remote sensing

MDPI

Technical Note

A Complete YOLO-Based Ship Detection Method for Thermal Infrared Remote Sensing Images under Complex Backgrounds

Liyuan Li [1,2], Linyi Jiang [1,2], Jingwen Zhang [1,2], Siqi Wang [1,2] and Fansheng Chen [1,3,4,*]

[1] State Key Laboratory of Infrared Physics, Shanghai Institute of Technical Physics, Chinese Academy of Sciences, 500 Yu Tian Road, Shanghai 200083, China; liliyuan@mail.sitp.ac.cn (L.L.); jianglinyi@mail.sitp.ac.cn (L.J.); zjw2020@mail.ustc.edu.cn (J.Z.); daqi23333@mail.ustc.edu.cn (S.W.)
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou 310024, China
[4] International Research Center of Big Data for Sustainable Development Goals (CBAS), Beijing 10094, China
* Correspondence: cfs@mail.sitp.ac.cn

Abstract: The automatic ship detection method for thermal infrared remote sensing images (TIRSIs) is of great significance due to its broad applicability in maritime security, port management, and target searching, especially at night. Most ship detection algorithms utilize manual features to detect visible image blocks which are accurately cut, and they are limited by illumination, clouds, and atmospheric strong waves in practical applications. In this paper, a complete YOLO-based ship detection method (CYSDM) for TIRSIs under complex backgrounds is proposed. In addition, thermal infrared ship datasets were made using the SDGSAT-1 thermal imaging system. First, in order to avoid the loss of texture characteristics during large-scale deep convolution, the TIRSIs with the resolution of 30 m were up-sampled to 10 m via bicubic interpolation method. Then, complete ships with similar characteristics were selected and marked in the middle of the river, the bay, and the sea. To enrich the datasets, the gray value stretching module was also added. Finally, the improved YOLOv5 s model was used to detect the ship candidate area quickly. To reduce intra-class variation, the 4.23–7.53 aspect ratios of ships were manually selected during labeling, and 8–10.5 μm ship datasets were constructed. Test results show that the precision of the CYSDM is 98.68%, which is 9.07% higher than that of the YOLOv5s algorithm. CYSDM provides an effective reference for large-scale, all-day ship detection.

Keywords: thermal infrared remote sensing; ship detection; deep learning; intra-class variation

Citation: Li, L.; Jiang, L.; Zhang, J.; Wang, S.; Chen, F. A Complete YOLO-Based Ship Detection Method for Thermal Infrared Remote Sensing Images under Complex Backgrounds. *Remote Sens.* **2022**, *14*, 1534. https://doi.org/10.3390/rs14071534

Academic Editor: Józef Lisowski

Received: 11 March 2022
Accepted: 18 March 2022
Published: 22 March 2022

1. Introduction

With the rapid development of space remote sensing technology, high-resolution and large-scale remote sensing images (RSIs) acquired from space-borne sensors are becoming increasingly enriched, which is promoting widespread use of RSIs. Among their uses, ship detection is of value in civil fields, including maritime transportation, navigation safety, fishery management, ship rescue, and ocean monitoring; and military fields, such as target searching, naval construction, navigation safety, and port monitoring. Therefore, more and more attention has been paid to automatic marine ship monitoring recently.

Existing marine ship target detection methods rely on visible light, synthetic aperture radar (SAR), and infrared imaging technologies. Optical RSIs are the main data source during ship detection due to advantages of broad sensing area, rich spectral features, and high resolution, but it is difficult to identify targets in poor light reflection conditions, especially at night. SAR observations achieve all-weather and long-distance detection, but they are vulnerable to the returns of waves, islands, radio frequency, and atmospheric noise during marine target detection. By measuring the infrared radiation changes caused

by the difference in target temperature, thermal infrared (TI) imaging converts the invisible infrared light into visible content. It has very important applications in hotpot area monitoring, camouflage target disclosure, and military target detection, and is paid attention to by the world's major military powers. Compared with visible-light imaging, TI imaging has the advantage of strong smoke penetration and all-day operation. Differently from SAR, TI imaging passively receives radiation, with good concealment and stronger security. Therefore, TI imaging and target detection have excellent applicability in complex sea status. Based on above problems, a novel all-day ship detection method for thermal infrared remote sensing images (TIRSIs) is proposed in this paper.

When a surface temperature is higher than absolute zero, electromagnetic waves are emitted. Meanwhile, with a change in temperature, the radiation intensity and wavelength distribution of electromagnetic waves change. According to Venn's displacement law, as the absolute black-body temperature T decreases, the peak wavelength λ of spectral radiation tends to increase in size. A ship's surface temperature is near room temperature, so the 8–10.5 μm TI band was selected to find the spectral radiation peak. In 8–10.5 μm, with little solar irradiance, the reflections of the sea surface and sky are the main background thermal radiance, so TIRSIs have great potential applicability for ship detection.

Data show that the average internal temperature of the ocean is 3.8 °C, and the global average surface temperature of the ocean is 17.4 °C. Due to the large temperature difference between day and night, the grayscale intensities of the ship and background are reversed during the day and night. In addition, the brightness of the ship varies with the changes in its motion state and environment, which further increases the difficulty for ship detection. The traditional ship detection methods are based on the manual extraction of features, generally accompanied by low accuracy, human experience dependence, and poor anti-interference, so they struggle to achieve ideal effects.

Among deep learning methods, convolutional neural networks (CNNs) based on texture feature extraction have become the mainstream in ship detection field. With multiple convolution kernels used in down-sampling, multi-level features are extracted by deep CNN. However, the down-sampling in the classical CNN structure has defects during information transmission. Specifically, although the down-sampling can expand the receptive field and reduce the amount of computation in data processing, it leads to a loss of scene information in spatial dimensions. This kind of loss is hard to ignore, especially for small-scale target detection, because it is theoretically impossible to reconstruct the information of a small target after multi-layer sampling. At present, the challenges of TI ship detection are summarized as follows:

1. The security level of TIRSIs is high, and TI ship datasets are lacking.
2. The brightness of the ship varies with the changes in its motion and the state of its environment.
3. Ship detection with TIRSIs is interfered with by complex scenes, which may include the returns of waves, islands, radio frequency, atmospheric noise, and clouds.
4. The variability of ship shape is great, and the textural information of small ships is insufficient.

In view of the above problems, a complete YOLO-based ship detection method (CYSDM) for TIRSIs with complex backgrounds is proposed in this paper, and TI ship datasets were made with the SDGSAT-1 thermal imaging system. As shown in Figure 1, the overall framework of CYSDM contains: preprocessing, TI datasets (overcoming intra-class variation), networks training (the improved lightweight YOLOv5s model), and prediction.

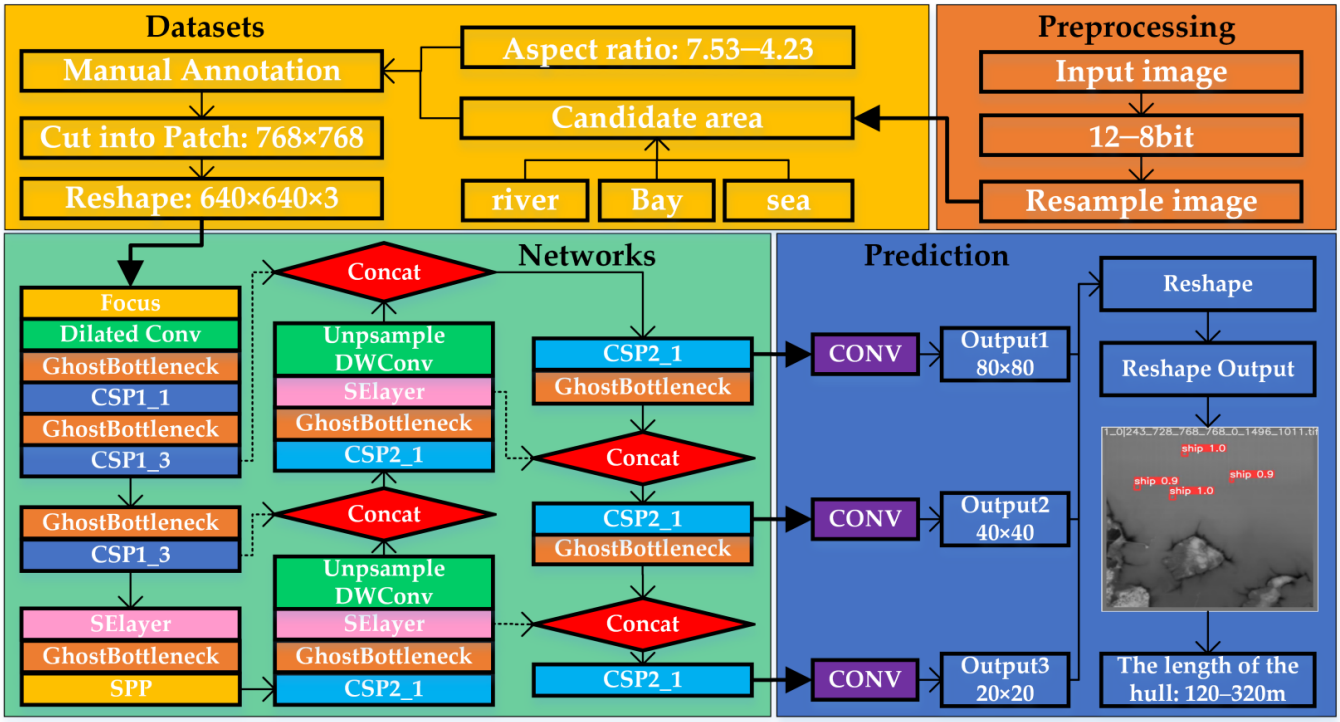

Figure 1. The overall framework of CYSDM: preprocessing, thermal infrared datasets (overcoming intra-class differences), the architecture of networks (the improved lightweight YOLOv5s model), and prediction.

First, the TIRSIs with a resolution of 30 m were up-sampled to 10 m by bicubic interpolation to avoid the loss of texture features caused by large-scale deep convolution. Then, complete ships with similar characteristics were selected and marked, in the middle of a river, a bay, and a sea. To enrich the datasets, the gray value stretching module was also added. Finally, the improved YOLOv5s model was used to detect a candidate area quickly. To reduce intra-class variation, the 7.53–4.23 aspect ratios of ships are manually selected during labeling, and 8–10.5 μm TI ship datasets were constructed. Test results show that the accuracy of the CYSDM is 98.68%, which is 9.07% higher than that of the YOLOv5s algorithm. The main contributions of this paper are as follows:

1. In response to the lack of TI ship datasets, 8–10.5 μm TI ship datasets were established. To reduce intra-class variation, complete vessels with aspect ratios of 7.53–4.23 in the middle of a river, a bay, and the sea were specifically selected.
2. In view of the variability of ship size, SE layer and dilated convolution modules with an enlarged receptive field were designed for the top of the network to retain more semantic features. The lightweight improved YOLOv5s algorithm should be used for ship detection of large-scale TIRSIs.
3. Ships of different scales in complex scenes, covered by clouds and fog, are detected efficiently by CYSDM. The proposed method provides an effective reference for large-scale, all-day ship detection.

The structure of this paper is as follows. In Section 2, we introduce the previous related research in this filed. The coarse and fine detection methods proposed in this work are elaborated in detail in Section 3. The experimental results and analysis are described in Section 4. Finally, in Section 5, we summarize the content of this study.

2. Previous Related Research

Hull and wake detection are two main methods used by maritime ship perception algorithms on remote sensing images. However, ship wake does not always exist, so

hull detection is more widely used. In general, feature extraction algorithms for ships are broadly classified into traditional and intelligent methods.

In traditional ship detection methods based on prior knowledge, binarization, threshold segmentation, or morphology methods [1–3] are adopted first to divide targets and background, and then geometric and texture features are extracted. The background and the target ships with obvious geometric features (fixed size, high compactness, and uniform bow shape) are segmented accurately by traditional methods. However, when the ship is blocked by clouds or docks, the geometric segmentation methods are not applicable, and texture features need to be added. However, when the ship is obscured by clouds, texture features are added during segmentation because the geometric features are not obvious. The method of extracting texture features (energy, inertia, entropy, and correlation) by statistics, geometry, modeling, and signal processing is beneficial to the analysis of overall image features because of its rotation invariance. However, due to the different resolutions of RSIs, the texture features of the same target vary greatly.

Classical operator ship detection methods contain candidate region extraction and fine target discrimination. In candidate region extraction, texture description of the local binary mode is important for HOG, and SIFT [4] is used to smooth the background during manual feature extraction. Then, support vector machine (SVM), extreme learning machine (ELM), k-nearest neighbor (KNN), linear discriminant analysis (LDA), or adaptive boosting is used for fine classification. The detection results of traditional methods based on low-level manual features are good for specific datasets in calm seas, but it is difficult to detect infrared targets in complex scenes due to weak generalization. In addition, the on-orbit applications of the above algorithms are limited by great computational complexity.

In 2014, with the rapid development of deep learning technology, CNN-based target feature extraction algorithms were widely used for ship detection [5]. Zhou et al. [6] proposed an improved pyramid network module for adaptive feature fusion for SAR images to select the best feature for multi-scale target detection tasks. By using semantic knowledge, including the relations among features, attributes, tags and classes, Xue et al. [7] proposed a deep multimodal relation learning method for inverse synthetic aperture radar (ISAR) images to effectively deal with complex multimodal recognition problems and improve the accuracy and speed of the whole system. For ISAR images, Rajkumar et al. [8] combined three model-based deep learning methods to vote at feature level and decision level, so as to combine the advantages of the two methods and achieve higher performance. Compared with infrared imaging, SAR images have higher resolution and richer texture information, making them more conducive to target detection.

By capturing multi-scale context information, Han et al. [9] proposed a multi-vision small-object detector that uses visible-light RSIs for the accurate detection of aircraft, cars, and ships at high speed with 24 FPS on a NVIDIA 1080Ti GPU of DELL. He et al. [10] proposed an offshore ship detection method that uses weighted voting and rotation-scale-invariant poses for high-resolution satellite images. However, it is difficult to extract the rotation angle, position, and scale factor of the ship from infrared images because of their low resolution.

Li et al. [11] proposed a detection method for low-contrast infrared targets of unknown size and polarity (bright or dark) that uses background subtraction and logarithmic histogram transformation to enhance the target. This method heavily relies on the difference in contrast between target and background, but infrared images always have the problems of low contrast and the serious non-uniformity. In addition, a large number of labeled samples are required during training; it is challenging to generate a CNN model with limited infrared samples.

Supervised learning is a good choice for machine learning problems, but high-quality datasets are necessary. When sufficient training datasets are not available, semi-supervised learning is a potential solution. Partially labeled and mostly unlabeled data can be used by semi-supervised learning. Based on a regional proposal network and weak, semi-supervised, deep sparse learning, a fast and efficient weak semi-supervised method for

ISAR was proposed by Xue et al. [12]. Self-supervision methods can be regarded as special forms of unsupervised learning methods with supervision. Ciocarlan et al. [13] utilized self-supervised learning to learn the features of large annotated datasets of Sentinel 2 images, and used small sample transfer learning to detect ships. However, this model requires large batch sizes and a long time for training. In addition, unsupervised learning cannot eliminate possible biases in system predictions, unless unsupervised models are specifically trained by additional datasets.

In contrast to man-made objects, the change in gray scale distribution on a smooth sea is slow, so the segmentation between sea and land is utilized for the preprocessing. The segmentation method based on median filter was used to reduce the interference of speckle noise and the overlapping between land and docks by Xu et al. [14]. However, the sea–land segmentation methods are limited by being specialized for seas, and are not suitable for ship detection in rivers. The discrimination of dim, small targets in infrared images not only relies on the geometric texture features, but also depends on the scene where the targets appear [15,16]. A scene classification method is proposed, which is used to distinguish similar scenes in multiple images and classify them correctly. The above target detection methods are limited by the accuracy of the extraction of image features and scene classification, which increases the uncertainty of prediction results. Considering the above problems, a ship detection method for 8–10.5 μm remote sensing images is proposed.

3. Materials and Methods

3.1. Thermal Infrared Ship Datasets

Research on common ship sizes and ship datasets are summarized as Tables 1 and 2. Due to the scarcity of infrared RSIs for ship detection, TI ship datasets of 8–10.5 μm were constructed by collecting the data of the SDGSAT-1 thermal imaging system. To reduce intra-class variations, complete ships with similar characteristics were selected specifically. According to statistics, complete ships with 7.53–4.23 aspect ratios in the middle of a river, on the sea surface, and on shore were annotated.

Table 1. The summary of common ship size.

Category	The Length/m	The Width/m	Mean Aspect Ratio
Aircraft Carrier	160.0–343.0	33.0–76.9	2.08–4.46
Amphibious Ship	70.0–250.0	8.5–40.0	1.75–6.25
Cruiser	142.0–247.5	15.8–27.5	5.16–8.95
Destroyer	112.0–171.0	10.2–16.4	6.83–10.42
Frigate	81.0–138.0	9.1–14.3	5.66–9.65
Mean	113.0–229.9	15.3–35.0	4.29–7.96

Table 2. The summary of ship datasets.

Datasets	Satellite	Bands	Resolution	Annotations
The Annotated Datasets	SDGSAT-1-TIS	8–10.5 μm	30 m	Ship (length: 120–320 m, aspect ratio: 7.53–4.23)
DOTA-v1.5 [17]	Google Earth, JL-1, GF-2	0.45–0.89 μm	0.04–1 m	17 classes (including Ship)
DOTA-v2.0 [18]	Aerial	0.45–0.89 μm	0.04–1 m	17 classes (including Ship)
NWPU VHR-10 [19]	optical images	0.38–0.76 μm	0.04–1 m	10 classes (including Ship)
DIOR [20]	SAR	0.38–0.76 μm	0.04–1 m	20 classes (including Ship)
HRSID [21]	Google Earth, JL-1, GF-2	1300–1.76 mm	0.5–3 m	Ship (SAR images of different resolutions, polarities, sea areas, and ports)
SSDD [22]	Aerial	1300–1.76 mm	0.5–3 m	
SAR-Ship-Dataset [23]	GF-3, Sentinel-1	1300–1.76 mm	3–25 m	

To make display, annotation, and training easy, the TIF is linearly converted from 12 to 8 bits. Then, the TIRSIs with a resolution of 30 m are up-sampled to 10 m to avoid the loss of texture characteristics caused by large-scale deep convolution. The methods of up-sampling mainly include: nearest, bilinear, and bicubic interpolations, and spectral

techniques [24]. The image quality after the first two processing methods is not high because of the properties of the low-pass filter, so the bicubic interpolation was chosen for up-sampling. After up-sampling, the large images of 30,000 × 30,000 were cropped to small patches of 768 × 768 with labels. The number of dataset patches was 1008, and each small patch contained 1–5 ships in the sea surface scenes. In the river scenes, the number of ships in one patch can reach more than 8.

The texture features of infrared dim small targets were very scarce, so it was impossible to accurately label ship targets. Therefore, ship occurrence area images on Google Maps were taken as references during annotations, as shown in Figure 2a. Compared with visible-light images, infrared targets are easily submerged in complex backgrounds due to the lack of texture information and low contrast with the background. In the infrared images, the sizes and aspect ratios of the complete ships and the ships adjacent to the shore are very different.

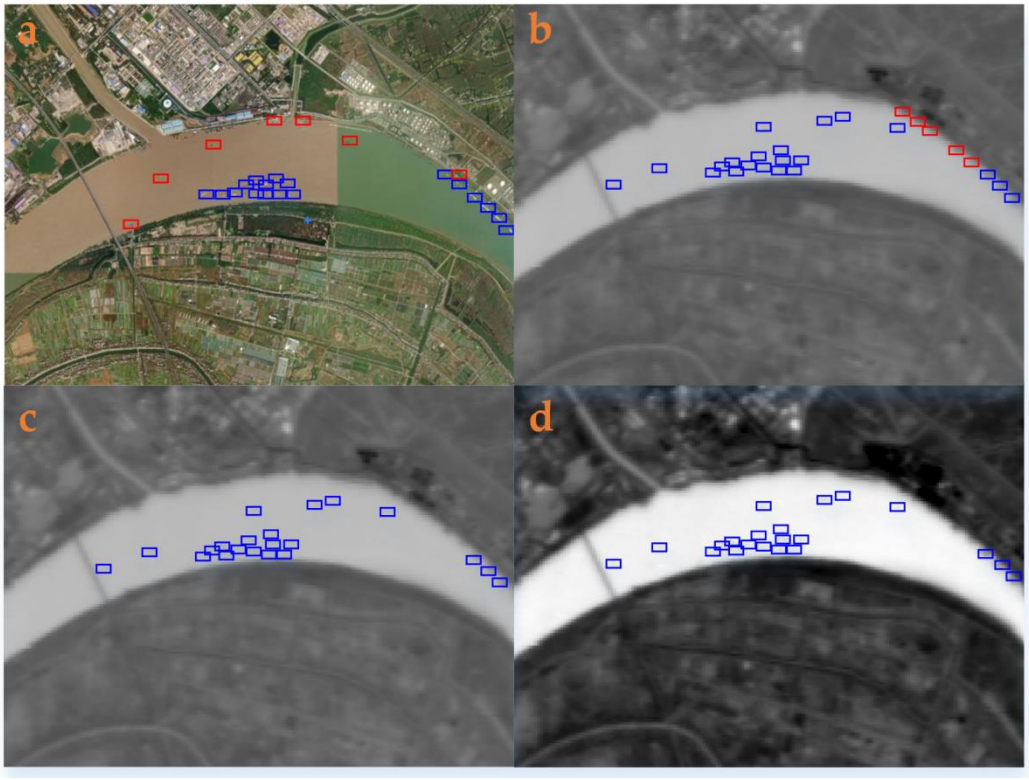

Figure 2. Jiajiang River local RSIs: (**a**) Google Maps images. TIRSIs: (**b**) annotations all types of vessels, (**c**) annotations of select vessels (intra-class variations considered), (**d**) data of images after contrast stretching (in TIRSIs: a blue box is a complete ship, a red box is a partial ship).

In infrared images, the geometric characteristics of complete ships at sea and ships adjacent to the shore are very different. Therefore, in the labeling process, in order to overcome the problem of a low detection rate caused by excessive intra-class differences, the complete ships with 7.53–4.23 aspect ratios were manually selected for truth labeling, as shown in Figure 2c. Due to the great difference in geometric features between the ships in the middle of a river and ships adjacent to the shore, complete ships with 7.53–4.23 aspect ratios were manually selected during labeling to overcome the problem of low accuracy caused by excessive intra-class differences, as shown in Figure 2c. To enrich the TI datasets, a linear contrast stretching module is added, as shown in Figure 2d. Due to the large size of the RSIs, TI images were cropped to 768 × 768 pixels. To ensure the integrality of vessels, there were overlapping areas during cutting. Except for small fuzzy ships, all ship targets in TIRSIs were manually marked as ground truth, and an 8–10.5 μm ship dataset was constructed.

3.2. The Improved YOLOv5s Algorithm

Driven by the excellent development of GPU, targets detection methods based on deep CNN have been greatly promoted. According to whether the proposed regions are generated or not, the intelligent methods are roughly classified into one-stage [25–28] and two-stage [29–31] detection methods. The advantage of two-stage models is high positioning accuracy, whereas one-stage detection models have the advantage in speed. In practical applications of real-time ocean observation and timely ship rescue, reducing prediction time is as important as improving detection accuracy. To ensure real-time detection by a model deployed on edge computing devices, an improved model based on a one-stage YOLOv5s was created, and Dilated Conv [32], depthwise convolution (DWConv) [33], and SELayer [34] modules were added.

The architecture of the networks is shown in Figure 1. Additionally, the details of the model, including the number of modules, the number of cycles, parameters, module names, and arguments, are summarized in Table 3. First, mosaic data enhancement, adaptive anchor calculation, and adaptive image scaling letterbox are used before the input to enrich the datasets and improve robustness. Focus, GhostBottleneck, and CSP structure [35] are utilized in the backbone; and the Dilated Conv module [32] and SELayer module [34] were designed for feature extraction of targets of different sizes. In Focus, the input image of 640 × 640 × 3 is firstly sliced into 320 × 320 × 12, and then after being convoluted by 32 convolution kernels, the feature maps of 320 × 320 × 32 are finally output. The Neck, FPN [36], PAN [27], and DWConv modules are used during up-sampling to extract multi-scale information at different stages in the modified network.

Table 3. The summary of the improved Yolov5s networks.

From	*n*	Parameters	Module	Arguments
−1	1	3520	Focus	[3, 32, 3]
−1	1	704	DWConv	[32, 64, 3, 2]
−1	1	3440	GhostBottleneck	[64, 64, 3, 1]
−1	1	18,784	GhostBottleneck	[64, 128, 3, 2]
−1	3	32,928	GhostBottleneck	[128, 128, 3, 1]
−1	1	2048	SELayer	[128, 16]
−1	1	66,240	GhostBottleneck	[128, 256, 3, 2]
−1	3	115,008	GhostBottleneck	[256, 256, 3, 1]
−1	1	8192	SELayer	[256, 16]
−1	1	5632	DWConv	[256, 512, 3, 2]
−1	1	656,896	SPP	[512, 512, [5, 9, 13]]
−1	1	32,768	SELayer	[512, 16]
−1	1	131,584	Conv	[512, 256, 1, 1]
−1	1	0	Upsample	[None, 2, 'nearest']
[−1, 6]	1	0	Concat	[1]
−1	1	361,984	C3	[512, 256, 1, False]
−1	1	33,024	Conv	[256, 128, 1, 1]
−1	1	0	Upsample	[None, 2, 'nearest']
[−1, 4]	1	0	Concat	[1]
−1	1	90,880	C3	[256, 128, 1, False]
−1	1	147,712	Conv	[128, 128, 3, 2]
[−1, 14]	1	0	Concat	[1]
−1	1	394,752	C3	[640, 256, 1, False]
−1	1	590,336	Conv	[256, 256, 3, 2]
[−1, 10]	1	0	Concat	[1]
−1	1	1,313,792	C3	[768, 512, 1, False]
[17, 20,23]	1	131,325	Detect	[80, [[116, 90, 156, 198, 373, 326], [30, 61, 62, 45, 59, 119], [10, 13, 16, 30, 33, 23]], [128, 128, 256]]

From represents which layer to start with, and *n* represents cycle index.

In the head of the network, the standard convolution (Conv) in the YOLOv5s model was replaced by the Dilated Conv module. To not reduce the spatial information, the

corresponding receptive field index is increased by the Dilated Conv. In other words, more space feature fusion is integrated, and multi-scale spatial information is extracted. Dilated Conv is shown in Equation (1), where F is a discrete function, $*_l$ is dilated convolution, K is a discrete filter, s is step length, l is the input stride, and t is an independent variable. $F(s)$ is a discrete convolution of step s, and $K(t)$ is a discrete filter with independent variable t. The formula of dilated convolution receptive field is shown as Equation (2), where RF_{i-1} is the receptive field of the upper layer, and k is the convolution kernel's size. The positions not occupied by the standard convolution kernels are filled with 0, and the dilated convolution kernel size is shown in Equation (3).

The parameter ratio of DWConv (*Param_DWConv*) to Conv (*Param_Conv*) is shown in Equation (4). The convolution kernel used is 3×3, so the parameters of DWConv include about $1/9$ those of Conv. As more and more features are extracted, instead of Conv, the computing resources are greatly saved by using DWConv to improve detection speed.

$$(F *_l K)(s + lt) = \sum_{s+lt=p} F(s)K(t) \tag{1}$$

$$RF_i = RF_{i-1} + (k-1) \times s \tag{2}$$

$$kernel\_dilation = dilation * (kernel - 1) + 1 \tag{3}$$

$$\frac{Param\_DWConv}{Param\_Conv} = \frac{C_{in} \times k \times k + C_{in} \times 1 \times 1 \times C_{out}}{C_{in} \times k \times k \times C_{out}} = \frac{1}{C_{out}} + \frac{1}{k \times k} \tag{4}$$

The improved lightweight ship detection model based on the YOLOv5s model was used to verify the dataset annotated by us. The model has the following advantages. Aiming at the problem of failed detection caused by large hull length differences, standard convolution with a small receptive field is used at the bottom of the network. Meanwhile, the Dilated Conv with large receptive field was put at the top of the network to retain more semantic features, which facilitates feature extraction of targets of different sizes. Additionally, the Conv were replaced by DWConv to reduce the number of parameters during down-sampling. The SELayer module was added to filter for more important features.

4. Experimental Analysis

4.1. Evaluation Criteria

Precision and recall are two evaluation criteria we used. True positive (*TP*), true negative (*TN*), false positive (*FP*), false negative (*FN*) are shown in Table 4. Precision represents the probability of correctly predicting positive samples among positive predictions, as shown in Equation (5). Recall represents the probability of correctly categorizing a true positive sample, as shown in Equation (6). Mean average precision (*mAP*) is the area enclosed under the precision–recall curve calculated by integration, as shown in Equation (7). The complexity of CNN without deviation is measured by parameters and floating-point operations (FLOPs), as shown in Equations (8) and (9), where $k_H \times k_W \times C_{in}$ represents convolution kernel size, C_{out} represents output channels, g represents group convolution number, and $C_{out} \times H_{out} \times W_{out}$ represent outputs. In conclusion, high *Precision*, *Recall*, and *mAP* are desired, along with low *FLOPs*.

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

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

$$mAP = \int_0^1 p(r)dr \tag{7}$$

$$Parameters = k_H \times k_W \times C_{in}/g \times C_{out} \tag{8}$$

$$FLOPs = (2 \times k_H \times k_W \times C_{in}/g - 1) \times C_{out} \times H_{out} \times W_{out} \tag{9}$$

Table 4. Confusion matrix.

Ground Truth	Predicted Class	
	Ship	**Non Ship**
Ship	*TP*	*FN*
Non Ship	*FP*	*TN*

4.2. Comparative Experiments

The data represent that the total length of the largest Chinese container ship, namely, Cosco Asia, is 349 m. The heaviest cargo ship (loaded) on the Yangtze River, namely, Baosheng Changyang, is 120 m long and 21 m wide. The range of ship lengths detected by the proposed method is 12–32 pixels, that is, 120–320 m, and the aspect ratio can be 7.53–4.23. The experiment was performed on a desktop computer with 470.57.02 NVIDIA-SMI, 470.57.02 driver version, 11.4 CUDA version, NVIDIA RTX 3070 GPU of Dell, and Pytorch framework. The SGD optimizer was utilized to update the network's weights. The initial learning rate was set to 0.01 for the first 300 iterations and 0.001 for the last 300 iterations. The weight decay of the optimizer was 0.001, and the momentum was 0.98.

To evaluate the validity of the algorithm, TIRSI datasets of a variety of scenes, including oceans, harbors, rivers, and islands, captured under different light and weather conditions, were tested. As shown in Figure 3, with the increase in epoch number, the mAP of the proposed method (CYSDM) became gradually more stable and was higher than YOLOv5s and the improved YOLO-based networks without considering the intra-class variation (IC-CYSDM). As shown in Table 5, the precision of the proposed method was 7.45% and 0.89% higher than the precision of YOLOv5s and the improved YOLO-based networks without considering the intra-class variation (IC-CYSDM), respectively. Additionally, the FLOPs of CYSDM totaled only 54.38% of those of YOLOv5s.

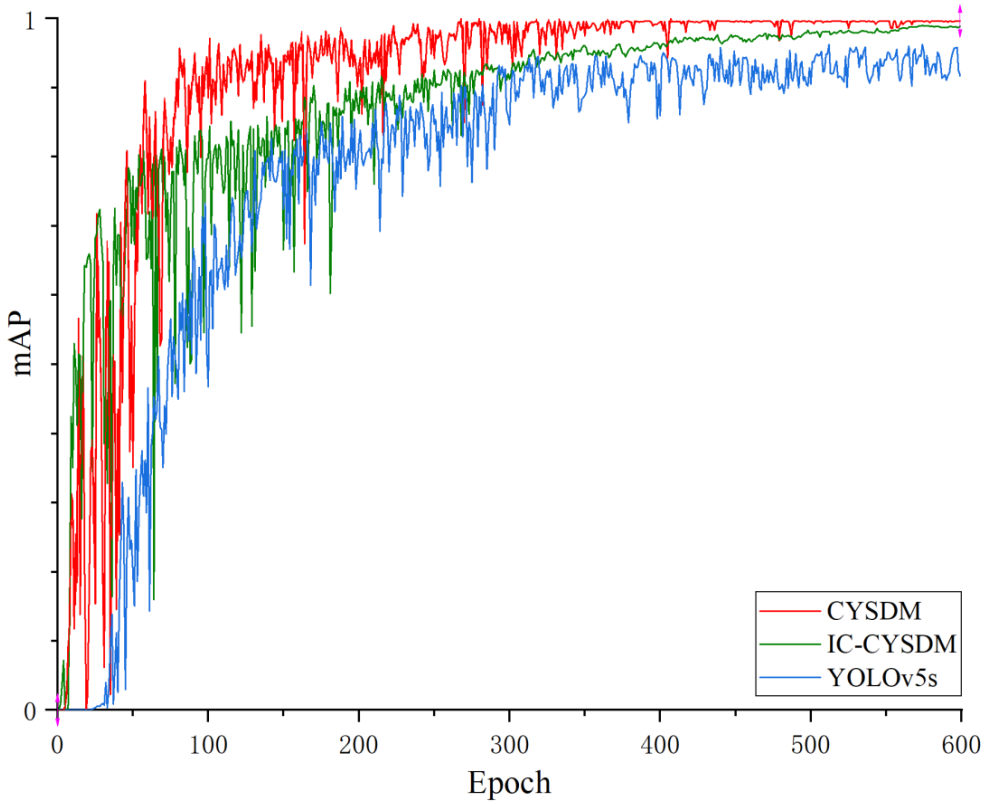

Figure 3. The mAP of the proposed method (CYSDM), the improved YOLO-based networks without considering the intra-class variation (IC-CYSDM), and the YOLOv5s algorithm.

Table 5. Experimental ship detection results for several models.

Model	Image Size	Batch Size	Precision (%)	Recall (%)	GFLOPs	Layers	Parameters
CYSDM	640	8	98.68	98.67	9.3	390	4.14 M
IC-CYSDM	640	8	96.19	96.07	9.3	390	4.14 M
YOLOv5s	640	8	89.61	91.99	17.1	283	7.28 M
YOLOv3 [37]	320	8	85.91	82.33	38.8	252	20.40 M
Faster R-CNN [31]	320	8	83.26	84.89	46.7	-	31.25 M
SSD512 [38]	320	4	87.34	86.75	19.6	-	138.0 M

GFLOPs = 10^9 FLOPs, 1 M = 10^6.

With the computing limitations of the desktop computer, the image size was set to 320 and the batch size to eight while training YOLOv3. Additionally, image size was set to 640 for the training of lightweight models CYSDM, IC-CYSDM, and YOLOv5s. As the image size increases, more texture and context discriminating features are captured. In a certain range, with an increase in the batch size, the shock is smaller during training to accurately reduce the loss. The predictions of the proposed method (CYSDM) in different scenes, including the sea, a bay, a river, and under cloud cover, are shown in Figure 4.

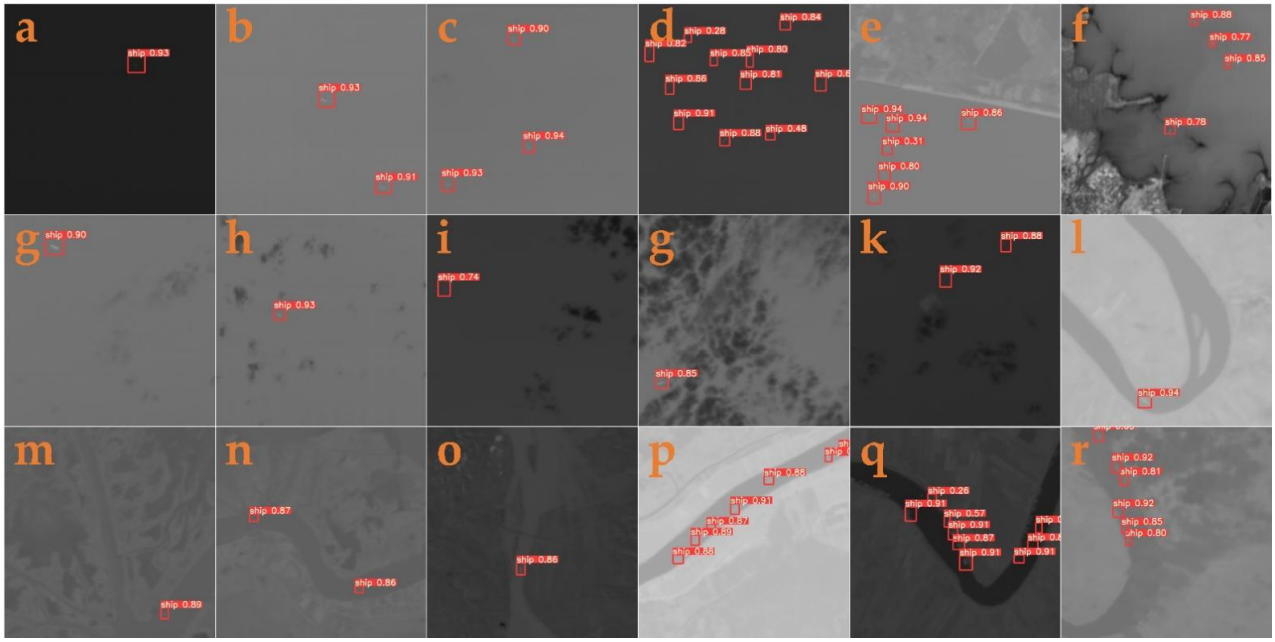

Figure 4. The predictions of the proposed method (CYSDM) in different scenes, including: (**a**–**d**) a calm sea, (**e**,**f**) a bay, (**g**–**k**) the sea under cloud cover, (**l**–**q**) a river, (**r**) a river under cloud cover.

5. Discussion

Visible remote sensing images have the advantage of high resolution and have been widely used in ship detection. However, visible RSIs are based on the reflection of light, and it is difficult to see and identify targets in the completely dark environment or conditions where the light is not strong enough. SAR technology can work all day and has a long detection range. However, radar observation is susceptible to interference from echoes of waves, islands and land masses, radiofrequency noise, and atmospheric noise, which makes it difficult to detect targets on the sea surface. By measuring changes in infrared radiation caused by differences in targets temperature and radiation, infrared thermal imaging converts invisible infrared light into visible content. It has special application value in hotspot area monitoring, camouflage target disclosure, and military target detection, so it is paid attention to by the major military powers. Compared with visible light, infrared has

the advantages of strong smoke penetration and all-day usability. Differently from SAR, infrared imaging passively receives radiation, providing good concealment and strong security. Therefore, target detection with infrared images has excellent applicability in complex sea conditions.

Our approach is based on supervised training. High quality, balanced, standardized, and thoroughly cleaned datasets are required; otherwise, supervised learning will yield poor results. In the 30 m resolution TI images, the geometric features of ships close to the riverbank are very different from those of the ships in the middle of the river or at sea. The aspect ratios of different types of ships were summarized and calculated thanks to literature research. According to that research, which we did before making the datasets, only ships with aspect ratios of 4.23–7.53 were annotated. This was not noticed in our initial work, and based on the datasets originally annotated, the IC-CYSDM model was obtained. To increase the reliability of the datasets, the aspect ratios of the labeled ships were carefully chosen to obtain good final datasets, which were used to train the CYSDM model. Through the experimental comparison, large differences in ships' aspect ratios, namely, large intra-class variation, led to poorer detection efficiency by the network. The experimental results after eliminating intra-class variation show that the proposed method (CYSDM) is suitable for ship detection in complex scenes, including seas, bays, and rivers, all with cloud cover. Additionally, it provides a reference for sensitive TI ship target searching for all parts of the day. However, the detection results of the proposed method were poor for ships close to the shore and adjacent vessels. To solve those problems, multi-frame information will be utilized in the future work.

6. Conclusions

The experimental results show that the proposed method is superior to YOLOv3 and Yolov5s, and the labeled TI ship datasets were tested with different models for their sea, bay, river, and cloud-cover images. The accuracy of CYSDM in this study was 98.68%, which is 2.49% higher than that of IC-CYSDM and 9.07% higher than that of YOLOv5s. Through extensive literature research, the challenges of on-orbit infrared ship detection were identified: (1) The fuzzy segment leads to false positives at the junction of land and sea. (2) The prediction results of infrared ship targets are easily interfered with by atmospheric noise, clouds, or reefs. (3) It is difficult to locate vessels of various shapes and sizes simultaneously. (4) In view of the limitations of computing resources, the accuracy and computational complexity of the algorithm need to be balanced. (5) There are few TI datasets, especially with ships, and a lot of manpower and material resources are required during annotation. Therefore, the TI ship datasets will continue to be enriched and refined, and the lightweight model will be properly deployed to a hardware platform in the future.

Author Contributions: Conceptualization, L.L. and S.W.; methodology, L.L.; software, L.L.; validation, L.J. and S.W.; formal analysis, J.Z.; investigation, L.J.; resources, F.C.; data curation, F.C.; writing—original draft preparation, L.L.; writing—review and editing, L.L.; visualization, J.Z.; supervision, L.L.; project administration, F.C.; funding acquisition, F.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by the Strategic Priority Research Program of the Chinese Academy of Sciences, grant number XDA19010102, and National Natural Science Foundation of China under grant number 61975222.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank the SDG BIG DATA Center and National Space Science Center for providing us with data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dellinger, F.; Delon, J.; Gousseau, Y.; Michel, J.; Tupin, F. SAR-SIFT: A SIFT-like algorithm for SAR images. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 453–466. [CrossRef]
2. Qin, X.; Zhou, S.; Zou, H.; Gao, G. A CFAR detection algorithm for generalized gamma distributed background in high-resolution SAR images. *IEEE Geosci. Remote Sens. Lett.* **2012**, *10*, 806–810.
3. Li, Z.; Itti, L. Saliency and gist features for target detection in satellite images. *IEEE Trans. Image Process.* **2011**, *20*, 2017–2029. [PubMed]
4. Qi, S.; Ma, J.; Lin, J.; Li, Y.; Tian, J. Unsupervised ship detection based on saliency and S-HOG descriptor from optical satellite images. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 1451–1455.
5. Chen, C.; Wang, B.; Lu, C.; Trigoni, N.; Markham, A. A Survey on Deep Learning for Localization and Mapping: Towards the Age of Spatial Machine Intelligence. *arXiv* **2020**, arXiv:2006.12567.
6. Zhou, K.; Zhang, M.; Wang, H.; Tan, J. Ship Detection in SAR Images Based on Multi-Scale Feature Extraction and Adaptive Feature Fusion. *Remote Sens.* **2022**, *14*, 755. [CrossRef]
7. Xue, B.; Tong, N. Real-World ISAR Object Recognition Using Deep Multimodal Relation Learning. *IEEE Trans. Cybern.* **2020**, *50*, 4256–4267. [CrossRef]
8. Theagarajan, R.; Bhanu, B.; Erpek, T.; Hue, Y.K.; Schwieterman, R.; Davaslioglu, K.; Shi, Y.; Sagduyu, Y.E. Integrating deep learning-based data driven and model-based approaches for inverse synthetic aperture radar target recognition. *Opt. Eng.* **2020**, *59*, 051407. [CrossRef]
9. Han, W.; Kuerban, A.; Yang, Y.; Huang, Z.; Liu, B.; Gao, J. Multi-Vision Network for Accurate and Real-Time Small Object Detection in Optical Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
10. He, H.; Lin, Y.; Chen, F.; Tai, H.-M.; Yin, Z. Inshore Ship Detection in Remote Sensing Images via Weighted Pose Voting. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3091–3107. [CrossRef]
11. Li, Y.; Li, Z.; Xu, B.; Dang, C.; Deng, J. Low-Contrast Infrared Target Detection Based on Multiscale Dual Morphological Reconstruction. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
12. Xue, B.; Tong, N. DIOD: Fast and Efficient Weakly Semi-Supervised Deep Complex ISAR Object Detection. *IEEE Trans. Cybern.* **2019**, *49*, 3991–4003. [CrossRef] [PubMed]
13. Ciocarlan, A.; Stoian, A. Ship Detection in Sentinel 2 Multi-Spectral Images with Self-Supervised Learning. *Remote Sens.* **2021**, *13*, 4255. [CrossRef]
14. Xu, J.; Sun, X.; Zhang, D.; Fu, K. Automatic Detection of Inshore Ships in High-Resolution Remote Sensing Images Using Robust Invariant Generalized Hough Transform. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 2070–2074.
15. Cheng, X.; Lu, J.; Feng, J.; Yuan, B.; Zhou, J. Scene recognition with objectness. *Pattern Recognit.* **2018**, *74*, 474–487. [CrossRef]
16. Han, Y.; Yang, X.; Pu, T.; Peng, Z. Fine-Grained Recognition for Oriented Ship Against Complex Scenes in Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–18. [CrossRef]
17. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983.
18. Ding, J.; Xue, N.; Xia, G.S.; Bai, X.; Yang, W.; Yang, M.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; et al. Object Detection in Aerial Images: A Large-Scale Benchmark and Challenges. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**. [CrossRef]
19. Cheng, G.; Zhou, P.; Han, J. Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 7405–7415. [CrossRef]
20. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [CrossRef]
21. Wei, S.; Zeng, X.; Qu, Q.; Wang, M.; Su, H.; Shi, J. HRSID: A High-Resolution SAR Images Dataset for Ship Detection and Instance Segmentation. *IEEE Access* **2020**, *8*, 120234–120254. [CrossRef]
22. Yang, X.; Sun, H.; Fu, K.; Yang, J.; Sun, X.; Yan, M.; Guo, Z. Automatic Ship Detection in Remote Sensing Images from Google Earth of Complex Scenes Based on Multiscale Rotation Dense Feature Pyramid Networks. *Remote Sens.* **2018**, *10*, 132. [CrossRef]
23. Wang, Y.; Wang, C.; Zhang, H.; Dong, Y.; Wei, S. A SAR Dataset of Ship Detection for Deep Learning under Complex Backgrounds. *Remote Sens.* **2019**, *11*, 765. [CrossRef]
24. Ghaderpour, E.; Pagiatakis, S.D.; Hassan, Q.K. A Survey on Change Detection and Time Series Analysis with Applications. *Appl. Sci.* **2021**, *11*, 6141. [CrossRef]
25. Li, L.; Li, X.; Liu, X.; Huang, W.; Hu, Z.; Chen, F. Attention Mechanism Cloud Detection with Modified FCN for Infrared Remote Sensing Images. *IEEE Access.* **2021**, *9*, 150975–150983. [CrossRef]
26. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 21–26 July 2017; pp. 2999–3007. [CrossRef]
27. Bochkovskiy, A.; Wang, C.; Liao, H. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
28. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020.

29. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
30. Girshick, R.; Fast, R.C.N.N. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448. [CrossRef]
31. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
32. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv* **2016**, arXiv:1511.07122.
33. Andrew, H.; Mark, S.; Grace, C.; Liang-Chieh, C.; Bo, C.; Mingxing, T.; Weijun, W.; Yukun, Z.; Ruoming, P.; Vijay, V. Searching for MobileNetV3. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27–28 October 2019; pp. 1314–1324. [CrossRef]
34. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef]
35. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. GhostNet: More Features from Cheap Operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 1577–1586.
36. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
37. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
38. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In *European Conference on Computer Vision*; Computer Vision (ECCV) Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2016; Volume 9905. [CrossRef]