Article

Not peer-reviewed version

A Novel Target Detection Method with Dual-Domain Multi-Frequency Feature in Side-Scan Sonar Images

Wen Wang , Yifan Zhang , Houpu Li * , Xue Gong , Lei Liu , Yixin Kang

Posted Date: 2 November 2023

doi: 10.20944/preprints202309.2024.v2

Keywords: target detection; side-scan sonar images; seabed object; frequency domain

Article

A Novel Target Detection Method with Dual-Domain Multi-Frequency Feature in Side-Scan Sonar Images

Wen Wang [1], Yifan Zhang [2], Houpu Li [1,*], Xue Gong [3], Lei Liu [1] and Yixin Kang [1]

[1] School of Electrical Engineering, Naval University of Engineering, Wuhan, Hubei, 430000, China
[2] School of Geoscience and Info-physics, Central South University, Changsha, Hunan, 410023, China
[3] School of Electronic Engineering, Naval University of Engineering, Wuhan, Hubei, 430000, China
[*] Correspondence: 215011021@csu.edu.cn

Abstract: Side-scan sonar (SSS) detection is a key method in applications such as underwater environmental security and subsea resource development. The use of acoustic images for seabed target detection has gradually become a mainstream underwater detection method. However, many existing detection approaches primarily concentrate on tracking the evolution path of optical image object detection tasks, resulting in complex structures and limited versatility. To tackle this issue, we introduce a pioneering Dual-Domain Multi-Frequency Network (D^{2}MFNet) meticulously crafted to harness the distinct characteristics of SSS image detection. In D^{2}MFNet, aiming at the underwater detection requirements of small scenes, we introduce a novel method for optimize and improve the detection sensitivity of different frequency ranges and propose a Multi-Frequency Combined Attention Mechanism (MFCAM). This mechanism amplifies the relevance of dual-domain features across different channels and space. Moreover, recognizing that SSS images can provide richer insights after frequency domain conversion, we introduce a Dual-Domain Feature Pyramid Network (D^{2}FPN). By incorporating frequency domain information representation, D^{2}FPN significantly augments the depth and breadth of feature information in underwater small datasets. Our methods are seamlessly designed for integration into existing networks, offering plug-and-play functionality with substantial performance enhancements. We have conducted extensive experiments to validate the efficacy of our proposed techniques, and the results showcase their state-of-the-art performance. MFCAM improves the mAP by 16.9% in the KLSG dataset and 15.5% in the SCTD dataset. The mAP of D^{2}FPN was improved by 8.4% in the KLSG dataset and by 9.8% in the SCTD dataset. We will make our code and models publicly available at https://dagshub.com/estrellaww00/D2MFNet.

Keywords: target detection; side-scan sonar images; seabed object; frequency domain

1. Introduction

The superior underwater transmission and convenient characteristics of side-scan sonar (SSS) make it the most effective and commonly used technical means for conventional underwater exploration tasks [1–6], such as ensuring underwater safety, channel obstacle scanning, archaeology, etc. With the further development of technologies, using SSS images to detect underwater targets has become an essential way of detecting the underwater environment [7–9].

In SSS image detection field, methods [10–12] are mainly based on one-stage or two-stage detection methods, such as Faster R-CNN [10], Cascade RCNN [13], SSD [14], Yolo series [11,15–18], et al. Considering the characteristics of small targets, less target amount, and the seawater noise, many methods follow the improved strategy of general target detection. The backbone and neck, the basic part of existing detection methods [19], are improved based on self-structured and existing advanced modules [8,20,21]. EMRN [22] proposes a multi-resolution features dimension uniform module to fix dimensional features from images of varying resolutions. In addition, methods different from the above routines are used in target detection algorithms, such as low-rank sparse matrix factorization [23], vision transformer [24], pulse-coupled neural network [25], et al. What distinguishes the SSS image detection field from other fields is that the amount of data is small, and the target in the image is accompanied by noise and shadows. In general, there is less feature information overall. At the same time, the above methods do not fully mine the feature information of the target, resulting in the methods being generally atopic and not universal. Moreover, these methods do not take into account the phenomenon that the image has additional rich feature information in the frequency domain.

In order to introduce additional information expression of the image in the frequency domain, in the traditional image processing field, scholars [26–29] consider feature extraction of target images from the time and frequency domains. Wang et al. [26] noted that CNN has the ability to capture high-frequency components that cannot be perceived by humans. HPGN [30] proposes a novel pyramid graph network targeting features, which is closely connected behind the backbone network to explore multi-scale spatial structural features. In the field of general-purpose target detection, Xu et al. [31], Liu et al. [32] and Qin et al. [33] had tried to combine the frequency method with the detection method. HSGM [34,35] proposes a hierarchical similarity graph module to relieve the conflict of backbone networks and mine the discriminative features. In contrast, the existing research methods in the field of sonar image target detection all perform feature extraction from a single domain in the time domain, without considering the additional feature representation of sonar images in the frequency domain, relative to the time domain.

This phenomenon is partly due to the fact that after the domain transformation of the image, the entire image as an array no longer has a spatial association. As a result, the method of domain conversion to obtain more feature information cannot be directly applied to various methods of target detection. Some methods have borrowed the concept of frequency to make some attempts: Zhu et al. [36] to separate the acoustically highlighted area from the surrounding environment based on frequency analysis. Wang et al. [37] had constructed a novel network by enhancing and fusing the different frequency characteristics of SSS images. However, by its very nature, frequency domain conversion is not used to obtain characteristic information.

In addition to the problem of methodology, data as the root cause focus of target detection tasks deserves attention. The release of public datasets [38] has brought qualitative improvements to the general target detection field, such as PASCAL [39], COCO [40] and ImageNet [41]. However, there is a lack of relevant public datasets to study in the field of SSS image target detection. The existing ones usually are much smaller in size, less numerous, and lack benchmarks, such as SCTD [42] and SeabedObjects-KLSG [43].

To solve the above problems, in this paper, we propose the dual-domain multi-frequency network (D^{2}MFNet). The D^{2}MFNet consists of MFCAM and D^{2}FPN. Considering that sonar images have more feature expressions in the frequency domain that are different from the time domain features, the MFCAM and D^{2}FPN module is constructed to form the D^{2}FENet that combines the feature expressions in the time and frequency domains. The main work is as follows:

1. **Dual-domain Multi-frequency Network**:

   Our pioneering approach incorporates frequency analysis into underwater sonar image detection, revitalizing the task with limited data through feature fusion and frequency-based attention mechanisms in D^{2}MFNet.
2. **Multi-frequency Combined Attention Mechanism**:

   We collect two public SSS datasets: part of SeabedObjects-KLSG for the classification task and SCTD for the detection task. The KLSG dataset is reorganized to adapt detection tasks with labeling in VOC format. Then the lack of standard benchmarks in both datasets is made up, benchmark results is provided that other scholars can refer to.
3. **Dual-Domain Feature Pyramid Method**:

   Innovating target detection, our D^{2}FPN method transcends the limitations of traditional frequency domain conversion by introducing feature fusion in the frequency domain, allowing for high-frequency information filtering and diverse unconventional feature map conversions to achieve unique and differentiated feature extraction

4. **Benchmark Dataset**:

   We curate and standardize two public SSS datasets – a portion of SeabedObjects-KLSG for classification and SCTD for detection. By adapting KLSG for detection tasks and providing benchmark results, we address the need for standardized benchmarks, offering valuable references for fellow researchers

The remainder of this paper is given as follows. Section 2 gives a brief introduction to the related works. Section 3 introduces the proposed SSS detection methods and benchmark dataset. Section 4 is about experimental results and analysis. Finally, Section 5 concludes the paper.

2. Related Work

2.1. Target Detection for SSS Images

As an acoustic image, the detection difficulty of SSS images lies in the characteristics of large noise, weak-and-small targets, and the small amount of data, which usually results in low precision of the detection method [44–46]. To solve these problems, the existing side-scan sonar target detection methods are usually based on the one-stage or two-stage detection methods, and the algorithm is improved for private datasets. The multi-branch shuttle neural network (MBSNN) [47], plays a role in AUV navigation tasks, improves the backbone and neck of the state-of-the-art Yolo5 with multi-branch shuttle network and BiFPN [20]. The multilevel feature fusion network (MLFFNet) [8] uses several improved modules to solve the problems in sonar images such as seafloor reverberation noise interference, low pixel ratio in foreground objects, and poor imaging resolution. TransYOLO [21] presents an anchor-free method based on a transformer feature fusion network and ellipse quality evaluation.

Different from the above target detection methods, some scholars introduce some unconventional methods for SSS image target detection. For example, Sun et al. [24] had proposed a dual path vision transformer network (DP-VIT) to accurately detect targets in forward-look sonar and side-scan sonar. Zhou et al. [25] had proposed an automatic underwater detection method based on the pulse-coupled neural network (PCNN) by using forward-looking sonar images. GiT [12] proposes a structure where graphs and transformers interact constantly, enabling close collaboration between global and local features. Li et al. [48] had proposed a transfer learning method for sonar image classification and target detection called the texture feature removal network, to deal with the problem of few targets in sonar images. Cheng et al. [49] had proposed a multi-domain collaborative transfer learning (MDCTL) method with multi-scale repeated attention mechanism (MSRAM) for improving the accuracy of underwater sonar image classification.

2.2. Application of Frequency Domain for SSS Images

In the field of traditional image processing, it is common to consider the characteristics of images in both the time domain and frequency domain, such as image filtering and downsampling [50]. The usual frequency domain conversion methods are DCT [51], DFT [52], DWT [53], and FFT [54] derived from DFT.

Converting an image from the time domain to the frequency domain has two benefits: simplification of calculations and processing of information. When calculating, there are two broad categories of simplifications: convolution operations into multiplication operations, and linear differential equations into linear algebraic equations. In information processing tasks, time-frequency conversion can measure images at high and low frequencies. Among them, the low-frequency component is mainly a comprehensive measure of the intensity of the entire image, and the high-frequency component is mainly a measure of the edge and outline of the image. Meanwhile, Wang et al. [26] had discussed the generalization ability of CNN from the frequency distribution of data. It is noted that CNN has the ability to capture high-frequency components that cannot be perceived by humans. And this phenomenon can be used to explain a variety of assumptions that cannot be understood by humans, such as generalization ability and adversarial sample robustness.

Therefore, in the field of general target detection, some scholars [31,33,55] have tried to verify the use of frequency domain images in target detection methods and improve the detection methods. Xu et al. [31] had converted images to the frequency domain, and then grouped all components of the same frequency into a channel. While multiple channels are generated, only the most important channels are kept and inference. Liu et al. [32] had proposed a novel multi-level wavelet CNN (MWCNN) model, in which wavelet transforms are embedded in the CNN architecture to reduce the resolution of the feature map while increasing the receptive field. This method can be applied not only to image restoration tasks but also to any CNN that requires pooling operations. Based on frequency analysis, Qin et al. [33] had mathematically proved that traditional global average pooling is a special case of frequency domain feature decomposition. Multi-spectral channel attention, termed as FcaNet, generalizes the compression of the channel attention mechanism in the frequency domain.

In sonar image tasks, some scholars have tried to use the addition frequency processing method to carry out deep mining of image information. Zhu et al. [36] used a protrusion detection technique based on frequency analysis to separate the acoustically highlighted area from the surrounding environment. This segmentation roughly locates the diver's target and generates a region of interest (ROI). Wang et al. [37] had constructed a novel recurrent pyramid frequency feature fusion network (RPFNet) by enhancing and fusing the different frequency characteristics of SSS images, using the residual structure and attention mechanism which effectively extracted fine-grained features, reduced background information interference, and improved nonlinear feature representation capabilities.

Although the above sonar image detection method uses the concept of frequency, its essence is only borrowed from this concept and does not convert the image into the frequency domain. The main reason is that the points on the spectrogram do not have a one-to-one correspondence with the points on the image, so the frequency domain information cannot be directly used in the network method.

2.3. Public Database and Benchmark for SSS Images

Data is the input to deep learning methods. It is important but easy to overlook. While building accurate algorithms and the application of computational methods is part of the process to achieve the task, a good machine learning project, from structuring the model to landing testing, is based on using the right datasets. For a machine learning model, the quality of the output largely depends on the amount and quality of the input.

The Pattern Analysis, Statistical Modelling and Computational Learning (PASCAL) Visual Object Classes (VOC) project [39] is a competition led by Mark Everingham and others, which is a very large competition from 2005 to 2012, initially mainly used for target detection, and finally contains 5 competitions, classification, detection, segmentation, action classification, person layout. ImageNet [41] is to build a large database of computer vision research, with keywords selected from WordNet. The complete Imagenet dataset has more than 14 million images, covering more than 20,000 categories of annotations and more than one million bounding box annotations, each category is about 500 1000 images. Common Objects in Context (COCO) [40] takes scene understanding as the goal, especially selecting more complex daily scenes, compared to the establishment of Pascal to promote the target detection task, the establishment of COCO is to promote the positioning and segmentation task in the natural background, so the target in the image is calibrated by a very accurate segmentation mask.

In the classification, detection, and segmentation tasks using SSS images, some datasets are established to make up for the lack of data in this field: WH-Dataset and QD-Dataset [56] are designed to validate the advantage of the method Wang et al. had proposed. The self-made sonar common target detection dataset (SCTD) [42] is designed for detection tasks with labeling in VOC format. The SeabedObjects-KLSG dataset [43] is a real SSS image dataset that focuses on multiple classes of sonar images of drowners, shipwrecks, aircraft, mines, and seabeds.

However, not all of the above datasets are available. Among them, the public SSS images dataset that can be obtained so far are SCTD and part of KLSG. And only SCTD are designed for target detection tasks.

3. Methods

In this section, we present in detail the structure of D2MFNet and the benchmark dataset. Firstly, the composition of the main network structure, consisting of MFCAM and D2FPN, is described. Among them, MFCAM focuses on information on different frequency components to better experience the information of the target's grayscale changes and edge contours. D2FPN introduces frequency domain features to further mine image feature information. Finally, the collected data set is processed, and the benchmark is built and analyzed.

3.1. Multi-Frequency Combined Attention Module

To obtain more feature information in a situation where sonar images' quality is quite terrible that even human eyes cannot figure out the specific target type, we first elaborate on the Fast Fourier Transform (FFT), which is the underlying algorithm of the following module, and then construct the D2FPN to realize feature extract and fusion in time and frequency domain.

3.1.1. Fast Fourier Transform

Fast Fourier Transform (FFT) is a fast Fourier transform that simplifies the computational complexity of Discrete Fourier Transform (DFT). The two-dimensional DFT is a transformation method that converts an image from the spatial domain to the frequency domain. And the formula of the two-dimensional Discrete Fourier Transform (2D DFT) is as follows:

$$F(u,v) = \sum_{x=0}^{M_x-1} \sum_{y=0}^{M_y-1} f(x,y) e^{-j2\pi(\frac{ux}{M_x} + \frac{vy}{M_y})}, \tag{1}$$

which $f(x,y)$ represents the spatial domain matrix of size $M_x \times M_y$, $F(u,v)$ represents the Fourier transform of $f(x,y)$, u and v can be used to determine the frequency of the sine-cosine, $u,v = 0,1,2,\ldots,M_u - 1|M_v - 1$, and $M_u \times M_v$ is the frequency domain matrix size after the transform.

Correspondingly, the inverse 2D DFT can be written as:

$$f(x,y) = \frac{1}{M_u \cdot M_v} \sum_{u=0}^{M_u-1} \sum_{v=0}^{M_v-1} F(u,v) e^{j2\pi(\frac{ux}{M_x} + \frac{vy}{M_y})}, \tag{2}$$

which $x,y = 0,1,2,\ldots,M_x - 1|M_y - 1$.

The size of the frequency domain matrix is the same as the size of the original spatial domain matrix. Each point in the frequency domain matrix represents a function with frequencies u,v, and the combination of these functions in the spatial domain is the original function $f(x,y)$.

With 2D FFT, we choose three images from different classes in SCTD to show the result after frequency domain transfer. As shown in Figure 1, the amplitude and the phase spectrum after spectral centralization are shown. Spectrum centralization makes the spectrum distribution in the frequency domain show the law of low middle and high encirclement. The amplitude spectrum has obvious signal structure characteristics, only contains the periodic structure contained in the image itself, and does not indicate where it is. The phase spectrum resembles a random pattern, and the movement of an object in space is equivalent to the phase shift in the frequency domain, making the phase spectrum equally important.

doi:10.20944/preprints202309.2024.v2

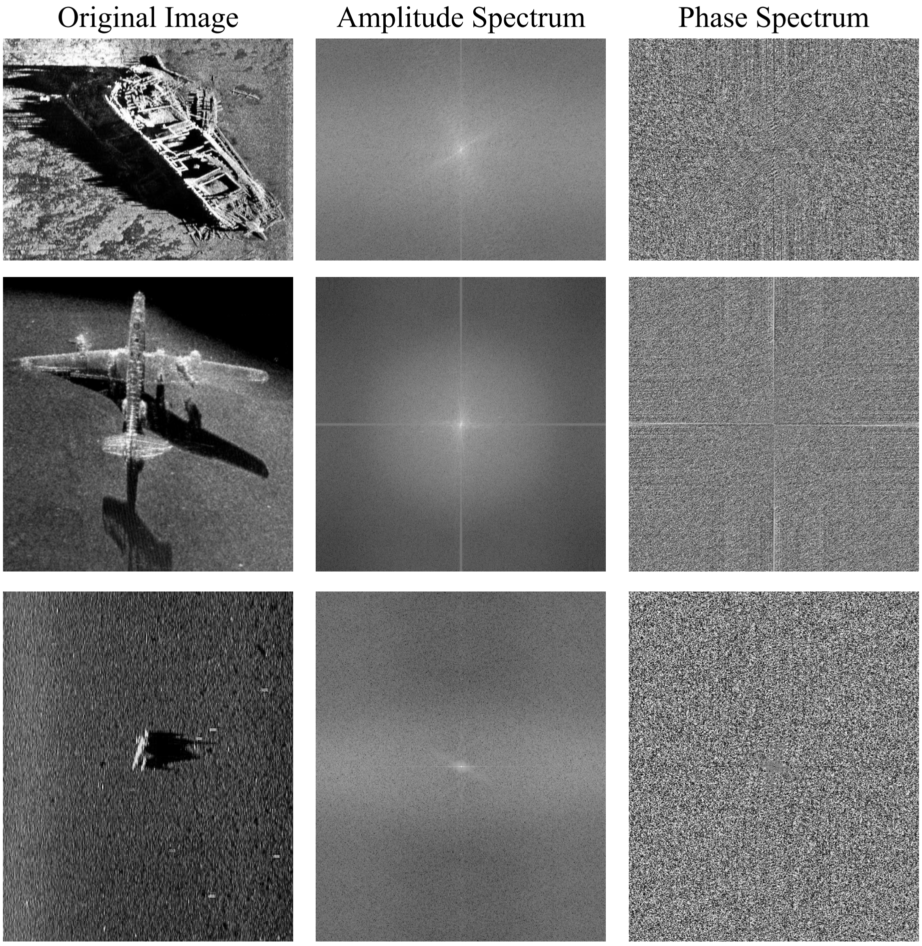


Figure 1. Centralized spectrogram after frequency domain conversion of SSS image, including amplitude map and phase map.

The high-frequency components in the frequency domain of the image correspond to the detailed information of the image, and the low-frequency components of the image correspond to the contour information of the image. The high-frequency component represents the abrupt part of the signal, while the low-frequency component determines the overall image of the signal. In the spectrogram, you can see points with different brightness, where a large brightness proves a large gradient which is the high-frequency component and a small brightness proves a small gradient at that point which is the low-frequency component.

3.1.2. construct of MFCAM

The high-frequency components in the frequency domain of the image correspond to the detailed information of the image. And the low-frequency components of the image correspond to the contour information of the image. In the task of target detection, the common methods only use low-frequency information to learn image features. This is a disadvantage when SSS images have less data. In this case, both channel and spatial attention mechanisms are essential, and additional attention information from high-frequency components is required. Therefore, based on the CBAM structure [57], we added attention structures with different frequency ranges as shown in Figure 2 to achieve a more accurate detection effect of underwater targets.

doi:10.20944/preprints202309.2024.v2

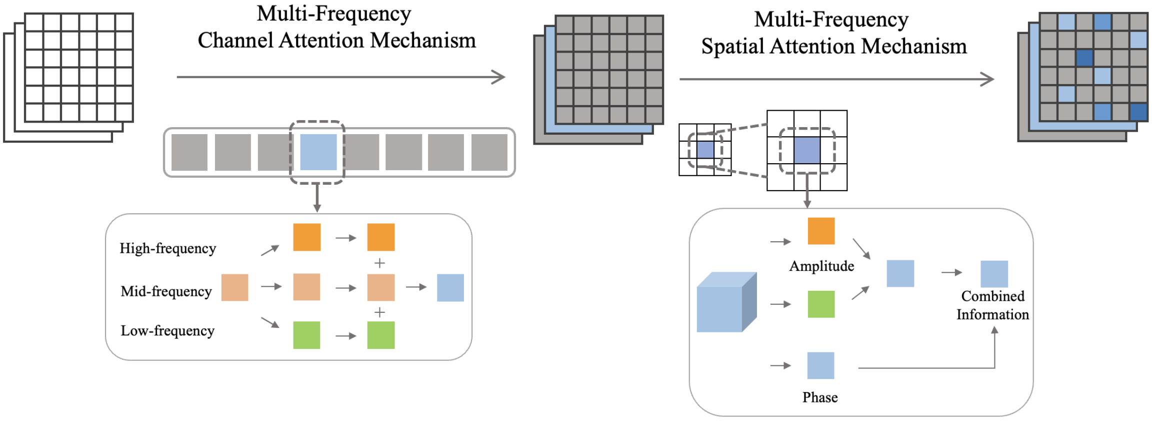


Figure 2. The overall structure of the multi-frequency combined attention mechanism, including channel attention module and spatial attention module.

Multi-frequency channel attention module.

In the attention mechanism of fusing global and local information, the channel attention mechanism can better fuse the global information to express the overall features of the feature map in different channels during feature extraction. Based on the existing channel attention mechanism structure, the frequency domain feature extraction method is added to extract and combine the channel weights in the high-frequency, low-frequency and other frequency ranges, which can better integrate the limited feature information.

In the structure shown in the Figure 3, when extracting the channel weights, the original feature map is converted into the frequency domain, and filtering operations in different frequency ranges are performed to extract the feature maps of high-frequency, low-frequency and other frequency ranges. On this basis, the filtered feature map is under the global average pooling, the channel feature weights of different channels and different frequency ranges are extracted, and then the weights of the three frequency ranges are added to obtain the final channel feature weights. The one-dimensional channel weights are multiplied by the original feature map one by one, to obtain the feature map processed by the multi-frequency channel attention mechanism module, and enter the next module.

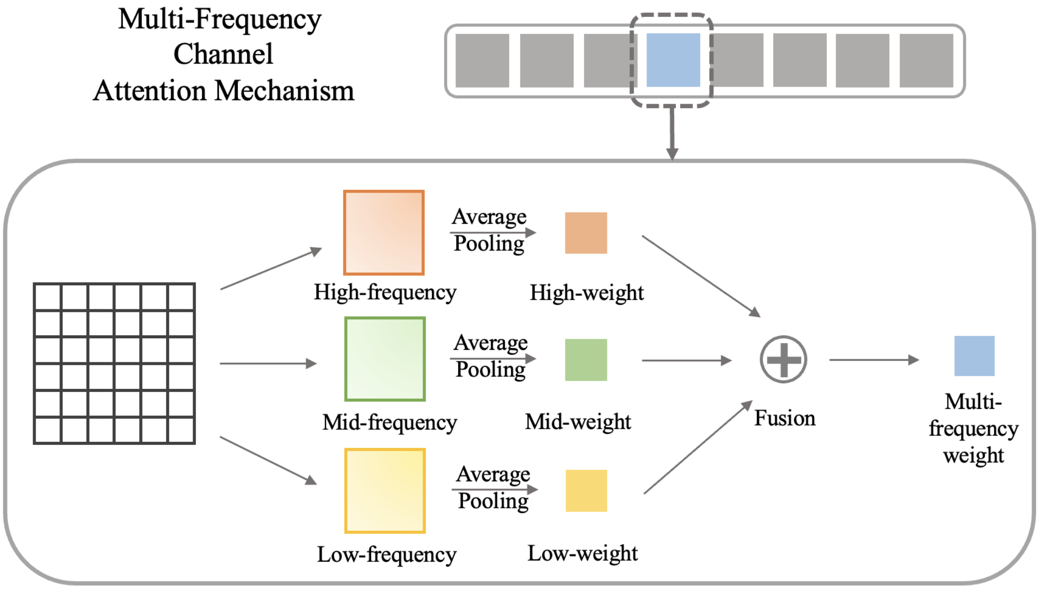


Figure 3. The structure of multi-frequency channel attention mechanism, in which information from high, low, and other frequency ranges is extracted.

Multi-frequency spatial attention module.

Different from the channel attention mechanism, the spatial attention mechanism adds corresponding weights to the feature map in a channel according to the set area size to achieve the effect of extracting local more important feature information in the same space. Based on the existing spatial attention mechanism structure, and considering the existence of different feature information in amplitude and phase after frequency domain conversion, according to the different distribution characteristics of amplitude and phase information, the two are combined to better express the key feature information in the region. The structure of the multi-frequency spatial attention mechanism based on FFT is proposed in this section.

In the structure shown in Figure 4, the frequency domain conversion is performed first, and the amplitude information and phase information are extracted separately. Set the window size of 2*2 or 3*3, and perform maximum pooling and minimum pooling of amplitude information to extract the characteristic weights of high-frequency and low-frequency ranges. At the same time, the phase information is under average pooling using the set window size. Multiply the weights with the feature maps of the original maps to obtain three feature maps with different treatments in the same space. The two feature maps of amplitude processing are combined and then combined with the feature maps of phase processing to form the final feature map processed by the multi-frequency spatial attention mechanism module, thus ending the overall process of the attention mechanism.

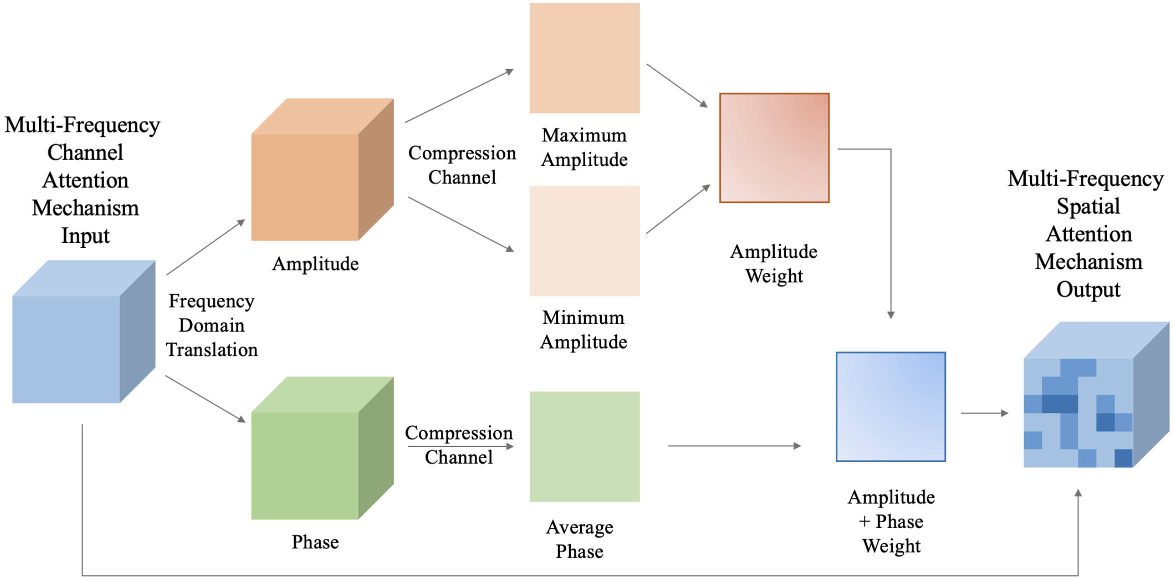


Figure 4. The structure of multi-frequency spatial attention mechanism, in which the high-frequency and low-frequency information in the amplitude, as well as the phase spatial structure information, are extracted.

3.2. Dual-Domain Feature Pyramid Network

3.2.1. Construct of D^2FPN

After the frequency domain conversion of the image, it has high-frequency information that cannot be perceived by the human eye, which is the characteristic information that is not used in the current acoustic image target detection method. However, due to the particularity of one-to-one correspondence between pixels after frequency domain conversion, the high-frequency information of the image cannot be directly obtained, and the further fusion between time-domain feature information and high-frequency feature information cannot be carried out directly. At the same time, the use of high-frequency information alone cannot produce the effect of increasing the target characteristic information, and it is extremely difficult to convert the detected targets in the high-frequency into targets in the time domain.

As shown in Figure 5, the D^2FPN proposed in this section converts the frequency domain of the specific feature map at the feature fusion nodes, and filters it to obtain its high-frequency information, while other feature maps perform frequency domain conversion, and after the amplitude information and phase information are fused separately in the frequency domain, it is converted into a time domain image, so as to complete the multi-feature map fusion. The whole structure is divided into bottom-up and top-down feature fusion, and large-scale information is fused into small-scale and small-scale information into large-scale. In this structure, the high-frequency information after frequency domain conversion is introduced into the original feature map of the bottom-up feature fusion process and the top-down fusion process.

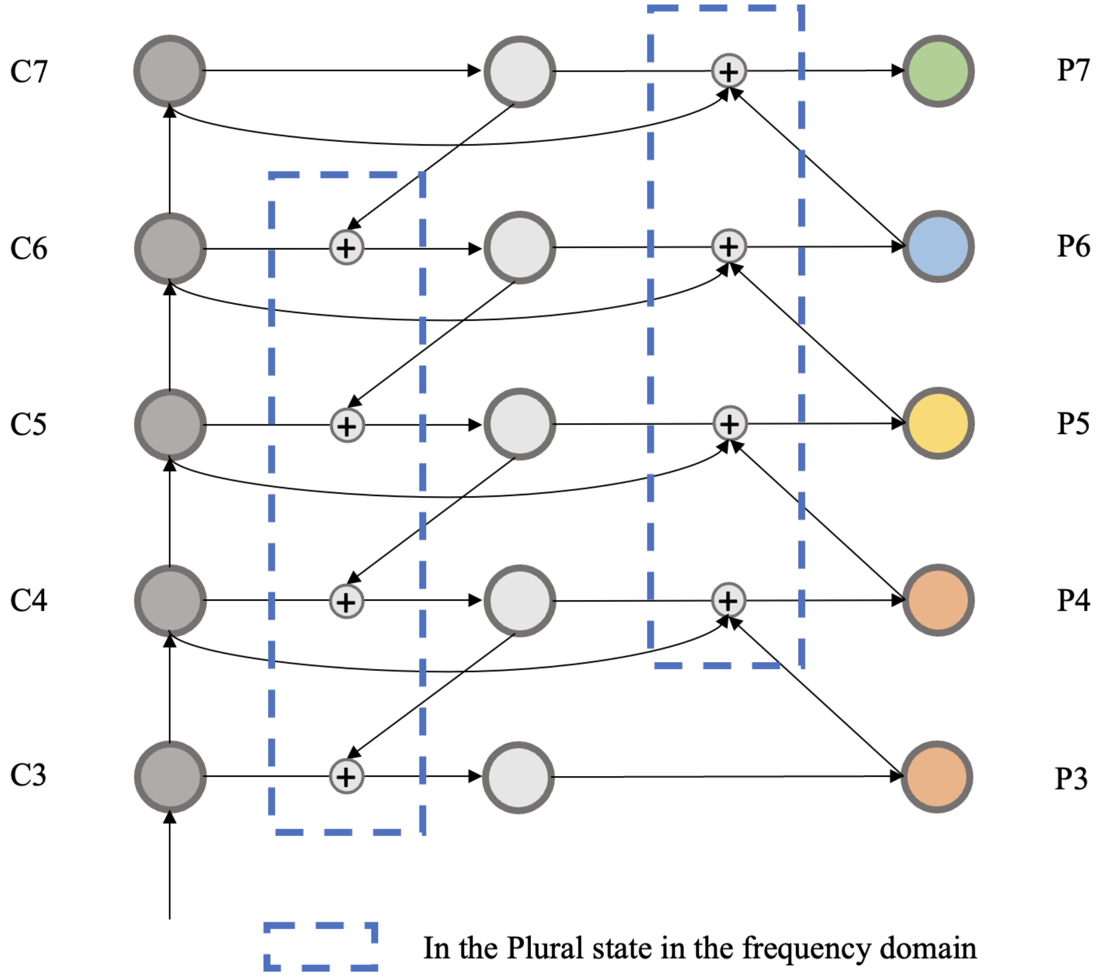


Figure 5. This is a single-cycle structure of dual-domain pyramid feature network structure, in which the blue dotted box is the part that needs to extract high-frequency information in the plural state.

The BiFPN structure enhances the global perception of the model, while the D^2FPN structure proposed in this section enhances the high-frequency information perception ability of the detection model on the basis of the original global perception.

3.2.2. Structural Detail Flow

Due to the large number of settings and fusions of this structure, the implementation process is also more complicated. Figure 6 shows the implementation process of a specific code module, which has the following operation process:

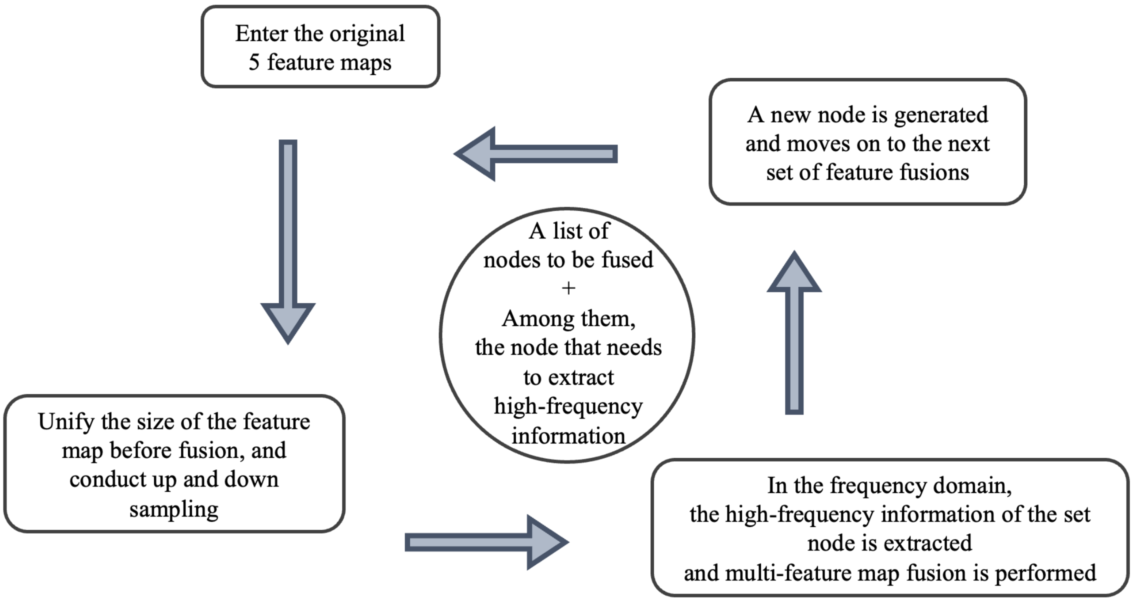


Figure 6. The processing cycle of the dual-domain pyramid revolves around the nodes that need to be fused.

1) Extract high-frequency information:

The frequency distribution of the uncentralized spectrogram is represented by intermediate low-frequency and high-frequency around them. Set the high-frequency filtering ratio to 0.15 which means the amplitude value in the quadrilateral with 0.75 sides and centered on the original image is set to zero. And then compound the amplitude with the phase information, at this time the value in the image is under the plural state.

(2) Up and down sampling:

The feature map that needs to be fused is calibrated for size consistency. The output image size of the set fusion nodes is compared with the input graph. The input image size is downsampled if the size is greater than the output size, and the upsampling operation is completed by bilinear interpolation if the input image size is smaller than the output image size. This unifies all image sizes to be fused.

(3) Fusion multi-feature map:

The set image nodes that need to be fused will be converted at the same time after (2) operation, and the node sequence is shown in Figure 7. The features that need to be extracted from high-frequency information are operated in (1), and each feature map is in the plural status. The amplitude and phase information of each feature map are separated, fused separately, and combined into a plural state. Finally, time domain conversion is carried out to complete the fusion of multiple feature maps.

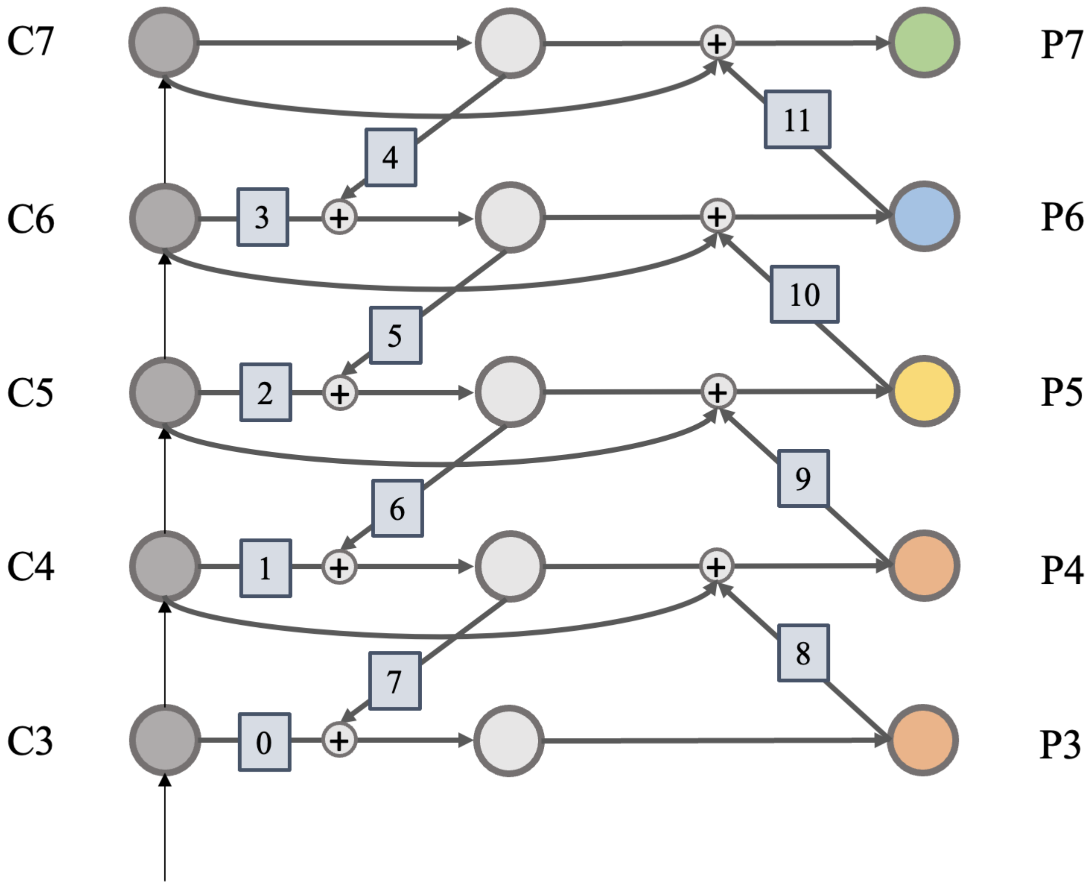


Figure 7. The fusion nodes are labeled to clarify the source of the feature map required for fusion when D^2FPN is processing.

(4)Perform multiple repetitions of structures:

After a single structure is completed, multiple cycles can be made to strengthen its global perception as well as high-frequency perception. The number of loops is set in the training configuration.

3.3. Dual-Domain Multi-Frequency Network for Detection

Due to the small scale of acoustic image data, and in some cases, the human eye cannot identify the data quality of the target, the data scale and its quality become the task difficulties that this section focuses on. This section proposes a dual-domain multi-frequency acoustic image target detection network model, which adds the multi-frequency combined attention mechanism module and dual-domain pyramid module to the backbone network of the detection model feature extraction and the neck structure before the head structure of the detection method.

In addition, the multi-frequency combined attention mechanism is combined with the ResNet backbone network to emphasize the feature information of the multi-frequency range in the process of feature extraction. The ResNet backbone network has four feature output layers, and the multi-frequency combined attention mechanism module will be multi-frequency enhanced in the last three output feature layers.

The dual-domain pyramid module is used as a separate module after the feature extraction backbone network to enhance global perception and high-frequency perception.

3.4. Benchmark Dataset

As introduced in related work, since the number and scale of SSS image datasets currently disclosed are far less than those of target detection in other fields, we search and use two publicly available SSS image datasets: SCTD and KLSG for related research. At the same time, method experiments using these two datasets can prove that our proposed D^2MFNet is more robust.

The first dataset, SCTD, is a dataset designed for SSS image target detection tasks, which already contains the target annotation information required by this model, and does not perform additional processing for annotation. SCTD contains a total of 357 images in three categories, including 271 ships, 35 humans, and 57 aircraft. The second dataset, KLSG, contains two categories and a total of 447 images, including 395 ships and 62 aircraft, but KLSG is a dataset designed for SSS image classification tasks and does not have annotated content related to object targets. We used the VOC annotation format as a standard to annotate the target objects required for the target detection task for KLSG.

Since the SCTD doesn't have evaluation metrics analysis based on multiple detection methods, and the original KLSG is a classification task dataset with no such evaluation metrics content, we put two datasets together for benchmark building and comparison. To establish the datasets benchmark for general target detection methods comparison and analysis, we choose these famous one-stage and two-stage general target detection methods with generally better results, such as Faster R-CNN, Cascade RCNN, Sparse R-CNN, SSD512, Retina Net, YOLOv5, YOLOv7, Deformable DETR, DAB-DETR, DINO and the D^2MFNet we proposed. AP50(Average Precision with Intersection over Union greater than 50%) and mAP(mean Average Precision) were used as the main evaluation metrics in this benchmark. And the metrics we use in this section and in the experiment section will be introduced and detailed in the Evaluation Metrics subsection. The benchmark test results are shown in Table 1

At the same time, during training the methods for the datasets, we consider that the smaller size of the dataset, the greater impact on the method training. We also briefly verify this concern in the case of training with pre-trained parameters and with no pre-trained parameters. The result is that under training without pre-trained parameters, the AP value of most models will be lower than 0.1%, which is more difficult to analyze and compare. Therefore, we use pre-trained parameters such as vgg16 and resnet50 in the one-stage and two-stage methods training to achieve a more visible data comparison effect.

We don't use data amplification methods, such as cropping, flipping, Mixup, Cutout, Mosaic, etc. This is because using data amplification in very small datasets makes the method training process more prone to overfitting. In fact, such data amplification in a very small data set only makes multiples for a small number of target feature replications, resulting in a large number of duplicate features. However, this effect is relatively small in ordinary large datasets and the positive effect of data amplification in such data sets is far greater than the negative effect.

Moreover, in order to reduce the adverse effect of overfitting on the metrics evaluation, the data in the table are tested and inferred using the convergent model. And there are multiple versions in the YOLO series of methods, such as YOLOv5 has four versions of YOLOv5-s, YOLOv5-l, YOLOv5-m, YOLOv5-x, and YOLOv7 also has five versions of YOLOv7-l, YOLOv7-x, YOLOv7-w, YOLOv7-e, YOLOv7-d. In the benchmark, basic settings such as YOLOv5-s and YOLOv7-l are used.

Table 1. The benchmark of SCTD and KLSG with pre-trained methods by using methods carried out recent years and by using our methods.

				KLSG			SCTD			
Method	**Year**	**Backbone**	**Input size**	**ship**	**airplane**	**mAP**	**ship**	**airplane**	**human**	**mAP**
Faster RCNN [10]	2015	ResNet50	1000, 600	0.432	0.233	0.333	0.497	0.452	0.312	0.420
Cascade RCNN [13]	2018	ResNet50	1000, 600	0.464	0.312	0.388	0.459	0.457	0.433	0.450
Sparse RCNN [58]	2021	ResNet50	1000, 600	0.106	0.091	0.099	0.052	0.033	0.001	0.028
SSD512 [14]	2016	VGG16	512,	0.353	0.006	0.180	0.141	0.000	0.149	0.097
Retina Net [59]	2017	ResNet50	1000, 600	0.126	0.044	0.085	0.047	0.025	0.000	0.024
YOLOv5 [16]	2020	CSPDarknet	512,	0.114	0.101	0.107	0.188	0.030	0.397	0.205
YOLOv7 [17]	2022	YOLOv7	512,	0.138	0.105	0.121	0.115	0.014	0.117	0.082
Deformable DETR [60]	2021	ResNet50	1000, 600	0.257	0.170	0.213	0.214	0.061	0.202	0.159
DAB-DETR [61]	2022	ResNet50	1000, 600	0.051	0.010	0.031	0.214	0.061	0.202	0.159
DINO [62]	2023	ResNet50	1000, 600	0.438	0.079	0.258	0.498	0.124	0.106	0.243
D^2MFNet	ours	ResNet50	1000, 600	**0.784**	**0.418**	**0.601**	**0.786**	**0.675**	**0.630**	**0.697**

As shown in Table 1, compared with the general object detection methods with better results in recent years, the D2MFNet proposed in this paper has obvious advantages. There are certain laws of difference between categories, between datasets, between methods, and between two-pair or three-pair combinations among them.

(1) Between categories:

It is obvious that the AP50 of the ship category with more data is higher, while the AP50 of the aircraft and human categories is much smaller. One of the main reasons is that the number of ship categories is much larger than other categories, and its effective target characteristics are much more.

(2) Between datasets:

In Faster R-CNN, Cascade RCNN, and YOLOv5, the SCTD dataset has better results. The reason may be that SCTD is an RGB dataset, and its original input channel is 3, which is more in line with the structure of conventional general target detection methods, including the basic method, Cascade RCNN, used in this paper. The experimental results in the Benchmark illustrate this point. There is also the possibility that the categories distribution of SCTD is more reasonable. However, since both datasets are small-scale datasets, it is not possible to further verify the proportion of this cause in this phenomenon, and considering the small sample size, cross-validation is not possible, so this possible problem is ignored. At the same time, it is guessed that the KLSG was generated as an RGB dataset when it was acquired. When producing SSS images, many SSS equipment manufacturers prefer to show SSS images in the form of color drawings, which can let the naked eye better distinguish the target and can better promote their own products. But the SSS images should be grayscale images, and it was converted into a grayscale image when the dataset was made. The process of changing from a monochromatic image to a multi-layer image and then to a monochromatic image will result in poor dataset quality.

(3) Between methods:

The overall mAP of the one-stage methods is lower than that of the two-stage methods. It is mainly because the one-stage methods do not produce candidate regions, they directly perform the class probability and position coordinate value operation of the object, so the final detection result can be directly obtained after a single detection. The one-stage methods have a faster detection speed, but the accuracy rate is worse than the two-stage methods. However, due to the small amount of data, the one-stage methods do not have an advantage in speed in the experiments in this paper. What's more, poor methods such as Sparse R-CNN may be designed based on large-scale data to ensure accuracy while reducing training parameters and time, and the experimental results are also in line with our expectations, which are most likely due to data scale and category distribution.

These are all possible factors that affect the method training results.

4. Experiment

4.1. Implementation Details

We built D^{2}MFNet based on the MMDetection open-source framework version 2.28.1. The environment used for model training and testing is ubuntu20.04 system, the hardware conditions are Intel(R) Xeon(R) Gold 6130 CPU and NVIDIA V100-32GB GPU, and the programming environment is Python 3.8, Pytorch 1.11.0, CUDA 11.3.

4.2. Evaluation Metrics

Based on the general target detection evaluation index [19] when training VOC datasets [39], we effectively evaluate our proposed method and compare it with the existing state-of-the-art method.

Intersection over Union (IoU) can be understood as the degree of coincidence between the bounding box and the ground truth. The calculation method is the intersection of the detection result and the ground truth than their union:

$$IoU = \frac{BoundingBox \cap GroundTruth}{BoundingBox \cup GroundTruth}, \tag{3}$$

where *BoundingBox* is the box predicted by the method, and *GroundTruth* is the box marked in the original picture.

Recall and precision are used to evaluate the effectiveness of target detection methods. The formula is as follows:

$$Precision = \frac{TP}{TP + FP}, \tag{4}$$

where TP means True Positive which represents the number of positive samples that are recognized as positive samples, and FP means False Positive which represents the number of negative samples that are incorrectly identified as positive samples. Precision is the percentage of True positives in the identified image.

$$Recall = \frac{TP}{TP + FN}, \tag{5}$$

where FN means False Positive which represents the amount of positive sample that is incorrectly identified as a negative sample. Recall is the proportion of all positive samples in the test set that are correctly identified as positive.

Typically, the Precision-recall (PR) curve or receiver operating characteristic (ROC) curve is used for visual analysis of model effects. The PR curve uses Precision, so both indicators of the PR curve focus on positive examples. The SSS image dataset used in this paper belongs to the category imbalance problem, which is mainly concerned with positive cases, so the PR curve is widely considered to be better than the ROC curve in this case.

The process of calculating and plotting the Precision-recall curve is as follows: changing the recognition threshold so that the system can identify the first K images in turn, and the change in the threshold will also cause the Precision and Recall values to change, to obtain the curve.

Average Precision (AP) is the area below the PR curve, and generally speaking, the better a classifier, the higher the AP value:

$$AP = \frac{1}{N}\sum_{i=1}^{N} P_i \cdot \Delta R_i, \tag{6}$$

where N represents the number of classes, P_i represents the Precision when the abscissa is R_i on the precision-recall curve of class i, ΔR_i represents the change in the recall value of the class i from $R_i - 1$ to R_i.

The mean Average Precision (mAP) is the average of APs in multiple categories. The size of mAP must be in the [0,1]. This indicator is one of the most important of the target detection methods.

4.3. Comparison with State-of-the-art Methods

Including the proposed multiple frequencies combined channel spatial combination attention mechanism module and the frequency domain-time domain dual-domain feature pyramid and the detection model of dual-domain multi-frequency acoustic image target detection network is constructed based on the Cascade RCNN framework.

4.3.1. MFCAM Implementation

The experiment was analyzed from three aspects: different datasets, different categories and different methods. And the experiment does not use dataset enhancement methods to amplify the dataset. The experimental results are shown in Table 2.

Table 2. Experimental results of multi-frequency attention mechanism in KLSG and SCTD datasets compare with that of CBAM and no attention mechanism.

Method	Module	Dataset	mAP	ship	aircraft	human
Cascade RCNN	-	KLSG SCTD	0.388 0.450	0.464 0.459	0.312 0.457	- 0.433
Cascade RCNN	CBAM	KLSG SCTD	0.429 0.501	0.540 0.701	0.317 0.478	- 0.325
Cascade RCNN	MFCAM	KLSG SCTD	0.598 **0.656**	**0.803** 0.719	0.393 **0.600**	- **0.649**

As shown in Table 2, there is a significant improvement in average category accuracy overall. In the module that does not apply to the attention mechanism, the mAP is 0.388, which is lower than that of 0.429 and 0.598 in the CBAM and MFCAM experiments using the attention mechanism. In terms of the target detection recognition rate of aircraft, ships and humans, it is also better to use the attention mechanism. and even in some target detection, the method of not using the attention mechanism has false alarms. In the detection experiments of CBAM and MFCAM, it is obvious that MFCAM has a higher mAP and sample recognition rate. In the dataset SCTD, MFCAM presents the same experimental results. Therefore, the comparative experiments of different methods on different datasets show that the multi-frequency channel and spatial attention mechanism proposed in this section can significantly improve the target detection effect of Cascade-RCNN.

Table 2 still has some phenomena different from the above rules, which can be divided into two aspects of data set differences and category differences:

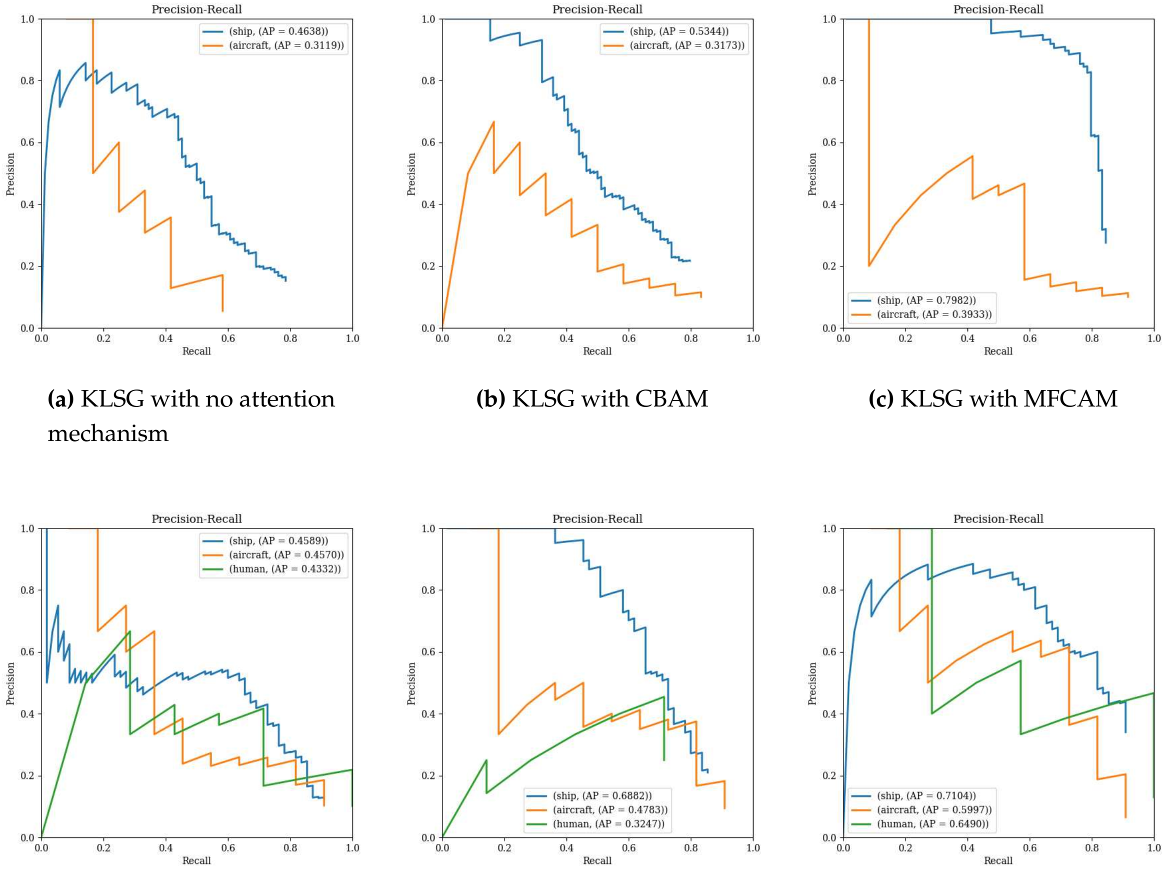
(1) Dataset differences:

Compared with the precision differences of each category in KLSG, the accuracy difference between SCTD samples is smaller, which is most likely due to the fact that SCTD takes into account the distribution of samples when the dataset is made, and KLSG as a classification task dataset does not explore this aspect, so this phenomenon is caused.

(2) Category differences:

The precision difference between different categories is also more significant, and the precision of the ship category is significantly higher than that of aircraft and humans. This phenomenon is largely due to the uneven distribution of class samples, and because aircraft targets and ships are more similar in the wreckage, more target misdetection samples will be generated.

In order to further analyze the model detection effect of different modules, the Precision-Recall (PR) curve is used to show the effect of the analysis model. The PR curve is shown in Figure 8.



(a) KLSG with no attention mechanism

(b) KLSG with CBAM

(c) KLSG with MFCAM

(d) SCTD with no attention mechanism

(e) SCTD with CBAM

(f) SCTD with MFCAM

Figure 8. The PR curve of detection effect with attention mechanism. (a) is the curve of using the KLSG dataset without attention mechanism, (b) is the curve of using the KLSG dataset with the CBAM method, (c) is the curve of using the KLSG dataset with the multi-frequency combined attention mechanism, (d) is the curve of using SCTD dataset without attention mechanism, (e) is the curve of using the SCTD dataset with the CBAM method, (f) is the curve of using the SCTD dataset with the multi-frequency combined attention mechanism.

In Figure 8, it can be seen that the resulting law is consistent with Table 2, and the three model state experiments in which different categories in the same dataset do not use the attention mechanism, the CBAM method and the MFCAM method are gradually improved. At the same time, due to the small number of data samples, the curve fluctuates greatly, forming a jagged curve. However, the PR curve of the human class did not improve the accuracy with the change of attention mechanism, because the amount of data was too small and much smaller than the other two categories, resulting in unbalanced samples and low stability in classes.

4.3.2. D^2FPN Implementation

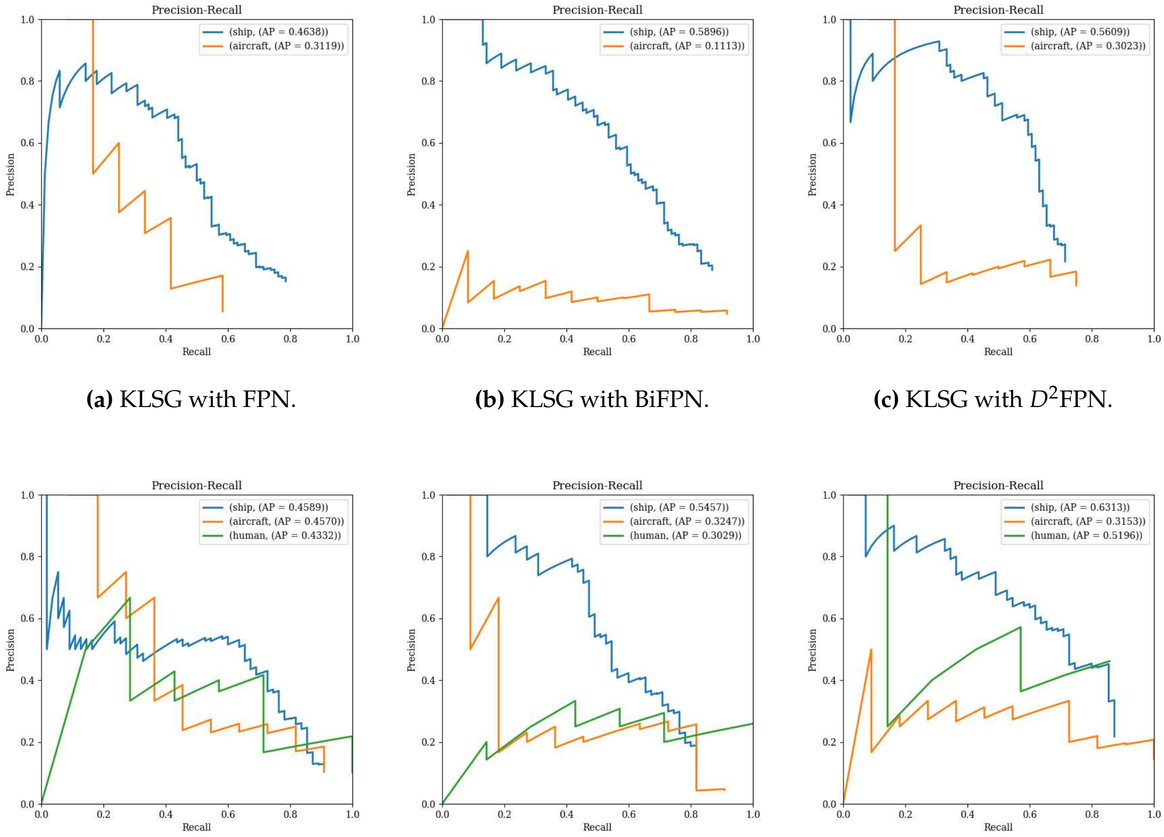
Compared with the traditional feature pyramid network and BiFPN, the dual-domain feature pyramid network is improved in multi-scale feature fusion, so this section designs different methods and comparative experiments between different datasets to verify the detection effect of the proposed method. The experimental setup is the same as in the section above, and the experimental results are the average precision of the model convergence point. The experiment was analyzed from three aspects: different datasets, different categories and different modules. The experimental results are shown in Table 3.

Table 3. Experimental results of dual-domain feature pyramids in KLSG and SCTD datasets compare with that of FPN and BiFPN.

Method	**Module**	**Dataset**	**mAP**	**ship**	**aircraft**	**human**
Cascade RCNN	FPN	KLSG SCTD	0.388 0.450	0.464 0.459	0.312 0.457	- 0.433
Cascade RCNN	BiFPN	KLSG SCTD	0.350 0.391	0.590 0.546	0.111 0.325	- 0.303
Cascade RCNN	D^2FPN	KLSG SCTD	0.434 **0.489**	0.565 **0.631**	0.302 0.315	- **0.520**

In the two different datasets, the effect of BiFPN is within 0.1 different from that of the experiment without module addition, and the effect of using it alone is poor. D^2FPN shows a better target detection effect in SCTD dataset. From the analysis of the mAP results, it is clear that Cascade RCNN without using any modules performs better than BiFPN and is inferior to the D^2FPN module proposed in this section. In ship detection, the performance of the KLSG and SCTD datasets is basically comparable. For different methods, the detection precision of the D^2FPN module proposed in this paper is higher than that of BiFPN, and the precision of BiFPN is higher than that of Cascade RCNN without any modules. In aircraft detection, although the detection precision of the three different experimental setups is not high, the D^2FPN is comparable to the BiFPN precision and higher than that of the Cascade RCNN. In human detection, KLSG has no experimental results because it does not contain such a class in the dataset, and the precision of D^2FPN detection is better than that of Cascade RCNN, and BiFPN has the worst precision. According to the overall experimental results, the proposed D^2FPN method can be adapted to the complex detection tasks of a variety of targets and has a great improvement compared with the feature pyramid network method at this stage.

Figure 9 shows the consistent result law in Table 3, in the same data set of different categories in the three model state experiments using FPN, BiFPN and the dual-domain feature pyramid network proposed in this paper. BiFPN has not improved, while the dual-domain feature pyramid network has been improved.



(a) KLSG with FPN.

(b) KLSG with BiFPN.

(c) KLSG with D^2FPN.

(d) SCTD with FPN.

(e) SCTD with BiFPN.

(f) SCTD with D^2FPN.

Figure 9. The PR curve of detection effect with feature pyramid network. (a) is the curve of using the KLSG dataset without using special FPN, (b) is the curve of using the KLSG dataset with the BiFPN method, (c) is the curve of using the KLSG dataset with the dual-domain feature pyramid network, (d) is the curve of using the SCTD dataset without using special FPN, (e) is the curve of using the SCTD dataset with the BiFPN method, (f) is the curve of using the SCTD dataset with the dual-domain feature pyramid network.

4.4. Ablation Study

The dual-domain multi-frequency feature fusion target detection method includes the multi-frequency fusion attention mechanism and the dual-domain feature pyramid network. When there are multiple modules in a method that can be independent of each other, the impact of the interaction between a single module and multiple modules on the method results can be tested separately, which can better adjust the model as a whole. This section sets up an ablation experiment, as shown in Table 4, to verify that each module contributes to the improvement of model accuracy.

Table 4. The index evaluation of the multi-frequency combined attention mechanism and the dual-domain feature pyramid network is carried out under the two datasets of KLSG and SCTD.

Method	AM Module	FPN Module	Dataset	mAP	ship	aircraft	human
Cascade RCNN	-	-	KLSG	0.388	0.464	0.312	-
	-	-	SCTD	0.450	0.459	0.457	0.433
	MFCAM	-	KLSG	0.598	0.803	0.393	-
	MFCAM	-	SCTD	0.656	0.719	0.600	0.649
	-	D^2FPN	KLSG	0.434	0.565	0.302	-
	-	D^2FPN	SCTD	0.489	0.631	0.315	0.520
	MFCAM	D^2FPN	KLSG	0.601	0.784	0.418	-
	MFCAM	D^2FPN	SCTD	**0.697**	0.786	**0.675**	0.630

Although the effect of D^2FPN is slightly inferior to MFCAM on both datasets, when D^2MFNet combines the two modules, its effect does not contradict, but still improves on the basis of MFCAM, which uses the best effect in a single module. The table is sufficient to prove that the MFCAM, D^2FPN, and D^2MFNet proposed in this paper have good performance in the SSS dataset without data augmentation.

5. Conclusion

This paper mainly comes from the environmental recognition task in underwater scenes, and the optical imaging methods that can be used in shallow seas but cannot be used in both medium and deep seas. As a result, the use of sonar images for underwater target detection tasks has become the main choice of scholars. However, the number of publicly available underwater SSS image datasets is small, and sea trials are more difficult in a short period of time.

On this basis, in order to obtain more feature information expression from SSS images, we use the method of frequency analysis to mine the depth features in a small number of image data in the frequency domain. Considering that the frequency characteristics of different ranges in the frequency domain are also different, we apply unique attention weights to different frequency ranges, named MFCAM, to achieve the effect of further mining information. At the same time, the D^{2}MFNet proposed in this paper concludes a dual-domain feature pyramid network that combines domain transformation and features, named D^{2}FPN, which solves the problem that the frequency domain conversion cannot correspond to the pixel position of the original image. The experimental results showed that these two modules work effectively both using alone or together, and our methods are state-of-the-art. Therefore, our results have practical implications for related research. In addition, we collected and processed some publicly available data and provided a benchmark for other scholars to refer to.

Our work doesn't change the backbone of the detection method which can make us dig for more information on the feature expression. In the next step, we will continue to carry out relevant work research on the basis of this article.

Author Contributions: Conceptualization, methodology, validation and writing—original draft preparation, Wen Wang; methodology, writing—review and editing, Yifan Zhang; supervision, funding acquisition and formal analysis, Houpu Li; data curation and visualization, Xue Gong; visualization, Lei Liu; software, Yixin Kang. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Science Fund for Outstanding Young Scholars grand number 42122025, Natural Science Foundation of China grant number 41974005 and 42374050.

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

References

1. Character, L.; Ortiz JR, A.; Beach, T.; Luzzadder-Beach, S. Archaeologic Machine Learning for Shipwreck Detection Using Lidar and Sonar. *Remote Sensing* **2021**, *13*. doi:10.3390/rs13091759.

2. Borrelli, M.; Legare, B.; McCormack, B.; dos Santos, P.P.G.M.; Solazzo, D. Absolute Localization of Targets Using a Phase-Measuring Sidescan Sonar in Very Shallow Waters. *Remote Sensing* **2023**, *15*. doi:10.3390/rs15061626.
3. Li, J.; Chen, L.; Shen, J.; Xiao, X.; Liu, X.; Sun, X.; Wang, X.; Li, D. Improved Neural Network with Spatial Pyramid Pooling and Online Datasets Preprocessing for Underwater Target Detection Based on Side Scan Sonar Imagery. *Remote Sensing* **2023**, *15*. doi:10.3390/rs15020440.
4. Xi, J.; Ye, X.; Li, C. Sonar Image Target Detection Based on Style Transfer Learning and Random Shape of Noise under Zero Shot Target. *Remote Sensing* **2022**, *14*. doi:10.3390/rs14246260.
5. Du, X.; Sun, Y.; Song, Y.; Sun, H.; Yang, L. A Comparative Study of Different CNN Models and Transfer Learning Effect for Underwater Object Classification in Side-Scan Sonar Images. *Remote Sensing* **2023**, *15*. doi:10.3390/rs15030593.
6. Meng, J.; Yan, J.; Zhao, J. Bubble Plume Target Detection Method of Multibeam Water Column Images Based on Bags of Visual Word Features. *Remote Sensing* **2022**, *14*. doi:10.3390/rs14143296.
7. Fernandes, J.d.C.V.; de Moura Junior, N.N.; de Seixas, J.M. Deep Learning Models for Passive Sonar Signal Classification of Military Data. *Remote Sensing* **2022**, *14*. doi:10.3390/rs14112648.
8. Wang, Z.; Guo, J.; Zeng, L.; Zhang, C.; Wang, B. MLFFNet: Multilevel Feature Fusion Network for Object Detection in Sonar Images. *IEEE Transactions on Geoscience and Remote Sensing* **2022**, *60*, 1–19. doi:10.1109/TGRS.2022.3214748.
9. Zhang, P.; Tang, J.; Zhong, H.; Wu, H.; Li, H.; Fan, Y. Orientation Estimation of Rotated Sonar Image Targets via the Wavelet Subimage Energy Ratio. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2022**, *15*, 9020–9032. doi:10.1109/JSTARS.2022.3215068.
10. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2017**, *39*, 1137–1149. doi:10.1109/TPAMI.2016.2577031.
11. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
12. Shen, F.; Xie, Y.; Zhu, J.; Zhu, X.; Zeng, H. Git: Graph interactive transformer for vehicle re-identification. *IEEE Transactions on Image Processing* **2023**.
13. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving Into High Quality Object Detection. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 6154–6162. doi:10.1109/CVPR.2018.00644.
14. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. Computer Vision – ECCV 2016; Leibe, B.; Matas, J.; Sebe, N.; Welling, M., Eds.; Springer International Publishing: Cham, 2016; pp. 21–37.
15. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement, 2018, [arXiv:cs.CV/1804.02767].
16. Jocher, G.; Chaurasia, A.; Stoken, A.; Borovec, J.; Kwon, Y.; Michael, K.; Fang, J.; Yifu, Z.; Wong, C.; Montes, D.; others. ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation. *Zenodo* **2022**.
17. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7464–7475.
18. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO Series in 2021, 2021, [arXiv:cs.CV/2107.08430].
19. Shen, F.; Wang, Z.; Wang, Z.; Fu, X.; Chen, J.; Du, X.; Tang, J. A Competitive Method for Dog Nose-print Re-identification. *arXiv preprint arXiv:2205.15934* **2022**.
20. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
21. Yuanzi, L.; Xiufen, Y.; Weizheng, Z. TransYOLO: High-Performance Object Detector for Forward Looking Sonar Images. *IEEE Signal Processing Letters* **2022**, *29*, 2098–2102. doi:10.1109/LSP.2022.3210839.
22. Shen, F.; Zhu, J.; Zhu, X.; Huang, J.; Zeng, H.; Lei, Z.; Cai, C. An Efficient Multiresolution Network for Vehicle Reidentification. *IEEE Internet of Things Journal* **2021**, *9*, 9049–9059.
23. Chen, B.; Yang, Z.; Yang, Z. An algorithm for low-rank matrix factorization and its applications. *Neurocomputing* **2018**, *275*, 1012–1020. doi:https://doi.org/10.1016/j.neucom.2017.09.052.

24. Sun, Y.; Zheng, H.; Zhang, G.; Ren, J.; Xu, H.; Xu, C. DP-ViT: A Dual-Path Vision Transformer for Real-Time Sonar Target Detection. *Remote Sensing* **2022**, *14*. doi:10.3390/rs14225807.

25. Zhou, T.; Si, J.; Wang, L.; Xu, C.; Yu, X. Automatic Detection of Underwater Small Targets Using Forward-Looking Sonar Images. *IEEE Transactions on Geoscience and Remote Sensing* **2022**, *60*, 1–12. doi:10.1109/TGRS.2022.3181417.

26. Wang, H.; Wu, X.; Huang, Z.; Xing, E.P. High-Frequency Component Helps Explain the Generalization of Convolutional Neural Networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

27. Shen, F.; Shu, X.; Du, X.; Tang, J. Pedestrian-specific Bipartite-aware Similarity Learning for Text-based Person Retrieval. Proceedings of the 31th ACM International Conference on Multimedia, 2023.

28. Li, M.; Wei, M.; He, X.; Shen, F. Enhancing Part Features via Contrastive Attention Module for Vehicle Re-identification. 2022 IEEE International Conference on Image Processing (ICIP). IEEE, 2022, pp. 1816–1820.

29. Shen, F.; Du, X.; Zhang, L.; Tang, J. Triplet Contrastive Learning for Unsupervised Vehicle Re-identification. *arXiv preprint arXiv:2301.09498* **2023**.

30. Shen, F.; Zhu, J.; Zhu, X.; Xie, Y.; Huang, J. Exploring spatial significance via hybrid pyramidal graph network for vehicle re-identification. *IEEE Transactions on Intelligent Transportation Systems* **2021**, *23*, 8793–8804.

31. Xu, K.; Qin, M.; Sun, F.; Wang, Y.; Chen, Y.K.; Ren, F. Learning in the Frequency Domain. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

32. Liu, P.; Zhang, H.; Lian, W.; Zuo, W. Multi-Level Wavelet Convolutional Neural Networks. *IEEE Access* **2019**, *7*, 74973–74985. doi:10.1109/ACCESS.2019.2921451.

33. Qin, Z.; Zhang, P.; Wu, F.; Li, X. FcaNet: Frequency Channel Attention Networks. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 783–792.

34. Shen, F.; Peng, X.; Wang, L.; Zhang, X.; Shu, M.; Wang, Y. HSGM: A Hierarchical Similarity Graph Module for Object Re-identification. 2022 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2022, pp. 1–6.

35. Shen, F.; Wei, M.; Ren, J. HSGNet: Object Re-identification with Hierarchical Similarity Graph Network. *arXiv preprint arXiv:2211.05486* **2022**.

36. Zhu, J.; Yu, S.; Gao, L.; Han, Z.; Tang, Y. Saliency-based diver target detection and localization method. *Mathematical Problems in Engineering* **2020**, *2020*, 1–14.

37. Wang, Z.; Zhang, S.; Zhang, C.; Wang, B. RPFNet: Recurrent Pyramid Frequency Feature Fusion Network for Instance Segmentation in Side-Scan Sonar Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2023**, pp. 1–17. doi:10.1109/JSTARS.2023.3266383.

38. Shen, F.; Lin, L.; Wei, M.; Liu, J.; Zhu, J.; Zeng, H.; Cai, C.; Zheng, L. A large benchmark for fabric image retrieval. 2019 IEEE 4th International Conference on Image, Vision and Computing (ICIVC). IEEE, 2019, pp. 247–251.

39. Everingham, M.; Eslami, S.M.A.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision* **2015**, *111*, 98–136.

40. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. Computer Vision – ECCV 2014; Fleet, D.; Pajdla, T.; Schiele, B.; Tuytelaars, T., Eds.; Springer International Publishing: Cham, 2014; pp. 740–755.

41. Yang, K.; Qinami, K.; Fei-Fei, L.; Deng, J.; Russakovsky, O. Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy. Conference on Fairness, Accountability, and Transparency, 2020. doi:10.1145/3351095.3375709.

42. Zhang, P.; Tang, J.; Zhong, H.; Ning, M.; Liu, D.; Wu, K. Self-Trained Target Detection of Radar and Sonar Images Using Automatic Deep Learning. *IEEE Transactions on Geoscience and Remote Sensing* **2022**, *60*, 1–14. doi:10.1109/TGRS.2021.3096011.

43. Huo, G.; Wu, Z.; Li, J. Underwater Object Classification in Sidescan Sonar Images Using Deep Transfer Learning and Semisynthetic Training Data. *IEEE Access* **2020**, *8*, 47407–47418. doi:10.1109/ACCESS.2020.2978880.

44. Wu, H.; Shen, F.; Zhu, J.; Zeng, H.; Zhu, X.; Lei, Z. A sample-proxy dual triplet loss function for object re-identification. *IET Image Processing* **2022**, *16*, 3781–3789.

45. Xu, R.; Shen, F.; Wu, H.; Zhu, J.; Zeng, H. Dual modal meta metric learning for attribute-image person re-identification. 2021 IEEE International Conference on Networking, Sensing and Control (ICNSC). IEEE, 2021, Vol. 1, pp. 1–6.
46. Xie, Y.; Shen, F.; Zhu, J.; Zeng, H. Viewpoint robust knowledge distillation for accelerating vehicle re-identification. *EURASIP Journal on Advances in Signal Processing* **2021**, *2021*, 1–13.
47. Wang, J.; Feng, C.; Wang, L.; Li, G.; He, B. Detection of Weak and Small Targets in Forward-Looking Sonar Image Using Multi-Branch Shuttle Neural Network. *IEEE Sensors Journal* **2022**, *22*, 6772–6783. doi:10.1109/JSEN.2022.3147234.
48. Li, C.; Ye, X.; Xi, J.; Jia, Y. A Texture Feature Removal Network for Sonar Image Classification and Detection. *Remote Sensing* **2023**, *15*. doi:10.3390/rs15030616.
49. Cheng, Z.; Huo, G.; Li, H. A Multi-Domain Collaborative Transfer Learning Method with Multi-Scale Repeated Attention Mechanism for Underwater Side-Scan Sonar Image Classification. *Remote Sensing* **2022**, *14*. doi:10.3390/rs14020355.
50. Gioux, S.; Mazhar, A.; Cuccia, D.J. Spatial frequency domain imaging in 2019: principles, applications, and perspectives. *Journal of biomedical optics* **2019**, *24*, 071613–071613.
51. Khayam, S.A. The discrete cosine transform (DCT): theory and application. *Michigan State University* **2003**, *114*, 31.
52. Briggs, W.L.; Henson, V.E. *The DFT: an owner's manual for the discrete Fourier transform*; SIAM, 1995.
53. Shensa, M.J.; others. The discrete wavelet transform: wedding the a trous and Mallat algorithms. *IEEE Transactions on signal processing* **1992**, *40*, 2464–2482.
54. Brigham, E.O. *The fast Fourier transform and its applications*; Prentice-Hall, Inc., 1988.
55. Qiao, C.; Shen, F.; Wang, X.; Wang, R.; Cao, F.; Zhao, S.; Li, C. A Novel Multi-Frequency Coordinated Module for SAR Ship Detection. 2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, 2022, pp. 804–811.
56. Wang, Z.; Zhang, S.; Huang, W.; Guo, J.; Zeng, L. Sonar Image Target Detection Based on Adaptive Global Feature Enhancement Network. *IEEE Sensors Journal* **2022**, *22*, 1509–1530. doi:10.1109/JSEN.2021.3131645.
57. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. Proceedings of the European Conference on Computer Vision (ECCV), 2018.
58. Sun, P.; Zhang, R.; Jiang, Y.; Kong, T.; Xu, C.; Zhan, W.; Tomizuka, M.; Li, L.; Yuan, Z.; Wang, C.; Luo, P. Sparse R-CNN: End-to-End Object Detection with Learnable Proposals. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14449–14458. doi:10.1109/CVPR46437.2021.01422.
59. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2999–3007. doi:10.1109/ICCV.2017.324.
60. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable {DETR}: Deformable Transformers for End-to-End Object Detection. International Conference on Learning Representations, 2021.
61. Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; Zhang, L. DAB-DETR: Dynamic Anchor Boxes are Better Queries for DETR. International Conference on Learning Representations, 2022.
62. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.; Shum, H.Y. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. The Eleventh International Conference on Learning Representations, 2023.

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.