Journal of *Marine Science and Engineering*

MDPI

Article

Underwater Unsupervised Stereo Matching Method Based on Semantic Attention

Qing Li [1,2], Hongjian Wang [1,*], Yao Xiao [1], Hualong Yang [1], Zhikang Chi [1] and Dongchen Dai [1]

[1] College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001, China; 15545131870@163.com (Q.L.); xiaoyao9@hrbeu.edu.cn (Y.X.); long1067726236@gmail.com (H.Y.); czk9152020@163.com (Z.C.); y791986653@163.com (D.D.)

[2] College of Intelligent Science and Engineering, Yantai Nanshan University, Yantai 264000, China

* Correspondence: cctime99@163.com

Abstract: A stereo vision system provides important support for underwater robots to achieve autonomous navigation, obstacle avoidance, and precise operation in complex underwater environments. This article proposes an unsupervised underwater stereo matching method based on semantic attention. By combining deep learning and semantic information, it fills the challenge of insufficient training data, enhances the intelligence level of underwater robots, and promotes the progress of underwater scientific research and marine resource development. This article proposes an underwater unsupervised stereo matching method based on semantic attention, targeting the missing training supervised dataset for underwater stereo matching. An adaptive double quadtree semantic attention model for the initial estimation of semantic disparity is designed, and an unsupervised AWLED semantic loss function is proposed, which is more robust to noise and textureless regions. Through quantitative and qualitative evaluations in the underwater stereo matching dataset, it was found that D1 all decreased by 0.222, EPE decreased by 2.57, 3px error decreased by 1.53, and the runtime decreased by 7 ms. This article obtained advanced results.

Keywords: semantic attention; underwater stereo matching; adaptive double quadtree; unsupervised AWLED semantic loss function

check for updates

Citation: Li, Q.; Wang, H.; Xiao, Y.; Yang, H.; Chi, Z.; Dai, D. Underwater Unsupervised Stereo Matching Method Based on Semantic Attention. *J. Mar. Sci. Eng.* **2024**, *12*, 1123. https://doi.org/10.3390/jmse12071123

Academic Editor: Rafael Morales

Received: 24 May 2024
Revised: 26 June 2024
Accepted: 2 July 2024
Published: 4 July 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

Underwater binocular stereo matching technology is a challenging research field that is mainly used for 3D reconstruction and machine vision systems in underwater environments. The underwater environment has unique visual characteristics, such as light attenuation, scattering, and the influence of water particles, which greatly increase the difficulty of stereo matching.

Underwater stereo vision systems can be used to monitor and evaluate the health status of coral reefs, undersea vegetation, and other ecosystems, helping scientists carry out environmental protection and ecological restoration work [1–4]. Through 3D reconstruction technology, researchers can accurately measure and analyze the size, shape, and behavior of underwater organisms, which is of great significance for biodiversity research and species conservation [5,6]. Underwater stereo vision systems can also be applied to the exploration of deep-sea mineral resources, providing accurate terrain and geological information for deposit positioning and evaluation [7–9]. In the field of underwater archaeology, underwater stereo vision systems can assist archaeologists in the three-dimensional reconstruction and analysis of sunken ships, ancient buildings, and other cultural sites, which can protect these precious cultural resources without the need for actual excavation [10–13]. For the construction and maintenance of underwater infrastructure such as bridges, dams, pipelines, and submarine cables, underwater stereo vision systems can provide accurate 3D models to assist engineers in structural integrity assessments and damage detection [14–16]. In the development of underwater robots (such as autonomous

underwater vehicles, AUVs), the underwater stereo vision system is a key technology for achieving autonomous navigation, obstacle avoidance, and the precise operation of robots, providing detailed three-dimensional information about the surrounding environment for robots and enhancing their operational capabilities in complex underwater environments [17–19]. Through these applications, not only does underwater binocular stereo matching technology promote the deepening of scientific research but it promotes the development and innovation of related technologies as well, making extremely significant contributions to environmental protection, resource development, and human understanding and utilization of the ocean.

First and foremost, the research status of underwater binocular stereo matching shows a diversified trend. Although traditional methods have generated promising results, their use for accurate underwater image matching is still limited. Deep learning models, such as CNNs and GANs, handle this more effectively by learning complex patterns and adapting to underwater challenges, providing more robust solutions [20,21]. For instance, stereo matching algorithms based on deep learning [22–25] have achieved good results in underwater scenes, and through training and optimization of neural networks, the accuracy and robustness of matching can be enhanced.

Additionally, underwater binocular stereo matching faces many challenges. Due to factors such as complex lighting, a turbid water quality, and the limited surface texture of underwater objects, the quality of underwater images is poor, making it difficult to accurately obtain depth information through binocular stereo matching. In addition, the complex shape and uncertain motion state of underwater objects also increase the difficulty of matching. Therefore, how to increase the accuracy and stability of underwater binocular stereo matching is one of the current research focuses.

Furthermore, the development direction of underwater binocular stereo matching is also worth paying attention [26–28]. With the continuous progress of artificial intelligence and computer vision technology [29,30], underwater binocular stereo matching is expected to achieve higher accuracy and efficiency [31–34]. Future research can combine multi-sensor information, such as sonar, LiDAR, etc., to promote the ability to obtain depth information. Meanwhile, it is also possible to explore the motion trajectory and shape reconstruction of underwater objects, providing more information and support for the navigation and operation of underwater robots.

2. Related Work

Recent research has shown that underwater stereo matching methods based on image semantic segmentation have made significant progress in improving the accuracy of stereo matching [35–39]. Utilizing semantic segmentation technology to identify the semantic information of different regions in underwater images, this method can provide more semantic context and constraints for the stereo matching process, thereby improving the accuracy and stability of underwater stereo matching. Fangfang Liu et al. [40] discussed the application of image semantic segmentation technology in underwater scenes, with a focus on improving the segmentation accuracy of underwater images. By introducing an unsupervised color correction method (UCM) module, the DeepLabv3+semantic segmentation network was extended, and an upsampling layer was added to preserve more target features and object boundary information. Compared with the original method, the segmentation accuracy was improved by 3%. This study emphasizes the importance of exploring marine resources through advanced image processing techniques. Junhao Liu et al. [9] introduced a new method for improving cross modal image text retrieval utilizing image-to-text and text-to-image generation models within the “dual teacher one learning” learning framework. Xinchen Y, et al. [24] introduced a module consisting of style adaptation, semantic adaptation, and parallax range adaptation to adapt land depth estimation models to underwater environments. By synthesizing programmatic underwater stereo images from ground data, semantic domain differences were minimized and the problem of disparity range mismatch was solved. Compared with existing methods, this method has

achieved excellent performance in underwater depth estimation. Jiawei Zhang et al. [23] proposed a method that combines image enhancement techniques with an improved semi global block matching (SGBM) method to improve stereo matching accuracy in underwater environments using semantic segmentation results. Xiaowei Yang et al. [41] proposed learning geometric information through a separate processing branch called edge flow, integrating edge clues into stereo matching. They also introduced multi-scale cost measures in hierarchical cost aggregation to capture structures and global representations, thereby enhancing the accuracy of scene understanding and disparity estimation. In addition, a disparity refinement network with dilated convolution was applied to further improve the accuracy of the final disparity estimation. Xuewen Yang et al. [42] designed an underwater self-supervised monocular depth estimation framework that utilizes the relationship between underwater light attenuation and depth changes to enhance depth extraction, emphasizing the importance of addressing underwater perception challenges and contributing to the improvement of underwater robot perception and ocean exploration. Based on current research in the field of underwater image processing, although some progress has been made in segmentation accuracy, cross modal retrieval, and depth estimation, there are still some common limitations. Specifically, the accuracy improvement of the method is limited, its adaptability is limited, and there is also room for improvement in information extraction and perception. These shortcomings need to be further expanded and discussed in detail to promote the improvement of methods and enhance the efficiency and accuracy of underwater image processing tasks.

By combining deep learning techniques with semantic segmentation networks, semantic information in images can be effectively extracted, providing richer constraints and contextual information for stereo matching. This article investigates the unsupervised stereo matching method based on semantic attention, which can help algorithms automatically learn the importance of different regions and improve the quality of stereo matching results. The innovation of this method lies in the combination of the powerful characteristics of deep learning and the ability of semantic segmentation networks utilizing semantic information and attention mechanisms to guide the unsupervised stereo matching process, thereby improving the accuracy and overall performance of matching. The workflow diagram of the underwater unsupervised stereo matching method based on semantic attention proposed in this article is shown in Figure 1.

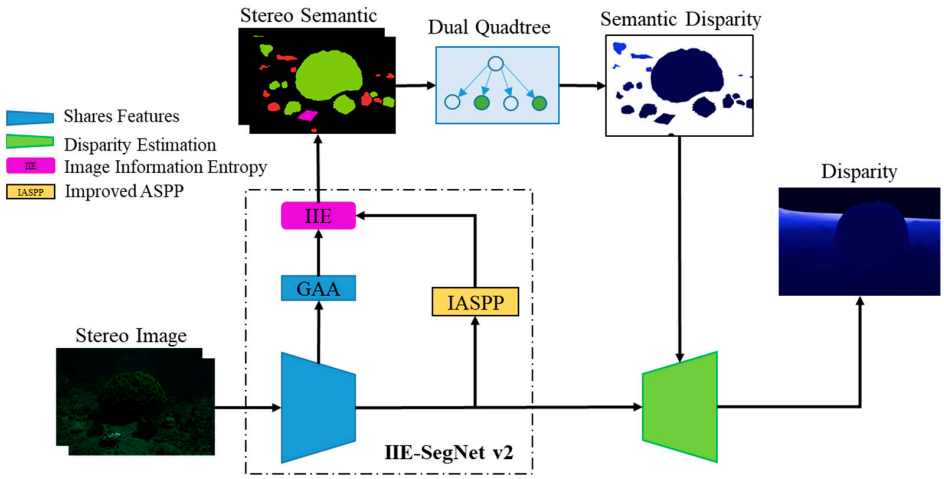


Figure 1. Workflow of underwater unsupervised stereo matching based on semantic attention.

3. Research Methods

3.1. IIE-SegNet-v2

The input underwater images are processed through the feature extraction module of the convolutional layer of VGG16 to obtain the output features of five pooling layers. The interval feature image information entropy is calculated for the pooling layer features

of the first, second, and third layers. The fifth pooling layer is connected to an improved ASPP module and upsampled to the scale of the previous layer along with the interval information entropy of each pooling layer. Finally, the semantic segmentation map of the input underwater image is obtained (see Figure 2).

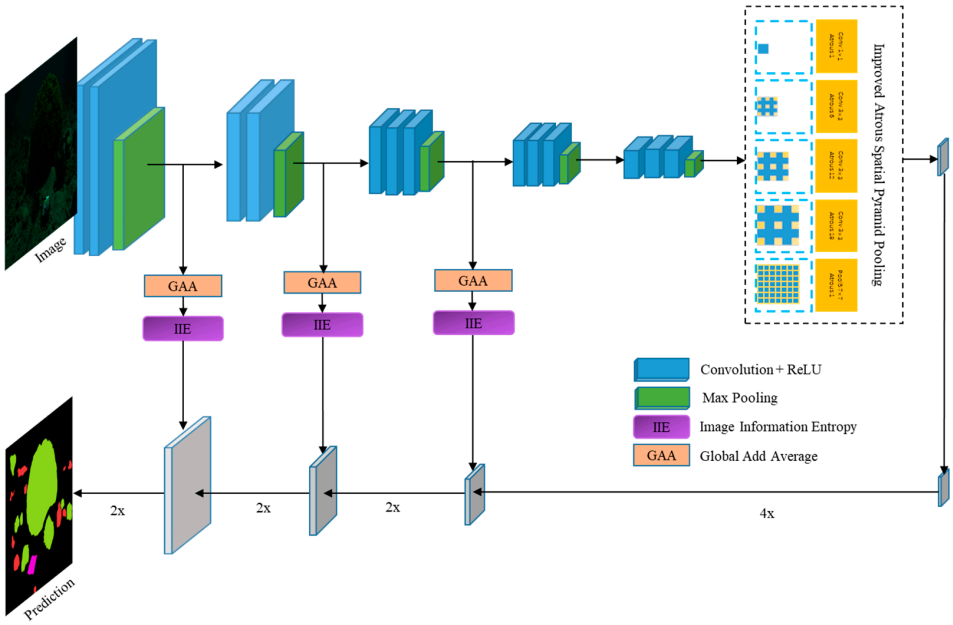


Figure 2. IIE-SegNet-v2 architecture.

3.2. Adaptive Double Quadtree Attention Model

This article designs a network structure based on adaptive double quadtree to calculate the disparity of semantic features provided by binocular semantic segmentation. The complete two-dimensional semantic segmentation image is modeled as an adaptive double quadtree, and the positions of semantic features are learned and evaluated. The feature representations on rough level quadtree nodes contain a lot of background information, while fine level quadtree nodes contain more precise semantic information, as shown in Figure 3.

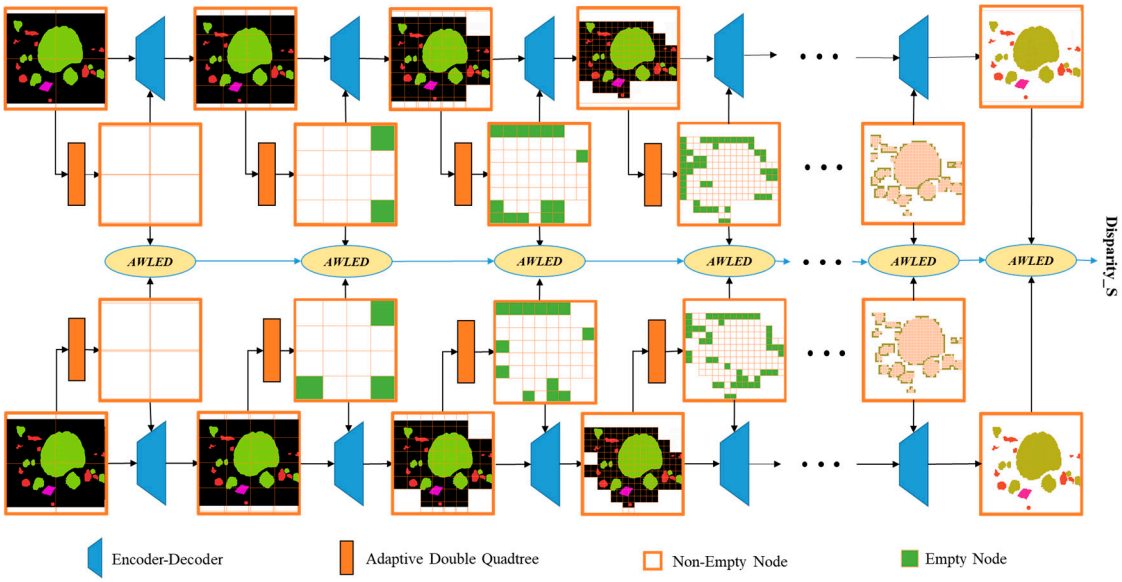


Figure 3. Adaptive double quadtree stereo matching network architecture.

3.2.1. Quadtree

In order to develop a quadtree segmentation distance threshold for underwater stereo matching datasets, the binocular semantic segmentation images extracted from the underwater stereo matching dataset were examined, and a strong positive correlation (0.75) was observed between the ratio of feature energy to the image size and the distance to the correct category. By performing linear regression on the data, the threshold function in Formula (1) can be obtained.

$$r = 82.057 \cdot \frac{\|v\|}{s} - 1.748 \tag{1}$$

Among them, r is the threshold used for the Mahalanobis distance, $\|v\|$ is the energy of the transformed feature vector, and s is the number of pixels in the image block.

Firstly, the position of the first rectangular box is determined by the 0th and 1st moments of the image, and the size of the first rectangular box is determined by right entering the image size. Starting from the top left corner of this rectangular box as the first point, iteratively split each node to form a quadtree structure for its child nodes. This quadtree structure generates 4 child objects for each parent node until it reaches the point where there are no separable parent nodes. Calculate the distance threshold for each child node using Formula (1) as a function of the feature energy and block size of the child nodes. Then, measure the distance from the parent node to each child node in the feature space. If the distance to all four child nodes is within the corresponding threshold of the child nodes, declare a semantic feature in the parent node, as shown in Figure 2. Mark the parent node as a terminal node and delete the child node. If there is no single semantic feature, perform paired grouping testing of subobjects. When the distance between the two nearest children is lower than the corresponding threshold, the children merge. When there are no child nodes approaching, all child nodes remain as parent nodes, and the quadtree iteration continues to group on each child node.

3.2.2. Adaptive Weighted Euclidean Distance

Adaptive weighted local Euclidean distance (AWLED) calculates depth values based on the weighted Euclidean distance (WED) [43] of the labels corresponding to the left and right views. Assuming that the mean of X in the sample set is m and the standard deviation is s, then the "standardized variable" of X is represented as follows:

$$X^* = \frac{X - m}{s} \tag{2}$$

The standard Euclidean distance (SED) [44] formula is as follows:

$$d_{12} = \sqrt{\sum_{k=1}^{n} \left(\frac{x_{1k} - x_{2k}}{s_k} \right)^2} \tag{3}$$

Among them, s_k represents the standard deviation of each dimension. If the reciprocal of variance is considered as a weight, this method can also be called the weighted Euclidean distance.

$$d_{1r} = \sum_{ij} \sqrt{\left(S_{ij}^l - S_{ij}^r \right)^2} = \sum_{ij} \sqrt{\delta_1 \left(x_i^{s_l} - x_i^{s_r} \right)^2 + \delta_2 \left(y_j^{s_l} - y_j^{s_r} \right)^2} \tag{4}$$

Here, δ_1 and δ_2 are weights, where $\delta_1 > \delta_2$ and $\delta_1 + \delta_2 = 1$.

3.2.3. Adaptive Quadtree

Generally, the weighted Euclidean distance is effective but slow, and so a window-based weighted local Euclidean distance is used. Due to the diverse matching range from large to small in this study, the adaptive window was chosen.

$$d_S = \sum_{ij} a_k w_{ij} \sqrt{\delta_1 \left(x_i^{s_l} - x_i^{s_r}\right)^2 + \delta_2 \left(y_j^{s_l} - y_j^{s_r}\right)^2} \tag{5}$$

Among them, a_k is the adaptive parameter, and w_{ij} is the window size. $\left(x_i^l, y_j^l\right)$ is the left image coordinate, and $\left(x_i^r, y_j^r\right)$ is the right image coordinate.

The final semantic disparity calculated by the adaptive weighted local Euclidean distance is shown as follows:

$$D_{\text{seg}} = \frac{1}{n} e^{-\lambda} \sum_{i=0}^{n} \beta_i d_{S_i} \tag{6}$$

Here, β_i is the weight, λ is the hyperparameter, and the size of n is determined by the number of adaptive quadtrees.

There are multiple local minima for a single pixel loss, and there is a significant deviation between these local minima and the ideal optimal value. This thesis uses an adaptive window approach to more accurately estimate the disparity error of pixels, as shown in Figure 4. This approach is more robust than traditional stereo matching processes because it can add up all costs based on the correlations between adjacent pixels, resulting in more accurate results. In Figure 3, it can be seen that the disparity error of the method of this thesis is close to the 1×1 window, but the disadvantage of being a large window is that small objects and details cannot be restored in the final disparity. The results obtained from the adaptive window in this article are much smaller in error than those obtained from a fixed proportion window. The adaptive window achieves a pixel by pixel calculation effect and significantly reduces the computational complexity.

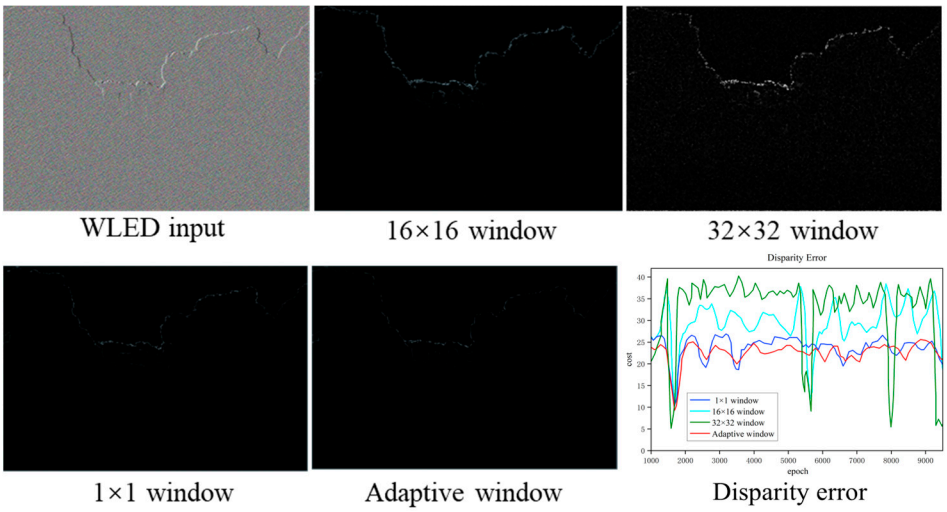


Figure 4. Parallax errors under different windows.

3.3. Overall Network Structure

Dual stream IIE SegNet v2 [45], as shown in the green dashed box in Figure 5, obtains the semantic segmentation map of the left image and the semantic segmentation map of the right image through IIE SegNet v2, respectively, and shares weights with double stream IIE SegNet v2.

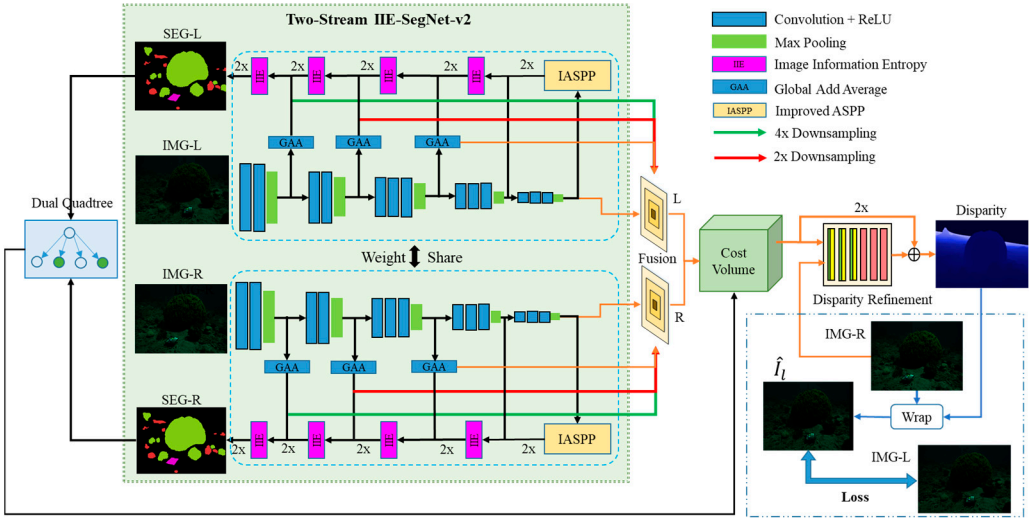


Figure 5. Underwater stereo matching unsupervised network based on semantic attention.

Firstly, the CNN layer of the dual stream IIE SegNetv2 encodes the local and global context, and the GAA model outputs from Pool 1 to Pool 3, along with the output from Pool 5, form multi-scale features for 1/8 of the low-resolution cost volume, providing multi-scale features for left and right images. Then, semantic features are provided by the dual stream IIE SegNet v2 and fed into the adaptive double quadtree for semantic attention matching. The obtained disparity is fed into 1/8 of the low-resolution cost volume for the initial disparity estimation, and finally fed into the disparity refinement network.

3.4. Unsupervised AWLED Semantic Loss

Unsupervised stereo matching algorithms have slight differences in the calculation of loss functions compared to supervised learning. The loss function of supervised learning is obtained by comparing the disparity calculated from the left and right images with the true disparity value. However, it is difficult to obtain the true value of disparity in stereo matching, and so the loss function in this article is the unsupervised loss obtained by comparing the left image with the reconstructed left image.

The loss function of unsupervised stereo matching is that one image in a stereo image pair can be reconstructed from another image using a disparity map. The smaller the difference between the two, the more accurate the disparity map is. The loss L_p is obtained by distorting the right image I^{r} by disparity D to reconstruct the left image $\hat{I}^l$ and the left image truth value I^l, which are calculated using the $L1$ normal form as follows:

$$L_p = \frac{1}{N}\sum_{i,j}\delta^p_{i,j}\left\|\hat{\mathrm{I}}^l_{i,j} - \mathrm{I}^l_{i,j}\right\|_1 \tag{7}$$

Here, N represents the number of pixels, and $\delta^p_{i,j}$ is to avoid outliers such as image edges or occluded areas, or the absence of corresponding pixels. If the luminosity at pixel (i,j) is greater than the threshold ε, $\delta_{i,j} = 0$, otherwise $\delta_{i,j} = 1$.

Semantic information clues can also guide disparity learning as a loss term. Based on the predicted semantic disparity map D_{seg}, feature distortion is used on the right segmentation map F^r_s to obtain the reconstructed left semantic segmentation map $\hat{F}^l_s$, and the left semantic segmentation map truth label F^l_s is used as guidance to learn the pixel classification. Finally, the semantic cue guidance loss L_s is measured between the classification distortion map and the truth labels.

$$L_s = \frac{1}{N_s}\sum_{i,j}\left\|\delta^S_{i,j}(F^l_s - \hat{F}^l_s)\right\|_1 \tag{8}$$

Among them, N_s is the number of semantic pixels, and $\delta_{i,j}^S$ is to avoid outliers, such as image edges or occluded areas, or the absence of corresponding pixels. If the semantic information clue at pixel (i, j) is greater than the threshold ξ, $\delta_{i,j}^S = 1$, otherwise $\delta_{i,j}^S = 0$.

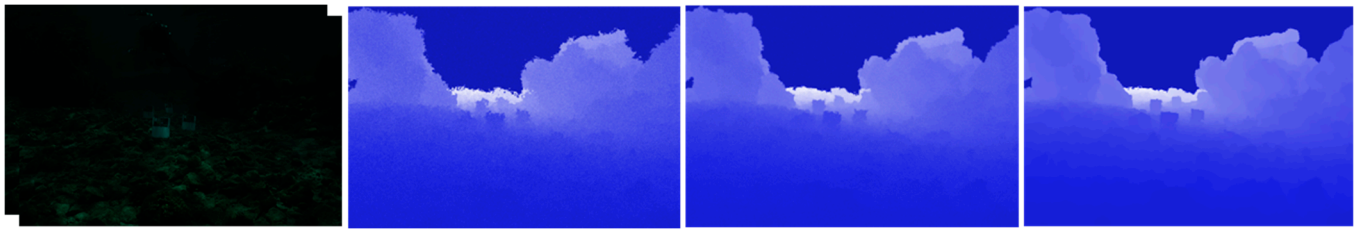
Because it is an unsupervised learning stage, when training the disparity network, the semantic loss L distorts the right semantic segmentation map F_s^r feature obtained from the semantic disparity D_{seg} obtained by AWLED, resulting in the reconstructed left semantic segmentation map F_s^l. Then, the left segmentation image truth F_s^l is used for supervision, and the semantic loss L also applies additional object perception constraints to guide disparity training in resolving local differences and ambiguities.

$$L = L_s + \lambda L_r \tag{9}$$

$$L_r = \frac{1}{N_s}\sum\left(\left|\nabla_x^2 D_{seg}\right| e^{-|\nabla_x^2 F_s^l|} + \left|\nabla_y^2 D_{seg}\right| e^{-|\nabla_y^2 F_s^l|}\right) \tag{10}$$

Among them, N_s is the number of semantic pixels, λ is the penalty factor, and ∇_x^2 and ∇_y^2 are the second-order derivatives along the x-axis and y-axis directions. This thesis utilizes the left and right consistency property of stereo matching as a pixel consistency constraint to achieve unsupervised learning of disparity maps.

As shown in Figure 6, the addition of semantic attention based the allocation calculation here is much smoother than the disparity error of MADNet [46] and Xchen Y, et al. [24]. The loss proposed in this thesis preserves complex and small structures and edges well, without getting stuck in local minima and losing details.



(**a**) (**b**) (**c**) (**d**)

Figure 6. Parallax error plots under different loss functions. (**a**) Input stereo image, (**b**) MADNet [46], (**c**) Xchen Y, et al. [24], and (**d**) Proposed method.

4. Experiment and Result Analysis

4.1. Underwater Stereo Matching Dataset

The underwater stereo matching dataset [47] contains 57 stereo pairs from four different sites in Israel, two in the Red Sea (representing tropical water), and two in the Mediterranean Sea (temperate water). In the Red Sea, the sites were a coral reef ('Katzaa'), which is 10–15 m deep (fifteen pairs), and a shipwreck ('Satil'), 20–30 m deep (eight pairs). In the Mediterranean Sea, both sites were rocky reef environments separated by 30 km, Nachsholim at a 3–6 m depth (13 pairs) and Mikhmoret at a 10–12 m depth (21 pairs). The dive sites' geographical locations are displayed in Figure 7 (right panel). In addition, all of the images were taken using a pair of DSLR cameras (Nikon D810 (Nikon Corporation, Phra Nakhon Si Ayutthaya, Thailand) with an AF-SNIKKOR 35 mm f/1.8G ED lens, encased in a Hugyfot housing with a dome port) on a rigid rig, as shown in Figure 7 (upper left panel).

4.2. Network Training

The implementation of an unsupervised stereo matching network for robots based on semantic segmentation in this article is based on Pytorch. This work uses the "multiple" learning rate strategy here. The parameters for training and testing the computer are an Intel Xeon E5-2630 V4 CPU (10 cores; 20 threads; lithography: 14 nm; processor base frequency: 2.20 GHz; maximum Turbo frequency: 3.0 GHz; third level cache: 25 m), and

the GPU is NVIDIA GeForce GTX 1080ti (NVIDIACUDA kernel: 3584; standard memory configuration: 11 GB).

The full resolution training/testing of underwater stereo matching datasets is crucial for unsupervised stereo matching systems to obtain an accurate depth. However, during the training process, the model cannot perform full resolution training of 2212 × 1476 in 11 GB of memory. In order to still train at full resolution, a region with 512 × 512 pixels is randomly selected and cropped from the same region in both the left and right images. This does not change parallax, and so models trained on a small parallax can be directly validated at full resolution during testing. When training the dataset, adjust the basic learning rate to 0.01, the power to 0.9, and the momentum and weight decay to 0.09 and 0.001, respectively, to ensure the accuracy and reliability of the model. These parameters of the learning strategy are maintained in an unsupervised learning process.

Figure 7. Creating the dataset.

4.3. Testing and Evaluation

4.3.1. Evaluation Indicators

1. Avg All is the endpoint error (EPE) for all regions:

$$\frac{1}{N_{all}} \sum_{(x,y)\in N_{all}} \left\{ \left| d_{est}(x,y) - d_{gt}(x,y) \right| \right\} \tag{11}$$

2. Evaluate all regions of the first frame image (D1 all):

If the disparity or flow endpoint error $\left| d_{est} - d_{gt} \right| < 3$px or $\frac{\left| d_{est}(x,y) - d_{gt}(x,y) \right|}{d_{gt}} < 5\%$, it is considered a correct estimate.

4.3.2. Qualitative and Quantitative Evaluations

To characterize accuracy, the loss function and the calculation of 3 pixel accuracy are defined as predicting the deviation between the left image and the right image, as shown in Figure 8.

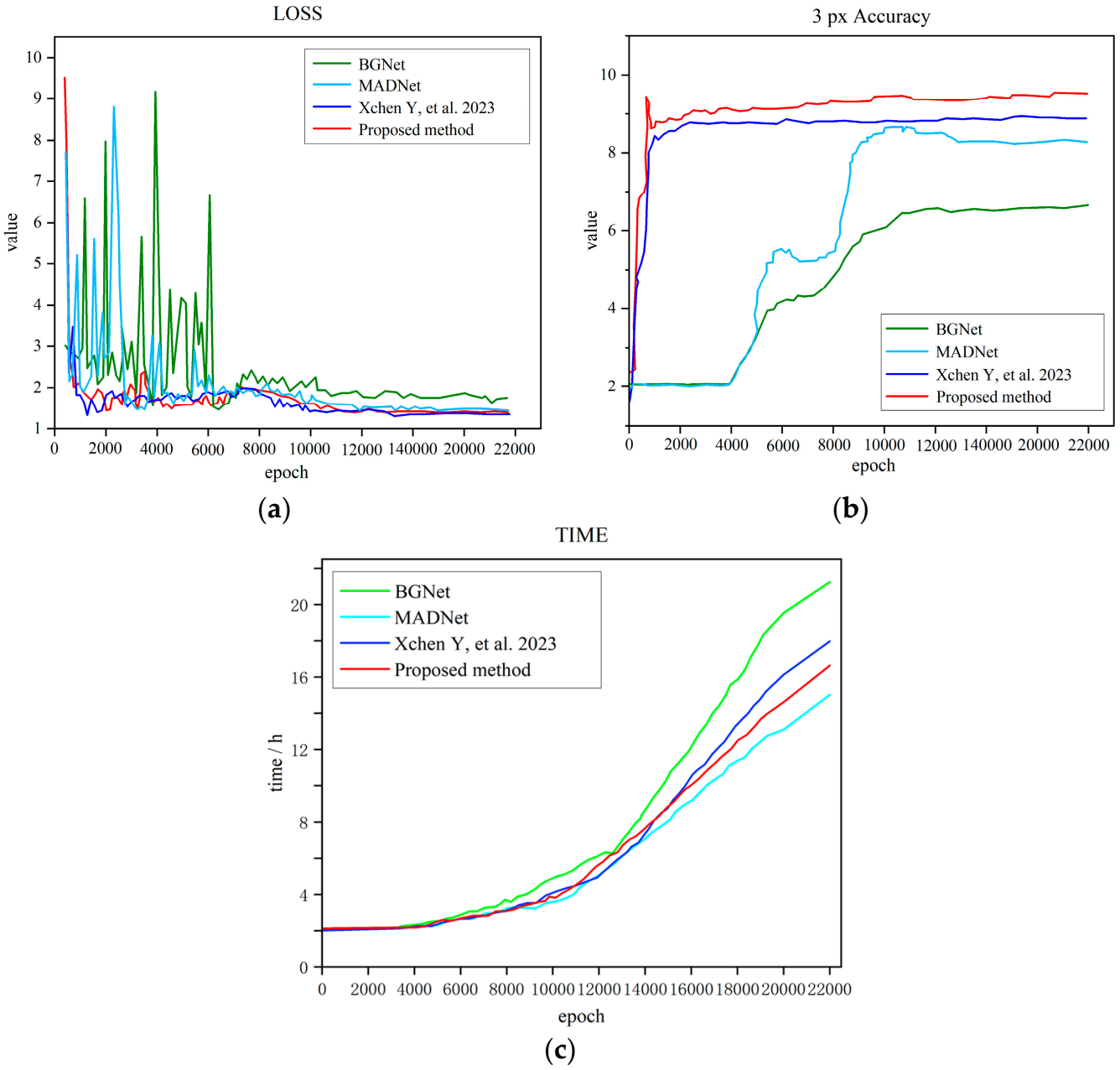


Figure 8. Loss function and 3 px accuracy [24]. (**a**) Training loss, (**b**)3 px accuracy, (**c**) Training time.

Through the comprehensive analysis of Figure 8a–c, it can be seen that compared with the depth deviation of the models trained by our method, BGNet, MADNet, and Xchen Y, et al. [24], the This work method can effectively reduce the depth error of the model output, and the training loss value can converge quickly. The 3 pixel accuracy quickly stabilizes, while the other three methods have varying degrees of deviation. The time is not significantly different among the four methods when the loss reaches stable convergence after 12,000 iterations.

When viewing the disparity map, special attention should be paid to the various noise differences, disparity deviations, enlarged edges of the disparity map, flying pixels of the disparity map, resolution, holes, and other evaluation methods that exist in the observed results map. Qualitative results should be observed and analyzed, as shown in Figure 9.

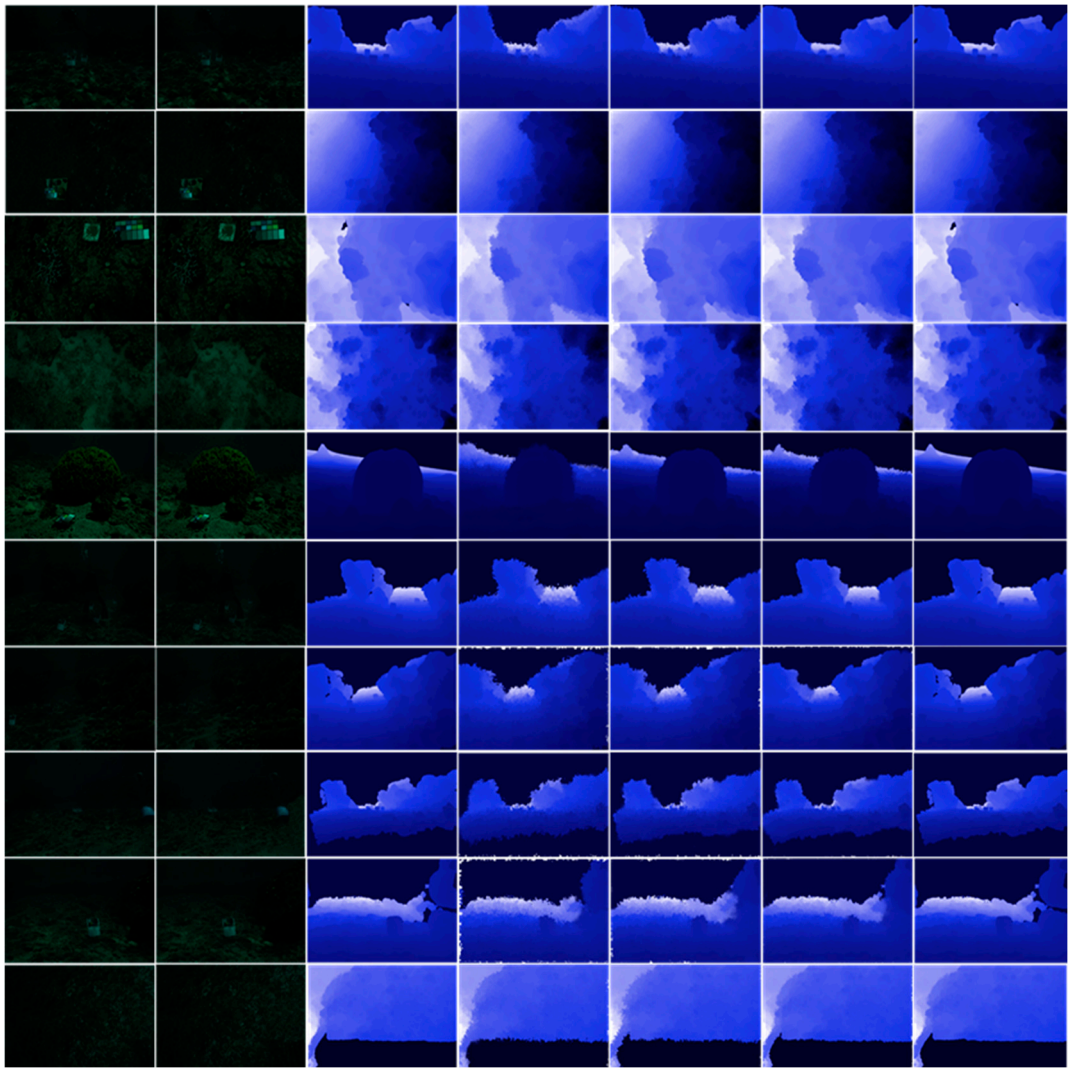


Figure 9. Qualitative evaluation of underwater stereo matching datasets. (**a**) Left_I, (**b**) Right_I, (**c**) GT, (**d**) BGNet [48], (**e**) MADNet [46], (**f**) Xchen Y, et al. [24], and (**g**) Proposed method.

As shown in Figure 9, BGNet, MADNet, Xchen Y, et al. [24], and proposed method were compared on an underwater stereo matching dataset. From the qualitative experimental results, it can be intuitively analyzed that in terms of noise, proposed method is superior to other methods. In terms of the resolution analysis, proposed method is superior to other methods. In terms of edge preservation, proposed method is also superior to other methods. Based on the comprehensive qualitative evaluation results, proposed method achieved the optimal result.

As shown in Table 1, proposed method-basic represents the unsupervised stereo matching method without incorporating the semantic attention model based on the dual quadtree. Proposed method-SS represents the method that directly uses left–right semantic segmentation for the disparity prediction without passing through the dual quadtree model. Meanwhile, proposed method represents the unsupervised stereo matching method proposed in this paper based on semantic attention. A comparative analysis of proposed method-basic, proposed method-SS, and proposed method reveals that the proposed method model excels in the absolute error rate, EPE, and 3-pixel accuracy, with a shorter runtime, demonstrating the best performance. In comparison, the proposed method-SS model slightly outperforms proposed method-basic in the absolute error rate but shows

better performance for EPE and 3-pixel accuracy, despite the longer runtime. Considering both performance and efficiency, the proposed method model outperforms overall.

Table 1. Quantitative evaluation of the underwater stereo matching dataset.

Model	D1-All	EPE	3px Error	Runtime
UWStereoNet [49]	0.857	17.96	35.12	3200 ms
MUNet [50]	0.594	9.76	12.87	1200 ms
BGNet [48]	0.727	12.93	16.32	263 ms
MADNet [46]	0.801	15.26	29.65	202 ms
Xchen Y, et al. [24]	0.547	6.55	10.11	263 ms
Proposed method-basic	0.672	5.98	15.69	232 ms
Proposed method-SS	0.505	5.08	9.84	368 ms
Proposed method	0.325	3.98	8.58	256 ms

The comprehensive quantitative comparison and analysis in Table 1 indicates that the proposed method model performs exceptionally well in underwater stereo matching tasks. Specifically, proposed method D1-all achieves 0.325, which is significantly lower than other models, such as UWStereoNet (0.857), MUNet (0.594), BGNet (0.727), MADNet (0.801), and Xchen Y et al. (0.547). In terms of the EPE metric, proposed method records a value of 3.98, which is markedly better than the other models, demonstrating more accurate stereo matching results. Furthermore, proposed method excels in 3 px accuracy with a score of 8.58, outperforming its competitors. Regarding runtime efficiency, proposed method demonstrates a rapid processing time of 256 milliseconds, showcasing superior efficiency compared to other models like UWStereoNet (3200 ms), MUNet (1200 ms), BGNet (263 ms), and Xchen Y et al. [24]. (263 ms), but is slightly inferior to MADNet (202 ms). Overall, proposed method not only exhibits outstanding precision and accuracy but also possesses clear advantages in efficiency and speed.

In addition, as shown in the qualitative analysis in Figure 9 and quantitative analysis in Table 1, the test results of this algorithm on the underwater binocular stereo matching dataset show that D1 all has been improved, with excellent performance in terms of 3-pixel error and EPE, but is slightly less than MADNet in terms of time. The effectiveness of this algorithm can be clearly seen through comprehensive quantitative and qualitative analyses.

5. Conclusions and Future Work

This thesis proposes an underwater unsupervised stereo matching method based on semantic attention to solve the problem of missing training supervised datasets in underwater stereo matching. On the basis of the semantic segmentation network IIE SegNet v2, this article designs an adaptive dual quadtree-based semantic attention module and unsupervised AWLED semantic loss, and solves the problem of mismatching in weak lighting, noise, and textureless areas in underwater environments; improves the stereo matching accuracy of underwater targets; and demonstrates robustness against noise and textureless areas.

The research results of this article are satisfactory and provide an effective depth acquisition method for underwater stereo matching. Through experimental verification of the underwater stereo matching dataset, better test results were achieved. Compared with the optimal evaluating indicators in advanced methods, D1 all decreased by 0.222, EPE decreased by 2.57, 3 px error decreased by 1.53, and the time increased by 54 ms, effectively improving the matching accuracy.

The next step of research will focus more on building adaptive learning systems and applying reinforcement learning techniques to achieve autonomous learning and optimization of models in constantly changing environments. Through continuous learning

and enhanced decision-making capabilities, these technologies will drive practical applications and performance improvements in the fields of stereo matching, enabling the system to handle complex underwater scenes more flexibly and intelligently and achieve better results.

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

Funding: This study was supported by National Science and Technology Innovation Special Zone Project (No. 21-163-05-ZT-002-005-03), the National Key Laboratory of Underwater Robot Technology Fund (No. JCKYS2022SXJQR-09), and a special program to guide high-level scientific research (No. 3072022QBZ0403).

Data Availability Statement: Data are contained within the article.

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

References

1. Kuppuswamy, R. Method to Profile the Maintenance Needs of a Fleet of Rotating Machine Assets using Partial Discharge Data. In Proceedings of the 2020 Electrical Insulation Conference (EIC), Knoxville, TN, USA, 22 June–3 July 2020; pp. 268–289.
2. Corti, N.; Bonali, F.L.; Pasquaré Mariotto, F.; Tibaldi, A.; Russo, E.; Hjartardóttir, Á.R.; Einarsson, P.; Rigoni, V.; Bressan, S. Fracture Kinematics and Holocene Stress Field at the Krafla Rift, Northern Iceland. *Geosciences* **2021**, *11*, 101. [CrossRef]
3. Hożyń, S.; Żak, B. Stereo Vision System for Vision-Based Control of Inspection-Class ROVs. *Remote Sens.* **2021**, *13*, 5075. [CrossRef]
4. Gerlo, J.; Kooijman, D.G.; Wieling, I.W.; Heirmans, R.; Vanlanduit, S. Seaweed Growth Monitoring with a Low-Cost Vision-Based System. *Sensors* **2023**, *23*, 9197. [CrossRef]
5. Zuo, Y.; Guan, H.; Duan, F.; Wu, T. A Light Field Full-Focus Image Feature Point Matching Method with an Improved ORB Algorithm. *Sensors* **2023**, *23*, 123. [CrossRef]
6. Torkaman, H.; Fakhari, M.; Karimi, H.; Taheri, B. New Frequency Modulation Strategy with SHE for H-bridge Multilevel Inverters. In Proceedings of the 4th International Conference on Electrical Energy Systems (ICEES), Hangzhou, China, 5–7 July 2024; pp. 157–161.
7. Madison, N.; Schiehll, E. The Effect of Financial Materiality on ESG Performance Assessment. *Sustainability* **2021**, *13*, 36–52. [CrossRef]
8. Pan, Y.L.; Chen, J.C.; Wu, J.L. A Multi-Factor Combinations Enhanced Reversible Privacy Protection System for Facial Images. In Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 5–9 July 2021; pp. 1–6.
9. Liu, J.; Yang, M.; Li, C.; Xu, R. Improving Cross-Modal Image-Text Retrieval with Teacher-Student Learning. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 3242–3253. [CrossRef]
10. Zuo, Y.; Yao, H.; Xu, C. Category-Level Adversarial Self-Ensembling for Domain Adaptation. In Proceedings of the 2020 IEEE International Conference on Multimedia and Expo (ICME), London, UK, 6–10 July 2020; pp. 1–6.
11. Zhang, T.; Ding, B.; Hu, Q.; Liu, Y.; Zhou, D.; Galo, W.; Fukuda, H. Research on Regional System Planning Method of Rural Habitat in Gully Regions of the Loess Plateau, under the Background of Rural Vitalization Strategy in China. *Sustainability* **2020**, *12*, 3317. [CrossRef]
12. Shearmana, A.; Zendulkováa, D. Use of National and International Databases for Evaluation of International Project Award Potential. In Proceedings of the 14th International Conference on Current Research Information Systems, CRIS2018, Umeå, Sweden, 13–16 June 2019; pp. 102–111.
13. Jeong, T.; Yun, J.; Oh, K.; Kim, J.; Woo, D.W.; Hwang, K.C. Shape and Weighting Optimization of a Subarray for an mm-Wave Phased Array Antenna. *Appl. Sci.* **2021**, *11*, 6803. [CrossRef]
14. Amer, M.; Laninga, J.; McDermid, W.; Swatek, D.R.; Kordi, B. Very Light Pollution DC Flashover Characteristics of Short Samples of Polymer Insulators. In Proceedings of the 2020 IEEE Conference on Electrical Insulation and Dielectric Phenomena (CEIDP), Virtual, 18–30 October 2020; pp. 143–146.
15. Hu, X.; He, C.; Walton, G.; Fang, Y. Face Stability Analysis of EPB Shield Tunnels in Dry Granular Soils Considering Nonuniform Chamber Pressure and a Dynamic Excavation Process. *Int. J. Geomech.* **2021**, *21*, 04021074. [CrossRef]
16. Bynum, M.; Staid, A.; Arguello, B.; Castillo, A.; Knueven, B.; Laird, C.D.; Watson, J.-P. Proactive Operations and Investment Planning via Stochastic Optimization to Enhance Power Systems' Extreme Weather Resilience. *J. Infrastruct. Syst.* **2021**, *27*, 04021004. [CrossRef]
17. Zhao, Z.; Zhang, H.; Yu, Y. Method for Calculating Text Similarity of Cross-Weighted Products Applied to Power Grid Model Search. In Proceedings of the 2020 IEEE 4th Conference on Energy Internet and Energy System Integration (EI2), Wuhan, China, 30 October–1 November 2020; pp. 3863–3867.

18. Akkoyun, F.; Ercetin, A.; Aslantas, K.; Pimenov, D.Y.; Giasin, K.; Lakshmikanthan, A.; Aamir, M. Measurement of Micro Burr and Slot Widths through Image Processing: Comparison of Manual and Automated Measurements in Micro-Milling. *Sensors* **2021**, *21*, 4432. [CrossRef]
19. Sousa, P.R.; Magalhães, L.; Resende, J.S.; Martins, R.; Antunes, L. Provisioning, Authentication and Secure Communications for IoT Devices on FIWARE. *Sensors* **2021**, *21*, 5898. [CrossRef]
20. Moghimi, A.; Welzel, M.; Celik, T.; Schlurmann, T. A Comparative Performance Analysis of Popular Deep Learning Models and Segment Anything Model (SAM) for River Water Segmentation in Close-Range Remote Sensing Imagery. *IEEE Access* **2024**, *12*, 52067–52085. [CrossRef]
21. da Silva Rocha, É.; Endo, P.T. A Comparative Study of Deep Learning Models for Dental Segmentation in Panoramic Radiograph. *Appl. Sci.* **2022**, *12*, 3103. [CrossRef]
22. Afkir, Z.; Guermah, H.; Nassar, M.; Ebersold, S. Machine Learning Based Approach for Context Aware System. In Proceedings of the IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), Napoli, Italy, 12–14 June 2019; pp. 43–48.
23. Zhang, J.; Han, F.; Han, D.; Su, Z.; Li, H.; Zhao, W.; Yang, J. Object measurement in real underwater environments using improved stereo matching with semantic segmentation. *Measurement* **2023**, *218*, 113–147. [CrossRef]
24. Ye, X.; Zhang, J.; Yuan, Y.; Xu, R.; Wang, Z.; Li, H. Underwater Depth Estimation via Stereo Adaptation Networks. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 5089–5101. [CrossRef]
25. Böer, G.; Gröger, J.P.; Badri-Höher, S.; Cisewski, B.; Renkewitz, H.; Mittermayer, F.; Strickmann, T.; Schramm, H. A Deep-Learning Based Pipeline for Estimating the Abundance and Size of Aquatic Organisms in an Unconstrained Underwater Environment from Continuously Captured Stereo Video. *Sensors* **2023**, *23*, 3311. [CrossRef]
26. Xi, Q.; Rauschenbach, T.; Daoliang, L. Review of Underwater Machine Vision Technology and Its Applications. *Mar. Technol. Soc. J.* **2017**, *51*, 75–97. [CrossRef]
27. Zhanga, X.; Zhangb, Z. Research on stereo matching algorithm of underwater binocular detection. In Proceedings of the Third International Conference on Computer Vision and Pattern Analysis (ICCPA), Hangzhou, China, 31 March–2 April 2023; Volume 12754, pp. 1–11.
28. Liabc, Y.; Sun, K. Review of Underwater Visual Navigation and Docking: Advances and Challenges. In Proceedings of the Sixth Conference on Frontiers in Optical Imaging and Technology: Imaging Detection and Target Recognition, Nanjing, China, 22–24 October 2024; Volume 13156, pp. 1–8.
29. Fayaz, S.; Parah, S.A.; Qureshi, G.J.; Kumar, V. Underwater Image Restoration: A state-of-the-art review. *IET Image Process* **2021**, *15*, 269–285. [CrossRef]
30. Saad, A.; Jakobsen, S.; Bondø, M.; Mulelid, M.; Kelasidi, E. StereoYolo+DeepSORT: A Framework to Track Fish from Underwater Stereo Camera in Situ. In Proceedings of the International Conference on Machine Vision, Edinburgh, UK, 10–13 October 2024; Volume 13072, pp. 1–7.
31. Ishibashi, S. The Stereo Vision System for an Underwater Vehicle. In Proceedings of the OCEANS 2009-EUROPE, Bremen, Germany, 11–14 May 2009; pp. 1–6.
32. John, Y.; Ying, C.; Chen, C. Underwater Image Enhancement by Wavelength Compensation and Dehazing. *IEEE Trans. Image Process.* **2012**, *21*, 1756–1769.
33. Deng, Z.; Sun, Z. Binocular Camera Calibration for Underwater Stereo Matching. *J. Phys. Conf. Ser.* **2020**, *1550*, 032047. [CrossRef]
34. Zhuang, S.; Zhao, Q.; Wang, G.; Song, Y. Analysis of Binocular Visual Perception Technology of Underwater Robot. In Proceedings of the International Conference on Image Processing and Intelligent Control, Kuala Lumpur, Malaysia, 5–7 May 2023; Volume 12782, pp. 1–9.
35. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 1–18.
36. Zhuang, Z.; Li, R.; Jia, K.; Wang, Q.; Li, Y.; Tan, M. Perception-Aware Multi-Sensor Fusion for 3D LiDAR Semantic Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 16260–16270.
37. Chen, L.; Dou, X.; Peng, J.; Li, W.; Sun, B.; Li, H. EFCNet: Ensemble Full Convolutional Network for Semantic Segmentation of High-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *9*, 8011705. [CrossRef]
38. Xu, L.; Ouyang, W.; Bennamoun, M.; Boussaid, F.; Xu, D. Multi-class Token Transformer for Weakly Supervised Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 4300–4309.
39. Zhang, Z.; Wang, B.; Yu, Z.; Zhao, F. Attention Guided Enhancement Network for Weakly Supervised Semantic Segmentation. *Chin. J. Electron.* **2023**, *32*, 896–907. [CrossRef]
40. Liu, F.; Fang, M. Semantic Segmentation of Underwater Images Based on Improved Deeplab. *J. Mar. Sci. Eng.* **2020**, *8*, 188. [CrossRef]
41. Yang, X.; Feng, Z.; Zhao, Y.; Zhang, G.; He, L. Edge supervision and muti-scale cost volume for stereo matching. *Image Vis. Comput.* **2022**, *117*, 104336. [CrossRef]

42. Yang, X.; Zhang, X.; Wang, N.; Xin, G.; Hu, W. Underwater self-supervised depth estimation. *Neurocomputing* **2022**, *514*, 362–373. [CrossRef]
43. Sharma, A.; Gupta, V. A novel approach for depth estimation using weighted Euclidean Distance in stereo matching. *Comput. Vis. Image Underst.* **2017**, *162*, 74–84.
44. Gupta, S.; Sharma, R. An efficient image retrieval method based on standard Euclidean distance. *J. Vis. Commun. Image Represent.* **2018**, *51*, 166–175.
45. Li, Q.; Wang, H.; Li, B.Y.; Yanghua, T.; Li, J. IIE-SegNet: Deep semantic segmentation network with enhanced boundary based on image information entropy. *IEEE Access* **2021**, *9*, 40612–40622. [CrossRef]
46. Poggi, M.; Tonioni, A.; Tosi, F.; Mattoccia, S.; Di Stefano, L. Continual Adaptation for Deep Stereo. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 4713–4727. [CrossRef] [PubMed]
47. Berman, D.; Levy, D.; Avidan, S.; Treibitz, T. Underwater single image color restoration using haze-lines and a new quantitative dataset. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 2822–2835. [CrossRef] [PubMed]
48. Xu, B.; Xu, Y.; Yang, X.; Jia, W.; Guo, Y. Bilateral Grid Learnimg for Stereo Matching Networks. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 12492–12499.
49. Skinner, K.A.; Zhang, J.; Olson, E.A.; Johnson-Roberson, M. UWStereoNet: Unsupervised learning for depth estimation and color correction of underwater stereo imagery. In Proceedings of the International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 7947–7954.
50. Ye, X.; Li, Z.; Sun, B.; Wang, Z.; Xu, R.; Li, H.; Fan, X. Deep joint depth estimation and color correction from monocular underwater images based on unsupervised adaptation networks. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 3995–4008. [CrossRef]

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.