M^{2}SODAI: Multi-Modal Maritime Object Detection Dataset With RGB and Hyperspectral Image Sensors

Jonggyu Jang
POSTECH
jgjang@postech.ac.kr

Sangwoo Oh
KRISO
swoh@kriso.re.kr

Youjin Kim
Samsung Electronics
youjin8022@gmail.com

Dongmin Seo
Semyung University
dseo@semyung.ac.kr

Youngchol Choi
KRISO
ycchoi@kriso.re.kr

Hyun Jong Yang*
POSTECH
hyunyang@postech.ac.kr

Abstract

Object detection in aerial images is a growing area of research, with maritime object detection being a particularly important task for reliable surveillance, monitoring, and active rescuing. Notwithstanding astonishing advances in computer vision technologies, detecting ships and floating matters in these images is challenging due to factors such as object distance. What makes it worse is pervasive sea surface effects such as sunlight reflection, wind, and waves. Hyperspectral image (HSI) sensors, providing more than 100 channels in wavelengths of visible and near-infrared, can extract intrinsic information about materials from a few pixels of HSIs. The advent of HSI sensors motivates us to leverage HSIs to circumvent false positives due to the sea surface effects. Unfortunately, there are few public HSI datasets due to the high cost and labor involved in collecting them, hindering object detection research based on HSIs. We have collected and annotated a new dataset called "**Multi-Modal Ship and flOating matter Detection in Aerial Images (M^{2}SODAI)**", which includes synchronized image pairs of RGB and HSI data, along with bounding box labels for 5,764 instances per category. We also propose a new multi-modal extension of the feature pyramid network called DoubleFPN. Extensive experiments on our benchmark demonstrate that the fusion of RGB and HSI data can enhance mAP, especially in the presence of the sea surface effects. The source code and dataset are available on the project page: https://sites.google.com/view/m2sodai.

1 Introduction

With the growing maritime traffic intensity, detecting and localizing ships and floating matters have become core functionalities for reliable monitoring, surveillance, and active rescuing [3, 7]. Conventionally, there have been sea surface maritime surveillance systems based on buoys and ships [56]. These systems are cost-efficient; however, their sensing range is relatively narrow. By virtue of their wide sensing range, aerial surveillance systems have received considerable research interest, the absolute majority of which leverage optical cameras. Although optical cameras can obtain high-resolution RGB images, the competence of optical sensors is degraded under dire but commonplace environmental conditions such as solar reflection or waves, *i.e.*, *sea surface effects*.

Hyperspectral image (HSI) sensors, which acquire imagery in hundreds of contiguous spectral bands, are emerging as a substitute or supplement of RGB sensors [42, 32]. Abundant spatio-spectral

*corresponding author

37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks.

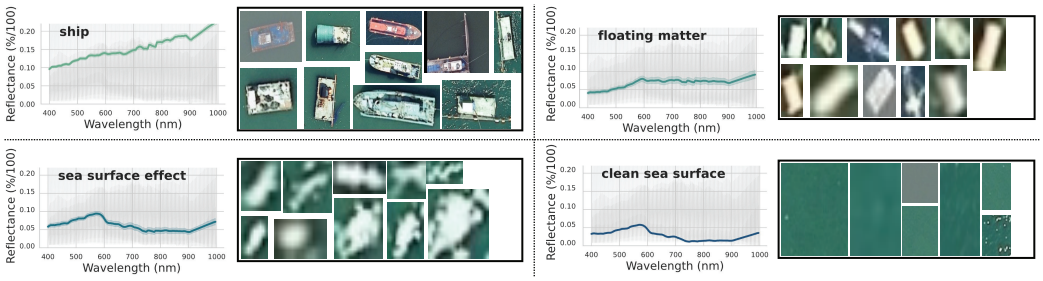


Figure 1: M^{2}SODAI dataset spectral analysis. From the top of the figure, we depict the hyperspectral reflectance intensity patterns and cropped RGB data of i) ship, ii) floating matter, iii) sea surface effect, and iv) clean sea surface. The figure shows that the floating matters and sea surface effects are similar in the RGB image; however, they have different reflectance intensity patterns in the HSI data.

Table 1: M^{2}SODAI dataset vs. related datasets for RGB and HSI data. Among all the datasets, the M^{2}SODAI dataset is the only dataset with i) bounding-box-annotated, ii) synchronized multi-modal, and iii) aerial RGB and HSI data.

dataset	#instances /class	#images	#classes	RGB data (width)	HSI data (width)	multi-modality	annotation	view	Year	description
VEDAI[38]	327	1,268	9	✓ (512, 1,024)	-	-	bounding box	aerial	2015	object detection
COWC[34]	2,007	32,700	1	✓ (2,048)	-	-	bounding box	aerial	2016	vehicle detection
CARPK[21]	89,777	1,448	1	✓ (1,280)	-	-	bounding box	aerial	2017	vehicle detection
DOTA-v1.0[50]	12,552	2,806	15	✓ (800-13,000)	-	-	bounding box	aerial	2018	object detection
VisDrone[59]	5,420	10,209	10	✓ (2,000)	-	-	bounding box	aerial	2018	object detection
iSAID[49]	43,696	2,806	15	✓ (800-13,000)	-	-	polygon	aerial	2019	object detection
FGSD[7]	131	2,612	43	✓ (930)	-	-	bounding box	aerial	2020	ship detection
DOTA-v2.0[8]	99,647	11,268	18	✓ (800-20,000)	-	-	bounding box	aerial	2021	object detection
India Pines[2]	-	1	16	-	✓ (145)	-	pixel-wise	aerial	2015	remote sensing
HAI[31]	-	65,000	-	✓ (500)	✓ (500)	✓ (Sync)	-	aerial	2021	dehazing
Samson[58]	-	1	3	-	✓ (952)	-	pixel-wise	aerial	2022	remote sensing
MDAS[22]	-	23	859	✓ (15,000)	✓ (300)	✓ (Sync)	pixel-wise	aerial	2022	remote sensing
HS-SOD[23]	120	60	1	-	✓ (1,024)	-	polygon	terrestrial	2018	object detection
ODHI[52]	832 (RGB), 207 (HSI)	2048 (RGB), 454 (HSI)	8	✓ (∼696)	✓ (∼696)	× (Async)	bounding box	terrestrial	2021	real/fake detection
M^{2}SODAI (ours)	**5,764**	**1,257**	**2**	**✓ (1,600)**	**✓ (224)**	**✓ (Sync)**	**bounding box**	**aerial**	**2023**	**object detection**

snapshots of HSIs provide inherent reflective properties of materials even with just a few pixels, which is not possible with RGB or any other types of images. Figure 1 shows the RGB and HSI data examples of the ships, floating matters, sea surface effects, and clean sea surface, where reflection intensity patterns of objects and backgrounds are plotted in the left parts. In the wavelengths of the near-infrared (NIR) region, water exhibits a pattern of sharply decreasing reflectance between 700 nm and 900 nm, unlike target objects [54]. That is, even at low resolution, HSI sensor data can identify unique object characteristics, differentiating the targets from the background[2].

However, most of the object detection datasets on aerial images are about optical images [7, 38, 34, 21, 50, 49, 8], and there are only handful HSI datasets publicly available. Even for other tasks, such as remote sensing, datasets with aerial HSIs are scarce because collecting HSI data is costly and labor-intensive [2, 58, 22]. In this work, we build a new Multi-Modal Ship and flOating matter Detection in Aerial Images (M^{2}SODAI) dataset, which contains synchronized pairs of aerial RGB and HSI data. For the data collection, we used an off-the-shelf HSI sensor taking 127 spatio-spectral channels for each snapshot on the wavelength from 400 nm to 1000 nm in steps of 4.5 nm. The spatial resolutions of the RGB and HSI data are 0.1 m and 0.7 m, respectively, at the altitude of 1 km.

For object detection in aerial images, one major drawback of HSIs is their relatively low spatial resolution (several meters) compared with optical images (tens of centimeters). Thus, HSI sensors have been commonly used in remote sensing systems which do not require high-resolution [14, 13, 1, 25]. As hardware technologies for HSI sensors evolve, their resolution has increased, facilitating

[2]For a detailed analysis, please see Appendix E.

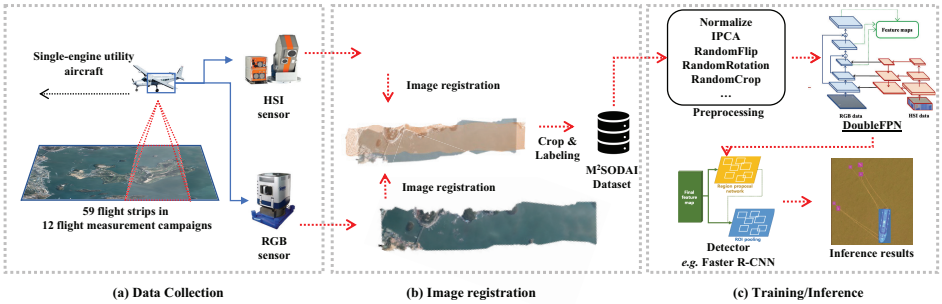
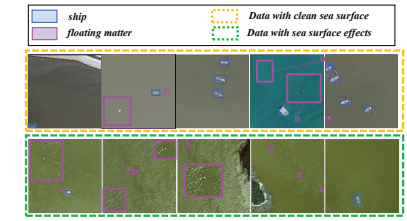


Figure 3: Illustration of the overall procedure of our work. (a) We collect the sensor data using a single-engine utility aircraft equipped with RGB and HSI sensors. (b) Due to the offset between RGB and HSI sensor data, we register RGB and HSI data into the same coordinate and then construct our dataset by cropping and labeling data. (c) The data are preprocessed and forwarded to the DoubleFPN layer. Finally, after the DoubleFPN layer, the detector estimates the bounding boxes of the target objects from the DoubleFPN output.

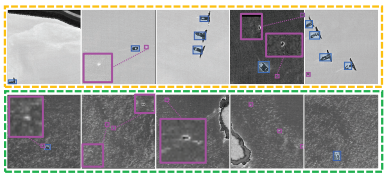
bounding-box-based deep learning research for object detection in HSIs [52]. Nonetheless, the resolution of off-the-shelf HSI sensors is not yet high enough for airborne surveillance systems. Therefore, it is a better choice to use HSIs as a supplement, not a substitute, to optical images in the case of far-field object detection. Further analysis of related works is provided in Appendix A.

The salient contributions of our work are listed as follows:

- **M^{2}SODAI dataset:** The M^{2}SODAI dataset is the first multi-modal, bounding-box-labeled, and synchronized aerial dataset, featured by 11,527 instances, 1,257 images, and synchronized RGB-HSI data. In Tab. 1, we compare the M^{2}SODAI dataset with the related public datasets on RGB and HSI data. Amongst the related datasets, the HAI [31] and MDAS [22] datasets only provide synchronized multi-modal aerial data; however, they come with no annotation [31] or low-resolution pixel-wise annotation [22].
  - **Raw data processing:** We add a contrast enhancer to the method proposed in [24] for more accurate data synchronization of RGB and HSI. For more details, please refer to Sec. 2 and Appendix D. Figure 2 illustrates randomly selected pairs of RGB and HSI data from our dataset.
- **Multi-modal benchmark and learning framework:** We conducted an object detection benchmark with our dataset, where the graphical and numerical results ensure the HSI data can enhance the detection accuracy, especially for data with sea surface effects. In order to fuse the RGB and HSI data, we propose an extension of the feature pyramid network (FPN) [26], DoubleFPN. For a more general benchmark, we build other fusion methods based on DetFusion [44] and UA-CMDet [43].



(a) Annotated RGB data.

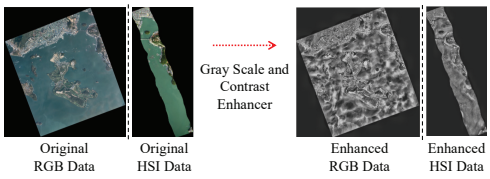


(b) Annotated HSI data.

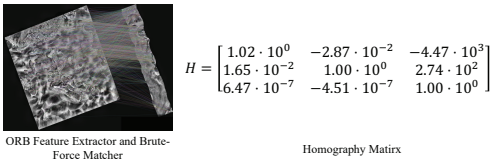
Figure 2: Examples of collected RGB and HSI data. (a): Typical RGB data in M^{2}SODAI dataset. (b): Infrared visualization of the synchronized HSI data in the dataset (ratio of the 25-th and 72-nd channels). Here, we show ten examples of images with and without sea surface effects. In the images with sea surface effects, the HSI data have more recognizable features than the RGB data.

2 M^{2}SODAI Dataset

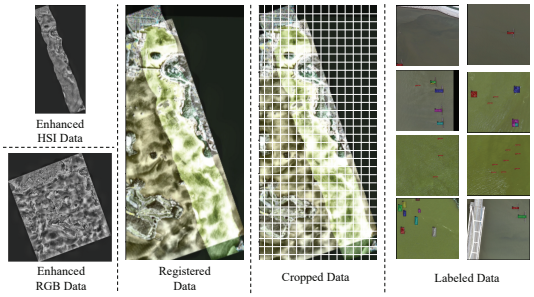
Overview of the proposed dataset construction procedure Figure 3 illustrates the overall procedure of the proposed scheme. In the first stage, we collect RGB and HSI sensor data using a single-engine utility aircraft equipped with RGB and HSI sensors. In the second stage, an image



(a) Gray scale transformation and a contrast enhancer.



(b) Feature matching and corresponding homography matrix.



(c) Image registration and dataset construction.

Figure 4: Illustration of our dataset construction procedure. (a) The RGB and HSI sensor data are transformed by a gray scaler and a contrast enhancer. (b) The matched feature of transformed sensor data and the corresponding homography matrix are obtained. (c) The sensor data are registered, cropped, and labeled.

registration method is used to coincide the pixels of RGB and HSI data. After the image registration, we construct our dataset by cropping by fixed size and annotating target objects (ships and floating matters) in the RGB and HSI data. Note that our dataset consists of HSI, RGB, and corresponding bounding box annotation data. Further details of our dataset are available in Appendix B. In the third stage, we train our DoubleFPN architecture and evaluate the trained model using the M^{2}SODAI dataset.

Data collection Our focus is to create a public dataset consisting of synchronized maritime aerial RGB and HSI data. To this end, we built a data collection system by leveraging a single-engine utility aircraft (Cessna Grand Caravan 208B). An HSI sensor (AsiaFENIX, Specim, Oulu, Finland) and an RGB sensor (DMC, Z/I Imaging, Aalen, Germany) are equipped on the bottom of the aircraft, the direction of which is downward. The raw data was acquired through 59 flight strips in 12 flight measurement campaigns, which cover a total area of 299.7 km^2. During the flight strips, the aircraft maintains its speed of 260 km/h and altitude of 1 km.

Table 2: Specification of RGB and HSI sensor. The resolutions of the sensors are corresponding to the aircraft's altitude of 1 km.

	HSI Sensor	**RGB Sensor**
Name	AisaFENIX (@Specim)	DMC (@Z/I Imaging)
Spectrum	400-1000 nm (in steps of 4.5 nm) 127 channels	Blue: 400-580 nm Green: 500-650 nm Red: 590-675 nm
Altitude	1 km	
Field of View	40°	74°
Resolution	0.7 m	0.1 m

Table 2 shows the detailed specifications of the sensors used in the data collection. The HSI sensor (AsiaFENIX) scans the wavelength range from 400 nm to 1000 nm in steps of 4.5 nm, a total of 127 spectrum bands. The wavelength range includes visible spectrum and NIR spectrum, generally used for remote sensing and machine vision tasks. The RGB sensor (DMC) captures high-resolution RGB data in three channels: Red (590-675 nm), Green (500-650 nm), and Blue (400-580 nm). We note that RGB and HSI data are collected simultaneously, in which the spatial resolutions of RGB and HSI sensors are approximately 0.1 m and 0.7 m, respectively.

Image registration and annotation In the previous step, we introduce the methodology of collecting the raw RGB and HSI data. Since the size of the raw data is too large for object detection (HSI: 3,220^2 pixels, and RGB: 22,520^2 pixels on average), we cropped the raw data into a fixed size. We note that RGB and HSI data are cropped in size of $1600 \times 1600 \times 3$ and $224 \times 224 \times 127$, respectively. However, the problem is that the coordinates of the collected RGB and HSI pairs are not matched. Hence, we employ an image registration method to correct pixel offsets between RGB and HSI pairs. In Fig. 4, our data processing procedure is depicted.

1. We transform the raw RGB and HSI data into grayscale images (Fig. 4a) [24].

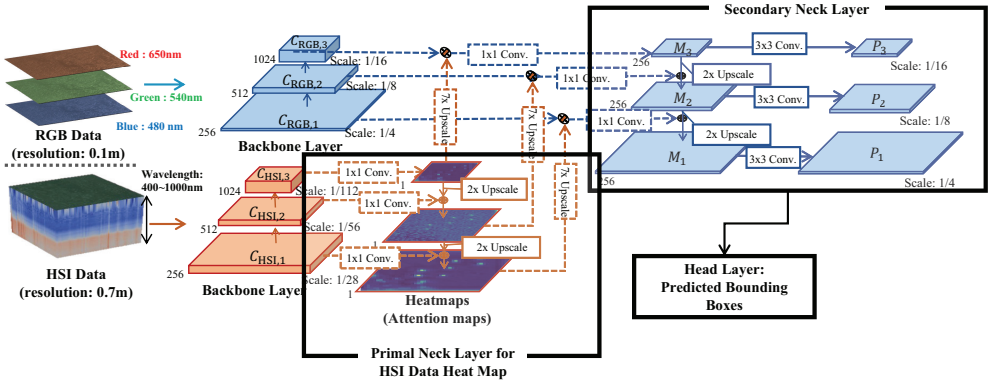


Figure 5: Schematic diagram of the DoubleFPN-based object detection architecture. The DoubleFPN object detection architecture consists of three sub-architectures: backbone, neck, and head layers. In the backbone layer, the feature maps of each input data are extracted, *i.e.*, bottom-up pathway. In the neck layer, the DoubleFPN fuses feature maps, *i.e.*, top-down pathway. In the head layer, the object detector estimates the classes and bounding boxes of the objects.

2. We apply contrast-limited adaptive histogram equalization (CLAHE)-based contrast enhancer to the grayscale RGB data and grayscale HSI data (Fig. 4a).
3. To estimate the homography matrix between the enhanced RGB data and enhanced HSI data, we carry out the oriented FAST and rotated BRIEF (ORB) feature descriptor [41] to both data, thereby extracting features of the data (Fig. 4b).
4. We use a Brute-force matcher to find the matched feature among the ORB features; then, the homography matrix is computed from least square optimization for synchronizing the matched features.
5. We crop the registered data in the same size and generate corresponding bounding box annotation data (Fig. 4c).

For object detection, we annotated the bounding boxes on the instances of two classes: 1) floating matter and 2) ships in both RGB and HSI data. We note that the following instances are labeled as floating matters: buoys, rescue tubes, small lifeboats, surfboards, and humans (mannequins[3]) with life vests. Also, for the ship class, we annotated bounding boxes on steamboats, cruise ships, fishing boats, sailboats, rafts, and other ship categories. We refer to the infrared visualization map of the HSI data for bounding box annotation. For labeling, two of the authors annotated target instances by using Labelme [47], in which the minimum size box containing each object was set as the policy, and multiple cross-checks were performed. For more details on raw data processing, please see Appendix D.

Dataset splits After the data processing, we obtained 1,257 pairs of synchronized RGB and HSI data, where the total number of instances in the dataset is 11,527. For experiments, we randomly divided the dataset into 1,007 training data, 125 validation data, and 125 test data.

3 Method: DoubleFPN

The feature fusion methods are categorized into i) early fusion, ii) middle fusion, and iii) late fusion [11]. The early fusion methods fuse sensor data before the backbone layers, thereby fully leveraging joint features of raw data. However, the common representations of different sensor data are challenging. On the other hand, the late fusion methods combine sensor data just before the final detector, whereas they have a potential loss for finding the correlation of sensor data. In our study, the aim is a compromise proposal of early and late fusion methods, *i.e.*, *middle fusion*.

Here, the training/inference procedure in Fig. 3 is addressed. In the canonical FPN structure [40], a pyramid structure for feature extraction is proposed to resolve the issues of memory inefficiency and

[3]Distressing a real person was done with a mannequin for safety reasons.

Table 3: AP (%) benchmark result on the M^{2}SODAI dataset with the DoubleFPN and the uni-modal baseline methods. All the results are obtained by using ResNet-50 backbone and Faster R-CNN detector. In addition to the AP-based metrics, we show types of neck layers and use of the RGB and HSI data.

neck layer	multi-modal	RGB data	HSI data	mAP	$AP_{@.5}$	$AP_{@.75}$	Ship	Float. Mat.	AP_s	AP_m	AP_l
DoubleFPN(ours)	✓	✓	✓	**44.4**	**84.8**	39.3	55.7	**33.1**	**35.2**	41.7	**61.4**
FPN (RGB) [26]	×	✓	×	38.8	77.0	33.3	52.4	25.2	18.3	**44.8**	55.6
FPN (HSI) [26]	×	×	✓	7.8	23.2	2.9	15.8	0.0	-	-	-
UA-CMDet [43]	✓	✓	✓	42.9	84.0	**40.0**	**55.9**	29.8	20.8	43.0	60.8
DetFusion [44]	✓	✓	✓	42.0	84.3	35.4	53.5	30.5	24.2	41.9	61.1
Early fusion	✓	✓	✓	42.9	83.0	37.6	54.2	31.5	18.9	44.1	59.7

*Best: **bold and underline**, second-best: underline.

low inference speed of the general feature map extraction architecture. However, the input of the FPN is a fixed-scale single image, and the output is feature maps sized proportionally to the input image.

For our dataset, the feature extraction network should be capable of handling RGB and HSI data with different scales. More importantly, HSI data itself does not have sufficiently high resolution to detect aerial objects, even though it can capture unique features of materials. Hence, we propose an extension of the canonical FPN to jointly extract feature maps by fusing two data. The detailed schematic diagram of the DoubleFPN is depicted in Fig. 5. We note that the DoubleFPN architecture can be generally implemented with other detectors, such as RetinaNet and FCOS [27, 45].

Dimensionality reduction and preprocessing Let us denote the size of RGB data and HSI data as $H_{\text{rgb}} \times W_{\text{rgb}} \times 3$ and $H_{\text{hyp}} \times W_{\text{hyp}} \times C_{\text{hyp}}$, respectively. We note that $C_{\text{hyp}} = 127$ in our dataset. Since several spectral features are necessary for object detection, we leverage the incremental PCA method. As a result, we observe that the cumulative variance of the first 30 principal components occupies more than 99.9% of the total variance. Hence, we use 30 principal components in our object detection instead of fully leveraging 127 channels.

Backbone layer In Fig. 5, the backbone layers are feed-forward CNNs that extract feature maps of the inputs, *i.e.*, *bottom-up pathway*. As in the figure, each pair of RGB and HSI data is fed into the separate backbone layer, in which the CNN layers for RGB and HSI data have N different scales. The output feature maps at each level are scaled by 1/2 of that at the previous level. Here, we denote the i-th feature map of RGB and HSI data as $C_{\text{RGB},i}$ and $C_{\text{HSI},i}$.

Neck layer In the neck layer, the DoubleFPN forwards N fused feature maps from N RGB feature maps and N HSI feature maps. In the primal neck layer, the HSI feature maps are converted into attention maps to represent weights for the high-resolution RGB features. At the top of the primal neck layer, the feature map $C_{\text{HSI},N}$ is fed into 1×1 convolution layer with one channel with Sigmoid activation function. In the top-down pathway of the primal neck layer, the i-th feature map $C_{\text{HSI},i}$ is forwarded into 1×1 convolution layer with one channel and is added with the 2x up-scaled previous attention map. Let us define the i-th attention map as H_i. Then, the i-th attention map H_i is up-scaled seven times and is multiplied with the i-th RGB feature map $C_{\text{RGB},i}$ for $i = 1, .., N$. In the secondary neck layer, the fused feature maps $H_i \cdot C_{\text{RGB},i}$ are forwarded into the canonical FPN structure. Consequently, after 3×3 convolution layers, we get a set of fused feature maps with different scales.

Head layer The head layer predicts the bounding boxes and classes of the objects from the output of the DoubleFPN. For the experiments, we introduce an application of our method to Faster R-CNN [40]. For further implementation details, please refer to Appendix F.

4 Experiments

Since none of the other datasets in Tab. 1 provides synchronized RGB, HSI, and bounding-box-annotation data, we evaluate the DoubleFPN on the M^{2}SODAI dataset.

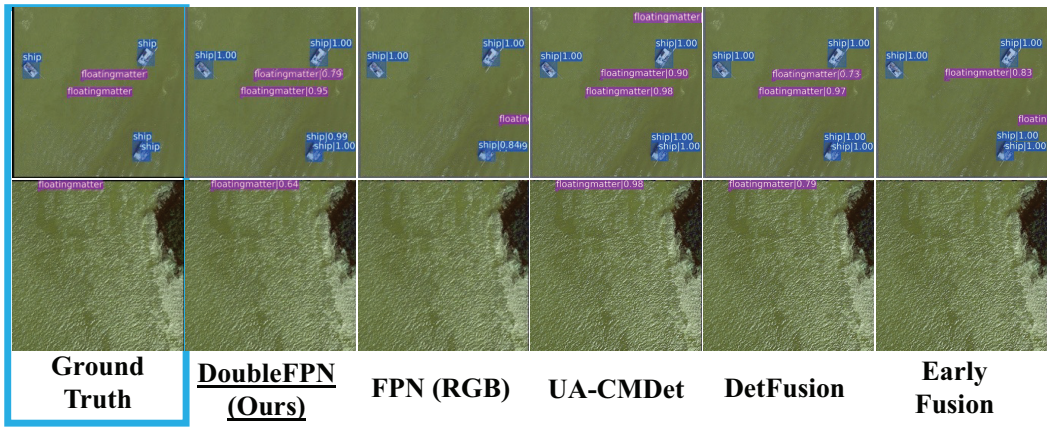


Figure 6: Detection results on data with sea surface effects. The first figure depicts the ground truth of the bounding box. The other figures show the detected bounding boxes of the objects, *e.g.*, floating matter, ship.

4.1 Setups and Implementation Details

Implementation details Our experiment is carried out on two NVIDIA RTX 3090 GPUs. The overall object detection model is trained for 73 epochs, in which the stochastic gradient descent parameters are: the learning rate of $2 \cdot 10^{-2}$, the momentum of 0.9, and the weight decay of $1 \cdot 10^{-4}$. In addition, the batch size is set to be one per GPU[4]. For fairness in the performance analysis, we evaluate all methods based on the ResNet-50 backbone model [19]. Since the ResNet-50 model provides five-stage feature maps, each of the backbone networks in the uni-modal methods provides five feature maps. In the DoubleFPN, the backbone layer for RGB and HSI data input forwards five and four feature maps, respectively, *i.e.* $N = 5$. For other experiment parameters, we follow the default parameters of the canonical FPN [40]. In the evaluation, we employ the standard COCO metrics average precision (AP) metrics: mAP (averaged AP over IoU thresholds from 0.5 to 0.95), $\text{AP}_{@.5}$, $\text{AP}_{@.75}$, AP_s (area$\in(0,32^2]$), AP_m (area$\in(32^2,96^2]$), and AP_l (area$\in(96^2,\infty)$).

4.2 Performance Analysis of DoubleFPN

Table 3 shows the evaluation result on the test set of the M^{2}SODAI dataset. As a baseline detector, we use a widely used uni-modal object detector, Faster R-CNN [40] for all benchmark results[5]. For comparison, we add an early fusion method with simple convolution layers and late fusion methods modified from DetFusion [44] and UA-CMDet [43].

Comparison with uni-modal object detection We first compare the DoubleFPN method and the uni-modal methods, which use either RGB or HSI data. First, we can see that the DoubleFPN method outperforms all other uni-modal methods in most of the metrics. This means that the DoubleFPN method significantly reduces the number of false positive bounding boxes by using the HSI data as a complement to RGB data. Second, when HSI data is used as a substitute for RGB data, the performance of object detection is significantly lower than that of the methods using only RGB data. This is because HSIs have relatively lower resolution than RGB images, so it is difficult to infer accurate shapes of bounding boxes even if they know whether the target objects exist or not. As a result, the benchmark results in Tab. 3 show that HSIs are suitable as a complement to RGB images, but are not yet sufficient enough as a substitute.

[4]The largest batch size in our GPU configuration.

[5]Although we have tried to train with recent object detectors such as TOOD [10] (mAP of 38.8 % with ResNet-50 backbone) and VFNet [55] (mAP 40.9 % of with ResNet-50 backbone), Faster R-CNN (mAP of 44.4 %) performs better under the training from the scratch settings. We note that for enhanced performance with HSI data, there's a need for either a representative pre-trained backbone layer or an improved training method for the latest detectors from the ground up.

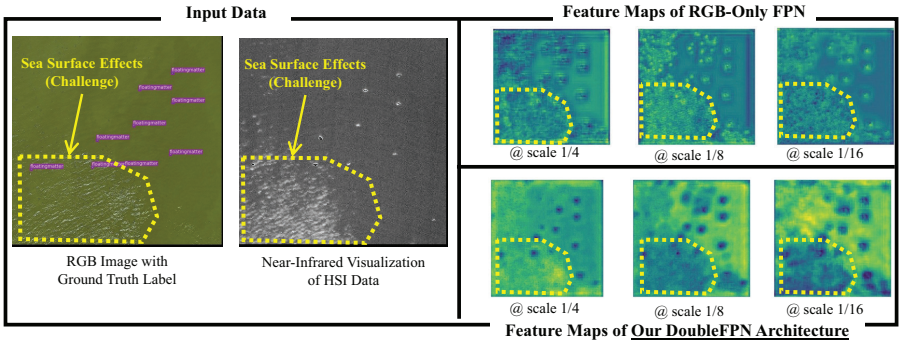


Figure 7: Feature maps with sea surface effects. The strongest feature among channels is selected for each pixel for visualization.

Comparison with multi-modal methods In Tab. 3, the DoubleFPN method outperforms the early fusion method, where the reason would be that our method uses the HSI feature maps as attention maps on RGB feature maps, whereas the early fusion method directly fuses the feature maps in the backbone layer. More specifically, compared to the late fusion (UA-CMDet and DetFusion) methods, our method has a higher mAP, since late fusion methods cannot jointly fuse the feature maps of RGB and HSI data.

Comparison with and without sea surface effects Table 4 shows the detection performance on the data with sea surface effects. By comparing Tabs. 3 and 4, the AP metrics of multi-modal methods are steady regardless of the sea surface effects; however, the AP metrics of the uni-modal methods have more degradation with many false positive bounding boxes if there are sea surface effects. This shows that multi-modal detection can perform more robust object detection for maritime object detection by leveraging the HSI data. For visualization, in Fig. 6, we show some samples of the object detection results and ground truth annotations on the data with sea surface effects. From the figure, we can see that the multi-modal methods propose more accurate bounding box estimations. For more examples of the benchmark results, please refer to Appendix G.

Table 4: Benchmark result for data with sea surface effects.

neck layer	RGB/HSI	mAP	$AP_{@.5}$	$AP_{@.75}$	Ship	Float. Mat.
DoubleFPN (ours)	✓/✓	**42.2** (↓ 2.2)	82.3	31.2	41.7	**42.8**
FPN (RGB)	✓/×	35.1 (↓ 4.4)	73.9	30.0	42.2	28.0
UA-CMDet	✓/✓	38.9 (↓ 3.0)	82.2	33.4	43.4	34.4
DetFusion	✓/✓	39.7 (↓ 2.3)	**83.3**	30.2	**43.5**	35.9
Early fusion	✓/ ✓	40.4 (↓ 2.5)	79.4	**34.4**	42.0	38.9

*Best: **bold and underline**, second-best: underline.
**(): mAP differences for the overall sea data results.

4.3 Visualization Analysis

Figure 7 visualizes feature maps of the DoubleFPN and the RGB-only canonical FPN. As depicted in the figure, the input data have strong sea surface effects in the bottom-left corner, which are the challenge. To the right of the input data image, we depict the feature maps of the RGB-only FPN, which are vulnerable to the sea surface effects. For example, the feature maps of the RGB-only FPN are not clear. On the other hand, in the lower part, the feature maps of the DoubleFPN are drawn, where the DoubleFPN fuses the RGB and HSI backbone outputs in order from low resolution to high resolution. As RGB and HSI data are fused, the feature maps of the DoubleFPN become clearer. Therefore, DoubleFPN can estimate the bounding boxes more accurately than RGB-only FPN method by delivering clearer feature maps to the detector.

5 Discussion

Summary Our work addresses the problem of maritime object detection in aerial images using two types of data: RGB and HSI. To this end, we created the M^2SODAI dataset, which is the first dataset composed of bounding box annotations, RGB, and HSI data. We propose a multi-modal object detection framework that fuses high-resolution RGB and low-resolution HSI data. Our extensive experiments confirm the robustness of our object detection model on maritime object detection.

Limitations The limitations of our work are three-fold. 1) There is room for performance enhancement by having pre-trained backbone networks HSI data and multi-modal detectors instead of Faster R-CNN. 2) When we collect the data, the weather is always sunny. A future research direction is to enhance the object detection performance by proposing a new neural network architecture or to collect data in various weather conditions (*e.g*. foggy, rainy, etc.) or main/sub categories (*e.g*. buoys, rescue boats, cars, buildings, etc.). Hopefully, the atmospheric correction, typically applied during HSI data collection, can adjust for unwanted weather conditions to simulate sunny conditions, thereby allowing our data to serve as a more general representation[15]. Additionally, as the dataset has been gathered in South Korea, there may exist potential biases in the data, such as variations in the object's distribution, the condition of the oceans, and the types of ships that are commonly used in the region. 3) The data collection scenario presented in this paper requires actual aircraft and expensive HSI sensors, resulting in significant financial costs. We believe that this paper will inspire relatively low-cost drone-based data collection methods and maritime surveillance systems with HSI data. 4) Techniques for image fusion that incorporate extra HSI demonstrate greater delays in comparison to those methods that rely solely on RGB. As a result, we have advocated for forthcoming research into a 3D CNN-based feature mapping for HSI, emphasizing an approach that is both more computationally streamlined and adept at extracting essential features.

Societal impact and ethics consideration First, the M^2SODAI dataset offers a new perspective on maritime object detection, which can bring about positive societal effects in various applications such as maritime safety and national defense. A typical negative societal impact during aerial data collection is capturing sensitive areas, such as military zones or private areas. We have carefully reviewed this aspect, and we ensure that our flight areas are limited to non-military zones and non-private areas as shown in Fig. B.1.

Usefulness of M^2SODAI In our benchmark, M^2SODAI has demonstrated the ability to enhance object detection accuracy using HSI data to complement existing high-resolution optical images. We have strong confidence that this dataset will not be limited to object detection tasks but can also be sufficiently utilized for other tasks, such as RGB to HSI reconstruction tasks.

Acknowledgment

This research was supported by a grant from Endowment Project of "Development of Open Platform Technologies for Smart Maritime Safety and Industries" funded by Korea Research Institute of Ships and Ocean engineering (PES4880), and Korea Institute of Marine Science and Technology Promotion (KIMST) funded by the Ministry of Oceans and Fisheries (RS-2022-KS221606).

References

[1] Rick Archibald and George Fann. Feature selection and classification of hyperspectral images with support vector machines. *IEEE Geosci. Remote Sens. Lett.*, 4(4):674–677, October 2007.

[2] Marion F. Baumgardner, Larry L. Biehl, and David A. Landgrebe. 220 band AVIRIS hyperspectral image data set: June 12, 1992 Indian Pine Test Site 3, Sep 2015. URL `https://purr.purdue.edu/publications/1947/1`.

[3] Domenico D Bloisi, Fabio Previtali, Andrea Pennisi, Daniele Nardi, and Michele Fiorini. Enhancing automatic maritime surveillance systems with visual information. *IEEE Trans. Intell. Transp. Syst.*, 18(4): 824–833, April 2017.

[4] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[5] Alvaro Luis Bustamante, Jose M Molina, and Miguel A Patricio. Information fusion as input source for improving multi-agent system autonomous decision-making in maritime surveillance scenarios. In *Proc. IEEE Inter. Conf. on information fusion (FUSION)*, pages 1–8, 2014.

[6] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Mmdetection: Open mmlab detection toolbox and benchmark. *CoRR*, abs/1906.07155, 2019. URL `http://arxiv.org/abs/1906.07155`.

[7] Kaiyan Chen, Ming Wu, Jiaming Liu, and Chuang Zhang. FGSD: A dataset for fine-grained ship detection in high resolution satellite images. *CoRR*, abs/2003.06832, 2020. URL `https://arxiv.org/abs/2003.06832`.

[8] Jian Ding, Nan Xue, Gui-Song Xia, Xiang Bai, Wen Yang, Michael Ying Yang, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(11):7778–7796, Nov. 2022. doi: 10.1109/tpami.2021.3117983.

[9] Fahimeh Farahnakian, Mohammad-Hashem Haghbayan, Jonne Poikonen, Markus Laurinen, Paavo Nevalainen, and Jukka Heikkonen. Object detection based on multi-sensor proposal fusion in maritime environment. In *IEEE Inter. Conf. on machine learning and applications (ICMLA)*, pages 971–976, December 2018.

[10] Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R Scott, and Weilin Huang. Tood: Task-aligned one-stage object detection. In *ICCV*, 2021.

[11] Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Glaeser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Trans. Intell. Transp. Syst.*, 2020.

[12] Yvonne Fischer and Alexander Bauer. Object-oriented sensor data fusion for wide maritime surveillance. In *Proc. Inter. WaterSide security conference (WSS)*, pages 1–6, Nov. 2010.

[13] Sara Freitas, Carlos Almeida, Hugo Silva, José Almeida, and Eduardo Silva. Supervised classification for hyperspectral imaging in UAV maritime target detection. In *Proc. IEEE Inter. Conf. on Autonomous Robot Systems and Competitions (ICARSC)*, pages 84–90, 2018.

[14] Sara Freitas, Hugo Silva, José Miguel Almeida, and Eduardo Silva. Convolutional neural network target detection in hyperspectral imaging for maritime surveillance. *Int. J. Adv. Rob. Syst.*, 16(3):1729881419842991, May 2019.

[15] Bo-Cai Gao, Marcos J Montes, Curtiss O Davis, and Alexander FH Goetz. Atmospheric correction algorithms for hyperspectral remote sensing data of land and ocean. *Remote Sense. of Environ.*, 113: S17–S24, 2009.

[16] Ross Girshick. Fast R-CNN. In *Proc. IEEE Inter. Conf. on computer vision (ICCV)*, pages 1440–1448, December 2015.

[17] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. IEEE Conf. on computer vision and pattern recognition (CVPR)*, pages 580–587, June 2014.

[18] Mohammad-Hashem Haghbayan, Fahimeh Farahnakian, Jonne Poikonen, Markus Laurinen, Paavo Nevalainen, Juha Plosila, and Jukka Heikkonen. An efficient multi-sensor fusion approach for object detection in maritime environments. In *Proc. IEEE Inter. Conf. on intelligent transportation systems (ITSC)*, pages 2163–2170, 2018. doi: 10.1109/ITSC.2018.8569890.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL `http://arxiv.org/abs/1512.03385`.

[20] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking ImageNet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4918–4927, 2019.

[21] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H. Hsu. Drone-based object counting by spatially regularized regional proposal network. In *Proc. IEEE Inter. Conf. on computer vision (ICCV)*, pages 4165–4173, 2017. doi: 10.1109/ICCV.2017.446.

[22] Jingliang Hu, Rong Liu, Danfeng Hong, Andrés Camero, Jing Yao, Mathias Schneider, Franz Kurz, Karl Segl, and Xiao Xiang Zhu. MDAS: A new multimodal benchmark dataset for remote sensing. *Earth Syst. Sci. Data Discuss.*, pages 1–26, 2022. doi: 10.5194/essd-2022-155.

[23] Nevrez Imamoglu, Yu Oishi, Xiaoqiang Zhang, Guanqun Ding, Yuming Fang, Toru Kouyama, and Ryosuke Nakamura. Hyperspectral image dataset for benchmarking on salient object detection. In *Proc. Inter. Con. on Quality of Multimedia Experience (QoMEX)*, pages 1–3, 2018. doi: 10.1109/QoMEX.2018.8463428.

[24] Juheon Lee, Xiaohao Cai, Carola-Bibiane Schönlieb, and David A. Coomes. Nonparametric image registration of airborne lidar, hyperspectral and photographic imagery of wooded landscapes. *IEEE Trans. Geosci. Remote Sens.*, 53(11):6073–6084, 2015. doi: 10.1109/TGRS.2015.2431692.

[25] Jun Li, José M Bioucas-Dias, and Antonio Plaza. Spectral–Spatial hyperspectral image segmentation using subspace multinomial logistic regression and markov random fields. *IEEE Trans. Geosci. Remote Sens.*, 50(3):809–823, March 2012.

[26] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proc. IEEE Conf. on computer vision and pattern recognition (CVPR)*, pages 936–944, 2017. doi: 10.1109/CVPR.2017.106.

[27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2999–3007, 2017. doi: 10.1109/ICCV.2017.324.

[28] Ryan Wen Liu, Weiqiao Yuan, Xinqiang Chen, and Yuxu Lu. An enhanced CNN-enabled learning method for promoting ship detection in maritime surveillance system. *Ocean Eng.*, 235:109435, September 2021.

[29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot MultiBox detector. In *Proc. European Conf. on computer vision (ECCV)*, pages 21–37. Springer, 2016.

[30] Spencer Low, Oliver Nina, Angel D. Sappa, Erik Blasch, and Nathan Inkawhich. Multi-modal aerial view object classification challenge results - pbvs 2023. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 412–421, June 2023.

[31] Aditya Mehta, Harsh Sinha, Murari Mandal, and Pratik Narang. Domain-aware unsupervised hyperspectral reconstruction for aerial image dehazing. In *Proc. IEEE winter Conf. on applications on computer vision (WACV)*, pages 413–422, 2021.

[32] F Melgani and L Bruzzone. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Trans. Geosci. Remote Sens.*, 42(8):1778–1790, August 2004.

[33] Sebastian Moosbauer, Daniel König, Jens Jäkel, and Michael Teutsch. A benchmark for deep learning based object detection in maritime environments. In *Proc. IEEE Conf. on computer vision and pattern recognition workshops (CVPRw)*, pages 916–925, 2019. doi: 10.1109/CVPRW.2019.00121.

[34] T. Nathan Mundhenk, Goran Konjevod, Wesam A. Sakla, and Kofi Boakye. A large contextual dataset for classification, detection and counting of cars with deep learning. In *Proc. European Conf. on computer vision (ECCV)*, pages 785–800, 2016. ISBN 978-3-319-46487-9.

[35] Mrunalini Nalamati, Nabin Sharma, Muhammad Saqib, and Michael Blumenstein. Automated monitoring in maritime video surveillance system. In *Proc. Inter. Conf. on image and vision computing New Zealand (IVCNZ)*, pages 1–6, November 2020.

[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Inter. Conf. on Machine Learning (ICML)*, pages 8748–8763, 2021.

[38] Sebastien Razakarivony and Frederic Jurie. Vehicle detection in aerial imagery : A small target detection benchmark. *J. Vis. Commun. Image R.*, 34:187–203, 2016. ISSN 1047-3203. doi: https://doi.org/10.1016/j.jvcir.2015.11.002.

[39] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *Proc. IEEE*, 2016.

[40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time object detection with region proposal networks. In *Proc. neural information processing systems (NeurIPS)*, volume 39, pages 1137–1149, June 2015.

[41] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: An efficient alternative to SIFT or SURF. In *Proc. IEEE Inter. Conf. on computer vision (ICCV)*, pages 2564–2571, 2011.

[42] Chintan A. Shah, Manoj K. Arora, Stefan A. Robila, and Pramod K. Vashney. ICA mixture model based unsupervised classification of hyperspectral imagery. In *Proc. applied imagery pattern recognition workshop (AIPR), 2002. Proceedings.*, pages 29–35, October 2002.

[43] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6700–6713, 2022. doi: 10.1109/TCSVT.2022.3168279.

[44] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Detfusion: A detection-driven infrared and visible image fusion network. In *Proceedings of the ACM International Conference on Multimedia*, pages 4003–4011, 2022.

[45] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proc. IEEE Inter. Conf. on computer vision (ICCV)*, 2019.

[46] Alberto Villa, Jón Atli Benediktsson, Jocelyn Chanussot, and Christian Jutten. Hyperspectral image classification with independent component discriminant analysis. *IEEE Trans. Geosci. Remote Sens.*, 49 (12):4865–4876, December 2011.

[47] Kentaro Wada. Labelme: Image Polygonal Annotation with Python. URL `https://github.com/wkentaro/labelme`.

[48] Nan Wang, Bo Li, Xingxing Wei, Yonghua Wang, and Huanqian Yan. Ship detection in spaceborne infrared image based on lightweight CNN and multisource feature cascade decision. *IEEE Trans. Geosci. Remote Sens.*, 59(5):4324–4339, May 2021.

[49] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. iSAID: A large-scale dataset for instance segmentation in aerial images. In *Proc. IEEE Conf. on computer vision and pattern recognition workshops (CVPRw)*, pages 28–37, 2019.

[50] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proc. IEEE Conf. on computer vision and pattern recognition (CVPR)*, pages 3974–3983, 2018. doi: 10.1109/CVPR.2018.00418.

[51] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *arXiv preprint arXiv:1611.05431*, 2016.

[52] Longbin Yan, Min Zhao, Xiuheng Wang, Yuge Zhang, and Jie Chen. Object detection in hyperspectral images. *IEEE Signal Process. Lett.*, 28:508–512, 2021. doi: 10.1109/LSP.2021.3059204.

[53] Lu Yan, Masahiro Yamaguchi, Naoki Noro, Yohei Takara, and Fuminori Ando. A novel two-stage deep learning-based small-object detection using hyperspectral images. *Opt. Rev.*, 26(6):597–606, December 2019.

[54] Qiguang Yang, Xu Liu, and Wan Wu. A hyperspectral bidirectional reflectance model for land surface. *Sensors*, 20(16), 2020. ISSN 1424-8220. doi: 10.3390/s20164456. URL https://www.mdpi.com/1424-8220/20/16/4456.

[55] Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sunderhauf. Varifocalnet: An iou-aware dense object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8514–8523, 2021.

[56] Yang Zhang, Qing-Zhong Li, and Feng-Ni Zang. Ship detection for visual maritime surveillance from non-stationary platforms. *Ocean Eng.*, 141:53–63, September 2017.

[57] Wang Zhengzhou, Y I N Qinye, L I Hongguang, and H U Bingliang. Surface ship target detection in hyperspectral images based on improved variance minimum algorithm. In *Proc. Inter. Conf. on digital image processing (ICDIP)*, pages 141–147. SPIE, 2016.

[58] Feiyun Zhu. Spectral unmixing datasets with ground truths. *CoRR*, abs/1708.05125, 2017. URL http://arxiv.org/abs/1708.05125.

[59] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(11):7380–7399, 2022. doi: 10.1109/TPAMI.2021.3119563.