

Article

Deep-Learning-Based Stereo Matching of Long-Distance Sea Surface Images for Sea Level Monitoring Systems

Ying Yang ^{1,2} , Cunwei Lu ² and Zhenhua Li ^{1,*}¹ School of Physics, Nanjing University of Science and Technology (NJUST), Nanjing 210092, China; yang@bene.fit.ac.jp² Information Electronics Engineering, Fukuoka Institute of Technology, Fukuoka 811-0295, Japan; lu@fit.ac.jp

* Correspondence: lizhenhua@njust.edu.cn

Abstract: Due to the advantages of coastal areas in the fields of agriculture, transport, and fishing, increasingly more people are moving to these areas. Sea level information is important for these people to survive after extreme sea level events. With the recent improvements in computing and storage capacities, image analysis as a new measuring method is being rapidly developed and widely applied. In this paper, a multi-camera-based sea level height measuring system was built along Japan's coast and a deep-learning-based stereo matching method has been proposed for this system to complete 3D measurements. In this system, cameras are set with long base distances to ensure the long-distance monitoring system's precision, which causes a huge difference between the fields of view of the left and right cameras. Since most common network structures complete stereo matching by depth-wise cross-correlation between left and right images, they rely too much on the high-quality rectification of two images and fail on our long-distance sea surface images. We established a feature detection and matching network to realize sea wave extraction and sparse stereo matching for the system. Based on our previous result using the traditional method, the initial disparity was computed to reduce the search range of stereo matching. A training set with 785 pairs of sea surface images and 10,172 pairs of well-matched sea wave images was constructed to supervise the network. The experimental results verified that the proposed method can realize sea wave extraction and mask generation. It can also realize sparse matching of sea surface images regardless of poor rectification.

Keywords: image measurement; deep learning; stereo matching; sea surface

Citation: Yang, Y.; Lu, C.; Li, Z. Deep-Learning-Based Stereo Matching of Long-Distance Sea Surface Images for Sea Level Monitoring Systems. *J. Mar. Sci. Eng.* **2024**, *12*, 961. <https://doi.org/10.3390/jmse12060961>

Academic Editor: Sergei Chernyi

Received: 19 April 2024

Revised: 28 May 2024

Accepted: 29 May 2024

Published: 7 June 2024



Copyright: © 2024 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

Recently, with billions in investments, coastal areas are developing fast, and sea level information has become increasingly important for coastal societies [1,2]. Over time, the sea level changes on different time scales from hours to centuries, which is caused by different factors such as tides, air pressure, wind, etc. [3]. In the spatial domain, due to some local factors, such as latitude, sea water volume, etc., the changes are different. Existing measurement systems rely on the installation of measuring devices at fixed locations for long-term or short-term observation and measurement [4–6], which is expensive for long-term maintenance. Furthermore, their observation and measurement coverage are limited to the vicinity of the installation. Thus, a system that can measure over a wide range stably is urgently required. In this paper, a binocular-vision-principle-based image measurement system is introduced for sea level height monitoring; the non-contact scanning measurements conducted by the system extend the monitoring range while reducing maintenance costs and improving its own survival rate after extreme events. The stereo matching method, as the most critical technical issue in this system, is the focus of this paper.

Traditional stereo matching is usually divided into four major steps: cost calculation, cost aggregation, disparity calculation, and disparity optimization. Recently, with the improvement in stereo benchmarks [7,8] and computing capacities [9], convolution neural networks (CNNs) have been applied to replace some parts of the stereo matching

pipeline [10–12] or the whole pipeline [13–15]. However, the neural-network-based stereo matching method also suffers from two kinds of uncertainties, namely aleatoric and epistemic uncertainties [16]. Aleatoric uncertainty is related to the input data, indicating regions that may be hard to match. Epistemic uncertainty captures the uncertainty in the model, and is suitable for identifying out-of-distribution data when the variations in the test domain are too complex to be covered in the training domain. To reduce these uncertainties, a two-step training strategy is employed in most learning methods [17,18]: pre-training on mature stereo benchmarks and fine-tuning on the target datasets. In this paper, we also use this training strategy to train our network. There are several problems we were faced with.

Firstly, long-distance and large-scale shooting causes disparities, varying by a wide range within a stereo image pair. Secondly, sea water is constantly in motion; left and right cameras that are milliseconds out of sync may result in an offset between the left and right stereo images, making it difficult to rectify the stereo image pair. Almost all existing neural-network-based stereo matching algorithms employ a horizontal cost-generation module to generate cost volumes by different well-designed similarity calculation formulas. This makes horizontal similarity the only benchmark for stereo matching, and this will undoubtedly worsen the stereo matching results when there is ineradicable uncertainty offset between the left and right images. Thus, dense stereo matching based on the cost volume was discarded, and sparse stereo matching based on sparse local key points was chosen.

In this paper, we adopted a coarse-to-fine matching manner. This is a two-step mechanism with sea wave detection and matching modules to locate the sea waves, and takes them as key points to realize sparse matching. A new sea wave training dataset with 785 pairs of sea surface images and 10,172 pairs of well-matched sea wave images was built to detect the location of sea waves and supervise the matching process. Furthermore, we proposed a disparity initialization module within the sparse matching network part, before searching for the corresponding key point. This offers an initial offset in vertical and horizontal directions to reduce the search range. It can tackle the problems of (1) large disparity ranges and (2) offset within stereo image pairs from the sea surface.

The remainder of this article is organized as follows. Section 2 describes the related work, including sea level measurement systems and deep-learning-based stereo matching. Section 3 briefly describes the binocular-vision-principle-based sea level monitoring system and its workflow. A detailed description of the proposed network framework is presented in Section 4. Experimental results are discussed in Section 5. Finally, Section 6 concludes this article.

2. Related Work

2.1. Sea Level Measurement

There are two types of sea level measurement methods: traditional methods using coastal equipment and long-distance measurements using offshore equipment [4]. Acoustic reflection gauges, stilling-well gauges [6], and tide poles [5] are coastal measuring equipment for tide level measurement. However, the interaction between coastal currents and coastal topography will cause local sea level changes on the sea surface, leading the measured height being higher or lower than the actual value by several centimeters, and measurement range is limited to the equipment location. In addition to these limits, for early tsunami warnings, sea level height information in the deep ocean is in high demand, which led to the introduction of offshore equipment. The main instruments for this are bottom pressure gauges [19] and buoys. The DARTII system [20] established by the National Oceanic and Atmospheric Administration and the submarine cable system [21] established by the Japan Meteorological Agency are examples of systems that have actually been put to use. Since the measurement range of these systems is limited to their setup area, it is possible that large prediction errors may occur within unequipped areas [22], such as the 3.11 Great Earthquake in Japan, where the predicted sea level height was revised after the earthquake three times. This fact has forced agencies to consider investing in

new equipment. As tsunamis or floods are rare events, investing in their measurement equipment alone is not cost-effective; thus, establishing a tsunami measurement system with long-term and stable sea level monitoring capabilities is the aim of our team, which could be used for other applications such as tide monitoring, typhoon detection and so on when there are no tsunamis.

To date, the video monitoring technique represents a supplemental solution to traditional measurement approaches, with low installation and management costs, non-contact measurements and a high sampling of monitoring targets. Nico Valentini [23] established a video monitoring system along the Apulian coast which can capture and process images automatically and release the results on a website. Sea wave 3D measurements by a stereo system became more common after the Wave Acquisition Stereo System (WASS) was proposed [24]. A variational stereo-technology-based video observational system for sea surface 3D reconstruction was developed in [25,26]. Some sea surface reconstruction methods compute disparity maps using local features [24,27,28]. Bergamasco et al. proposed a sea wave 3D reconstruction pipeline in 2017 [29]. All the steps required for constructing disparity maps from stereo image pairs are specified; however, this system cannot be used for long-distance measurements because it focuses on 3D reconstruction in short ranges. Most stereo systems are used to compensate for the lost details of sea waves during the 3D reconstruction process.

2.2. Deep-Learning-Based Stereo Matching

Stereo matching can be divided into two categories according to the matching density: sparse matching and dense matching [30]. Deep stereo matching usually focuses on the generation of a dense disparity map. Zbontar and LeCun [11] proposed the use of convolution neural networks for the computation of cost volumes for binocular stereo matching in 2015. The FlowNetCorr network structure [13] extends the application of the network to optical flow prediction. Then, the FlowNetCorr network architecture was improved to construct DisNet [14] for stereo matching; a scene flow prediction database called SceneFlow was also built, which laid the foundation for applying the end-to-end network architecture to stereo matching. Along this line, a series of improved algorithms were proposed to further boost the performance of stereo matching, such as residual learning, semantic cue and edge information preservation, adaptive aggregation, etc. [31–35]. However, the high computation and memory costs of 3D CNNs often prevent these models from being applied to large-scale cost volumes. Recently, many iterative methods [36–38] have been proposed to improve the efficiency of matching tasks. These networks need adequate and complete datasets to supervise their training process. However, without a well-rectified sea surface image dataset, they are hard to realize for our proposed system.

In some sense, the sparse stereo matching problem can be viewed as a multi-object detection and tracking problem. Most object detection methods make predictions relative to some initial guesses [39,40]. They depend heavily on the exact way these initial guesses are set. Recently, based on the application of Transformer [41] to object detection [42], this hand-crafted process was removed from the object detection network structure [43,44]. Most existing methods for object tracking are based on training discriminate classifiers [45] from templates provided by object detection. Recently, evolution of the Siamese network [46] has changed the technique to offline pair matching; then, this network was improved from different aspects [47–50]. However, all of these improved networks rely on bounding box initialization.

3. Sea Level Monitoring System

Our system was designed to be deployed along the Japan coast near the Pacific Ocean with 60 stereo imaging systems, 17 detection centers, and 10 base stations; the details of the deployment can be found in our previous work [51]. Stereo cameras are the main measurement equipment, the monitoring distance is up to 20 km and the cameras are set on spin platforms, enabling them to scan the sea surface to extend the monitoring

coverage. Figure 1 shows the configuration of the proposed stereo image measurement system. To increase the monitoring distance, the baseline length of the system is set to about 30 m. The rotations of the stereo cameras are controlled by sea surface images captured through client calculators.

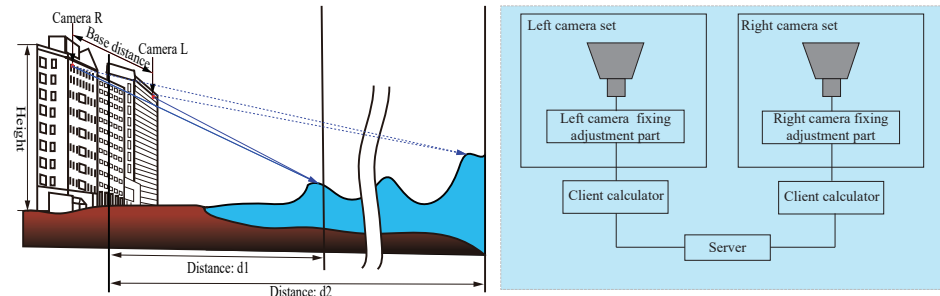


Figure 1. Configuration of the stereo sea level monitoring system (on the left is the schematic diagram and on the right is the main equipment of the proposed system).

This system is a special stereo measurement system in the respect that the monitoring distance is much larger and the monitored object is the sea surface. Here, we describe its pipeline.

Figure 2 shows the workflow of the system. Firstly, calibration is the foundation of highly accurate measurements. Intrinsic calibration has been proven to be easy to obtain within a dedicated laboratory while extrinsic calibration is difficult due to the large monitoring distance. Extrinsic calibration is executed based on the proposed calibration bar. Secondly, fast and accurate stereo matching between left and right image pairs is also important for accurate 3D measurements. In this paper, a neural-network-based sparse matching method is presented. As dense stereo matching is hard to realize and 3D reconstruction is not necessary for sea level measurements, we calculate the sea wave's 3D coordinates to determine the sea level height of the surface. This is reasonable because visible waves on the sea surface are mostly formed by sea winds and waves above and below the average sea level height coexist; when the calculated number of waves is sufficient, their average height can reflect the average sea level height. Thus, matching of sea waves is conducted instead of traditional dense stereo matching. Thirdly, the 3D filtering techniques are implemented to remove erroneous matches.

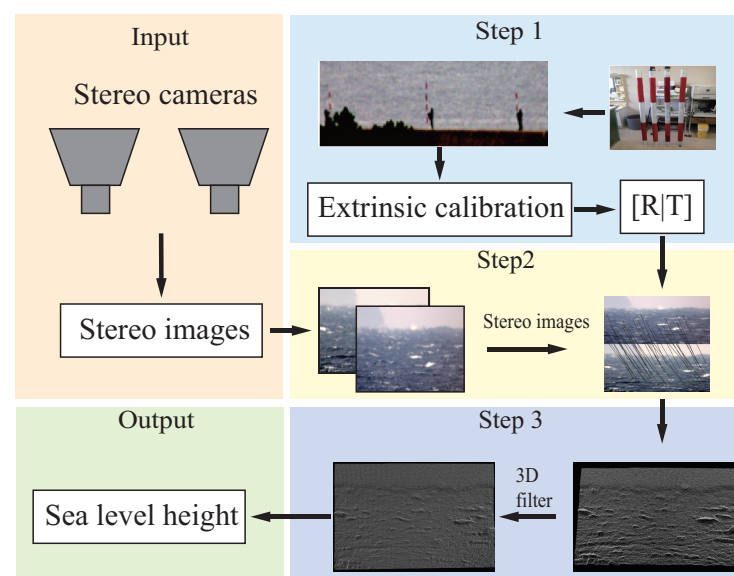


Figure 2. Workflow of the measurement system pipeline.

4. Transformer and Siamese Joint Sparse Matching Network

Sparse stereo matching is focused on in this section. It is completed by a two-stage neural network. To realize accurate and automatic matching of sea surface images, we design two modules in the network structure: a Transformer-based detection module for automatic sea wave detection and a Siamese-network-based matching module for pixel-wise matching. We first introduce the detection module in Section 4.1 and then describe the matching module in Section 4.2.

4.1. Sea Wave Detection Module

Transformer was first introduced by Vaswani et al. [41] as a new attention-based building block for machine translation. It is combined with parallel decoding for set prediction for its ability to perform global computations. Figure 3 shows the whole structure of the detection module. It is a fine-tuning model based on the DETR model [42] with our sea wave dataset, which was established based on our previous work [52]; the details of the dataset are described later in this subsection.

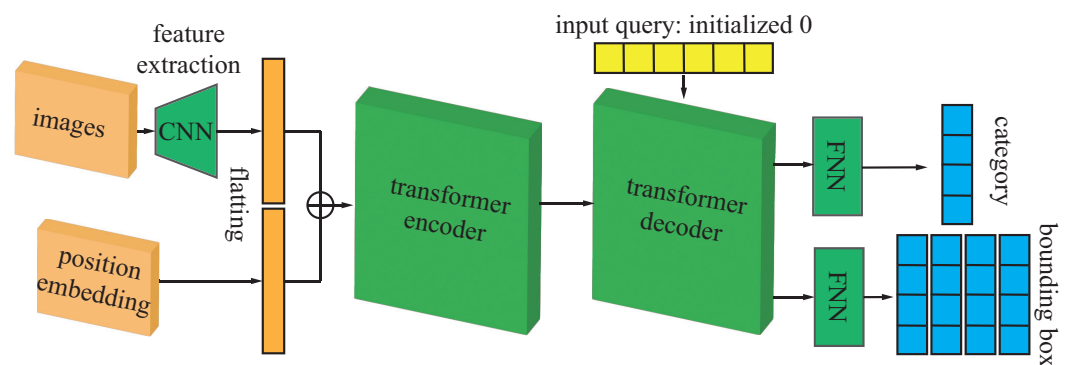


Figure 3. Sea wave detection module.

As the extracted feature map is flattened to feed into the transformer encoder, the position information is lost. To utilize the position information, the position of one pixel is represented by a 256-dimensional vector (128 elements represent x coordinate encoding and 128 elements represent y coordinate encoding), and is embedded with the extracted feature map to feed into the transformer. Position embedding is conducted using trigonometric functions (sine and cosine) to limit the value of position encoding in $[-1, 1]$. The angular frequency of each dimension in x or y coordinates is $1/10,000^{2i/128}$, where i ranges from 0 to 64. In $2i$ dimensions, the encoding value is calculated by the sine function, and in $2i + 1$ dimensions, it is calculated by the cosine function, as the following Equation (1) shows:

$$\begin{cases} PE(pos, 2i) = \sin(\frac{pos}{10,000^{2i/128}}) \\ PE(pos, 2i + 1) = \cos(\frac{pos}{10,000^{2i/128}}) \end{cases} \quad (1)$$

Here, pos is the location of the image pixel; for 2D images, one pixel has two location information values x and y , representing its coordinates along the horizontal and vertical directions.

The loss function has two items: one is classification loss and the other is bounding box loss. For the prediction with index i , we define the probability of class c_i as $\hat{p}_i(c_i)$ and the predicted box as $\hat{b}_i$; its corresponding ground truth is c_i with bounding box b_i . With this notation, we define the loss function in the following Equation (2):

$$\mathcal{L}(y, \hat{y}) = \sum_{i=1}^N \left[-\hat{p}_i(c_i) + \mathcal{L}_{box}(b_i, \hat{b}_i) \right] \quad (2)$$

where $\hat{y}$ is the predicted set $\hat{y} = \{\hat{y}_i\}_{i=1}^N$ and y represents its corresponding ground truth set $y = \{y_i\}_{i=1}^N$. The negative of the predicted possibility $-\hat{p}_i(c_i)$ of $\hat{y}_i$ being the c_i class is used as the classification loss. The second item is bounding box loss $\mathcal{L}_{box}(\cdot, \cdot)$; it is a linear combination of two parts to score the bounding box. One part is the Euclidean distance l_2 , and the other one is the IoU loss [53] to calculate the overlap ratio between b_i and $\hat{b}_i$. It is defined as in the following Equation (3),

$$\mathcal{L}_{box}(b_i, \hat{b}_i) = \lambda_{L_1}(b_i - \hat{b}_i)^2 + \lambda_{iou}\mathcal{L}_{iou}(b_i, \hat{b}_i) \quad (3)$$

Before calculating the loss function, we need to match the objects in the predicted query and target query. For high-accuracy sea level height measurements, we require the detected sea wave number to range from 100 to 200 according to our measurement principle. Therefore, the length of the predicted query is set to 200, which is fed into the transformer decoder to search for the classification and bounding box. The loss function calculates the costs between each pair of objects in two queries. We use the winner-takes-all criterion to select the best pair-wise matches; the number of best pair matches is equal to the number of objects in the target query. This can be efficiently computed via the Hungarian algorithm [54].

The training process of this module is supervised by the coco panoptic dataset (the largest panoptic segmentation dataset of common objects, featuring over 200K labeled images of objects such as different kinds of animals, appliances, food, and much more; it contains sufficient categories) and our established sea wave dataset. There are 12,614 samples in the whole dataset, 11,829 samples in the coco panoptic dataset, and 785 samples in the sea wave dataset. The samples can be divided into 134 categories, one is sea waves, with the label set to 201, and the labels of the other 133 category objects are set randomly from 0 to 200.

The establishment of the sea wave dataset is based on our previous work [52]. There are three types of sea surface images taken from two locations, Fukuoka Institute of Technology (FIT) and Fukuoka Kenritsu Suisan High School (FKSH), with three different monitoring distances: 14–20 km, 4–10 km, and 8–14 km. Figure 4 shows part of the dataset structure.

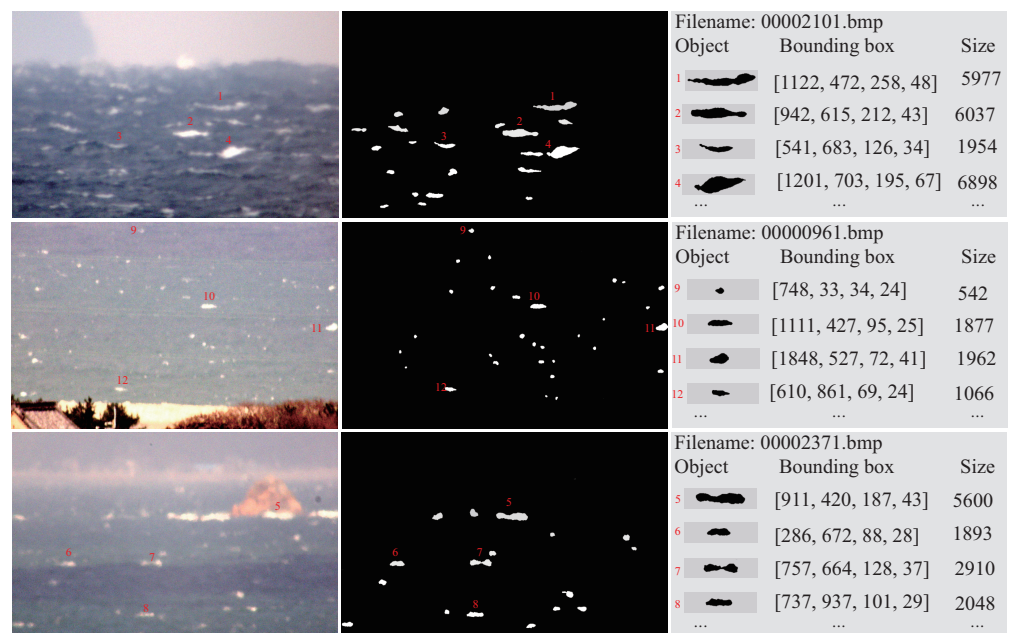


Figure 4. Sea wave detection dataset. The left column shows the 14–20 km, 4–10 km, and 8–14 km original images (from top to bottom), the middle column shows their corresponding mask images (captured in our previous work) and the right column shows part of their annotation files containing the bounding boxes and sizes of sea waves in the original images.

In this figure, from left to right, the columns contain original images, masks and annotation files. During the training process, the original images are fed into the CNN to extract feature maps, and then the extracted feature maps are fed into the transformer to calculate the prediction query as Figure 4 shows. The prediction query and target query (mask and annotation files) are substituted into Equation (2) to calculate loss value. Minimizing the loss value can optimize the parameters of the network.

4.2. Sea Wave Matching Module

Traditional matching methods focus on finding a bounding box or object key point of a template, which is low-fidelity. In this paper, we advocate for pixel-wise matching based on a convolution neural network. This is a simple development of a Siamese network. In this subsection, we will introduce this matching module from three aspects: the network architecture, loss function and dataset establishment.

Figure 5 shows the architecture of this network. A Siamese CNN backbone is used to extract feature maps from template z and search image x . Template z is a cropped section of size $h \times w$ centered on the object obtained by a detection module. The search image x is a larger image centered on the estimated position of the object in another image; in this paper, it was obtained in our previous work [55]. To compare the difference between the template and search image, a depth-wise cross-correlation represented by $\star d$ is selected to produce a multi-channel response. Each element in the cross-correlation results is calculated via a convolution operation; the kernel is a template feature map and the original data are search image feature maps. Therefore, each element in it measures the similarity between the template and search image at a specific location, representing the response of the search image (in a candidate window) for the template feature map; we called this the response of a candidate window (RoW).

Obviously, the RoW encodes the necessary information from classification scores and from the overlap of two images. Thus, it is possible to generate a target bounding box and a binary mask in the search image by adding three search branches according to SiamFC, Siam RPN, and SiamMask. Figure 5 shows these three branches: CNN_1 is tasked with discriminating each RoW between the target and background; CNN_2 outputs the bounding box of the target; and CNN_3 generates a target mask in each RoW. All of these branches are set up by two layered convolutional neural networks according to different tasks; the final output channels are then defined. For the classification branch, the output channel is in $2k$ dimensions, $1k$ for positive scores and $1k$ for negative scores. For the bounding box regression branch, it is in $4k$ dimensions, representing the center (c_x, c_y) and size h, w of the bounding box. The mask generation branch has $63 \times 63k$ dimensions; 63×63 is the produced mask image size.

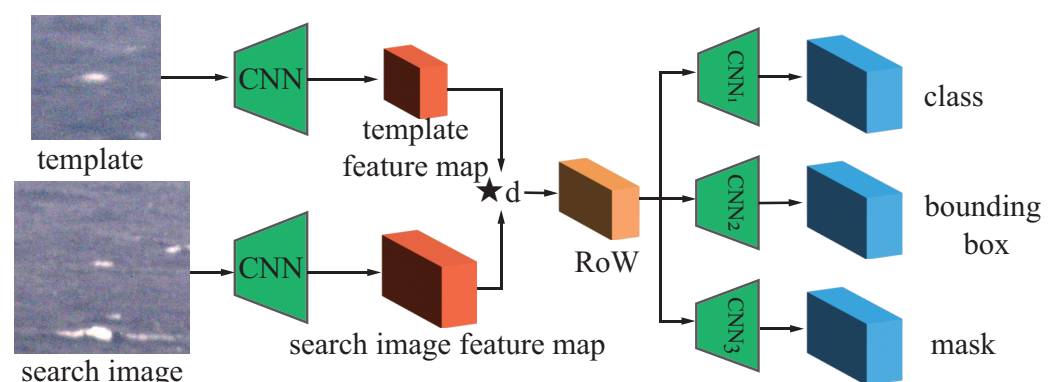


Figure 5. Architecture of the sea wave matching module.

For each branch, different loss functions are defined according to their different tasks. The classification branch is trained using cross-entropy loss:

$$\mathcal{L}_{cls} = \frac{1}{N} \sum_{i=1}^N [-y_i \log p_i + (1 - y_i) \log(1 - p_i)] \quad (4)$$

where p_i is the prediction of the classification branch for the i -th anchor, y_i is the ground truth of this anchor, and N is the total number of anchors within the search image. The bounding box regression branch is trained with the L_1 loss to avoid over-penalizing outliers. The normalized distance between an anchor and the ground-truth box is defined as:

$$\delta[0] = \frac{T_x - A_x}{A_w}, \delta[1] = \frac{T_y - A_y}{A_h}, \delta[2] = \ln \frac{T_w}{A_w}, \delta[3] = \ln \frac{T_h}{A_h} \quad (5)$$

where A_x, A_y, A_w, A_h denote the center point and the size of the anchor box and T_x, T_y, T_w, T_h denote those of the ground truth box. The regression loss of the bounding box is defined as:

$$\mathcal{L}_{reg} = \frac{1}{2N} \sum_{i=1}^N \sum_{j=0}^3 (y_i + 1) \text{smooth}_{L_1}(\delta[j] - q[j]) \quad (6)$$

Here, y_i is the ground-truth label for the i -th anchor and $q[j]$ is the normalized distance between the predicted bounding box and the i -th anchor. The mask branch is trained by a binary logistic regression loss; it is defined in the following Equation (7):

$$\mathcal{L}_{mask} = \sum_{n=1}^N \left(\frac{1 + y_n}{2wh} \sum_{ij} \log(1 + e^{-c_n^{ij} m_n^{ij}}) \right) \quad (7)$$

where $y_n \in \{\pm 1\}$ is the ground truth label of the n -th predicted mask of the n -th element in the RoW. $c_n^{ij} \in \{\pm 1\}$ denotes the label corresponding to pixel (i, j) of the object mask in the ground truth mask, m_n^{ij} is the corresponding predicted mask and $w \times h$ denotes the size of the mask.

The total loss function of the matching module is the linear combination of three branch losses, defined in the following equation:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{reg} + \lambda_3 \mathcal{L}_{mask} \quad (8)$$

where λ_1, λ_2 and λ_3 are the weights of different losses during the training process; their values are set to 0.1, 0.7 and 0.2 according to our experience. The training dataset of this module is made up of 10,172 pairs of well-matched sea waves. Positive and negative pairs of template and search image samples are obtained through the well-designed selection and cropping of sea wave pairs. They are input into the network, and the predicted classification scores, bounding boxes, and masks are used to calculate the total loss according to Equation (8). Minimization of the total loss will allow the module to converge in the ground truth direction. The training process of sea wave matching is fine-tuned based on previous researchers' work [50].

The structure of the dataset is shown in Figure 6. The template is cropped according to the red rectangle at the origin and the mask image of the left camera's field of view. The search image is larger than the template and is cropped according to the yellow rectangle in the origin image of the right camera's field of view. To improve the robustness of the network, the location of the yellow rectangle is randomly selected while its size is constant.

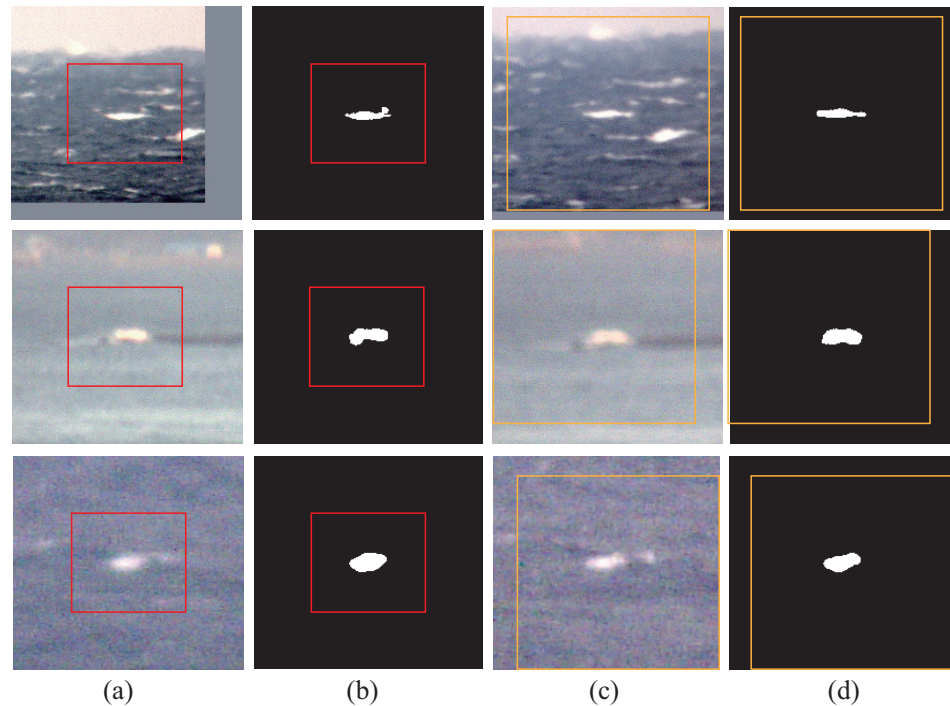


Figure 6. Part of the sea wave matching dataset. From the top to the bottom, the figure shows sea wave images and masks in the left and right cameras' fields of view with 14–20 km, 8–14 km, and 4–10 km shooting distances. Column (a) shows the sea waves in the left camera's field of view, (b) is their corresponding masks, (c) is the same sea wave in the right camera's field of view, and (d) is the corresponding masks. The red rectangles show the template areas in the left images and mask images, while the yellow rectangles show the search areas in the right images and mask images.

4.3. Disparity Initialization and Matching

From the former subsections, we know the detection module finds the sea waves in left images and the matching module finds the same sea wave in right images. During the matching process, the size of the search image is slightly larger than the template size; traversing the whole right image to cut search images is time-consuming.

However, for sea level height monitoring, real-time matching is required. We solve this problem by adding a disparity initialization, which can produce the approximate location of the target sea wave in the right image based on the location of the detected sea wave in the left image. Disparity initialization is conducted within the sea surface image based on our previous work [55], assuming the pixel coordinate in the vertical direction is y . From the top of the image to the bottom, y increases, and we found that disparity d increases when y increases; in other words, the relationship between y and d is approximately proportional, as Figure 7 shows.

Given a pair of stereo sea surface images, we can manually select two pairs of well-matched points to approximately calculate the proportional relationship. The relationship is used to select search image location to eliminate the process of traversing the whole image.

The matching process can be summarized in the following steps: (1) for each left image, the object detection module is used to obtain M templates $\mathcal{D} = \{d_i\}_{i=1}^M$, (2) the search images $\mathcal{S} = \{s_i\}_{i=1}^M$ are selected in right image based on disparity initialization, (3) the templates whose search images are out of the right camera's field of view are deleted and the template set is updated to $\hat{\mathcal{D}} = \{\hat{d}_i\}_{i=1}^{\hat{M}}$, and (4) the matching module is utilized to produce the specific mask, location and bounding box of the object in each search image. During this matching process, traditional assignment problems between two image objects are eliminated and replaced by disparity initialization, which is much more efficient.



Figure 7. The proportional relationship between disparity and the y coordinate.

5. Experiments

In the experiment section, we first describe the implementation details of two modules in Section 5.1; then, we evaluate our approach on sea surface images captured by our tsunami measurement system during three different periods, 18–23 August 2018, 8–16 March 2017 and 29 February–6 March 2016, from two sites, FKSH and FIT, at three monitoring distances, 8–14 km, 4–10 km, and 14–20 km. The evaluation is conducted from two aspects: sea wave detection (Section 5.2) and sea wave matching (Sections 5.3 and 5.4).

5.1. Implementation

For both detection and matching modules, feature extraction is achieved using the ResNet-50 [56] architecture, where the extracted feature map has 2048 channels. Then, it passes through a 2D CNN with a 1×1 kernel to reduce its channels to obtain a 256-channel feature map f . For object detection, the 256-channel feature map f is flattened and fed into the transformer to detect the objects and their location information fd . For object matching, 256-channel feature maps of the template and search image are fed into the depth-wise cross-correlation module to calculate their similarity map fm . By applying a linear transformation with (256, 251) weight shape to fd , we can classify detected objects. Applying a series of linear transformations with (256, 256), (256, 256), (256, 4) weight shapes to fd , we can obtain the bounding boxes of detected objects. Feeding fm into three connection layers and in each layer, there are two CNNs with 1×1 kernels and strides equal to 1, we can produce the bounding box and mask of the template in a search image. Table 1 shows the details of the architecture.

Table 1. The architecture of the overall network.

Module		Detection		Matching	
Layers					
backbone		ResNet-50			
layer1		conv2d [2048, 256] flattening position embedding		conv2d [2048, 256]	
layer2		Transformer		deep-wise cross correlation	
layer3	linear [256, 251]	linear [256, 256] linear [256, 256] linear [256, 4]	conv2d [256, 256] conv2d [256, 10]	conv2d [256, 256] conv2d [256, 20]	conv2d [256, 256] conv2d [256, 3969]
output	object classification	bounding box	match score	bounding box	mask

1. [input_channel, output_channel] means the channels of input and output features of the specific network,

2. linear represents the linear transformation, 3. conv2d represents the 2D convolution network

5.2. Evaluation of Sea Wave Detection

We first show the sea wave detection results in Figure 8, which have been produced by the method described in this paper. The network was trained for 70 epochs with the training dataset (described in Section 4.1) and the learning rate is 1×10^{-4} . The training time of this module is 8 days, 13 h, 57 min and 48 s. In (a), the detected masks of all sea waves are shown, where the image was captured from a 14–20 km monitoring distance, and (b) shows the detection result, where all the masks in (a) are marked in the original image. From the result, we find that 22 sea waves are detected, 20 sea waves are correctly detected, and 4 sea waves are missed.

We also conducted comparisons between the traditional extraction method [52] and the method described in this paper. Figure 9 shows three pairs of images from this comparison on three images from three monitoring distances: 14–20 km, 8–14 km and 4–10 km. Column (a) shows the detection results of our detection module and (b) shows the results of traditional method. Traditional methods focus on solving the problem of uneven illumination in long-distance sea surface images to extract sea waves, while the neural network ignores this problem, and we find that it can detect the sea waves missed by traditional methods, and at the same time, the edges of detected sea waves are much smoother.

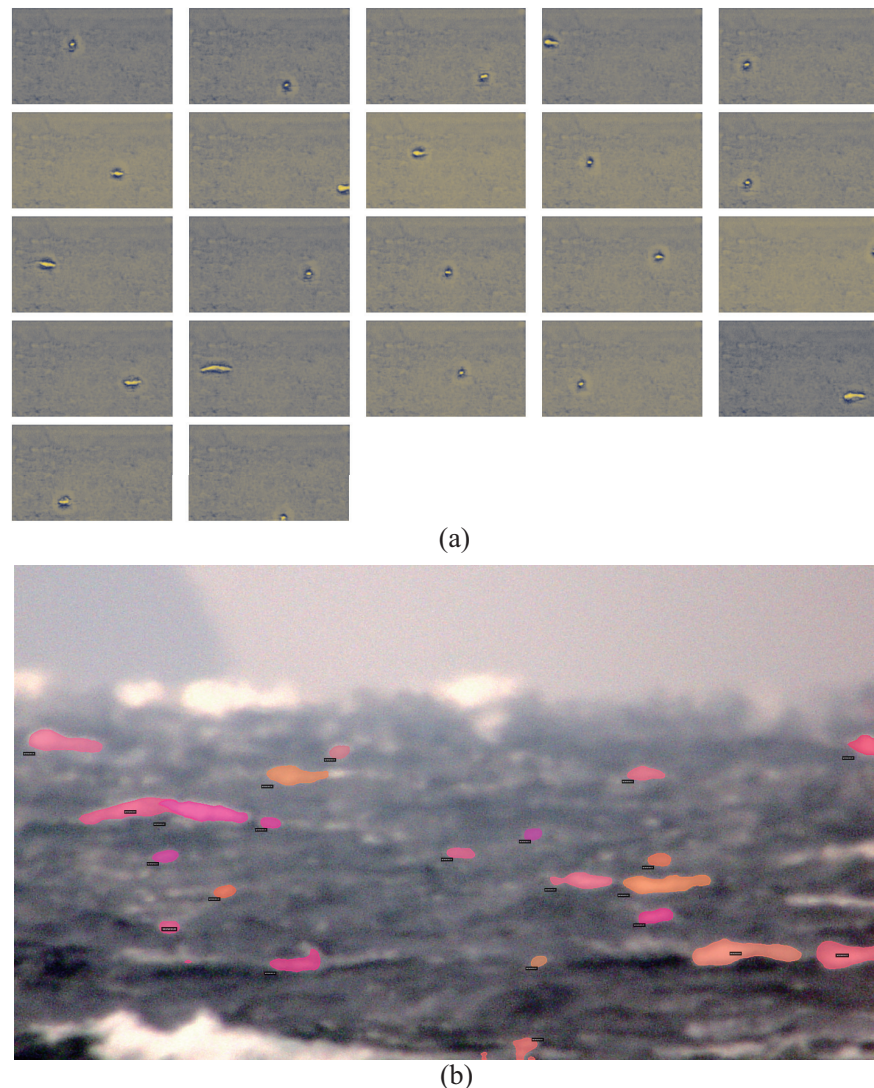


Figure 8. Sea wave detection result: (a) shows the masks of detected sea waves, (b) shows the mark result.

To show that the neural-network-based detection module outperforms the traditional detection method, we processed 67 images shot at different times with different monitoring distances (10 images from 14–20 km away, 33 images from 8–14 km away, 24 images from 4–10 km away), and the detection results are shown in the following Table 2 and Figure 10. From the table, the detection ratio of the proposed detection method is 86.47%, 11.95% higher than a traditional extraction method (here, our previously proposed dynamic threshold algorithm). In Figure 10, the neural network detection results are marked in blue. It is easy to find that with the same false positive rate, the proposed network achieves a higher true positive rate most of the time. In these results, only the numbers of detected sea waves are collected and the quality of the detected results is not evaluated; in other words, if there is a mask within one sea wave area, we collect it and ignore the edge, size, shape and other features of the mask.



Figure 9. Comparison of sea wave detection results, column (a) shows the results described in this paper, column (b) shows results of [52]. From top to bottom are the images shoot from 14–20 km, 8–14 km and 4–10 km.

Table 2. The numerical comparison of detection results.

Methods Images	Original Number	Neural Network Detection	Traditional Detection
14–20 km (10 images)	218	181	150
8–14 km (33 images)	615	528	497
4–10 km (24 images)	823	723	587
total	1656	1432	1234
ratio	1.00	86.47	74.52

1. Original number denotes the number of manually collected sea waves that can be clearly observed in original images; 2. Traditional detection represents the detection results of our previously proposed dynamic threshold method.

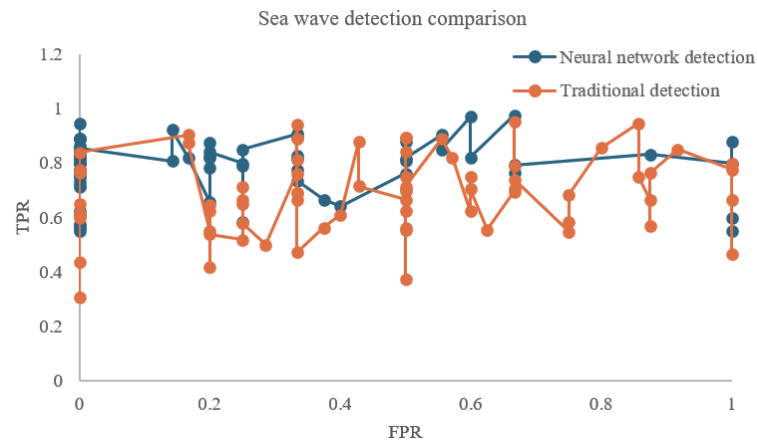


Figure 10. ROC curves of sea wave detection results; the blue line shows the results described in this paper and the red line shows the results of [52]. The results are collected from 67 images shot at different times with different monitoring distances.

5.3. Matching Mask

We show part of the mask extraction results of the matching module in Figure 11; the module was trained with 630 epochs with the training dataset (described in Section 4.2), and the learning rate was manually adjusted from 1×10^{-1} to 1×10^{-3} . The whole training time of this module is 1 day, 10 h, 40 min and 3 s. The threshold of IoU (intersection over union between the predicted and ground truth masks) is set to 0.0 and 0.5; a predicted mask with an IoU larger than 0.5 is recorded as good predictions and masks with an IoU larger than 0.0 are recorded as positive prediction; otherwise, they are negative predictions. The ground truth is manually collected based on our previous research. Good, positive, and negative prediction samples of three long-distance sea surface images are shown, and from top to bottom, mask generation results from three monitoring distances of 14–20 km, 8–14 km and 4–10 km are shown. From the IoU results, we found that the size of the mask generated by the matching module is always smaller than the manually marked mask, and the location approaches the center of the sea wave; this may be caused by the dissimilarity in the nature of the sea wave pairs captured by left and right cameras.

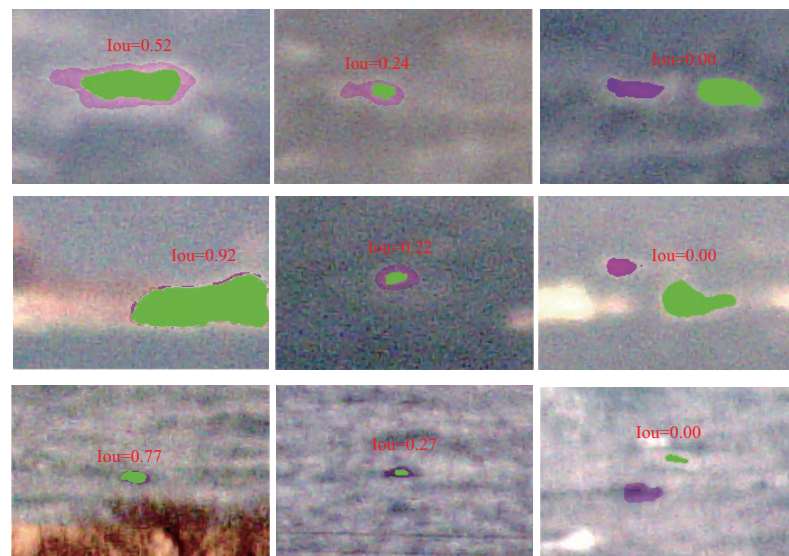


Figure 11. Samples of different IoU results. Sea wave masks are generated by a matching module (green) and a manual marking process (lavender); the color of overlap is between green and lavender. From top to bottom are the sea waves from images shot at a distance of 14–20 km, 8–14 km, and 4–10 km.

Table 3 shows the numerical results aggregated from 53 original images. We can quantify the effectiveness of the matching module for pixel-wise mask generation. A total of 77.6% sea waves were correctly matched; however, only 15.0% of sea waves obtain masks with an IoU > 0.5, and this is lower than our expectation. However, comparing the matching results with the traditional method's results, we find that it can generate a higher positive accuracy and a mask with a clear shape and smooth edges. We will describe this in the next subsection.

Table 3. The produced matching mask quality (IoU).

Items Images	Ground Truth Number	Well (IoU > 0.5)	Positive (IoU > 0.0)	Negative (IoU = 0)
14–20 km (13 image pairs)	133	27	91	42
8–14 km (15 image pairs)	150	15	133	17
4–10 km (25 image pairs)	343	53	262	81
total	626	95	486	140
ratio	1.00	15.0	77.6	22.4

1. The ground truth number was manually collected; 2. Positive number represents the number of predictions for which the IoU is larger than 0.0.

5.4. Matching Results

The major goal of this paper is completing sparse stereo matching via a neural network. In the previous subsections, we showed the object detection and mask generation results. In this subsection, we will show the most important result of this paper—the stereo matching result—which is produced based on object detection and mask generation results. To validate the effectiveness of the proposed sparse matching method, we compared the proposed method matching result with the RANSAC + SURF method matching result and the result of our previously proposed method [51]. There are two evaluation terms, precision and recall, which are defined using the following formulas, which are slightly different from the definitions in [51]:

$$precision = \frac{nP}{(nP + nN)} \times 100 \quad (9)$$

$$recall = \frac{nTP}{nTP + nFN} \times 100 \quad (10)$$

where nP and nN are the numbers of correctly and wrongly matched points; they record all the detected points regardless of whether these points belong to sea waves or not. nTP is the number of correctly matched sea waves. nFN is the number of correct sea waves that are not detected. nTP and nFN are the variables related to the number of sea waves.

To illustrate the comparative result intuitively, we show six representative images' matching results taken at different times. Figure 12, ①, ②, ③, ④, ⑤ and ⑥ are the six groups of comparison results of the representative image pairs. The (a) column shows the RANSAC + SURF matching results, the (b) column shows the matching results from [51] and the (c) column shows the results of our proposed method in this paper. ① and ② were taken from 29 February to 6 March in 2016 from FKSH, and the distance was 14–20 km. ③ and ④ were taken from 18 to 23 August in 2018 from FIT, and the distance was 8–14 km. ⑤ and ⑥ were taken from 8 to 15 March in 2017 from FIT; the monitoring distance was 4–10 km.

We also conducted a quantitative comparison of these image pairs; the performance of each method is shown in Table 4. The ground truth is established by manual checking. The average precision of the RANSAC + SURF algorithm is 88.1%. This means that this method can match the detected key points correctly. However, the average recall of this method is 20.5%, showing that the algorithm missed many sea waves, which will then influence the sea level measurement precision. The average precision of our proposed

method is 87.4%. This is the lowest among the three methods because the proposed method will find a match for each detected object, even for those whose corresponding sea wave is out of the field of view of the camera, leaving out these sea waves. The proposed method can achieve an almost 100% precision according to our experiments, as Figure 13 shows.

In Figure 13, column (a) shows the matching results of our proposed neural network; the red rectangles mark the areas where correct sea wave matches are out of the view of the other camera. Column (b) shows the improved results if we eliminate these kinds of incorrect matches using manually selected parameters; there are almost no incorrect matches. The recall of the proposed method is larger than the RANSAC + SURF algorithm but lower than that of the [51] method. However, the biggest highlight of the proposed method is that it can generate pixel-wise matching masks during the matching process.

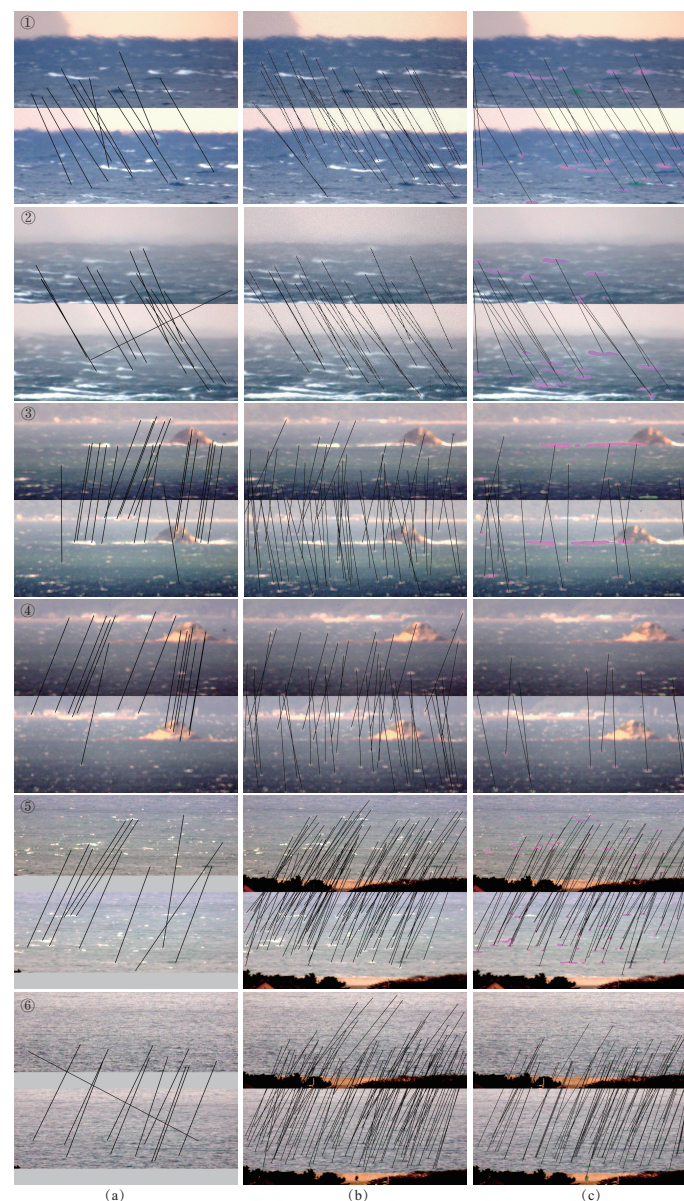


Figure 12. Comparison of RANSAC + SURF, the method in ref. [51] and the method proposed in this paper. Rows ① and ② are the images taken from 29 February to 6 March 2016 (14–20 km); rows ③ and ④ are the images taken from 18 to 23 August 2018 (8–14 km); rows ⑤ and ⑥ are the images taken from 8 to 15 March 2017 (4–10 km). Column (a) shows the results of RANSAC + SURF; column (b) shows the results of [51]’s method; and column (c) shows the results of the proposed method.

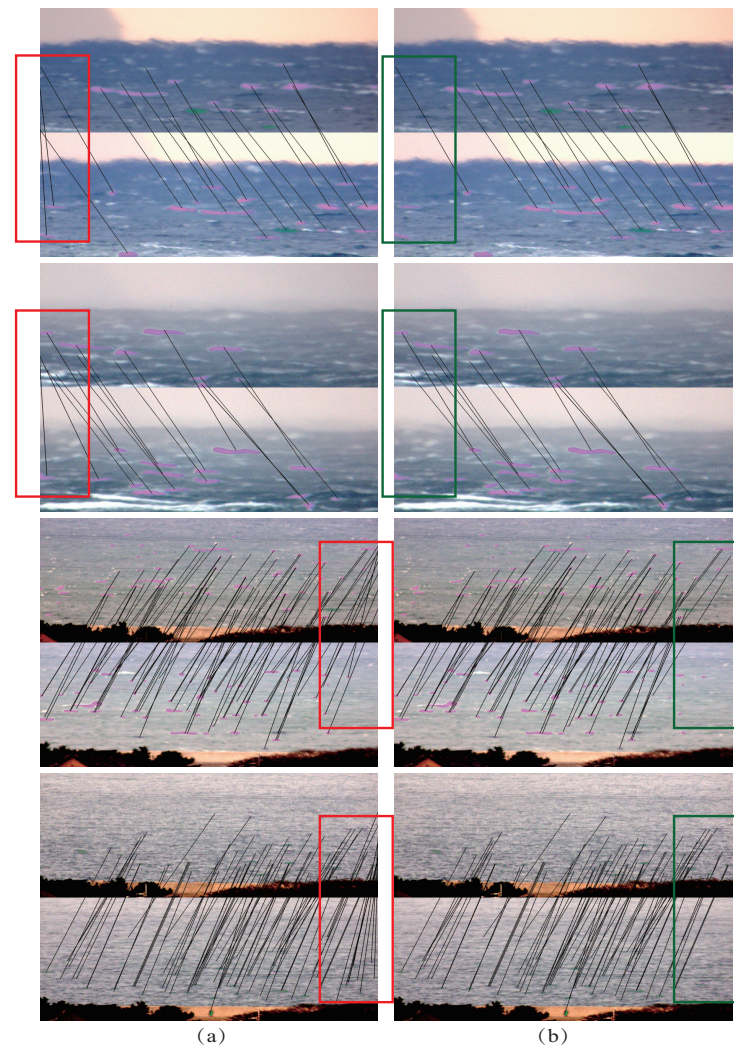


Figure 13. Results of the proposed method and the improvements after manual parameter adjustment. Column (a) shows the original images; they are the same as the result in Figure 12. Red rectangles mark the main source of incorrect matches, column (b) shows the improved results, and the green rectangles mark the disappearance of the incorrect matches.

Table 4. Comparison of sparse matching results.

NO	RANSAC + SURF		[51] Method		Proposed Method	
	Precision	Recall	Precision	Recall	Precision	Recall
①	72.7	28.6	100.0	92.6	76.4	46.4
②	92.3	48.0	95.7	88.0	86.7	52.0
③	100.0	22.0	92.5	98.0	93.3	28.0
④	100.0	2.4	88.1	90.2	100	24.4
⑤	81.8	9.6	99.0	100.0	87.3	67.4
⑥	81.8	9.7	96.7	95.7	80.8	64.1
Average	88.1	20.5	95.3	97.4	87.4	47.5

In our previous work, matching was conducted between left and right sea waves, the center of the bounding box is taken as the sea wave key point, and the bounding boxes of left and right sea waves are captured independently using a dynamic threshold. Thus, they are independent and easily influenced by the extraction threshold, which will influence the final matching accuracy. In this paper, matching is conducted between the left sea wave and right searched area and the features of the sea waves and searched areas are extracted by a Siamese network with sharing weights; the key point of the right sea wave is the point

which has the largest similarity with the left sea wave. This can solve the problem caused by different extraction thresholds.

We also compared the running times of the RANSAC + SURF method, our proposed traditional method [51] and the neural network method. Three experiments were carried out on three sets of image pairs: 61 pairs of images from 14–20 km away, 79 pairs of images from 8–14 km away, and 66 pairs of images from 4–10 km away. The execution equipment for RANSAC + SURF and the traditional method was a desktop with a 3.4 GHz Intel core CPU and 6 GB of memory with C++ code, while the execution equipment for the neural network method was a desktop with a 3.6 GHz Intel(R) Xeon(R) CPU and NVIDIA RTX A5500 GPU. Figure 14 shows the comparison results. Observing Figure 14, we find that the neural network method is slower than the other two methods, and the traditional method is still the fastest. In the future, we plan to accelerate the neural network through structure adjustment, parameter optimization, etc., to allow it to achieve a speed comparable to the traditional method.

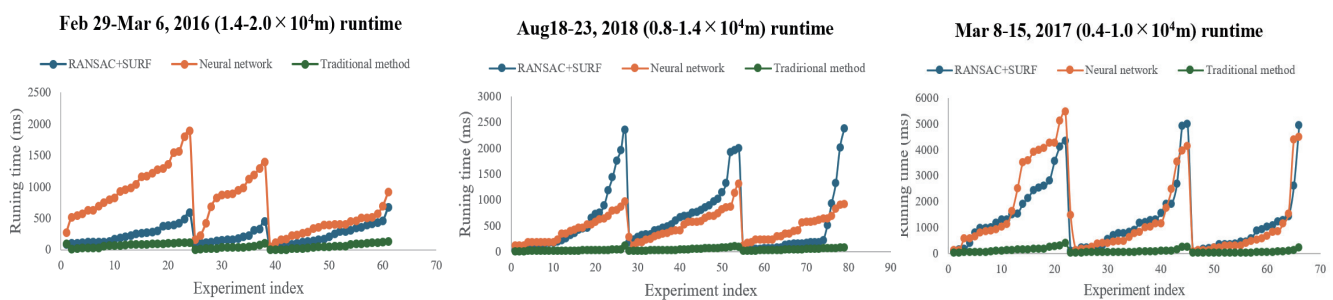


Figure 14. Comparison of the running times of the RANSAC + SURF method, our proposed traditional method [51] and the neural network method.

6. Discussion

In this paper, we introduced a neural network architecture for sparse stereo matching for a long-distance sea level height monitoring system, which can replace the traditional feature extraction and stereo matching system processes. According to the detection results, it achieves a better recognition rate and quality than the traditional feature extraction method without considering the uneven illumination of long-distance sea surface images. The matching module also conducts pixel-wise matching of the input template and search image, and more than 70% of sea waves can be correctly matched with an IoU > 0.

The matching precision of the proposed matching module is the lowest compared with the two traditional methods due to the mandatory matching result output of the designed module that does not consider sea waves whose corresponding target is out of the right image's field of view. In the future, we will explore a method to allow output = $\emptyset$; by eliminating these types of mismatches, the precision of the proposed matching module may be up to 100%. Currently, it seems that the result of the neural network method is still no better than the traditional algorithm, but it is just at the beginning stage; many more network structures and training methods need to be explored to find an appropriate architecture for long-distance sea level height monitoring systems, and deep-learning-based dense matching is still being explored by our lab. We hope that our work will inspire further studies on applying neural networks to multi systems that still rely on traditional image processing algorithms.

Author Contributions: Conceptualization, methodology and validation, Y.Y., methodology implementation and supervision, C.L. and Z.L., writing—original draft preparation, Y.Y., writing—review and editing, C.L. and Z.L., sea surface image resource, C.L. and Y.Y. All authors have read and agreed to the published version of the manuscript

Funding: This research was funded by NSFC under grants 62221004, 61971225 and 62175110, the NJUST under No. TSXK2022D00x, JSPS KAKENHI Grant Numbers JP17K01331, and the MEXT-Supported Program for the Strategic Research Foundation at Private Universities S1311050.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article.

Acknowledgments: The authors want to express their sincere gratitude to the other members of Lulab for accompanying them when performing surface photography experiments. Also, sincere thanks go to the English teachers Samantha Hawkins and Sam Tuza for conducting grammar and spelling checks of the manuscript.

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

References

1. Pompe, J.J.; Rinehart, J.R. Mitigating damage costs from hurricane strikes along the southeastern U.S. Coast: A role for insurance markets. *Ocean Coast. Manag.* **2008**, *51*, 782–788. [\[CrossRef\]](#)
2. Villamayor, B.M.R.; Rollon, R.N.; Samson, M.S.; Albano, G.M.G.; Primavera, J.H. Impact of Haiyan on Philippine mangroves: Implications to the fate of the widespread monospecific *Rhizophora* plantations against strong typhoons. *Ocean Coast. Manag.* **2016**, *132*, 1–14. [\[CrossRef\]](#)
3. Pugh, D.; Woodworth, P. *Sea-Level Science: Understanding Tides, Surges, Tsunamis and Mean Sea-Level Changes*; Cambridge University Press: Cambridge, UK, 2014.
4. Intergovernmental Oceanographic Commission. *Manual on Sea Level Measurement and Interpretation*; Volume I-Basic procedures; UNESCO: Paris, France, 1985.
5. Miguez, B.M.; Testut, L.; Wöppelmann, G. The Van de Casteele test revisited: An efficient approach to tide gauge error characterization. *J. Atmos. Ocean. Technol.* **2008**, *25*, 1238–1244. [\[CrossRef\]](#)
6. Bunt, T.G. XI. Description of a new tide-gauge, constructed by Mr. TG Bunt, and erected on the eastern bank of the river Avon, in front of the Hotwell House, Bristol, 1837. *Philos. Trans. R. Soc. Lond.* **1838**, *128*, 249–251.
7. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 3354–3361.
8. Menze, M.; Geiger, A. Object scene flow for autonomous vehicles. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3061–3070.
9. Koomey, J.; Berard, S.; Sanchez, M.; Wong, H. Implications of historical trends in the electrical efficiency of computing. *IEEE Ann. Hist. Comput.* **2010**, *33*, 46–54. [\[CrossRef\]](#)
10. Zagoruyko, S.; Komodakis, N. Learning to compare image patches via convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4353–4361.
11. Zbontar, J.; LeCun, Y. Computing the stereo matching cost with a convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1592–1599.
12. Luo, W.; Schwing, A.G.; Urtasun, R. Efficient deep learning for stereo matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5695–5703.
13. Dosovitskiy, A.; Fischer, P.; Ilg, E.; Hausser, P.; Hazirbas, C.; Golkov, V.; Van Der Smagt, P.; Cremers, D.; Brox, T. FlowNet: Learning optical flow with convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2758–2766.
14. Mayer, N.; Ilg, E.; Hausser, P.; Fischer, P.; Cremers, D.; Dosovitskiy, A.; Brox, T. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4040–4048.
15. Gidaris, S.; Komodakis, N. Detect, replace, refine: Deep structured prediction for pixel wise labeling. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5248–5257.
16. Kendall, A.; Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
17. Rao, Z.; He, M.; Dai, Y.; Zhu, Z.; Li, B.; He, R. Nlca-net: A non-local context attention network for stereo matching. *APSIPA Trans. Signal Inf. Process.* **2020**, *9*, e18. [\[CrossRef\]](#)
18. Liang, Z.; Guo, Y.; Feng, Y.; Chen, W.; Qiao, L.; Zhou, L.; Zhang, J.; Liu, H. Stereo matching using multi-level cost volume and multi-scale feature constancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 300–315. [\[CrossRef\]](#)
19. Spencer, R.; McGarry, C.; Harrison, A.; Vassie, J.; Baker, T.; Smithson, M.; Haranzogo, S.; Woodworth, P. The ACCLAIM programme in the South Atlantic and Southern oceans. *Int. Hydrogr. Rev.* **1993**, *70*, 7–21.

20. Meinig, C.; Stalin, S.E.; Nakamura, A.I.; González, F.; Milburn, H.B. Technology developments in real-time tsunami measuring, monitoring and forecasting. In Proceedings of the OCEANS 2005 MTS/IEEE, Washington, DC, USA, 17–23 September 2005; IEEE: Piscataway, NJ, USA, 2005; pp. 1673–1679.
21. Agency, J.M. Prediction of Tsunami. Available online: <http://www.data.jma.go.jp/svd/eqev/data/tsunami/ryoteki.html> (accessed on 11 September 2022).
22. Jin, D.; Lin, J. Managing tsunamis through early warning systems: A multidisciplinary approach. *Ocean Coast. Manag.* **2011**, *54*, 189–199. [\[CrossRef\]](#)
23. Valentini, N.; Saponieri, A.; Damiani, L. A new video monitoring system in support of Coastal Zone Management at Apulia Region, Italy. *Ocean Coast. Manag.* **2017**, *142*, 122–135. [\[CrossRef\]](#)
24. Benetazzo, A. Measurements of short water waves using stereo matched image sequences. *Coast. Eng.* **2006**, *53*, 1013–1032. [\[CrossRef\]](#)
25. Gallego, G.; Yezzi, A.; Fedele, F.; Benetazzo, A. A variational stereo method for the three-dimensional reconstruction of ocean waves. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 4445–4457. [\[CrossRef\]](#)
26. Gallego, G.; Yezzi, A.; Fedele, F.; Benetazzo, A. Variational stereo imaging of oceanic waves with statistical constraints. *IEEE Trans. Image Process.* **2013**, *22*, 4211–4223. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Wanek, J.M.; Wu, C.H. Automated trinocular stereo imaging system for three-dimensional surface wave measurements. *Ocean Eng.* **2006**, *33*, 723–747. [\[CrossRef\]](#)
28. Brandt, A.; Mann, J.; Rennie, S.; Herzog, A.; Criss, T. Three-dimensional imaging of the high sea-state wave field encompassing ship slamming events. *J. Atmos. Ocean. Technol.* **2010**, *27*, 737–752. [\[CrossRef\]](#)
29. Bergamasco, F.; Torsello, A.; Sclavo, M.; Barbariol, F.; Benetazzo, A. WASS: An open-source pipeline for 3D stereo reconstruction of ocean waves. *Comput. Geosci.* **2017**, *107*, 28–36. [\[CrossRef\]](#)
30. Wedel, A.; Rabe, C.; Vaudrey, T.; Brox, T.; Franke, U.; Cremers, D. Efficient dense scene flow from sparse or dense stereo data. In Proceedings of the European Conference on Computer Vision, Marseille, France, 12–18 October 2008; Springer: Berlin/Heidelberg, Germany, 2008; pp. 739–751.
31. Liang, Z.; Feng, Y.; Guo, Y.; Liu, H.; Chen, W.; Qiao, L.; Zhou, L.; Zhang, J. Learning for disparity estimation through feature constancy. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2811–2820.
32. Pang, J.; Sun, W.; Ren, J.S.; Yang, C.; Yan, Q. Cascade residual learning: A two-stage convolutional neural network for stereo matching. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 887–895.
33. Song, X.; Zhao, X.; Fang, L.; Hu, H.; Yu, Y. Edgestereo: An effective multi-task learning network for stereo matching and edge detection. *Int. J. Comput. Vis.* **2020**, *128*, 910–930. [\[CrossRef\]](#)
34. Xu, H.; Zhang, J. Aanet: Adaptive aggregation network for efficient stereo matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 1959–1968.
35. Yang, G.; Zhao, H.; Shi, J.; Deng, Z.; Jia, J. Segstereo: Exploiting semantic information for disparity estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 636–651.
36. Lipson, L.; Teed, Z.; Deng, J. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In Proceedings of the 2021 International Conference on 3D Vision (3DV), London, UK, 1–3 December 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 218–227.
37. Teed, Z.; Deng, J. Raft: Recurrent all-pairs field transforms for optical flow. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; , Proceedings, Part II 16; pp. 402–419.
38. Wang, F.; Galliani, S.; Vogel, C.; Pollefeys, M. IterMVS: Iterative probability estimation for efficient multi-view stereo. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8606–8615.
39. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; Volume 28.
40. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
41. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
42. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; Proceedings, Part I 16; pp. 213–229.
43. Stewart, R.; Andriluka, M.; Ng, A.Y. End-to-end people detection in crowded scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2325–2333.
44. Ren, M.; Zemel, R.S. End-to-end instance segmentation with recurrent attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6656–6664.

45. Bolme, D.S.; Beveridge, J.R.; Draper, B.A.; Lui, Y.M. Visual object tracking using adaptive correlation filters. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 2544–2550.
46. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H. Fully-convolutional siamese networks for object tracking. In Proceedings of the Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, 8–10 and 15–16 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; Proceedings, Part II 14; pp. 850–865.
47. He, A.; Luo, C.; Tian, X.; Zeng, W. Towards a better match in siamese network based visual object tracker. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018.
48. Zheng, J.; Ma, C.; Peng, H.; Yang, X. Learning to track objects from unlabeled videos. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 13546–13555.
49. Yan, B.; Peng, H.; Wu, K.; Wang, D.; Fu, J.; Lu, H. Lighttrack: Finding lightweight neural networks for object tracking via one-shot architecture search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 15180–15189.
50. Hu, W.; Wang, Q.; Zhang, L.; Bertinetto, L.; Torr, P.H. Siammask: A framework for fast online object tracking and segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 3072–3089. [[PubMed](#)]
51. Yang, Y.; Lu, C.; Li, Z. Long-Distance Sea Wave Sparse Matching Algorithm for Sea Level Monitoring System. *J. Mar. Sci. Eng.* **2023**, *11*, 391. [[CrossRef](#)]
52. Yang, Y.; Lu, C. Long-distance sea wave extraction method based on improved Otsu algorithm. *Artif. Life Robot.* **2019**, *24*, 304–311. [[CrossRef](#)]
53. Rezaatofghi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.
54. Kuhn, H.W. The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **1955**, *2*, 83–97. [[CrossRef](#)]
55. Yang, Y.; Lu, C. A stereo matching method for 3D image measurement of long-distance sea surface. *J. Mar. Sci. Eng.* **2021**, *9*, 1281. [[CrossRef](#)]
56. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.