Resolving Complex Backgrounds and Multi-scale Challenges: The ST-FSFF Approach

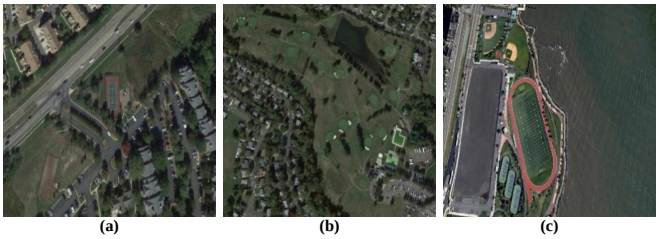
Yanling Li, Jiaman Li, Tianyu Zhao, Qingqi Liang, Zhipeng Yang, and Chongyang Chen

***Abstract*—Accurate identification of objects in remote sensing images is essential for both civilian surveillance and military defense. The primary challenges of remote sensing object detection lie in (1) complicated environmental contexts and (2) multi-scale object distribution, which hinder accurate recognition and localization of objects. Although Convolutional Neural Networks (CNN) have been extensively adopted for remote sensing detection tasks, their performance remains constrained by limited receptive fields and insufficient global contextual awareness. To overcome these limitations, our study presents an advanced ST-FSFF architecture to overcome these constraints. Specifically, the architecture incorporates a Swin Transformer-based backbone for comprehensive feature extraction, coupled with our newly developed Full-Scale Feature Fusion mechanism that optimally combines multi-level features to enhance both semantic understanding and positional accuracy. A specialized Region Proposal Network scans these multi-scale feature maps to identify potential targets, followed by a detection module for final classification and positioning. Experimental evaluations on the MAR20, NWPU VHR-10, and RSOD benchmark datasets demonstrate the superior performance of our method, with accuracies of 91.5%, 94.3%, and 97.3%, respectively, which outperform mainstream detection methods.**

***Index Terms*—Remote sensing, Deep learning, Object detection, Swin Transformer, Feature fusion.**

I. Introduction

REMOTE sensing images are a vital source of data, playing a crucial role in both civilian and military fields. The tasks of identifying and localizing targets in remote sensing images demand highly accurate detection algorithms. Recent studies highlight the dominance of Convolutional Neural Networks (CNN) in remote sensing object detection, driven by their ability to model complex feature hierarchies. Guo et al. [1] presented a CNN-based solution for multi-scale object detection. Their unified framework combines features from different resolutions to handle size variations in remote sensing data. FMSSD [2] adopted a spatial pyramid structure with dilated convolutions arranged in parallel. While this effectively expands the receptive field, it significantly increases computational demands. Both methods introduce targeted improvements for addressing the multi-scale characteristics of RSI and achieve outstanding performance. However, they lack consideration of the contextual information surrounding the targets. To tackle dense small object detection in remote sensing imagery, Wu et al. [3] developed BCS-YOLOv8s, an enhanced version of YOLOv8s specifically optimized for clustered small target scenarios. For remote sensing image analysis, Yan et al. [4] developed a deformable R-CNN variant that adaptively utilizes IoU information, enhancing detection performance particularly for small objects in multi-class scenarios. Li [5] systematically improves the YOLOv5 network architecture by introducing a PConv-based C3-Faster lightweight module, Ghost convolution using feature reparameterization, and Squeeze-and-Excitation attention mechanism for systematic lightweighting.



(a) (b) (c)

Fig. 1. Characteristics of remote sensing images (a) and (b) complex background environments (c) scale variations.

However, background interference and scale variation in remote sensing images make object detection highly challenging [6]. (1) Complex background environments, as shown in Fig. 1 (a) and (b), remote sensing images often cover vast geographic areas, where complex scenes and backgrounds frequently resemble the foreground, leading to interference with foreground detection. (2) Scale variations, as illustrated in Fig. 1 (c), and significant differences in object scales exist within the same remote sensing scene, making it challenging to balance the detection of objects at varying scales.

Feature extraction efficiency is significantly enhanced in Swin Transformer through its hierarchical structure and shifted-window attention mechanism, which optimally models interactions between local and global feature representations. Wang et al. [7] presented a technique for integrating CNN features into ViT, enhancing its ability to learn both global context and local details, thereby boosting classifi-

Manuscript received February 3, 2025; revised July 20, 2025. This study was partially supported by the Excellent Course Program for Postgraduate Education of Henan Province, China (YJS2025JC30); the Science and Technology Research Project of Henan Province, China (252102211025); and the Key Research and Development Project of Henan Province, China (241111212200).

Yanling Li is a professor at the School of Computer and Information Technology, Xinyang Normal University, Xinyang 46400, Henan, China. (e-mail: ly175@163.com).

Jiaman Li is a postgraduate student at the School of Computer and Information Technology, Xinyang Normal University, Xinyang 46400, Henan, China. (corresponding author to provide phone: +8613673367327; e-mail: jiaman0813@163.com).

Tianyu Zhao is a postgraduate student at the School of Computer and Information Technology, Xinyang Normal University, Xinyang 46400, Henan, China. (e-mail: zty0201520@163.com).

Qingqi Liang is a postgraduate student at the School of Computer and Information Technology, Xinyang Normal University, Xinyang 46400, Henan, China. (e-mail: qqliang2022@126.com).

Zhipeng Yang is a teacher at the School of Computer and Information Technology, Xinyang Normal University, Xinyang 46400, Henan, China. (e-mail: yangzp@xynu.edu.cn).

Chongyang Chen is a teacher at the School of Computer and Information Technology, Xinyang Normal University, Xinyang 46400, Henan, China. (e-mail: cychen@xynu.edu.cn).

cation accuracy. Liang et al. [8] developed an enhanced Swin Transformer-based framework that fuses convolutional and attention-based features to boost the detection of small-scale objects in cluttered remote sensing imagery. Xue et al. [9] designed a triple change detection framework named TCD-Net, which tackles the difficulties of change detection in remote sensing imagery by combining multi-frequency features with a full-scale Swin Transformer architecture. This method enhances feature representation in dynamic regions and enables cross-scale context modeling, thereby mitigating the inherent shortcomings of conventional detectors when used in remote sensing applications.

Multi-scale feature fusion empowers the model to simultaneously capture target characteristics across diverse scales, thereby significantly enhancing detection accuracy in complex backgrounds and for objects of varying sizes. By hierarchically integrating feature representations from multiple levels, this approach amplifies the model's ability to discern small targets, distant objects, and fine-grained details through cross-resolution information propagation. Such capability ensures robust adaptability to scale variations inherent in remote sensing scenarios. Zhang et al. [10] improved the network's ability to extract and characterise target features by creating a two-layer digital semi-synthetic backbone network structure and introducing deformable convolutions, coordinate attention mechanisms and a new CAB module. Zhao et al. [11] introduced an attention-guided fusion module within a YOLOX-based framework, aiming to enhance contextual perception by enlarging the receptive field across multiple feature scales. Wang et al. [12]enhanced the GD fusion module using a PConv-based FasterNet Block for better spatial feature extraction, and incorporated EMA attention to improve detection accuracy. Hou et al. [13] developed an enhanced detection method for remote sensing imagery based on the YOLOv9 framework. By integrating the C3, CD, and CGA modules, the DSConvRepNCSPELAN4 module, and the CARAFE module, the model achieves significant performance enhancements in detecting objects under complex backgrounds and multi-scale variations. Gong et al. [14] developed SGMFNet, an innovative architecture for remote sensing detection that synergistically combines spatial global attention with hierarchical multi-scale feature integration. This method effectively improves detection performance under complex backgrounds and significant scale variation conditions. Liu et al. [15] designed an improved detection approach for remote sensing images that leverages attention mechanisms and multi-scale feature integration. Built on a refined Faster R-CNN design, the proposed method shows improved performance in detecting small targets under complex scene conditions. Zhang et al. [16] developed SGMFNet, a self-attention-driven detection model with multi-scale feature integration, tailored to overcome background complexity, scale variance, and object crowding in remote sensing images.

Based on the observations above, we use Swin Transformer to extract global features and solve the issue of complex background interference in images. Its hierarchical design and shifted window mechanism preserve local detail features while gradually building global feature representations. This approach overcomes the limitation of traditional CNN with restricted receptive fields. A network architecture based on full-scale feature fusion (FSFF) is proposed to address the problem of significant scale differences. The model combines features from various depths to strengthen its capacity for representing targets at different spatial scales. The FSFF achieves richer semantic and more precise spatial multi-scale feature representations through a cross-scale information interaction mechanism. The design of this fusion method is based on two main motivations: (1) Compared to single-scale features, multi-scale features can effectively cover targets of different sizes. Shallow-layer representations, characterized by high-resolution feature maps and limited receptive fields, are more effective in identifying small-scale targets. In contrast, deeper-layer features, with coarser spatial resolution and broader receptive coverage, are advantageous for the recognition of large objects. Therefore, fusing multi-scale features enables a more comprehensive detection of objects of various sizes. (2) Shallow layers preserve precise spatial localization yet lack high-level semantics, whereas deeper layers encode robust semantic concepts at the cost of reduced positional accuracy. By fusing features of different scales, the semantic information from deep networks can be fully utilized while preserving the spatial sensitivity of shallow networks. This method strengthens the localization ability of shallow features while simultaneously improving the semantic representation of deep features, thereby significantly boosting the model's detection performance across targets of varying scales.

Based on the Swin Transformer and FSFF, we propose an innovative network architecture, ST-FSFF, aimed at addressing the interference from cluttered environments and varying target scales in remote sensing object detection through an efficient feature fusion strategy. This study makes the following contributions:

(1) In our framework, global features are obtained using the Swin Transformer backbone, which employs self-attention to model spatial dependencies over extended image regions. This helps address the challenge of complex background interference.

(2) The FSFF is proposed, which can efficiently integrate target information at different scales to generate rich multi-scale feature representations. It substantially enhances both the precision and stability of target detection in remote sensing imagery.

(3) Based on Swin Transformer and FSFF, we constructed an innovative ST-FSFF network for solving the complex background problem and target scale inconsistency problem, which enhances the feature extraction capability of target information.

II. RELATED WORKS

A. Traditional Object Detection Algorithms

Prior to deep learning, object detection largely utilized manually engineered descriptors, such as Scale-Invariant Feature Transform (SIFT) and Speeded-Up Robust Features. These methods describe object characteristics by focusing on localized visual descriptors within the image. The discriminative features extracted from images serve as inputs to machine learning classifiers, including support vector machines (SVM) and random forest algorithms. For instance, Liu et al. [17] employed the Canny algorithm for robust edge

feature extraction, leveraging its multi-stage gradient computation framework. Thereafter, the Hough transform was applied to detect line segments, enabling the determination of runway positions and the detection and localization of target positions. Tao et al. [18] improved the SIFT feature descriptor and incorporated prior knowledge to determine the location of objects in remote sensing images. Yao et al. [19] analyzed the limitations of existing pixel-wise methods and used the Hough Transform to determine the presence of potential airports. Subsequently, they extracted SIFT features from candidate regions based on salient region detection and classified them to determine the location of target areas. Therefore, Xu et al. [20] constructed a ship-shaped template set derived from the Hough Transform technique. They employed a sliding window strategy on real images to evaluate the correlation between each candidate region and predefined shape characteristics, enabling the identification of potential target occurrences. Zhang et al. [21] adopted a multi-scale sliding window strategy to produce candidate regions with varying dimensions and aspect ratios. Visual features were then extracted from each region, and a cascaded SVM classifier was applied to assign confidence scores, facilitating the identification of potential targets.

B. Deep Learning-Based Object Detection Algorithms

The proliferation of deep learning has established convolutional neural networks (CNN) as a predominant paradigm for object detection in remote sensing imagery, driving significant research advancements in this domain. Sun et al. [22] proposed an improvement scheme based on the YOLOv7 model, which enhances the model's ability to detect specific targets by introducing a CBAM attention mechanism, a small target detection layer, and a CoordConv module. This architecture explicitly models three critical characteristics of remote sensing targets: (1) non-uniform spatial distribution, (2) multi-scale presence , and (3) arbitrary orientation variations, consequently achieving state-of-the-art detection accuracy. Dong et al. [23] proposed a remote sensing object detection framework based on the Receptive Field Expansion Block (RFEB). By implementing RFEB modules on top of the Feature Pyramid Network structure, the system gains the ability to dynamically modify receptive fields. This adaptation improves both contextual understanding and the flow of deep semantic characteristics across different pyramid levels, resulting in better multi-scale object detection capabilities. Li et al. [24] designed the Adjacent Context Collaborative Network. Detection performance for salient objects is significantly enhanced through the integration of the Adjacent Context Collaborative Module within the encoder-decoder architecture, enabling more precise feature extraction and localization. A parameter-free mask was proposed in [25] to distinguish foreground instances from background in hierarchical features. The interference from complex backgrounds is alleviated in LSKNet [26] through optimized spatial receptive fields derived from prior knowledge, particularly improving small object localization. The coordinate attention in [27] enhanced both localization and classification precision. Hu et al. [28] developed an attention-based multi-scale network for ship detection in complex scenes. A single-stage detection framework was developed by Ma et al. [29], where features of interest are enhanced through saliency-based amplification and scale-aware associations.

C. Feature Fusion

Feature fusion integrates multi-source or cross-layer feature representations to produce enriched feature embeddings with enhanced discriminative capabilities. This process improves the model's representational capacity, enabling it to capture more valuable information in complex scenes or multi-scale objects, thereby improving object detection accuracy and robustness. Zhao's adaptation [30] equips YOLOX with scale-aware attention modules that dynamically adjust to capture both local details and global scene context. Song et al. [31] incorporated an Adaptive Instance Normalization block during feature fusion to enhance cross-domain adaptability, significantly improving detection model robustness. The framework further integrates an attention mechanism to refine localization of fine-grained patterns. Effective aggregation of cross-level features was achieved by Wang et al. [32] through a novel fusion module with channel-wise attention mechanisms. Zhao et al. [33] introduced an attention-based fusion mechanism tailored for aircraft detection, which enhances both texture details and semantic representations through deep integration. Chen et al. [34] developed a fusion framework that facilitates spatial–semantic interaction via efficient bidirectional coupling and adaptive weighting strategies.

III. Method

As demonstrated in Fig. 2, We present an innovative remote sensing image object detection network named ST-FSFF, which consists of four main components: (1) Backbone network (Swin Transformer), (2) FSFF, (3) Region Proposal Network (RPN), and (4) Detection Network. First, the Swin Transformer receives the input image and extracts multi-scale feature representations through a series of convolutional operations, effectively capturing both local and global information within the image. Second, the FSFF module fully integrates semantic and spatial information at different scales, generating feature representations across various scales. Subsequently, the RPN automatically generates high-quality candidate regions (i.e., object-bounding boxes). The multi-scale feature maps, enriched through feature extraction and fusion, are fed into the RPN module. Using the Region of Interest (RoI) Align operation, the RPN precisely extracts candidate object regions from the feature maps and generates the corresponding class and bounding box information. Ultimately, feature maps refined via RoIAlign are input into the detection head's fully connected layers for classification and box regression, resulting in final predictions. The backbone and FSFF module are elaborated upon in the following parts.

A. Swin Transformer

To mitigate feature corruption caused by complex backgrounds in remote sensing imagery, our framework adopts Swin Transformer's hierarchical attention mechanism as the foundational feature extractor. Integrating self-attention with hierarchical representation learning, the Swin Transformer synergistically captures both fine-grained local patterns and comprehensive global contexts through its shifted-window

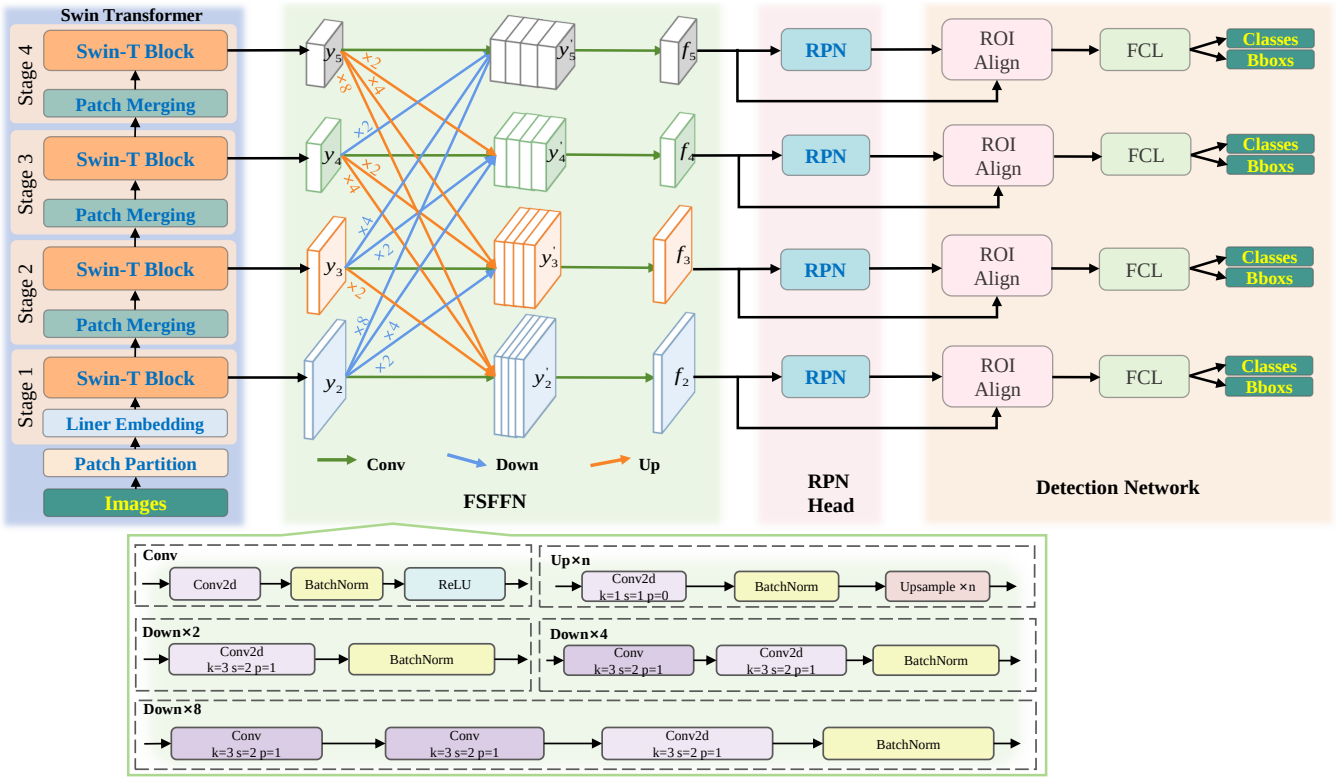


Fig. 2. ST-FSFF Network Architecture Diagram.

architecture. This approach expands the network's receptive field, enhances efficiency, and improves the capture of global and contextual information. These improvements lead to better object detection performance.

The Swin Transformer adopts a hierarchical architecture consisting of four progressive stages, illustrated on the left in Fig. 2. Initially, the RGB image is partitioned into distinct, non-overlapping patches, which are then processed by the patch splitting unit and mapped into embedding representations. The resulting patch embeddings are subsequently forwarded into the Swin Transformer block to perform feature encoding. In stages 2 through 4, the patch tokens are initially fed into the patch merging operation, where neighboring features are aggregated to produce lower-dimensional representations before being processed by the subsequent Swin Transformer blocks. This hierarchical design allows the network to generate feature maps at resolutions comparable to those of standard convolutional architectures.

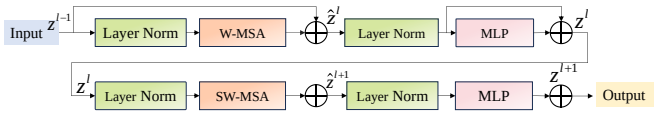


Fig. 3. Two Successive Swin Transformer Blocks.

Fig. 3 shows the structure of two consecutive Swin Transformer Blocks [35]. A typical Swin Transformer Block integrates several key components, including Layer Normalization (LayerNorm), window-based and shifted window-based multi-head self-attention mechanisms (W-MSA and SW-MSA), residual pathways, and a multi-layer perceptron (MLP). The shifted window mechanism serves as a pivotal innovation in the Swin Transformer design. The Swin Transformer extracts features across four sequential stages. Prior to this, the patch partition unit segments the input image into smaller regions. At the initial stage, each patch is linearly embedded to a 96-channel representation before being processed by a Swin Transformer Block. From stage two onward, the architecture incorporates patch merging operations followed by successive Swin Transformer Blocks to build hierarchical features.

To construct hierarchical representations, the patch merging operation reduces the spatial resolution of feature maps while increasing the number of channels before each Swin Transformer Block. This approach promotes multi-scale feature modeling by condensing spatial dimensions. At the core of Swin Transformer computations lies a multi-head self-attention scheme, which enables the modeling of long-range interactions by attending to spatial relationships across positions. Within the block structure, two attention mechanisms—window-based (W-MSA) and shifted-window (SW-MSA)—are alternated between layers. W-MSA limits the attention scope to fixed, non-overlapping windows, reducing computational cost, while SW-MSA introduces relative spatial shifts between adjacent windows to facilitate inter-window information exchange. In contrast, traditional global attention mechanisms, like MSA, process the entire image but incur higher computational overhead. Through this hierarchical and localized attention design, Swin Transformer enhances its capacity to encode contextual dependencies effectively.

Leveraging the window partitioning strategy, the feature transformation between adjacent Swin Transformer Blocks (layer l to $l+1$) follows:

$$\hat{z}^l = W - MSA(LN(z^{l-1})) + z^{l-1}, \tag{1}$$

$$z^l = MLP(LN(\hat{z}^l)) + \hat{z}^l, \tag{2}$$

$$\hat{z}^{l+1} = SW - MSA(LN(z^l)) + z^l, \tag{3}$$

$$z^{l+1} = MLP(LN(\hat{z}^{l+1})) + \hat{z}^{l+1}. \tag{4}$$

where z^l represents the output of the MLP and $\hat{z}^l$ denotes the output features of the W-MSA module.

In this mechanism, attention is calculated only within each window and not across windows. The formulas for computing attention scores in W-MSA and SW-MSA are as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_K}})V. \tag{5}$$

where Q, K, and V represent the query, key, and value matrices, respectively. These matrices are obtained through linear transformations of the input matrix after the LayerNorm layer, and all three have a dimension of d_k.

This study adopts the Swin Transformer architecture, which is divided into four stages comprising a total of 12 Swin Transformer blocks. The number of blocks in stages 1 to 4 is 2, 2, 6, and 2, respectively.

B. Full-Scale Feature Fusion

In deep learning approaches, multi-scale features are typically employed to capture objects of varying scales. Low-resolution features are more sensitive to larger objects, while high-resolution features are more effective for smaller ones. Therefore, constructing multi-scale feature representations is crucial.

To address the challenge of cross-scale defect characterization, we introduce Full-Scale Feature Fusion (FSFF), a computationally efficient architecture that hierarchically combines defect signatures across multiple magnification levels. As shown in Fig. 2, FSFF can be divided into three steps. First, remote sensing images of size $H \times W \times C$ are input into the ST-FSFF network to generate multi-scale feature maps $\{y_i, i = 2, 3, 4, 5\}$, where y_i represents the feature map generated by stage i, and i denotes the stage index.

Second, multi-scale feature fusion fully exchanges feature information at different scales. We take the calculation of the fused feature y_3' as an example. The input is composed of four feature maps $\{y_i\}_{i=2}^5$, and the output y_3' is the sum of the four input features after transformation. Among them, it y_2 undergoes a two-fold down-sampling operation, reducing the scale and increasing the channel dimension to ensure that its scale and channel dimension are the same as those of f_3. y_3 is processed by a Conv module, which includes a 1×1 convolution, batch normalization, and a ReLU activation function. For y_4 and y_5, the network uses two-fold and four-fold up-sampling operations, respectively, to enlarge and compress the channel dimension. Then, the four processed feature maps are concatenated to obtain the fused feature y_3'. Similar to the calculation of y_3', the calculations of y_2', y_4', and y_5' can be easily deduced formally:

$$\begin{cases} y_2' = C(y_2) + Up(y_3) + Up(y_4) + Up(y_5), \\ y_3' = C(y_3) + Down(y_2) + Up(y_4) + Up(y_5), \\ y_4' = C(y_4) + Down(y_2) + Down(y_3) + Up(y_5), \\ y_5' = C(y_5) + Down(y_2) + Down(y_3) + Down(y_4). \end{cases} \tag{6}$$

where $C(\cdot)$, $Up(\cdot)$, $Down(\cdot)$ represent the convolution operation, the up-sampling operation, and the down-sampling operation, respectively.

Finally, the FSFF applies the Conv operation to obtain multi-scale feature representations $\{f_i\}_{i=2}^5$, formally:

$$f_i = C(y_i'), i = 2, 3, 4, 5. \tag{7}$$

IV. Experiments

This section outlines the experimental configuration, covering the datasets involved, assessment indicators, and technical implementation specifics. We then discuss the evaluation criteria for the object detection and classification tasks in this study, analyzing the training and testing results of the model using these metrics.

A. Datasets

In our experimental analysis, we adopt three widely recognized datasets: MAR20, NWPU VHR-10, and RSOD. The MAR20 dataset is a comprehensive benchmark tailored for military aircraft recognition in high-resolution remote sensing imagery. It comprises 3,842 images (800×800 pixels) gathered from 60 military airport locations distributed across countries such as the United States and Russia. Sourced via Google Earth, it includes 22,341 annotated objects categorized into 20 distinct aircraft classes, such as SU-35, C-130, and TU-160, labeled from A1 to A20. The NWPU VHR-10 dataset is a widely used public benchmark for object detection in the remote sensing domain. It provides 800 high-resolution optical images containing 10 categories of everyday objects, including airplane (A), baseball diamond (BD), basketball court (BC), bridge (B), ground track field (GTF), harbor (H), ship (S), storage tank (ST), tennis court (TC) and vehicle (V). Meanwhile, the RSOD dataset, published by Wuhan University in 2017, contains 976 labeled remote sensing images obtained from Google Earth and Skymap. It covers four major object types: aircraft, oil tanks, overpasses, and playgrounds, encompassing a total of 6,950 annotated targets. The three datasets with diverse scenes and varying target scales and background complexity provide challenging test environments for the algorithms.

To partition the data, we adhered to the standard split scheme defined by each dataset. Specifically, 70% of the images were allocated for model training to enable comprehensive feature learning across varied scenarios. Another 20% were assigned for testing, facilitating a robust evaluation of the model's generalization ability. The remaining 10% were utilized for validation purposes, allowing for performance monitoring and fine-tuning during training to mitigate overfitting and optimize learning.

B. Evaluation Metrics and Experimental Setup

In object detection, objects in an image may vary in location and category. This requires evaluating both the classification and localization performance of the model. Standard evaluation metrics for image classification are not directly applicable to object detection tasks. Thus, we employ the mean Average Precision (mAP), a widely accepted metric in object detection. The evaluation includes precision (P), which quantifies the fraction of correct positive detections,

TABLE I
MAR20 DATASET PERFORMANCE EVALUATION

Methods	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	A20	mAp
Yolov5	85.4%	81.5%	87.6%	78.3%	80.5%	90.5%	90.2%	87.5%	87.9%	90.8%	85.8%	**89.2**%	67.2%	88.2%	47.8%	89.1%	90.5%	74.5%	81.3%	80.0%	82.7%
Yolov7[36]	85.8%	81.5%	86.8%	76.3%	72.2%	89.9%	89.8%	89.4%	89.1%	90.7%	86.2%	87.4%	64.9%	88.3%	47.0%	87.8%	90.4%	64.9%	83.9%	76.8%	81.5%
Yolov8	86.1%	81.7%	88.1%	69.6%	75.6%	89.9%	90.5%	89.5%	89.8%	90.9%	87.6%	88.4%	67.5%	88.5%	46.3%	88.2%	90.5%	70.5%	78.7%	80.2%	81.9%
Yolov10	85.0%	83.6%	87.4%	70.6%	79.6%	90.6%	89.7%	89.8%	90.4%	90.8%	85.5%	88.1%	68.4%	88.3%	42.4%	88.9%	90.5%	62.3%	78.2%	77.7%	81.4%
Yolov11	86.3%	80.8%	88.9%	82.5%	76.0%	90.1%	89.8%	87.3%	89.2%	90.8%	84.3%	86.2%	65.7%	87.4%	44.1%	87.5%	90.3%	56.4%	82.8%	76.7%	81.2%
DINO[37]	**87.0**%	82.0%	88.5%	70.0%	76.0%	90.2%	90.5%	87.0%	88.5%	90.7%	88.0%	88.5%	68.0%	89.0%	48.0%	88.5%	90.3%	70.0%	79.0%	81.0%	82.0%
Faster-RCNN	85.5%	82.2%	87.0%	78.0%	80.0%	90.3%	90.1%	89.0%	87.5%	90.6%	85.0%	89.0%	67.0%	88.0%	47.5%	89.0%	90.2%	73.0%	81.5%	80.5%	82.5%
ST-FSFF (ours)	86.2%	**90.2**%	**94.9**%	**95.7**%	**86.6**%	**95.7**%	**92.1**%	**93.9**%	**94.4**%	**99.2**%	**89.8**%	87.8%	**76.4**%	**96.1**%	**86.6**%	**97.5**%	**98.0**%	**92.1**%	**86.2**%	**90.0**%	**91.5**%

TABLE II
NWPU VHR-10 DATASET PERFORMANCE EVALUATION

Methods	A	BD	BC	B	GTE	H	S	ST	TC	V	mAp
Yolov5	16.9%	65.6%	56.5%	72.8%	57.7%	41.5%	60.2%	84.4%	2.3%	91.7%	54.9%
Yolov7[36]	35.4%	58.6%	65.8%	83.6%	60.4%	51.4%	56.8%	83.4%	3.4%	95.8%	59.7%
Yolov8	27.4%	55.6%	63.8%	83.0%	62.4%	40.4%	55.8%	85.4%	3.0%	95.2%	57.2%
Yolov10	53.6%	66.6%	88.1%	72.4%	70.5%	40.1%	62.6%	84.2%	2.1%	93.2%	63.4%
Yolov11	52.2%	91.0%	92.9%	**90.7**%	84.3%	44.1%	67.5%	91.9%	3.5%	95%	71.3%
DINO[37]	97.5%	96.4%	88.2%	94.6%	**99.9**%	91.0%	86.9%	**92.7**%	86.8%	82.5%	91.7%
Faster-RCNN	93.5%	96.1%	85.3%	80.8%	96.8%	91.0%	85.0%	84.3%	**95.4**%	94.4%	90.3%
ST-FSFF (ours)	**99.9**%	**98.8**%	**95.2**%	87.7%	**99.9**%	**98.7**%	**89.0**%	85.0%	93.0%	**96.2**%	**94.3**%

TABLE III
RSOD DATASET PERFORMANCE EVALUATION

Methods	aircraft	oil tank	overpass	playground	mAP
Yolov5	70.8%	90.2%	78.7%	98.1%	84.5%
Yolov7[36]	71.7%	90.3%	81.0%	**99.8**%	85.7%
Yolov8	76.1%	90.3%	81.3%	98.8%	86.6%
Yolov10	80.3%	89.6%	88.7%	99.1%	89.4%
Yolov11	94.5%	96.0%	84.6%	92.5%	91.9%
DINO[37]	95.4%	95.8%	74.9%	83.2%	87.3%
Faster R - CNN	95.7%	97.7%	90.1%	95.1%	94.7%
ST - FSFF (ours)	**96.7**%	**99.4**%	**93.6**%	99.4%	**97.3**%

and recall (R), indicating the proportion of actual positives correctly retrieved. These values are computed using true positives (TP), false positives (FP), and false negatives (FN), based on validation and test results. A TP refers to a prediction that overlaps sufficiently with a ground truth box. FP arises when a predicted box exists where no ground truth is present, while FN indicates missed detections of actual objects. The equations for these evaluation metrics are presented as follows:

$$P = \frac{TP}{TP + FP}, \tag{8}$$

$$R = \frac{TP}{TP + FN}, \tag{9}$$

$$AP = \int_0^1 P(R)d(R), \tag{10}$$

$$mAP = \frac{\sum_{i=1}^{n} AP_i}{n}. \tag{11}$$

C. Comparative Experiments

To assess its capabilities, our method is evaluated in comparison with cutting-edge deep learning algorithms. Experiments were conducted on the MAR20, NWPU VHR-10, and RSOD datasets, and the results are presented in Table I-Table III.

As shown in Table I, our method achieves a mAP of 91.5%, surpassing all baseline methods by a significant margin (e.g., +8.8% over YOLOv5 and +9.0% over Faster R-CNN). The results substantiate ST-FSFF's overall advantage in processing varied object categories and complex scenes, exhibiting significantly superior performance over comparative models across most categories (particularly A3-A11 and A13-A20). Slightly lower than the DINO method in categories A1 and A12. The reason is that these targets usually have obvious local saliency and more homogeneous background structure, which is suitable for recognition by local feature-dominated detection methods based on the dense anchor frame strategy (e.g., DINO). In contrast, ST-FSFF places greater emphasis on enhancing feature representation through multi-scale semantic fusion and contextual modeling, demonstrating superior performance when handling complex scenes and multi-scale objects. This demonstrates the overall superiority of ST-FSFF in handling diverse object categories and complex scenes. ST-FSFF exhibits remarkable improvements in detecting objects under occlusion and in small-scale scenarios. For instance, it attains 99.2% on A10 (occluded objects) and 89.8% on A11 (small objects). These results provide empirical evidence that validates the efficacy of the proposed methodology. The ST-FSFF framework employs a Swin Transformer backbone. It utilizes sliding window self-attention (SW-MSA) mechanism. This approach maintains local structural features while enhancing global context modeling. Consequently, it effectively reduces interference from cluttered environments on object recognition accuracy.

Table II summarizes the detection performance on the NWPU-VHR-10 dataset. Comparative experiments demonstrate our method's superior performance over YOLO-series approaches, particularly in complex backgrounds. For heavily occluded targets, the proposed approach achieves detection accuracies of 98.7% (H) and 89.0% (S), outperforming all baseline methods. The best results were not achieved on B, ST, or TC. Traditional detectors based on convolutional structures (e.g., Faster-RCNN) have a natural advantage in effectively extracting high-frequency information, such as edges and textures, from ST and TC, which have regular target contours, clear shapes, and less background interference. Notably, the accuracy gap on category B is minimal, suggesting that ST-FSFF effectively represents and discriminates targets with medium structural complexity. Additionally, all of the methods presented in this paper outperform other

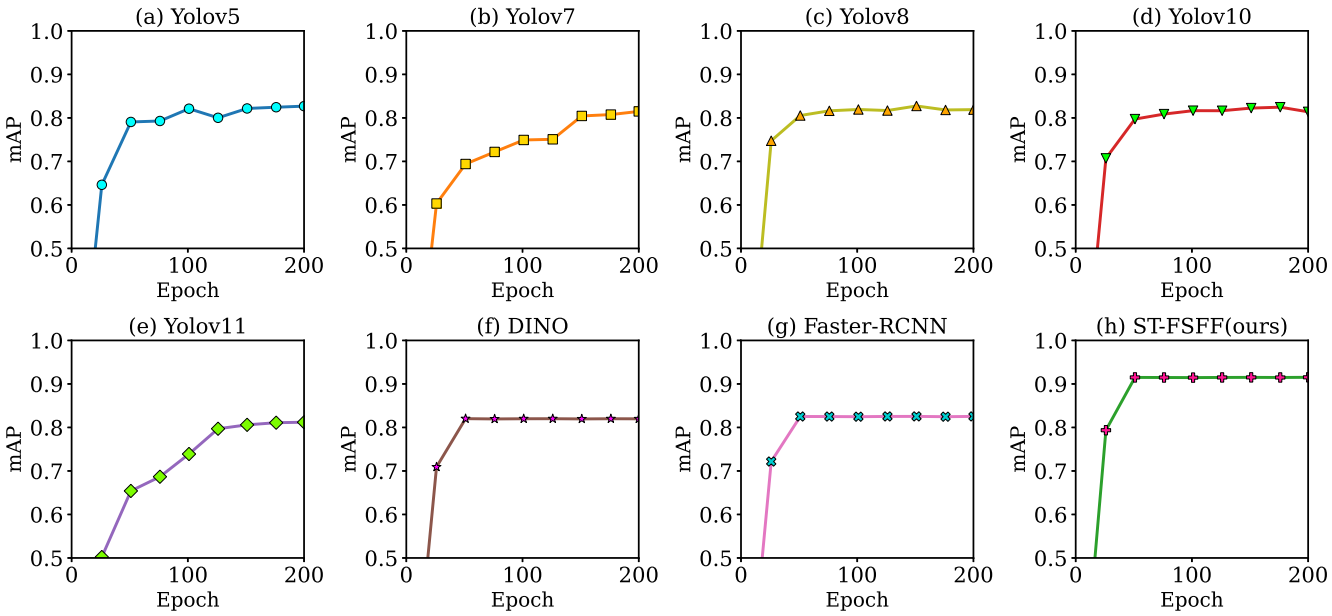


Fig. 4. Multiple methods' mAP comparison on the MAR20 dataset.

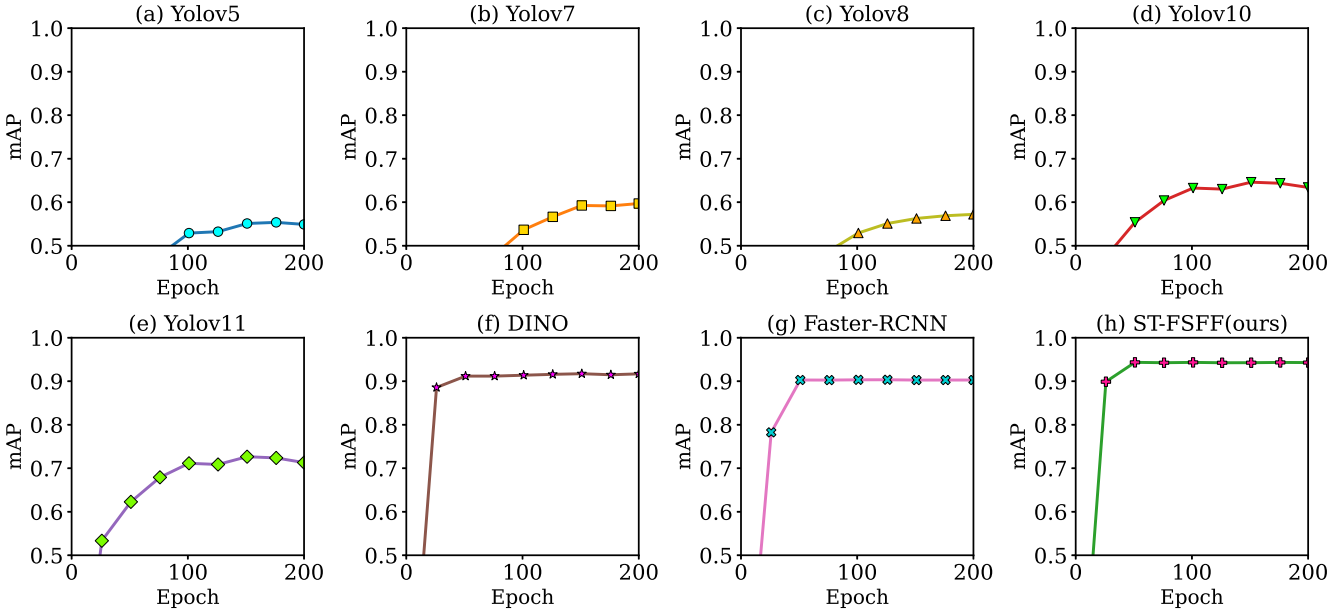


Fig. 5. Multiple methods' mAP comparison on the NWPU VHR-10 dataset.

methods, demonstrating their stability in detecting targets in complex remote sensing scenes.

Table III presents the quantitative results on the RSOD dataset, where our approach achieves a 97.3% mAP, outperforming all other comparison methods. Among them, Faster R-CNN achieves 94.3% mAP, DINO achieves 87.3% mAP, and the latest YOLOv11 achieves 91.9%, which are all lower than our method. The method in this paper is only 0.4% lower than Yolov7 in the paly ground category, a phenomenon that stems from differences in the design of the model structure. YOLOv7 excels at detecting targets with consistent appearance and limited morphological variations, leveraging high-resolution feature preservation and specialized anchor optimization strategies. In contrast, ST-FSFF incorporates global attention mechanisms with multi-scale feature fusion to significantly improve adaptability and robustness in complex scenarios.

To further demonstrate the performance trend of each model during training, we added curves to Fig. 4-Fig. 6 representing the variation in mAP for each comparison model on the three datasets, respectively. The variation curves offer more intuitive insights into the training dynamics and stability of each model than the tables, which only report the final accuracy.

To improve readability and avoid curve overlap and visual clutter caused by excessive data points, we used an equal-interval sampling strategy for visualisation. Specifically, we selected representative data points from the full training log every 25 epochs and kept the final results at the end of training. This approach strikes a balance between presenting performance trends in full and ensuring graphical distinguishability. Most object detection models (e.g., two-

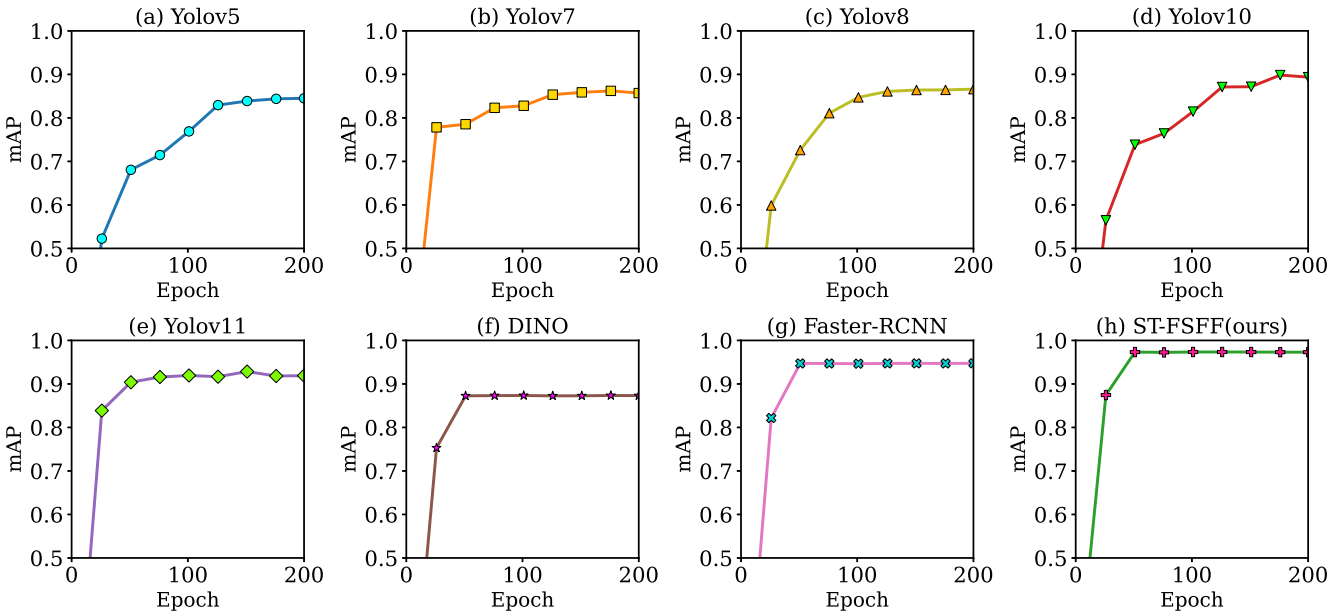


Fig. 6. Multiple methods' mAP comparison on the RSOD dataset.

stage methods) converge within 50 epochs. However, we trained all models for 200 epochs to ensure fair performance comparison. This avoids incomplete evaluation from early stopping and guarantees consistent comparison under equal training epochs. The figure clearly demonstrates that our proposed ST-FSFF model achieves significantly faster convergence than conventional methods across all datasets. Quantitatively, ST-FSFF attains superior final mAP values on all benchmark datasets. These consistent results substantiate that our approach surpasses state-of-the-art detection models in terms of both precision and training efficiency.

For comparative visualization of detection performance, representative samples from all three datasets were selected to illustrate the outputs of different algorithms, with results depicted in Fig. 7-Fig. 9. The visualization results demonstrate that ST-FSFF achieves higher detection accuracy, effectively reducing false positives and missed detections. ST-FSFF provides more precise object localization and clearer boundary identification. These visualization results are consistent with the quantitative analyses in Table I - Table III, further demonstrating the advantages of ST-FSFF in remote sensing target detection tasks.

D. Ablation Experiments

We conducted systematic module evaluation by progressively integrating the Swin Transformer and FSFF components into the Faster R-CNN baseline. The ablation study results on the MAR20, NWPU VHR-10, and RSOD datasets are presented in Table IV-Table VI, respectively.

The results of the ablation experiments show that the synergy between Swin Transformer and FSFF module significantly improves the model performance.On the three datasets of MAR20, NWPU VHR-10 and RSOD, there are obvious limitations in P, R and mAP metrics when using Swin Transformer or FSFF alone. Relying on Swin Transformer alone is difficult to effectively handle multi-scale targets, while FSFF alone is susceptible to complex background interference due to the lack of global semantic understanding.

TABLE IV
ABLATION EXPERIMENTS WITH MAR20

Swin Transformer	FSFF	Metrics P	R	mAP
✓		86.0%	86.3%	87.9%
	✓	86.4%	87.6%	89.4%
✓	✓	89.6%	90.3%	91.5%

TABLE V
ABLATION EXPERIMENTS WITH NWPU VHR-10

Swin Transformer	FSFF	Metrics P	R	mAP
✓		87.8%	88.4%	91.9%
	✓	88.4%	86.1%	92.6%
✓	✓	89.6%	90.3%	94.3%

TABLE VI
ABLATION EXPERIMENTS WITH RSOD

Swin Transformer	FSFF	Metrics P	R	mAP
✓		89.2%	88.5%	92.2%
	✓	88.9%	87%	94.1%
✓	✓	93.4%	91.3%	97.3%

From the perspective of PR curve differences in Fig. 10-Fig. 12, the combined "Swin Transformer + FSFF" curve consistently approaches the top-right corner on both MAR20 and RSOD datasets, strongly validating the synergistic model's advantage in achieving optimal balance between high recall and high precision. For the NWPU VHR-10 dataset, Fig. 11 shows that while the "Swin Transformer + FSFF (red curve)" and "FSFF" (blue curve) achieve

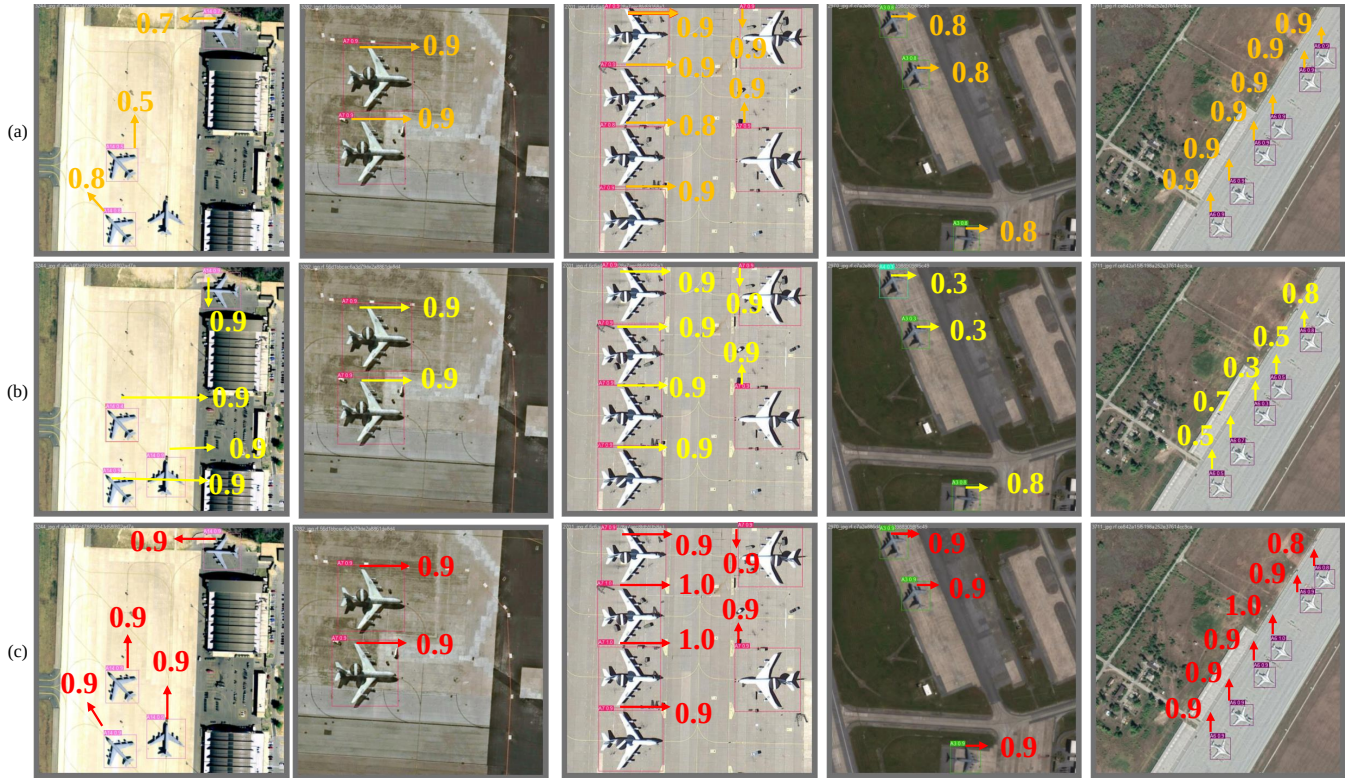


Fig. 7. Representative detection results on the MAR20 dataset using three selected methods: (a) YOLOv5, (b) Faster R-CNN, and (c) the proposed ST-FSFF.

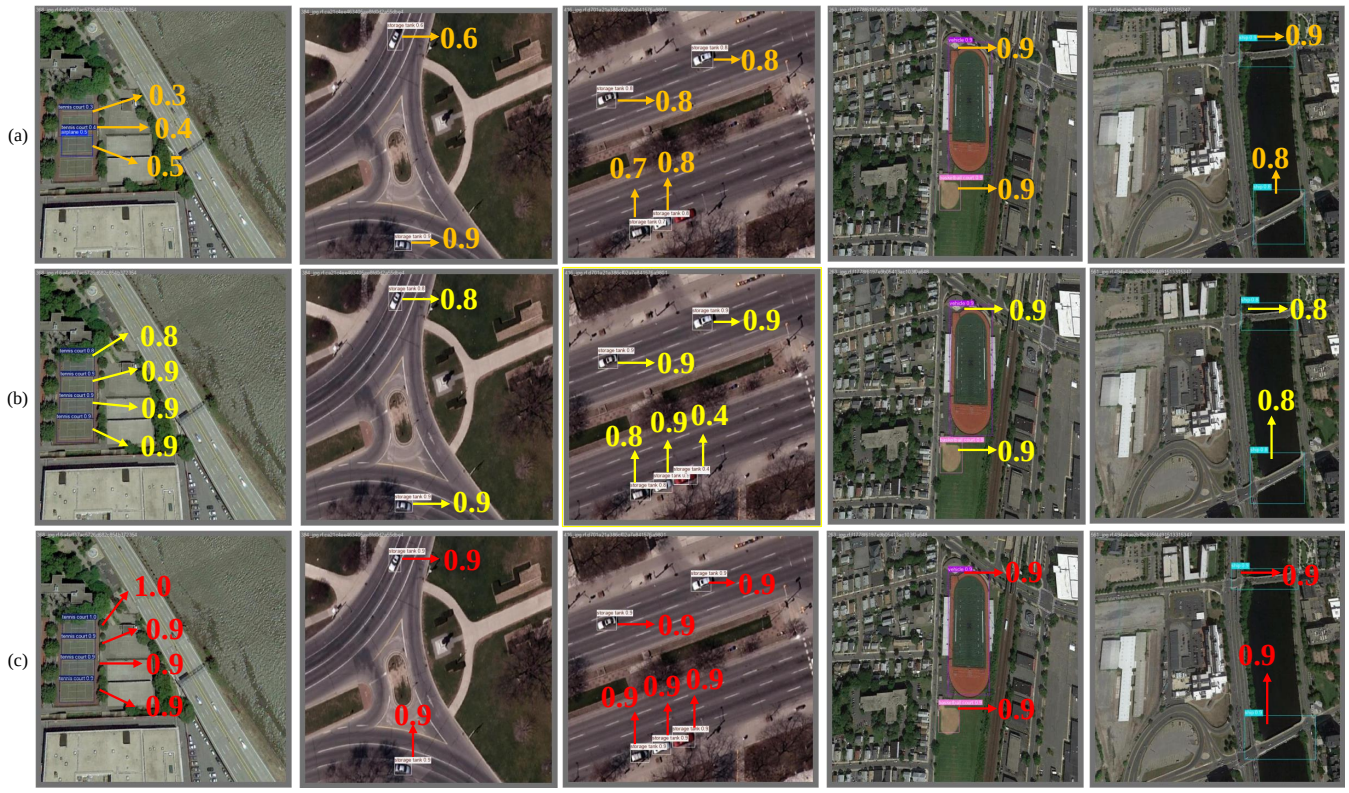


Fig. 8. Representative detection results on the NWPU VHR-10 dataset using three selected models: (a) DINO, (b) Faster R-CNN, and (c) the proposed ST-FSFF.

comparable performance in certain intervals (R = 0.5-0.9), the red curve demonstrates significantly higher precision than the blue curve in the critical high-recall region (R > 0.9). This clearly validates the advantage of the global attention mechanism in high-recall scenarios.

Analysis of the mAP curves (Fig. 13-Fig.15) demonstrates that integrating Swin Transformer with FSFF achieves more comprehensive feature representation. Swin Transformer establishes a robust global semantic foundation, while FSFF supplements multi-scale spatial details. This synergistic com-

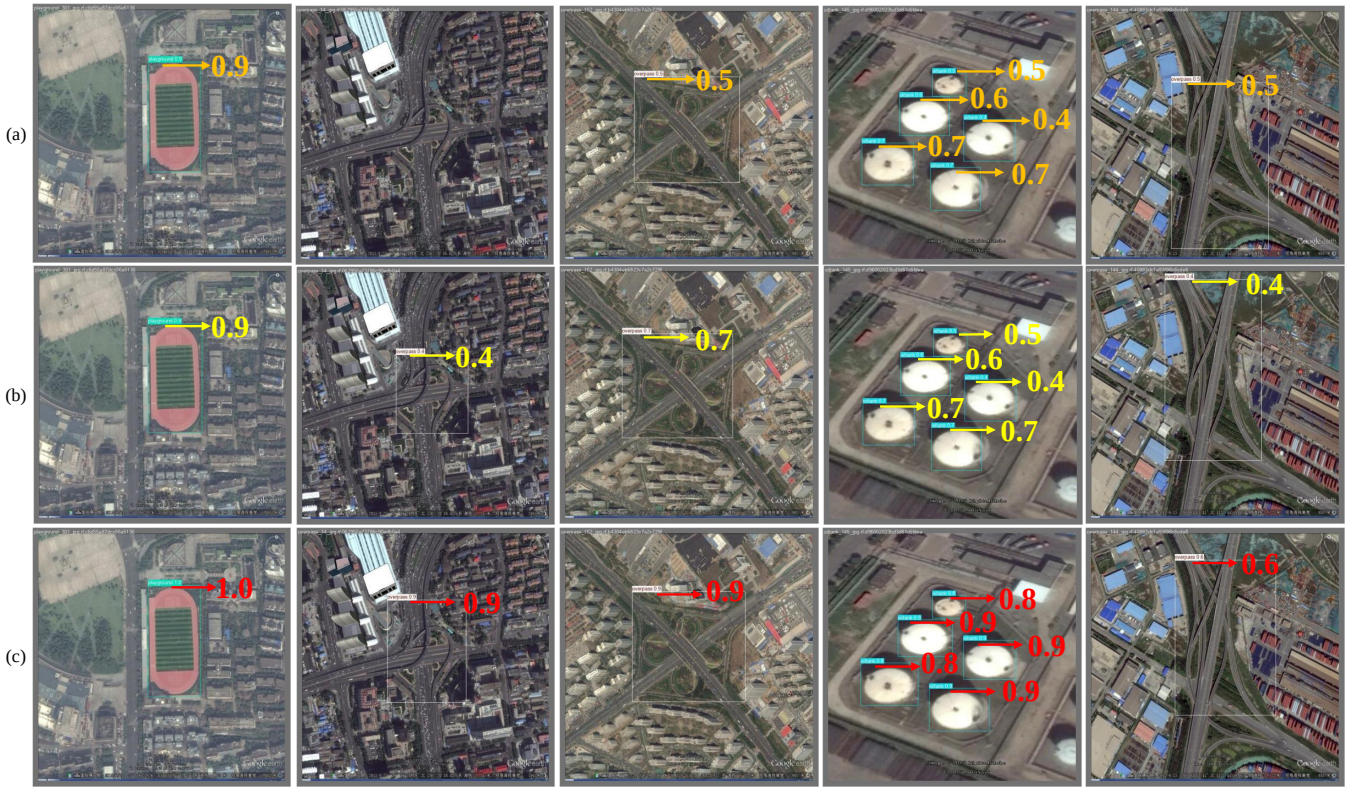


Fig. 9. Representative detection results on the RSOD dataset using three selected models: (a) YOLOv11, (b) Faster R-CNN, and (c) the proposed ST-FSFF.

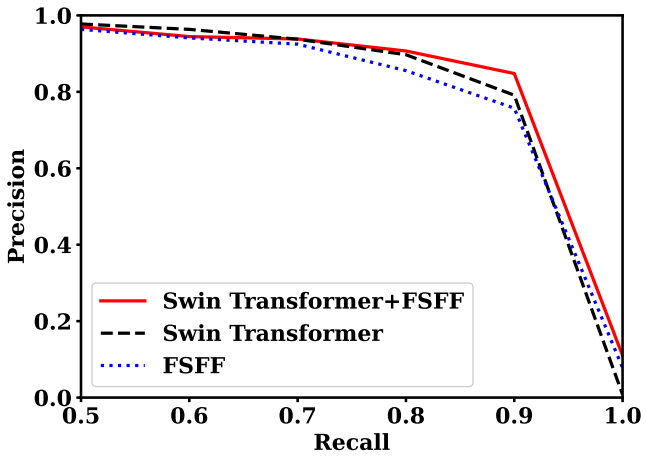


Fig. 10. PR curves comparison for module ablation study on MAR20 dataset.

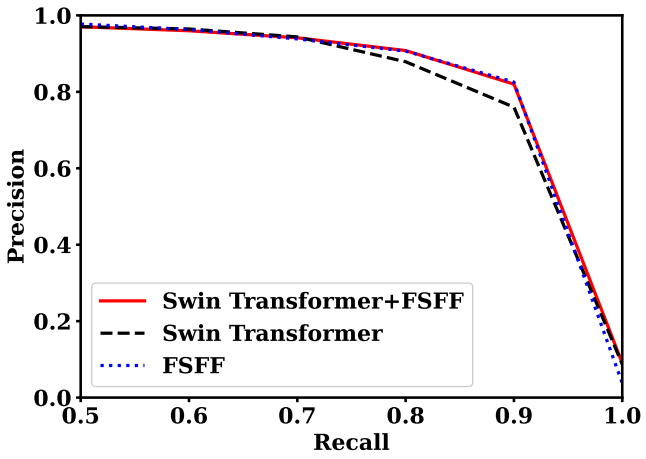


Fig. 11. PR curves comparison for module ablation study on NWPU VHR-10 dataset.

bination yields substantial mAP improvements: from 89.4% to 91.5% on MAR20, 92.6% to 94.3% on NWPU VHR-10, and 94.1% to 97.3% on RSOD. Moreover, the model demonstrates faster convergence and superior precision-recall trade-offs. These results thoroughly validate the scientific rationale and necessity of the dual-module collaboration in ST-FSFF network design.

V. Conclusion

This paper proposes ST-FSFF, a novel two-stage framework for object detection in remote sensing images. ST-FSFF integrates several key components to enhance performance. Firstly, the introduction of the Swin Transformer captures spatial and frequency domain features while extracting remote dependency features with global semantics. This effectively suppresses the interference from complex backgrounds in remote sensing images, providing more discriminative features for subsequent tasks. Secondly, to address the issue of varying object scales, the FSFF method is proposed. Empirical analysis based on multiple remote sensing benchmarks confirms both the robustness of our approach and its superior performance compared to recent state-of-the-art methods.

References

[1] W. Guo, W. Yang, H. Zhang, and G. Hua, “Geospatial object detection in high resolution satellite images based on multi-scale convolutional

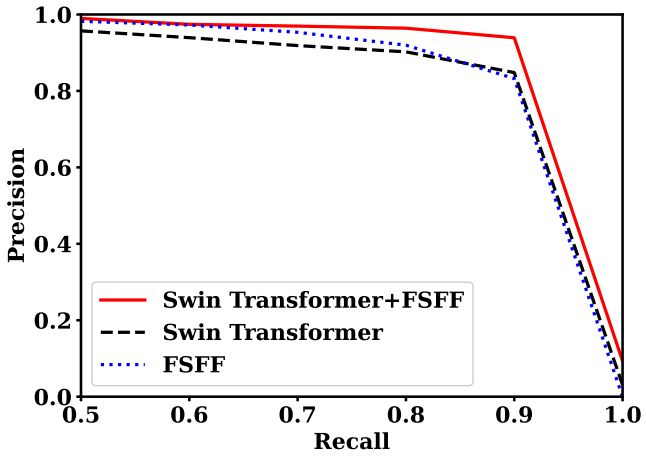


Fig. 12. PR curves comparison for module ablation study on RSOD dataset.

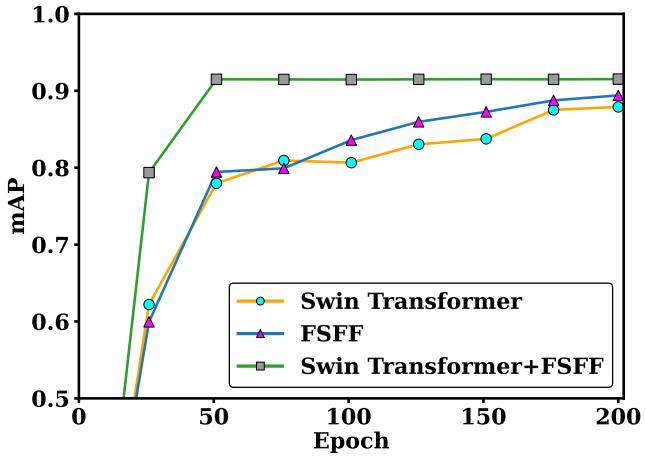


Fig. 13. Comparison of mAP for ablation experiments on the MAR20 dataset.

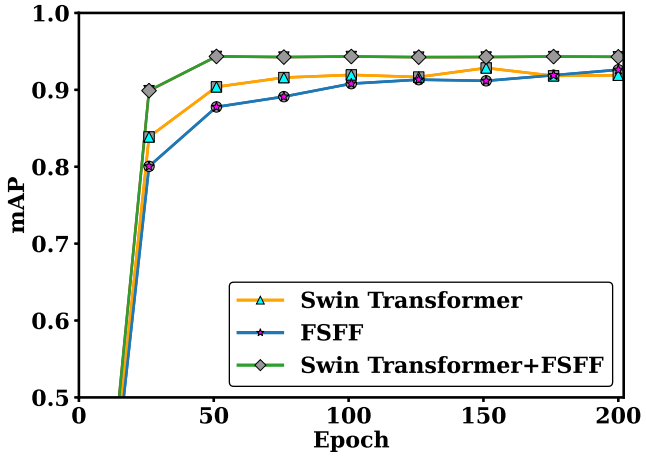


Fig. 14. Comparison of mAP for ablation experiments on the NWPU VHR-10 dataset.

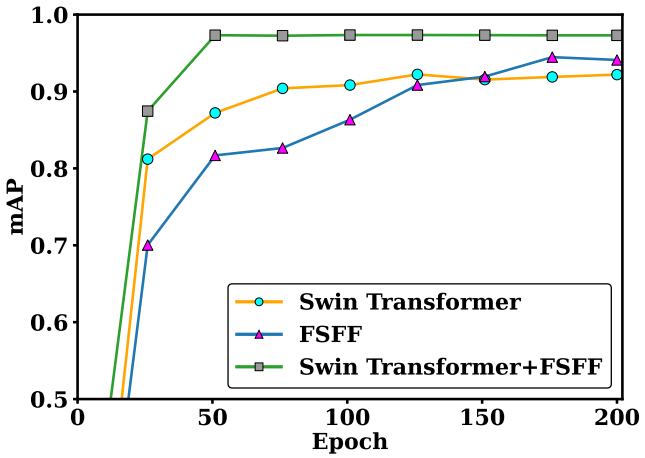


Fig. 15. Comparison of mAP for ablation experiments on the RSOD dataset.

neural network," Remote Sensing, vol. 10, no. 1, p. 131, 2018.

[2] P. Wang, X. Sun, W. Diao, and K. Fu, "Fmssd: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery," IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 5, pp. 3377–3390, 2019.

[3] W. Du, X. Ouyang, N. Zhao, and Y. Ouyang, "Bcs-yolov8s: A detecting method for dense small targets in remote sensing images based on improved yolov8s," IAENG International Journal of Computer Science, vol. 52, no. 2, pp. 417–426, 2025.

[4] J. Yan, H. Wang, M. Yan, W. Diao, X. Sun, and H. Li, "Iou-adaptive deformable r-cnn: Make full use of iou for multi-class object detection in remote sensing imagery," Remote Sensing, vol. 11, no. 3, p. 286, 2019.

[5] S. Li, X. Zhang, and R. Shan, "Enhanced yolov5 for efficient marine debris detection." Engineering Letters, vol. 32, no. 8, 2024.

[6] J. Pan and Y. Zhang, "Small object detection in aerial drone imagery based on yolov8." IAENG International Journal of Computer Science, vol. 51, no. 9, pp. 1346–1354, 2024.

[7] G. Wang, H. Chen, L. Chen, Y. Zhuang, S. Zhang, T. Zhang, H. Dong, and P. Gao, "P 2fevit: Plug-and-play cnn feature embedded hybrid vision transformer for remote sensing image classification," Remote Sensing, vol. 15, no. 7, p. 1773, 2023.

[8] W. Liang, J. Tan, H. He, H. Xu, and J. Li, "Detection of small objects from uav imagery via an improved swin transformer," In IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2024, pp. 9134–9138.

[9] D. Xue, T. Lei, S. Yang, Z. Lv, T. Liu, Y. Jin, and A. K. Nandi, "Triple change detection network via joint multifrequency and full-scale swin-transformer for remote sensing images," IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–15, 2023.

[10] X. Zhang and Z. Zhang, "Traffic sign detection algorithm based on improved yolov7," In International Conference on Image, Signal Processing, and Pattern Recognition (ISPP 2023), vol. 12707. SPIE, 2023, pp. 1258–1266.

[11] W. Zhao, Y. Kang, H. Chen, Z. Zhao, Z. Zhao, and Y. Zhai, "Adaptively attentional feature fusion oriented to multiscale object detection in remote sensing images," IEEE Transactions on Instrumentation and Measurement, vol. 72, pp. 1–11, 2023.

[12] B. Wang, H. Cui, X. Yu, Z. Su, and Y. Zheng, "Research on gangue detection method based on GD-YOLO," Engineering Letters, vol. 33, no. 1, pp. 59–68, 2025.

[13] D. Hou and Y. Zhang, "Object detection model for remote sensing images based on yolov9," IAENG International Journal of Computer Science, vol. 52, no. 3, pp. 840–847, 2025.

[14] X. Gong and D. Liu, "Sgmfnet: a remote sensing image object detection network based on spatial global attention and multi-scale feature fusion," Remote Sensing Letters, vol. 15, no. 5, pp. 466–477, 2024.

[15] Y. Liu and Y. Xiao, "Remote sensing object detection method based on attention mechanism and multi-scale feature fusion," In 2022 41st Chinese Control Conference (CCC). IEEE, 2022, pp. 7155–7160.

[16] Y. Zhang, C. Wu, T. Zhang, Y. Liu, and Y. Zheng, "Self-attention guidance and multiscale feature fusion-based uav image object detection," IEEE Geoscience and Remote Sensing Letters, vol. 20, pp. 1–5, 2023.

[17] D. Liu, L. He, and L. Carin, "Airport detection in large aerial optical imagery," In 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 5. IEEE, 2004, pp. V–761.

[18] C. Tao, Y. Tan, H. Cai, and J. Tian, "Airport detection from large ikonos images using clustered sift keypoints and region information," IEEE Geoscience and Remote Sensing Letters, vol. 8, no. 1, pp. 128–132, 2010.

[19] X. Yao, J. Han, L. Guo, S. Bu, and Z. Liu, "A coarse-to-fine model for airport detection from remote sensing images using target-oriented visual saliency and crf," Neurocomputing, vol. 164, pp. 162–172, 2015.

[20] J. Xu, K. Fu, and X. Sun, "An invariant generalized hough transform based method of inshore ships detection," In 2011 International Symposium on Image and Data Fusion. IEEE, 2011, pp. 1–4.
[21] Z. Zhang, J. Warrell, and P. H. Torr, "Proposal generation for object detection using cascaded ranking svms," In CVPR 2011. IEEE, 2011, pp. 1497–1504.
[22] M. Sun and Y. Tian, "Research on traffic sign object detection algorithm based on deep learning." Engineering Letters, vol. 32, no. 8, 2024.
[23] X. Dong, R. Fu, Y. Gao, Y. Qin, Y. Ye, and B. Li, "Remote sensing object detection based on receptive field expansion block," IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1–5, 2021.
[24] G. Li, Z. Liu, D. Zeng, W. Lin, and H. Ling, "Adjacent context coordination network for salient object detection in optical remote sensing images," IEEE Transactions on Cybernetics, vol. 53, no. 1, pp. 526–538, 2022.
[25] C. Li, G. Cheng, G. Wang, P. Zhou, and J. Han, "Instance-aware distillation for efficient object detection in remote sensing images," IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–11, 2023.
[26] Y. Li, Q. Hou, Z. Zheng, M.-M. Cheng, J. Yang, and X. Li, "Large selective kernel network for remote sensing object detection," In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 16 794–16 805.
[27] X. Yang, X. Zhang, N. Wang, and X. Gao, "A robust one-stage detector for multiscale ship detection with complex background in massive sar images," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–12, 2021.
[28] J. Hu, X. Zhi, S. Jiang, H. Tang, W. Zhang, and L. Bruzzone, "Supervised multi-scale attention-guided ship detection in optical remote sensing images," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–14, 2022.
[29] W. Ma, N. Li, H. Zhu, L. Jiao, X. Tang, Y. Guo, and B. Hou, "Feature split–merge–enhancement network for remote sensing object detection," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–17, 2022.
[30] W. Zhao, Y. Kang, H. Chen, Z. Zhao, Z. Zhao, and Y. Zhai, "Adaptively attentional feature fusion oriented to multiscale object detection in remote sensing images," IEEE Transactions on Instrumentation and Measurement, vol. 72, pp. 1–11, 2023.
[31] B. Song, P. Liu, J. Li, L. Wang, L. Zhang, G. He, L. Chen, and J. Liu, "Mlff-gan: A multilevel feature fusion with gan for spatiotemporal remote sensing images," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–16, 2022.
[32] X. Wang, S. Wang, C. Ning, and H. Zhou, "Enhanced feature pyramid network with deep semantic embedding for remote sensing scene classification," IEEE Transactions on Geoscience and Remote Sensing, vol. 59, no. 9, pp. 7918–7932, 2021.
[33] Y. Zhao, L. Zhao, Z. Liu, D. Hu, G. Kuang, and L. Liu, "Attentional feature refinement and alignment network for aircraft detection in sar imagery," arXiv preprint arXiv:2201.07124, 2022.
[34] R. Chen, Z. Cai, and W. Cao, "Mffn: An underwater sensing scene image enhancement method based on multiscale feature fusion network," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–12, 2021.
[35] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," In Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10 012–10 022.
[36] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 7464–7475.
[37] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," arXiv preprint arXiv:2203.03605, 2022.