

Appendix

Datasets

In this paper, extensive experiments are conducted on four publicly available benchmark datasets for UAV object detection and one well-recognized benchmark dataset for remote sensing object detection.

VisDrone. The VisDrone-2019-DET dataset consists of 6,471 training images, 548 validation images, and 3,190 test images, all captured by drones at varying altitudes across diverse locations. Each image is annotated with bounding boxes corresponding to ten predefined object categories: pedestrian, person, car, van, bus, truck, motorbike, bicycle, awning-tricycle, and tricycle. In this paper, We used the VisDrone-2019-DET training set and validation set for training and testing, respectively.

UAVDT. The UAVDT dataset comprises 23258 training images and 15069 testing images. Its scenes are similar to those in VisDrone2019, encompassing three categories: car, bus, and truck. We choose the training set for training and the test set for testing.

CODrone. The CODrone dataset contains 10,004 high-resolution images, partitioned into training (5,002 images), validation (2,000 images), and test (3,002 images) sets. Each image features annotations using oriented bounding boxes for objects across 12 distinct categories: car, truck, traffic-sign, person, motor, bicycle, traffic-light, tricycle, bridge, bus, boat, and ship. The imagery was captured under diverse lighting conditions—including normal daylight, low light, and nighttime—and from six unique viewpoints. These viewpoints resulted from combinations of three flight altitudes (30 m, 60 m, and 100 m) and two camera angles (30° and 90°). We choose the training set for training and the test set for testing.

UAVVaste. UAVVaste is a dataset designed specifically for aerial rubbish detection. It consists of 772 images and 3716 hand-labeled annotations of waste in urban and natural environments such as streets, parks, and lawns. We choose the training set for training and the test set for testing.

SIMD. SIMD is a medium-scale dataset specifically constructed for multi-scale and multi-category object detection, with a primary focus on vehicle detection in satellite imagery. It comprises 5,000 RGB images, acquired from 79 geographic locations across Europe and the United States through Google Earth, and adopts a 4:1 training-to-test split (4,000 training images and 1,000 test images). In total, the dataset contains 45,096 annotated instances covering 15 categories, predominantly including vehicles (e.g., cars, trucks, buses, and long vehicles), alongside various aircraft models and boats. We choose the training set for training and the test set for testing.

Evaluation Metrics

The standard COCO metrics we use to evaluate and compare the performance of various methods, including the AP(averaged over uniformly sampled IoU thresholds ranging from 0.50-0.95 with a step size of 0.05), and AP50 (AP at an IoU threshold of 0.50). Additionally, to comprehensively evaluate the model, metrics such as GFLOPs, and parameters are employed to determine the model’s complexity. The GFLOPs are calculated based on an input resolution of 640×640 .

Basic module comparison

To more intuitively illustrate the differences between our proposed SFSNet and existing feature extraction networks tailored for aerial images, we present a comparative diagram of the basic module (selective spatial-frequency module) of our method and those of other representative methods (LSKNet, PKINet and StripNet), as illustrated in Figure 6.

LSKNet employs large kernel decomposition with spatial selection mechanisms to dynamically adapt receptive fields. PKINet introduces Inception-style parallel multi-kernel convolutions and context anchor attention for robust scale variation handling and long-range dependency capture. StripNet adopts orthogonal large-strip convolutions to model high-aspect-ratio aerial objects through serialized horizontal and vertical operations. Despite their contributions, these methods exhibit notable limitations: LSKNet introduces background noise through large kernel convolutions, PKINet suffers from computational overhead and feature redundancy due to multiple large-kernel structures, and StripNet demonstrates limited generalization to diverse object geometries and global context modeling. SSFNet, proposed in this paper, adopts dual-domain learning and selection mechanisms to enable efficient and adaptive modeling of the diverse contextual demands of various objects.

Additional experiments

Results on SIMD Dataset. To evaluate the generalizability of our proposed method, we further performed comparative experiments on the remote sensing object detection dataset SIMD. As illustrated in Table 11, SFS-DETR-S achieves an AP of 66.1% and an AP_{50} of 81.2%. Compared with the baseline model RT-DETR-R18, the model attains relative improvements of 2.4% and 2.6% in AP and AP_{50} , fully demonstrating the generalization ability of our proposed method.

Comparative experiment for FPS. In this paper, we computed the Frames Per Second (FPS) of baseline models and SFS-DETR models using the PyTorch implementation in 32-bit Floating Point Precision, as shown in Table 12. The results show that SFS-DETR meets the real-time requirements.

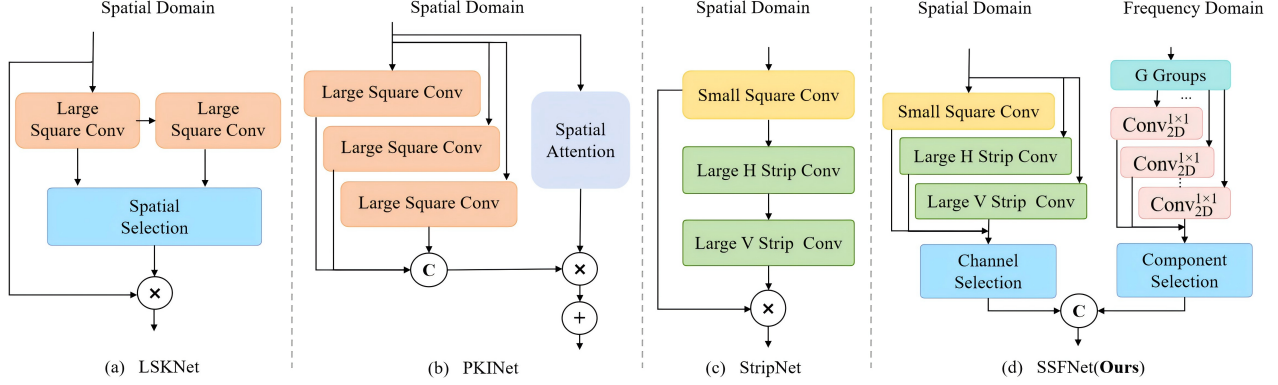


Figure 6. Structural comparison between our proposed selective spatial-frequency module and other representative methods, including LSKNet, PKINet and StripNet.

Table 11. Experimental Results on SIMD Dataset.

Model	AP	AP ₅₀
RT-DETR-R18	63.7	78.6
YOLOv8-L	63.1	78.1
YOLOv9-M	62.2	76.6
EMSD-DETR	64.3	79.4
SFS-DETR-S(Ours)	66.1	81.2

Table 12. Comparison of FPS with baseline and SSF-DETR on the VisDrone dataset.

Model	Params	GFLOPs	FPS
RT-DETR-R18	20	60.0	144
RT-DETR-R50	42	136.0	82
SFS-DETR-S	21.1	68.5	96
SFS-DETR	38.7	122.0	68

Table 13. Experiments with different ratio in Inner-MPDIoU.

IoU	AP	AP ₅₀
Inner-MPDIoU(ratio = 0.85)	31.1	50.5
Inner-MPDIoU(ratio = 0.75)	31.2	50.7
Inner-MPDIoU(ratio = 0.80)	31.3	50.9

Ablation experiment for Inner-MPDIoU ratio. In this paper, we introduce the Inner-MPDIoU loss function, which synergistically combines the advantages of Inner-IoU and MPDIoU. To achieve the optimal parameter selection, we conducted a series of ablation experiments, as detailed in Table 13. The experimental results indicate that setting the Inner-MPDIoU ratio to 0.8 is a suitable choice.

Visualization. To further showcase the advantages of SFS-DETR in aerial scenes, we present visualizations of feature maps and detection results on the CODrone dataset,

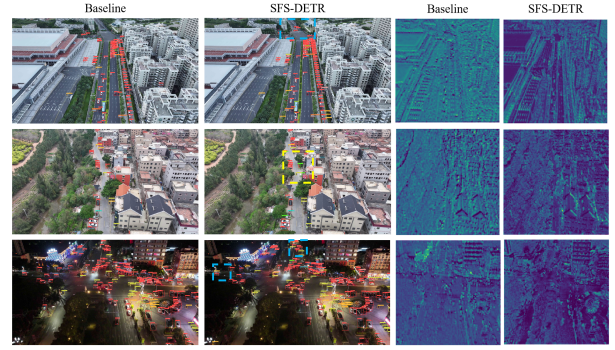


Figure 7. Visualizations of the detection results of baseline and our proposed method on CODrone.

as illustrated in Figure 7. Specifically, the baseline model exhibits missing detections for distant and extremely small targets, accompanied by certain false detections, as indicated by the blue boxes and yellow boxes in Figure 7. From the feature map comparisons, it can be observed that our proposed method more robustly captures the overall shapes and fine-grained boundaries of objects. These visualization results further verify the superior detection performance of our method in complex aerial scenes.