Journal of *Marine Science and Engineering*

MDPI

Article

Spatial Orientation Relation Recognition for Water Surface Targets

Peiyong Gong [1], Kai Zheng [1,*], Yi Jiang [2], Huixuan Zhao [1], Xiao Liang [3], Zhiwen Feng [4] and Wenbin Huang [4]

[1] Marine Electrical Engineering College, Dalian Maritime University, Dalian 116026, China; gpy_dlmu@foxmail.com (P.G.)
[2] Information Science and Technology College, Dalian Maritime University, Dalian 116026, China
[3] Naval Architecture and Ocean Engineering College, Dalian Maritime University, Dalian 116026, China
[4] Marine Design and Research Institute of China, Shanghai 200011, China
* Correspondence: kzh@dlmu.edu.cn

Abstract: Recently, extensive research efforts have concentrated on comprehending the semantic features of images in the field of computer vision. In order to address the spatial orientation relations among water surface targets (WSTs) in an image, which is a fundamental semantic feature, this paper focused on the recognition of spatial orientation relations. We first developed the water surface target spatial orientation vector field (WST-SOVF) algorithm, a novel end-to-end methodology, to recognize these spatial orientation relations among WSTs in an image. The WST-SOVF algorithm encodes the spatial orientation relation into the learning framework of a new deep convolutional neural network model, which comprises two distinct branches: the T-branch and the S-branch, both designed for the spatial feature extraction. The T-branch employs keypoint estimation to identify central points and classify the WST categories, while the S-branch constructs a spatial orientation vector field between WSTs, where each pixel in the field encodes the spatial orientation angle between two separated WSTs and collectively determines the category of spatial orientation. A fusion module was also designed to integrate the spatial feature obtained from both branches, thereby generating a comprehensive triple list that provides not only all the WSTs and their spatial orientation relations, but also their associated confidence levels. We performed a comparative evaluation of our WST-SOVF algorithm based on Huawei's "Typical Surface/Underwater Target Recognition" dataset and the results demonstrated the outstanding performance of WST-SOVF algorithm.

Keywords: water surface target; spatial orientation relation; spatial orientation vector field; deep convolutional neural network; recognition

check for updates

Academic Editor: Rafael Morales

Received: 31 January 2025
Revised: 24 February 2025
Accepted: 27 February 2025
Published: 28 February 2025

Citation: Gong, P.; Zheng, K.; Jiang, Y.; Zhao, H.; Liang, X.; Feng, Z.; Huang, W. Spatial Orientation Relation Recognition for Water Surface Targets. *J. Mar. Sci. Eng.* **2025**, *13*, 482. https://doi.org/10.3390/jmse13030482

Copyright: © 2025 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

The detection and recognition of water surface targets (WSTs) remain crucial for ship navigators and unmanned navigation systems, significantly enhancing overall navigation safety. Effective look-out can provide essential spatial information about WSTs, assisting the assessment of situational awareness and collision risks. Furthermore, identifying the spatial relations among these targets can improve the accuracy of target detecting and tracking, thus contributing to a comprehensive evaluation of the navigation environment and the potential risk of collision. Currently, the visual sensors installed on vessels serve as powerful tools for navigators in the detection and identification of WSTs, as advancements in computer vision have demonstrated remarkable capabilities in object detection and recognition tasks. In addition to giving information on the categories and sizes of the WSTs,

which can be efficiently processed by modern computer vision algorithms, the spatial relations among these targets are also represented in the images captured by the visual sensors. This enables the extraction of spatial information concerning the vessel and the WSTs through computational methods.

The recognition and detection of WSTs based on image have been extensively studied in the literature. Traditional approaches mainly primarily depend on edge detection, color analysis, and morphological processing. Widely used image processing algorithms, such as Canny edge detector and Sobel operator [1], identify the edges of targets by evaluating the contrast between the water surface and the target itself. Nonetheless, these approaches demonstrate limited effectiveness when faced with complex backgrounds such as waves and splashes as well as low-contrast situations. Recently, numerous automated feature extraction methods based on the complex neural networks have been proposed to improve the efficiency of image recognition. A significant milestone is the discovery of convolutional neural networks (CNNs), which have become one of the mainstream methods of water surface image processing in recent years [2]. CNNs enable the recognition and detection of WSTs through a multi-layered architecture. Several convolutional layers are employed in CNNs to automatically extract local features from the image such as surface fluctuations and target edges. Activation functions and pooling layers subsequently enhance feature representations while reducing computational complexity. In the fully connected layers, the network integrates the extracted features, ultimately using the output layer to classify and localize water surface targets. CNNs have achieved significant advancements in the recognition and detection of WSTs, demonstrating exceptional capability in handling complex backgrounds and low-contrast images. For instance, networks such as VGGNet [3] and GoogleNet [4] excel at extracting precise feature information from large datasets of WST images, enabling accurate target classification. Deep networks like Faster R-CNN [5] allow CNNs to effectively detect water surface targets, such as ships and buoys, maintaining high recognition accuracy even in the presence of challenging backgrounds like waves and splashes. Furthermore, the YOLO [6–11] model has been successfully applied to real-time target detection, facilitating the rapid localization and classification of water surface targets, which substantially enhances the efficiency of real-time monitoring. Models such as fully convolutional networks (FCNs) [12] and U-Net [13] employ pixel-level image segmentation to successfully isolate targets from the water surface background, thereby improving detection accuracy in complex environments including those with waves and splashes. General speaking, CNNs, through their automated feature learning, end-to-end training processes, and deep network architectures, have significantly enhanced the accuracy, efficiency, and robustness of water surface target recognition and detection.

Although CNNs have shown significant achievement in WST recognition tasks, they still face significant challenges. A notable issue observed in many neural network models is that deep semantic information in images has not yet received sufficient attention during the object recognition process. Particularly, it may lead to inaccurate results when the spatial relations between features are not properly represented. Figure 1 presents a misclassification case, where Figure 1a shows an image containing a sailboat, and Figure 1b illustrates an image containing a wooden boat and a piece of canvas. We evaluated a classification experiment on both Figure 1a,b with the three widely adopted classification models: VGG16 [3], ResNet50 [14], and EfficientNet [15]. All three models classified both Figure 1a,b as “sailboat.” However, the canvas features were positioned inappropriately in Figure 1b, resulting in an obvious misclassification as a “sailboat.” This failure highlights a fundamental shortcoming in conventional CNNs: without explicit modeling of spatial dependencies between features, such architectures struggle to differentiate contextually dissimilar objects (e.g., a sailboat versus a boat with a displaced canvas).



(**a**) Sailboat (**b**) Irrational sailboat

Figure 1. Example images of misclassification.

Figure 2 presents a representative example of a WST image. Compared with natural or urban scene images, WST images exhibit distinctive features, with particularly prominent spatial position relations among these features. Firstly, the background of WST images is typically simple, with targets being highly salient. In open water regions such as lakes or oceans, the background is usually homogeneous, leading to a higher contrast between the targets (e.g., ships, buoys) and the background. Secondly, while the targets in WST images typically exhibit a natural two-dimensional spatial distribution, this distribution often reflects an underlying three-dimensional spatial arrangement. For example, targets *B* and *C* appear in an "up-down" relation in the WST image, but actually correspond to "front-back" spatial position relations in reality. Similarly, while target *C* may appear "below right" of target *A* in the WST image, it corresponds to a "right-front" relation in three-dimensional space. The two-dimensional relations in WST images facilitate intuitive inference of the relative positions and distances of targets in three-dimensional space. Overall, WST images, characterized by their simple backgrounds, salient targets, and two-dimensional spatial distribution, excel in representing spatial positional relations. Compared with images from other scenes, WST images can better highlight target spatial relations under specific conditions. However, despite the clear advantages of WST images in representing spatial relations, the current exploration of deep semantic features, such as spatial positioning and context, remains underdeveloped and warrants further research.

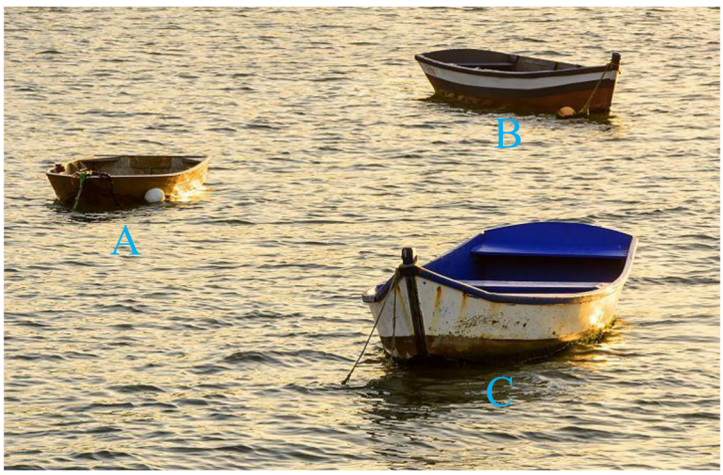


Figure 2. Example images of the water surface scene.

Recent advancements in artificial intelligence have led to significant developments in spatial relation recognition within WST systems, encompassing a diverse spectrum of methodological approaches, aiming to address the inherent limitations in water surface

target recognition and detection while providing innovative frameworks for spatial information analysis. These methods encompass both traditional deep learning architectures and emerging hybrid solutions, with implementations ranging from CNNs and capsule networks to more sophisticated frameworks, collectively advancing the capabilities of spatial understanding in water surface scenarios. Ref. [16] advanced the field by developing a sophisticated object detection framework that demonstrated a remarkable capability in analyzing ship safety plans. Their innovative approach enables the automated verification of safety equipment and signage placement against regulatory requirements, representing a significant step forward in maritime safety compliance. Building upon the YOLOv7 architecture, Ref. [17] introduced the groundbreaking LWS-YOLOv7, a lightweight model specifically engineered for water-surface object detection. Their notable contribution includes the enhancement of the localization loss function and the integration of the w-CIoU function, substantially improving the model's performance on small object detection while optimizing positive and negative sample label allocation. A significant advancement in underwater robotics was achieved by [18], who pioneered a YOLO-based 3D perception algorithm for underwater vehicle-manipulator systems. This sophisticated approach revolutionizes surface and underwater object detection and localization precision, marking a crucial development in underwater manipulation capabilities. Ref. [19] made substantial strides in maritime surveillance by developing an innovative adaptive target detection and tracking model for unmanned surface vehicles. Their research addressed the complex challenges posed by maritime environments, offering enhanced tracking capabilities amid dynamic conditions and diverse environmental factors. Ref. [20] proposed the capsule network, which uses vector input and output. Each capsule corresponds to an entity, with the vector's length representing the entity's probability, and the vector's direction representing the entity's attribute. As a class of neural networks, the capsule network can evaluate the relative relations between targets and their spatial orientations. Ref. [21] introduced an image captioning model that could automatically produce descriptive text from an image. This model enables machines to recognize targets in an image and understand how they relate to each other. Actually, image captioning is a general semantic understanding algorithm that can recognize the WST's spatial position relation under certain conditions.

While these traditional deep learning approaches have shown promising results in maritime object detection and spatial recognition, recent advances in deep learning architectures have opened new avenues for improving detection accuracy and robustness. Particularly noteworthy are the developments in enhanced hybrid attention deep learning and spectrum-based hybrid deep learning, which have brought novel perspectives to maritime surveillance, object detection, and spatial relation recognition. Ref. [22] proposed an enhanced hybrid attention-based neural network combining CNNs, BiLSTMs, and attention mechanisms to analyze the spatial-temporal characteristics of ship heave motion, achieving superior prediction accuracy in complex maritime spatial environments. Their research significantly advanced spatial motion prediction through the comprehensive analysis of multi-dimensional spatial features (heave displacement, pitch movements, and wave interactions), establishing a robust framework for maritime spatial relationship understanding. Refs. [23,24] introduced AEFFNet and HAFDN architectures, both leveraging enhanced hybrid attention mechanisms. AEFFNet focuses on small object detection in UAV imagery with innovative feature fusion strategies, while HAFDN combines channel, local, and global spatial attention for micro-video venue recognition. Although not directly tested on maritime targets, their advances in enhanced hybrid attention deep learning provide valuable guidance for improving maritime object detection and spatial recognition. Furthermore, the research by [25,26] on spectrum-based hybrid deep learning, through their exploration of hybrid sensing techniques that incorporated two different spectrum sensing

approaches, provides valuable insights that could be potentially applied to maritime object detection and spatial relationship recognition.

The current research on spatial feature understanding is mainly based on target recognition, and they neither provide a transparent scheme of spatial position relation recognition nor any performance test results. Indeed, as far as the authors know, there are no models to recognize the spatial relation between objects in an image. This highlights the need for further exploration into integrating deep semantic features with spatial relations to improve target recognition and detection accuracy.

Fortunately, although the above research ignored the spatial relation between features, it implies the fact that CNNs can learn the spatial position between the WSTs in an image [27]. Therefore, this paper considered the recognition of spatial orientation between WSTs in images by using a new one-stage algorithm. We first sorted out the relative spatial orientation of the WSTs in the image relative to the reference object, and the triplet (a three-element structure consisting of object A, object B, and the spatial orientation relation type between A and B) was introduced to characterize the spatial orientation relation between the targets. As shown in Figure 3, these spatial orientations can further be converted into four categories according to people's perception of the WST's relative spatial orientation in the image, where Figure 3a–d represents the left and right (L&R) spatial orientation, front and back (F&B) spatial orientation, front left and rear right (FL&RR) spatial orientation, left rear and right front (LR&RF) spatial orientation, respectively. Thus, we can transform the recognition problem of the spatial orientation between WSTs in the image into a classification problem. Furthermore, a human being usually scans from one WST to another in the eyes through the straight line between them when we observe their spatial orientation. This perception mechanism motivated us to encode the spatial orientation into a CNN model. This paper proposed a new algorithm, the WST spatial orientation vector field (WST-SOVF) algorithm, which is the first to recognize four categories of the spatial orientation between WSTs in images. Similar to the way a human being can perceive the spatial orientation between the WSTs in eyes, the WST-SOVF algorithm establishes the spatial orientation vector field between WSTs. The ground-truth of each pixel of the vector field is encoded by the spatial orientation angle of the WST relative to the reference object. All of the pixels in the vector field finally "vote" to determine the category of the spatial orientation relation between a group of WSTs. For the judgment of the WST category, we followed the detection scheme of key points and realized the target category's judgment by estimating the WST's central area. Consequently, a new DCNN with two branches based on ResNet50 [14] was developed. One branch of this DCNN was used to identify the category of WSTs, and the other was used to predict the spatial orientation vector field between each group of WSTs. Moreover, in order to obtain the optimal parameters of the WST-SOVF algorithm, a pair of coupled regression loss functions corresponding to the WST category's judgment and the spatial orientation vector prediction was designed based on the structure of the two-branch neural network. We demonstrate abundant experimental results on Huawei's "Typical Surface/Underwater Target Recognition" dataset and show the efficiency and robustness of the proposed algorithm.



(**a**) L&R (**b**) F&B (**c**) FL&RR (**d**) LR&RF

Figure 3. Four categories of fundamental WST spatial orientation relations.

The rest of the paper is organized as follows. Section 2 discusses the representation of the spatial orientation relation between WSTs in images. Section 3 illustrates the proposed WST-SOVF algorithm in detail. In Section 4, the experimental results demonstrate the superior performance of the proposed WST-SOVF algorithm on a variety of water surface scene images. Section 5 provides an in-depth discussion of the experimental results and limitations of the algorithm. Finally, concluding remarks with the future direction of research are summarized in Section 6.

2. Spatial Orientation Relation Between WST in an Image

Currently, the recognition and detection techniques for WSTs primarily focus on extracting the regional features of the WSTs, while deep semantic features have yet to receive significant attention. As a fundamental semantic characteristic, the spatial patterns between WSTs are rarely discussed. A water surface image may contain complex spatial patterns. As shown in Figure 4, Figure 4a shows a typical WST spatial pattern. The lighthouse is on the right side of the sailboat if the sailboat is considered as the reference; the sailboat is on the left side of the lighthouse if the lighthouse is taken as the reference. This spatial pattern is quite common for the separate objects in the image. Compared with the spatial pattern in Figure 4a, Figure 4b presents a more complex spatial pattern with overlapping water surface objects. The image shows the runabout overlaid head of the cabin boat, indicating that the runabout is in front of the cabin boat. It also provides a clear representation of their spatial relation. Figure 4c further exhibits various spatial patterns in the image. At first glance, the image illustrates the two red lifeboats are on the cruise ship, which is a spatial containment pattern. Additionally, when considering the relative position of the lifeboat in the horizontal direction of the image, it becomes apparent that the lifeboat is to the right of the cruise ship. This spatial pattern is often ignored. Figure 4d shows a scene with multiple open boats and cabin boats containing complex spatial patterns. Any open boat or cabin boat can be the reference, then the spatial pattern between the WST and the reference can be obtained. Therefore, for a given combination of water surface objects, we can observe the specific spatial pattern. These combinations may yield multiple spatial patterns.



(**a**) left and right (**b**) overlap (**c**) containment (**d**) complex spatial patterns

Figure 4. WST spatial patterns in the images.

Although there may be various spatial patterns between water surface objects in an image, this paper only concentrated on the spatial orientation patterns between two separated WSTs. However, there was no strict definition on the representation of the spatial orientation relations between two separated WSTs. As shown in Figure 5, there are three WSTs, *A*, *B*, and *C*, in the pixel coordinate system. To represent the spatial relation between targets *A* and *C*, if target *A* is taken as the reference origin, the spatial relation between the two can be described as "*C* is to the right-front of *A*". Conversely, when using target *C* as the reference origin, the relation is described as "*A* is to the left-rear of *C*". In reality, these two representations of the spatial relations between *A* and *C* are equivalent, with the apparent ambiguity arising from the symmetry inherent in human language. However, machine recognition typically requires determinacy and uniqueness in the representation of spatial relation to enable reliable subsequent operations such as classification, decision-making, and prediction. Therefore,

in order to eliminate the ambiguity between the spatial orientations of targets described in natural language, this paper introduced four fundamental categories of spatial orientation based on the orientation angle between the two water surface objects (i.e., L&R (left and right), F&B (front and back), FL&RR (front left and rear right), and LR&RF (left rear and right front)) through the following formula:

$$\begin{cases} \text{Category 1} & \text{L\&R}, & \partial(O_{ref}, O_{tar}) \in & [-15^{\circ}, 15^{\circ}] \\ \text{Category 2} & \text{F\&B}, & \partial(O_{ref}, O_{tar}) \in & [75^{\circ}, 90^{\circ}] \cup [-75^{\circ}, -90^{\circ}] \\ \text{Category 3} & \text{FL\&RR}, & \partial(O_{ref}, O_{tar}) \in & (15^{\circ}, 75^{\circ}) \\ \text{Category 4} & \text{LR\&RF}, & \partial(O_{ref}, O_{tar}) \in & (-15^{\circ}, -75^{\circ}) \end{cases} \tag{1}$$

where O_{ref} is the center point of the object that is closer to the vertical axis of the pixel coordinate system, O_{tar} is the center point of the another object, $\partial(O_{ref}, O_{tar})$ is the angle between the vector $\overrightarrow{O_{ref}O_{tar}}$ and the horizontal axis in the pixel coordinate system. According to (1), the spatial orientation between two objects shall be invariant, even if the roles of the two targets are interchanged. Consequently, the recognition problem of spatial orientation can be formulated as a classification challenge involving four fundamental spatial orientations between water surface objects within an image.

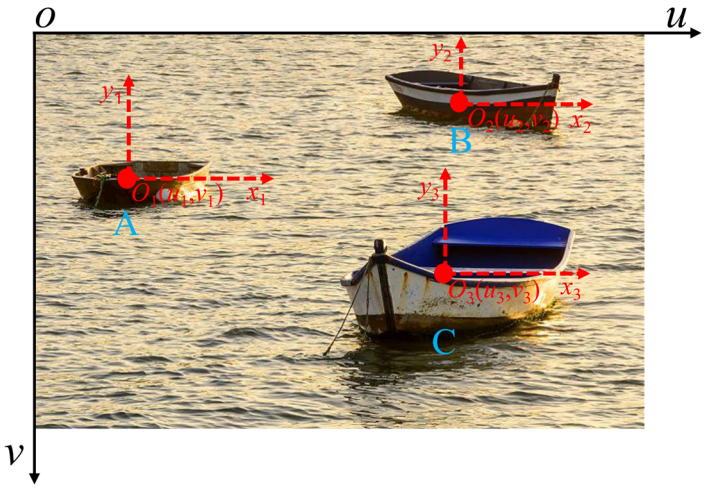


Figure 5. The relation between the pixel coordinate system and image coordinate system.

In representing the spatial orientation relations between objects, it is essential not only to capture the category of relations among a set of objects, but also to clearly define the category between the two involved objects. To address this, the paper introduced the use of triplets to represent spatial orientation relations between WSTs, specifically in the form of <object A, object B, their spatial orientation relation>. In this triplet, the first two elements, A and B, denote the two objects involved in the spatial relation, while the third element specifies the category of their spatial orientation. The order of A and B in the triplet does not affect the relation, allowing for flexibility in representation. This triplet provides a clear, intuitive, and structured way to express the spatial relations between WSTs in an image. Such representation enhances the interpretability of the reasoning process, as each relation can be individually interpreted and verified, making the machine's understanding of the image more transparent and facilitating debugging and refinement of the system.

3. WST Spatial Orientation Relation Recognition Algorithm

3.1. Toward Spatial Orientation Relations Representation

To determine the spatial orientation category between two separated water surface objects to finally output a triplet to characterize the recognition result, this paper investigated two critical aspects of the problem (i.e., the types of water surface objects involved and their corresponding spatial orientations). Actually, the identification of water surface objects can be effectively addressed using various state-of-the-art object detection algorithms such as CenterNet [28–31], CornerNet [29], OpenPose [32], and YOLO [6–11]. The spatial orientations can subsequently be recognized by using the relative localizations obtained after all water surface objects in the image have been identified and localized. However, such a two-phase methodology cannot provide an end-to-end network capable of comprehensively understanding the spatial features of an image. To overcome the limitations inherent in the two-stage framework, it is essential for the recognition algorithm to simultaneously identify both categories. In other words, the algorithm is required to recognize the object and then obtain the category of the spatial orientation relation between the objects, along with the corresponding confidence levels for each water surface object pair in an image, represented as {<object A1, object B1, their spatial orientation>, . . ., <object An, object Bm, their spatial orientation>} [33,34].

Our algorithm was conceptualized as a pixel-level deep learning network featuring two parallel branches connected after an hourglass structure, with the objective of independently identifying WSTs and the spatial orientation relations between these targets. The final output, through a fusion module, generates a representation of the target spatial orientation relations as triplets. As shown in Figure 6, the WST recognition branch leverages current state-of-the-art keypoint detection algorithms to predict a Gaussian distribution (marked with yellow salient areas) at the center of each WST, with the heatmap channel corresponding to the Gaussian distribution indicating the category of the WST. Additionally, to recognize the spatial orientation relations between WSTs, we propose that the pixels within the region between target centers can fully determine these relations. Each pixel in this region, along with the reference target, forms a vector, where the vector orientation represents the "trend" of the spatial orientation relations between the targets. The distribution of all vector "trends" in this region forms the category of the spatial orientation relation, as indicated by the red vector arrows in Figure 6. This region is referred to as the spatial orientation relations vector field of WST *A* and *C*. We can effectively design a spatial orientation relations recognition branch to predict the pixel distribution within this vector field, and subsequently determine the type of field based on the pixel distribution, which in turn allows us to classify the WSTs. Finally, a fusion module is introduced to integrate the predictions from both branches, yielding the output of the spatial orientation relation triplets.

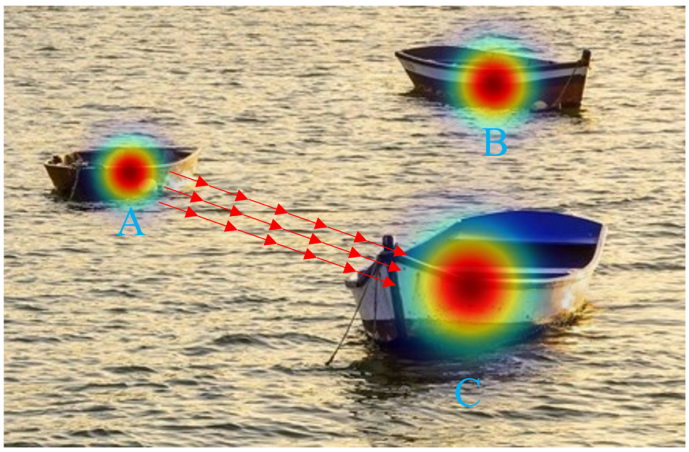


Figure 6. Conceptualization of spatial orientation relation recognition.

3.2. Target Category Recognition Branch Design

The WST category recognition branch (T-branch) was designed to recognize the categories of the WSTs within an image. Inspired by the DCNNs developed for object detection [35,36], which typically extract the feature maps through detecting keypoints associated with the target, the architecture of the T-branch is similarly constructed to generate Gaussian distribution in the various heatmap layers centered around the WSTs, which can be utilized to supervise the prediction of Y.

Let $I \in R^{w \times h \times 3}$ represent an input image of width w and height h, and let $FM \in R^{\frac{w}{z} \times \frac{h}{z} \times 64}$ denote the refined feature map generated by the hourglass structure, where $z = 4$ is the output stride. The FM is processed through the T-branch, which comprises two continuous convolutional layers, to obtain the heatmap $Y \in [0, 1]^{\frac{w}{z} \times \frac{h}{z} \times c}$. In our case, $c = 8$ denotes the categories of water surface objects, which included "lighthouse", "sailboat", "buoy", "cargo ship", "naval vessels", "passenger ship", "submarine", and "fishing boat".

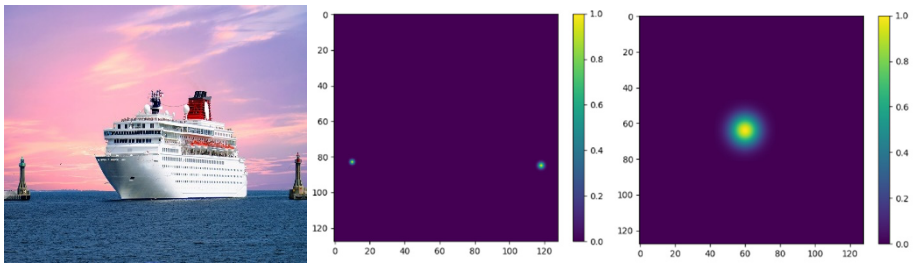
Accurately predicting a WST from a single pixel in the heatmap Y poses significant challenges. Therefore, our objective was to generate a Gaussian distribution centered at the detected WST's location within Y, thereby enhancing the heatmap's features for target localization. Let $C_t = (C_{tx}, C_{ty})$ denote the coordinates of the water surface object's center point in Y. The Gaussian distribution [28–31] around C_t can be obtained as follows:

$$T_c(x, y) = \exp\left(-\frac{(x - C_{tx})^2 + (y - C_{ty})^2}{2\sigma_p^2}\right) \quad (2)$$

where (x, y) denotes the coordinates of the pixel in the heatmap, and σ_p represents a target size-adaptive standard deviation. The subscript c indicates the channel of the Gaussian kernel, with T_c corresponding to the c-th channel associated with the class of the water surface object. The parameter σ_p ensures that the radius of the Gaussian circle is proportional to the water surface object size; typically, σ_p can be set to 1/3 of the water WST size.

The prediction pixel value $Y_{x,y,c}$ denotes the confidence level of a detected WST within the heatmap Y, where (x, y) indicates the coordinates of the pixel in the heatmap while c refers to both the channel of the heatmap and the class of the detected WST. In other words, there are as many heatmap channel types as there are template categories. In other words, the number of heatmap channels corresponds directly to the number of WST categories. The larger the value of $Y_{x,y,c}$, the closer it approaches the center point of the prediction Gaussian distribution $T_x(x,y)$. Notably, $Y_{x,y,c} = 0$ corresponds to the background.

Figure 7 displays the outputs of the T-branch for the original image presented in Figure 7a, which contains two classes of WSTs: two "lighthouse" and one "passenger ship". Figure 7b,c illustrates the ground-truth representations of the "lighthouse" channel and "passenger ship" channel, respectively, generated via the Gaussian kernel centered at the WST's central point. In Figure 7b, the heatmap represents the "lighthouse" class, displaying two Gaussian distributions, T_x (x, y), centered around the "lighthouse", while Figure 7c corresponds to the "passenger ship" class, where the Gaussian distribution, T_x (x, y), was centered on the "passenger ship". Notably, the radius of the Gaussian circle was directly proportional to the size of the WST. Consequently, the heatmap where the Gaussian circle is located indicates the WST category, while the location of the Gaussian circle and the associated pixels can be further employed to recognize the respective spatial orientations between the WSTs.



(**a**) Original image (**b**) "Lighthouse" class (**c**) "Passenger ship" class

Figure 7. The output of the T-branch.

3.3. Spatial Orientation Relations Recognition Branch Design

The spatial orientation relations recognition branch (S-branch) was designed to identify the spatial orientation relations between the WSTs within the image. It generates two spatial orientation vector fields: FB-field and LR-field, between each pair of WSTs, which can be utilized to evaluate the spatial orientation relations between them.

Specifically, the S-branch has a neural network comprising two convolutional layers. The second convolutional layer employs two convolutional kernels to map the feature map FM to a feature map F, which contains spatial orientation vector fields. $F \in [0,1]^{\frac{W}{Z} \times \frac{H}{Z} \times C}$, where W and H represent the width and height of the input image, $Z = 4$ denotes the output stride, and $C = 2$ indicates the number of channels in F. Let $C_t = \{C_{t1}, C_{t2}, C_{t3}, \ldots, C_{tn}\}$ represent the n detected center points of the WSTs in the T-branch. Our objective was to generate two spatial orientation vector fields for each pair of WSTs in F. Consequently, F produces a set F^c of areas (each area is referred to as a field), $F^c = \left\{F_1^{h,c}, F_2^{h,c}, \ldots, F_J^{h,c}\right\}$, where $J = C_n^2$ denotes the number of fields, $h = 4$ indicates that the width of each field is four pixels, and $c \in \{0, 1\}$ indicates that there are only two types of fields, namely the FB-field and LR-field, which correspond to the front and back (F&B) and left and right (L&R) orientations, respectively. For each field, $F_J^{h,c} = \{f_1, f_2, \ldots f_m\}$ denotes the score of this spatial orientation vector field (either FB-field or LR-field), where m indicates the number of pixels within the field.

Figure 8 demonstrates how to generate each spatial orientation relation vector field $F_J^{h,c}$. Taking the encoding of the vector field between the reference object A and the target object B as an example, the center point of reference object A is designated as the origin, thereby establishing an image coordinate system. The ground truth for pixel scores within this field can be encoded according to the following formula:

$$f_m^c = \begin{cases} \left| \sin\left(\arccos\left(\frac{\vec{ax} \cdot \vec{af_m}}{\left\|\vec{ax}\right\| \times \left\|\vec{af_m}\right\|}\right)\right) \right|, & c = 0 \\ \left| \sin\left(\arccos\left(\frac{\vec{ay} \cdot \vec{af_m}}{\left\|\vec{ay}\right\| \times \left\|\vec{af_m}\right\|}\right)\right) \right|, & c = 1 \end{cases} \tag{3}$$

where $c = 0$ indicates the pixel in FB-field, $c = 1$ indicates the pixel in LR-field, $\vec{ax}$ and $\vec{ay}$ represent unit vectors along the x-axis and y-axis, respectively. The vector $\vec{af_m}$ denotes the displacement vector formed by the pixel within the field relative to the coordinate origin. Pixels located outside the field are considered irrelevant to spatial orientation and are assigned 0. Conversely, pixels within the field are subject to superposition, with their values capped at a maximum of 1. This encoding algorithm was designed to ensure that as the spatial orientation relations between WSTs approximate the "F&B" relation, the pixel value in the FB-field will increase, while the pixel value in the LR-field will decrease, and vice versa.

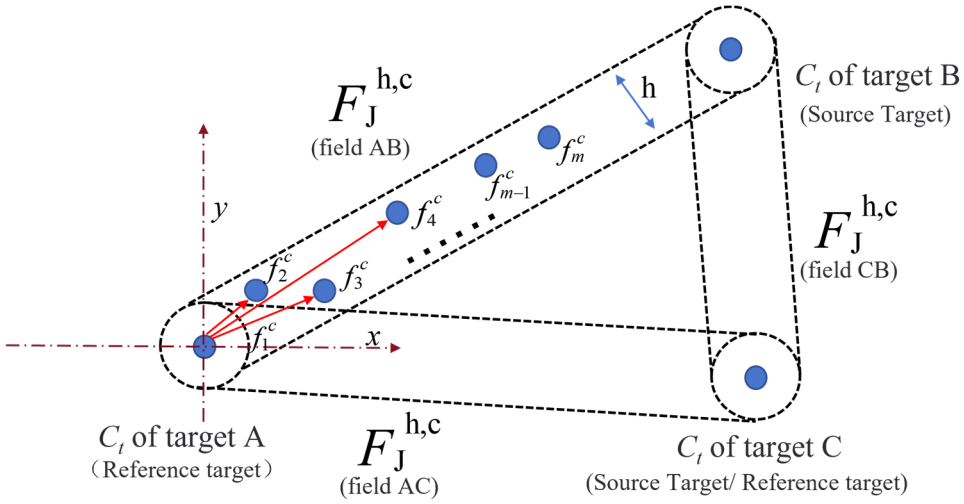
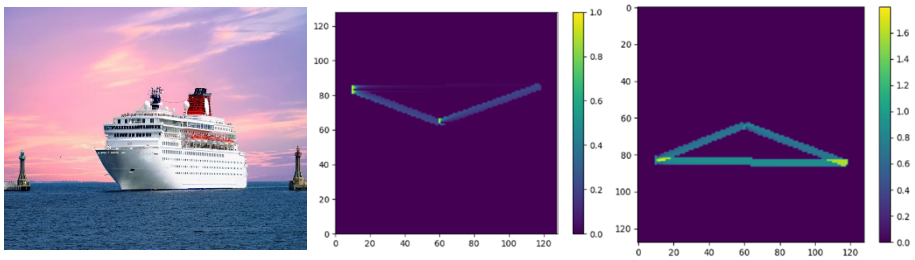


Figure 8. Schematic diagram of SOVF encoding.

Figure 9 illustrates the output F from the S-branch corresponding to the original image shown in Figure 9, which is identical to Figure 7a. In this image, the spatial orientation relations are as follows: between the "lighthouse" on the left and the "lighthouse" on the right, the relation is "L&R"; between the "lighthouse" on the left and the "passenger ship", the relation is also "L&R"; and between the "lighthouse" on the right and the "passenger ship", the relation remains "L&R". Figure 9b,c shows the anticipated ground-truth FB-field and LR-field generated by the S-branch. In total, six spatial orientation vector fields were produced in F, consisting of three FB-fields and three LR-fields. Notably, the pixel scores in the LR-field were significantly higher than those in the FB-field.



(**a**) Original image (**b**) FB-field (**c**) LR-field

Figure 9. The output of the S-branch.

3.4. Fusion Module Design

To obtain the triplet list that contained all the spatial connections within an image, a fusion module was required to synthesize the output Y from the T-branch and F from the S-branch, along with the corresponding confidence levels for each pair.

The fusion module was implemented as follows. Initially, a 3×3 max pooling layer was used on the predicted heatmap Y. The subsequent step involved the selection of center points according to the value of heatmap Y that exceeded a threshold designated as $K = C$ (we set $C = 0.5$), indicating the presence of detected WSTs. The channel containing the center points was utilized to determine the category of each WST, with k representing the confidence associated with that category. Following this, based on the center points of the detected WSTs in Y, pairwise relations between WSTs were established, and FB-fields and LR-fields could also be constructed from the predicted output F.

The pixel scores derived from field can be utilized to predict the spatial orientation relations between the water surface objects. We can calculate the sum of the scores within the spatial orientation vector fields in both FB-fields and LR-fields as follows:

$$\begin{cases} S_{fb} = SUM(F_J^{h,0}) \\ S_{lr} = SUM(F_J^{h,1}) \end{cases} \tag{4}$$

where S_{fb} and S_{lr} represent the votes from all pixels in the FB-fields and LR-fields, respectively. Finally, based on the pixel positions of the center point predicted on Y, the confidence associated with a specific spatial relation can be obtained from the following formula:

$$\{R : conf\} = \begin{cases} \{\text{L\&R} : \frac{S_{lr}}{S_{lr}+S_{fb}}\}, & S_{lr} > S_{fb} \text{ and } \frac{S_{lr}}{S_{fb}} > T \\ \{\text{F\&B} : \frac{S_{fb}}{S_{fb}+S_{lr}}\}, & S_{fb} > S_{lr} \text{ and } \frac{S_{fb}}{S_{lr}} > T \\ \{\text{FL\&RR} : 1 - \left|0.5 - \frac{S_{lr}}{S_{lr}+S_{fb}}\right| - \left|0.5 - \frac{S_{fb}}{S_{lr}+S_{fb}}\right|\}, & S < \frac{S_{fb}}{S_{lr}} < T \text{ and } [\, C_{rty} > C_{sty}] \\ \{\text{LR\&RF} : 1 - \left|0.5 - \frac{S_{lr}}{S_{lr}+S_{fb}}\right| - \left|0.5 - \frac{S_{fb}}{S_{lr}+S_{fb}}\right|\}, & S < \frac{S_{fb}}{S_{lr}} < T \text{ and } [\, C_{rty} < C_{sty}] \end{cases} \tag{5}$$

where R represents the category of spatial orientation relations while *conf* denotes the confidence of the corresponding spatial relation; both are represented using key-value pairs. Additionally, S and T are two thresholds, configured as $S = 0.8$ and $T = 1.2$. C_{rty} and C_{sty} denote the Y-axis coordinates of the centers of the reference object (object that is closer to the vertical axis of the pixel coordinate system) and the target, respectively. It is clear that our predictions for "FL&RR" and "LR&RF" require leveraging the pixel positions of the center point predicted on Y.

Thus, the fusion module integrates the features extracted from both the T-branch and S-branch. Utilizing the planar positions of these central points, the spatial orientation relations and confidence levels between each pair of WSTs can be derived from the feature map F produced by the S-branch. This process yields the final output of the target spatial orientation relations in the form of a triplet set.

3.5. Architecture of Spatial Orientation Relations Recognition Network

As illustrated in Figure 10, the architecture of the WST-SOVF algorithm followed the generic encoder-decoder framework so that an hourglass-shaped deep convolutional neural network (DCNN) consisting of a chain of convolution and up-convolution layers was generated. Any effective deep neural network (DNN) architecture, such as VGG [3], ResNet [14], InceptionNet [4], DenseNet [37], etc. can serve as the backbone of the encoder module for feature extraction from the images. The decoder comprises up-convolutional layers that diverge into two branches: the T-branch and the S-branch, for the extraction of two classes of feature maps. The upsample module consists of three up-convolutional layers that transform the output of the backbone into a higher-resolution feature map. The T-branch is specifically designed for WST recognition, which generates multiple heatmap layers, with the quantity of these layers corresponding to the number of WST categories. Simultaneously, the spatial orientation recognition branch, referred to as the S-branch, generates a dual-channel heatmap, facilitating the prediction of vector fields for each WST pair. The subsequent fusion module was designed to integrate the WST recognition output Y with the spatial orientation recognition output F, enabling the compilation of a comprehensive triplet list of all WST pairs.

The network architecture details of our spatial orientation relations recognition model are illustrated in Figure 11. The first part adopted ResNext50, which had been pre-trained for image classification. For the feature extraction phase, the input color image (512 × 512 size)

was highly compressed into latent features through a series of deeply stacked convolutional blocks. As a result, the spatial dimensions of these features were reduced to 1/32 of the original resolution, while the number of channels remained substantial. The second part was the upsample module, which enhances a standard residual network by incorporating three up-convolutional networks, thereby facilitating a higher-resolution output. This up-convolutional phase ensures that the output size of the heatmap (512 × 512 pixels) is adequate to meet the coding requirements for the WST center point region and the WST-SOVF. The third part of the network consisted of a branching structure that divided into two branches, each containing two layers of convolution. One branch, namely the T-branch, is responsible for generating the heatmap corresponding to the Gaussian center point of the image target category [38], effectively performing a decoding task to reconstruct the global layout of the target center point region. The other branch (i.e., S-branch) predicts the spatial orientation vector fields between WSTs. The fusion module was the last part, which establishes the spatial connections among all water surface objects in the image by synthesizing the outputs from both the T-branch and the S-branch. Thus, the output of the entire network is a triplet list that includes the object pairs and their corresponding spatial orientation relations.

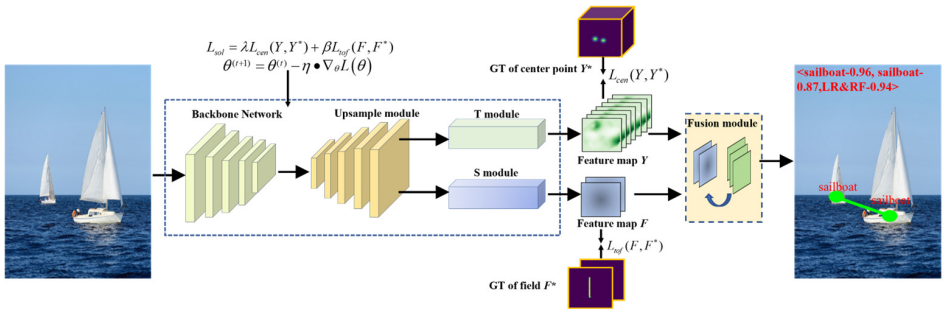


Figure 10. Overview network architecture of the spatial orientation relations recognition model.

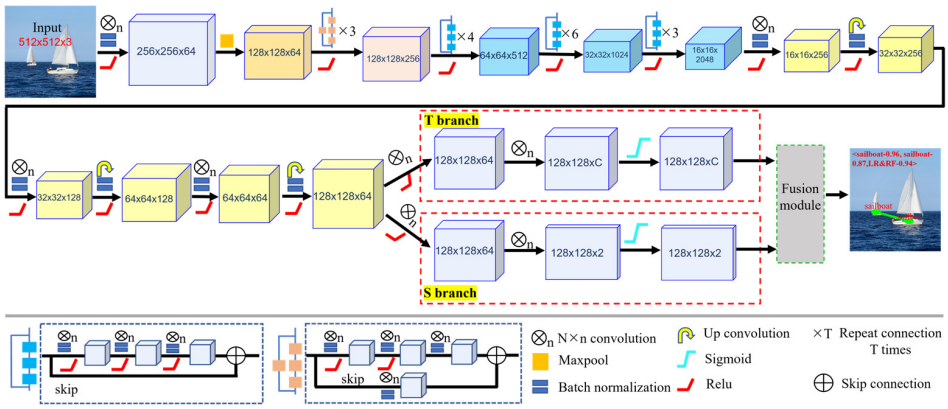


Figure 11. Details of the network architecture of our spatial orientation relation recognition model.

3.6. Loss Function Design

Optimizing the discrepancy between the predicted and ground-truth spatial orientations significantly enhances the model accuracy across diverse objects. In order to optimize the trainable parameters of the proposed network, this study introduced a novel spatial loss function that was included in the overall loss computation and backpropagation process. The total adopted loss function can be defined as:

$$L_{sol} = \lambda L_{cen}(Y, Y^*) + \beta L_{tof}(F, F^*) \quad (6)$$

where L_{cen} (Y, Y^*) and L_{tof} (F, F^*) represent the target center point loss and the spatial orientation vector field loss, respectively. The parameters λ and β are two balancing factors for L_{cen} (Y, Y^*) and L_{tof} (F, F^*), which were determined as 2 and 1, respectively, through extensive experimentation. Furthermore, we computed the following two quantities:

(1) Target center point loss L_{cen} (Y, Y^*): L_{cen} (Y, Y^*) contributes to the target category recognition. L_{cen} (Y, Y^*) has a similar form as the loss function employed in CenterNet [28,30] and CornerNet [29,31], that is

$$L_{cen}(Y, Y^*) = \frac{-1}{N} \sum_{x,y,c} \begin{cases} (1 - Y_{x,y,c}) \log(Y_{x,y,c}) & \text{when } Y^*_{x,y,c} = 1 \\ (Y_{x,y,c})^4 \left(1 - Y^*_{x,y,c}\right)^4 \log(1 - Y_{x,y,c}) & \text{otherwise} \end{cases} \tag{7}$$

where $Y_{x,y,c}$ denotes the confidence level of a detected target within the heatmap Y, (x, y) indicates the coordinates of the pixel in the heatmap Y, c refers to both the channel of the heatmap and the class of the detected target,$Y^*_{x,y,c}$ denotes the ground-truth of the pixel $Y_{x,y,c}$, and N is the number of targets in the heatmap Y^*. This loss function enables the predicted values of all pixels surrounding the target center point to follow a Gaussian distribution. Additionally, the pixel values outside this region diminish toward zero as the distance increases.

(2) Spatial orientation vector field loss L_{tof} (F, F^*): L_{tof} (F, F^*) was specifically designed to accurately model the region of the spatial orientation vector field, taking into account the spatial orientation vectors both within and outside the field [39–41]. For the vectors located inside the field, we expected that each pixel would approximate the ground truth closely. Thus, L_{tof} (F, F^*) was chosen to represent the mean square error relative to the ground-truth. Conversely, for pixels outside this region, a logarithmic function was employed to facilitate a more rapid convergence toward zero.

$$L_{tof}(F, F^*) = \frac{1}{M} \sum_{i=1}^{M} \begin{cases} (F^*_{x,y,c} - F_{x,y,c})^{\alpha} & F^*_{x,y,c} > 0 \\ -\log(1 - F_{x,y,c}) & F^*_{x,y,c} = 0 \end{cases} \tag{8}$$

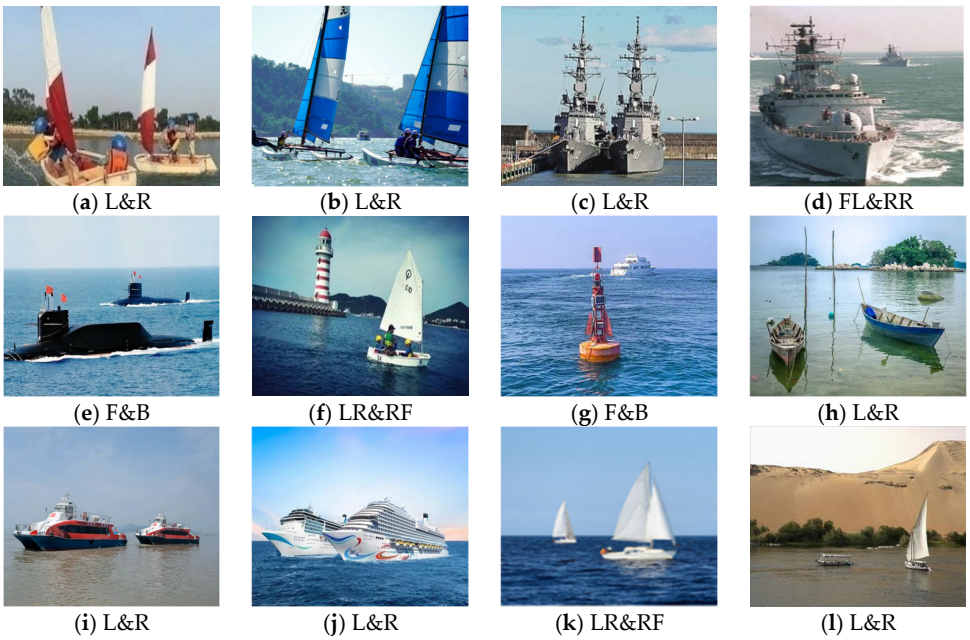
where $F_{x,y,c}$ denotes the value of the pixels in the feature map F, (x, y) indicates the coordinates of the pixel in the feature map F, c refers to both the channel of the feature map and the class of the spatial orientation vector fields, $F^*_{x,y,c}$ denotes the ground-truth of the pixel $F_{x,y,c}$, and M is the number of spatial orientation vector fields in the feature map F. It is obvious that $F^*_{x,y,c}$ inside the fields is non-zero, and the value of the pixels outside the fields are all 0. The hyperparameter α was set to 4, determined through comprehensive experimental evaluations.

4. Experiments and Analysis

4.1. Dataset and Experimental Setup

As shown in the literature, deep neural networks generally benefit from big data training, so a well-structured dataset with rich spatial features is essential for optimizing the parameters of our algorithm. For the training and evaluation of the WST-SOVF algorithm, we selected images from Huawei's "Typical Surface/Underwater Target Recognition" (SUTR) dataset, which was publicly released during a nationwide artificial intelligence competition hosted by Huawei in China in 2022. This publicly available dataset offers a rich set of images including eight distinct WST types. We partitioned this dataset into three subsets for training, testing, and validation, specifically allocating 3780 images for the training set, 378 images for the testing set, and 378 images for the validation set, thereby adhering to a 10:1:1 ratio.

During the testing phase, it was crucial to filter the test set to retain only those images that contained two WSTs, annotating them with their corresponding spatial orientation relation categories. Notably, the SUTR dataset contains complex spatial orientation relations among WSTs, making it a suitable choice for training the WST-SOVF algorithm. However, within the test set, the predominant spatial orientation relations observed among the WSTs were primarily of the "F&B" (front and back) and "L&R" (left and right) types, and most targets in the same image were mostly of the same category. Figure 12 below offers a representative sample of the data contained within the test set.The experimental training and testing were conducted on a laptop, equipped with an intel i5-11400 CPU and an NVIDIA GeForce RTX 3070 laptop GPU (Shenzhen Shenzhou Innovation Technology Co., Ltd. in Shenzhen, China.), featuring 8 GB of video memory. The experimental environment comprised Python 3.8, PyTorch 1.7.0, and CUDA 11.1 for parallel computing, and cuDNN 8.0.4 as the neural network accelerator library.



(**a**) L&R (**b**) L&R (**c**) L&R (**d**) FL&RR

(**e**) F&B (**f**) LR&RF (**g**) F&B (**h**) L&R

(**i**) L&R (**j**) L&R (**k**) LR&RF (**l**) L&R

Figure 12. Demonstration of the test set based on SUTR construction.

4.2. Model Training

Our WST-SOVF algorithm is a fully end-to-end CNN model. The training process of this CNN model was divided into two stages, namely the frozen stage and the unfrozen stage. To optimize resource utilization and reduce training time, we froze the weights of the ResNet50 [14] backbone feature extraction network at the beginning. Then, we trained the unfrozen network structure parameters for 100 epochs. Finally, all network structures were jointly trained for an additional 200 epochs. During the frozen training stage, only the parameters of the latter half of the model were adjusted, resulting in a small video memory occupancy. Consequently, the batch size was set to 8 for this stage. In the unfrozen training stage, the parameters of all models were updated, requiring a larger video memory, which prompted us to halve the batch size during training. The Adam optimizer was employed as the model training optimizer throughout the entire process, with the learning rate being adaptively adjusted according to changes in loss. The maximum learning rate was set at 50^{-4}, and the minimum learning rate was set at 50^{-6}.

During the training process of the WST-SOVF model, images were input into the network to generate predicted outputs for both the T-branch and the S-branch. The ground truth for the T-branch was established based on the coding methodology of the target Gaussian center

point, while the ground truth for the S-branch was constructed in accordance with the coding guidelines applicable to fields. The prediction heatmap of the T-branch was compared with its corresponding ground-truth to construct the loss function L_{cen}, which pertains to the center point region. Similarly, the prediction heatmap of the S-branch along with its ground truth were utilized to establish the loss function L_{tof}. Training all network weights with back-propagation resulted in the classification prediction of the WST category and the spatial orientation between WSTs. The process of encoding the ground-truth for both branches is described in Section 3.

Figure 13 demonstrates the ground-truth generated for several typical instances during training. In Figure 13(a_1–d_1), each represents an original training image, while Figure 13(a_2–d_2) displays the encoded heatmaps corresponding to the T-branch associated with the original training images. These figures only show the heatmap of the category channel where the target is located. The third and fourth rows in Figure 13 show the encoded heatmaps generated by the S-branches during training. Figure 13(a_3–d_3) represents the FB-field, whereas Figure 13(a_4–d_4) illustrates the LR-field.

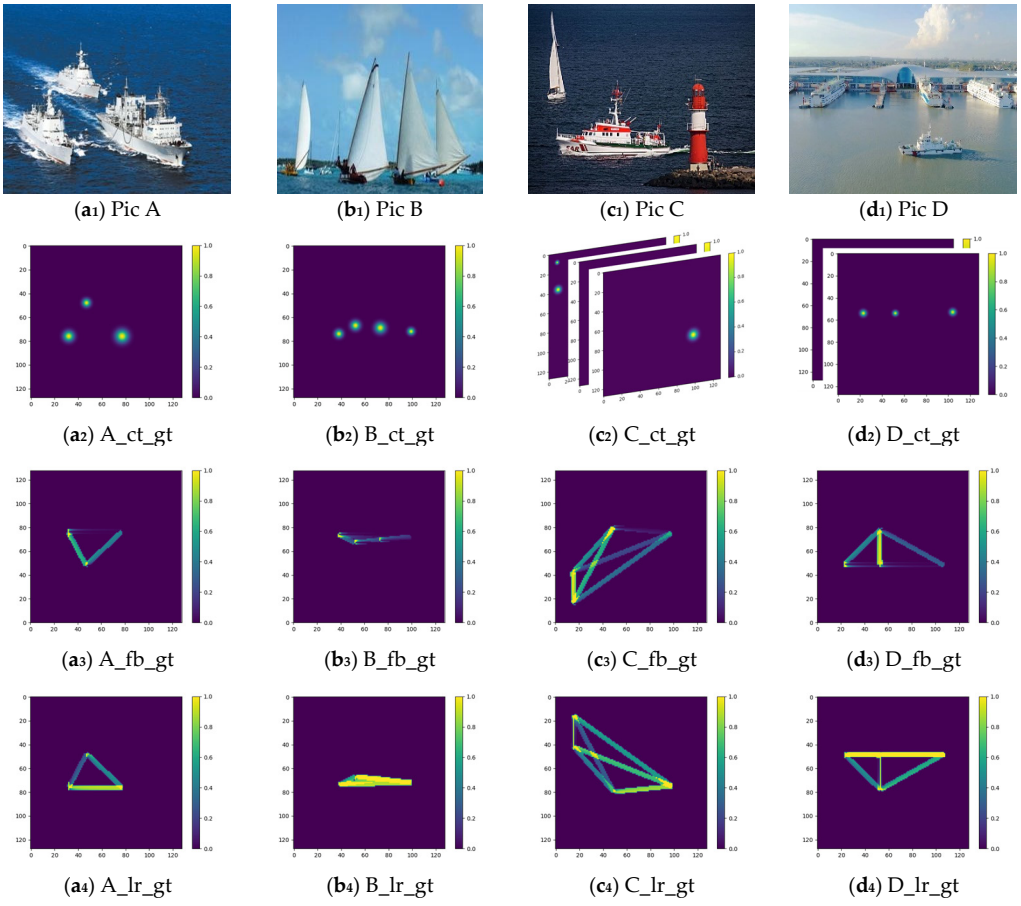


Figure 13. Visualization of the ground-truth constructed by the SUTR dataset.

4.3. Spatial Orientation Recognition Results

Figure 14 presents the recognition results of the trained WST-SOVF algorithm. The images in the first row of Figure 14, specifically Figure 14(a_1–d_1), are the input images. The second row comprises the corresponding results from the T-branch, only focusing on the channels of the heatmap related to the target category. Particularly, Figure 14(c_2) displays the heatmaps for two channels, as the original image contained targets from both categories: "sailboat" and "naval vessels". Furthermore, the images in the third row present the heatmaps associated with the "F&B" relation fields, while their "L&R" relation fields are provided in the heatmaps of the fourth row. Figure 14(a_5–d_5) visualizes the recognition of the spatial orientation relations,

which are labeled on top of the images. The label in the image shows the detected water surface object categories and their spatial orientation relation as well as the corresponding confidence level. Key parameters obtained in the decoding process are summarized in Table 1. It can be observed that spatial orientation relation of water surface objects in the image were consistent with the results presented in the heat map. For the instance presented in Figure 14(a_1), the spatial orientation relation between the two "naval vessels" was the "F&B" class, which was highlighted in the heatmap representing this spatial relation (see Figure 14(a_3)). In contrast, the heatmap appeared considerably sparse in Figure 14(a_4), which represents "L&R" relations.

Additionally, Figure 15 presents further recognition results for the images in the test set. Three instances in Figure 15h noted which images contained more than two classes of water surface objects. For these cases, a triple list was provided to indicate the detected water surface object categories along with their corresponding spatial orientation relations.

These results demonstrate that the proposed WST-SOVF algorithm can accurately recognize both the WST categories and the spatial orientation relations. Moreover, this end-to-end algorithm can simultaneously provide the corresponding confidence levels. In more complicated scenarios where the image contains multiple water surface object categories, the algorithm was able to recognize all the water surface objects along with their respective special orientation relations.

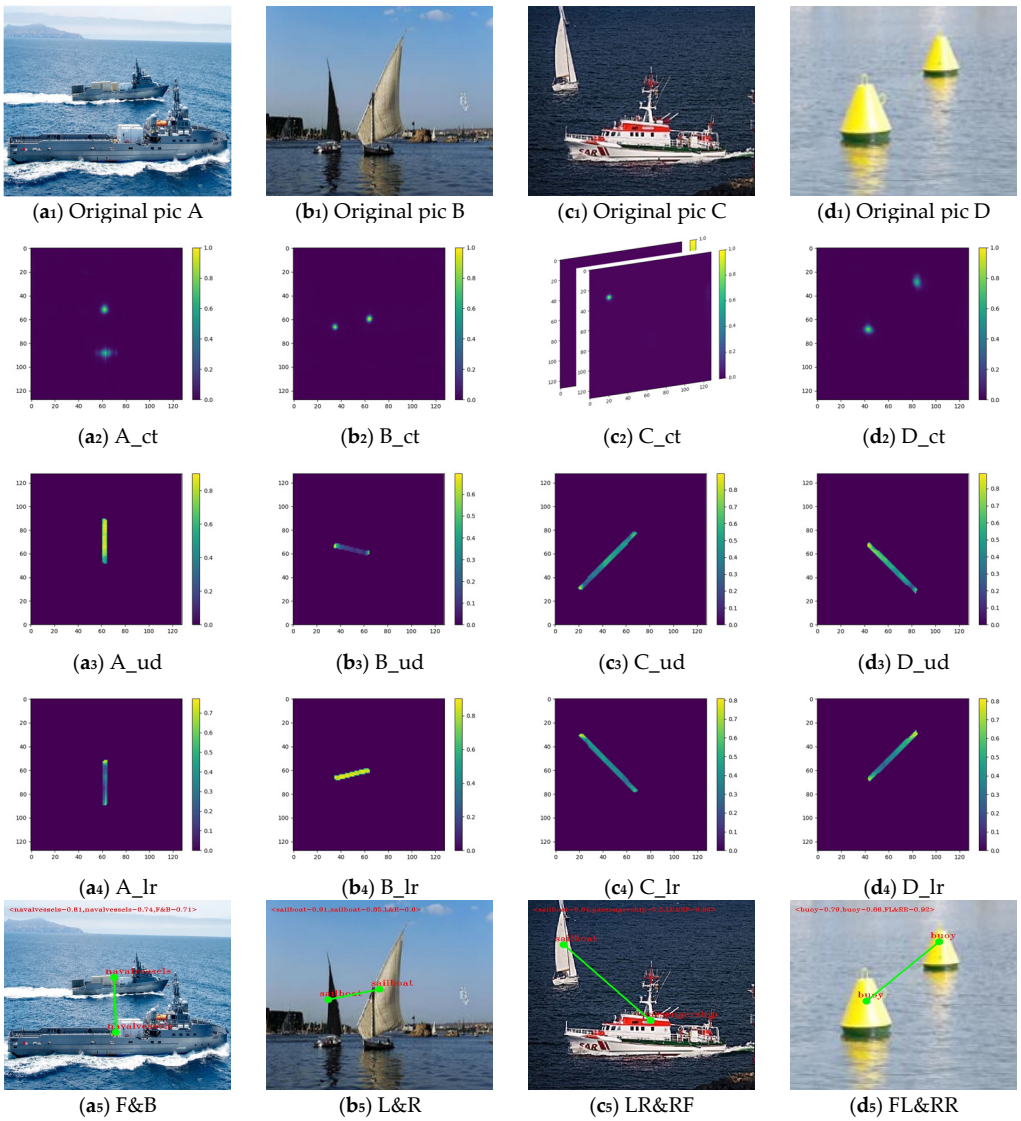


Figure 14. Classification results of the WST-SOVF algorithm for spatial orientation relations.

Table 1. Results of key parameters in forward inference.

Results	Image (a_1)	Image (b_1)	Image (c_1)	Image (d_1)
$\{R : conf\}$	{4, 0.81} {4, 0.74}	{1, 0.91} {1, 0.85}	{1, 0.84} {5, 0.50}	{2, 0.79} {2, 0.66}
S_{fb}	220.9502	16.0086	126.4875	123.9426
S_{lr}	36.6966	98.7706	118.4239	104.7614



(**a**) L&R (**b**) FL&RR (**c**) L&R (**d**) L&R

(**e**) F&B (**f**) LR&RF (**g**) L&R, F&B, FL&RR (**h**) L&R

(**j**) F&B (**j**) F&B (**k**) L&R (**l**) L&R

Figure 15. Recognition results for the spatial orientation relation in the test set.

4.4. Evaluation of the WST-SOVF Algorithm

In this section, we evaluate the performance of the proposed WST-SOVF algorithm using a range of metrics to assess its effectiveness in the context of the SUTR dataset. Specifically, we employed various metrics, including the confusion matrix and its related metrics such as precision, recall, and *F*1, to evaluate the performance of the WST-SOVF algorithm in recognizing four types of spatial orientation ("F&B", "L&R", "FL&RR", "LR&RF") relations [41–43].

The confusion matrix [42–44] was constructed with the model's predicted category quantity statistics on the horizontal axis and the actual label quantity statistics on the vertical axis. The diagonal entries indicate the number of predictions made by the model that aligned with the data labels. A higher value on the diagonal signifies superior prediction results in this category. Figure 16 illustrates the confusion matrix of the WST-SOVF algorithm on the test set.

Utilizing the confusion matrix, a variety of evaluation metrics can be derived, including the precision (which measures the accuracy of the algorithm's predictions), recall (which assesses the algorithm's capacity to identify all pertinent positive instances), and *F*1 (which is the harmonic mean of precision and recall, providing a unified metric that balances both the accuracy of positive predictions and the algorithm's effectiveness in identifying all relevant instances) [43–45].

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

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

$$F1 = \frac{2TP}{TP + FP + TP + FN} \tag{11}$$

where TP denotes the proportion of positive samples that were correctly predicted, FP denotes the proportion of negative samples that were predicted incorrectly, TN represents the proportion of negative samples that were correctly predicted, and FN represents the proportion of positive samples that were predicted incorrectly. Table 2 presents the evaluation metrics for each category. The average precision of the WST-SOVF algorithm on the test set was 97.5%, while the average recall was 92.0%. The algorithm demonstrated considerable detection performance for both the "L&R" and "F&B" spatial orientation relations. Nevertheless, there remains potential for enhancement in the detection outcomes for other categories.

Additionally, it should be noted that the precision and $F1$ scores across all four spatial orientation relation categories were quite impressive, but the recall for the "FL&RR" and "LR&RF" categories were relatively lower. We attributed this to two main factors. First, in the test set, "FL&RR" and "LR&RF" had fewer test samples compared with the other categories, and categories with smaller sample sizes typically exhibit lower recall. Second, both "FL&RR" and "LR&RF" share a common characteristic of "inclination", which made it difficult for the model to distinguish between them. Specifically, analysis of the confusion matrix in Figure 16 revealed that five samples from "FL&RR" were misclassified as "L&R", and four samples from "LR&RF" were similarly misclassified as "L&R", indicating feature overlap with the "L&R" category. This overlap likely stems from similarities in spatial orientation relation representation, particularly when samples exhibit transitional spatial arrangements that the model tends to classify into the more abundant "L&R" category.

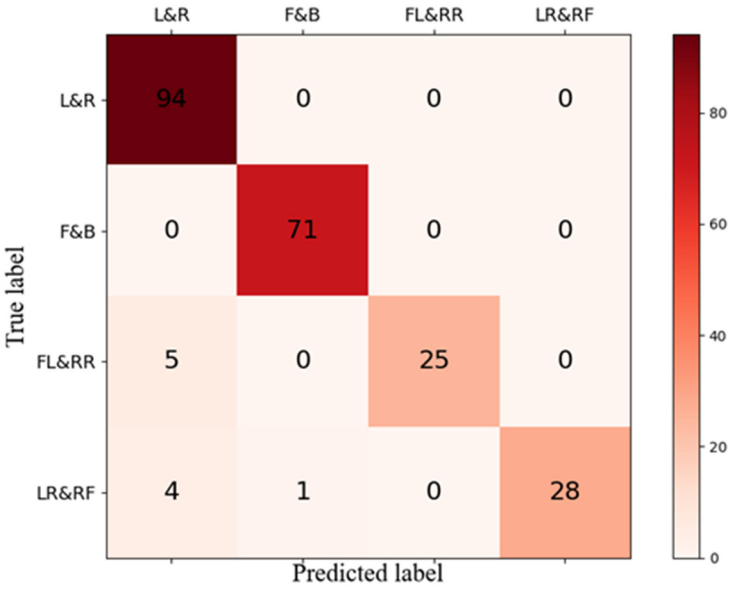


Figure 16. Confusion matrix for the DCNN model of the WST-SOVF algorithm evaluation.

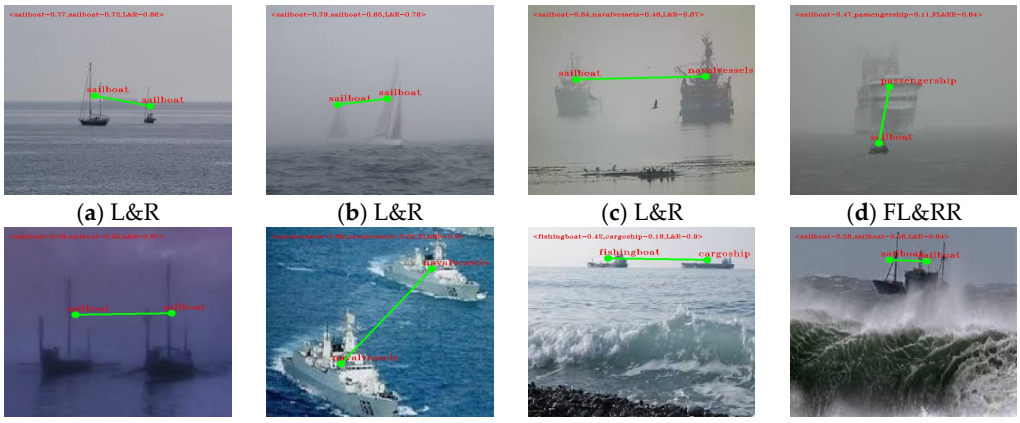
Table 2. Results of evaluation metrics for each category.

Category	Precision	Recall	*F1*
L&R	0.913	1.000	0.954
F&B	0.986	1.000	0.993
FL&RR	1.000	0.833	0.909
LR&RF	1.000	0.848	0.918
Average	0.975	0.920	0.944

4.5. Additional Experiments and Analysis

To validate the robustness of our WST-SOVF algorithm, we conducted qualitative experiments across scenarios of varying complexity including complex marine environments and occlusion cases as well as generalization experiments on other broader datasets.

For robustness experiments in complex marine environments, we selected two common and representative elements: waves and fog, to evaluate our WST-SOVF algorithm. We collected a number of maritime images under varying fog visibility levels (according to international visibility classification, Level 1–8) and different sea states (Beaufort scale 0–9) to conduct spatial orientation relation recognition. Figure 17 presents some representative results. We found that the WST-SOVF algorithm performed effectively under low-level fog conditions (Level 1–3) and calm to moderate sea states (Beaufort scale 0–4), as shown in Figure 17a,f, successfully completing spatial orientation relation recognition. However, under dense fog conditions (Level 6–8) or rough sea states (Beaufort scale 7–9), the WST-SOVF algorithm encountered certain challenges, as demonstrated in Figure 17c–e,g,h, which primarily manifested in lower confidence for target recognition. The performance degradation could mainly be attributed to the reduced image contrast in dense fog and increased noise from wave motion in rough seas. When these challenging conditions occurred simultaneously, the WST-SOVF algorithm's effectiveness was further compromised. Overall, these results indicate that while WST-SOVF shows promise in moderate conditions, its performance in complex marine environments remains modest and requires further improvement.



(**a**) L&R (**b**) L&R (**c**) L&R (**d**) FL&RR

(**e**) L&R (**f**) FL&RR (**g**) L&R (**h**) L&R

Figure 17. Recognition results in complex marine environments.

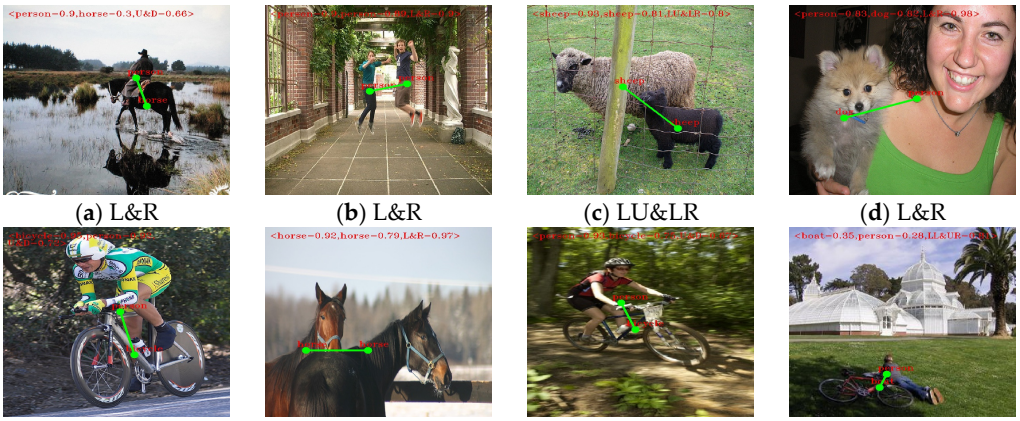
To evaluate the robustness of our WST-SOVF algorithm under occlusion scenarios, we collected a set of water surface target images containing occlusion cases for experimental validation, with some representative qualitative results presented in Figure 18. The experimental results demonstrated that our WST-SOVF algorithm maintained robust performance when the occlusion ratio was below 0.5 (i.e., when the occluding foreground object covers less than 50% of the background object's area). This robust performance under moderate occlusion can be attributed to our architectural design choices. Specifically, our recognition branch incorporated state-of-the-art anchor-free detection principles, which have proven to be more resilient to partial occlusions. The anchor-free approach allows the algorithm to effectively capture and utilize visible object features even when parts of the target object are obscured, enabling reliable spatial orientation relation recognition under moderate occlusion conditions. However, while the current implementation shows promising results, there is room for improvement in handling more severe occlusion scenarios.



(**a**) L&R (**b**) L&R (**c**) F&B (**d**) L&R

Figure 18. Recognition results in occlusion scenarios.

To validate the generalization capability of our WST-SOVF algorithm, we conducted additional experiments on the Pascal Visual Object Classes (PASCAL VOC) dataset [46]. The Pascal VOC dataset serves as a well-established benchmark in computer vision, containing diverse object categories across various real-world scenarios. This dataset includes over 11,000 images with 20 different object categories and rich spatial relations between objects, making it an ideal testbed for spatial relation orientation recognition algorithms. By training and testing our WST-SOVF algorithm on this dataset, we aimed to evaluate its ability to generalize beyond water surface environments to more diverse terrestrial scenarios. Specifically, we evaluated our WST-SOVF algorithm on the PASCAL VOC 2012 dataset without modifying any architectural settings or hyperparameters. The WST-SOVF algorithm demonstrated particularly strong performance in the "L&R" and "F&B" spatial orientation relation categories, achieving precision scores of 0.959 and 0.950, and recall scores of 0.934 and 0.826, respectively. Figure 19 visualizes several representative spatial orientation relation recognition results on the PASCAL VOC 2012 test set. These results are notably impressive considering that the algorithm was originally designed for water surface targets, but achieved such competitive performance on general object categories in the PASCAL VOC dataset. These experimental results on broader datasets demonstrate the strong generalization capability of our WST-SOVF algorithm beyond its original design domain.



(**a**) L&R (**b**) L&R (**c**) LU&LR (**d**) L&R

(**e**) U&D (**f**) L&R (**g**) U&D (**h**) LL&UR

Figure 19. Recognition results in other broader datasets.

It is worth noting that due to the inherent differences between water surface and non-water surface scene characteristics, the spatial orientation relation categories we used in PASCAL VOC images ("L&R", "U&D", "LL&UR", and "LU&LR") corresponded to the four spatial orientation relation categories introduced in this paper. Specifically, "U&D" (up and down) corresponded to "F&B", "LL&UR" corresponded to "FL&RR", and "LU&LR" (left upper and lower right) corresponded to "LR&RF". These mappings reflect the fundamental differences in spatial characteristics between water surface and general scene images.

5. Discussion

Our WST-SOVF algorithm demonstrated excellent performance in water surface target spatial orientation relation recognition, achieving an impressive average precision of 97.5% and average recall of 92.0% across all categories. The algorithm exhibited particularly strong detection capabilities for the "L&R" and "F&B" spatial orientation relations, with high precision and $F1$ scores across all four spatial relation categories. Beyond the primary dataset, our approach showed remarkable generalization capabilities, performing effectively on the Pascal VOC dataset with minimal adaptations. This cross-domain effectiveness validates the robustness of our architectural design, particularly its anchor-free detection principles, which maintained reliable performance even under moderate occlusion conditions (below 50%).

Despite these achievements, our algorithm exhibits several meaningful challenges. Primarily, the WST-SOVF algorithm's performance in complex marine environments, particularly those characterized by dense fog and rough seas, still has scope for improvement. Our qualitative experiments revealed that while the algorithm performed effectively under mild conditions (fog levels 1–3, sea states 0–4), its performance tended to decrease in challenging conditions with dense fog (levels 6–8) or rough seas (Beaufort scale 7–9). This performance degradation manifests as lower confidence scores and increased misclassifications, limiting the algorithm's practical utility in real-world maritime surveillance applications where adverse weather conditions are common. The suboptimal performance in complex water surface scenarios can be attributed to several factors. Dense fog substantially reduces image contrast and obscures critical visual features, compromising the algorithm's ability to accurately detect and classify objects. Similarly, rough sea conditions introduce significant motion-induced noise and create unstable backgrounds with complex wave patterns, disrupting the feature representation learning process. Moreover, complex marine states significantly affect environmental illumination, imaging quality, and object morphology. Most notably, these challenging conditions can cause considerable variations in color characteristics, which serves as a crucial feature for both maritime object detection and spatial relation recognition. Our future work will focus on improving the WST-SOVF's robustness in complex water surface environments through enhanced image preprocessing, multi-modal sensing integration, and expanded training datasets. Furthermore, inspired by traditional water surface target detection methods that leverage color features effectively, we plan to explore the integration of color-based semantic information into our deep learning framework. Particularly, we aim to utilize the natural color contrast between targets and the blue water background, which could provide valuable implicit cues for spatial orientation relation recognition.

Subsequently, while the WST-SOVF algorithm demonstrated satisfactory performance in handling moderate occlusion situations, another technical challenge emerged when dealing with severe occlusion (occlusion rate exceeding 50%) and overlapping scenarios. Occlusion handling remains a persistent challenge in computer vision, as it represents a complex spatial relationship among objects. However, we have initiated research to address this limitation by extending the WST-SOVF algorithm. Our approach incorporates clustering-based concepts to group occluded foreground objects, occluded background objects, and the occlusion context together, enabling the recognition of such spatial relations as a unified set.

Additionally, the computational efficiency of the WST-SOVF algorithm merits careful consideration. Our empirical evaluation demonstrated that the algorithm achieved a processing throughput of 15–20 frames per second when deployed on an NVIDIA RTX 3070 GPU (Shenzhen Shenzhou Innovation Technology Co., Ltd. in Shenzhen, China.), with a model complexity of 124 M parameters. While this computational performance suffices for con-

ventional water surface target surveillance applications, the processing latency may present challenges in scenarios requiring a stringent real-time response. The model's parametric complexity reflects an intentional design choice that prioritizes detection reliability in challenging marine environments over computational efficiency. Future optimizations could potentially explore model compression techniques and hardware acceleration frameworks to enhance the WST-SOVF algorithm's computational performance.

6. Conclusions

A new novel end-to-end algorithm, namely the WST-SOVF algorithm, was proposed to recognize spatial relations between WSTs in images, which is a fundamental spatial semantic characteristic of images. The WST-SOVF algorithm was designed as an hourglass-shaped deep convolutional neural network model with two separated branches, which established a WST spatial orientation vector field to represent the orientation angles between the WSTs and integrated keypoint recognition to encode their spatial orientation relations. We trained the model on the Huawei's "Typical Surface/Underwater Target Recognition" dataset, where it achieved superior performance in detecting four fundamental types of spatial orientation relations. It is important to note that this study primarily concentrates on fundamental spatial orientation relations between two separate WSTs. Actually, more complex spatial patterns, potentially involving higher-dimensional spatial relations—such as overlapping, intersection, surrounding and occlusion—constitute the spatial semantic characteristics of objects within an image. The recognition of these spatial relations enhances the accuracy of WST perception and will assist navigators or navigation systems to cope with complex navigation situations, and improving these spatial relations remains a key direction for our future research.

Author Contributions: Methodology, P.G.; Validation, K.Z.; Writing—original draft, P.G.; Writing—review & editing, K.Z.; Supervision, K.Z., Y.J., H.Z., X.L., Z.F. and W.H.; Funding acquisition, K.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the National Natural Science Foundation of China (No. 52071047); partially supported by the National Key Research, and Development Program of China (No. 2021YFB3901501); partially supported by the Distinguished Young Scholar Project of Dalian City (No. 2024RJ012); partially supported by the Fundamental Research Funds for the Central Universities (No. 3132023512); and partially supported by the Dalian City Science and Technology Plan (Key) Project (No. 2024JB11PT007).'

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Dataset available on request from the authors.

Acknowledgments: The authors gratefully acknowledge the reviewers for their precious time and effort dedicated to this manuscript.

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

References

1. Canny, J. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1986**, *PAMI-8*, 679–698. [CrossRef]
2. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
3. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
4. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.

5. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
6. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
7. Redmon, J.; Farhadi, A. YOLOv3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
8. Li, X.; Xu, L.; Wei, M.; Liu, Y.; Wang, Z. An underwater crack detection method based on improved YOLOv8. *Ocean Eng.* **2024**, *313*, 114051. [CrossRef]
9. Wang, S.; Li, Y.; Qiao, S. ALF-YOLO: Enhanced YOLOv8 based on multiscale attention feature fusion for ship detection. *Ocean Eng.* **2024**, *308*, 113538. [CrossRef]
10. Pan, H.; Li, G.; Feng, H.; Zhang, Z.; Chen, J. Surface defect detection of ceramic disc based on improved YOLOv5s. *Heliyon* **2024**, *10*, e22626. [CrossRef]
11. Zhao, L.; Liu, J.; Ren, Y.; Wang, G.; Li, Z. YOLOv8-QR: An improved YOLOv8 model via attention mechanism for object detection of QR code defects. *Comput. Electr. Eng.* **2024**, *118*, 108595. [CrossRef]
12. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 640–651.
13. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference (MICCAI), Munich, Germany, 5–9 October 2015; pp. 234–241.
14. 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
15. Tan, M.; Le, Q.V. EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
16. Kong, M.C.; Roh, M.I.; Kim, K.S.; Lee, J.H.; Lee, B.Y. Object detection method for ship safety plans using deep learning. *Ocean Eng.* **2022**, *246*, 110587. [CrossRef]
17. Li, Z.; Ren, H.; Yang, X.; Wang, D.; Sun, J. LWS-YOLOv7: A lightweight water-surface object-detection model. *J. Mar. Sci. Eng.* **2024**, *12*, 861. [CrossRef]
18. Chen, Y.; Zhao, F.; Ling, Y.; Zhang, S. YOLO-based 3D perception for UVMS grasping. *J. Mar. Sci. Eng.* **2024**, *12*, 1110. [CrossRef]
19. Guo, Y.; Shen, Q.; Ai, D.; Liu, Z.; Wang, H. Sea-IoUTracker: A more stable and reliable maritime target tracking scheme for unmanned vessel platforms. *Ocean. Eng.* **2024**, *299*, 113902. [CrossRef]
20. Sabour, S.; Frosst, N.; Hinton, G.E. Dynamic routing between capsules. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3856–3866.
21. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.
22. Liu, Y.; Cheng, X.; Han, K.; Liu, Z.; Feng, B. Investigation into the prediction of ship heave motion in complex sea conditions utilizing hybrid neural networks. *J. Mar. Sci. Eng.* **2024**, *13*, 1. [CrossRef]
23. Nian, Z.; Yang, W.; Chen, H. AEFFNet: Attention enhanced feature fusion network for small object detection in UAV imagery. *IEEE Access* **2025**, *11*, 26494–26505. [CrossRef]
24. Wang, B.; Huang, X.; Cao, G.; Yang, L.; Wei, X.; Tao, Z. Hybrid-attention and frame difference enhanced network for micro-video venue recognition. *J. Intell. Fuzzy Syst.* **2022**, *43*, 3337–3353. [CrossRef]
25. Nuanmeesri, S. Spectrum-based hybrid deep learning for intact prediction of postharvest avocado ripeness. *IT Prof.* **2025**, *26*, 55–61. [CrossRef]
26. Gall, M.; Gardill, M.; Fuchs, J.; Horn, T. Learning representations for neural networks applied to spectrum-based direction-of-arrival estimation for automotive radar. In Proceedings of the 2020 IEEE/MTT-S International Microwave Symposium (IMS), Los Angeles, CA, USA, 4 August–30 September 2020; pp. 1031–1034.
27. Wu, D.; Su, B.; Hao, L.; Liu, Y.; Zhang, X. A feature detection network based on self-attention mechanism for underwater image processing. *Ocean. Eng.* **2024**, *311*, 114025. [CrossRef]
28. Law, H.; Deng, J. CornerNet: Detecting objects as paired keypoints. *Int. J. Comput. Vis.* **2020**, *128*, 642–656. [CrossRef]
29. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
30. Zhou, S.; Wang, L.; Chen, Z.; Zheng, H.; Lin, Z.; He, L. An improved YOLOv9s algorithm for underwater object detection. *J. Mar. Sci. Eng.* **2025**, *13*, 230. [CrossRef]
31. Zou, C.; Yu, S.; Yu, Y.; Gu, H.; Xu, X. Side-scan sonar small objects detection based on improved YOLOv11. *J. Mar. Sci. Eng.* **2025**, *13*, 162. [CrossRef]
32. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.E.; Sheikh, Y. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *43*, 172–186. [CrossRef]
33. Peyre, J.; Laptev, I.; Schmid, C.; Sivic, J. Weakly-supervised learning of visual relations. In Proceedings of the Ieee International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5179–5188.

34. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
35. Howard, A.; Sandler, M.; Chen, B.; Wang, W.; Chen, L.C.; Tan, M.; Chu, G.; Vasudevan, V.; Zhu, Y.; Pang, R.; et al. Searching for MobileNetV3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 1314–1324.
36. Zhou, X.; Zhuo, J.; Krähenbühl, P. Bottom-up object detection by grouping extreme and center points. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 6568–6577.
37. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
38. Wang, Y.; Peng, H.; Xiong, Y.; Zhu, L.; Wu, Z. Spatial relationship recognition via heterogeneous representation: A review. *Neurocomputing* **2023**, *533*, 116–140. [CrossRef]
39. Bazarevsky, V.; Grishchenko, I.; Raveendran, K.; Zhu, T.; Zhang, F.; Grundmann, M. BlazePose: On-device real-time body pose tracking. *arXiv* **2020**, arXiv:2006.02099.
40. Huang, J.; Zhu, Z.; Guo, F.; Huang, G. The devil is in the details: Delving into unbiased data processing for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 5700–5709.
41. Fang, H.S.; Xie, S.; Tai, Y.W.; Lu, C. RMPE: Regional multi-person pose estimation. In Proceedings of the Ieee International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2334–2343.
42. Hand, D.J.; Till, R.J. Efficient multiclass ROC approximation by decomposition via confusion matrix perturbation analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 1122–1137.
43. Wang, H.; Wang, X.; Dou, A. Study on the precision evaluation method for a specific category in the classification of remote sensing image. In Proceedings of the 2012 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Munich, Germany, 22–27 July 2012; pp. 5008–5011.
44. Rouabeh, H.; Abdelmoula, C.; Masmoudi, M. Performance evaluation of decision tree and neural network techniques for road scene image classification task. In Proceedings of the International Image Processing, Applications and Systems Conference (IPAS), Hammamet, Tunisia, 5–7 November 2014; pp. 1–5.
45. Chen, J.; Liu, H. Laboratory water surface elevation estimation using image-based convolutional neural networks. *Ocean. Eng.* **2022**, *248*, 111180. [CrossRef]
46. Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1627–1645. [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.