YOLOv7-sea: Object Detection of Maritime UAV Images based on Improved YOLOv7

Hangyue Zhao
Beijing University of Posts and Telecommunications
zhaohy21315@bupt.edu.cn

Hongpu Zhang
Beijing University of Posts and Telecommunications
zhp@bupt.edu.cn

Yanyun Zhao
Beijing University of Posts and Telecommunications
Beijing Key Laboratory of Network System and Network Culture, China
zyy@bupt.edu.cn

Abstract

Object detection algorithms play an important role in maritime search and rescue missions, where they are designed to detect people, boats and other objects in open water. However, the SeaDronesee dataset has the characteristics of small targets and large sea surface interference, which brings great challenges to general object detectors. To address these issues, we propose an improved detector YOLOv7-sea. Based on YOLOv7[2], we add a prediction head to detect tiny-scale people or objects. Besides, we integrate Simple, Parameter-Free Attention Module (SimAM) to find attention regions in the scene. To achieve further improvements to our proposed YOLOv7-sea, we provide some useful strategies such as data augmentation, Test time augmentation (TTA), and bundled box fusion (WBF). On the ODv2 challenge dataset, the AP result of YOLOv7-sea is 59.00% , which is about 7% higher than the baseline model (YOLOv7).

1. Introduction

In the past few years, computer vision applications in marine and freshwater fields have developed rapidly. For accurate navigation in heavy traffic or close to the coast, computer vision is increasingly important. MaCVi 2023 workshop aims to promote the use of modern computer vision methods in a variety of air and surface water fields. The workshop includes four challenges designed to advance the development of computer vision algorithms for Search and Rescue (SAR) missions and autonomous ship navigation. We mainly discuss the object detection task in this article.

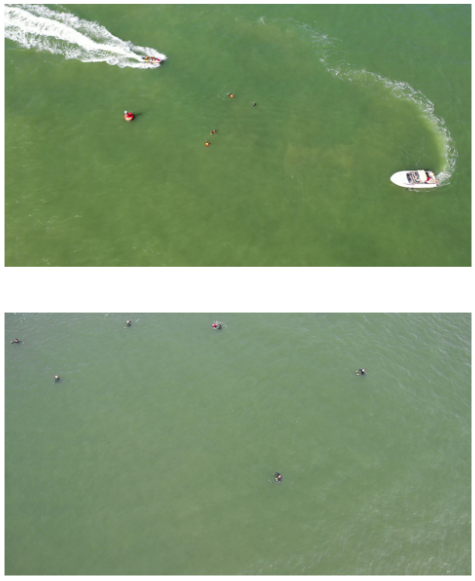

Figure 1. Instances of the SeaDronesee dataset. The target size in the picture is small and difficult to distinguish. At the same time there is disturbance on the waves.

SeaDronesSee Object Detection is aimed at detecting humans, boats and other objects in open water. Although a lot of progress has been made in object detection, most of the previous models are designed for natural scene images, and they have good results for medium and large objects. Us-

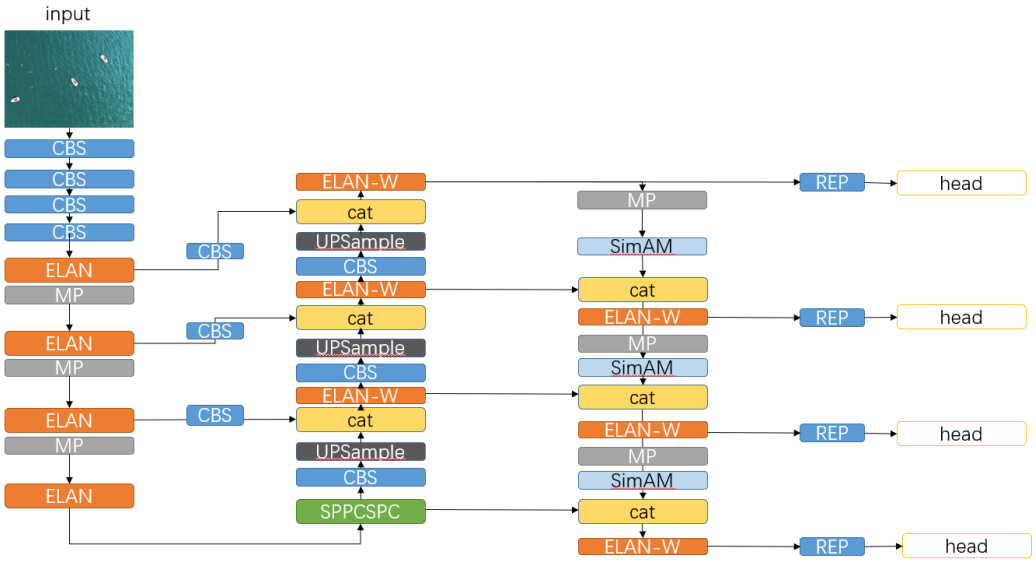

Figure 2. The architecture of YOLOv7-sea.

ing them directly to detect objects in maritime drone capture scenarios may not be suitable. Due to the large number of small objects and the confusing waves on the sea surface in the data set, the task of object detection in maritime SAR is far from solved. For example, the best performing model of the SeaDronesSee object detection track currently only achieves 36% mAP. Compare that to the common COCO benchmark with the best performer achieving over 60% mAP[1].

The YOLO model is the detection network of the clas-sic one-stage algorithm, which has the characteristics of fast running speed and small memory footprint. On the basis of inheriting the advantages of the original YOLO model, YOLO v7[2] has better detection accuracy and faster inference speed due to its more advanced network structure and training strategy skills.

Based on the excellent characteristics of the YOLO network, we propose an improved YOLOv7-sea. We introduce the SimAM [3]attention module to improve the feature extraction ability of the network and build a feature fusion method that is conducive to the identification of small object difficulties. Subsequently, we introduce an extra head for tiny objects detection. Furthermore, we also apply some strategies to improve detection performance, including Test Time Augmentation (TTA) and Weighted Box Fusion (WBF)[4]. Compared with YOLOv7, our improved method can better detect objects in the images captured by maritime UAVs and has better performance on SAR tasks.

In summary, our paper has the following contributions:

1. Based on YOLOv7 network, we propose an object detection scheme to detect objects in sea surface images captured by UAVs.

2. We add a prediction head to YOLOv7 to help predict small objects.

3. We integrate the SimAM module[3] into YOLOv7, which can help the network find regions of interest in images.

4. We provide a useful package of tips for object detection task, including Test Time Augmentation (TTA) and Weighted Box Fusion (WBF).

5. On the SeadroneSee OD v2 test dataset, our proposed YOLOv7-sea achieves 59% (AP). Finally, our model ranks 3rd in the UAV Object Detection v2 challenge.

2. Related work

2.1. General object detection

The current CNN-based detection network can be structurally divided into a one-stage (One-Stage) algorithm and a two-stage (Two-Stage) algorithm, which are classified according to whether they have a proposal step for a region of interest. GIRSHICK et al. successively proposed a series of Two-Stage detectors such as R-CNN (2014)[6], Fast R-CNN (2015) [5], Faster R-CNN (2017)[7]. In recent years, with the design of single-stage detectors, the YOLO series[8][9][10] has attracted extensive attention due to its high accuracy and fast speed. YOLO[8]was first proposed by Joseph et al. in 2016. After that, people continued to improve it and gradually derived subsequent versions, and the detection performance improved steadily. YOLOv7[2] is the most advanced improved model of the current YOLO series. It adopts a more efficient ELAN module on the ba-

sis of YOLOv5, and proposes a method for auxiliary head training, which has high accuracy and performance. Therefore, we choose YOLOv7 as our baseline model than other methods.

2.2. Data Augmentation

Data augmentation is to allow limited data to produce value equivalent to more data without substantially increasing data, making the model more robust to images obtained from different environments. Data enhancement of single data is mainly divided into geometric operation and color transformation. Common data enhancements of color transformation include noise, blur, color transformation, erasure, filling, etc. Geometric transformations include flipping, rotating, shifting, cropping, deforming, scaling and other operations. Unlike single-sample data augmentation, multi-sample data augmentation methods utilize multiple samples to generate new samples. Several researchers have proposed methods that use multiple images clustered together for data augmentation, namely MixUp[11], Copy paste[12], and Mosaic[10]. Mixup is a data augmentation method based on the principle of neighborhood risk minimization proposed by Facebook Artificial Intelligence Research Institute and MIT. It uses linear interpolation to obtain new sample data. Mosaic data enhancement uses four pictures to splicing the four pictures, each picture has its corresponding frame, and a new picture is obtained after splicing the four pictures. In YOLOv7-sea, we mainly use Mosaic, MixUp, HSV, Translate, Scale, etc.

2.3. UAV image object detection

With the rapid development of UAV platforms, applications in monitoring, crowd counting and other fields are advancing rapidly. However, object detection in UAV images is more challenging compared to ground images due to variations in viewpoint, scale and high density of objects. In [13], transformer blocks are introduced in both the backbone and the head to enhance UAV RGB image feature extraction based on YOLOv5. Furthermore, many methods [15][17] generate a set of sub-images based on cropping methods, which can increase the size of the object and enlarge the dataset. In [14], CBMA[19] is embedded into the CSP block of YOLOv4 to achieve the attention objective of feature maps.

3. Method

3.1. Overall Architecture

An overview of YOLOv7-sea is shown in Figure 2, which is mainly based on YOLOv7. The whole architecture consists of three parts. First, the backbone of ELAN from YOLOv7 is adopted to extract feature maps. To make the network learn useful information better, we integrate SimAM attention module[3]. This can highlight the key target features contained in the shallow network and weaken irrelevant information, improving the detection performance of the algorithm for small targets. Since the SeadroneSee dataset contains many very small instances, we add a prediction head to the neck and head sections. Finally, other effective techniques are also employed to achieve better accuracy and robustness, including Test Time Augmentation (TTA) and Weighted Box Fusion (WBF).

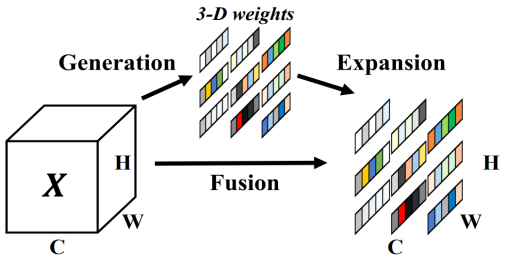

Figure 3. SimAM with full 3D weights for attention [3].

3.2. Extra Prediction head

The seadrone dataset contains many very small instances, making network detection very difficult. Different from the three detection heads in YOLOv7, four detection heads are the ability to detect extremely small swimmers and aquatic equipment, especially those captured at higher positions. The four-head structure can make the detection more stable and reduce the negative impact caused by the severe object scale variance. As shown in Fig. 3, our added prediction head is generated from low-level, high-resolution feature maps, and the performance of small object detection is better despite the increased computational and memory cost.

3.3. SimAM Attention Module

The imagery captured by drones is wide-ranging and disturbed by sea conditions, so effective focus on important areas is critical. Attention modules are widely used in deep learning to enhance feature extraction and focus on useful target objects. However, most current attention modules usually assign weights along the channel dimension to improve the performance of the model, bringing extra parameters to the model. SimAM[3] is a parameter-free attention mechanism that can flexibly assign 3D attention weights to feature maps, thereby enhancing the model's ability to extract features. It does not require additional parameters when deriving attention weights from feature maps, ensuring lighter weight and higher efficiency. The attention mechanism is to find important neurons by measuring the linear separability between neurons, and assign these neu-

rons higher priority. We embed SimAM into the proposed modified YOLOv7 model to improve the performance of object detection. SimAM is derived from neuroscience theory. To differentiate the importance of neurons and successfully achieve attention, the energy function is used to define the linear separability between neuron t and all other neurons in the same channel. The energy function of each neuron is defined as follows[3]:

$$e_t\left(\omega_t, b_t, y, x_i\right) = \frac{1}{M-1}\sum_{i=1}^{M-1}\left(-1-\left(\omega_t x_i + b_t\right)\right)^2 + \left(1-\left(\omega_t t + b_t\right)\right)^2 + \lambda\omega_t^2 \tag{1}$$

where t and x_i are the target neuron and other neurons in the channel, ω_t and b_t are the weights and biases of the linear transformation of t, i is the index in the spatial dimension, λ is the hyper-parameter and $M = HW$ is the number of all neurons on a single channel. The transformation weights and bias are expressed as follows:

$$\omega_t = \frac{2\left(t-\mu_t\right)}{\left(t-\mu_t\right)^2 + 2\sigma_t^2 + 2\lambda} b_t \tag{2}$$

$$\mathrm{b_t} = -\frac{1}{2}\left(t+\mu_t\right)\omega_t \tag{3}$$

where μ_t and σ_t^2 are mean and variance of all neurons except t :

$$\mu_t = \frac{1}{M-1}\sum_{i=1}^{M-1} x_i \tag{4}$$

$$\sigma_\mathrm{t}^2 = \frac{1}{M-1}\sum_{i=1}^{M-1}\left(x_i-\mu_t\right)^2 \tag{5}$$

By calculating the analytical solutions of ω_t, b_t, and the mean and variance of all neurons in the channel, the minimum energy formula is obtained as:

$$\mathrm{e_t^*} = \frac{4\left(\hat{\sigma}^2+\lambda\right)}{(t-\hat{\mu})^2 + 2\hat{\sigma}^2 + 2\lambda} \tag{6}$$

where μ and $\hat{\sigma}^2$ stand for mean and variance:

$$\hat{\mu} = \frac{1}{M}\sum_{i=1} x_i \tag{7}$$

$$\hat{\sigma}^2 = \frac{1}{M}\sum_{i=1}^{M}\left(x_i-\hat{\mu}\right)^2 \tag{8}$$

It can be seen from equation (6) that the smaller the energy function value, the greater the linear separability between neuron t and other neurons. The entire attention module is implemented under the guidance of this energy function, avoiding excessive heuristics and adjustment work. By operating on a single neuron and integrating this linear separability into an end-to-end framework, the neural network has been improved. According to the experiments in the paper[3], after integrating SimAM into different models on different classification and detection datasets, the performance of the model is improved, proving the effectiveness of this module. And our experiments in this paper show that the introduction of the SimAM non-parametric attention mechanism in the YOLOv7 structure helps the model to more effectively extract the feature information of the objects during the detection process without increasing the original network parameters.

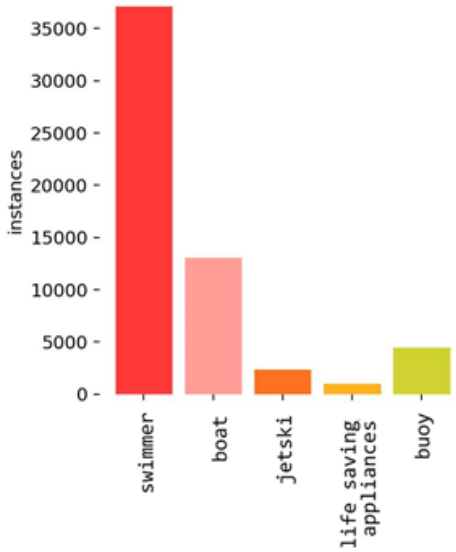

Figure 4. The number of labels of each category.

4. Experiments

4.1. Datasets and Settings

SeaDronesSee is a large dataset designed to help develop systems for SAR using UAVs in maritime scenarios[1]. Its object detection dataset contains 14,227 images, of which 8,930 are in the training set, 1,547 in the validation set, and 3,750 in the test set. The task is to detect object categories like swimmers, boats, jet skis, buoys, lifesaving equipment (life jackets/belts). Ground-truth bounding boxes are available, and the evaluation protocol is based on standard mean average precision. Due to the application scenario, the task provides an additional sub-track where only class-agnostic models are evaluated, similar to detecting any non-water use case.

For detection, we train and infer on 8 NVIDIA Tesla V100 GPUs. Because the proposed YOLOv7-sea and

YOLOv7 share the backbone and most of the head, YOLOv7 is pretrained on COCO dataset[18]. We only train the model on the SeaDronesee training set for 60 epochs, and the first 3 epochs are used for warm-up. Our model uses SGD as the optimizer with a default weight decay of 0.0005 and momentum of 0.937. The size of the input image for our model is very large with a side length of 2400.

4.2. Evaluation Metric

In this paper, the threshold of the intersection over union (IOU) between the prediction frame and the target frame is selected to be greater than 0.5 as the criterion for judging target detection, and the average precision (AP) and the mean average precision (mAP) are used as the evaluation indicators of the model. Specifically, AP is the average of all 10 thresholds of IoU in the range [0.50: 0.95], with a uniform stride of 0.05 for all classes, and is used as the main metric for ranking.

4.3. Implementation Details

Crop image. Object detection in UAV images is more challenging compared to ground images. There are many small objects in the images captured by the drone, such as the size of the object is less than 32 pixels. For small targets, the larger the input image scale, the better the detection performance. Directly magnify the image several times and input it into the network for training, which requires a particularly large memory. To save memory and improve efficiency, we crop the training image into four parts. Training on cropped images saves memory and enlarges the dataset size. The enlarged image makes the small objects more obvious and improves the sensitivity of the network to small objects.

Test time Augmentation. During the inference phase, we use test time augmentation (TTA), which is an application of data augmentation to the test dataset, to improve the performance of our method. We first scale the image to twice, and then test with images of different scales, where we downscale the scaled image to 1x, 0.83x, 0.75x, 0.67x, and 0.5x using 5 different scales. We randomly flipped images down to 0.83x and 0.75x. Finally, we feed 5 images of different scales to YOLOv7-sea and use NMS to fuse the test predictions.

4.4. Experimental results

In this section, we evaluate SeaDronesee on the test set and compare with other methods. Table 1 reports all experimental results. We ended up with a good score on the test set challenge, which is much higher than the results submitted by the committee. In the end, we got the 3rd place with 59% mAP.

Method	AP	AP50	AP75	AR1	AR10
Baseline	41.81	72.33	41.25	36.33	48.91
Yolov7	53.61	83.12	56.72	43.79	60.41
ours	59	90.72	64.15	46.41	67.98

Table 1. The comparison of the performance on SeaDronesee dataset

Method	AP	AP50
Baseline	54	87.8
+head	55.8(+1.4)	45.2
+SimAM	57(+1.2)	92.3
+Model fusion	58.2(+1.2)	93.3

Table 2. Ablation Study on SeaDronesee validation dataset

4.5. Ablation

In order to verify the contribution of the added head, SimAM block to improve the detection performance, respectively, we conduct ablation experiments on the SeaDronesee dataset. As shown in Table 2, we gradually add modules from each layer to the baseline to demonstrate the contribution of these modules for improving model performance. The first row shows the performance of the baseline. From the second row to the last row, AP/AP50 gradually increases from 54/87.8 to 58.2/93.3.

Effect of extra prediction head. Adding an extra head to the original structure will add extra parameters, increasing the number of network layers from the original 105 to 131. But it will make the detector more sensitive on small objects. The AP of small objects (area less than 32^2) on the validation set increases from 0.396 to 0.427, demonstrating the effectiveness of adding additional detection heads.

Effect of SimAM module. SimAM does not bring additional parameters, and the overall parameter quantity of the network has not changed. After introducing the SimAM module, the mAP of the network has increased by 1.2% compared with the original model, and the detection performance has been greatly improved. This shows that the SimAM module can flexibly assign 3D attention weights to feature maps, further enhancing the feature processing capabilities of the network.

Effect of Multi-model Fusion. Model fusion can integrate the learning capabilities of each model, so that the final result can complement each other and improve the generalization ability of the final model. In this challenge we employ Weighted Box Fusion (WBF) to combine the predictions of object detection models. We used multiple training models, including YOLOv7-sea, YOLOv7, YOLOv7e6. As shown in Table 2, the fused model has better performance on the validation set.

5. Conclusion

The SeaDronesee dataset has the characteristics of small target and large sea surface interference, which brings great challenges to general object detectors. In response to these problems, we add a head for tiny object detection, SimAM attention block and some experienced tricks in YOLOv7, and propose a YOLOv7-sea with stronger visual information. Based on them, our algorithm significantly outperforms existing state-of-the-art detectors and achieves AP59% on the SeaDroneSee OD V2 task. We hope this report helps developers and researchers gain better experience in analyzing scenarios captured by maritime drones and in search and rescue missions.

References

[1] Varga, Leon Amadeus, et al. "Seadronessee: A maritime benchmark for detecting humans in open water." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022.

[2] Wang, Chien-Yao, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors." arXiv preprint arXiv:2207.02696 (2022).

[3] Yang, Lingxiao, et al. "Simam: A simple, parameter-free attention module for convolutional neural networks." International conference on machine learning. PMLR, 2021.

[4] Solovyev R, Wang W, Gabruseva T. Weighted boxes fusion: Ensembling boxes from different object detection models[J]. Image and Vision Computing, 2021, 107: 104117.

[5] Girshick, Ross. "Fast r-cnn." Proceedings of the IEEE international conference on computer vision. 2015.

[6] Uijlings J R R, Van De Sande K E A, Gevers T, et al. Selective search for object recognition[J]. International journal of computer vision, 2013, 104(2): 154-171.

[7] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.

[8] Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[9] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[10] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[11] Zhang H, Cisse M, Dauphin Y N, et al. mixup: Beyond empirical risk minimization[J]. arXiv preprint arXiv:1710.09412, 2017.

[12] Ghiasi, Golnaz, et al. "Simple copy-paste is a strong data augmentation method for instance segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

[13] Zhu, Xingkui, et al. "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[14] Zhang, Zixiao, et al. "ViT-YOLO: Transformer-based YOLO for object detection." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[15] Zhang, Xindi, Ebroul Izquierdo, and Krishna Chandramouli. "Dense and small object detection in uav vision based on cascade network." Proceedings of the IEEE/CVF Interna-tional Conference on Computer Vision Workshops. 2019.

[16] Qiao S, Chen L C, Yuille A. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

[17] Ozge Unel, F., Burak O. Ozkalayci, and Cevahir Cigla. "The power of tiling for small object detection." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019.

[18] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." European conference on computer vision. Springer, Cham, 2014.

[19] Woo, Sanghyun, et al. "Cbam: Convolutional block attention module." Proceedings of the European conference on computer vision (ECCV). 2018.

[20] Cai Z, Vasconcelos N. Cascade r-cnn: Delving into high quality object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6154-6162.

[21] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016: 21-37.

[22] Lin, Tsung-Yi, et al. "Feature pyramid networks for object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.