OS-SHIP-1K: A CYCLEGAN-BASED OPTICAL-SAR MULTIMODAL SHIP DETECTION DATASET

Anonymous authors
Paper under double-blind review

ABSTRACT

Over the past decade, multimodal remote sensing image fusion techniques have developed rapidly. By integrating multi-source remote sensing data, it is possible to obtain more comprehensive and accurate observational information while compensating for the limitations of single-modality imagery. However, in the field of optical–synthetic aperture radar (SAR) multimodal ship detection, challenges in achieving spatiotemporal consistency during data acquisition, coupled with constraints related to data privacy and national defense security, have resulted in a scarcity of publicly available datasets, severely hindering technological advances. To address this issue, this paper introduces the first publicly available dataset for optical–SAR multimodal ship detection, termed OS-Ship-1K, based on CycleGAN. The dataset comprises 1,000 pairs of aligned optical and SAR ship images, annotated with two categories: *Ship* and *Ships*. OS-Ship-1K covers both inshore and offshore scenarios while meeting the detection requirements for both sparse and dense targets. Furthermore, we conduct a comprehensive evaluation of 14 single-modal detectors and 6 multimodal fusion detectors on OS-Ship-1K to establish baseline standards. We hope that the release of the OS-Ship-1K dataset will attract broader attention and engagement from the research community, thereby driving new breakthroughs in optical–SAR multimodal ship detection. The dataset will be released after acceptance.

1 INTRODUCTION

In recent years, the global space-air-ground integrated multi-sensor observation technology has undergone comprehensive development (Xu et al., 2023). Particularly, the widespread application of satellites, aircraft, and unmanned aerial vehicles has significantly enhanced the acquisition capability of multi-source remote sensing images. As a key technique in multi-source remote sensing image processing, multimodal remote sensing image fusion has consequently attracted extensive attention (Yuan & Wei, 2024).

Multimodal remote sensing image fusion (Liu et al., 2025) refers to the integration of remote sensing data from different sensors or imaging modalities to generate higher-quality and more informative image data. Compared to a single sensor, the joint processing of multi-source data can provide more reliable, comprehensive, and accurate observation results. For instance, optical and SAR images exhibit significant complementarity in multimodal fusion. Optical images offer high resolution and rich multispectral information (Jiang et al., 2024), making them easy for human observation, but they are greatly affected by weather and lighting conditions. In contrast, SAR images, obtained through active microwave imaging, provide all-weather and all-time advantages (Sun et al., 2025), unaffected by clouds, fog, or light, yet typically suffer from lower resolution and higher noise levels, requiring specialized interpretation (Zhou et al., 2025). The fusion of optical and SAR modalities can both compensate for the limitations of single-modality and integrate multimodal information, making it widely applied in fields such as disaster monitoring, national defense security, environmental management, agricultural production, and urban planning (Zhou et al., 2024) (Wang et al., 2023).

Ship object detection has been widely applied in maritime safety, shipping management, and maritime rescue (Wang et al., 2025) (Zhang et al., 2025), holding significant research importance and strategic value in military and economic development aspects. It serves as one of the key applications of multimodal remote sensing image fusion technology. However, optical-SAR multimodal

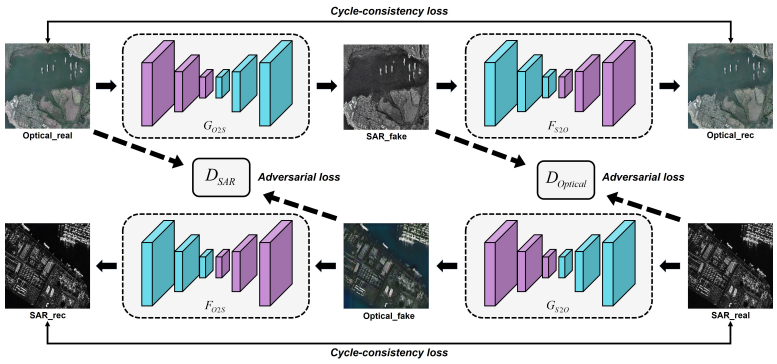

Figure 1: The basic principle for constructing the optical-SAR multimodal ship detection dataset. Here, G and F denote generators, with subscripts indicating the direction of image translation. D_{SAR} and $\mathrm{D}_{Optical}$ represent the discriminators.

ship detection research is constrained by challenges including difficulties in data acquisition and registration, high and specialized annotation costs, as well as restrictions on data privacy and national defense security. Consequently, there are no publicly available optical-SAR multimodal ship detection datasets currently, and related studies remain primarily limited to single-modality approaches (Li et al., 2024a), which have largely hindered the development of multimodal fusion technology in this field.

Recent studies (Zhao et al., 2024a) (Guo et al., 2024) (Yang et al., 2022) have shown that unpaired image-to-image (I2I) translation networks can achieve cross-modal mappings without requiring paired training data. This clever approach circumvents the challenge of acquiring paired data, providing a highly promising solution for constructing an optical-SAR multimodal ship detection dataset. To promote the development of this field, this paper proposes constructing an optical-SAR multimodal ship detection dataset based on unpaired I2I translation networks, aiming to fill the data gap and provide robust data support for related research. Figure 1 illustrates the basic principle of constructing the optical-SAR multimodal ship detection dataset based on CycleGAN. The main contributions of this study are as follows:

- This paper systematically evaluates the performance of mainstream unpaired I2I translation networks in the optical-SAR ship image translation task. Based on CycleGAN, we construct the first publicly available dataset for optical-SAR multimodal ship detection, named OS-Ship-1K, effectively filling a data gap in this field.
- Using the OS-Ship-1K dataset, this paper conducts a comprehensive analysis of 14 single-modal and 6 multimodal fusion detectors, and provides reliable baseline results to facilitate and support future research.

2 Related Work

2.1 Paired Multimodal Remote Sensing Image Datasets

A multimodal remote sensing image paired dataset refers to a paired collection consisting of two or more data from different modalities, typically composed of images acquired from different sensors, where each pair covers the same geographical area and is aligned in space and time as closely as possible (Jia et al., 2021). For example, the SEN1-2 dataset (Schmitt et al., 2018) contains 282,384 pairs of 256×256 pixel images, including SAR imagery from the Sentinel-1 satellite and optical imagery from the Sentinel-2. The QXS-SAROPT dataset (Huang et al., 2021) comprises 20,000 pairs of high-resolution optical and SAR images, with SAR images collected from the Gaofen-3 satellite and optical images from Google Earth, covering the port cities of San Diego in the United States, and Shanghai and Qingdao in China. The DroneVehicle dataset (Sun et al., 2022b) provides 28,439 pairs of visible and infrared drone-captured images for vehicle detection. The VEDAI dataset (Razakarivony & Jurie, 2016) publicly releases 1,210 pairs of visible-light and infrared im-

ages, specifically for vehicle detection in aerial remote sensing. By integrating the complementary characteristics of different modal data, these datasets support applications in remote sensing such as image registration, data fusion, object detection, and change detection (Xia et al., 2018). However, in the ship detection field, datasets like HRSC2016 (Liu et al., 2017), FAIR1M (Sun et al., 2022a), ISDD (Han et al., 2022), HRSID (Wei et al., 2020), SSDD (Zhang et al., 2021a), and SARDet-100K (Li et al., 2024b) are all limited to a single modality. Currently, there is no publicly available multi-modal paired ship detection dataset, which severely restricts the development of multimodal fusion technology in this domain.

2.2 Unpaired I2I Translation Networks

I2I translation is a critical task in computer vision and machine learning, aiming to map an image from one domain to another while preserving its original semantic content and key features. According to the supervision method of the training data, I2I translation networks can be divided into paired and unpaired methods. Paired I2I translation networks rely on paired training samples, where each input image X_i corresponds to a target image Y_i, and a typical representative method is Pix2Pix (Isola et al., 2017). This model is built on the generative adversarial network (GAN) framework and employs a U-Net architecture for the generator and a PatchGAN discriminator. The generator is responsible for generating the target image from the source image, while the discriminator evaluates the authenticity of the generated image. This method can achieve high-quality generation results, but requires significant time and cost to acquire paired data. In contrast, unpaired I2I translation methods do not require strict paired training data. They only need image collections from two domains, X and Y, to learn the mapping between them. CycleGAN (Zhu et al., 2017), as a pioneering work in unpaired I2I translation networks, innovatively introduced cycle consistency loss to ensure that images remain consistent during the transformation process from the source domain to the target domain and back to the source. This mechanism effectively constrains the model to learn reasonable mappings between the source and target domains, eliminating the reliance on paired data, and thus has been widely applied. Given the current lack of publicly available paired optical-SAR ship target images, this paper focuses on researching unpaired I2I translation methods.

3 OS-Ship-1K Dataset

This section first describes the training image sources, experimental environment, implementation details, and evaluation metrics for unpaired I2I translation networks. Next, it provides an in-depth analysis of the quantitative and visual results of each translation model. Finally, it offers a detailed introduction to the optical-SAR multimodal ship detection dataset based on CycleGAN.

3.1 Unpaired I2I Translation Networks Training Data Collection

To enhance the performance of unpaired I2I translation networks, the construction of the optical-SAR training dataset needs to meet the following key requirements: First, data scale and balance are essential, with the training data including a large and evenly distributed set of optical and SAR images to improve the robustness and stability of the translation model. Second, scene diversity is critical, requiring the training images to encompass a variety of scenes and cover ship targets as extensively as possible to enhance the model's generalization ability. Finally, to improve the efficiency of the translation model in learning cross-domain mappings, it is advisable to select optical and SAR images with minimal feature differences to reduce learning difficulty. Therefore, this paper adheres strictly to these standards and integrates multiple existing optical and SAR ship detection datasets. Figure 2 illustrates the sources and distribution of the optical-SAR training dataset (Li et al., 2020) (Zhang et al., 2019) (Zhang et al., 2021b) (Gallego et al., 2018) (Zhipeng, 2023) (Zhang et al., 2020b) (Gallego et al., 2018) (Wang et al., 2019) (Lei et al., 2021) (Li et al., 2024b), with detailed information provided in Appendix A.

3.2 Experimental Setup

Experimental environment and training details. All experiments were conducted on a single 4090 GPU, with PyTorch version 2.1.1, CUDA 11.8, and Ubuntu 18.04 as the server operating system. In the experiments, the model input image size was set to 640×640 and randomly cropped to 512×512,

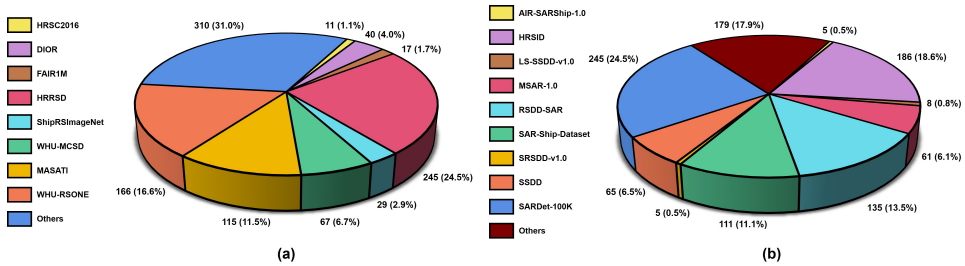

Figure 2: Composition of the Optical-SAR training dataset. (a) Optical image dataset. (b) SAR image dataset. Note: *Others* refers to images generated through copy-paste operations.

with a batch size of 1. Considering the sensitivity of translation models to training parameters, we strictly followed the official default configurations unless otherwise specified.

Image quality evaluation metrics. We adopt four mainstream image quality evaluation metrics to assess the image translation performance: structural similarity index metric (SSIM), peak signal-to-noise ratio (PSNR), Fréchet inception distance (FID), and learned perceptual image patch similarity (LPIPS). SSIM evaluates the similarity between two images based on luminance, contrast, and structural information, aligning well with human visual perception. PSNR quantifies image distortion by calculating the mean squared error between two images but tends to overlook structural information. FID primarily focuses on the distribution similarity between generated and real images, effectively reflecting the overall performance of the model. LPIPS is a perceptual similarity metric used to measure the difference between two images from a human vision perspective. The detailed calculation formulas are provided in Appendix B.1.

3.3 Comparison of Unpaired I2I Translation Networks

Quantitative comparison. Currently, there is no definitive conclusion on the optimal direction for optical-SAR image translation. Therefore, this paper systematically evaluates the performance of 7 mainstream unpaired I2I translation networks—CycleGAN (Zhu et al., 2017), MUNIT (Huang et al., 2018), GcGAN (Fu et al., 2019), CUT (Park et al., 2020), NICE-GAN (Chen et al., 2020), QS-Attn (Hu et al., 2022), and UNSB (Kim et al., 2023), in the tasks of optical-to-SAR (O2S) image translation and SAR-to-optical (S2O) image translation. As shown in Table 1, the translation from O2S images generally outperforms that in the reverse direction. This is particularly evident in CUT and UNSB, where all metrics in the O2S translation task surpass those in the reverse conversion task. Among all networks, CycleGAN performs the best. We attribute this to its strong cycle consistency loss, which provides an effective self-supervised signal, enabling the model to preserve geometric structural information of images during translation. This characteristic is crucial for remote sensing tasks that require strict preservation of ground object structures. CUT, relying on contrastive learning, excels in SSIM, FID, and LPIPS metrics for the O2S translation task, though its PSNR is relatively low. MUNIT and UNSB achieve translation by decoupling image content and style; although this approach is effective in certain style transfer tasks, it may lead to performance degradation in O2S translation tasks requiring strict geometric structure preservation due to information loss. Additionally, we speculate that the design principles of GcGAN, NICE-GAN, and QS-Attn may not align with the specific requirements of the O2S image translation task, resulting in suboptimal experimental outcomes.

Method	Optical-to-SAR				SAR-to-Optical			
	SSIM↑	PSNR↑	FID↓	LPIPS↓	SSIM↑	PSNR↑	FID↓	LPIPS↓
CycleGAN (Zhu et al., 2017)	**0.5336**	**19.52**	**154.8**	0.4846	**0.4536**	**19.96**	**155.3**	**0.4762**
MUNIT (Huang et al., 2018)	0.2068	14.91	262.5	0.6571	0.2258	14.68	256.4	0.6545
GcGAN (Fu et al., 2019)	0.1816	13.79	228.7	0.6232	0.4045	16.96	271.9	0.5761
CUT (Park et al., 2020)	0.5143	15.87	186.1	**0.4222**	0.3772	15.75	257.9	0.5605
NICE-GAN (Chen et al., 2020)	0.4110	16.76	217.2	0.5645	0.4272	16.56	213.3	0.5262
QS-Attn (Hu et al., 2022)	0.4216	15.00	193.0	0.4487	0.3789	16.53	276.7	0.5637
UNSB (Kim et al., 2023)	0.3894	15.45	230.7	0.5643	0.3689	14.66	251.8	0.5830

Table 1: Quantitative comparison of unpaired I2I translation networks. Here, bold indicates the best, underline indicates the second best. For each model, the better-performing translation direction is marked in red, while the worse-performing is marked in blue.

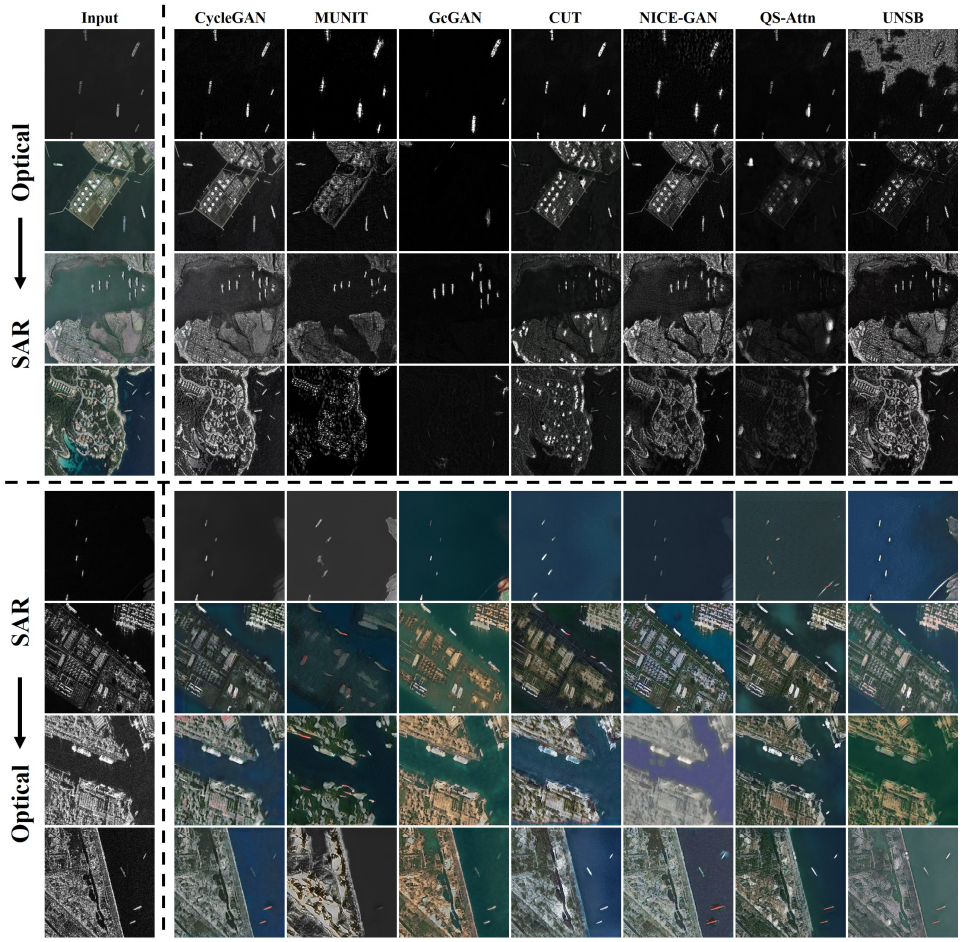

Figure 3: Comparison of generation results from unpaired I2I translation networks. Bounded by vertical dashed lines, the first column represents the model input, while the remaining columns represent the outputs of the unpaired I2I translation networks. Bounded by horizontal dashed lines, the upper half displays the O2S image translation results, while the lower half displays the S2O image translation results.

Qualitative visual comparison. This study aims to construct a paired optical-SAR multimodal ship detection dataset using unpaired I2I translation networks. Therefore, the translation networks need to preserve the geometric information of the images while restoring the real modalities as much as possible. From the overall translation results shown in Figure 3, it is evident that the quality of S2O image translation is generally lower than that of O2S translation, with the former exhibiting more pronounced image blurring and content distortion. This is consistent with the data analysis results in Table 1, indicating that S2O translation is more challenging. Analyzing the translation from O2S images, CycleGAN performs the best, clearly retaining the outlines of ships and coastlines while maintaining sharp texture details. Although MUNIT preserves image geometric information to some extent, it suffers from ship deformation (third row) and target loss (fourth row) issues. GcGAN and QS-Attn exhibit poor translation performance, with generated images appearing blurry and suffering from severe blank areas, resulting in substantial loss of original image information. CUT shows inadequate learning of the brightness and darkness relationships in SAR images, leading to poor realism in the translation images. Additionally, NICE-GAN and UNSB achieve relatively good translation results but lack stability, for example, failing to accurately reconstruct the sea surface background (first row). On the other hand, in the S2O image translation task, CycleGAN, CUT, and QS-Attn perform notably well, effectively restoring optical images with natural colors and clear details, preserving both ship and water region features. In contrast, MUNIT and GcGAN show poor translation performance, characterized by distorted image content, unnatural textures, and insuffi-

cient overall realism. Although NICE-GAN and UNSB produce optical images with rich colors, they struggle to accurately retain details, such as blurred ship outlines and overexposure issues.

Based on the above analysis, CycleGAN demonstrates superior performance among all models, with robust and realistic image translation results, effectively meeting the requirements for constructing an optical-SAR multimodal ship detection dataset. In comparison, the performance of other models varies widely, none reaching the level of CycleGAN. Therefore, this paper ultimately selects CycleGAN for constructing the optical-SAR multimodal ship detection dataset.

3.4 Analysis of the OS-Ship-1K Dataset

Based on CycleGAN, this paper converts real optical images into corresponding SAR images and selects 1,000 high-quality optical-SAR pairs to construct the OS-Ship-1K multimodal ship detection dataset. In OS-Ship-1K, all images are resized to 512×512 resolution and annotated with two target categories: *Ship* (single ship) and *Ships* (clustered multiple ships). As shown in Figure 4, the scale distributions of the *Ship* and *Ships* are mainly concentrated within 100×100 pixels. Furthermore, according to the target size definitions in the COCO dataset (Lin et al., 2014), the *Ship* category includes 64.0% small targets (2,906 instances), 35.4% medium targets (1,610 instances), and 0.6% large targets (26 instances), totaling 4,542 instances. For the *Ships* category, small targets account for 45.8% (217 instances), medium targets for 50.4% (239 instances), and large targets for 3.6% (18 instances), with a total of 474 instances. This indicates that OS-Ship-1K is a small target detection dataset. Finally, we divide the OS-Ship-1K dataset in a 6:2:2 ratio into 600 training images, 200 validation images, and 200 test images. For more details, refer to Appendix C.

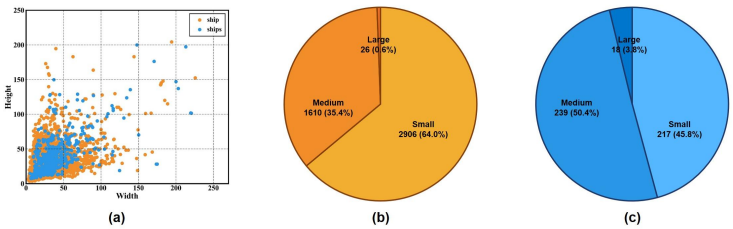

Figure 4: Statistical analysis of the OS-Ship-1K dataset. (a) shows the scale distribution of *Ship* and *Ships*. (b) and (c) respectively show the proportions and quantities of small, medium, and large targets for the *Ship* and *Ships*.

In summary, OS-Ship-1K is a novel multimodal ship detection dataset that effectively fills the data gap in the optical-SAR multimodal ship detection field, providing robust data support for the development of related technologies. We believe that OS-Ship-1K will not only drive progress in multimodal ship detection technology but also play a significant role in other remote sensing image processing tasks, such as image matching.

4 Baseline Experiments

This section first introduces the training details and object detection evaluation metrics, and then provides a comprehensive assessment of the performance of single-modal and multimodal fusion detection methods on the OS-Ship-1K dataset.

4.1 Implementation Details

Training details. We conducted the experimental analysis under the same software and hardware environment as described in Section 3.2. In the experiments, we uniformly set the input image size to 512×512, the batch size to 8, and trained the models using pretrained weights. For single-modal detectors, we trained Faster R-CNN, Cascade R-CNN, Dynamic R-CNN, FoveaBox, YOLOX, TOOD, DETR, DINO, and Deformable-DETR on the MMdetection framework. These detectors all employed ResNet50 as the backbone network, SGD as the optimizer, were trained for 12 epochs, and had the learning rate gradually reduced from an initial value of 0.02 to 0.0002 following the MultiStepLR strategy. Additionally, YOLOv3, YOLOv5, YOLOv8, YOLOv11, and RT-DETR were trained on the Ultralytics framework. These detectors were all medium-sized, trained with SGD for 100 epochs, with an initial learning rate of 0.01, and a minimum learning rate of 0.0001.

On the other hand, for multimodal fusion detection networks, we adopted the publicly available code from the respective authors and strictly followed the recommended settings in the papers to ensure the accuracy and reproducibility of the experimental results.

Object detection evaluation metrics. In addition to the common object detection evaluation metrics such as precision, recall, average precision (AP), and mean average precision (mAP), we also introduce AP_S, AP_M, and AP_L to measure the detection accuracy for small, medium, and large objects, respectively. Furthermore, we evaluate the complexity of different detectors using parameters (Params) and floating-point operations (FLOPs). The specific formulas see Appendix B.2.

4.2 Performance of Single-Modal and Multimodal Fusion Detectors on the OS-Ship-1K Dataset

This section comprehensively evaluates the performance of single-modal and multimodal fusion detection methods on the OS-Ship-1K dataset from three perspectives: quantitative results, P-R curves, and qualitative visual analysis. Among the single-modal detectors, 14 representative algorithms are compared, including two-stage methods: Faster R-CNN (Ren et al., 2016), Cascade R-CNN (Cai & Vasconcelos, 2018), and Dynamic R-CNN (Zhang et al., 2020a); one-stage detectors: FoveaBox (Kong et al., 2020), YOLOv3 (Redmon & Farhadi, 2018), YOLOv5 (Jocher, 2020), YOLOv8 (Jocher et al., 2023), YOLOv11 (Khanam & Hussain, 2024), YOLOX (Ge et al., 2021), and TOOD (Feng et al., 2021); and end-to-end methods: DETR (Carion et al., 2020), DINO (Zhang et al., 2022), RT-DETR (Zhao et al., 2024b), and Deformable-DETR (Zhu et al., 2020). In addition, 6 multimodal fusion detectors are selected: SuperYOLO (Zhang et al., 2023), CDC-YOLOFusion (Wang et al., 2024), DEYOLO (Chen et al., 2024), CFT (Qingyun et al., 2021), ICAFusion (Shen et al., 2024), and TFDet (Zhang et al., 2024).

Quantitative results. Table 2 presents the quantitative results of various detectors on the OS-Ship-1K dataset. Overall, detection networks perform significantly better on the optical modality dataset than on the SAR modality. This is primarily due to the fact that optical modality data consists of real images, providing richer and more easily learnable feature information. Additionally, the pretrained weights are more aligned with optical ship detection scenarios, enabling the models to maximize their performance. Consequently, the detectors exhibit superior performance on the optical dataset. Furthermore, whether on the optical or SAR modality, one-stage detectors generally outperform two-stage and end-to-end detectors while offering notable computational efficiency advantages, with the YOLO series standing out. Specifically, on the optical modality dataset, YOLOv5 achieves 73.4% in mAP_{50} and 55.8% in mAP, YOLOv8 reaches 73.8% and 56.4%, and YOLOX achieves 72.4% and 56.0%, all ranking at the top level. On the SAR modality, YOLOv5 achieves 66.9% and 51.6% in mAP_{50} and mAP, respectively, YOLOv8 records 63.7% and 47.7%, and YOLOX reaches 69.2% and 51.5%, also leading the performance. In contrast, while two-stage detectors offer robust detection performance, they fall short of one-stage detectors in both performance and efficiency, with higher parameters and computational complexity. Transformer-based end-to-end detectors theoretically possess strong feature learning capabilities, but their performance is suboptimal due to the limited scale of the OS-Ship-1K dataset. Among multimodal fusion detectors, optical features effectively assist SAR target detection, leading to excellent performance on the SAR modality ship detection. Notably, ICAFusion and TFDet demonstrate outstanding detection performance, with ICAFusion achieving 73.9% in mAP_{50}, surpassing the best-performing single-modal detectors, YOLOv8. However, the performance of other multimodal detectors remains limited. This is likely due to the fact that the multimodal fusion detection networks used in the experiments were primarily designed for optical-infrared modalities, and their network architectures and fusion strategies differ from those required for the optical-SAR modalities, thereby impacting the detection results.

P-R curve analysis. Figure 5 provides a comparison of the P-R curves for various detectors on the OS-Ship-1K dataset, intuitively reflecting the comprehensive detection performance of different methods. From the overall distribution of the P-R curves, it is evident that the curves of each detector are relatively uniformly concentrated within a certain range, indicating robust detection performance. Further observation of the locally magnified area reveals that ICAFusion (in gray) exhibits the best P-R curve, particularly showing a more pronounced advantage on the SAR modality dataset. This suggests that the fusion of optical-SAR modalities can effectively enhance the performance of SAR ship target detection. Additionally, the P-R curves of detectors such as YOLOv5,

Modal	Type	Method	Ship	Ships	mAP_{50}	mAP	AP_S	AP_M	AP_L	Params(M)	FLOPs(G)
Optical	Two-Stage	Faster R-CNN (Ren et al., 2016)	69.0	32.9	68.3	50.9	42.2	62.2	51.7	41.4	63.2
		Cascade R-CNN (Cai & Vasconcelos, 2018)	70.5	35.0	70.3	52.8	44.7	63.9	38.5	69.2	91.0
		Dynamic R-CNN (Zhang et al., 2020a)	69.7	35.7	69.0	52.7	43.3	64.0	54.1	41.4	63.2
	One-Stage	YOLOv3 (Redmon & Farhadi, 2018)	70.5	33.3	67.9	51.9	48.7	59.9	22.9	103.7	282.2
		YOLOv5 (Jocher, 2020)	72.1	39.6	73.4	55.8	48.0	65.7	51.6	25.0	64.0
		YOLOv8 (Jocher et al., 2023)	71.9	40.8	73.8	**56.4**	**50.1**	65.9	34.2	25.8	78.7
		YOLOv11 (Khanam & Hussain, 2024)	71.7	35.8	69.5	53.7	47.0	62.9	25.8	20.0	67.7
		YOLOX (Ge et al., 2021)	72.0	39.9	72.4	56.0	49.1	65.8	44.8	54.1	49.7
		TOOD (Feng et al., 2021)	69.5	37.1	71.1	53.3	42.3	**66.2**	**55.9**	32.0	50.5
		FoveaBox (Kong et al., 2020)	66.5	33.5	66.9	50.0	37.1	65.0	48.1	36.2	51.7
	End-to-End	DETR (Carion et al., 2020)	58.0	17.5	59.2	37.8	29.0	49.7	36.8	41.6	72.8
		DINO (Zhang et al., 2022)	72.8	36.5	70.2	54.7	47.9	64.9	**55.9**	47.5	80.7
		RT-DETR (Zhao et al., 2024b)	68.7	27.2	61.7	48.0	43.6	56.1	21.0	41.9	125.6
		Deformable-DETR (Zhu et al., 2020)	49.8	20.8	57.2	35.3	25.4	49.3	41.6	40.1	51.8
SAR	Two-Stage	Faster R-CNN (Ren et al., 2016)	67.1	23.2	62.3	45.1	39.5	54.6	45.6	41.4	63.2
		Cascade R-CNN (Cai & Vasconcelos, 2018)	68.2	27.4	64.4	47.8	38.8	59.2	55.7	69.2	91.0
		Dynamic R-CNN (Zhang et al., 2020a)	67.3	28.1	64.9	47.7	40.3	58.2	43.2	41.4	63.2
	One-Stage	YOLOv3 (Redmon & Farhadi, 2018)	68.5	27.8	63.8	48.2	39.1	60.3	12.2	103.7	282.2
		YOLOv5 (Jocher, 2020)	70.7	32.6	66.9	51.6	47.2	60.2	35.8	25.0	64.0
		YOLOv8 (Jocher et al., 2023)	70.2	25.3	63.7	47.7	42.6	56.8	48.9	25.8	78.7
		YOLOv11 (Khanam & Hussain, 2024)	70.7	25.7	62.9	48.2	40.4	58.9	30.4	20.0	67.7
		YOLOX (Ge et al., 2021)	70.4	32.7	69.2	51.5	45.4	61.3	41.1	54.1	49.7
		TOOD (Feng et al., 2021)	67.2	31.7	65.6	49.4	40.9	61.2	42.7	32.0	50.5
		FoveaBox (Kong et al., 2020)	64.5	28.1	63.4	46.3	34.3	59.8	41.0	36.2	51.7
	End-to-End	DETR (Carion et al., 2020)	53.9	14.7	55.7	34.3	25.9	45.5	21.0	41.6	72.8
		DINO (Zhang et al., 2022)	71.4	25.9	63.0	48.6	38.1	61.7	47.0	47.5	80.7
		RT-DETR (Zhao et al., 2024b)	69.1	21.9	59.9	45.5	38.2	55.4	29.0	41.9	125.6
		Deformable-DETR (Zhu et al., 2020)	46.6	16.6	54.6	31.6	23.3	45.7	30.8	40.1	51.8
Fusion	One-Stage	SuperYOLO (Zhang et al., 2023)	67.3	26.6	63.5	46.9	41.7	55.3	12.9	**4.8**	**18.0**
		CDC-YOLOFusion (Wang et al., 2024)	68.5	27.8	69.9	48.1	40.1	58.6	32.4	82.8	-
		DEYOLO (Chen et al., 2024)	70.9	31.1	67.0	51.0	46.2	60.1	36.3	48.8	-
	-	CFT (Qingyun et al., 2021)	68.0	26.5	68.2	47.2	39.4	57.0	25.6	44.5	-
		ICAFusion (Shen et al., 2024)	72.7	38.2	**73.9**	54.9	49.6	65.0	42.2	59.9	-
		TFDet (Zhang et al., 2024)	70.9	38.0	72.4	54.4	48.4	62.7	38.0	36.6	-

Table 2: Performance comparison of detectors on the OS-Ship-1K dataset. The best results are highlighted in bold, and the second-best results are indicated with underlines.

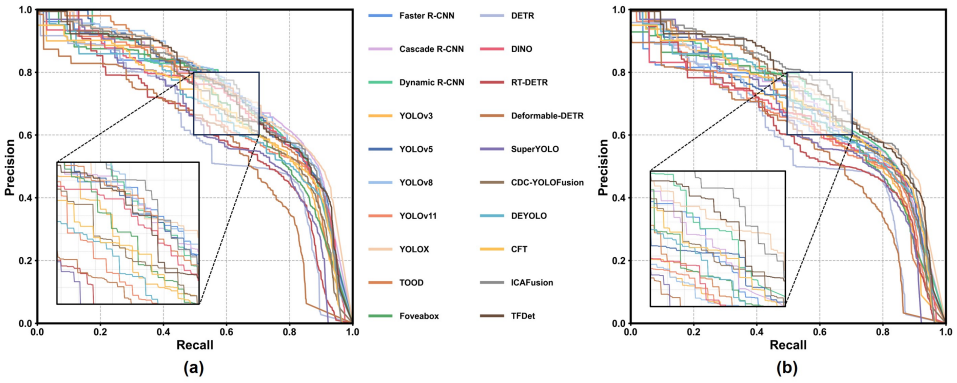

Figure 5: P-R curve comparisons of various detectors on the OS-Ship-1K dataset. (a) Optical modality dataset. (b) SAR modality dataset.

YOLOv8, YOLOX, and TFDet also perform well, aligning with the quantitative analysis results in Table 2, further validating the effectiveness of these methods.

Qualitative visual analysis. Figure 6 presents the visual detection results of different types of detectors on the OS-Ship-1K dataset. Overall, the detection performance in the optical modality outperforms that in the SAR. A deeper analysis reveals that in offshore sparse scenarios, most detectors suffer from misidentifying sea surface clutter as ships, whereas Faster R-CNN stands out by achieving accurate and error-free detection. However, in offshore dense detection scenarios, the detection performance of Faster R-CNN declines noticeably, while other detectors perform well, particularly ICAFusion, which correctly identifies 16 targets without any false positives or misses.

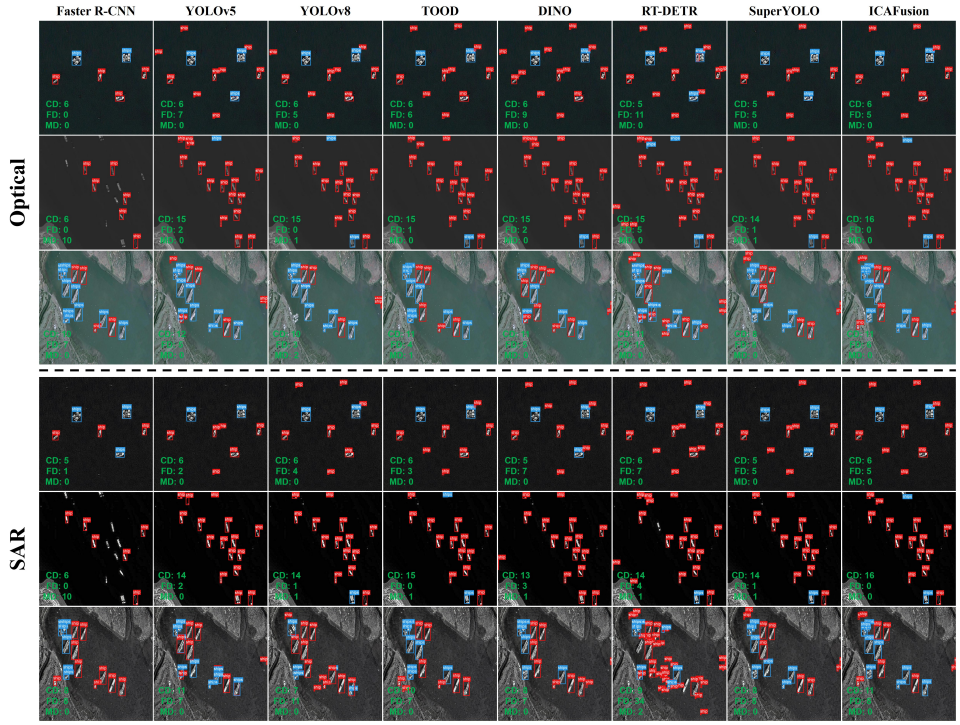

Figure 6: Visual comparison of various detectors on the OS-Ship-1K dataset. (a) Offshore sparse scene. (b) Offshore dense scene. (c) Inshore scene. Divided by a dashed line, the upper part shows the detection results for the optical modality, while the lower part shows the detection results for the SAR modality. Targets identified as the *Ship* are marked with red bounding boxes, targets identified as the *Ships* are marked with blue bounding boxes. The counts of correct detections (CD), false detections (FD), and missed detections (MD) are summarized in the lower-left corner of the image.

This highlights the performance variations among different detectors in specific scenarios, necessitating the selection of appropriate detectors based on actual application contexts. Furthermore, in inshore scenes, the number of false positives is markedly higher than that of missed detections, though all detectors exhibit robust performance. For instance, RT-DETR records 15 false positives in the optical modality but as many as 34 in the SAR modality. This reflects that in complex SAR images, detectors have weaker capabilities to distinguish objects from backgrounds, often misidentifying background clutter as targets. This issue also represents one of the most challenging aspects of the SAR ship object detection.

5 Conclusion

The scarcity of an optical-SAR multimodal ship detection dataset has severely hindered the development of multimodal fusion technology in this field. To address this, this paper constructs the first OS-Ship-1K dataset designed for optical-SAR multimodal ship detection tasks based on CycleGAN. We provide a detailed description of the dataset construction process, including the selection of training data for the unpaired I2I translation network, quantitative and qualitative analyses, and a comprehensive evaluation of the dataset. Additionally, this paper presents experimental benchmarks for 14 single-modal detection methods and 6 multimodal detectors on the OS-Ship-1K dataset, offering reference and support for future research. We hope that the release of the OS-Ship-1K dataset will attract more participation and attention from researchers, further promoting the development of optical-SAR multimodal ship detection technology.

REFERENCES

Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6154–6162, 2018.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Runfa Chen, Wenbing Huang, Binghui Huang, Fuchun Sun, and Bin Fang. Reusing discriminators for encoding: Towards unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8168–8177, 2020.

Yishuo Chen, Boran Wang, Xinyu Guo, Wenbin Zhu, Jiasheng He, Xiaobin Liu, and Jing Yuan. Deyolo: Dual-feature-enhancement yolo for cross-modality object detection. In *International Conference on Pattern Recognition*, pp. 236–252. Springer, 2024.

Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R Scott, and Weilin Huang. Tood: Task-aligned one-stage object detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3490–3499. IEEE Computer Society, 2021.

Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Kun Zhang, and Dacheng Tao. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2427–2436, 2019.

Antonio-Javier Gallego, Antonio Pertusa, and Pablo Gil. Automatic ship classification from optical aerial images with convolutional neural networks. *Remote Sensing*, 10(4):511, 2018.

Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

Zhe Guo, Zhibo Zhang, Qinglin Cai, Jiayi Liu, Yangyu Fan, and Shaohui Mei. Ms-gan: Learn to memorize scene for unpaired sar-to-optical image translation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:11467–11484, 2024.

Yaqi Han, Jingwen Liao, Tianshu Lu, Tian Pu, and Zhenming Peng. Kcpnet: Knowledge-driven context perception networks for ship detection in infrared imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–19, 2022.

Xueqi Hu, Xinyue Zhou, Qiusheng Huang, Zhengyi Shi, Li Sun, and Qingli Li. Qs-attn: Query-selected attention for contrastive learning in i2i translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18291–18300, 2022.

Meiyu Huang, Yao Xu, Lixin Qian, Weili Shi, Yaqin Zhang, Wei Bao, Nan Wang, Xuejiao Liu, and Xueshuang Xiang. The qxs-saropt dataset for deep learning in sar-optical data fusion. *arXiv preprint arXiv:2103.08259*, 2021.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 172–189, 2018.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3496–3504, 2021.

Lingjie Jiang, Baoxi Yuan, Jiawei Du, Boyu Chen, Hanfei Xie, Juan Tian, and Ziqi Yuan. Mffsodnet: Multiscale feature fusion small object detection network for uav aerial images. *IEEE Transactions on Instrumentation and Measurement*, 73:1–14, 2024.

G Jocher. Yolov5 by ultralytics. In *Online. Available: https://github.com/ultralytics/yolov5*, 2020.

G Jocher, A Chaurasia, and J Qiu. Yolov8 by ultralytics. In *Online. Available: https://github.com/ultralytics/ultralytics*, 2023.

Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024.

Beomsu Kim, Gihyun Kwon, Kwanyoung Kim, and Jong Chul Ye. Unpaired image-to-image translation via neural schr\" odinger bridge. *arXiv preprint arXiv:2305.15086*, 2023.

Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, Lei Li, and Jianbo Shi. Foveabox: Beyound anchor-based object detection. *IEEE Transactions on Image Processing*, 29:7389–7398, 2020.

Songlin Lei, Dongdong Lu, Xiaolan Qiu, and Chibiao Ding. Srsdd-v1. 0: A high-resolution sar rotation ship detection dataset. *Remote Sensing*, 13(24):5104, 2021.

Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307, 2020.

Ke Li, Di Wang, Zhangyuan Hu, Wenxuan Zhu, Shaofeng Li, and Quan Wang. Unleashing channel potential: Space-frequency selection convolution for sar object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17323–17332, 2024a.

Yuxuan Li, Xiang Li, Weijie Li, Qibin Hou, Li Liu, Ming-Ming Cheng, and Jian Yang. Sardet-100k: Towards open-source benchmark and toolkit for large-scale sar object detection. *Advances in Neural Information Processing Systems*, 37:128430–128461, 2024b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Kewei Liu, Tao Li, and Dongliang Peng. Aerial image object detection based on rgb-infrared multi-branch progressive fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

Zikun Liu, Liu Yuan, Lubin Weng, and Yiping Yang. A high resolution optical satellite image dataset for ship recognition and some new baselines. In *International conference on pattern recognition applications and methods*, volume 2, pp. 324–331. SciTePress, 2017.

Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European conference on computer vision*, pp. 319–345. Springer, 2020.

Fang Qingyun, Han Dapeng, and Wang Zhaokui. Cross-modality fusion transformer for multispectral object detection. *arXiv preprint arXiv:2111.00273*, 2021.

Sebastien Razakarivony and Frederic Jurie. Vehicle detection in aerial imagery: A small target detection benchmark. *Journal of Visual Communication and Image Representation*, 34:187–203, 2016.

Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

Michael Schmitt, Lloyd Haydn Hughes, and Xiao Xiang Zhu. The sen1-2 dataset for deep learning in sar-optical data fusion. *arXiv preprint arXiv:1807.01569*, 2018.

Jifeng Shen, Yifei Chen, Yue Liu, Xin Zuo, Heng Fan, and Wankou Yang. Icafusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognition*, 145: 109913, 2024.

Xian Sun, Peijin Wang, Zhiyuan Yan, Feng Xu, Ruiping Wang, Wenhui Diao, Jin Chen, Jihao Li, Yingchao Feng, Tao Xu, et al. Fair1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 184:116–130, 2022a.

Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6700–6713, 2022b.

Zhongzhen Sun, Xiangguang Leng, Xianghui Zhang, Zheng Zhou, Boli Xiong, Kefeng Ji, and Gangyao Kuang. Arbitrary-direction sar ship detection method for multi-scale imbalance. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

Chao Wang, Wenxuan Fang, Xiang Li, Jian Yang, and Lei Luo. Msod: A large-scale multi-scene dataset and a novel diagonal-geometry loss for sar object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

Shiyu Wang, Zhanchuan Cai, and Jieyu Yuan. Automatic sar ship detection based on multifeature fusion network in spatial and frequency domains. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–11, 2023.

Yuanyuan Wang, Chao Wang, Hong Zhang, Yingbo Dong, and Sisi Wei. A sar dataset of ship detection for deep learning under complex backgrounds. *remote sensing*, 11(7):765, 2019.

Zian Wang, Xianghui Liao, Jin Yuan, You Yao, and Zhiyong Li. Cdc-yolofusion: Leveraging cross-scale dynamic convolution fusion for visible-infrared object detection. *IEEE Transactions on Intelligent Vehicles*, 2024.

Shunjun Wei, Xiangfeng Zeng, Qizhe Qu, Mou Wang, Hao Su, and Jun Shi. Hrsid: A high-resolution sar images dataset for ship detection and instance segmentation. *Ieee Access*, 8: 120234–120254, 2020.

Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3974–3983, 2018.

Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132, 2023.

Xi Yang, Jingyi Zhao, Ziyu Wei, Nannan Wang, and Xinbo Gao. Sar-to-optical image translation based on improved cgan. *Pattern Recognition*, 121:108208, 2022.

Maoxun Yuan and Xingxing Wei. C^2former: Calibrated and complementary transformer for rgb-infrared object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

Hongkai Zhang, Hong Chang, Bingpeng Ma, Naiyan Wang, and Xilin Chen. Dynamic r-cnn: Towards high quality object detection via dynamic training. In *European conference on computer vision*, pp. 260–275. Springer, 2020a.

Jiaqing Zhang, Jie Lei, Weiying Xie, Zhenman Fang, Yunsong Li, and Qian Du. Superyolo: Super resolution assisted object detection in multimodal remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.

Tianwen Zhang, Xiaoling Zhang, Xiao Ke, Xu Zhan, Jun Shi, Shunjun Wei, Dece Pan, Jianwei Li, Hao Su, Yue Zhou, et al. Ls-ssdd-v1. 0: A deep learning dataset dedicated to small ship detection from large-scale sentinel-1 sar images. *Remote Sensing*, 12(18):2997, 2020b.

Tianwen Zhang, Xiaoling Zhang, Jianwei Li, Xiaowo Xu, Baoyou Wang, Xu Zhan, Yanqin Xu, Xiao Ke, Tianjiao Zeng, Hao Su, et al. Sar ship detection dataset (ssdd): Official release and comprehensive data analysis. *Remote Sensing*, 13(18):3690, 2021a.

Xin Zhang, Xue Yang, Yuxuan Li, Jian Yang, Ming-Ming Cheng, and Xiang Li. Rsar: Restricted state angle resolver and rotated sar benchmark. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7416–7426, 2025.

Xue Zhang, Xiaohan Zhang, Jiangtao Wang, Jiacheng Ying, Zehua Sheng, Heng Yu, Chunguang Li, and Hui-Liang Shen. Tfdet: Target-aware fusion for rgb-t pedestrian detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

Yuanlin Zhang, Yuan Yuan, Yachuang Feng, and Xiaoqiang Lu. Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8):5535–5548, 2019.

Zhengning Zhang, Lin Zhang, Yue Wang, Pengming Feng, and Ran He. Shiprsimagenet: A large-scale fine-grained dataset for ship detection in high-resolution optical remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:8458–8472, 2021b.

Wenbo Zhao, Nana Jiang, Xiaoxin Liao, and Jubo Zhu. Hvt-cgan: Hybrid vision transformer cgan for sar-to-optical image translation. *IEEE Transactions on Geoscience and Remote Sensing*, 2024a.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16965–16974, 2024b.

DONG Zhipeng. Research on object detection in high resolution remote sensing imagery based on convolutional neural networks. *Acta Geodaetica et Cartographica Sinica*, 52(9):1613, 2023.

Jie Zhou, Yongxiang Liu, Bowen Peng, Li Liu, and Xiang Li. Madinet: Mamba diffusion network for sar target detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Zheng Zhou, Lingjun Zhao, Kefeng Ji, and Gangyao Kuang. A domain-adaptive few-shot sar ship detection algorithm driven by the latent similarity between optical and sar images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–18, 2024.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

A Appendix

The specific sources and quantities of the optical-SAR training dataset. Among them, the optical image dataset totals 1,000 images, HRSC2016 accounts for 1.1% (11 images), DIOR for 4.0% (40 images), FAIR1M for 1.7% (17 images), HRRSD for 24.5% (245 images), ShipRSImageNet for 2.9% (29 images), WHU-MCSD for 6.7% (67 images), MASATI for 11.5% (115 images), WHU-RSONE for 16.6% (166 images), and Others for 31.0% (310 images). On the other hand, the SAR image dataset also totals 1,000 images, with specific sources and quantities as follows: AIR-SARShip-1.0 accounts for 0.5% (5 images), HRSID for 18.6% (186 images), LS-SSDD-v1.0 for 0.8% (8 images), MSAR-1.0 for 6.1% (61 images), RSDD-SAR for 13.5% (135 images), SAR-Ship-Dataset for 11.1% (111 images), SRSDD-v1.0 for 0.5% (5 images), SSDD for 6.5% (65 images), SARDet-100K for 24.5% (245 images), and Others for 17.9% (179 images).

B APPENDIX

B.1 IMAGE QUALITY EVALUATION METRICS

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{1}$$

Here, x and y denote the two images. where μ_x, σ_x^2, μ_y, σ_y^2, denote the mean and variance of the features of x and y,respectively, σ_{xy} is the covariance between x and y. $c1$ and $c2$ are two constants used to ensure that the denominator is not zero. A larger SSIM value indicates greater structural similarity between the real and generated images.

$$PSNR(x,y) = 10\log_{10}\left(\frac{MAX^2}{MSE}\right) \tag{2}$$

where MAX denotes the maximum gray value of x and MSE denotes the mean squared error between x and y. A higher PSNR value reflects better quality of the generated image.

$$FID(x,y) = \|\mu_x - \mu_y\|^2 + Tr(C_x + C_y - 2(C_xC_y)^{1/2}) \tag{3}$$

where μ_x, C_x, μ_y, C_y denote the mean and covariance matrix of the features of x and y, respectively. $Tr(\cdot)$ denotes the trace function. The lower the value of FID represents the more similar the distribution of images.

$$LPIPS(x,y) = \sum_j \frac{1}{W_jH_j}\sum_{h,w} w_j \left\| f_{h,w}^j(x) - f_{h,w}^j(y)\right\|_2^2 \tag{4}$$

where j denotes the layer index of the network, W_j and H_j represent the width and height of the feature map at the j-th layer, respectively. w_j is the learnable weight, and $f_{h,w}^j$ denotes the feature vector at the spatial position (h,w) in the j-th layer. A smaller LPIPS value indicates greater perceptual similarity between two images, showing better quality of the generated image.

B.2 OBJECT DETECTION EVALUATION METRICS

Precision and recall are core metrics for evaluating model performance. Precision represents the proportion of true positives among the samples predicted as positive by the model. Recall represents the proportion of all true positive samples correctly identified by the model. Their calculation formulas are as follows.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

Here, true positives (TP) represent the ship targets correctly identified by the model, false positives (FP) denote the background or other objects incorrectly identified as ships, and false negatives (FN) represent the real ship targets that the detector failed to identify.

Subsequently, by calculating precision and recall, the average precision (AP) can be derived. AP represents the average precision for a single category across different confidence thresholds, while the mean average precision (mAP) represents the average precision across all categories under different confidence thresholds, providing a more comprehensive evaluation of model performance. In particular, mAP_{50} refers to the average precision when the intersection over union threshold between predicted and ground-truth boxes is set to 0.5. Their calculation formulas are as follows.

$$AP = \int_0^1 P(R)dR \tag{7}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{8}$$

Here, $P(R)$ represents the area under the P-R curve, with recall as the horizontal axis and precision as the vertical axis. N is the total number of categories, and i denotes the individual categories.

C APPENDIX

Category	Train		Val		Test	
	Optical	SAR	Optical	SAR	Optical	SAR
Ship	2673	2673	924	924	945	945
Ships	269	269	99	99	106	106
All	2942	2942	1023	1023	1051	1051

Table 3: Statistics of instance numbers for each category in the training, validation, and test sets of the OS-Ship-1K dataset.

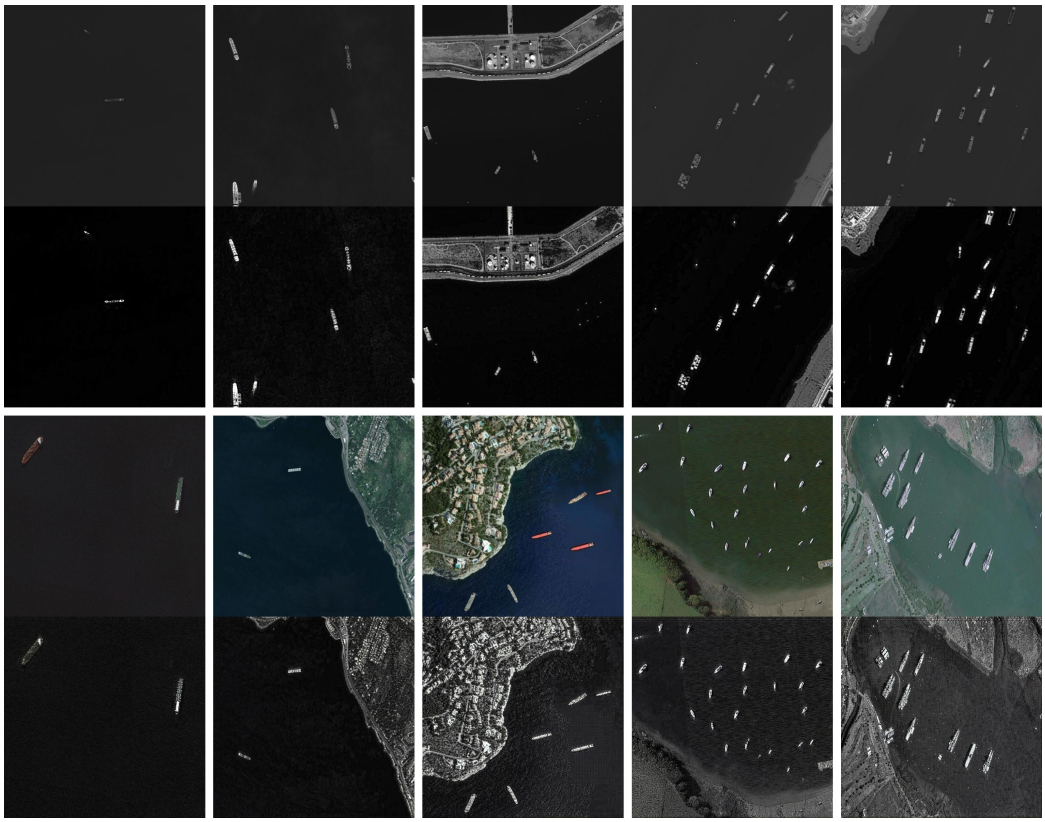

Figure 7: OS-Ship-1K dataset sample images