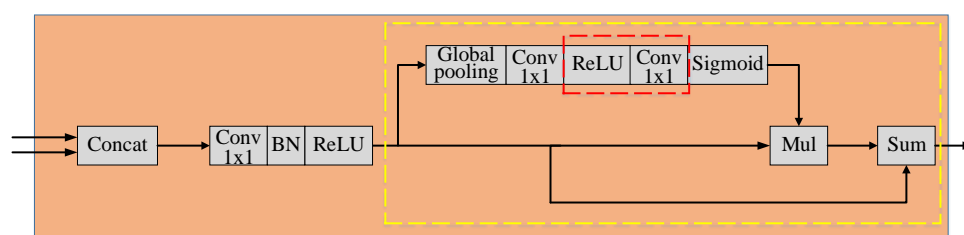


Figure 5. The fusion module.

3.2. Attention Model

It is a very common ice phenomenon that drift ice blocks may crash or rub with shore ice, some of which stop at the edge of the shore ice and freeze up to become a new part of shore ice. In this situation, it is not easy to distinguish drift ice and shore ice. Therefore, context information with spatial and semantic distinguishability is required. Inspired by DANet [29], we adopt a dual-attention module in the deep branch of our model, which contains a channel attention and a positional attention. The channel attention model can integrate the relevant features of all channels, so as to generate global association between channels and obtain stronger specific semantic response ability. The local correlation of spatial information can be learned by positional attention model, and the correlation between features of any position can be used to enhance the expression of their respective features. Then, the fusion of the above two kinds of enhanced attentive features will make the extracted features more effective to distinguish different kinds of objects.

Figure 6a shows the details of the channel attention module. Firstly, four operations, i.e., a global average pooling, a 1×1 convolution, a batch normalization, and a sigmoid function, are successively performed on the input feature map I to produce an attentive vector. Then, the attentive vector is used to weigh the input feature map, and the result is element-wise added with the input feature map to generate channel-wise attentive feature O .

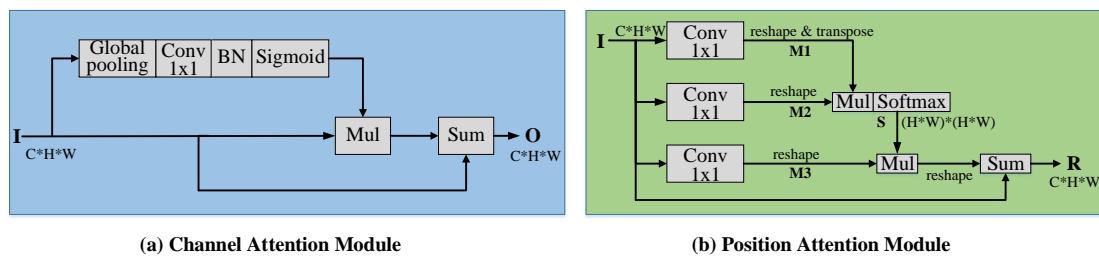


Figure 6. The attention module.

The structure of the position attention module is presented in Figure 6b. Given the input feature map $I \in \mathbb{R}^{C \times H \times W}$, I is input into three 1×1 convolution blocks to produce three feature maps $M1$, $M2$, and $M3$, where $\{M1, M2, M3\} \in \mathbb{R}^{C \times H \times W}$. Meanwhile, $M1$, $M2$, and $M3$ are reshaped to $\mathbb{R}^{C \times N}$, where $N = H \times W$. Subsequently, we multiply the transpose of $M1$ by $M2$, and utilize a softmax operation to produce a positional attention map $S \in \mathbb{R}^{N \times N}$. Then, we multiply the transpose of S by $M3$ and reshape the result to $\mathbb{R}^{C \times H \times W}$. Lastly, the feature $R \in \mathbb{R}^{C \times H \times W}$ is obtained by element-wise addition between the reshaped result and the feature map I .

3.3. Fusion Module

The outputs of the deep branch and shallow branch represent semantic context information and detailed information, respectively. Both of them can greatly help to process accurate segmentation. To combine these features of two different levels and play to their strengths, an effective feature fusion strategy is required. The feature fusion module in [23] is adopted, as shown in Figure 5. Firstly, we concatenate the output features of the deep branch and the shallow branch, then a 1×1 convolution, a batch normalization, and a ReLu function is performed on the concatenated feature maps successively. Finally, a sub-module shown in the yellow dotted box, which looks alike to the channel attention module, is calculated to produce the fusion result. The only differences between the sub-module and channel attention module are that batch normalization in the channel attention module is replaced by a ReLu function and a 1×1 convolution.

3.4. Sub-Pixel Up-Sampling Module

Sub-pixel up-sampling derives from super resolution research [28] and has been used in semantic segmentation and other tasks. As shown in Figure 7, assuming that the input is

a low-resolution image or feature map, the sub-pixel up-sampling is to extract the features from the input and finally fuse the extracted features to generate high-resolution (HR) images. Three convolution layers are adopted in detail. After that, the feature maps of channel number r^2 are obtained, in which r is the magnification factor. Then, through the sub-pixel convolution layer, r^2 channels of each pixel are transformed to a sub-pixel block with size $r \times r$ in the HR image. Finally, the feature map of $H \times W \times r^2$ is reshaped to a HR map of $rH \times rW \times 1$. To upsample feature maps with more than one channel, sub-pixel up-sampling strategy is applied to each channel.

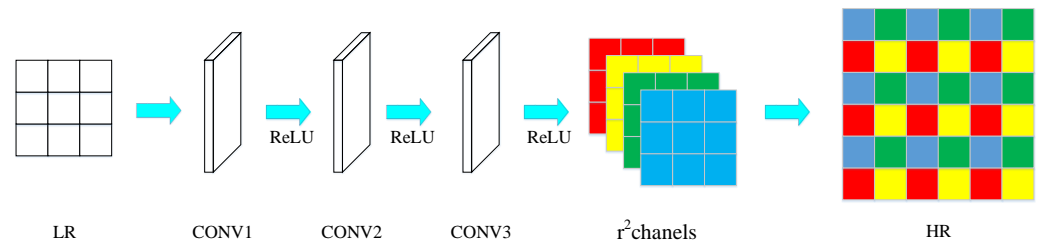


Figure 7. The sub-pixel up-sampling module.

3.5. Loss Function

In our model, auxiliary losses are adopted to supervise the training of the proposed network. The main loss is used to supervise the final prediction of the whole ICENETv2. Three particular auxiliary losses are utilized to supervise the outputs of two dual attention modules respectively after Res3 and Res4 and the output of the first fusion model. Categorical cross-entropy loss is adopted as the loss function for these four losses, as shown in Equation (1). Meanwhile, parameter α is adopted to balance the weight of the main loss and three auxiliary losses, as presented in Equation (2). The three auxiliary losses share a weight α , since the orders of magnitude of them are the same and they are all utilized to supervise intermediate feature maps without significant difference in importance. We conducted an ablation experiment on three auxiliary losses and an experiment to find a reasonable parameter α . The details refer to Section 4.3. Finally, α is set to 1 in our subsequent experiments. The joint loss allows the optimizer to optimize the model more conveniently.

$$L_{CE} = - \sum_{i=1}^n t_i \log(p_i) \quad (1)$$

where n indicates n classes, t_i is the truth label and p_i is the Softmax probability for the i th class.

$$L_{joint} = L_{CE_m} + \alpha \sum_{i=1}^3 L_{CE_i} \quad (2)$$

where α denotes the hyper-parameter of controlling the relative importance of the three auxiliary loss terms. L_{CE_m} is the main loss function. L_{CE_1} and L_{CE_2} are used to supervise the outputs of two dual attention modules after Res3 or Res4, respectively. L_{CE_3} is used to optimize the output of the first fusion. L_{CE_m} , L_{CE_1} , L_{CE_2} , and L_{CE_3} are calculated by Equation (1). L_{joint} is the joint loss function.

4. Experiments

In this section, firstly, the implementation details of the experiments are described. Secondly, the ablation experimental results performed on NWPU_YRCC2 are illustrated and analyzed. Thirdly, our ICENETv2 model is compared with some state-of-the-art methods. Finally, our model is utilized to solve an actual application, calculating the drift ice cover density.

4.1. Implementation Details

Our models are implemented based on Pytorch. In the training procedure, mini-batch stochastic gradient descent [30] is adopted, the momentum is initially set to 0.9, and decays by weight 1.83×10^{-2} with a batch size of 4. The training time is set to 200 epochs. The overall training time for 200 epochs takes about 30 h on two NVIDIA GeForce GTX 2080Ti cards and the training time per epoch of our model is about 9 minutes. After about 120 trained epochs, our model converges. Our NWPU_YRCC2 dataset contains 1525 annotated images. In the experiments, it is divided into a training set, validation set, and test set at a ratio of 6:2:2, then the segmentation results were colored for better visualization. The NWPU_YRCC dataset contains a total of 814 annotated images, including 570 images for training, 82 images for verification, and 244 images for testing. Note that the test set includes the verification set.

4.2. Evaluation

Pixel accuracy and mean intersection over union are usually adopted to measure the accuracy of semantic segmentation. Therefore, we used these two indicators to calculate and compare the performances of different methods. Assume that there are totally $n + 1$ categories (from 0 to n) and class 0 indicates a void class or background. p_{ij} is the amount of pixels that belong to class i but were predicted to be class j . p_{ii} is the amount of pixels of class i and were predicted to be class i . When $i \neq j$ and class i is regarded as positive, p_{ji} and p_{ij} is the number of false positives and false negatives, respectively. PA (pixel accuracy) and MIOU (mean intersection over the union) can be described as follows.

Pixel accuracy (PA). PA, described in Equations (3), is the ratio of the number of correctly classified pixels to the overall amount of pixels:

$$PA = \frac{\sum_{i=0}^n p_{ii}}{\sum_{i=0}^n \sum_{j=0}^n p_{ij}}. \quad (3)$$

Mean intersection over the union (MIOU). IoU is the ratio of the intersection area of the predicted segmentation with the ground truth to the union area of them for one particular class. MIOU is often adopted to measure the accuracy for the segmentation of more than one class. It denotes the average value of the IoUs of all classes. The specific description is shown in Equation (4):

$$MIOU = \frac{1}{n+1} \sum_{i=0}^n \frac{p_{ii}}{\sum_{j=0}^n p_{ij} + \sum_{j=0}^n p_{ji} - p_{ii}}. \quad (4)$$

4.3. Ablation Study

To verify the effect and contribution of three auxiliary losses, we conduct an ablation experiment on them as shown in Table 2. The network is trained and optimized by one main loss function and two selected auxiliary loss functions at a time. The results show that these three auxiliary losses all contribute to the final semantic segmentation and there is no significant difference in importance among them. These three auxiliary losses share a weight α . To find a reasonable parameter α , we also conducted an experiment shown in Table 3. It shows that when $\alpha = 1$, the method performs best. Therefore, we chose $\alpha = 1$ in our subsequent experiments.

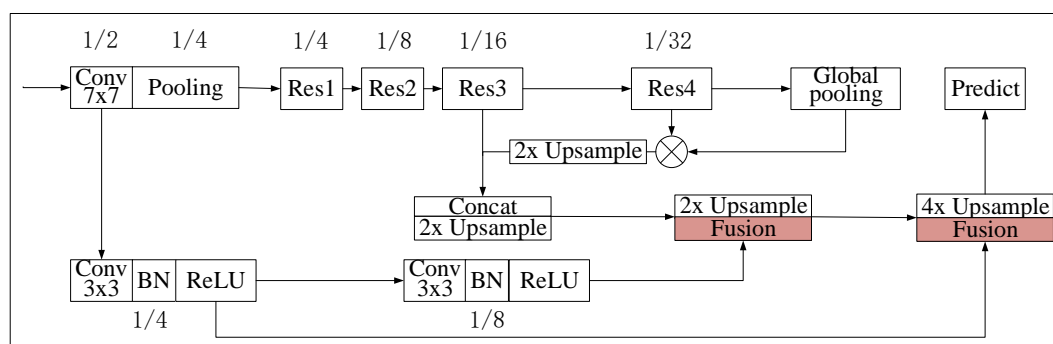
Table 2. Experiments on the effects of three auxiliary losses on the attention module. PA: Pixel accuracy. MIoU: Mean intersection over the union.

L_{CE_1}	L_{CE_2}	L_{CE_3}	IoU(%)				MIoU(%)	PA(%)
			Drift Ice	Shore Ice	Water	Others		
	✓	✓	79.889	83.742	89.255	77.686	82.643	91.645
✓		✓	80.983	82.100	89.318	78.753	82.789	91.634
✓	✓		81.868	83.401	89.538	76.951	82.940	91.846
✓	✓	✓	81.127	81.582	90.484	80.548	83.435	91.943

Table 3. The influence of different auxiliary loss functions weight α on the segmentation results.

α	IoU(%)				MIoU(%)	PA(%)
	Drift Ice	Shore Ice	Water	Others		
0.1	79.192	81.317	89.671	80.358	82.635	91.463
0.2	75.293	80.364	90.781	81.855	82.073	91.055
0.4	80.422	81.869	88.639	81.642	83.143	91.533
0.8	79.549	80.910	88.691	82.776	82.982	91.337
1	81.127	81.582	90.484	80.548	83.435	91.943

To verify the effects of the proposed model and their three principle sub-modules, we conducted another ablation experiment successively adding the channel attention, the position attention, sub-pixel up-sampling, and auxiliary loss to the baseline model. Table 4 shows the results of these ablation experiments. The first row shows the results of the baseline model, which is depicted in Figure 8. Compared with ICENETv2, the baseline does not contain the dual attention model and sub-pixel up-sampling model, and its loss only considers the main loss. In Table 4, ‘CAM’ and ‘PAM’ respectively denotes adding channel attention and position attention after the block Res3 and Res4. ‘CAM + PAM’ means adding both attention modules after the block Res3 and Res4. The notation ‘Sub-pixel’ means that part of up-sampling operations are sub-pixel up-sampling shown in Figure 4. The notation ‘Au_loss’ means that three auxiliary losses are also considered in the loss function.

**Figure 8.** The baseline model.

The experimental results show that MIoU has increased by 2.674% by adding only the position attention module in the baseline model and 3.760% by adding only the channel attention module. When the two attention modules are adopted simultaneously in the baseline model, the MIoU is increased by 4.249%. On this basis, sub-pixel and Au_loss modules are added respectively, and MIoU is improved by 2.449% and 1.922% respectively. When all the sub-modules of the ablation experiment are added to the baseline module, the MIoU achieves 83.435%.

Moreover, Figure 9 shows some visualization results to demonstrate the effectiveness of the two attention modules. By adding the position attention module or channel attention

module to the baseline model, the fine-grained semantic segmentation of drift ice is more accurate and some details and object boundaries are clearer.

Table 4. Ablations on the channel attention module, the position attention module, the auxiliary loss, and sub-pixel up-sampling module.

Baseline	Cam	Pam	Sub-Pixel	Au_Loss	IoU(%)				MIoU(%)	PA(%)
					Drift Ice	Shore Ice	Water	Others		
✓					74.496	78.996	86.413	64.940	76.211	88.594
✓	✓				77.323	76.461	87.674	78.425	79.971	89.776
✓		✓			73.630	82.086	83.550	76.274	78.885	88.933
✓	✓	✓			77.263	81.780	88.034	74.762	80.460	90.463
✓	✓	✓	✓		79.085	82.404	89.254	80.893	82.909	91.529
✓	✓	✓		✓	78.381	82.138	89.720	79.290	82.382	91.328
✓	✓	✓	✓	✓	81.127	81.582	90.484	80.548	83.435	91.943

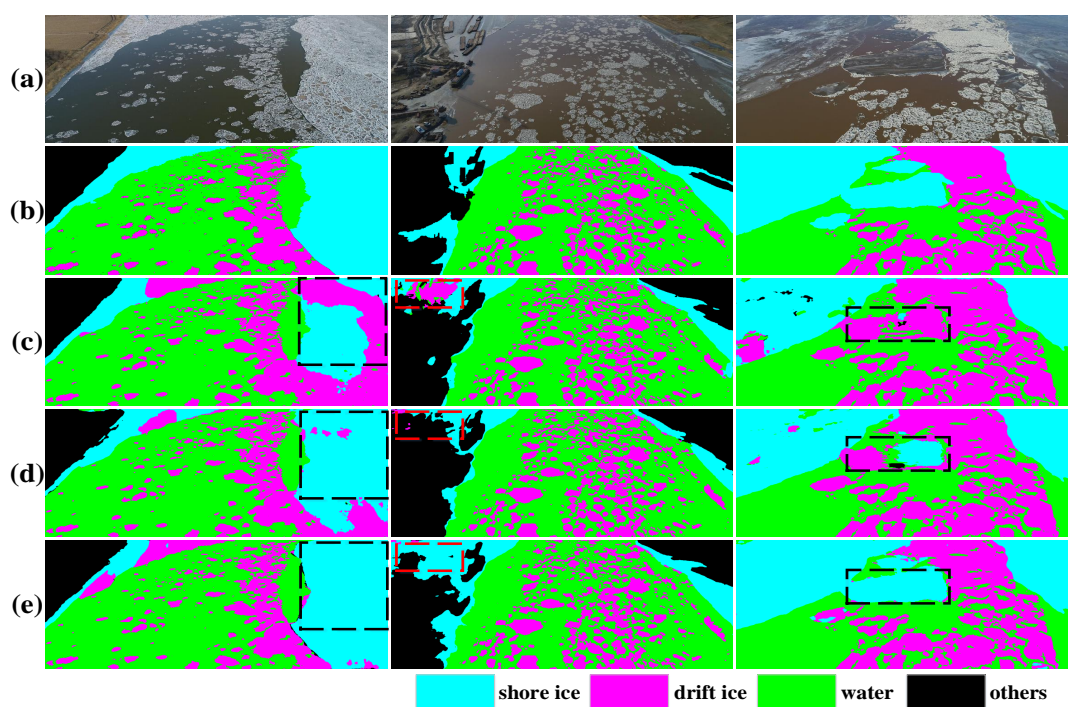


Figure 9. Visualization results of ablation study. (a,b) respectively represents the images and the corresponding labels. (c) represents the results obtained through the baseline model. (d) is the results of adding both PAM and CAM modules. (e) represents the prediction obtained by adding all sub-modules of the ablation experiment.

4.4. Comparison with the State-of-the-Art

The methods DeepLabV3 [20], DenseASPP[31], PSPNet [21], RefineNet [22], and BiSeNet [23] have achieved excellent performance on public datasets. They are the representative semantic segmentation methods of the state-of-the-art. Therefore, we compare the proposed ICENETv2 with them on NWPU_YRCC2 and NWPU_YRCC. The comparison results are presented in Table 5. They demonstrate that the ICENETv2 has a significant improvement in terms of MIoU on NWPU_YRCC2, meanwhile the single-class in terms of IoU for the drift ice is also the highest. Figure 10 presents some visualization comparison results. We can see that the results of the ICENETv2 have a better segmentation and recognition accuracy of small-scale targets. The experimental results also show that the performance of ICENETv2 on NWPU_YRCC is significantly higher than that of other

methods and slightly higher than that of ICENET. This can further verify the effectiveness of ICENETv2 on fine-grained segmentation.

Table 5. Comparison of our ICENETv2 with other methods.

Method	Dataset	IoU(%)				MIoU(%)	PA(%)
		Ice		Water	Others		
		Drift Ice	Shore Ice				
DeepLabV3 [20]	NWPU_YRCC2	62.127	69.938	77.363	61.020	67.612	85.902
DenseASPP [31]		68.924	75.479	81.628	66.522	73.138	88.769
PSPNet [21]		73.875	76.221	85.855	75.261	77.803	90.626
RefineNet [22]		74.442	81.125	87.116	75.427	79.528	91.634
BiseNet [23]		72.623	84.026	87.282	76.441	80.093	91.607
ICENET [12]		74.452	82.799	87.865	77.053	80.542	91.844
ICENETv2		81.127	81.582	90.484	80.548	83.435	91.943
DeepLabV3 [20]	NWPU_YRCC [12]		84.537	76.941	79.028	80.024	92.108
DenseASPP [31]			87.716	80.064	83.798	83.630	93.934
PSPNet [21]			88.196	81.483	83.774	84.374	93.966
RefineNet [22]			88.483	82.970	84.733	85.371	94.312
BiseNet [23]			89.301	83.464	87.814	86.497	95.058
ICENET [12]			91.583	84.891	88.253	88.112	95.932
ICENETv2			90.911	86.101	90.365	88.506	94.542

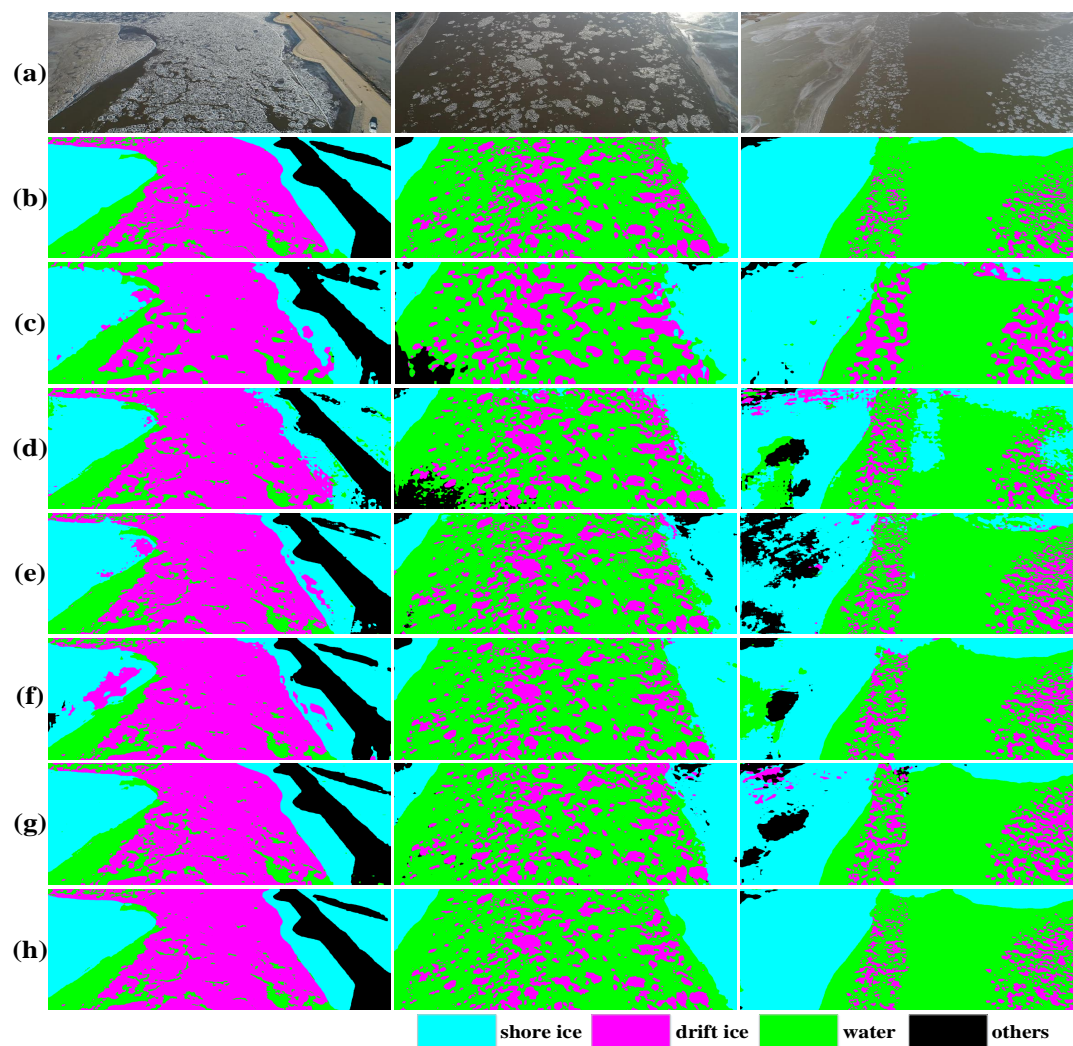


Figure 10. Comparison of different methods on the NWPU_YRCC2 dataset. (a,b) respectively represent the images and the corresponding labels. (c–h) respectively represent the results of algorithm DeepLabV3, DenseASPP, PSPNet, RefineNet, BiseNet, and ICENETv2 on the NWPU_YRCC2 dataset, arranged in ascending order by MIoU in Table 5.

4.5. Application on the Calculation of Drift Ice Cover Density

Drift ice cover density is one of the most important factors in predicting the freeze-up date of river and can provide more information for ice situation analysis. Now manual visual measurement is still adopted to calculate the drift ice cover density in many hydrological stations. This measurement way is greatly affected by human experience, and is usually prone to error. Based on the predicted fine-grained river ice semantic segmentation map, the drift ice cover density can be calculated automatically, as shown in Equation (5). The Drift_Ice_Num and River_Water_Num in Equation (5) represents the pixel number of water and drift ice respectively, in the predicted fine-grained river ice semantic segmentation map:

$$\text{Drift_Ice_Cover_Density} = \frac{\text{Drift_Ice_Num}}{\text{Drift_Ice_Num} + \text{River_Water_Num}}. \quad (5)$$

To accurately calculate the drift ice cover density, the UAV lens should be perpendicular to the river surface. At the same time, the bank on both side of the river should be included in the shooting process, so that the bank can be viewed. We selected five typical scenes to verify our calculation, as shown in Figure 11. The error between the drift ice cover density calculated by the predicted fine-grained river ice semantic segmentation map and that by the label is only 5.6%. This demonstrates that our method is accurate enough to meet actual application requirements.

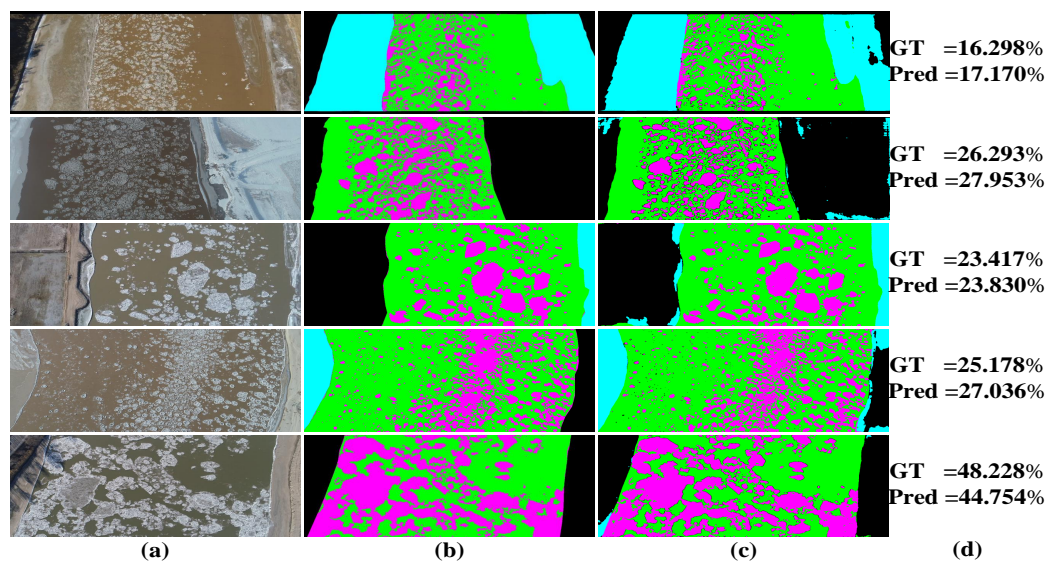


Figure 11. Experimental results of drift ice cover density. (a–c) respectively represent the images, the corresponding labels, and the prediction semantic segmentation maps. (d) shows the drift ice cover density calculated from the label and predicted semantic segmentation map.

4.6. Discussion

From the ablation experiments, it can be seen that ICENETv2 is 7.224% higher on MIoU than the baseline model and each sub-module has a certain contribution. The visualization results shown in Figure 8 also exhibit the improvement effect on the segmentation detail of the attention module and the sub-pixel up-sampling. By adding these sub-modules mentioned above to the baseline model, our ICENETv2 achieved the highest IoU in the drift ice category, reaching 81.127% IoU, indicating that our model is suitable for fine-grained segmentation.

The proposed ICENETv2 is compared on our NWPU_YRCC2 dataset with the state-of-the-art methods, including DeepLabV3 [20], DenseASPP [31], PSPNet [21], RefineNet [22], and BiSeNet [23]. The experimental results indicate that the proposed method achieves

significant improvements over the state-of-the-art methods in terms of mean IoU. Although our method is inspired by BiSeNet to some extent, it is carefully designed to adapt the characteristics of fine-grained segmentation. It adopts a new fusion structure to effectively fuse high-level semantic information and low-level finer information, and utilizes dual-attention [29] to highlight the distinguishable semantic features between drift ice and shore ice. A learnable up-sampling strategy [28] is used to further reconstruct the finer information, since the appearance of drift ice is diverse and sometimes its scale is prone to be small. In addition it has a joint loss function with three auxiliary losses to sufficiently train the network. These three parts are different from BiSeNet. The experimental results on NWPU_YRCC2 can demonstrate our design is effective to the problem of fine-grained ice segmentation. To further illustrate this conclusion, we also compare ICENET and ICENETv2 on both NWPU_YRCC2 and NWPU_YRCC, shown in Table 5. The performance of ICENETv2 on NWPU_YRCC is slightly higher than that of ICENET, reaching 88.506% MIOU. While, the performance of ICENETv2 on NWPU_YRCC2 is 2.893% higher on MIOU than that of ICENET. This comparison can further verify the effectiveness of our design on fine-grained segmentation.

From the application experiment, we can see that the accuracy of drift ice cover density calculated by the predicted semantic segmentation map is sufficient to meet the requirements of practical applications. This application of semantic segmentation to the calculation of drift ice cover density is very significant and innovative, since the manual visual measurement still adopted by many hydrological stations is inaccurate and error-prone. In the future, we will verify its application in more scenes and further improve the MIOU of the model. Moreover, we will focus on optimizing the code, reducing the computational complexity of their model, and improving the efficiency and speed of semantic segmentation, so that our proposed method can be trained more quickly and be run in lightweight portable computing devices.

5. Conclusions

In this research, a UAV visible image dataset named NWPU_YRCC2 was built for fine-grained river ice semantic segmentation. All images were collected in the Ningxia–Inner Mongolia reach of the Yellow River, since the ice phenomenon in this reach is very typical and diverse. Then, a novel network architecture named ICENETv2 was proposed for accurate fine-grained river ice semantic segmentation, which could efficiently fuse multiscale high-level semantic context features and low-level finer features. The experiments show that our ICENETv2 outperformed other methods on the NWPU_YRCC2 dataset. Furthermore, by using the predicted fine-grained river ice semantic segmentation map, the drift ice cover density could be calculated. Its error was only 5.6%, which is accurate enough to meet actual application requirements.

Author Contributions: The main idea was proposed by X.Z., J.J. and Y.Z. (Yanning Zhang). The dataset was captured, labeled, and analyzed by Y.W., N.W., M.F., J.J. and Y.Z. (Yang Zhou). The experiments were designed and carried out by Y.Z. (Yang Zhou), J.J. and N.W. The manuscript was written by X.Z., Y.Z. (Yang Zhou) and revised by Y.Z. (Yanning Zhang) and M.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (grant numbers 61971356, 61801395, 61971273, and 62071384) and the Natural Science Foundation of Shaanxi province (2020GM-137).

Acknowledgments: This research was supported by the National Natural Science Foundation of China (grant numbers 61971356, 61801395, 61971273, and 62071384) and the Natural Science Foundation of Shaanxi province (2020GM-137).

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

References

1. Agafonova, S.; Frolova, N.; Krylenko, I.; Sazonov, A.; Golovlyov, P. Dangerous ice phenomena on the lowland rivers of European Russia. *Nat. Hazards* **2017**, *88*, 171–188. [\[CrossRef\]](#)
2. Jia, H.; Chen, F.; Pan, D. Disaster Chain Analysis of Avalanche and Landslide and the River Blocking Dam of the Yarlung Zangbo River in Milin County of Tibet on 17 and 29 October 2018. *Int. J. Environ. Res. Public Health* **2019**, *16*, 4707. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Graf, R. Estimation of the Dependence of Ice Phenomena Trends on Air and Water Temperature in River. *Water* **2020**, *12*, 3494. [\[CrossRef\]](#)
4. Chu, T.; Lindenschmidt, K.E. Integration of space-borne and air-borne data in monitoring river ice processes in the Slave River, Canada. *Remote Sens. Environ.* **2016**, *181*, 65–81. [\[CrossRef\]](#)
5. Romanov, P. Global multisensor automated satellite-based snow and ice mapping system (GMASI) for cryosphere monitoring. *Remote Sens. Environ.* **2017**, *196*, 42–55. [\[CrossRef\]](#)
6. Ingram, M.; Arp, C.D.; Jones, B.M.; Ajadi, O.A.; Meyer, F.J. Analyzing floating and bedfast lake ice regimes across Arctic Alaska using 25 years of space-borne SAR imagery. *Remote Sens. Environ.* **2018**, *209*, 660–676. [\[CrossRef\]](#)
7. Beaton, A.; Whaley, R.; Corston, K.; Kenny, F. Identifying historic river ice breakup timing using MODIS and Google Earth Engine in support of operational flood monitoring in Northern Ontario. *Remote Sens. Environ.* **2019**, *224*, 352–364. [\[CrossRef\]](#)
8. Bourgault, D. Shore-based photogrammetry of river ice. *Can. J. Civ. Eng.* **2008**, *35*, 80–86. [\[CrossRef\]](#)
9. Ansari, S.; Rennie, C.; Seidou, O.; Malenchak, J.; Zare, S. Automated monitoring of river ice processes using shore-based imagery. *Cold Reg. Sci. Technol.* **2017**, *142*, 1–16. [\[CrossRef\]](#)
10. Kalke, H.; Loewen, M. Support vector machine learning applied to digital images of river ice conditions. *Cold Reg. Sci. Technol.* **2018**, *155*, 225–236. [\[CrossRef\]](#)
11. Singh, A.; Kalke, H.; Ray, N.; Loewen, M. River Ice Segmentation with Deep Learning. *arXiv* **2019**, arXiv:1901.04412.
12. Zhang, X.; Jin, J.; Lan, Z.; Li, C.; Fan, M.; Wang, Y.; Yu, X.; Zhang, Y. ICENET: A Semantic Segmentation Deep Network for River Ice by Fusing Positional and Channel-Wise Attentive Features. *Remote Sens.* **2020**, *12*, 221. [\[CrossRef\]](#)
13. Ren, J.; Hwang, B.; Murray, P.; Sakhalkar, S.; McCormack, S. Effective SAR sea ice image segmentation and touch floe separation using a combined multi-stage approach. In Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015; pp. 1040–1043.
14. Dang, X.; Wu, Y.; Fan, W.; Zhang, S. Discussion on sea ice segmentation of high resolution radar data. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 4861–4864.
15. Zhang, N.; Wu, Y.; Zhang, Q. Detection of sea ice in sediment laden water using MODIS in the Bohai Sea: A CART decision tree method. *Int. J. Remote Sens.* **2015**, *36*, 1661–1674. [\[CrossRef\]](#)
16. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Conf. Comput. Vis. Pattern Recognit.* **2015**, 3431–3440. doi:10.1109/LGRS.2018.2795531. [\[CrossRef\]](#)
17. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
18. 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. [\[CrossRef\]](#) [\[PubMed\]](#)
19. 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.* **2017**, *40*, 834–848. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
21. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
22. Lin, G.; Milan, A.; Shen, C.; Reid, I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1925–1934.
23. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 325–341.
24. Fu, C.; Popescu, I.; Wang, C.; Mynett, A.; Zhang, F. Challenges in modelling river flow and ice regime on the Ningxia-Inner Mongolia reach of the Yellow River, China. *Hydrol. Earth Syst. Sci.* **2014**, *18*, 1225–1237. [\[CrossRef\]](#)
25. Luo, D. Risk evaluation of ice-jam disasters using gray systems theory: The case of Ningxia-Inner Mongolia reaches of the Yellow River. *Nat. Hazards* **2014**, *71*, 1419–1431. [\[CrossRef\]](#)
26. Wu, C.G.; Wei, Y.M.; Jin, J.L.; Huang, Q.; Zhou, Y.L.; Liu, L. Comprehensive evaluation of ice disaster risk of the Ningxia-Inner Mongolia Reach in the upper Yellow River. *Nat. Hazards* **2015**, *75*, 179–197. [\[CrossRef\]](#)
27. Wikipedia Contributors. Yellow River—Wikipedia, The Free Encyclopedia. 2020. Available online: <https://en.wikipedia.org/wiki/File:Yellowrivermap.jpg> (accessed on 1 January 2021).
28. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883.

-
29. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3146–3154.
 30. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [[CrossRef](#)]
 31. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. Denseaspp for semantic segmentation in street scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3684–3692.