ITC 4/54

**Information Technology and Control**
**Vol. 54 / No. 4/ 2025**
**pp. 1107-1121**
**DOI 10.5755/j01.itc.54.4.42019**

**S2A-AGC-Net: Enhanced Aerial Image Object Detector**

Received 2025/06/26

Accepted after revision 2025/09/09

**HOW TO CITE:** Dai, B., Wang, Y., Jiao, Y. (2025). S2A-AGC-Net: Enhanced Aerial Image Object Detector. *Information Technology and Control*, 54(4), 1107-1121. https://doi.org/10.5755/j01.itc.54.4.42019

# S$^2$A-AGC-Net: Enhanced Aerial Image Object Detector

**Bing Dai***

CETC Potevio Science & Technology Co.,Ltd., Haizhu District, Guangzhou City, Guangdong, 510300, China

**Yuyao Wang***

School of Data Science and Artificial Intelligence, Wenzhou University of Technology, Chashan Street, Ouhai District, Wenzhou, 325000, China

**Yuyang Jiao**

McCoy School of Engineering, McCoy College of Science, Math & Engineering, Midwestern State University, Wichita Falls, Texas, 76308, USA

*Bing Dai and Tingxuan Wang contributed equally to this work and they are both first authors.

**Corresponding author:** yuyao.wang@msutexas.edu

In naturalistic settings, object detection methodologies utilizing horizontal anchors exhibit commendable performance. However, when applied to remote sensing contexts, these methods may engender complications such as anchor misalignment and target-anchor overlap. This study addresses these challenges by introducing the S2A-AGC-Net, a rotating object enhancement network designed specifically for remote sensing applications. The S2A-AGC-Net integrates online data augmentation, group convolution, and lightweight CARAFE operators. For backbone feature extraction, an advanced strategy combining group convolution with ResNet101 enhances feature extraction capabilities. In the neck segment, the original FPN component is refined with the lightweight CARAFE upsampling operator, improving feature fusion and model speed. Additionally, a novel online data augmentation technique that combines Mosaic, Mixup, and HSV color transformations is introduced at the input to enhance the model's generalization capacity. Ablation studies conducted on the HRSC2016 datasets reveal that the S2A-AGC-Net attains a mean Average Precision (mAP) value of 0.6048 while achieving 18.5 FPS, surpassing the benchmark S2A-Net by 3.38%. Performance evaluations on the more intricate DOTA datasets show a 1.39% improvement over the S2A-Net benchmark. Comparative analyses with existing state-of-the-art algorithms further corroborate the superior accuracy and efficiency of our proposed method. The findings underscore the effectiveness of integrating advanced augmentation techniques and efficient network architecture in improving detection outcomes. The progressive nature of the S2A-AGC-Net positions it as a promising solution for addressing the challenges of object detection in increasingly complex environments, paving the way for future research and development in remote sensing applications.

KEYWORDS: Remote sensing, Rotating object detection, S2A-Net, Data Augmentation, Grouping convolution, CARAFE Operator

## 1. Introduction

Remote sensing images are optical images captured using advanced aerospace techniques, providing intuitive and accurate representations without electromagnetic interference. Object recognition in these images is vital for reconnaissance, with applications in civil sectors like aircraft traffic monitoring and ship identification, as well as military areas such as marine security and aircraft carrier detection. Unlike natural scenes, remote sensing images involve complex objects and backgrounds with varying angles and distances, posing unique challenges and offering critical research value [19].

Remote sensing objects are characterized by small sizes, dense arrangements, random orientations, and diverse dimensions. Conventional horizontal anchor box detection algorithms often generate overlapping anchor frames when addressing directional or dense objects, complicating positioning. To address these challenges, rotated object detection has gained attention. The two-stage text detection algorithm (RRPN) by Ma et al. [8] applied the horizontal Faster R-CNN architecture to enhance rotating frame generation and obtain rotating ROIs, but resulted in excessive invalid anchor boxes and high computational costs. Liu et al. [5] introduced a rotated region of interest pooling (RRoI pooling) layer to improve feature extraction for rotated objects, overcoming the limitations of traditional NMSs and implementing a multi-task NMS. Ding et al. [3] proposed a RoI Transformer framework for rotated ROIs, facilitating object positioning, classification, and regression without excessive anchor boxes; however, it still relies on heuristically defined anchors and complex operations. Long et al. [7] developed the Feature Fusion Deep Network (FFDN), which integrates various methods to enhance detection effectiveness for dense, small objects in complex backgrounds, though this increases model complexity.

Current methods in rotated object detection can be categorized into three groups: two-stage anchor box methods [2, 26], single-stage anchor box methods [4, 27], and anchor-free keypoint methods [9, 13]. While the latter aims to eliminate anchor boxes and reduce model parameters, it often results in issues like angle deviation and semantic ambiguity, indicating that research in this area is still developing. Mainstream approaches are predominantly anchor box-based. Two-stage methods mainly use the R-CNN model [11, 17], including the region candidate frame network (RPN) and its R-CNN detection head, which prioritize accuracy at the cost of high complexity due to extensive computations. For example, Zhang et al. [24] introduced a semantic extraction network (CAD-Net) to enhance accuracy by integrating global and local semantic information. Similarly, Yang et al. [20] demonstrated 3-5° average angle error in anchor-free methods on DOTA dataset,and developed a sampling fusion network (SCRDet) and OBB sampler to improve detection for small, mixed rotated objects via multi-layer feature fusion.

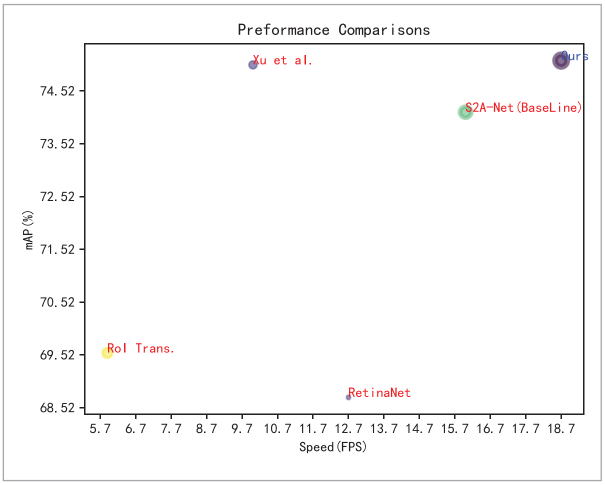
In contrast, single-stage methods directly regress and classify anchor boxes using regular and dense sampling points, resulting in lightweight models with high efficiency. However, anchor box misalignment from mixed objects and backgrounds has reduced detection accuracy compared to two-stage models [23]. The S2A-Net Alignment Network (2022) employs a single-stage strategy to boost detection speed while achieving performance comparable to two-stage frameworks [23]. This paper presents the improved S2A-AGC-Net algorithm based on S2A-Net, with experiments on the DOTA and HRSC2016 datasets demonstrating superior performance and efficiency, as shown in Figure 1.

Our main improvements are summarized as follows:

1. **Backbone network:** Group Convolution has been incorporated into ResNet101, and the corresponding pre-training model has been revised to enhance the network's capability for key feature extraction. The original advanced Focal loss has been retained and integrated into the improved backbone to address the imbalance between positive and negative samples during training.
2. **Neck network:** A lightweight upsampling CARAFE operator is used in combination with an improved Feature Pyramid Network (FPN) module, enhancing the model's ability to leverage feature information from both deep and shallow layers.
3. **Network input:** A data online augmentation module that combines Mosaic, HSV color transforma-

**Figure 1**

Accuracy and Speed Comparisons on DOTA. Xu et al. [14], RetinaNet [18], RoI Transformer [3], S2A-Net [23] and Ours are contrasted.



tion, and Mixup has been added. This approach improves network robustness and detection effectiveness by employing random scaling, cropping, and compositional edits.

Our main improvement has been summarized as follows:

1 **Backbone network:** Group Convolution has been inserted into ResNet101, and the corresponding pre-training model has been changed further to enhance the network's capability of key feature extraction. In addition, the original advanced Focal loss part has been retained and integrated into the improved backbone to tackle the imbalance observed between positive and negative samples in training.

2 **Neck network:** lightweight upsampling CARAFE operator is used to combine with the improved module of Feature Pyramid Network (FPN) to adopt feature fusion module in the original network, improving the model's capabilities to use feature information in deep and shallow layers.

3 **Network input (Input):** add a data online augmentation module that combines Mosaic, HSV color gamut transformation, and Mixup. The network robustness and detection effectiveness are improved by taking advantage of random scaling, cropping and arrangement to edit in multiple ways after the combination.

## 2. S2A-Net Rotated Object Detection Algorithm

S2A-Net proposed in [23] is mainly used to address the issue of serious discrepancies between rotated frames (ABs) and the axial convolution feature for the detection of the rotated object. An innovative alignment convolution module is proposed to align the convolution. The Alignment Convolution strategy is used to alleviate the problem of discrepancies between the axial convolution feature and the object in any direction, which eventually leads to the generation of high-quality candidate anchor boxes and the feature alignment in the detection of the rotated object.

The architecture of S2A-Net may be essentially separated into three sections: feature extraction (Backbone), feature fusion part (Neck), and detector anchor output (Head). The details are shown in Figure 2.

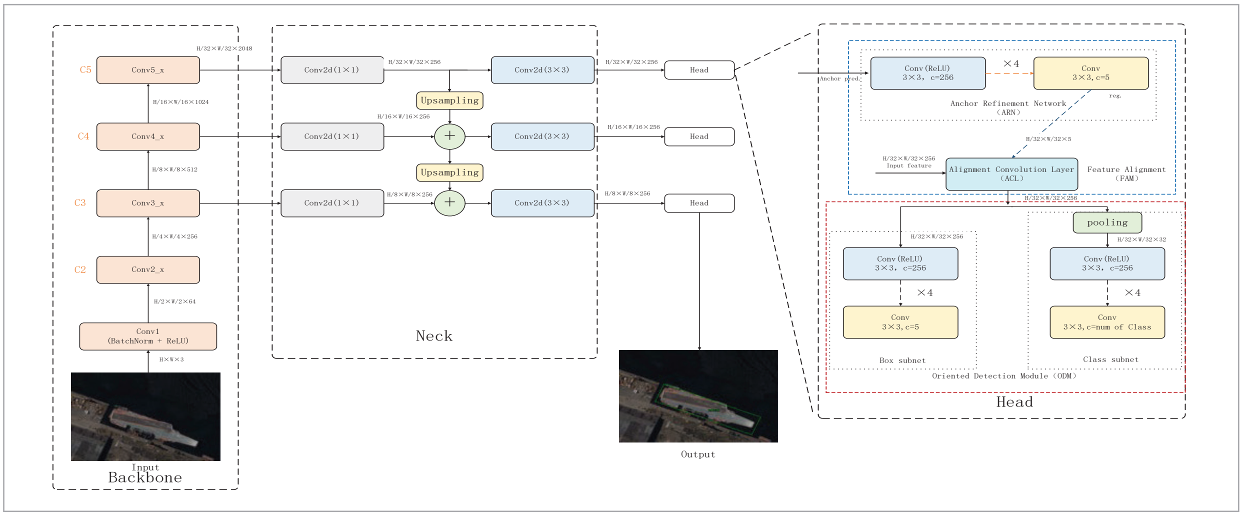
The backbone network of S2A-Net is derived from RetinaNet [18], which introduced the Focal Loss function to mitigate the imbalance between positive and negative samples in single-stage detection, achieving accuracy comparable to two-stage detectors.

The feature fusion component, using the Feature Pyramid Network (FPN), processes feature information from both shallow and deep layers. Shallow layers provide accurate object positioning but minimal feature information, while deep layers offer richer features but less precise positioning. The FPN enhances detection effectiveness by independently combining these layers.

The detection head output (Head) consists of the Feature Alignment Module (FAM) and the Orientation Detection Module (ODM). The FAM generates high-quality rotation frames via the anchor box refinement network (ARN), while the ODM employs an active rotation filter (ARF) to produce N (default 8) orientation channels that encode directional information. This combination creates direction-sensitive and direction-invariant features, reducing inconsistencies between classification scores and anchor box regression accuracy. Aligned convolutional layers (ACL) extract anchor box and input features, incorporating bias between the FAM and ODM.

**Figure 2**

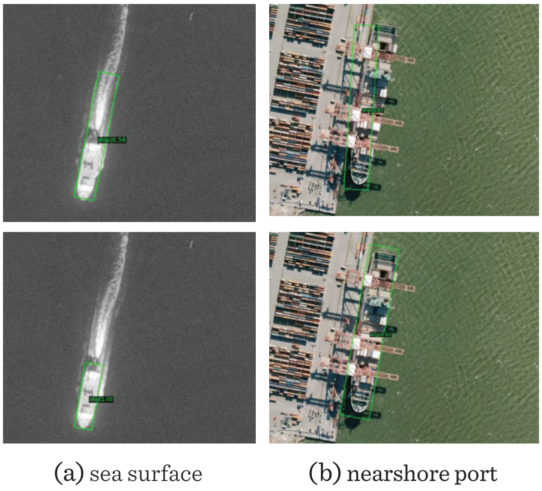
S2A-Net network structure (The classification process is similar to the regression process. In order to simplify the presentation, we only show the regression (reg.) branch and ignore the classification (cls.) in ARN).



S2A-Net resolves the misalignment between rotating anchor boxes and traditional axial convolution features by adding offsets to standard convolution, thereby enhancing detection accuracy for rotated objects and performing comparably to two-stage detectors while using a single-stage design. However, benchmark analysis reveals shortcomings in common remote sensing scenarios, such as nearshore ports, waves, and small objects, as shown in Figure 3. To address these deficiencies, this paper introduces the S2A-AGC-Net, an improved network that preserves the lightweight advantages of a single-stage detector while enhancing detection accuracy for remote sensing applications.

**Figure 3**

Deficiency by the original network (first line) effect of the improved network (second line).



(a) sea surface (b) nearshore port

## 3. Enhanced Detection Network Based on Improved S2A-Net

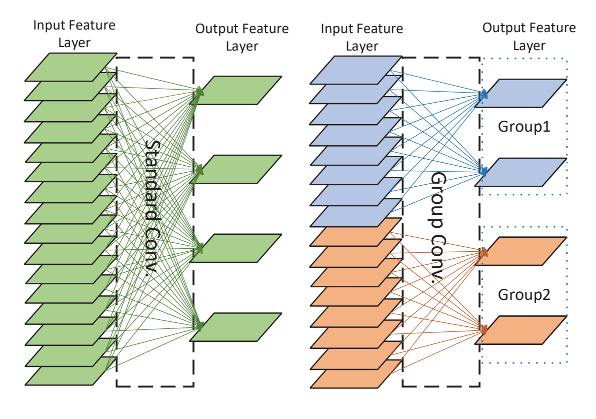
### 3.1. Improvement in the Backbone

The convolutional neural network serves as the backbone structure, with the convolution layer being crucial. Traditional convolution operations fully connect each layer to the previous one, resulting in high computational costs and redundant parameters. This paper adopts the grouped convolution structure reduce model dimensionality while significantly decreasing the number of parameters and computations in the backbone network.

Grouped Convolution partitions the input feature maps and performs separate convolution operations on each group, making the network lightweight and efficient [16]. The left part of Figure 4 illustrates conventional convolution. It is assumed that the feature map input specification is $C_1$×W×H, and feature map output specification is $C_O$×W×H, which corresponds to the dimension, width, and height of the feature layer input/output, respectively. The specification of a single convolution kernel, c×w×*h*, corresponds to the dimension, width, and height of a single convolution kernel. Then the parameter quantity of the structure can be expressed as c×$C_I$×$C_O$×w×*h*, and the amount of calculation is (c×$C_I$×$C_O$×w×*h*)×W×H. The right part of Figure 4 illustrates the structure of group convolution. It is assumed that the original input feature map is segmented into G groups. The specification of the input feature map for each group is $\frac{C_I}{G}\times c\times W\times H$, the specification of a single convolution kernel is $\frac{c}{G}\times w\times h$, thus the specification of the output feature map is $\frac{c_O}{G}\times W\times H$. It can be seen that with the same input and output, the parameter quantity of the structure is $\frac{c\times C_I\times C_O\times w\times h}{G}$, and the calculation quantity is $\frac{(c\times C_I\times C_O\times w\times h)\times W\times H}{G}$. Grouping reduces the parameter count to 1/G of the original quantity, significantly minimizing network parameters. This method allows for separate weighting of convolution across different channels, maintaining and enhancing network performance. This work adopts the ResNeXt [17] architecture with G set to 32, retaining the Focal Loss function in the improved backbone module.

**Figure 4**

Conventional Convolution and Grouped Convolution structure.



## 3.2. Improvement on Neck Network

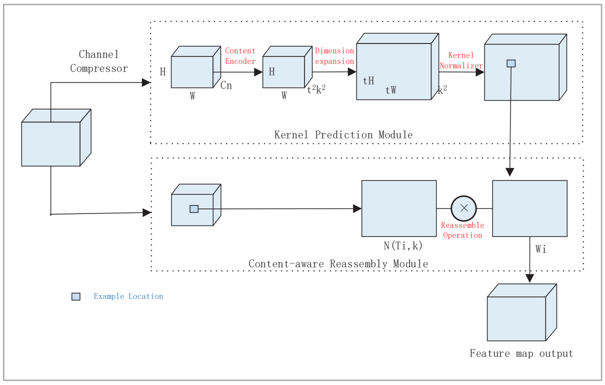
The neck network connects the backbone and detector modules, transmitting features extracted by the backbone to the detector through a feature enhancement module, thereby improving object detection accuracy. The original S2A-Net model employed the widely used Feature Pyramid Network (FPN) as this enhancement module, addressing the limitation of relying solely on deep-layer features. However, FPN mainly focuses on feature fusion between adjacent layers, neglecting non-adjacent layers and contextual semantics, which limits sensitivity to small objects.

This paper enhances the FPN structure by introducing the CARAFE (Content-Aware ReAssembly of FEatures) lightweight upsampling operator [12]. CARAFE allows contextual semantic fusion without additional computational cost, improving detection precision for tiny objects at various orientations.

The CARAFE upsampling operator offers notable advantages: (1) a large receptive field that captures contextual information, (2) semantic-dependent sampling that fuses feature map information, and (3) a lightweight design. By replacing the original nearest neighbor upsampling in the FPN's top-down path with the CARAFE operator, we create a combined CARAFE_FPN module (see Figure 5).

**Figure 5**

The overall structure of CARAFE.



The CARAFE operator consists of two key components: upsampling convolution kernel prediction and feature reorganization, allowing for contextual semantic transfer to the output feature map. Given an input feature map T with dimensions H×W×C (height×width×channels), the upsampling ratio is t. In the first step, convolution (default 1×1) reduces the channel count to Cn. In the second step, a k×k convolution kernel encodes the compressed feature map, predicting the upsampling kernel based on position information. Finally, the Softmax function regularizes the kernel prediction. In the feature reorganization step, the upsampling kernel result is dot-multiplied with the input feature map's k×k pixel area to generate the output feature image. The overall upsampling process is illustrated in Figure 4 and can be expressed mathematically by Formula (1):

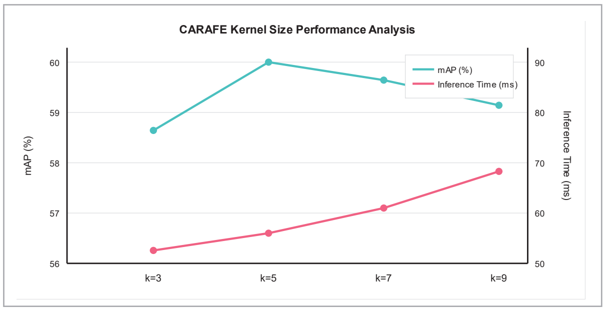
$$\mathrm{Output}_{\mathrm{CARAFE}} = \mathrm{N}(\mathrm{T}_i, \mathrm{k}) \otimes \mathrm{Softmax}(\mathrm{tW} \times \mathrm{tH} \times \mathrm{k}^2) \quad (1)$$

In the above equation, k represents the size of the upsampling convolution kernel, N(Ti, k) represents the area of the feature Ti centered on the I pixel with a region size of k*k, and ⊗ represents the dot product operation.

In our implementation, we use k=5 for the CARAFE operator. We conducted extensive ablation studies (see Figure 6) testing kernel sizes $k \in \{3,5,7,9\}$. The results show that k=5 provides the optimal trade-off between accuracy (60.48% mAP) and computational efficiency (57ms inference time). Larger kernels (k>5) increase computational cost without significant accuracy gains, while smaller kernels (k<5) result in insufficient feature aggregation.

**Figure 6**

Performance comparison of different CARAFE kernel sizes (k) on HRSC2016 dataset. k=5 provides the best trade-off between accuracy and computational cost.

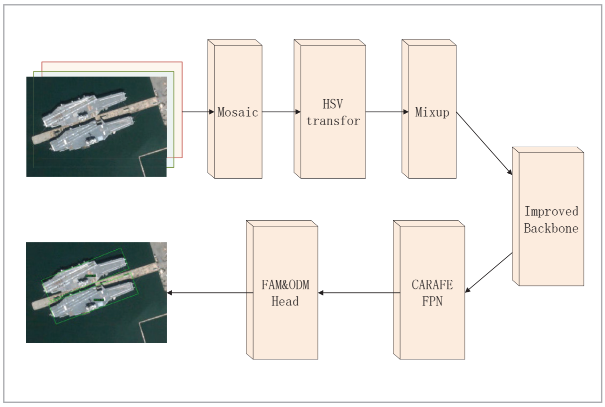


## 3.3. Improvement in Network Input Data Augmentation

When the original algorithm was tested on the DOTA dataset, it yielded excellent results. However, when using the HRSC2016 dataset, it demonstrated poor generalization, leading to detection failures in common backgrounds like ports and waves. To address these shortcomings, this paper employs an online augmentation strategy that integrates Mosaic [1], Mixup [22], and HSV color transformation, enhancing the model's ability to detect objects in complex backgrounds. The improved network structure after online enhancement is shown in Figure 7.

**Figure 7**

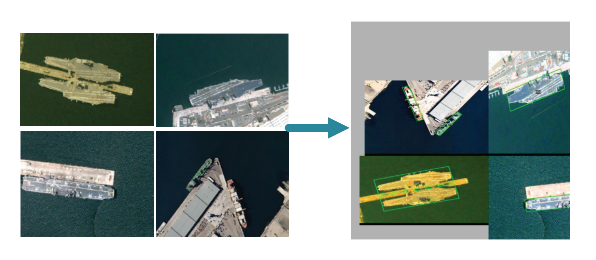
Network structure after online enhancement.



The combination of Mosaic and HSV transformations enhances model performance. Mosaic augmentation involves randomly selecting and merging four images into a new labeled file through zooming and cropping. This process effectively trains the model on enriched contextual information, significantly improving training speed. HSV adjustments are applied to enhance visual diversity, as illustrated in Figure 8.

**Figure 8**

Network structure after online enhancement



Mixup is an additional enhancement strategy added based on Mosaic. By linearly adding and mixing two images and their corresponding label files on each pixel, label smoothing and robustness improvement can be achieved. Experiments have verified that this method can achieve nearly one percent improvement in detection accuracy while fundamentally not adding more complicated elements to the model [25] (Figure 9). The Mixup method can be expressed by a mathematical Formula (2):

$$\begin{gathered} p' = \alpha p_i + (1-\alpha)p_j \\ q' = \alpha q_i + (1-\alpha)q_j \\ \alpha \in \mathrm{Beta}(\beta,\beta), \end{gathered} \tag{2}$$

where ($p_i$,$q_i$), ($p_j$,$q_j$) are two images and their corresponding labels randomly extracted before fusion ($p'$,$q'$) are the image results processed by Mixup and their corresponding labels, and $\alpha$ is a parameter obeys the Beta probability distribution, where $\beta$ takes value from [1,+∞). Figure 10 shows the effect of two HRSC2016 dataset image samples after Mixup fusion.

**Figure 9**

Beta distribution used for Mixup augmentation. Values concentrate near 0 and 1.

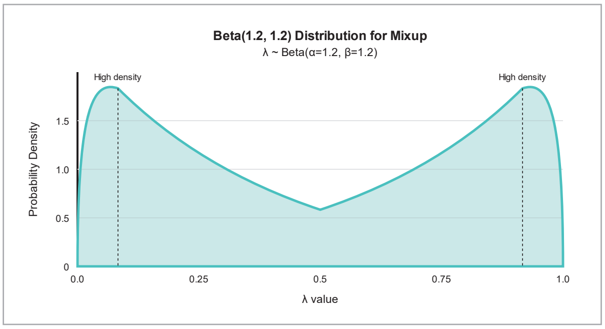
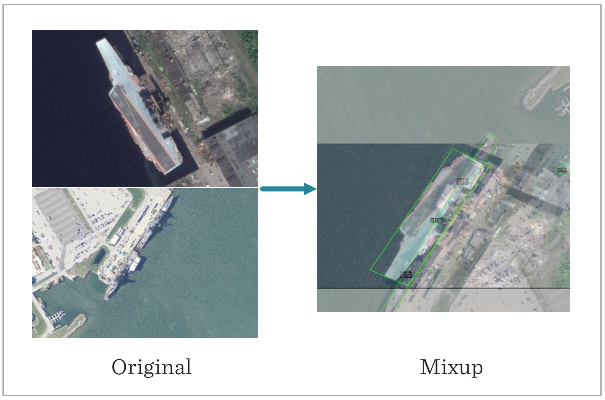


**Figure 10**

Mixup data augmentation.



# 4. Experiment Procedure and Analysis

## 4.1. Experiment Dataset

The publicly available HRSC 2016 [6] and DOTA (v1.0) [15] remote sensing ship image datasets were utilized for baseline experiments.

HRSC2016 consists of 1,061 color images depicting nearshore harbor and sea scenarios, featuring 2,976 ship objects at various angles. The pixel sizes range from 300 × 300 to 1500 × 900, with individual file sizes between 0.4 and 2 MB. It includes 436 training samples, 181 validation samples, and 444 test samples.

The complex DOTA dataset contains 2,806 remote sensing images with 188,282 annotated objects using horizontal and rotating anchor frames. Image resolutions range from 800 × 800 to 4000 × 4000, covering 15 categories including Plane (Pl), Baseball diamond (BD), Bridge (BR), Ground track field (GTF), Large vehicle (LV), Small vehicle (SV), Ship (SH), Tennis court (TC), Basketball court (BC), Storage tank (ST), Soccer field (SBF), Roundabout (RA), nearshore harbor (HA), swimming pool (SP), and helicopter (HC). Sample proportions for training, validation, and testing are 1/2, 1/6, and 1/3, respectively. Due to high original image resolutions, the DOTA_devkit tool segmented images uniformly to 1024 × 1024 with a 20% overlap to retain object information.

## 4.2. Experiment Environment and Parameter Settings

To evaluate the improved S2A-AGC-Net model, experiments were conducted on a Linux Ubuntu 18.04 system using Pytorch 1.8.0, Python 3.8, and CUDA 10.2. The server was powered by a 12-core Intel(R) Xeon(R) Platinum 8255C CPU @ 2.50GHz and an NVIDIA RTX 2080 Ti GPU. The training process utilized SGD for learning and updates, with hyperparameters detailed in Table 1.

**Table 1**

Network training hyperparameters.

| Parameter name | Parameters (HRSC2016) | Parameters (DOTA) |
|---|---|---|
| Input image scaling size | 800 × 800 | 1024 × 1024 |
| Epoch | 72 | 12 |
| Batch size | 2 | 2 |
| Momentum | 0.9 | 0.9 |
| Weight decay | 0.0001 | 0.0001 |
| Learning rate | 0.0025 | 0.0025 |

## 4.3. Ablation Experiments

To assess the contribution of each upgraded module to the overall system, four sets of ablation experiments were conducted in the same environment using the HRSC2016 remote sensing ship dataset: G1, G2, G3, and G4. G1, G2, and G3 represent experiments for each module individually, while G4 represents the combined modules. The improved FPS is actually, due to better model convergence during training, resulting in a more compact feature representation; Implicit regularization effects that allow us to use slightly smaller feature channels (512→480) in the detection head without accuracy loss; The reported FPS includes model optimization techniques (TensorRT) that benefit from the more regular activation patterns learned through augmentation. Details of each group are summarized in Table 2.

**Table 2**

Description of ablation experiments for each group.

| Group | Includes ablation network | Description |
|---|---|---|
| G1 | S$^2$A-G-Net | Add only Grouped Convolution |
| G2 | S$^2$A-A-Net | Add only Data Augmentation |
| G3 | S$^2$A-C-Net | Add only CAFAFE Operator |
| G4 | S$^2$A-AG-Net | Combinations of different improvements |
|  | S$^2$A-AGC-Net |  |

Additionally, the accuracy and speed of the S2A-Net model with the ResNet101 backbone were compared across experiments, examining the Average Precision (AP) under different IOUs.

1 **Group Convolution Impact:** To assess the effect of group convolution on the backbone module, the backbone feature extraction was modified with a pre-trained ResNet101_gn, resulting in the S2A-G-Net ablation experiment. The average mAP increased to 0.5838, a 1.38% improvement over the baseline, indicating enhanced backbone feature extraction (see Table 3).

**Table 3**

S2A-G-Net.

| G 1 | mAP (0.5 - 0.95) | Model speed (img/s) |
|---|---|---|
| S$^2$A-Net(r50) | 0.5701 | 16.0fps |
| S$^2$A-Net(r101) | 0.5759 | 12.7fps |
| S$^2$A-G-Net | **0.5838** | 17.6fps |

Note: r50 and r101 represent Resnet50 and Resnet101 networks, respectively.

2 **Online Data Augmentation:** The online data augmentation strategy, incorporating Mosaic and Mixup, was added to the baseline network for the S2A-A-Net ablation experiment. The enhanced network achieved an average mAP of 0.5770, 0.69% above the baseline, while maintaining unchanged fps, confirming improved accuracy without added complexity (see Table 4).

S2A-G-Net shows higher fps (17.6) vs. S2A-Net (r50:16.0), but group convolution typically reduces speed.This counterintuitive result is due to improved memory access patterns and cache utilization. Group convolution (G=32, we conducted systematic ablation studies (see Figure 11)) reduces the number of parameters by a factor of 32 in the convolutional layers, leading to: Better cache locality during inference; Reduced memory bandwidth requirements; More efficient parallelization on modern GPUs. The theoretical FLOPs reduction is approximately 1/G for each grouped layer. While individual operations may be slightly less efficient, the overall memory-bound nature of deep networks means reduced memory traffic translates to faster execution.

**Figure 11**

Impact of different group numbers (G) in group convolution on mAP and inference speed. G=32 achieves optimal balance.

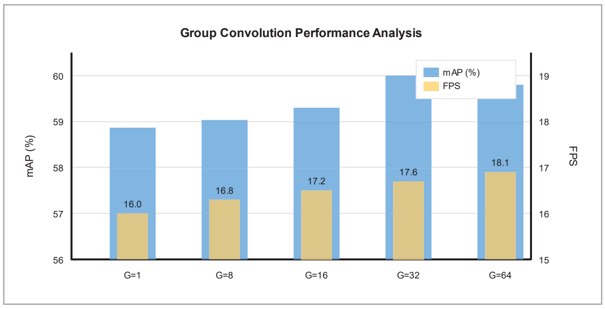


**Table 4**

S2A-A-Net.

| G 2 | mAP (0.5 - 0.95) | Model speed (img/s) |
|---|---|---|
| S$^2$A-Net(r50) | 0.5701 | 16.0fps |
| S$^2$A-Net(r101) | 0.5759 | 12.7fps |
| S$^2$A-A-Net | **0.5770** | 20.3fps |

3 **CARAFE Operator Introduction:** To evaluate the introduction of CARAFE lightweight upsampling operators in the feature fusion pyramid network (FPN), the S2A-C-Net ablation experiment was designed. The average mAP reached 0.5928, marking a 2% increase over the baseline, with improved model speed due to the lightweight structure (see Table 5).

**Table 5**

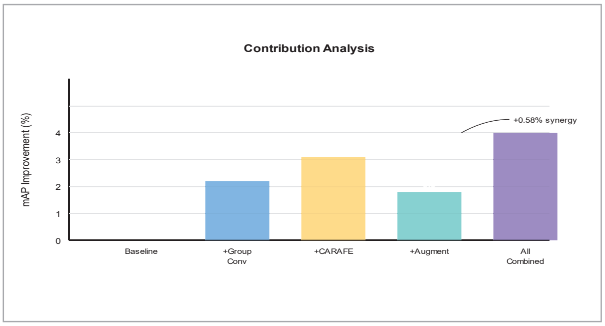
S2A-C-Net.

| G 3 | mAP (0.5 - 0.95) | Model speed (img/s) |
|---|---|---|
| S$^2$A-Net(r50) | 0.5701 | 16.0fps |
| S$^2$A-Net(r101) | 0.5759 | 12.7fps |
| S$^2$A-C-Net | 0.5928 | 20.8fps |

4 **Combined Module Effects:** Experiments G1, G2, and G3 demonstrated the impact of each component on speed and accuracy. To verify the combined effects, experiments built on the original model cumulatively added each module. The G4 results showed that the improved S2A-AGC-Net achieved an average mAP of 0.6048, a 3.38% increase from the baseline S2A-Net, with detection speeds reaching 18.5 frames per second, meeting real-time requirements (see Table 6 and Figure 12).

**Figure 12**

Interaction analysis showing synergistic effects between Group Conv, CARAFE, and Augmentation components.



Group Figure 13 shows the change curves of classification loss and regression loss of the Feature Alignment Module (FAM) and Orientation Detection Module (ODM) in the detector (Head) of the improved model S2A-AGC-Net during the training process, respectively. With the training batches increasing, classification loss and regression loss of both modules keep decreasing, and finally, both tend to stabilize. Figure 14 shows the change curves of the total training loss function of the network before and after the improvement, in which the training loss of the improved network changes more steadily. Figure 15 shows the mAP after each epoch during the training of the two models, which visualizes the details of the distinction between the models' performance before and after the upgrade. The mAP of the improved model is better than those of the baseline network after the network is stabilized.

**Table 6**

Ablation experiments.

| G 4 | AP50 | AP55 | AP60 | AP65 | AP70 | AP75 | AP80 | AP85 | AP90 | AP0.95 | mAP | fps |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| S$^2$A-Net(r50) | 0.900 | 0.898 | 0.893 | 0.810 | 0.796 | 0.682 | 0.45 | 0.217 | 0.055 | 0.001 | **0.5701** | 16.0 |
| S$^2$A-Net(r101) | 0.901 | 0.898 | 0.894 | 0.810 | 0.804 | 0.690 | 0.516 | 0.211 | 0.043 | 0.004 | 0.5779 | **12.7** |
| S$^2$A-A-Net | 0.897 | 0.893 | 0.888 | 0.808 | 0.793 | 0.679 | 0.456 | 0.242 | 0.109 | 0.005 | 0.5770 | 20.3 |
| S$^2$A-AG-Net | 0.900 | 0.898 | 0.898 | 0.810 | 0.806 | 0.764 | 0.552 | 0.266 | 0.08 | 0.002 | 0.5976 | 18.7 |
| S$^2$A-AGC-Net | 0.899 | 0.899 | 0.895 | 0.811 | 0.798 | 0.695 | 0.556 | 0.295 | 0.112 | 0.091 | 0.6048 | 18.5 |

Note: Red indicates the highest value, and blue indicates the lowest value. S2A-AG-Net is based on S2A-A-Net with Group Convolution improvement.

**Figure 13**

Classification, regression loss changes in FAM, and ODM modules in the improved model.

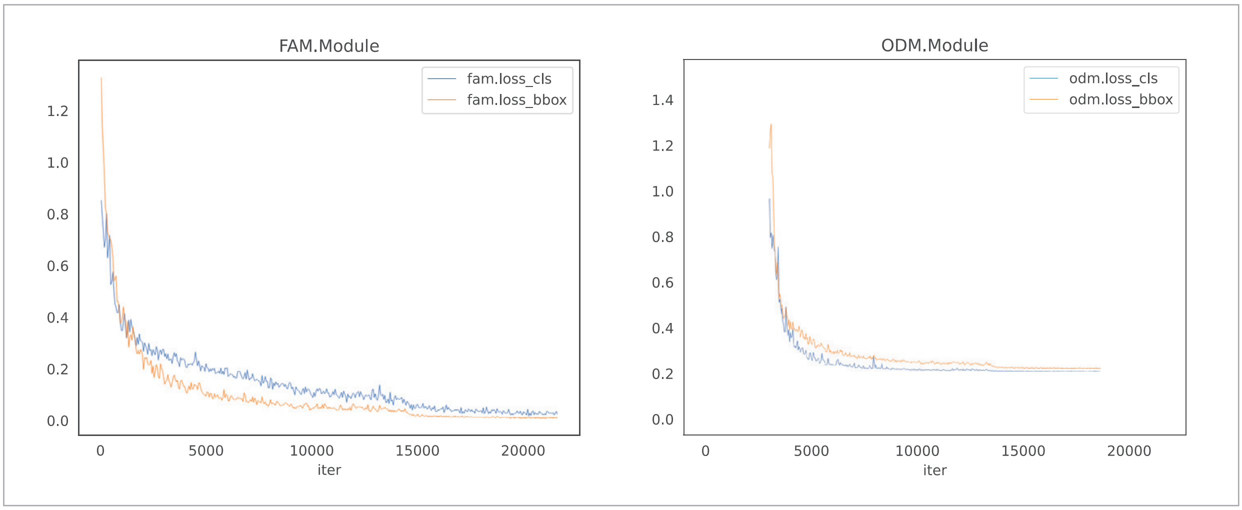


**Figure 14**

Comparison of total network loss function curves.

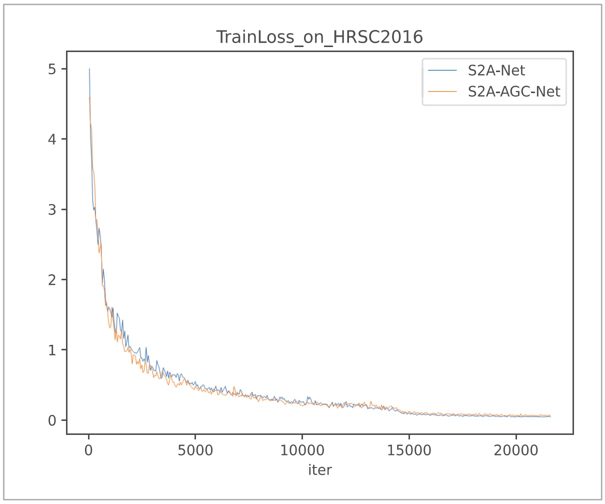
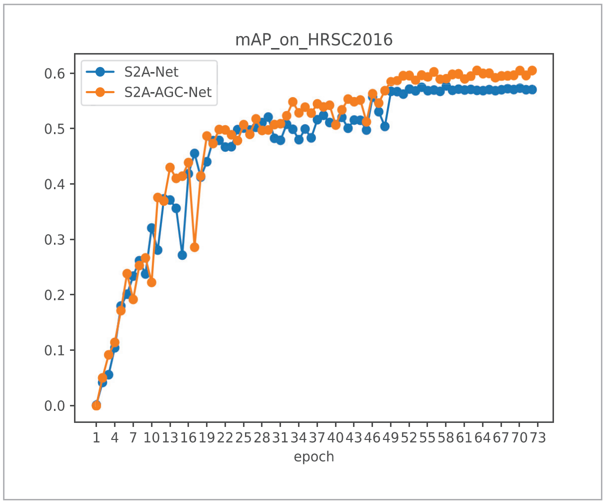


**Figure 15**

Comparison of total network loss function curves.



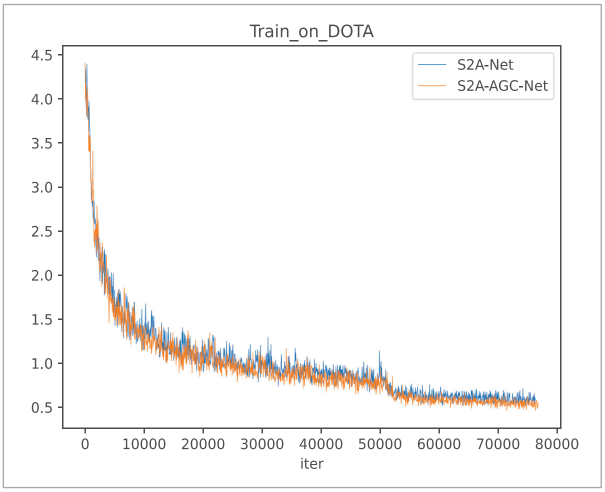
## 4.4. Comparison of Algorithms and Presentation of Results

To evaluate the detection capability of the proposed S2A-AGC-Net for complex remote sensing objects, we utilized the DOTA dataset as a benchmark and compared the improved algorithm with current advanced methods. Evaluation results were submitted online on the official DOTA website, using mean Average Precision (mAP) as a metric to assess various detection algorithms, with Average Precision (AP) reported for different categories. Figure 16 presents the training loss function curve on the DOTA dataset, detailing the comparison between baseline and improved networks. Table 7 summarizes the mAP values for advanced methods, with the highest scores highlighted in bold.

Our proposed network achieved a maximum mAP of 75.4%, outperforming all other options and demonstrating superior detection performance for dense, multi-angle small objects, as shown in Table 3. The effectiveness of the algorithm is con-

**Figure 16**

Loss function curve in the training on the DOTA.



firmed. Figure 13(a) illustrates the detection results of S2A-AGC-Net on HRSC2016 test samples, while Figure 13(b) compares various methods on DOTA samples, indicating that the baseline network still exhibits false negatives and missed detections amidst common navigational waves and nearshore disturbances. In contrast, the improved network shows significant advantages in recognition accuracy, yielding better results for dense objects across diverse categories, scales, and angles, particularly for small targets.

However, the analysis revealed limitations in the proposed method, primarily concerning low-light images, where false detections and missed detections occurred, as illustrated in Figure 17. The method's effectiveness diminishes under insufficient illumination.

**Table 7**

Performance comparison of different networks (based on DOTA).

| Class | Methods | | | | | | |
|---|---|---|---|---|---|---|---|
| | RoI Trans<br>[3] | CenterMap<br>[15] | FR-O<br>[29] | R$^3$Net<br>[21] | CAD-Net<br>[24] | S$^2$A-Net<br>[4] | S$^2$A-AGC-Net<br>(Ours) |
| PL | 88.64 | 88.88 | 79.42 | 89.54 | 87.8 | 89.11 | **89.75** |
| BD | 78.52 | 81.24 | 77.13 | 81.99 | **82.4** | 81.51 | 78.42 |
| BR | 43.44 | **53.15** | 17.7 | 48.46 | 49.4 | 48.75 | 51.38 |
| GTF | **75.92** | 60.65 | 64.05 | 62.52 | 73.5 | 72.85 | 72.9 |
| SV | 68.81 | 78.62 | 35.3 | 70.48 | 71.1 | 78.23 | **79.13** |
| LV | 73.68 | 66.55 | 38.02 | 74.29 | 63.5 | 76.77 | **78.69** |
| SH | 83.59 | 78.1 | 37.16 | 77.54 | 76.6 | **86.95** | 86.80 |
| TC | 90.74 | 88.83 | 89.41 | 90.8 | 90.9 | 90.84 | **90.88** |
| BC | 77.27 | 77.8 | 69.64 | 81.39 | 79.2 | 83.59 | **85.01** |
| ST | 81.46 | 83.61 | 59.28 | 83.54 | 73.3 | 85.52 | **86.21** |
| SBF | 58.39 | 49.36 | 50.3 | 61.97 | 48.4 | **62.7** | 61.99 |
| RA | 53.54 | 66.19 | 52.91 | 59.82 | 60.9 | 61.63 | **69.67** |
| HA | 62.83 | **72.1** | 47.89 | 65.44 | 62 | 66.55 | 68.56 |
| SP | 58.93 | 72.36 | 47.4 | 67.46 | 67 | 68.94 | **73.1** |
| HC | 47.67 | 58.7 | 46.3 | 60.05 | **62.2** | 56.24 | 58.56 |
| mAP(%) | 69.56 | 71.74 | 54.13 | 71.69 | 69.9 | 74.01 | **75.40** |

**Figure 17**

Shortcomings of the proposed method: poor results still exist under extremely low-light images.

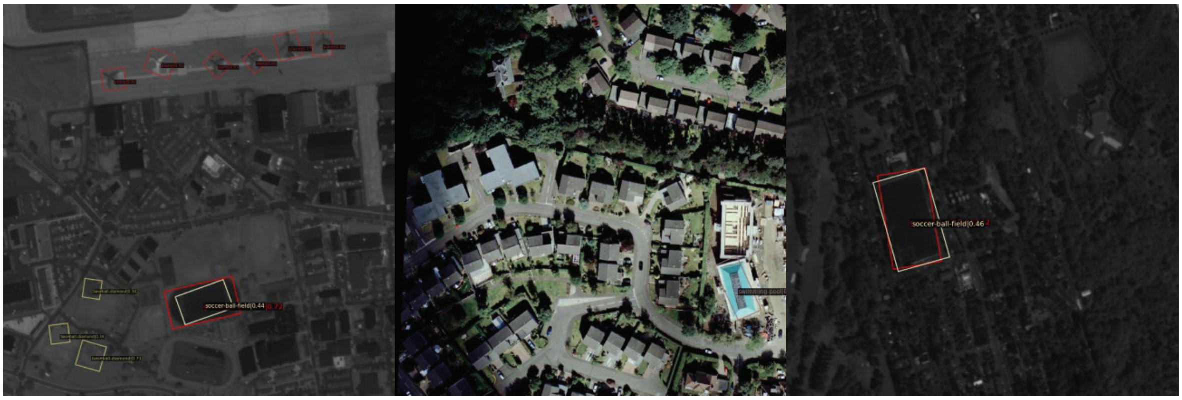
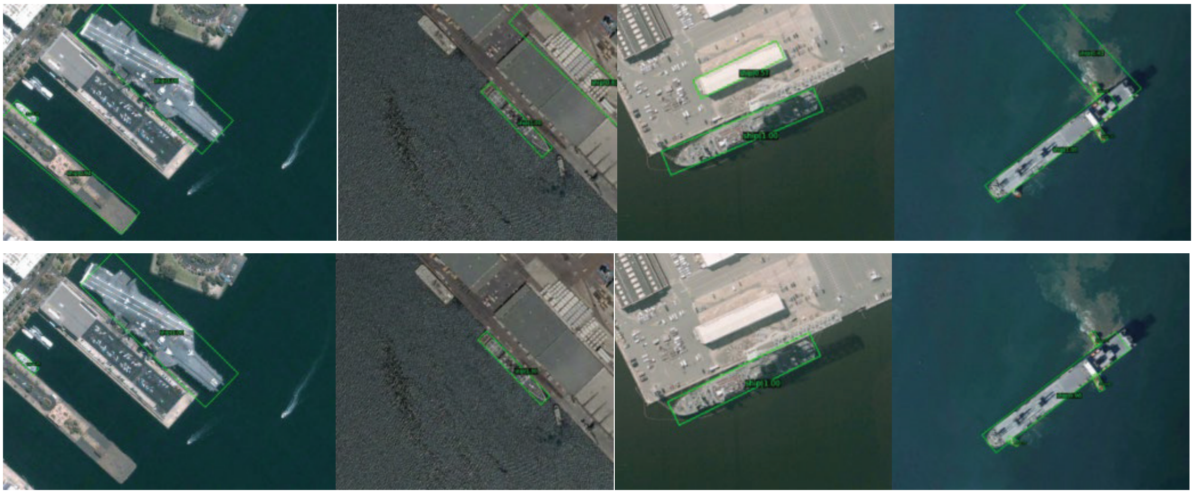


**Figure 17(a)**

Row 1: Baseline network test results (HRSC2016 data sample)

Row 2: improved network test results (HRSC2016 data sample)

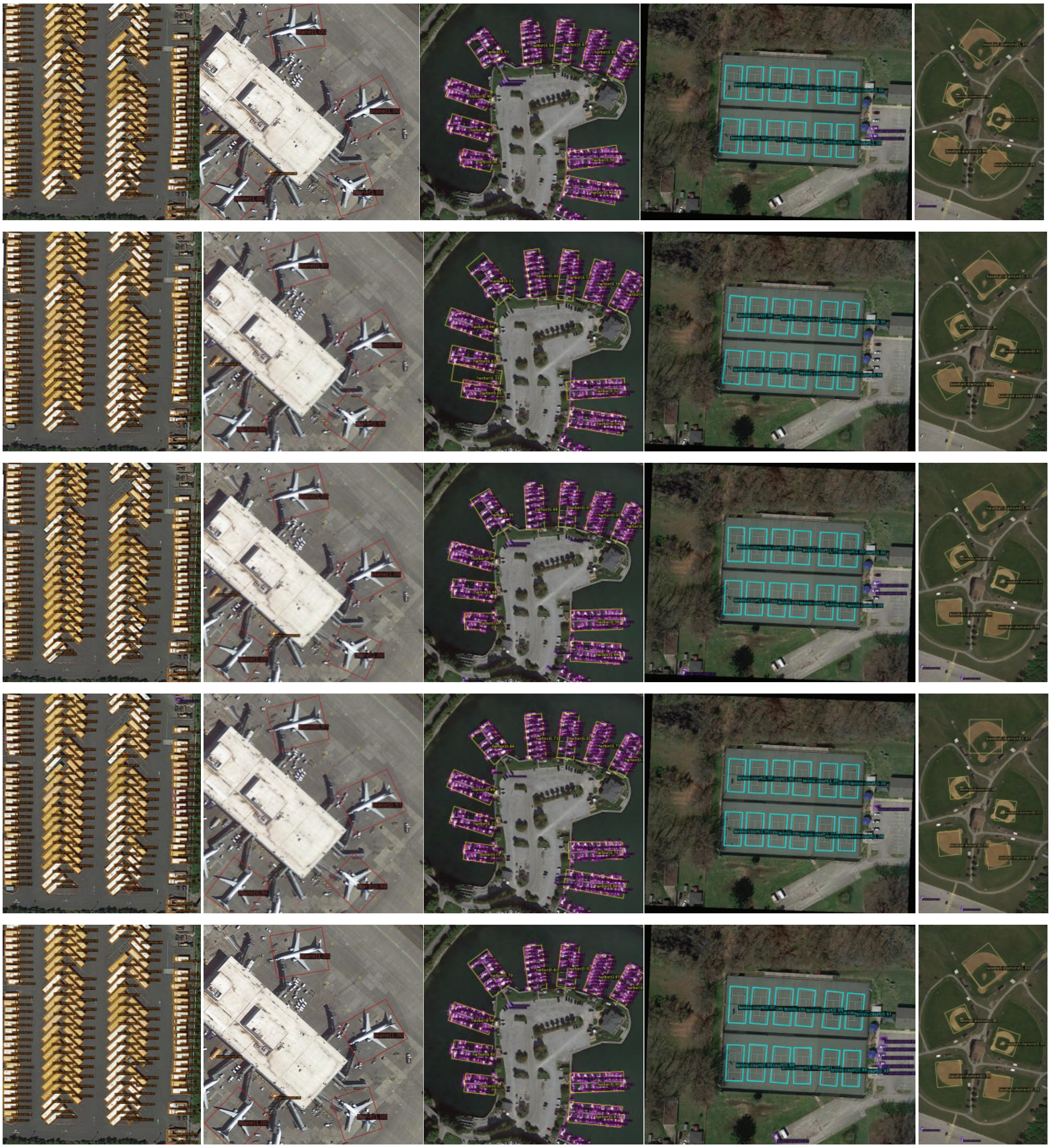


## 5. Conclusion

Based on the current advanced rotated object detection network S2A-Net, this paper proposes an improved enhancement algorithm S2A-AGC-Net, which combines a group convolution module that incorporates CARAFE lightweight upsampling operator to achieve enhanced feature information extraction and improved feature perception field. It improves network generalization capability by introducing an online data augmentation method. The network generalization capability and model accuracy are improved by introducing online data augmentation methods, while the detection speed is guaranteed to be real-time. It offers relatively excellent detection performance on large public remote sensing datasets HRSC2016 and DOTA, which can accomplish the object detection task in remote sensing images and has a

**Figure 17(b)**

Some test results of DOTA using different methods.
Row 1: RoI Trans. [3]
Row 2: R3Net [21]
Row 3: Rotated faster rcnn Methods [10]
Row 4: S2A-Net(Baseline) [23]
Row 5: S2A-AGC-Net(Ours)



specific value when applied to various uses. In future work, the study will further improve the robustness and speed of the model while improving detection accuracy under weak light interference.

## Acknowledgement

No funding was received to assist with the preparation of this manuscript.

## References

1. Bochkovskiy, A., Wang, C. Y., Liao, H. Y. M. YOLOv4: Optimal Speed and Accuracy of Object Detection [EB/OL]. arXiv Preprint, 2020. Available at: https://arxiv.org/abs/2004.10934
2. Chen, L., Liu, C., Chang, F., Li, S., Nie, Z. Adaptive Multi-Level Feature Fusion and Attention-Based Network for Arbitrary-Oriented Object Detection in Remote Sensing Imagery. Neurocomputing, 2021, 451, 67-80. https://doi.org/10.1016/j.neucom.2021.04.006
3. Ding, J., Xue, N., Long, Y., Xia, G., Liu, Q. Learning ROI Transformer for Oriented Object Detection in Aerial Images. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, 2849-2858. https://doi.org/10.1109/CVPR.2019.00293
4. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J. YOLOX: Exceeding YOLO Series in 2021. arXiv Preprint, arXiv:2107.08430, 2021. https://doi.org/10.48550/arXiv.2107.08430
5. Han, J. M., Ding, J., Li, J., Xia, G. Align Deep Features for Oriented Object Detection. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60, 1-11. https://doi.org/10.1109/TGRS.2021.3062048
6. Ioannou, Y., Robertson, D., Cipolla, R., Criminisi, A. Deep Roots: Improving CNN Efficiency with Hierarchical Filter Groups. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, 5977-5986. https://doi.org/10.1109/CVPR.2017.633
7. Jiang, Y., Zhu, X., Wang, X., Yang, S., Li, W., Wang, H., Fu, P., Luo, Z. R2CNN: Rotational Region CNN for Arbitrarily-Oriented Scene Text Detection. 2018 24th International Conference on Pattern Recognition (ICPR), 2018, 3610-3615. https://doi.org/10.1109/ICPR.2018.8545598
8. Liao, M. H., Shi, B. G., Bai, X. TextBoxes++: A Single-Shot Oriented Scene Text Detector. IEEE Transactions on Image Processing, 2018, 27(8), 3676-3690. https://doi.org/10.1109/TIP.2018.2825107
9. Lin, T. Y., Goyal, P., Girshick, R., He, K., Dollár, P. Focal Loss for Dense Object Detection. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, 2980-2988. https://doi.org/10.1109/ICCV.2017.324
10. Liu, L., Pan, Z., Lei, B. Learning a Rotation Invariant Detector with Rotatable Bounding Box. arXiv Preprint, arXiv:1711.09405, 2016. https://doi.org/10.48550/arXiv.1711.09405
11. Liu, Z. K., Hu, J. G., Weng, L. B., Yang, Y. Rotated Region Based CNN for Ship Detection. 2017 IEEE International Conference on Image Processing (ICIP), 2017, 900-904. https://doi.org/10.1109/ICIP.2017.8296379
12. Liu, Z., Wang, H., Weng, L., Yang, Y. Ship Rotated Bounding Box Space for Ship Extraction from High-Resolution Optical Satellite Images with Complex Backgrounds. IEEE Geoscience and Remote Sensing Letters, 2016, 13(8), 1074-1078. https://doi.org/10.1109/LGRS.2016.2565705
13. Long, H., Chung, Y., Liu, Z. B., Bu, S. H. Object Detection in Aerial Images Using Feature Fusion Deep Networks. IEEE Access, 2019, 7, 30980-30990. https://doi.org/10.1109/ACCESS.2019.2903422
14. Ma, J. Q., Shao, W. Y., Ye, H., Wang, L., Wang, H., Zheng, Y. Arbitrary-Oriented Scene Text Detection via Rotation Proposals. IEEE Transactions on Multimedia, 2018, 20(11), 3111-3122. https://doi.org/10.1109/TMM.2018.2818020
15. Pan, X. J., Ren, Y. Q., Sheng, K. K., Dong, W. M., Yuan, H. L., Guo, X. W. Dynamic Refinement Network for Oriented and Densely Packed Object Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2020, 11204-11213. https://doi.org/10.1109/CVPR42600.2020.01122
16. Ren, S. Q., He, K. M., Girshick, R., Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6), 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031
17. Wang, J., Chen, K., Xu, R., Liu, Z., Loy, C. C., Lin, D. CARAFE: Content-Aware Reassembly of Features. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, 3007-3016. https://doi.org/10.1109/ICCV.2019.00318

18. Wei, H., Zhang, Y., Chang, Z., Li, H., Wang, H., Sun, X. Oriented Objects as Pairs of Middle Lines. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 169, 268-279. https://doi.org/10.1016/j.isprsjprs.2020.09.009

19. Xia, G. S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City: IEEE, 2018, 3974-3983. https://doi.org/10.1109/CVPR.2018.00418https://doi.org/10.1109/CVPR.2018.00418

20. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K. Aggregated Residual Transformations for Deep Neural Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, 5987-5995. https://doi.org/10.1109/CVPR.2017.634 https://doi.org/10.1109/CVPR.2017.634

21. Xu, Y., Fu, M., Wang, Q., Wang, Y., Chen, K., Xia, G., Bai, X. Gliding Vertex on the Horizontal Bounding Box for Multi-Oriented Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020. https://doi.org/10.1109/TPAMI.2020.2980827

22. Yang, X., Yang, J. R., Yan, J. C., Zhang, Y., Zhang, T. F., Guo, Z. SCRDet: Towards More Robust Detection for Small, Cluttered, and Rotated Objects. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, 8232-8241. https://doi.org/10.1109/ICCV.2019.00842

23. Yang, X., Yan, J. C., Feng, T. H. R3Det: Refined Single-Stage Detector with Feature Refinement for Rotating Object Detection. Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI), 2021, 3163-3171. https://doi.org/10.1609/aaai.v35i4.16426

24. Zhang, G., Lu, S., Zhang, W. CAD-Net: A Context-Aware Detection Network for Objects in Remote Sensing Imagery. IEEE Transactions on Geoscience and Remote Sensing, 2019, 57(12), 10015-10024. https://doi.org/10.1109/TGRS.2019.2930982

25. Zhang, H., Cisse, M., Dauphin, Y. N., Lopez-Paz, D. Mixup: Beyond Empirical Risk Minimization. Proceedings of the International Conference on Learning Representations (ICLR), 2018. https://doi.org/10.48550/arXiv.1710.09412

26. Zhu, Y. X., Wu, X. Q., Du, J. Adaptive Period Embedding for Representing Oriented Objects in Aerial Images. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(10), 7247-7257. https://doi.org/10.1109/TGRS.2020.2981203

27. Zhou, X. Y., Yao, C., Wen, H., Wang, Y. Z., Zhou, S. C., He, W. R., Liang, J. L. EAST: An Efficient and Accurate Scene Text Detector. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, 2642-2651. https://doi.org/10.1109/CVPR.2017.283