International Journal on Semantic Web and Information Systems
Volume 19 • Issue 1

A Semantic Feature Enhancement-Based Aerial Image Target Detection Method Using Dense RFB-FE

Xinyang Li, Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China*
Jingguo Zhang, Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China

ABSTRACT

Aerial image target detection is a challenging task due to the complex backgrounds, dense target distribution, and large-scale differences often present in aerial images. Existing methods often struggle to effectively extract detailed features and address the issue of imbalanced positive and negative samples. To tackle these challenges, an aerial image target detection method (dense RFB-FE-CGAM) based on dense RFB-FE and channel-global attention mechanism (CGAM) was proposed. First, the authors design a shallow feature enhancement module using dense RFB feature multiplexing and expand convolution within an SSD network, improving detailed feature extraction. Second, they introduce CGAM, a global attention module, to enhance semantic feature extraction in backbone networks. Finally, they incorporate a focal loss function for joint training, addressing sample imbalance. In experiments, the method achieved an mAP of 0.755 on the DOTA dataset and recall/AP values of 0.889/0.906 on HRSC2016, confirming the effectiveness of dense RFB-FE-CGAM for aerial image target detection.

KEYWORDS

Aerial Images, Channel-Global Attention Mechanism, Dense RFB, Multi-Scale Features, Semantic Feature Enhancement, Single Shot Multibox Detector

1. INTRODUCTION

With the iterative update of electronic communication technology and the continuous maturity of cloud computing (Bhardwaj, et al., 2022; Kumar, et al., 2022), big data (Stergiou, et al., 2021), knowledge graph (Zhou, et al., 2023), sensors (Srivastava, et al., 2022), network security (Li, et al., 2022), edge computing (Lv, et al., 2022), and artificial intelligence technology (Wang, et al., 2020; Jelusic, et al., 2022), the UAV industry has entered a rapid development stage. Drones have been applied in various fields such as power detection, environmental protection, biological detection, logistics and transportation, disaster rescue, data collection, and mobile communication (Razakarivony., & Julie., 2016). In the coming years, the deep integration of drone technology with artificial intelligence (Li,

DOI: 10.4018/IJSWIS.331083 *Corresponding Author

D., et al., 2019; Nhi, et al., 2022), image processing (Chu, et al., 2022; Qian, et al., 2022; Zheng, et al., 2022), network security (Alomani, et al., 2022; Gaurav, et al., 2022) and other technologies will not only further overcome the problems of drones in current industrial production, it will also promote the landing of UAV applications in new fields (Betti., & Tucci., 2023; Ahmed, et al., 2022; Sun, et al., 2020; Luo, et al., 2022; Zhang, et al., 2021). The wide application of drones in society has significantly improved production efficiency and also considerably reduced the consumption of human, material, and financial resources. Drones are becoming increasingly important in today's society (Luo, et al., 2022).

Currently, deep learning-based (Sayour, et al., 2022; Kadry, et al., 2021) object detection algorithms can maintain high detection performance. For common scenes, such as those with a relatively single background, a small number of targets, a large target size, and a horizontal image shooting perspective, classical object detection algorithms can maintain high detection accuracy while ensuring detection speed (Fu, et al., 2021; Kim, et al., 2008; Betti., 2022). Aerial imagery is divided into satellite imagery captured by satellites and UAV imagery captured using UAVs; aerial imagery captured by satellites is characterized by large size, fixed shooting angle, and a small percentage of targets in the image (Huang. et al.,; Zhong, et al., 2017; Han, et al., 2019; Cao, et al., 2020; Elhagry, et al., 2022).

UAV images are more complex and richer because of the limitations of the shooting equipment, environment, and other factors. Compared with satellite images, UAV images are more widely used in civil and military fields, thus, exploring the potential information of UAV images is important for in-depth applications of UAVs in various fields (Zhang, et al., 2020; Dikbayir., & BÜLBÜL., 2020; Chen, et al., 2020). The application of UAVs in the civilian sector is relatively established, with the following specific application scenarios:

1. **Application of UAVs to environmental management:** Drones can be equipped with infrared cameras with night vision shooting, and the drones use the characteristics of strong mobility and a wide shooting field of view to monitor the target area of haze around the clock. Environmental governance personnel can quickly locate the source of pollution according to the images captured by drones to eliminate pollution (Chen, et al., 2022).
2. **Application of drones in electric power inspection:** When drones inspect power lines, they use the positioning equipment onboard and combine it with target detection technology to quickly report potentially dangerous areas to maintenance personnel, who can quickly overhaul the power lines according to the alarm location to minimize the probability of damage to the power lines (Tian, et al., 2020; Xu., & Wu., 2020; Divya, 2019; Zou., & Shi., 2017).
3. **Application of drones to agriculture:** With the continuous expansion of the agricultural planting scale, today's agricultural development has entered an era of intelligent agriculture. The common application scenarios of drones in agriculture are as follows: drones replace traditional machines to spray pesticides on large crops; for certain crops prone to pests and diseases, drones can be combined with biological monitoring technology to monitor the crop-growing area and achieve early warnings of large-scale agricultural pests and diseases (Wang, et al., 2018; Yang, et al., 2021; Zhao, et al., 2021; Estalayo, et al., 2006).
4. **Application of UAVs during major disasters:** In the event of a sudden earthquake, the communication system in the disaster area is suddenly paralyzed, and drones can carry relevant communication equipment to fly to the center of the disaster area and provide emergency communication capabilities. In the event of a forest fire, drones can monitor the fire in close proximity in the shortest possible time, find trapped people in time, and then send a distress signal to the control center to help firefighters quickly locate and extinguish the fire.
5. **Application of drones in crime prevention:** In a complex urban environment, drones can use their characteristics of not being bound by space to track and monitor certain criminals in real-time and provide relevant information to public security personnel in a timely manner. Drones play

an important role in the military. At the national border, drones can replace the army to conduct all-weather patrols and monitor illegal border crossing and smuggling of foreign personnel; in the context of modern electronic warfare, stealth drones can effectively avoid enemy interception and perform the task of close surveillance or attack; near important military bases, drones can carry infrared cameras to perform real-time vigilance tasks to prevent enemy sneak attacks.

In summary, UAVs play an irreplaceable role in modern society (Chen., & Liu., 2017; Ouyang., & Wang., 2019). One of the main reasons UAVs are widely used in both civilian and military fields is that UAVs carry shooting equipment that can provide timely and information-rich images, which can be processed using computer vision-related technologies such as target detection and target tracking to obtain more accurate potential information. In the future, UAV technology and computer vision will remain the focus of research scholars and industry, and the interaction of the two technologies will not only continue to deepen the application of UAVs in existing fields but also lead to new application directions for UAVs (Huang, et al., 2023).

However, current methods in the field of aerial image target detection still face several challenges, especially when dealing with issues such as complex backgrounds, dense target distributions, and large-scale variations. Challenges include poor robustness against background interference, difficulty accurately modeling dense or differently scaled targets, and more. Compared to satellite imagery, UAV images have a wider range of applications in both civilian and military fields. Therefore, it is crucial to explore the potential information within UAV images for the extensive utilization of drones in various domains. In order to effectively improve the performance of aerial image-target detection, a target detection method based on dense RFB Feature Enhance (RFB-FE) and channel-global attention mechanism (CGAM) is proposed (called dense RFB-FE-CGAM). This approach integrates advanced techniques such as shallow feature enhancement, global attention mechanisms (Chopra, et al., 2022), and focal loss functions to effectively address the complexity challenges in the field of unmanned aerial vehicle image target detection. By introducing a shallow feature enhancement module, it enhances the ability to handle complex backgrounds. The adoption of a global attention mechanism strengthens semantic feature extraction in scenarios with densely distributed targets. Additionally, joint training and focal loss functions effectively handle large-scale differences. This comprehensive approach not only improves the accuracy and robustness of UAV image target detection but also extends its wide applicability across various domains.

In summary, the main contributions are as follows:

1. A shallow feature enhancement module was designed based on the single shot multibox detector (SSD) network, this module enhances feature extraction capability by connecting the outputs of different branches and utilizing feature multiplexing to link the output of the previous branch with the input of the next branch, thereby enriching the scale and diversity of the perceptual field.
2. The CGAM was designed by introducing the global attention module (GAM), it utilizes dilated convolution instead of pooling layer operations in SAM, effectively extracting deep semantic information by reducing the loss of fine details during downsampling, and it can better capture the relationship between targets and context through global attention.
3. The focal loss function was introduced for joint training, which will effectively avoid the positive and negative sample imbalance problem, thereby improving the precision of target detection.

2. RELATED WORKS

With the enhancement and high-speed development of UAV remote sensing technology, collected aerial images have the advantages of a large area, rich content, and undisturbed information acquisition, and have gained widespread attention. As simple low-altitude flight tools, UAVs have achieved

remarkable results in geological exploration, environmental investigation, victim detection, and other fields, and their development prospects are immeasurable. Reference (Li, et al., 2019) proposed a detection method for attention-feature pyramid networks using channel attention modules and dot product attention modules. Reference (Chen, et al., 2021) proposed cascaded attentional networks to suppress the background noise in a feature pyramid network from coarse to fine. Reference (Li, et al., 2020) proposed a multi-layer attention network that combines positional attention and channel attention; reference (Yang, et al., 2020) proposed a dual-path feature attention network to guide the network to focus on the target region; reference (Zhu, et al., 2021) added the transformer model to YOLOv5. However, the use of global attention in extremely large aerial images with distant target information distributions can introduce interference information and redundant computations. In ref. (Liu, et al., 2021), a swine transformer model was proposed, and the sliding window design enabled the self-attentive-based transformer model to have linear computational complexity while retaining the ability to extract global information, which is highly effective for intensive predictive tasks. However, the sliding window is divided manually, which is limited by the need to redesign the window size and retrain it when the feature map scale changes. Reference (Gawande, et al., 2022) used selective search methods to reduce the time consumption of a sliding window brute-force search. Reference (Abdullah, et al., 2022) used a chunking and cutting method to segment large images into small images, performed subsequent detection and recognition on each small image separately, and finally stitched all detection results. In ref. (Wu, et al., 2019), the sensitivity of convolutional neural network feature extraction to angles was weakened by introducing rotation-invariant quantum modules, and the effect of directional diversity on detection was mitigated; in ref. (Ying, et al., 2019), the size of feature images was increased by introducing shallow and deep feature up-sampling; in ref. (Yang, et al., 2020), an attention mechanism was introduced to weaken the background and enhance the target information, and an additional mask calculation was introduced, which led to an increase in network computation. Reference (Maesako., & Zhang., 2022) proposed an additive target-area masking method (AVIS) to suppress the computational load.

Although deep learning has made significant progress in small target detection in remote sensing images, there are still shortcomings. For example:

1. Given the inherent characteristics of remote sensing images, the multi-scale and diversity of remote sensing image targets, especially for arbitrarily densely arranged small targets, existing detection networks lack an effective combination of deep and shallow features, which can easily overlook detailed information. Therefore, the feature extraction ability of detection networks needs to be improved.
2. The background of remote sensing images is complex and easily affected by natural factors such as lighting and clouds, resulting in an imbalance between positive and negative samples, which will reduce the precision of target detection.

To solve these problems faced by existing aerial image target detection methods, based on the SSD algorithm, a new target detection method using the dense RFB-FE and CGAM is proposed. Specifically, the dense RFB-FE and CGAM were used to improve the shallow and deep level feature extraction capabilities of the network, respectively, and the focus classification Loss function was used to effectively solve the feature imbalance problem between positive and negative samples.

3. PROPOSED AERIAL IMAGE-TARGET DETECTION METHOD

3.1 Dense RFB-FE CGAM Model Architecture

The perception field size of the shallow feature layer of the SSD algorithm is 92×92, and considering that the input image size is 800×800, the SSD algorithm can only establish connections among local

feature points and cannot capture dependencies between features over longer distances, and the six detection branches of the SSD algorithm are independent of each other, the features extracted from each layer lack contextual information. This study improved the SSD algorithm, as shown in Figure 1. In this study, a multi-scale ship target detection algorithm based on dense RFB and CGAM is proposed with the following main improvements: First, a dense RFB shallow feature enhancement module (dense RFB-FE) is designed to enhance the shallow detail information of the network; second, CGAM is designed to effectively extract the deep semantic information (Chu, et al., 2022); and third, we the loss function is designed, and the focused classification loss function is added.

3.2 Channel Global Attention Mechanism

Remote sensing images are characterized by complex background information, an excessive presence of small targets, and various target sizes. When performing multiple downsampling convolution operations to extract feature information, small targets occupy fewer pixels, and repeated iterations of the background generate redundant information. Therefore, this study improved the convolutional block attention module (CBAM). The original CBAM module is a combination of the channel attention module (CAM) and spatial attention module (SAM). However, because of the limited coverage of the perceptual field, the pooling operation causes insufficient extraction of target feature information from the deep feature map, particularly for small targets, when the backbone network extracts feature information after multiple downsampling operations. With the proposed cavity convolution, the ordinary 3×3 convolution can only cover an area of nine with the same number of parameters, whereas the cavity convolution can cover an area of 25 or more based on a 3×3 convolution, which can be equivalent to 5×5, 7×7, etc. As the cavity rate increases, the difficulty in recovering detailed information during the upsampling process decreases. In addition, because the remote sensing image contains complex background information when the backbone network continuously downsamples to extract feature information, the background information generates redundant information owing to repeated iterations. This redundant information interferes with the focus-of-attention mechanism in the depth feature map, leading to false or incorrect detections. By contrast, Baudes' nonlocal mean method is different from traditional denoising methods that use local information to filter images and can effectively suppress the noise present in the image. Based on the aforementioned analysis, a CGAM based on CBAM is designed by introducing a global attention module (GAM), which uses hole convolution instead of the pooling layer operations in SAM. In this study, we used a hole rate similar to the sawtooth waveform and set the hole rate d as 1, 2, and 3, enabling the convolution kernel to cover a larger area under the same parameters. This study changed the connection order between modules and utilized contextual information to establish global dependencies between channels and each pixel in the global feature map to suppress redundant information interference.

Figure 2 shows the CGAM structure, Figure 3 shows the CAM structure, and Figure 4 shows the GAM structure. The two one-dimensional vectors obtained after pooling were shared using a multilayer perceptron (MLP) network. To reduce the computational effort and model parameters, a

Figure 1. Structure of ship target detection network

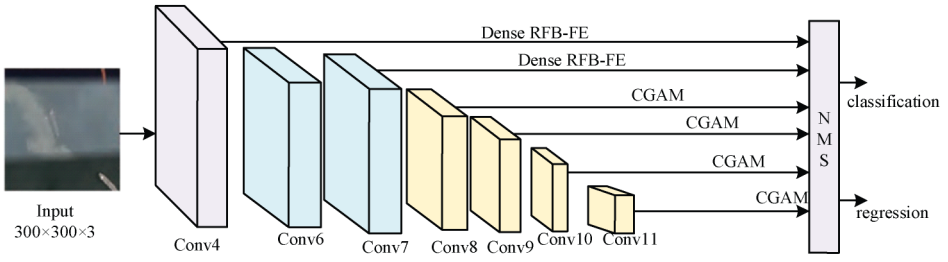


Figure 2. CGAM structure

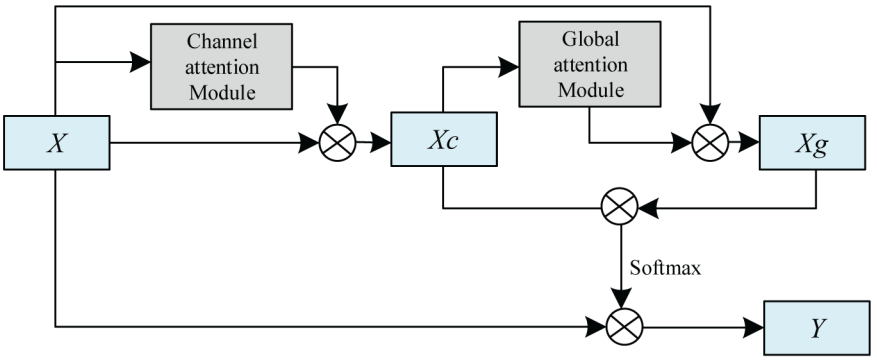


Figure 3. CAM structure

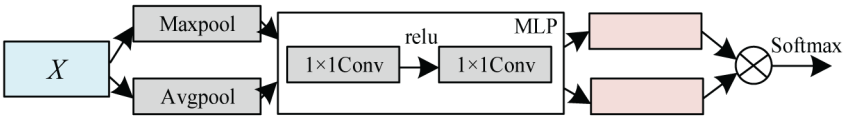
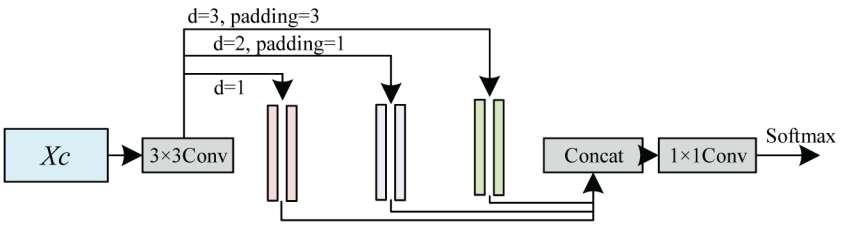


Figure 4. GAM structure



1×1 convolution was used instead of a fully connected layer, and the number of channels was first reduced by a 1×1 convolution, accelerated by a ReLU activation function, and then restored by a 1×1 convolution. The output results were summed in the channel dimensions and normalized to 0-1 using the sigmoid activation function to obtain a channel attention weight Zc of size 1×1×C.

Specifically, CGAM introduces a Global Attention Module, whose task is to capture global contextual information within the image. It achieves this by utilizing dilated convolution, which expands the receptive field to better understand the overall structure and relationships within the image. The dilation introduces holes in the convolution kernel, allowing it to cover more pixels and thus better capture target features. Furthermore, by employing the GAM and dilated convolution, CGAM establishes global dependencies, meaning it can better comprehend the relationships between different regions in the image. This aids in capturing semantic information within the image effectively.

In summary, CGAM serves to enhance the feature extraction capabilities of the backbone network, especially during multiple downsampling convolution operations, particularly when dealing with small targets and complex backgrounds. It achieves this by expanding the receptive field, introducing global contextual information, and optimizing convolution operations. This improvement contributes

to enhancing the performance of computer vision tasks such as target detection and classification, especially when dealing with images containing complex structures and targets of varying sizes.

3.3 Dense RFB-FE Module

In this study, the dense RFB-FE module was designed with the idea of a dense connection, which connects the output of each branch with the input of the next branch through feature multiplexing to extract target features, enrich the scale variety of the sensory field, further increase the sensory field, and adapt to feature extraction and effective detection of multi-scale targets. As shown in Figure 5, the detection branch of $19 \times 19 \times 1024$, for example, is no longer limited to increasing the number of branches, expansion rate, and convolution kernel by the dense RFB-FE module. The output of the first branch is connected to the second input as $19 \times 19 \times 256 + 1024$ through feature multiplexing, and the outputs of both the first and second branches are added to the third branch as $19 \times 19 \times 256 + 1024 + 256$. In other words, the output feature map of the previous branch is convolved by the null in the subsequent branch, which further expands the perceptual field and enhances detailed feature extraction and localization.

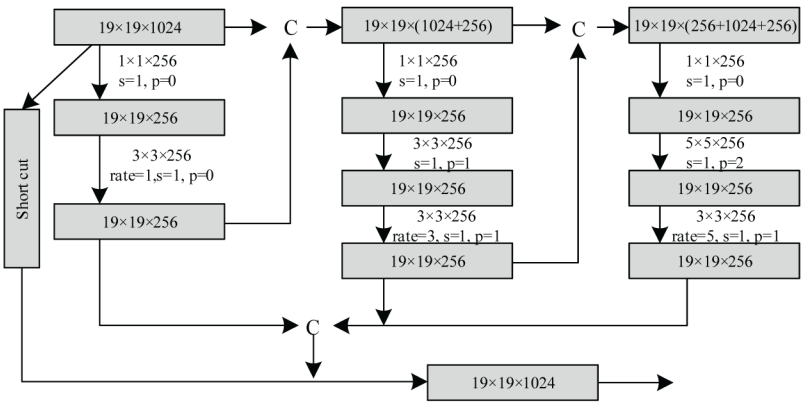
Each layer of this module applies densely connected Receptive Field Blocks, enhancing the network's capability to extract richer features. This module aids in capturing both local and global background information in aerial images, helping the network learn more discriminative features, which is crucial for accurate target detection. Furthermore, through improved feature extraction, it enhances the overall detection accuracy of the method, enabling the network to accurately locate and classify targets in aerial images. Additionally, the dense RFB-FE module addresses robustness against scale variations and complex backgrounds by capturing multiscale patterns, allowing the network to effectively capture contextual information to handle scale changes and improve detection performance.

3.4 Loss Function Design

For the classification task, the focal loss function was used:

$$L_{fl}(x,c) = -\sum_{c=1}^{l}\sum_{i\in Pos}^{N} \alpha_t (1-c_{i,c})^{\gamma} \log(\hat{x}_{i,c}) - \sum_{c=0}\sum_{j\in N_{eg}} (1-\alpha_t)\log(\hat{x}_{j,c}) \quad (1)$$

Figure 5. Structure of dense RFB-FE module (19 × 19 × 1024)



where, $\hat{x}_{i,c}$ denotes the probability that the category is background and correct, and the loss function hyperparameters are $\alpha_t \in [0, 1]$ and $\gamma \in [0, 5]$. In this study, the focus classification Loss function was designed. In the other words, we empirically adopted $\gamma = 2$, and the model focused on negative samples for training, which effectively solved the positive and negative sample imbalance problem, thereby improving the detection precision.

The primary reason for using the focal loss function is to address the imbalance between positive and negative samples in aerial image target detection. This loss function focuses on challenging samples, mitigates the disparity in the number of positive and negative samples, and enhances the model's robustness, thereby effectively improving target detection performance, especially in scenarios where positive samples are relatively scarce but of greater significance.

The regression task used the Smooth L1 function, where, l and g contained the four elements of the object position $\left(\hat{g}_j^{cx}, \hat{g}_j^{cy}, \hat{g}_j^{w}, \hat{g}_j^{h}\right)$:

$$L_{loc}(x, l, g) = -\sum_{i \in P_{os,m}} \sum_{\in \{cx,cy,w,h\}} x_{i,j}^{k} smooth_{L1}(l_i^m - \hat{g}_j^m) \tag{2}$$

$$smooth_{L1}(x) = \begin{cases} 0.5x^2 \ \ if \ \ (|\ x\ | < 1) \\ |\ x\ | - 0.5 \ \ otherwise \end{cases} \tag{3}$$

4. EXPERIMENT AND ANALYSIS

4.1 Experimental Environment

The experimental conditions are listed in Table 1.

4.2 Dataset and Details

The DOTA dataset (Xia, et al., 2018) has 2806 aerial images containing different scene sizes with resolutions ranging from 800 × 800 to 4000 × 4000, including 188282 instances in 15 categories. The labeling method is a quadrilateral of arbitrary shape and orientation determined by four points and can be applied to the detection tasks of a horizontal bracket box and an oriented bracket box.

HRSC2016 (Liu, et al., 2017) is a large dataset collected for ship detection, and it was extracted from six important ports on Google Earth. The training, validation, and test sets comprised 436, 181, and 444 images, respectively.

Table 1. Experimental platform settings

Experimental Environment	Specific Information
Operating system	Ubuntu18.06
Memory	64GB
Language	Python3.8
Development tool	PyCharm
GPU	GTX2080Ti
Development platform	Tensorflowl.8.0

In the experiments, the initial anchor scales were set to 2, and the aspect ratio was set to 1. The strides were set to 8, 16, 32, 64, and 128, respectively. In the loss function, the weights for different loss components were all set to 1. We utilized the Adam optimizer for network training with an initial learning rate of 2.5×10^{-5}, and it was decreased by a factor of 10 at each decay step. The batch size was set to 2. We employed a warm-up strategy with 3 epochs for network pre-training, with a learning rate of 1×10^{-5}. For the DOTA dataset, the training lasted for 30 epochs, and random flips were applied for data augmentation. For the HRSC2016 dataset, the training consisted of 20 epochs without any data augmentation. In this case, only the first category, which is all objects treated as ships, was considered.

4.3 Evaluation Indexes

In this study, we adopted AP as the accuracy evaluation index of the model detection target. The index is calculated as follows:

$$P = \frac{TP}{TP + FP} \tag{4}$$

$$R = \frac{TP}{TP + FN} \tag{5}$$

$$AP = \int_0^1 PRdR \tag{6}$$

TP denotes the number of correctly identified targets; FP denotes the number of incorrectly identified targets; and FN denotes the number of unidentified targets.

4.4 Comparison Experiments

4.4.1 Comparison of Different Algorithms on DOTA and HRSC2016 Datasets

To demonstrate the target detection capability of the proposed method, the proposed method was compared with that of YOLOv2 (Chen, et al., 2020), Faster RCNN (Elhagry, et al., 2022), TPH-YOLOv5 (Zhu, et al., 2021), MFIAR-Net (Yang, et al., 2020) and AVIS (Maesako., & Zhang., 2022), and the image pixel size was set to 800 × 800 for fairness reasons. The comparison results of all methods on the DOTA dataset are listed in Table 2, and Table 3 lists the comparative experimental results of all methods on the HRSC2016 dataset. Among these six models, the proposed Dense RFB-FE-CGAM outperformed the YOLOv2, Faster RCNN, TPH-YOLOv5, MFIAR-Net and AVIS methods in the aerial image target detection task.

On the DOTA dataset, this method achieved a mAP of 0.755. DOTA is a widely recognized dataset for aerial image target detection, encompassing various object categories and a range of object sizes. Attaining a high mAP indicates that this method can accurately detect and locate objects of different categories while handling scale variations. These results demonstrate the competitiveness of this approach in aerial image target detection and underscore its capability to perform well in complex scenarios.

On the HRSC2016 dataset, this method demonstrates impressive recall and precision rates. HRSC2016 is a dataset specifically focused on large ship targets, which often exhibit complex shapes and textures. By achieving high recall and precision rates, this method validates its effectiveness in

Table 2. Comparison of the experimental results on the DOTA dataset

Method	PL	BR	HA	GTF	SV	LV	RA	SP
YOLOv2	0.835	0.483	0.446	0.567	0.702	0.802	0.623	0.733
Faster RCNN	0.862	0.528	0.517	0.657	0.837	0.859	0.652	0.755
TPH-YOLOv5	0.814	0.501	0.508	0.653	0.846	0.782	0.633	0.702
MFIAR-Net	0.865	0.528	0.482	0.612	0.732	0.754	0.612	0.688
AVIS	0.855	0.529	0.511	0.693	0.766	0.792	0.627	0.746
Dense RFB-FE-CGAM (ours)	0.871	0.502	0.523	0.683	0.809	0.838	0.685	0.736
Method	**SH**	**SBF**	**TC**	**BC**	**ST**	**BD**	**HC**	**mAP**
YOLOv2	0.817	0.535	0.824	0.778	0.828	0.818	0.588	0.692
Faster RCNN	0.835	0.602	0.836	0.829	0.856	0.849	0.612	0.739
TPH-YOLOv5	0.842	0.597	0.902	0.812	0.837	0.828	0.623	0.725
MFIAR-Net	0.793	0.578	0.841	0.788	0.808	0.799	0.653	0.702
AVIS	0.882	0.639	0.897	0.828	0.869	0.836	0.679	0.743
Dense RFB-FE-CGAM (ours)	0.891	0.618	0.910	0.881	0.876	0.851	0.651	0.755

Table 3. Comparison of the experimental results on the HRSC2016 dataset

Method	Recall	AP
YOLOv2	0.817	0.822
Faster RCNN	0.873	0.881
TPH-YOLOv5	0.847	0.872
MFIAR-Net	0.851	0.864
AVIS	0.878	0.893
Dense RFB-FE-CGAM (ours)	0.889	0.906

detecting and locating challenging ship targets. This is of significant importance for applications such as maritime border surveillance and maritime security.

Figures 6 and 7 provide a visual representation of the results obtained through our proposed method applied to two datasets: DOTA and HRSC2016. Our method employs a strategically designed shallow feature enhancement module based on the SSD network. This module utilizes dense RFB feature multiplexing and inflated convolution techniques to amplify the scale and diversity of the perceptual fields, aligning them more closely with the human eye's viewpoint map. This enhancement significantly bolsters the shallow network's proficiency in extracting intricate and detailed features.

Furthermore, we introduce a Channel-Global Attention Mechanism (CGAM) to our method, which substitutes null convolution for the pooling layer operation seen in SAM. This substitution mitigates the loss of fine-grained information during the downsampling process and augments the network's backbone feature extraction capability. Consequently, these enhancements contribute to superior detection performance.

Specifically, the Dense RFB-FE-CGAM method achieves deep feature extraction and multiscale target adaptability by introducing the CGAM structure and Dense RFB-FE module. Experimental results indicate that, compared to other methods, this approach yields superior performance in target

Figure 6. Visualization results of Dense RFB-FE-CGAM on the DOTA dataset

detection. These findings are of significant importance for further refining aerial image target detection methods and enhancing detection accuracy. Additionally, they provide valuable guidance and reference for future researchers in the field of aerial image analysis. Furthermore, with fewer parameters and faster inference speed, it further demonstrates the superiority of the Dense RFB-FE-CGAM model.

Furthermore, as shown in Table 4, our method demonstrates excellent performance in terms of computational resources and speed. Firstly, we utilize the Single Shot MultiBox Detector (SSD) network as the underlying model, which offers high computational efficiency. Secondly, we incorporate the Channel-Global Attention Mechanism (CGAM) and the Dense RFB-FE (Feature Enhancement) module, both of which are carefully designed and optimized to enhance feature extraction capabilities and detection performance. Compared to other methods, our approach delivers superior detection accuracy and faster processing speeds. By conducting experiments under the same hardware and environmental conditions for comparison, our method outperforms in both precision and speed in target detection. However, it's worth noting that specific performance may vary depending on factors such as hardware configuration, experimental environment, and dataset used.

4.4.2 Ablation Experiments

To verify the effectiveness of the proposed method for aerial image target detection, the corresponding ablation experiments were designed using the control variable method.

Figure 7. Visualization results of dense RFB-FE-CGAM on the HRSC2016 dataset

Table 4. Comparison of speed performance of the proposed and existing models

Method	FLOPS (G)	Inference time (s)	Params (M)
YOLOv2	124.6	0.05	54.26
Faster RCNN	215.6	0.028	41.15
TPH-YOLOv5	50.4	0.02	7.3
MFIAR-Net	107.3	0.024	39.71
AVIS	152.1	0.016	69.76
Dense RFB-FE-CGAM (ours)	179.7	0.014	23.16

Experiment I: RFB-FE method
Experiment II: Dense RFB-FE method
Experiment III: Dense RFB-FE-CBAM method
Experiment IV: Dense RFB-FE-CGAM (Without Focused Loss)
Experiment V: Dense RFB-FE-CGAM

In the conducted ablation experiments, we conduct a thorough comparison between our proposed Dense RFB-FE-CGAM model and the baseline stem network, and we have presented the results in Tables 5 and 6. It's noteworthy that the performance metrics of our proposed Dense RFB-FE-CGAM outperformed those of the base-stem network across both the DOTA dataset and the HRSC2016 dataset. This notable improvement can be attributed to several key factors inherent to our Dense RFB-FE-CGAM approach.

Firstly, our Dense RFB-FE-CGAM method incorporates a strategically designed shallow feature enhancement module, leveraging the strengths of the SSD network architecture. This module efficiently exploits dense RFB feature multiplexing and expanded convolution techniques, significantly augmenting the scale and diversity of the perceptual field, thereby aligning it more closely with the human eye's viewpoint map. This enhances perceptual field boosts the shallow network's capacity to extract intricate and informative features from the input data.

Furthermore, our approach introduces the innovative Channel-Global Attention Mechanism (CGAM), a global attention module. This module plays a crucial role in enhancing semantic feature extraction within the backbone networks, contributing significantly to improved target detection performance.

In addition to these innovations, we have integrated a focal loss function into our model's joint training process, effectively addressing the challenge of sample imbalance. This addition ensures that our model can focus more on critical samples, ultimately leading to improved performance.

Importantly, our experiments also include a comparison with an alternative scenario where the focusing classification loss function was not utilized. This comparison clearly demonstrated that our proposed Dense RFB-FE-CGAM model consistently achieves superior results, thus validating the effectiveness of the focusing classification loss function as an essential component of our approach.

5. CONCLUSION

This study proposes an aerial image target detection method based on dense RFB-FE and CGAM for the characteristics of a complex background, dense target distribution, and large-scale differences

Table 5. Results of ablation experiments on the DOTA dataset

Method	Recall	mAP
RFB-FE	0.747	0.734
Dense RFB-FE	0.751	0.743
Dense RFB-FE-CBAM	0.755	0.748
Dense RFB-FE-CGAM (Without Focused Loss)	0.760	0.751
Dense RFB-FE-CGAM (ours)	0.768	0.755

Table 6. Results of ablation experiments on the HRSC2016 dataset

Method	Recall	AP
RFB-FE	0.825	0.813
Dense RFB-FE	0.851	0.862
Dense RFB-FE-CBAM	0.875	0.890
Dense RFB-FE-CGAM (Without Focused Loss)	0.881	0.898
Dense RFB-FE-CGAM (ours)	0.889	0.906

in aerial images. The proposed method designs a shallow feature enhancement module based on an SSD network and a CGAM, using hole convolution instead of a pooling layer operation in the SAM to reduce the loss of detailed information during downsampling and introduce a focal loss function for joint training. The experiments demonstrated the effectiveness of the proposed method for aerial image target detection. And there is potential for expansion into other areas, such as agricultural field detection and management, among others.

However, the proposed model also has some limitations:

1. Aerial scenes have the characteristics of a small percentage of the target pixel area and a relatively scattered and sparse distribution, which reduces the detection precision of small targets. Without increasing the computing cost, the subsequent algorithm can use a cascade structure combining coarse detection and fine detection, and the detection precision will be tried to improve through parameter optimization and other methods.
2. While pursuing high model accuracy, ensuring the feasibility and performance control of the model in practical applications is of paramount importance. In our future work, we will propose the following strategies, including selecting appropriate model architectures, transfer learning, model pruning and quantization, model compression, edge computing platform selection, model lightweighting, model optimization, as well as system design and performance monitoring. These are all key steps contributing to achieving efficient, practical, and controllable drone applications. These considerations will have a significant impact on future research and real-world applications.
3. We took into consideration both model accuracy and parameter count when deciding to reduce the complexity of the model without significantly affecting accuracy. Specifically, we made each branch of the model handle its processing independently. However, this independence among branches can lead to a lack of contextual information between layers. As for how to capture comprehensive global context information, we will further investigate this in our future work.

Our future research directions include improving small target detection precision through a cascade structure, optimizing model feasibility and performance for practical applications, and exploring methods to capture comprehensive global context information within the model.

DATA AVAILABILITY

The data used to support the findings of this study are included within the article.

CONFLICTS OF INTEREST

The author declares that there is no competing interest for this work, and no funding was received.

FUNDING STATEMENT

This research received no external funding.

REFERENCES

Abdullah, T. A. (2022). *Plant Coverage Estimation and Missing Plant Detection of Rice Crops Using UAV Imagery* [Doctoral dissertation]. Lamar University-Beaumont.

Ahmed, M., Wang, Y., Maher, A., & Bai, X. (2022). Fused RetinaNet for small target detection in aerial images. *International Journal of Remote Sensing*, *43*(8), 2813–2836. doi:10.1080/01431161.2022.2071115

Almomani, A., Alauthman, M., Shatnawi, M. T., Alweshah, M., Alrosan, A., Alomoush, W., Gupta, B. B., Gupta, B. B., & Gupta, B. B. (2022). Phishing website detection with semantic features based on machine learning classifiers: A comparative study. *International Journal on Semantic Web and Information Systems*, *18*(1), 1–24. doi:10.4018/IJSWIS.297032

Betti, A. (2022) *A lightweight and accurate YOLO-like network for small target detection in Aerial Imagery.* arXiv preprint arXiv:2204.02325.

Betti, A., & Tucci, M. (2023). YOLO-S: A Lightweight and Accurate YOLO-like Network for Small Target Detection in Aerial Imagery. *Sensors (Basel)*, *23*(4), 1865–1872. doi:10.3390/s23041865 PMID:36850465

Bhardwaj, A., & Kaushik, K. (2022). Predictive analytics-based cybersecurity framework for cloud infrastructure. *International Journal of Cloud Applications and Computing*, *12*(1), 1–20. doi:10.4018/IJCAC.297106

Cao, C., Wu, J., Zeng, X., Feng, Z., Wang, T., Yan, X., Wu, Z., Wu, Q., & Huang, Z. (2020). Research on airplane and ship detection of aerial remote sensing images based on convolutional neural network. *Sensors (Basel)*, *20*(17), 4696–4708. doi:10.3390/s20174696 PMID:32825315

Chen, C. H., & Liu, K. H. (2017) Stingray detection of aerial images with region-based convolution neural network, *2017 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, 175-186. doi:10.1109/ICCE-China.2017.7991052

Chen, G., Yi, X., & Li, Z. (2020). Third-party construction target detection in aerial images of pipeline inspection based on improved YOLOv2 and transfer learning. *Jisuanji Yingyong*, *40*(4), 1062–1073.

Chen, H. W., Chen, C. Y., Nguyen, K. L. P., Chen, B. J., & Tsai, C. H. (2022). Hyperspectral sensing of heavy metals in soil by integrating AI and UAV technology. *Environmental Monitoring and Assessment*, *194*(7), 518. doi:10.1007/s10661-022-10125-5 PMID:35731279

Chen, L., Liu, C., Chang, F., Li, S., & Nie, Z. (2021). Adaptive Multi-LevelFeature Fusion and Attention-Based Network for Arbitrary-Oriented Object Detection in Remote Sensing Imagery. *Neurocomputing*, *451*(2), 67–80. doi:10.1016/j.neucom.2021.04.011

Chopra, M., Singh, S. K., Sharma, A., & Gill, S. S. (2022). A comparative study of generative adversarial networks for text-to-image synthesis. *International Journal of Software Science and Computational Intelligence*, *14*(1), 1–12. doi:10.4018/IJSSCI.300364

Chu, J., Zhao, X., Song, D., Li, W., Zhang, S., Li, X., & Liu, A. A. (2022). Improved Semantic Representation Learning by Multiple Clustering for Image-Based 3D Model Retrieval. *International Journal on Semantic Web and Information Systems*, *18*(1), 1–20. doi:10.4018/IJSWIS.297033

Dikbayir, H. S., & Bülbül, H. Ï. (2020). Deep Learning based vehicle detection from aerial images. *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA),* 956-960.

Divya, N. (2019) Image Specific Similar Target Detection in Aerial Images Using Gaussian Mixture Model. *2019 International Carnahan Conference on Security Technology (ICCST)*, 1-5. doi:10.1109/CCST.2019.8888422

Elhagry, A. (2022). *Investigating the Challenges of Class Imbalance and Scale Variation in Object Detection in Aerial Images*. arXiv:2202.02489.

Estalayo, E., Salgado, L., & Jaureguizar, F. (2006) Efficient image stabilization and automatic target detection in aerial FLIR sequences. *Automatic Target Recognition XVI*, 184-195.

Fu, L., Gu, W., Li, W., Chen, L., Ai, Y., & Wang, H. (2021). Bidirectional parallel multi-branch convolution feature pyramid network for target detection in aerial images of swarm UAVs. *Defence Technology*, *17*(4), 1531–1541. doi:10.1016/j.dt.2020.09.018

Gaurav, A., Gupta, B. B., & Panigrahi, P. K. (2022). A comprehensive survey on machine learning approaches for malware detection in IoT-based enterprise information system. *Enterprise Information Systems*, 1–25.

Gawande, U., Hajari, K., & Golhar, Y. (2022). SIRA: Scale illumination rotation affine invariant mask R-CNN for pedestrian detection. *Applied Intelligence*, *52*(9), 10398–10416. doi:10.1007/s10489-021-03073-z PMID:35039716

Han, L., Tao, P., & Martin, R. R. (2019). Livestock detection in aerial images using a fully convolutional network. *Computational Visual Media*, *5*(5), 221–228. doi:10.1007/s41095-019-0132-5

Huang, M., Zhang, Y., & Chen, Y. (n.d.). Small Target Detection Model in Aerial Images Based on TCA-YOLOv5m. *IEEE Access : Practical Innovations, Open Solutions*, *2*(21), 368–387. doi:10.1109/ACCESS.2022.3232293

Huang, T., Zhu, J., Liu, Y., & Tan, Y. (2023). UAV aerial image target detection based on BLUR-YOLO. *Remote Sensing Letters*, *14*(2), 186–196. doi:10.1080/2150704X.2023.2174385

Jelusic, P. B., Poljicak, A., Donevski, D., & Cigula, T. (2022). Low-Frequency Data Embedding for DFT-Based Image Steganography. *International Journal of Software Science and Computational Intelligence*, *14*(1), 1–11. doi:10.4018/IJSSCI.312558

Kadry, S., Taniar, D., Meqdad, M. N., Srivastava, G., & Rajinikanth, V. (2021, November). Assessment of Brain Tumor in Flair MRI Slice with Joint Thresholding and Segmentation. In *International Conference on Mining Intelligence and Knowledge Exploration* (pp. 47-56). Cham: Springer International Publishing.

Kim, Z. W., & Sengupta, R. (2008). Target detection and position likelihood using an aerial image sensor. *2008 IEEE International Conference on Robotics and Automation*, 59-64.

Kumar, S., Kumar, S., Ranjan, N., Tiwari, S., Kumar, T. R., Goyal, D., Sharma, G., Arya, V., & Rafsanjani, M. K. (2022). Digital watermarking-based cryptosystem for cloud resource provisioning. *International Journal of Cloud Applications and Computing*, *12*(1), 1–20. doi:10.4018/IJCAC.311033

Li, C., Xu, C., & Cui, Z. (2019). Feature-Attentioned ObjectDetection in Remote Sensing Imagery. *Proceedings ofthe 26th IEEE International Conference on Image Processing*, 3886-3890.

Li, D., Deng, L., Gupta, B. B., Wang, H., & Choi, C. (2019). A novel CNN based security guaranteed image watermarking generation scenario for smart city applications. *Information Sciences*, *479*, 432–447. doi:10.1016/j.ins.2018.02.060

Li, S., Qin, D., Wu, X., Li, J., Li, B., & Han, W. (2022). False alert detection based on deep learning and machine learning. *International Journal on Semantic Web and Information Systems*, *18*(1), 1–21. doi:10.4018/IJSWIS.313190

Li, Y. Y., Huang, Q., Pei, X., Jiao, L., & Shang, R. (2020). RADet: Refine FeaturePyramid Network and Multi-Layer Attention Network forArbitrary-Oriented Object Detection of Remote SensingImages. *Remote Sensing (Basel)*, *12*(3), 389–403. doi:10.3390/rs12030389

Liu, Z., Lin, Y., & Cao, Y. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *Proceedings of the 18th IEEE/CVF InternationalConference on Computer Vision*, 9992-10002. doi:10.1109/ICCV48922.2021.00986

Liu, Z., Yuan, L., & Weng, L. (2017). A High Resolution Optical Satellite ImageDataset for Ship Recognition and Some New Baselines. *ICPRAM*, 324-331.

Luo, X., Wu, Y., & Wang, F. (2022). Target detection method of UAV aerial imagery based on improved YOLOv5. *Remote Sensing (Basel)*, *14*(19), 5063–5070. doi:10.3390/rs14195063

Luo, X., Wu, Y., & Zhao, L. (2022). YOLOD: A target detection method for UAV aerial imagery. *Remote Sensing (Basel)*, *14*(14), 3240–3251. doi:10.3390/rs14143240

Lv, L., Wu, Z., Zhang, L., Gupta, B. B., & Tian, Z. (2022). An edge-AI based forecasting approach for improving smart microgrid efficiency. *IEEE Transactions on Industrial Informatics*, *18*(11), 7946–7954. doi:10.1109/TII.2022.3163137

Maesako, K., & Zhang, L. (2022). AVIS: An Innovative Image Preprocessing Method for Object Detection of Aerial Images. *2022 IEEE Wireless Communications and Networking Conference (WCNC)*, 920-925. doi:10.1109/WCNC51071.2022.9771814

Nhi, N. T. U., & Le, T. M. (2022). A model of semantic-based image retrieval using C-tree and neighbor graph. *International Journal on Semantic Web and Information Systems*, *18*(1), 1–23. doi:10.4018/IJSWIS.295551

Ouyang, L., & Wang, H. (2019). Aerial target detection based on the improved YOLOv3 algorithm. *2019 6th International Conference on Systems and Informatics (ICSAI),* 1196-1200.

Qian, W., Li, H., & Mu, H. (2022). Circular LBP Prior-Based Enhanced GAN for Image Style Transfer. *International Journal on Semantic Web and Information Systems*, *18*(2), 1–15. doi:10.4018/IJSWIS.315601

Razakarivony, S., & Jurie, F. (2016). Vehicle detection in aerial imagery: A small target detection benchmark. *Journal of Visual Communication and Image Representation*, *3*(6), 187–203. doi:10.1016/j.jvcir.2015.11.002

Sayour, M. H., Kozhaya, S. E., & Saab, S. S. (2022). Autonomous robotic manipulation: Real-time, deep-learning approach for grasping of unknown objects. *Journal of Robotics*, *2022*, 2022. doi:10.1155/2022/2585656

Srivastava, A. M., Rotte, P. A., Jain, A., & Prakash, S. (2022). Handling data scarcity through data augmentation in training of deep neural networks for 3D data processing. *International Journal on Semantic Web and Information Systems*, *18*(1), 1–16. doi:10.4018/IJSWIS.297038

Stergiou, C. L., Psannis, K. E., & Gupta, B. B. (2021). InFeMo: Flexible big data management through a federated cloud system. *ACM Transactions on Internet Technology*, *22*(2), 1–22. doi:10.1145/3426972

Sun, W., Yan, D., Huang, J., & Sun, C. (2020). Small-scale moving target detection in the aerial image by deep inverse reinforcement learning. *Soft Computing*, *24*(6), 5897–5908. doi:10.1007/s00500-019-04404-6

Tian, H., Zheng, Y., & Jin, Z. (2020). Improved RetinaNet model for the application of small target detection in the aerial images. *IOP Conference Series: Earth and Environmental Science*, 11569-11573. doi:10.1088/1755-1315/585/1/012142

Wang, H., Li, Z., Li, Y., Gupta, B. B., & Choi, C. (2020). Visual saliency guided complex image retrieval. *Pattern Recognition Letters*, *130*, 64–72. doi:10.1016/j.patrec.2018.08.010

Wang, X., Deng, Y., & Duan, H. (2018). Edge-based target detection for unmanned aerial vehicles using competitive Bird Swarm Algorithm. *Aerospace Science and Technology*, *7*(3), 708–720. doi:10.1016/j.ast.2018.04.047

Wu, X., Hong, D., Tian, J., Chanussot, J., Li, W., & Tao, R. (2019). ORSIm detector: A novel object detection framework in optical remote sensing imagery using spatial-frequency channel features. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(7), 5146–5158. doi:10.1109/TGRS.2019.2897139

Xia, G. S., Bai, X., & Ding, J. (2018). DOTA: A large-scale dataset for object detection in aerial images. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3974-3983. doi:10.1109/CVPR.2018.00418

Xu, D., & Wu, Y. (2020). Improved YOLO-V3 with DenseNet for multi-scale remote sensing target detection. *Sensors (Basel)*, *20*(15), 4276–4284. doi:10.3390/s20154276 PMID:32751868

Yang, F., Li, W., Hu, H., Li, W., & Wang, P. (2020). Multi-scale feature integrated attention-based rotation network for object detection in VHR aerial images. *Sensors (Basel)*, *20*(6), 1686–1695. doi:10.3390/s20061686 PMID:32197365

Yang, F., Ma, B., & Wang, J. (2021). Target detection of UAV aerial image based on rotational invariant depth denoising automatic encoder. *Journal of Northwestern Polytechnical University*, *38*(6), 1345-1351.

Ying, X., Wang, Q., Li, X., Yu, M., Jiang, H., Gao, J., Liu, Z., & Yu, R. (2019). Multi-attention object detection model in remote sensing images based on multi-scale. *IEEE Access : Practical Innovations, Open Solutions*, *7*(5), 94508–94519. doi:10.1109/ACCESS.2019.2928522

Zhang, M., Pang, K., Gao, C., & Xin, M. (2020). Multi-scale aerial target detection based on densely connected inception ResNet. *IEEE Access : Practical Innovations, Open Solutions*, *8*(1), 84867–84878. doi:10.1109/ACCESS.2020.2992647

Zhang, M., Wang, C., & Yang, J. (2021). Research on engineering vehicle target detection in aerial photography environment based on YOLOX. *2021 14th international symposium on computational intelligence and design (ISCID)*, 254-256.

Zhao, Y., Jia, J., Liu, D., & Qian, Y. (2021). He-yolo: Aerial target detection based on improved YOLOv3. *International Journal of Pattern Recognition and Artificial Intelligence*, *35*(13), 2150–2159. doi:10.1142/S0218001421500361

Zheng, Z., Zhou, J., Gan, J., Luo, S., & Gao, W. (2022). Fine-Grained Image Classification Based on Cross-Attention Network. *International Journal on Semantic Web and Information Systems*, *18*(1), 1–12. doi:10.4018/IJSWIS.315747

Zhong, J., Lei, T., & Yao, G. (2017). Robust vehicle detection in aerial images based on cascaded convolutional neural networks. *Sensors (Basel)*, *17*(12), 2720–2729. doi:10.3390/s17122720 PMID:29186756

Zhou, J., Zeng, W., Xu, H., & Zhao, X. (2023). Active Temporal Knowledge Graph Alignment. *International Journal on Semantic Web and Information Systems*, *19*(1), 1–17. doi:10.4018/IJSWIS.318339

Zhu, X., Lyu, S., & Wang, X. (2021). TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Headfor Object Detection on Drone-captured Scenarios. *18th IEEE/CVF International Conference on Computer Vision*, 2778-2788.

Zou, Z., & Shi, Z. (2017). Random access memories: A new paradigm for target detection in high resolution aerial remote sensing images. *IEEE Transactions on Image Processing*, *27*(3), 1100–1111. doi:10.1109/TIP.2017.2773199 PMID:29220314

Xinyang Li is an associate researcher at Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences. He received his BE and ME degrees in communication engineering from the Jinlin University in 2007 and 2009, respectively, and his Ph.D. degree in optical engineering from the University of Chinese Academy of Sciences in 2015. He is the author of more than 10 journal papers and has written one book. His current research interests include aerial imaging and measurement technology as well as servo control technology.