M4-SAR: A Multi-Resolution, Multi-Polarization, Multi-Scene, Multi-Source Dataset and Benchmark for optical-SAR Object Detection

Chao Wang[1], Wei Lu[2], Xiang Li[3], Jian Yang[1], and Lei Luo[1✉]

[1] Nanjing University of Science and Technology, Nanjing 210094, China
[2] Anhui University, Hefei 230601, China
[3] Nankai University, Tianjin 300350, China
Project page: https://github.com/wchao0601/M4-SAR
wchao0601@163.com, luoleipitt@gmail.com

Abstract. Single-source remote sensing object detection using optical or SAR images struggles in complex environments. Optical images offer rich textural details but are often affected by low-light, cloud-obscured, or low-resolution conditions, reducing the detection performance. SAR images are robust to weather, but suffer from speckle noise and limited semantic expressiveness. Optical and SAR images provide complementary advantages, and fusing them can significantly improve the detection accuracy. However, progress in this field is hindered by the lack of large-scale, standardized datasets. To address these challenges, we propose a new comprehensive dataset for optical-SAR fusion object detection, named **M**ulti-resolution, **M**ulti-polarization, **M**ulti-scene, **M**ulti-source **SAR** dataset (**M4-SAR**). It contains 112,174 instance-level aligned image pairs and nearly one million labeled instances with arbitrary orientations, spanning six key categories. To enable standardized evaluation, we develop a unified benchmarking toolkit that integrates six state-of-the-art multi-source fusion methods. Additionally, we propose E2E-OSDet, a novel end-to-end multi-source fusion detection framework that mitigates cross-domain discrepancies and establishes a robust baseline for future studies. Extensive experiments on M4-SAR demonstrate that fusing optical and SAR data can improve mAP by 5.7% over single-source inputs, with particularly significant gains in complex environments.

Keywords: Scene analysis and understanding · Dataset · Benchmark · Synthetic aperture radar (SAR) · Optical-SAR object detection

1 Introduction

With the rapid development of deep learning techniques, single-source object detection (e.g., optical [10, 13, 19, 24, 31, 32, 41, 44] or synthetic aperture radar (SAR) [9, 36, 51, 54]), and multi-source object detection (optical-SAR [23, 65])

✉ Corresponding Author

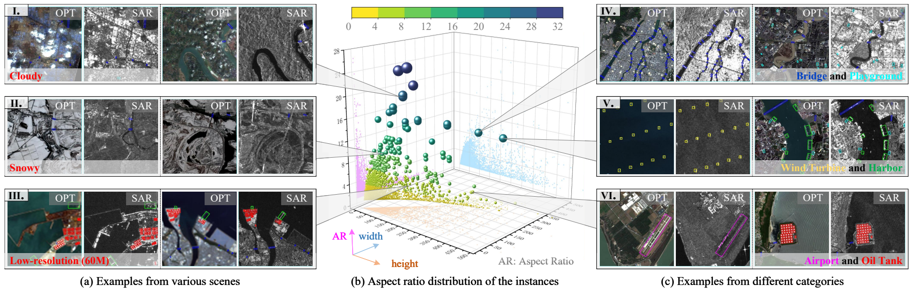

Fig. 1: Examples of scenes and six key target categories in the proposed M4-SAR dataset, accompanied by instance size and aspect ratio distributions.

have garnered significant attention in the remote sensing field. These methods have demonstrated remarkable effectiveness in applications, including disaster monitoring and urban planning. As shown in Fig. 1 (c), optical images provide high-resolution visual cues with rich texture, color, and spectral information, facilitating fine-grained feature extraction and semantic interpretation under clear atmospheric conditions. However, their detection performance degrades markedly under low-light, cloud-obscured, or low-resolution conditions (see Fig. 1 (a)). Despite these drawbacks, the intuitive appearance and broad availability of optical data have fueled significant advancements in optical image object detection [4,6,29,30,72]. In contrast, SAR sensors enable all-weather, all-day imaging, resilient to cloud cover and illumination changes. However, the inherent speckle noise and low contrast arising from microwave backscatter present significant obstacles to precise localization and classification. Moreover, constructing large-scale SAR datasets [1,16,20,25,58,60,71] is hindered by the high cost of data acquisition and the laborious fine-grained annotation. Although optical and SAR images exhibit strong complementarity in spatial resolution and environmental robustness, existing single-source [33,42,43,45,47,68] and multi-source object detection methods [3,14,62,66] face challenges in addressing cross-domain distribution shifts and suffer from the absence of a standardized evaluation benchmark. To enable robust detection in complex scenarios, there is an urgent need for a large-scale optical-SAR fusion object detection dataset and a standardized benchmarking toolkit.

Currently, a major bottleneck in optical-SAR fusion object detection lies in the lack of large-scale, high-quality, and standardized datasets. As shown in Table 1, most existing SAR datasets are limited to single-source data, suffering from limited scale, narrow category coverage, or inconsistent annotations. For example, HRSID [57], SADD [67], SSDD [69], SIVED [26], and ShipDataset [56] exhibit notable limitations in terms of category diversity and dataset scale. In recent years, the emergence of larger-scale datasets such as SARDet-100K [25], FAIR-CSAR [58], and RSAR [71] has partially addressed these limitations. However, these datasets are often synthesized by standardization of data from multiple sources, potentially introducing bias during data preprocessing and compro-

Table 1: Comparison between existing optical and SAR object detection datasets and the proposed M4-SAR dataset. O and S denote optical and SAR data, respectively.

Dataset	O	S	Resolution	Image Size	Cat.	Instances	Images	Ins/Img
HRSC2016 [29]	✓	✗	0.4M $\sim$ 2M	300 $\sim$ 1,500	1	2,976	1,061	2.80
LEVIR [72]	✓	✗	0.2M $\sim$ 1M	600×800	3	11,028	21,952	0.50
DIOR-R [4]	✓	✗	0.5M $\sim$ 30M	800×800	20	190,288	23,463	8.11
DOTA-v1.0 [6]	✓	✗	-	800 $\sim$ 4,000	15	188,282	2,806	67.10
HRSID [57]	✗	✓	-	800×800	1	16,969	5,604	3.03
SADD [67]	✗	✓	-	224×224	1	7,835	883	8.87
SSDD [69]	✗	✓	-	512×512	1	2,587	1,160	2.23
SIVED [26]	✗	✓	-	512×512	1	12,013	1,044	11.51
ShipDataset [56]	✗	✓	-	256×256	1	50,885	39,729	1.28
MSAR [59]	✗	✓	-	256×256	4	65,202	30,158	2.16
SARDet-100K [25]	✗	✓	-	512×512	6	245,653	116,598	2.11
FAIR-CSAR [58]	✗	✓	-	1,024×1,024	5	349,002	106,672	2.11
RSAR [71]	✗	✓	-	512×512	6	183,534	95,842	1.91
OGSOD-1.0 [53]	✓	✓	20M	256×256	3	48,589	18,331	2.65
OGSOD-2.0 [52]	✓	✓	20M	256×256	3	54,000	20,359	2.65
M4-SAR (Ours)	✓	✓	**10M, 60M**	**512×512**	**6**	**981,862**	**112,174**	**8.75**

mising the fairness of model evaluation. Moreover, constructing paired optical-SAR datasets is prohibitively expensive: acquiring SAR data is already costly, and the acquisition of spatio-temporally aligned optical images is even more difficult, further limiting the construction of multi-source datasets.

Although OGSOD-1.0 [53] and OGSOD-2.0 [52] represent pioneering efforts in optical-SAR multi-source data, they are primarily designed for cross-modal knowledge distillation. Moreover, their test sets consist solely of SAR images, which limits their applicability for end-to-end multi-source fusion object detection (Other related work can be found in Appendix 8.). To address this issue, we constructed the novel optical-SAR fusion object detection dataset (M4-SAR) covering multi-resolution, multi-polarization, multi-scene, and multi-source using publicly available optical and SAR data provided by Sentinel-1 [48], and Sentinel-2 [8] satellites. To overcome SAR annotation challenges, we propose a semi-supervised optical-assisted labeling strategy that utilizes optical images' semantic richness to substantially improve annotation quality. As illustrated in Fig. 1 and Table 1, M4-SAR includes six key target categories with 112,174 image pairs and 981,862 instances. M4-SAR surpasses existing datasets in instance volume and image numbers, establishing a robust benchmark for multi-source fusion object detection in complex scenarios.

Another important challenge is the lack of publicly available algorithms tailored for optical-SAR fusion object detection. Most existing studies are based on single-source datasets or evaluated on limited scenarios, making it difficult to conduct fair performance comparisons. To bridge this gap, we introduce Multi-Source Rotated Object Detector (MSRODet), an open-source toolkit that integrates state-of-the-art fusion detectors (e.g., MHFNet [70], CFT [37], CLANet [14], CSSA [3], CMADet [40], ICAFusion [39], and MMIDet [63]), providing a standardized evaluation framework for researchers utilizing the M4-SAR dataset. MSRODet not only enables fair and reproducible benchmarking across different fusion methods but also lays a solid foundation for the development of future multi-source fusion object detection algorithms.

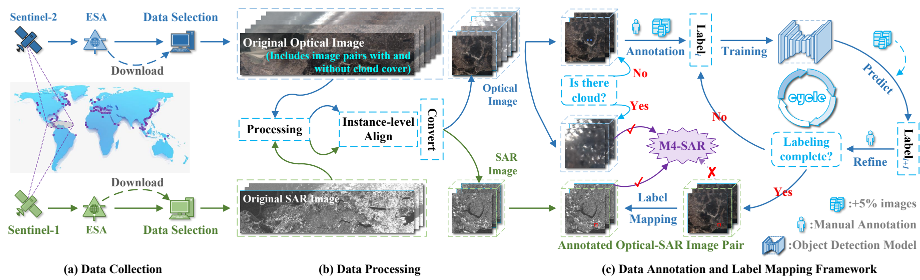

Fig. 2: Pipeline for M4-SAR dataset construction.

Moreover, extensive experiments demonstrate that the intrinsic differences in the imaging mechanisms of optical and SAR sensors lead to significant domain gaps (Fig. 4 (a)), thereby constraining the performance of optical-SAR fusion object detection. To address this issue, we propose the first end-to-end framework for optical-SAR fusion object detection, termed E2E-OSDet. This framework effectively fuses features from both modalities, alleviates cross-modal domain gaps, and results in significant improvements in detection performance.

The main contributions of this work are summarized as follows:

- **A Novel Optical-SAR Dataset:** M4-SAR is a novel optical-SAR fusion object detection dataset with multi-resolution, multi-polarization, and multi-scene characteristics, containing 112,174 image pairs and 981,862 instances. This dataset comprehensively covers complex scenes and diverse target categories frequently encountered in multi-source fusion detection tasks, thereby providing a robust foundation for advancing research in this field.
- **Multi-source Evaluation Toolkit (MSRODet) and E2E-OSDet:** We construct MSRODet, a comprehensive benchmarking toolkit incorporating state-of-the-art fusion object detection methods tailored for optical-SAR fusion object detection, which provides a standardized and reproducible evaluation platform based on the M4-SAR dataset. In addition, we propose E2E-OSDet, a novel end-to-end optical-SAR fusion object detection framework, which effectively mitigates significant cross-domain discrepancies between optical and SAR modalities at the feature level.

2 M4-SAR Dataset Construction

2.1 Overview of the Proposed Dataset Construction Pipeline

Fig. 2 illustrates the overall pipeline of the proposed dataset construction process. In the first stage, we acquire optical and SAR images from Sentinel-1 and Sentinel-2 satellites provided by the European Space Agency (ESA). In the second stage, the acquired images are pre-processed and aligned. After completing the alignment, we crop aligned image pairs into fixed-size patches to generate initial unlabeled samples. In the third stage, we introduce a semi-supervised annotation strategy guided by optical images to ensure high-quality labeling.

2.2 Data Collection

In Fig. 2 (a), we collected the required data via the open-access platform provided by the ESA. A map-based selection tool was used to specify areas of interest, covering representative coastal cities, including Tianjin, Los Angeles, Hong Kong, and London. We then retrieved SAR images along with their corresponding digital elevation model (DEM) data according to specific criteria. For optical imagery, Sentinel-2 images covering the same regions as the Sentinel-1 data were acquired, specifically including the red (R), green (G), and blue (B) spectral bands. To minimize variations caused by temporal discrepancies between the two modalities, we selected image pairs with a time interval of less than 10 days. This setting ensures that targets of interest (e.g., bridges, oil tanks) maintain structural consistency within the short interval, thereby enhancing the accuracy of subsequent instance-level alignment. The product type for optical images is L2A, with cloud coverage limited to 0–10% (the corresponding cloud-free (0%) images are used for high-quality annotation) to guarantee high visual fidelity. The SAR images are provided in the Single Look Complex (SLC) format to preserve full coherence information. The collected data span the period from 2020 to 2022 and encompass multiple spatial resolutions and polarization modes. The optical images have 10M, 20M, and 60M resolutions, and the SAR data cover both vertical-horizontal (VH) and vertical-vertical (VV) polarizations.

2.3 Data Processing

As shown in Fig. 2 (b), we combined the red, green, and blue channels to generate standard optical images using data processing software (ENVI). The optical images were provided in three spatial resolutions: 10M, 20M, and 60M. To highlight the complementary characteristics of optical and SAR images, we retained only the high-resolution (10M) and low-resolution (60M) variants. For the SAR data, we first applied multi-look averaging to the SLC images in both the azimuth and range directions to reduce speckle effects and improve radiometric resolution. Subsequently, filtering techniques were employed to suppress speckle noise. Using the corresponding DEM data, we geocoded and radiometrically calibrated the SAR images and conducted geometric alignment and correction to improve spatial localization precision. After preprocessing, the optical and SAR images were aligned based on their geographic coordinates. As the overlap between the two modalities was partial, the aligned image pairs were cropped into 1,024×1,024 patches, and only those with valid overlapping regions were retained for subsequent annotation. Notably, this initial cropping was intended to facilitate efficient manual annotation. After annotation was completed, we further partitioned the annotated images into non-overlapping 512×512 patches to meet the demands of different downstream applications.

2.4 Data Annotation and Label Mapping

As the acquisition times of Sentinel-1 [48] and Sentinel-2 [8] are not perfectly synchronized, instance-level alignment of dynamic targets (e.g., ships) becomes difficult, which compromises the accuracy of label transfer. To address this issue, we focus on static and semantically meaningful targets, such as bridges, harbors, oil tanks, playgrounds, airports, and wind turbines, and annotate them using oriented bounding boxes. SAR's unique imaging characteristics cause ambiguous target boundaries, challenging accurate annotation by non-experts. In contrast, optical images (cloudless) provide clearer visual cues that facilitate both target identification and annotation transfer. With precise alignment, annotations derived from optical images can be reliably transferred to SAR images, thereby significantly reducing manual annotation costs.

Given the scale of the dataset, fully manual labeling is both time-consuming and labor-intensive. To address this challenge, we propose a semi-supervised annotation strategy, as illustrated in Fig. 2 (c). The procedure involves the following steps: First, 5% of the entire cloudless optical images are manually annotated, ensuring that each selected image contains all six target categories: bridges, harbors, oil tanks, playgrounds, airports, and wind turbines. This subset establishes a high-quality foundation for subsequent annotation. Once manually labeled, a detector [17] is trained using this subset (with the same data used for both training and evaluation during the initial phase). The trained model is then employed to generate pseudo-labels for an additional 5% of the data, which are subsequently refined through manual correction and added to the labeled set. This updated set is used to retrain the detector in the next iteration. The annotation process is repeated iteratively until the full dataset is labeled. During this procedure, images without valid target instances are excluded to enhance annotation efficiency of the M4-SAR dataset.

During the annotation process, we use cloud-free images to ensure annotation accuracy. After annotation is complete, the cloud-free images are replaced with the corresponding cloud-containing images. Based on instance-level alignment, we achieve high-quality label transfer, enabling accurate projection of annotation information from optical images to their corresponding SAR counterparts. As a result, we construct M4-SAR, the large-scale dataset for optical-SAR fusion object detection, which offers a solid foundation for subsequent method evaluation and algorithm development. More dataset details are presented in Appendix 7.

2.5 Data Analysis

As shown in Table 1, our proposed M4-SAR dataset comprises 112,174 image pairs, 981,862 instances, and six representative target categories: bridges, harbors, oil tanks, playgrounds, airports, and wind turbines. Each image has a resolution of 512×512 pixels and contains an average of 8.75 instances, highlighting the dense object distribution typical in remote sensing scenes. As illustrated in Fig. 1 (c), rotated bounding boxes offer significantly higher localization accuracy compared to horizontal bounding boxes. The challenge analysis of category-specific attributes in the M4-SAR dataset is presented as follows:

– **Aspect Ratio Challenge:** The aspect ratio of an object plays a critical role in detection accuracy. Fig. 1 (b) illustrates the distribution of aspect ratios

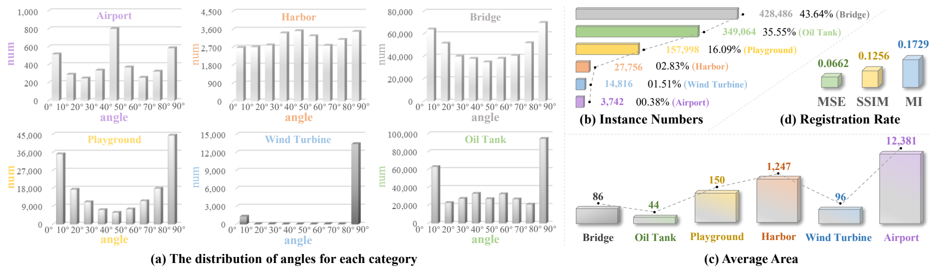

Fig. 3: Statistical visualization of category attributes in the proposed M4-SAR dataset. (a) Angle distribution of instances in each category. (b) Percentage of instances per category. (c) Average pixel area of instances per category.

across categories. When the aspect ratio approaches 1, targets tend to appear square, potentially leading to recognition ambiguities. Conversely, extremely large aspect ratios increase the model's sensitivity to object shapes, thus posing additional challenges to accurate detection.

- **Angle Diversity Challenge:** All target orientation angles are normalized to θ ($\theta \in [0, \pi/2)$, consistent with YOLO-OBB [17] rules.), and their distributions are visualized in Fig. 3 (a). Notably, offshore wind turbines tend to exhibit square-like appearances, with orientation angles predominantly distributed to 0° and 90° due to the absence of surrounding reference structures and the top-down perspective of satellite imagery. Although oil tanks are circular, their labeled angles are contextually defined based on surrounding spatial features. This highlights the importance of contextual cues in remote sensing images and reflects the task's sensitivity.
- **Category Imbalance Challenge:** Fig. 3 (b) displays the distribution of instances by category, revealing that bridges (43.64%) constitute the largest proportions in the dataset. This imbalance reflects the actual occurrence frequency of typical targets in coastal regions: static infrastructure such as bridges and oil tanks dominates satellite imagery. The relatively low number of harbor (2.83%) instances is primarily due to their extensive spatial coverage, with a single port typically corresponding to a single instance.
- **Scale Inconsistency Challenge:** Fig. 3 (c) illustrates the average pixel area of each category, highlighting the substantial variation in object scales across different categories. Such large discrepancies in object size (i.e., object area) introduce a critical and commonly encountered challenge for robust object detection, as models must simultaneously handle extremely small and large targets within a unified framework.
- **Modal Weak Alignment Challenge:** The optical and SAR images are aligned only using geographic coordinates, resulting in coarse rather than pixel-level correspondence. Together with non-simultaneous acquisitions and dynamic scene changes, this leads to notable cross-modal misalignment, as reflected by the low Mean Squared Error (MSE: 0.0662), Structural Similarity Index Measure (SSIM: 0.1256), and Mutual Information (MI: 0.1729)

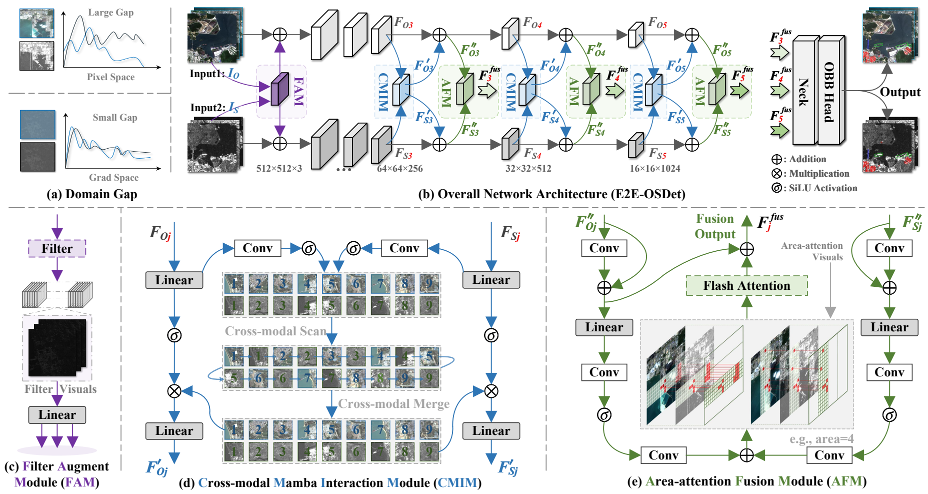

Fig. 4: (a) Domain gap between optical and SAR images. (b) Overall framework of the proposed end-to-end optical-SAR fusion object detection (E2E-OSDet). Architectural details of the proposed filter augment module (c), cross-modal Mamba interaction module (d), and area-attention fusion module (e).

scores shown in Fig. 3 (d). This weak alignment poses a significant challenge for multi-source fusion object detection.

3 Evaluation Tools and Methods

3.1 Optical-SAR Fusion Object Detection Tools (MSRODet)

Due to the lack of publicly available methods specifically designed for optical-SAR fusion object detection, conducting a fair performance comparison remains challenging. To address this limitation, we propose a Multi-Source Rotated Detector (MSRODet) built upon existing multi-source detection frameworks (e.g., RGB-infrared, RGB-depth). MSRODet integrates several representative methods, including MHFNet [70], CFT [37], CLANet [14], CSSA [3], CMADet [40], ICAFusion [39], and MMIDet [63], thereby providing a standardized and accessible benchmark for evaluating optical-SAR fusion object detection approaches. As shown in Fig. 4 (a), empirical analysis reveals that significant domain gaps remain a primary obstacle to effective optical-SAR fusion. These discrepancies primarily arise from fundamental differences in visual characteristics and imaging mechanisms between optical and SAR modalities. Consequently, designing detection frameworks specifically tailored to optical-SAR fusion is essential for enhancing overall detection performance.

3.2 End-to-End Optical-SAR Fusion Object Detection (E2E-OSDet)

Multi-source fusion object detection remains underexplored in the context of optical-SAR fusion, and significant limitations persist in addressing domain discrepancies between optical and SAR data. As shown in Fig. 4 (b), to address these challenges, we propose a novel end-to-end optical-SAR fusion object detection framework, termed E2E-OSDet, which systematically addresses cross-modal challenges across three levels: data input, domain alignment, and feature fusion.

Filter Augment Module (FAM): First, to reduce the domain gap between optical and SAR images, we propose a Filter Augment Module (FAM, Fig. 4 (c)) leveraging classical image filtering operators (e.g., HOG [5], Canny [2], Haar [49], Grad [21], and WST [35]). This module transforms the low-dimensional and sparse SAR representations into a more discriminative high-dimensional space, effectively mitigating the impact of domain differences.

Cross-modal Mamba Interaction Module (CMIM): Second, most existing methods focus on spatial-level fusion but often neglect the discrepancies in the feature domain. Mamba [11] possesses strong state-space modeling capabilities, which enables effective cross-modal sequence-level interactions and alleviates feature domain bias. However, current Mamba-based methods [7, 15, 64] are hindered by complex architectures and intricate interaction mechanisms, limiting their practical applicability. To address this, we propose a Cross-modal Mamba Interaction Module (CMIM, Fig. 4 (d)), which enables deep cross-modal interactions through simple and efficient operations.

Area-Attention Fusion Module (AFM): Finally, to further enhance the discriminability of fused features across different spatial regions, we introduce an Area-Attention Fusion Module (AFM). This module captures salient features of local targets and strengthens responses in critical regions, thereby enhancing the effectiveness of feature fusion. The detailed structure of AFM is illustrated in Fig. 4 (e). For additional E2E-OSDet details, please refer to Appendix 9.

4 Experiments and Analysis

4.1 Experiment Setup and Implementation Details

Experiments were conducted on two NVIDIA RTX 4090 GPUs. All fusion models (MSRODet, COMO, and the proposed E2E-OSDet) were trained over 300 epochs with the SGD optimizer, using a learning rate of 0.01, momentum of 0.937, and weight decay of 5×10^{-4}. The batch size was set to 64, and the input images were resized to a resolution of 512×512. To ensure fair performance evaluation, all methods were evaluated using a unified framework comprising the YOLOv11 [17] backbone and the oriented detection head from YOLOv8 [17]. These fusion models were trained from scratch, without leveraging any pre-trained weights. During evaluation, we adopted the standard COCO evaluation protocol to measure Average Precision (AP), including AP_{50}, AP_{75}, and mAP.

Table 2: Quantitative results on the M4-SAR. O and S denote optical and SAR data, respectively. Param. denotes the parameter count of the model. Inf.T indicates inference time per image. The best two results are highlighted in **Black** and <u>mark</u>.

Method	**O**	**S**	**Param. (M)** ↓	**FLOPs (G)** ↓	**Inf.T (ms)** ↓	**AP_{50}** ↑							**AP_{75}** ↑	**mAP** ↑
						Bri.	Har.	Oil.	Ply.	Apr.	Win.	*ALL*		
R-FCOS [47]	✓	✗	31.9	51.6	17.6	32.8	79.8	29.7	67.0	78.9	54.8	57.2	26.4	30.2
R-ATSS [68]	✓	✗	36.0	51.9	19.1	33.2	80.0	29.3	66.6	78.3	54.8	57.0	27.1	30.0
O-RepPoint [22]	✓	✗	36.6	48.6	16.1	39.7	71.9	21.2	56.3	65.5	53.1	51.3	17.9	26.7
RTMDet [34]	✓	✗	8.9	51.0	11.1	34.0	79.8	30.3	61.1	79.1	55.9	56.7	26.5	29.2
PSC [61]	✓	✗	31.9	51.7	18.3	30.7	79.7	25.9	57.4	78.1	49.4	53.5	26.2	28.5
LSKNet [24]	✓	✗	21.8	43.0	20.4	42.9	80.6	34.9	78.4	79.1	54.0	61.6	31.2	34.0
YOLOv5 [17]	✓	✗	8.1	12.8	7.2	61.4	88.0	56.6	76.4	91.3	91.8	77.6	59.8	52.9
YOLOv6 [18]	✓	✗	16.0	27.7	6.1	57.6	82.1	52.7	67.6	89.4	90.4	73.3	64.9	48.8
YOLOv8 [17]	✓	✗	10.0	15.7	6.7	63.6	88.6	57.9	80.0	91.3	92.5	79.0	61.6	54.4
YOLOv9 [55]	✓	✗	6.4	15.1	14.4	61.9	87.7	55.3	74.4	90.0	92.4	76.9	58.9	52.5
YOLOv10 [50]	✓	✗	7.5	14.5	7.8	65.2	89.9	58.8	82.3	90.8	93.5	80.1	62.8	55.7
YOLOv11 [17]	✓	✗	9.7	14.4	6.5	64.3	90.0	57.8	80.7	90.9	91.5	79.2	60.5	54.3
YOLOv12 [46]	✓	✗	9.4	14.4	13.4	62.7	88.6	57.1	80.1	91.0	92.7	78.7	61.1	54.0
YOLO26 [17]	✓	✗	10.5	15.7	9.1	63.0	88.2	58.3	80.6	90.3	90.1	78.4	62.8	55.4
YOLO-Master [27]	✓	✗	29.3	35.6	15.7	64.4	90.0	59.0	80.7	92.3	93.6	80.0	63.9	56.3
R-FCOS [47]	✗	✓	31.9	51.6	17.6	23.9	80.1	27.9	60.1	75.8	43.5	51.9	26.4	28.3
R-ATSS [68]	✗	✓	36.0	51.9	19.1	25.3	80.3	26.8	62.3	77.0	43.9	52.6	24.9	27.8
O-RepPoint [22]	✗	✓	36.6	48.6	16.1	26.2	78.6	21.9	53.7	59.9	48.8	48.2	13.9	20.4
RTMDet [34]	✗	✓	8.9	51.0	11.1	16.2	71.0	21.1	39.3	68.9	39.6	42.7	20.9	22.3
PSC [61]	✗	✓	31.9	51.7	18.3	22.3	80.0	20.6	51.9	75.7	39.4	48.3	24.5	26.5
LSKNet [24]	✗	✓	21.8	43.0	20.4	33.4	80.2	28.4	69.3	76.8	51.2	56.6	27.1	30.4
YOLOv5 [17]	✗	✓	8.1	12.8	7.2	40.5	85.3	41.0	64.3	85.5	89.9	67.7	44.0	41.3
YOLOv6 [18]	✗	✓	16.0	27.7	6.1	36.5	79.4	32.8	53.5	85.7	86.7	62.4	37.6	36.5
YOLOv8 [17]	✗	✓	10.0	15.7	6.7	44.9	85.8	42.5	67.9	87.5	89.2	69.6	46.1	43.2
YOLOv9 [55]	✗	✓	6.4	15.1	14.4	43.9	84.8	38.4	61.5	87.9	88.8	67.6	44.2	41.4
YOLOv10 [50]	✗	✓	7.5	14.5	7.8	49.8	87.0	45.5	73.4	87.9	90.4	72.3	50.8	46.2
YOLOv11 [17]	✗	✓	9.7	14.4	6.5	46.9	87.2	44.6	71.1	88.8	89.2	71.3	48.6	44.7
YOLOv12 [46]	✗	✓	9.4	14.4	13.0	46.4	87.7	43.6	70.8	90.0	90.7	71.5	49.2	45.4
YOLO26 [17]	✗	✓	10.5	15.7	9.1	48.8	88.3	48.9	75.8	88.9	88.9	73.3	53.8	48.2
YOLO-Master [27]	✗	✓	29.3	35.6	15.7	49.4	89.3	48.3	75.1	90.6	90.8	73.9	53.1	48.5
COMO [28]	✓	✓	22.6	24.3	37.1	72.3	90.8	60.4	88.7	90.5	96.3	83.2	66.3	57.9
MHFNet [70]	✓	✓	16.1	26.7	19.2	73.7	91.4	<u>61.8</u>	89.9	90.4	95.6	83.8	65.6	58.0
CFT [37]	✓	✓	53.8	31.8	40.6	<u>75.8</u>	<u>92.5</u>	61.3	91.6	90.3	96.3	84.6	<u>68.9</u>	<u>59.9</u>
CLANet [14]	✓	✓	48.2	53.4	29.1	74.8	92.2	60.7	91.3	91.6	<u>97.2</u>	84.6	68.5	59.6
CSSA [3]	✓	✓	13.5	21.4	12.3	73.3	91.7	59.3	88.9	91.6	95.8	83.4	66.4	58.0
CMADet [40]	✓	✓	41.5	31.8	46.7	70.9	90.7	52.0	86.4	91.7	97.1	81.5	63.5	55.7
ICAFusion [39]	✓	✓	29.0	26.7	23.6	74.7	91.9	60.9	91.0	<u>91.8</u>	96.7	84.5	67.3	58.8
MMIDet [63]	✓	✓	53.8	31.8	41.9	74.9	**92.6**	61.1	<u>91.7</u>	91.4	97.0	<u>84.8</u>	68.6	59.8
E2E-OSDet	✓	✓	24.7	40.7	20.9	**77.7**	90.7	**64.3**	**91.8**	**92.1**	**97.8**	**85.7**	**70.3**	**61.4**

4.2 Main Results

Experimental results (see Table 2) indicate that detection accuracy using SAR images alone is significantly lower than when using optical images alone. This gap primarily stems from the unique imaging mechanism of SAR: even when targets are visible in the image, SAR struggles to accurately estimate the shape of target bounding boxes. In contrast, all multi-source fusion methods outperform their corresponding single-modal baselines, clearly demonstrating that cross-modal fusion effectively leverages the complementarity between SAR and optical data to significantly reduce false positives and enhance detection accuracy. Among these, our proposed E2E-OSDet achieves the highest performance across all fusion methods, with a mAP of 61.4%, fully validating the importance of mitigating domain differences in cross-modal detection. Furthermore, E2E-OSDet exhibits high efficiency with only 24.7M parameters and achieves an inference

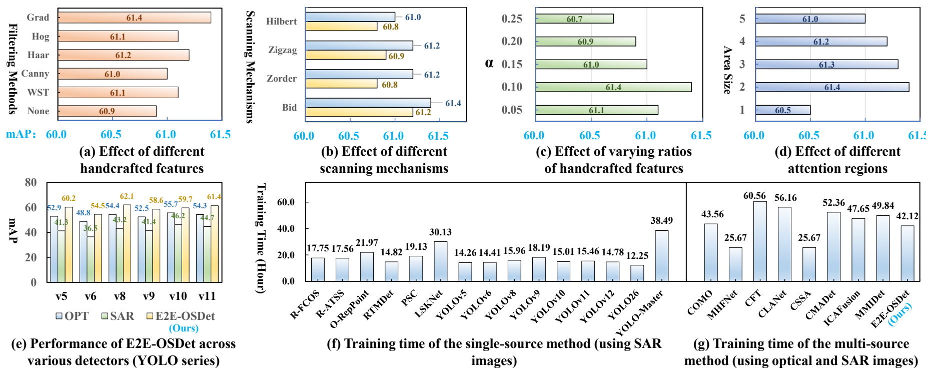

Fig. 5: Comprehensive performance evaluation under different parameter settings.

Table 3: Ablation analysis of E2E-OSDet on M4-SAR dataset.

No.	FAM	CMIM	AFM	Train-Params.	$AP_{50}\uparrow$	$AP_{75}\uparrow$	mAP $\uparrow$
(a)	✗	✗	✗	14.17M Baseline	83.4	66.4	58.0
(b)	✓	✗	✗	14.17M +0M	84.5	67.1	58.9
(c)	✗	✓	✗	19.01M +5.68M	84.9	68.6	59.6
(d)	✗	✗	✓	19.84M +4.84M	84.6	68.9	60.0
(e)	✓	✓	✗	19.84M +4.84M	85.5	69.9	61.0
(f)	✓	✗	✓	19.01M +5.68M	84.9	69.5	60.7
(g)	✗	✓	✓	24.69M +10.52M	85.2	70.0	60.9
(h)	✓	✓	✓	24.69M +10.52M	**85.7**	**70.3**	**61.4**

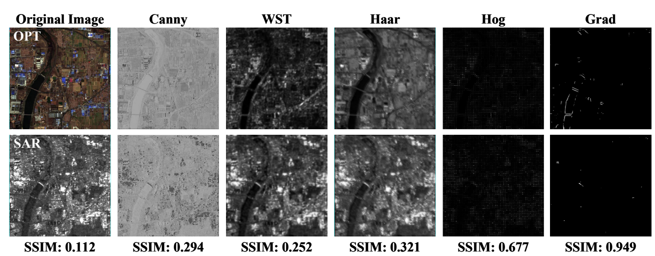

Fig. 6: Visualization results of handcrafted features (Canny, WST, Haar, Hog, Grad).

speed of 20.9ms. As shown in Fig. 5 (e), we also evaluated its generalization capability across different detection frameworks (YOLO series, including v5, v6, v8, v9, v10.), demonstrating strong transferability and robustness.

4.3 Ablation Experiment

As shown in Table 3 (a), we adopt CSSA [3] as our baseline model, which serves as the foundation for independently evaluating the contribution of each proposed module. The results (see Table 3 (a), (b), and (c)) indicate that all three modules lead to performance gains, among which the AFM module provides the most substantial improvement in detection accuracy. We further assess the interoperability of the modules. As shown in Table 3 (e), (f), and (g), the modules

Fig. 7: Qualitative comparison of the proposed E2E-OSDet and six fusion methods.

are compatible with one another, and their integration does not result in any performance degradation. Finally, we integrate all modules (see Table 3 (h)), resulting in additional gains in overall model performance.

Impact of Filter Augment: To assess the effectiveness of the FAM, we tested a range of conventional feature descriptors. As shown in Fig. 5 (a) and Fig. 6, the results indicate that incorporating handcrafted features significantly enhances detection performance. These results mapping pixels into a handcrafted feature space effectively reduces the distribution gap between optical and SAR data. Notably, the Grad feature yields superior performance due to its ability to narrow domain differences and capture multi-scale information.

Impact of Scanning Mechanisms: We further investigate the effectiveness of the CMIM module under different scanning mechanisms [12]. As illustrated in Fig. 5 (b)(blue bars), CMIM consistently enhances detection accuracy regardless of the underlying scanning strategy. This observation highlights the robustness and adaptability of CMIM, suggesting that the proposed module can serve as a versatile component that generalizes well across diverse architectural settings rather than being restricted to a particular scanning design.

Impact of Feature Proportion and Area Size: In addition, we investigate the influence of two key design choices: the proportion of handcrafted

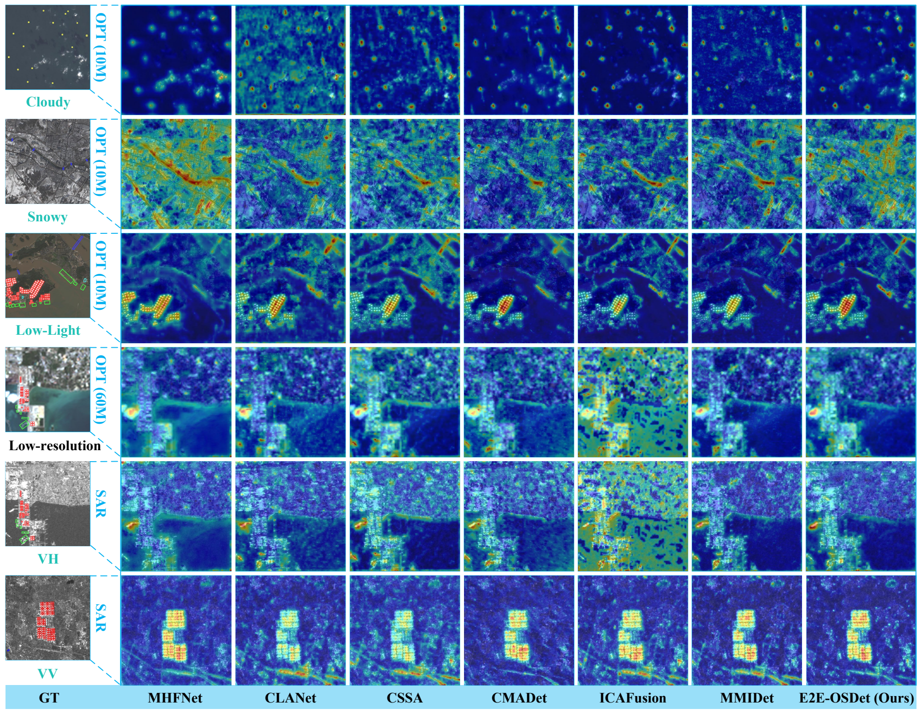

Fig. 8: Grad-CAM [38] heatmaps of the proposed E2E-OSDet and six fusion methods.

feature input and the adoption of area attention. As shown in Fig. 5 (c) and (d), varying these hyperparameters leads to only marginal performance differences, indicating that the overall framework is relatively insensitive to such configurations. This stability suggests that our proposed E2E-OSDet achieves consistent performance without requiring delicate hyperparameter tuning, further underscoring its practicality in real-world deployment.

4.4 Qualitative Analysis

As shown in Fig. 7 and Fig. 8, our M4-SAR effectively reveals the strengths and limitations of existing multi-source methods under realistic challenges such as weak cross-modal alignment, scale inconsistency, and angle diversity. Due to coarse geographic alignment and non-simultaneous acquisition, noticeable spatial discrepancies exist between optical and SAR images. Under such conditions, direct fusion methods (e.g., ICAFusion) are prone to mislocalization and orientation errors, while alignment-aware approaches (e.g., CMADet) demonstrate improved robustness. However, even these methods struggle in cases with severe modal shifts or complex backgrounds. In contrast, the E2E-OSDet produces more precise bounding boxes, more accurate angle predictions, and more compact feature activations, indicating stronger cross-modal interaction and background suppression capability. These results validate both the diagnostic value of our dataset in evaluating multi-source fusion strategies and the effectiveness of the E2E-OSDet. More experimental results can be seen in Appendix 10.

5 Limitations Analysis and Future Work

Despite the substantial advantages of the proposed M4-SAR in terms of scale, diversity, and scene coverage, certain limitations persist that require further exploration. First, the optical and SAR images are aligned using geographic coordinates rather than pixel-level registration. As a result, precise spatial correspondence between modalities is not guaranteed. This design better reflects realistic cross-sensor acquisition conditions, but it also means that M4-SAR is not suitable for tasks requiring accurate pixel-wise alignment, such as image registration, change detection with strict spatial consistency, or pixel-level fusion. Second, the M4-SAR dataset currently focuses on six representative coastal infrastructure categories. Although these classes exhibit significant variation in scale, aspect ratio, and orientation, the semantic diversity is still limited compared with large-scale natural image benchmarks. Future extensions could incorporate more fine-grained or dynamic categories to further broaden evaluation coverage. These limitations reflect practical trade-offs during dataset construction and highlight potential directions for future improvements.

Our proposed M4-SAR establishes a high-quality benchmark for optical–SAR fusion object detection, and the accompanying E2E-OSDet provides a strong baseline for future research. Building upon this foundation, several promising directions merit further investigation. First, expanding the dataset to include more dynamic and fine-grained categories would enhance semantic diversity and broaden its applicability. Second, designing lightweight multimodal fusion architectures could facilitate deployment in resource-constrained or edge computing scenarios. Third, incorporating temporal alignment strategies and multimodal temporal modeling may help mitigate the effects of non-simultaneous acquisitions, further improving cross-modal consistency and robustness.

6 Conclusion

This paper presents the M4-SAR dataset and the E2E-OSDet method to tackle two persistent challenges in optical-SAR fusion object detection: the scarcity of high-quality, large-scale datasets and the absence of a unified and effective algorithmic framework. We also develop MSRODet, a comprehensive open-source toolkit for multi-source rotated object detection that integrates mainstream fusion detection algorithms and enables fair performance comparisons on the M4-SAR dataset. We expect that M4-SAR, as a benchmark dataset, will substantially advance research in multi-source remote sensing detection, while E2E-OSDet, as the first fusion framework specifically designed for optical-SAR scenarios, will serve as a strong baseline for future studies.

Acknowledgements

This work was supported in part by the National Science Fund of China (No. 62276135, and U24A20330).

References

1. Bowald, D., Wheelwright, J., Nina, O., Sappa, A., Hammoud, R., Blasch, E., Inkawhich, N.: 3rd multi-modal aerial view image challenge: Sensor domain translation-pbvs 2025. In: CVPR. pp. 4640–4649 (2025)
2. Canny, J.: A computational approach to edge detection. IEEE TPAMI (6), 679–698 (1986)
3. Cao, Y., Bin, J., Hamari, J., Blasch, E., Liu, Z.: Multimodal object detection by channel switching and spatial attention. In: CVPR. pp. 403–411 (2023)
4. Cheng, G., Wang, J., Li, K., Xie, X., Lang, C., Yao, Y., Han, J.: Anchor-free oriented proposal generator for object detection. IEEE TGRS **60**, 1–11 (2022)
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. vol. 1, pp. 886–893 (2005)
6. Ding, J., Xue, N., Xia, G.S., Bai, X., Yang, W., Yang, M.Y., Belongie, S., Luo, J., Datcu, M., Pelillo, M., et al.: Object detection in aerial images: A large-scale benchmark and challenges. IEEE TPAMI **44**(11), 7778–7796 (2021)
7. Dong, W., Zhu, H., Lin, S., Luo, X., Shen, Y., Guo, G., Zhang, B.: Fusion-mamba for cross-modality object detection. IEEE TMM (2025)
8. Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., et al.: Sentinel-2: Esa's optical high-resolution mission for gmes operational services. RSE **120**, 25–36 (2012)
9. Geng, P., Lu, X., Hu, C., Liu, H., Lyu, L.: Focusing fine-grained action by self-attention-enhanced graph neural networks with contrastive learning. IEEE TCSVT **33**(9), 4754–4768 (2023)
10. Girshick, R.: Fast r-cnn. In: ICCV. pp. 1440–1448 (2015)
11. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. In: COLM (2024)
12. He, H., Bai, Y., Zhang, J., He, Q., Chen, H., Gan, Z., Wang, C., Li, X., Tian, G., Xie, L.: Mambaad: Exploring state space models for multi-class unsupervised anomaly detection. NeurIPS **37**, 71162–71187 (2024)
13. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV. pp. 2961–2969 (2017)
14. He, X., Tang, C., Zou, X., Zhang, W.: Multispectral object detection via cross-modal conflict-aware learning. In: ACM MM. pp. 1465–1474 (2023)
15. Hu, X., Tai, Y., Zhao, X., Zhao, C., Zhang, Z., Li, J., Zhong, B., Yang, J.: Exploiting multimodal spatial-temporal patterns for video object tracking. In: AAAI. vol. 39, pp. 3581–3589 (2025)
16. Inkawhich, N., Thorp, C., Wheelwright, J., Nina, O., Bowald, D., Sappa, A., Blasch, E.: 4th multi-modal aerial view image challenge: Sar classification-pbvs 2025. In: CVPR. pp. 4631–4639 (2025)
17. Jocher, G., Qiu, J., Chaurasia, A.: Ultralytics YOLO (Jan 2023), `https://github.com/ultralytics/ultralytics`
18. Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., Ke, Z., Li, Q., Cheng, M., Nie, W., et al.: Yolov6: A single-stage object detection framework for industrial applications. arXiv preprint arXiv:2209.02976 (2022)

19. Li, K., Wang, D., Hu, Z., Zhu, W., Li, S., Wang, Q.: Unleashing channel potential: Space-frequency selection convolution for sar object detection. In: CVPR. pp. 17323–17332 (2024)
20. Li, W., Yang, W., Hou, Y., Liu, L., Liu, Y., Li, X.: Saratr-x: Towards building a foundation model for sar target recognition. IEEE TIP (2025)
21. Li, W., Yang, W., Liu, T., Hou, Y., Li, Y., Liu, Z., Liu, Y., Liu, L.: Predicting gradient is better: Exploring self-supervised learning for sar atr with a joint-embedding predictive architecture. ISPRS **218**, 326–338 (2024)
22. Li, W., Chen, Y., Hu, K., Zhu, J.: Oriented reppoints for aerial object detection. In: CVPR. pp. 1829–1838 (2022)
23. Li, X., Du, Z., Huang, Y., Tan, Z.: A deep translation (gan) based change detection network for optical and sar remote sensing images. ISPRS **179**, 14–34 (2021)
24. Li, Y., Li, X., Dai, Y., Hou, Q., Liu, L., Liu, Y., Cheng, M.M., Yang, J.: Lsknet: A foundation lightweight backbone for remote sensing. IJCV pp. 1–22 (2024)
25. Li, Y., Li, X., Li, W., Hou, Q., Liu, L., Cheng, M.M., Yang, J.: Sardet-100k: Towards open-source benchmark and toolkit for large-scale sar object detection. NeurIPS **37**, 128430–128461 (2024)
26. Lin, X., Zhang, B., Wu, F., Wang, C., Yang, Y., Chen, H.: Sived: A sar image dataset for vehicle detection based on rotatable bounding box. RS **15**(11), 2825 (2023)
27. Lin, X., Peng, J., Gan, Z., Zhu, J., Liu, J.: Yolo-master: Moe-accelerated with specialized transformers for enhanced real-time detection. arXiv preprint arXiv:2512.23273 (2025)
28. Liu, C., Ma, X., Yang, X., Zhang, Y., Dong, Y.: Como: Cross-mamba interaction and offset-guided fusion for multimodal object detection. Information Fusion **125**, 103414 (2026)
29. Liu, Z., Yuan, L., Weng, L., Yang, Y.: A high resolution optical satellite image dataset for ship recognition and some new baselines. In: ICPRAM. vol. 2, pp. 324–331 (2017)
30. Llerena, J.M., Zeni, L.F., Kristen, L.N., Jung, C.: Gaussian bounding boxes and probabilistic intersection-over-union for object detection. arXiv preprint arXiv:2106.06072 (2021)
31. Lu, W., Chen, S.B., Ding, C.H., Tang, J., Luo, B.: Lwganet: A lightweight group attention backbone for remote sensing visual tasks. arXiv preprint arXiv:2501.10040 (2025)
32. Lu, W., Li, H.D., Wang, C., Chen, S.B., Ding, C.H., Tang, J., Luo, B.: Unravelnet: A backbone for enhanced multi-scale and low-quality feature extraction in remote sensing object detection. ISPRS **231**, 431–442 (2026)
33. Lu, W., Yang, X., Chen, S.B.: Lwganet: Addressing spatial and channel redundancy in remote sensing visual tasks with light-weight grouped attention. In: AAAI. vol. 40, pp. 7574–7582 (2026)
34. Lyu, C., Zhang, W., Huang, H., Zhou, Y., Wang, Y., Liu, Y., Zhang, S., Chen, K.: Rtmdet: An empirical study of designing real-time object detectors. arXiv preprint arXiv:2212.07784 (2022)
35. Mallat, S.: Group invariant scattering. CPAM **65**(10), 1331–1398 (2012)
36. MMRotate Contributors: OpenMMLab rotated object detection toolbox and benchmark (Feb 2022), `https://github.com/open-mmlab/mmrotate`
37. Qingyun, F., Zhaokui, W.: Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery. PR **130**, 108786 (2022)

38. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV. pp. 618–626 (2017)
39. Shen, J., Chen, Y., Liu, Y., Zuo, X., Fan, H., Yang, W.: Icafusion: Iterative cross-attention guided feature fusion for multispectral object detection. PR **145**, 109913 (2024)
40. Song, K., Xue, X., Wen, H., Ji, Y., Yan, Y., Meng, Q.: Misaligned visible-thermal object detection: a drone-based benchmark and baseline. IEEE TIV (2024)
41. Sun, Y., Lian, J., Yang, J., Luo, L.: Controllable-lpmoe: Adapting to challenging object segmentation via dynamic local priors from mixture-of-experts. In: ICCV. pp. 22327–22337 (2025)
42. Sun, Y., Wang, C., Yang, J., Luo, L.: Small but mighty: Dynamic wavelet expert-guided fine-tuning of large-scale models for optical remote sensing object segmentation. AAAI (2025)
43. Sun, Y., Xu, C., Yang, J., Xuan, H., Luo, L.: Frequency-spatial entanglement learning for camouflaged object detection. In: ECCV. pp. 343–360 (2024)
44. Sun, Y., Xuan, H., Yang, J., Luo, L.: Glconet: Learning multisource perception representation for camouflaged object detection. TNNLS (2024)
45. Sun, Y., Yang, J., Luo, L.: United domain cognition network for salient object detection in optical remote sensing images. IEEE TGRS **62**, 1–14 (2024)
46. Tian, Y., Ye, Q., Doermann, D.: Yolov12: Attention-centric real-time object detectors. arXiv preprint arXiv:2502.12524 (2025)
47. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: ICCV. pp. 9627–9636 (2019)
48. Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M., et al.: Gmes sentinel-1 mission. RSE **120**, 9–24 (2012)
49. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR. vol. 1, pp. I–I (2001)
50. Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., et al.: Yolov10: Real-time end-to-end object detection. NeurIPS **37**, 107984–108011 (2024)
51. Wang, C., Fang, W., Li, X., Yang, J., Luo, L.: Msod: A large-scale multiscene dataset and a novel diagonal-geometry loss for sar object detection. IEEE TGRS **63**, 1–13 (2025)
52. Wang, C., Luo, L., Fang, W., Yang, J.: Cross-modal gaussian localization distillation for optical information guided sar object detection. In: ICASSP. pp. 1–5 (2025)
53. Wang, C., Ruan, R., Zhao, Z., Li, C., Tang, J.: Category-oriented localization distillation for sar object detection and a unified benchmark. IEEE TGRS **61**, 1–14 (2023)
54. Wang, C., Sun, Y., Yang, J., Luo, L.: Localized background-aware generative distillation for enhanced remote sensing object detection. IEEE TCSVT (2026)
55. Wang, C.Y., Yeh, I.H., Mark Liao, H.Y.: Yolov9: Learning what you want to learn using programmable gradient information. In: ECCV. pp. 1–21 (2024)
56. Wang, Y., Wang, C., Zhang, H., Dong, Y., Wei, S.: A sar dataset of ship detection for deep learning under complex backgrounds. RS **11**(7), 765 (2019)
57. Wei, S., Zeng, X., Qu, Q., Wang, M., Su, H., Shi, J.: Hrsid: A high-resolution sar images dataset for ship detection and instance segmentation. IEEE Access **8**, 120234–120254 (2020)

58. Wu, Y., Suo, Y., Meng, Q., Dai, W., Miao, T., Zhao, W., Yan, Z., Diao, W., Xie, G., Ke, Q., et al.: Fair-csar: A benchmark dataset for fine-grained object detection and recognition based on single look complex sar images. IEEE TGRS (2024)
59. Xia, R., Chen, J., Huang, Z., Wan, H., Wu, B., Sun, L., Yao, B., Xiang, H., Xing, M.: Crtranssar: A visual transformer based on contextual joint representation learning for sar ship detection. RS **14**(6), 1488 (2022)
60. Ye, Y., Teng, X., Yang, H., Chen, S., Sun, Y., Bian, Y., Tan, T., Li, Z., Yu, Q.: 3mos: a multi-source, multi-resolution, and multi-scene optical-sar dataset with insights for multi-modal image matching. Visual Intelligence **3**(1), 19 (2025)
61. Yu, Y., Da, F.: Phase-shifting coder: Predicting accurate orientation in oriented object detection. In: CVPR. pp. 13354–13363 (2023)
62. Yuan, M., Wang, Y., Wei, X.: Translation, scale and rotation: cross-modal alignment meets rgb-infrared vehicle detection. In: ECCV. pp. 509–525 (2022)
63. Zeng, Y., Liang, T., Jin, Y., Li, Y.: Mmi-det: Exploring multi-modal integration for visible and infrared object detection. IEEE TCSVT (2024)
64. Zhan, Y., Zeng, Z., Liu, H., Tan, X., Tian, Y.: Mambasod: Dual mamba-driven cross-modal fusion network for rgb-d salient object detection. Neurocomputing **631**, 129718 (2025)
65. Zhang, C., Feng, Y., Hu, L., Tapete, D., Pan, L., Liang, Z., Cigna, F., Yue, P.: A domain adaptation neural network for change detection with heterogeneous optical and sar remote sensing images. IJAEOG **109**, 102769 (2022)
66. Zhang, J., Cao, M., Yang, X., Xie, W., Lei, J., Li, D., Yang, G., Huang, W., Li, Y.: E2e-mfd: Towards end-to-end synchronous multimodal fusion detection. In: NeurIPS (2024)
67. Zhang, P., Xu, H., Tian, T., Gao, P., Li, L., Zhao, T., Zhang, N., Tian, J.: Sefepnet: Scale expansion and feature enhancement pyramid network for sar aircraft detection with small sample dataset. IEEE JSTARS **15**, 3365–3375 (2022)
68. Zhang, S., Chi, C., Yao, Y., Lei, Z., Li, S.Z.: Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In: CVPR. pp. 9759–9768 (2020)
69. Zhang, T., Zhang, X., Li, J., Xu, X., Wang, B., Zhan, X., Xu, Y., Ke, X., Zeng, T., Su, H., et al.: Sar ship detection dataset (ssdd): Official release and comprehensive data analysis. RS **13**(18), 3690 (2021)
70. Zhang, W., Zhang, X., Xu, X., Wei, S., Shi, J., Wang, Y., Zeng, T.: Mhfnet: Multimodal hybrid fusion framework for misaligned sar-optical ship detection. ISPRS **231**, 151–166 (2026)
71. Zhang, X., Yang, X., Li, Y., Yang, J., Cheng, M.M., Li, X.: Rsar: Restricted state angle resolver and rotated sar benchmark. In: CVPR. pp. 7416–7426 (2025)
72. Zou, Z., Shi, Z.: Random access memories: A new paradigm for target detection in high resolution aerial remote sensing images. IEEE TIP **27**(3), 1100–1111 (2017)