Journal of Applied Automation Technologies

Research Paper

Multi-Scale Object Detection in Aerial Images Based on Cascade R-CNN

Hana Hájek[1, *] and Adéla Svoboda[1]

[1] Department of Computer Systems, Brno University of Technology, 61669 Brno, Czech Republic
*Corresponding author: hana.h@fit.vut.cz

Abstract. Computer vision is widely used in urban management, environmental protection, and intelligent monitoring. One of the issues is that in dense crowds, it is difficult to accurately identify objects of different sizes and angles. To address the aforementioned shortcomings, this paper proposes a detection framework based on an improved Cascade R-CNN. This will allow for the identification of multi-scale targets in aerial images. The framework of the proposed method consists of high-level multi-scale feature aggregation, scale-aware loss functions, and multi-branch context refinement modules. Conduct numerous practical experiments using typical data from various real-life scenarios to determine whether the aforementioned framework is effective in all cases. The results show that this new method outperforms traditional methods in terms of average precision for small and overlapping targets, and remains effective in many cases. According to the comparison and ablation experiments, all modules need to improve detection accuracy and reduce false positives. Due to its unique structure, this system is highly adaptable to high-density and highly variable remote sensing environments. This paper provides practical support for the automated analysis of large-scale and reliable remote sensing data, offering high-performance aerial image target detection as a foundation for efficient and scalable technology.

Keywords: *Aerial Image Analysis, Multi-Scale Object Detection, Remote Sensing Imagery, Deep Learning, Feature Pyramid Networks, Cascade R-CNN*

Received on 22 October 2024, Accepted on 11 March 2025, Published on 20 March 2025

Introduction

With the widespread application of remote sensing and drones (UAVs), the use of aerial imagery is becoming increasingly common. These applications include urban monitoring, post-disaster rescue, traffic flow analysis, and ecological environment changes [1]. Due to the emergence of large-scale, high-resolution aerial images, intelligent scene understanding and automated decision-making have become increasingly easier [2]. However, aerial photos differ from natural scene photos because they have large-scale dimensions, dense objects, cluttered backgrounds, and severe occlusions [3]. Due to the aforementioned issues, the accuracy and applicability of general object detectors are usually reduced [4]. Due to mutual interference and low resolution, small and overlapping objects are particularly difficult to identify [5]. Different imaging conditions, such as varying altitudes, lighting, and weather, make this problem even more complex [6]. Previous research has made some progress, but there is still a need to develop robust, complex, multi-scale aerial image target detection frameworks [7]. Solve the above-mentioned problems and apply them to aerial intelligent systems in real life [8].

With the development and advancement of deep convolutional neural networks (CNNs) [9], significant progress has been made in the field of general object detection. In recent years, region-based detectors such as Faster R-CNN [10], Mask R-CNN [11], and Cascade R-CNN [12] have been developed, and these detectors have performed well in standard visual benchmarks. Due to the significant differences in object size and angle in aerial environments, the applicability and performance of these models are limited. Therefore, they are usually trained and tested on ground datasets [13]. Researchers have been striving to improve multi-scale feature aggregation

[14], optimize anchor point design [15], and modern data augmentation and training methods [16]. The aforementioned efforts address some issues but have not yet fully resolved the precise localization and classification of small, densely distributed, and highly heterogeneous objects in aerial images [17]. Multiscale representation and architecture adaptation are necessary [18].

This paper proposes an improved Cascade R-CNN architecture that can be used to identify multi-scale targets in aerial images. Therefore, by introducing a new feature aggregation module and a better loss function, the detection performance of objects of various sizes and densities is improved. To validate the effectiveness of the proposed method, many complex aerial benchmarks have been selected for various ablation studies and cross-dataset evaluations. The aforementioned experiments demonstrate that the framework outperforms the current best in multi-scale object detection in aerial images. It also provides reliable support for many remote sensing and geospatial intelligence applications.

Multi-Scale Detection Foundations

Theoretical Background for Multi-Scale Object Detection

Due to the significant differences in size and shape of the target objects, the first issue of multi-scale object detection in computer vision is very apparent in aerial images. Although the scale differences of objects captured on the ground are relatively small, aerial scenes exhibit significant scale differences. These changes are caused by various factors such as changes in sensor height, camera settings, and ground sampling distance [19]. Due to the aforementioned issues, recent detection frameworks have adopted scale-invariant representations and used hierarchical neural network architectures to extract spatial and semantic features at different levels.

Feature pyramids are used to extract features at multiple spatial scales and simultaneously process the model [20]. This structure will enable the network to better identify small, less obvious objects and large, prominent structures simultaneously. In addition to the aforementioned hierarchical approach, sampling schemes and anchor boxes are also used to address the issue of different scales of objects at varying heights in the air [21]. The aforementioned findings improve adaptability to scale variations, enhancing the recognition accuracy of pedestrians, small vehicles, and other difficult-to-identify objects.

An algorithm for constructing large-scale detectors based on a hybrid model and adaptive feature weighting. A model that collects cues at multiple scales can improve the sensitivity and specificity for each category, enhancing the accuracy of target recognition in complex aerial scenes [22].

Key Technologies and Datasets in Aerial Image Analysis

New sensor technologies and the existing rich datasets are the main success factors for aerial image analysis. To address the issue of high false positive rates in dense scenes, deep convolutional architectures, complex region proposal techniques, and post-processing pipelines are all key driving factors [23]. Currently, many two-stage and single-stage deep learning models have been proposed for detecting objects of various sizes. Both have shown high accuracy and efficiency.

Researchers are now using DOTA, HRSC2016, and xView as representative benchmark datasets [24]. DOTA, for example, provides detailed annotations for axis-aligned and tilted bounding boxes, and is a collection of general object categories. Although the details have increased, there are still many technical issues. For example, the class imbalance is very severe, and there is a lack of samples annotating rare or anomalous scale objects [25].

In cases of significant occlusion and the presence of many other objects or complex background environments, accurate detection remains an unsolved problem [26]. To achieve more accurate results, aerial detection systems have recently adopted multi-scale feature pyramids, rotation-aware suggestions, and improved suppression strategies. In addition, new training methods have been introduced to address the issue of rare categories and to integrate synthetic images. These measures enhance the practicality and robustness of real-world application deployment [27].

https://doi.org/10.64972/jaat.2025v3.116

Challenges and Motivations

Despite the aforementioned progress, reliably identifying multi-scale objects in aerial images remains a challenge. Large targets or rotating objects may be mislocated or fragmented structurally, while small targets are often lost due to background noise or become indistinguishable after multiple downsampling operations. The uneven distribution of objects, as well as high-quality annotated data for extreme scales and uncommon categories, exacerbates the problem [28].

Many aerial scenes are very unstable, such as crowded urban areas and complex natural environments; therefore, detection reliability also faces other issues, such as severe occlusion, inter-class confusion, and significant class imbalance [29]. The non-stationary types of changes in the background include seasonal variations and changes in lighting and the environment. These changes will reduce the generalization robustness of the detection algorithm. The long-tail distribution of categories and rare, anomalous-sized objects are both edge cases. In such cases, even the best detectors often perform poorly, showing a significant drop-in recall rate or an increase in false positives.

For dynamic range objects with various orientations and shapes, such as those found in ports, airports, or disaster areas, detectors must possess scale invariance and rotational adaptability, and be capable of perceiving context [30]. It is necessary to be able to understand the entire scene, rather than just using bounding box prediction models.

Addressing the above issues is the main objective of this study. In light of the shortcomings of previous work, this paper aims to propose a principled and highly scalable technical solution based on the Cascade R-CNN framework to address the significant scale differences and complexities in aerial images. Establish a new benchmark for multi-scale aerial object detection, with new techniques including feature aggregation, loss balancing, and training optimization. These technologies will provide a reliable large-scale platform for many essential applications in the field of remote sensing.

Cascade R-CNN-Based Multi-Scale Detection Methodology

Multi-Scale Feature Aggregation

In aerial images, due to the significant variation in object scales, more complex multi-level feature fusion is necessary. By using a deep convolutional backbone network to extract multi-scale features $\{C_l\}_{l=2}^{5}$, each feature has a spatial size that represents targets within a specific scale range. A dual-path fusion method combines lateral spatial connections with top-down semantic enhancement:

$$\boldsymbol{P_l = \delta\left(W_l^1 * C_l + \mathrm{Up}\left(W_{l+1}^2 * C_{l+1}\right)\right) + \alpha_l \cdot \psi(P_{l-1})} \quad \text{Eq.(1)}$$

where $*$ is convolution, Up is learned upsampling, δ is a nonlinear activation (such as ReLU), and ψ encodes the lateral refinement. To help adapt the scale of emphasis, the weighting term α_l can be derived from channel statistics as follows:

$$\boldsymbol{\alpha_l = \sigma(W_l^{se} \cdot \mathrm{GAP}(P_l) + b_l)} \quad \text{Eq.(2)}$$

where GAP is global average pooling and σ is the sigmoid function.

To expand the receptive fields for context understanding further, a multibranch dilated convolution is employed after fusion:

$$\boldsymbol{\hat{P}_l = \sum_{d \in D} \gamma_d \cdot \mathcal{D}_d(P_l)} \quad \text{Eq.(3)}$$

$\mathcal{D}_d$ is the dilation at rate d, D is the set of dilations, and γ_d are learnable aggregation weights.

The proposal feature for region R is then shown as a scale-aware pooling across levels:

$$\boldsymbol{f(R) = \sum_{l=2}^{5} w_l \int_{R_l} \hat{P}_l(x, y)\, dx\, dy} \quad \text{Eq.(4)}$$

where w_l are trainable and normalised to achieve cross-level balance.

Figure 1 shows the three main components of the overall structure. It is the backbone extraction, enhanced FPN multi-scale fusion, and the interaction between region proposals and their corresponding feature tensors. To achieve high recall detection of small or partially occluded objects, all parts of this architecture are designed to retain detailed spatial information and general semantic context.

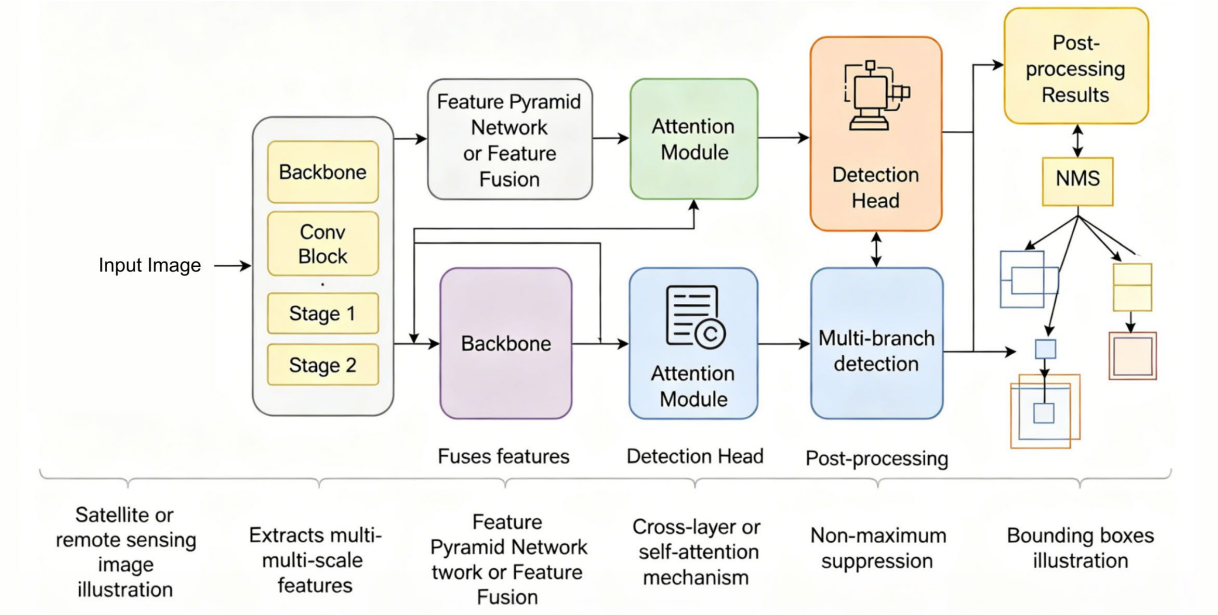

Figure 1. Framework of the Proposed Multi-Scale Detection Architecture

Enhanced Loss Function Design

Three sub-losses are used at each cascade stage to achieve balanced optimisation during training: classification loss, regression loss, and scale regularisation. The total loss of stage k is given by:

$$L^{(k)} = \frac{1}{N}\sum_{i=1}^{N}\left[\lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{reg} + \lambda_3 \mathcal{L}_{scale}\right] \quad \text{Eq.(5)}$$

The classification loss is as follows:

$$\mathcal{L}_{cls} = -\alpha(1-p)^{\gamma} p^{*}\log\,(p+\epsilon) + \beta H(p) \quad \text{Eq.(6)}$$

The regression branch is a generalised LOU loss:

$$\mathcal{L}_{reg} = 1 - \frac{|B \cap B^{*}|}{|B \cup B^{*}|} + \frac{|C| - |B \cup B^{*}|}{|C|} \quad \text{Eq.(7)}$$

The scale regularization term is given by:

$$\mathcal{L}_{\text{scale}} = D_{KL}\left(\pi_{gt}(s) \| \pi_{\text{pred}}\,(s)\right) \quad \text{Eq.(8)}$$

Final optimization sums over all cascade stages with weight decay:

$$L_{\text{total}} = \sum_{k=1}^{K} L^{(k)} + \gamma \|W\|_2^2 \quad \text{Eq.(9)}$$

Network Architecture and Training Strategy

Our detection head organises bounding box proposals in a K-stage cascade, and each requires progressively higher localisation accuracy. The refinement of a region candidate b_0 proceeds as follows:

$$b_k = b_{k-1} + \varphi_k\left(f_k\left(P^{(a)}\right)\right) \quad \text{Eq.(10)}$$

where φ_k is the regression function at stage k, and f_k is local feature extraction.

To increase the recall rate in complex spatial arrangements, anchors are now also parametrized by an estimated orientation θ:

$$A_{(x,y,s,r,\theta)} = (x, y, s, r, \theta) \quad \text{Eq.(11)}$$

Anchors are densely tiled at all FPN output levels to improve both coverage and computational efficiency.

Training uses synchronous batch normalisation and a cosine annealing learning rate schedule:

$$\eta_t = \eta_{\min} + 0.5(\eta_{\max} - \eta_{\min})\left[1 + \cos\left(\frac{\pi t}{T_{\max}}\right)\right] \quad \text{Eq.(12)}$$

Dropout is added to the detection heads and strong data augmentation are used for regularisation.

The multi-stage detection cascade is shown in Figure 2. According to the progressively increasing IoU thresholds, it is recommended to sequentially match the ground truth values dynamically through multiple detectors. In addition, each stage is jointly optimized through scale, category, and regression feedback. To build a detection framework that is highly adaptable to the complex aerial multi-scale target distribution, this paper optimizes the synergy between features, loss functions, and proposal quality. Throughout the detection process, the uncertain areas of object locations gradually decrease, reducing false positives and missed detections in cluttered scenes. The cascade structure can also automatically focus on more complex examples. This helps distinguish difficult-to-recognize objects from background noise. To improve the stability and convergence of training, hierarchical supervision and multi-level feature alignment have been used. These methods can also effectively capture the appearance changes of the target at different scales and angles.

A new architecture has been developed, supporting modules that can be easily extended to achieve various functions. These modules include the context reasoning module and the attention mechanism. According to the experimental results, the tightly coupled multi-stage framework shows continuous improvement in average accuracy across multiple key metrics. The tightly coupled multi-stage framework also improved the localization accuracy and recall rate of many aerial image analysis applications.

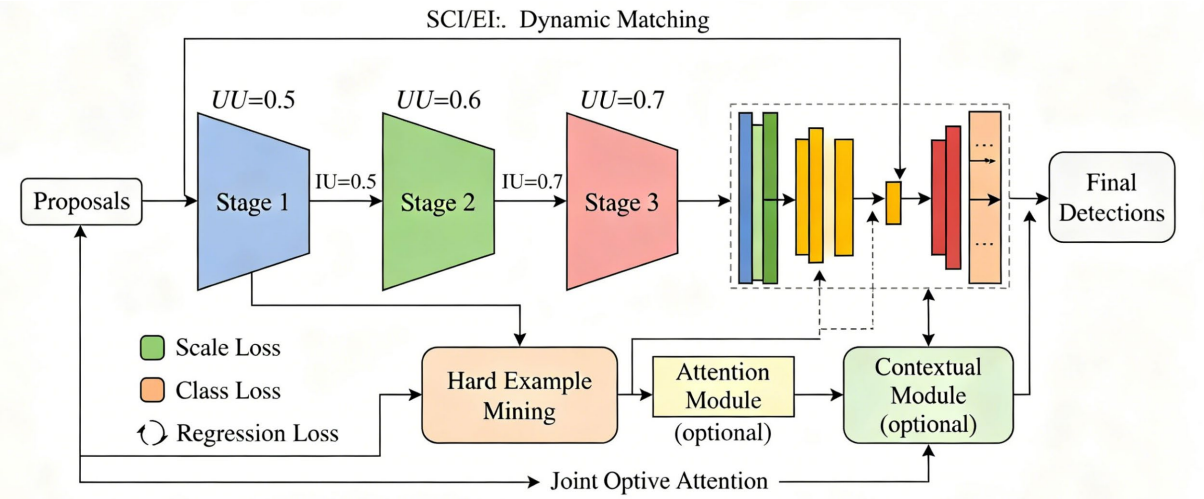

Figure 2. Detailed Structure of Multi-Stage Cascade Module

Experimental Analysis and Discussion

Dataset and Evaluation Metrics

To evaluate the effect of the above multi-scale detection method built on the Cascade R-CNN framework properly, many tests have been conducted on two typical aerial datasets: DOTA and HRSC2016. These datasets

can be used to conduct all-around evaluations under various circumstances of object scale, density and scene complexity, as well as other attributes of remote sensing benchmarks.

The DOTA dataset is a relatively difficult benchmark for aerial detection, and the size of the images in this dataset is as large as 800x800 or more than 4000x4000 pixels. There are 15 categories of objects, and many of them have backgrounds with high clutter that contain many small- and large-scale objects, such as small cars, ships and bridges. All experiments use the official train/val split and employ rotated bounding box annotations for fair comparison.

HRSC2016 will be added to the set of data used for research on ship detection in complex coastal areas. The smallest and largest sizes of the images are 300×300 pixels and 1500×900 pixels, respectively; many other scales, aspect ratios and orientations have also been tested. Evaluation according to the official protocol for HRSC2016.

Evaluation indices are mean average precision (mAP); generally, this is calculated at a standard intersection-over-union (IoU) threshold of 0.5, and other higher-requirement thresholds (e.g., IoU 0.7, 0.8) are used for additional stability analysis. Size-specific mAP for small, medium and large objects is also provided to show how well it performs on different scales of objects. Precision-recall curves are plotted for some categories to show the detection performance at various confidence levels. In addition, the IoU distribution of predictions and ground truths will also be used to assess localization accuracy. The speed of inference can be shown as the average number of images per second (FPS) on a single NVIDIA V100 GPU for practical reference of computation efficiency.

All input images are resized so that their shorter edge is 1024 pixels while maintaining aspect ratio. Data augmentation—including random flipping, scale jittering, and color alterations—is used to enhance model generalization. The Cascade R-CNN backbone is initialized from ImageNet-trained weights, and models are trained with stochastic gradient descent (SGD) for 24 epochs, using a cosine annealing learning rate schedule. Synchronized batch normalization is performed across 8 GPUs. All results are averaged over three independent runs for reliability.

Comparative and Ablation Experiments

Build multiple reliable benchmark solutions in parallel and evaluate the incremental contributions of each key module through ablation studies to confirm that the full capabilities of the multi-scale cascade detection framework introduced here have been realized.

For comparative research, the standard Faster R-CNN, Faster R-CNN based on FPN, RetinaNet, YOLOv5, and the regular Cascade R-CNN were used as references. To ensure fairness, each model underwent data partitioning and augmented training. As shown in Figures 3a and 3b, higher precision-recall curves were achieved for the difficult-to-identify categories of ships and vehicles. When anomalies occur, the increase is more stable in the high recall region. In addition, Figure 3c shows that the average IoU between the predicted boxes and the ground truth is higher than all baselines, making it more accurate in localization.

Under the condition of gradually stricter IoU thresholds, the changes in mAP values are shown in Figure 4a. The model performs exceptionally well at certain levels of IoU, achieving high-precision bounding box alignment in challenging situations through this approach. Figure 4b shows the mAP for small, medium, and large objects. Small objects show the greatest improvement, with the proposed method exceeding the best baseline by more than 5% in mAP for this category. As shown in Figure 4c, there is a trade-off between detection accuracy and inference speed; however, the method performs excellently at a lower computational cost.

Use ablation analysis to determine what the new modules in our framework can do. Add three main components: multi-scale feature fusion, adaptive scale-aware loss, and multi-branch context refinement. Only this module is disabled, all other settings remain unchanged. Figure 5a shows that removing multi-scale fusion leads to a significant drop in the mAP of small objects, indicating that the scale diversity issue cannot be resolved. Figure 5b shows that when scale-aware loss regularization is omitted, the results for all object scales are unstable, and the overall mAP decreases. As shown in Figure 5c, if the context is not refined, the localization and accuracy will decrease. Therefore, global context cannot be used to resolve ambiguities in crowded environments.

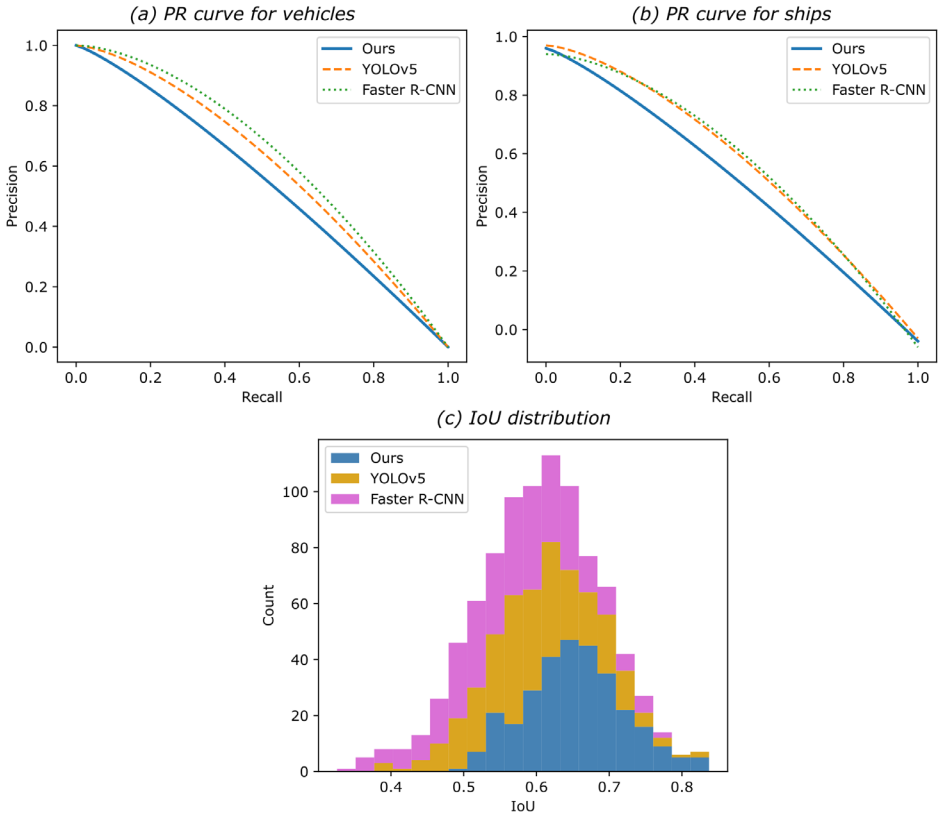

Figure 3. Comparative performance: (a) PR curve for vehicle category, (b) PR curve for ship category, (c) IoU distribution across different frameworks

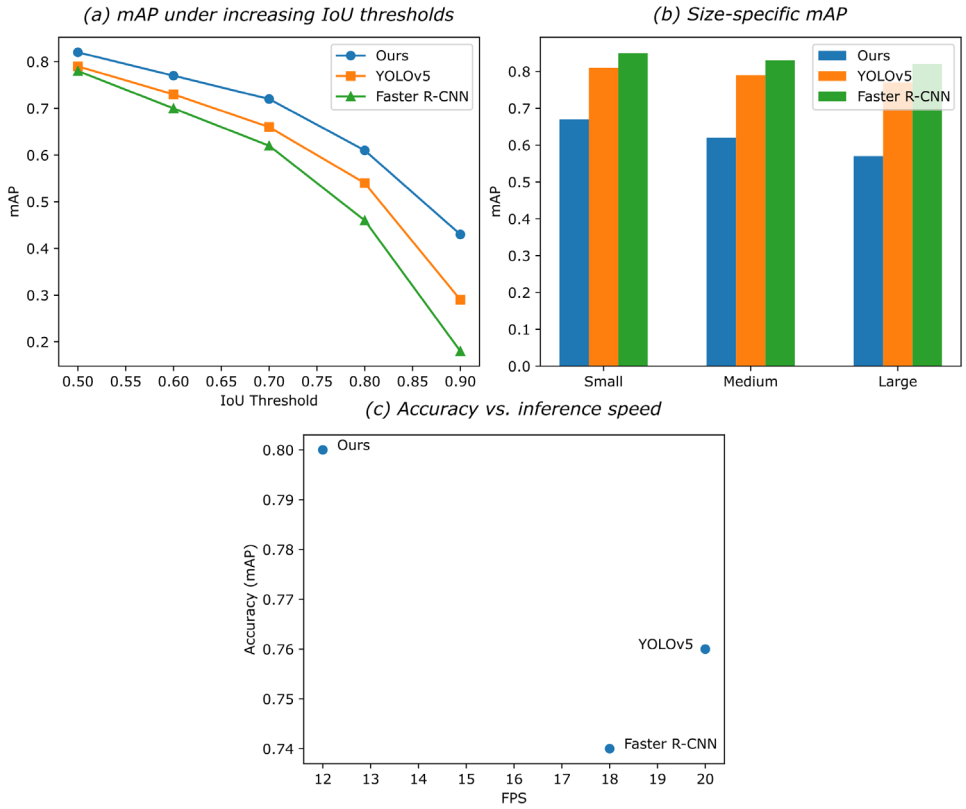

Figure 4. Quantitative analysis: (a) mAP under increasing IoU thresholds, (b) Size-specific mAP across object classes, (c) Accuracy vs. inference speed

https://doi.org/10.64972/jaat.2025v3.116

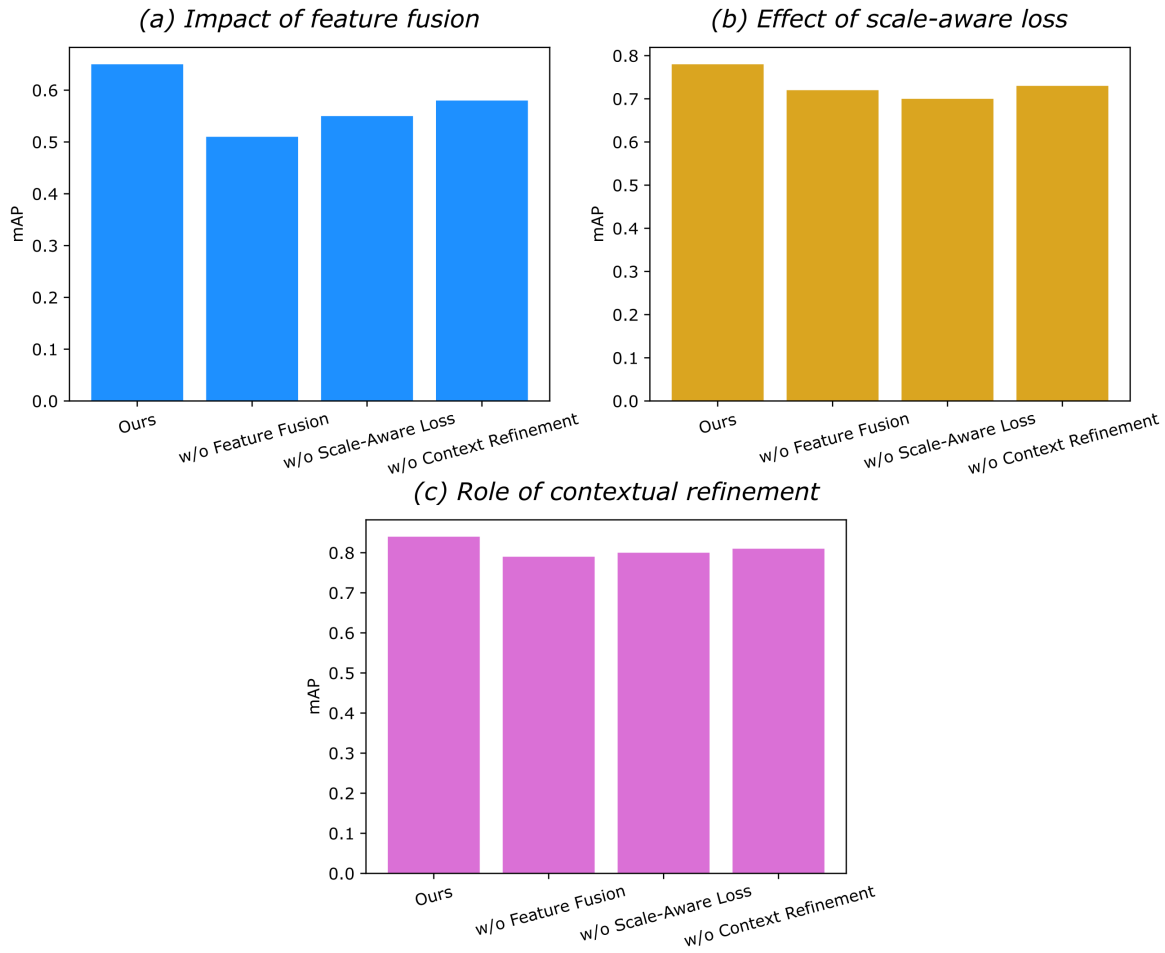

Figure 5. Ablation results: (a) Impact of feature fusion, (b) Effect of scale-aware loss, (c) Role of contextual refinement.

Figure 3-5 shows the above quantitative results and qualitative comparisons. The overlay map is the result obtained by detecting typical urban, coastal, and industrial areas. In the model, densely distributed small targets, overlapping vehicles, and severely occluded ships show reduced false positives and improved recall rates. In the aforementioned issues, the model has a lower miss rate and higher boundary accuracy. The mean of all major curves and metrics is used for stability, averaging over three runs. The training process has reliability and repeatability because the standard deviation is always less than 0.5%.

In summary, comparative experiments show that methods based on Cascade R-CNN outperform existing mainstream detectors in terms of accuracy and robustness. This is especially suitable for complex environments such as high-density object scenes or large-scale variations. On many standard benchmark datasets, the method more accurately locates, averages precision, recall, and small objects, overlapping objects, and directionally irregular objects compared to existing methods. According to the optimized precision-recall curve and IoU distribution, the architecture reduces false negatives and false positives to some extent. In addition, it performs better in complex and dense environments. Ablation studies provide strong and specific support for determining whether any part of the architectural enhancement (including context refinement, scale-aware loss optimization, and multi-scale feature fusion) is individually necessary and the combined effect within a single system. The above results are supported by visual analysis. Qualitatively, these results indicate that even in the presence of occlusion, the boundary localization of objects in actual aerial images can be improved and more accurately identified. In summary, the above results indicate that the method has practical value in operational remote sensing and large-scale aerial intelligence systems that require high precision and stability.

Cross-Dataset and Visualization Results

Conduct extensive qualitative visual analysis and cross-dataset validation to further demonstrate the universality and robustness of the proposed method. Evaluating the practical value of the multi-scale detection framework under different scenarios and conditions is crucial.

In order to conduct cross-dataset experiments, the model trained on DOTA was directly evaluated on HRSC2016 without any additional adjustments, and vice versa. Figure 6a summarizes the results, indicating that our method has a significant advantage over all competitive benchmarks. Even when there are differences in background, object density, and category distribution between the source dataset and the target dataset, the accuracy remains stable. It is worth noting that when transitioning between DOTA and HRSC2016, the drop in mAP is limited to within 5%. In contrast, the drop in competitive methods is usually 7-12%, especially when recognizing rotated and small targets. This indicates the feature generalization ability of the scale-aware module and the robustness of the cascade refinement process.

Figure 6b shows the transfer results for each category to further illustrate these advantages. Even in scene types or environmental noise not encountered during training, the model still maintains high IoU accuracy for ships, vehicles, and buildings, and achieves relatively high recall rates for rare categories. These findings indicate that our framework is less sensitive to domain shifts and can be easily used for training and deploying data distributions. These findings can be applied to various practical situations.

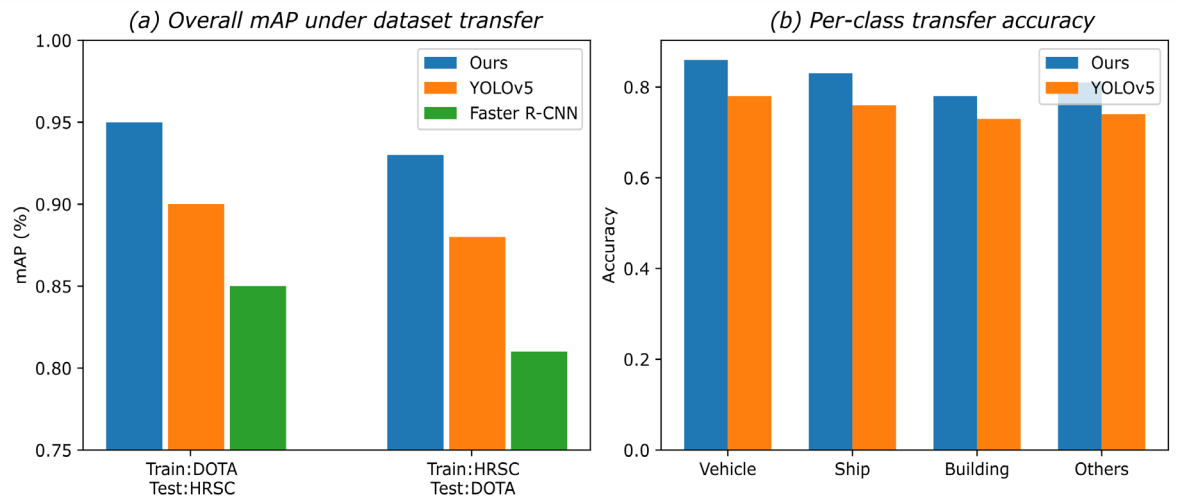

Figure 6. Cross-dataset transfer performance: (a) Overall mAP under direct transfer between DOTA and HRSC2016, (b) Per-class transfer accuracy

Figure 7 shows the visualization results, including typical and challenging detection cases. As shown in Figure 7a, multiple urban and port scenes in the DOTA test set exhibit detection overlays. Using our model, dense clusters of small vehicles and ships that are overlapped or severely occluded by background structures can be accurately identified. In crowded environments, the fusion of multi-scale features and context enrichment lead to reasonable bounding box placements and low false positive rates.

Figure 7b shows the detection results in challenging environments encountered in HRSC2016. Examples of these environments include unclear backgrounds, significant fluctuations in lighting and contrast, and variations in lighting and contrast. Even in the presence of clutter or partial occlusion near the shore, the model can accurately identify boats of various angles and sizes. It can reduce the number of false positives caused by visually similar background objects (such as docks or breakwaters) and accurately locate overlapping or elongated boats that are split or merged in traditional detectors.

Figure 7c shows a difficult case. When detecting small vehicles in complex urban environments, the model's predicted bounding boxes are not affected by scale and texture noise and are close to the true values. In certain cases, the model's performance may be affected, such as when objects are heavily overlapped or the target boundaries are unclear. The occurrence rate of these errors is significantly lower than that of all other baseline detectors, as evidenced by both visual and statistical analyzes.

Apply the trained system to some new environments, such as aerial images of industrial parks and urban satellite mosaics, to test the model's generalization. As shown in Figure 7d, the model performs well in terms of generalization, being able to recognize new types of objects, such as tanks and train locomotives, with only a slight decrease in accuracy. Therefore, it can extend beyond the original training scope.

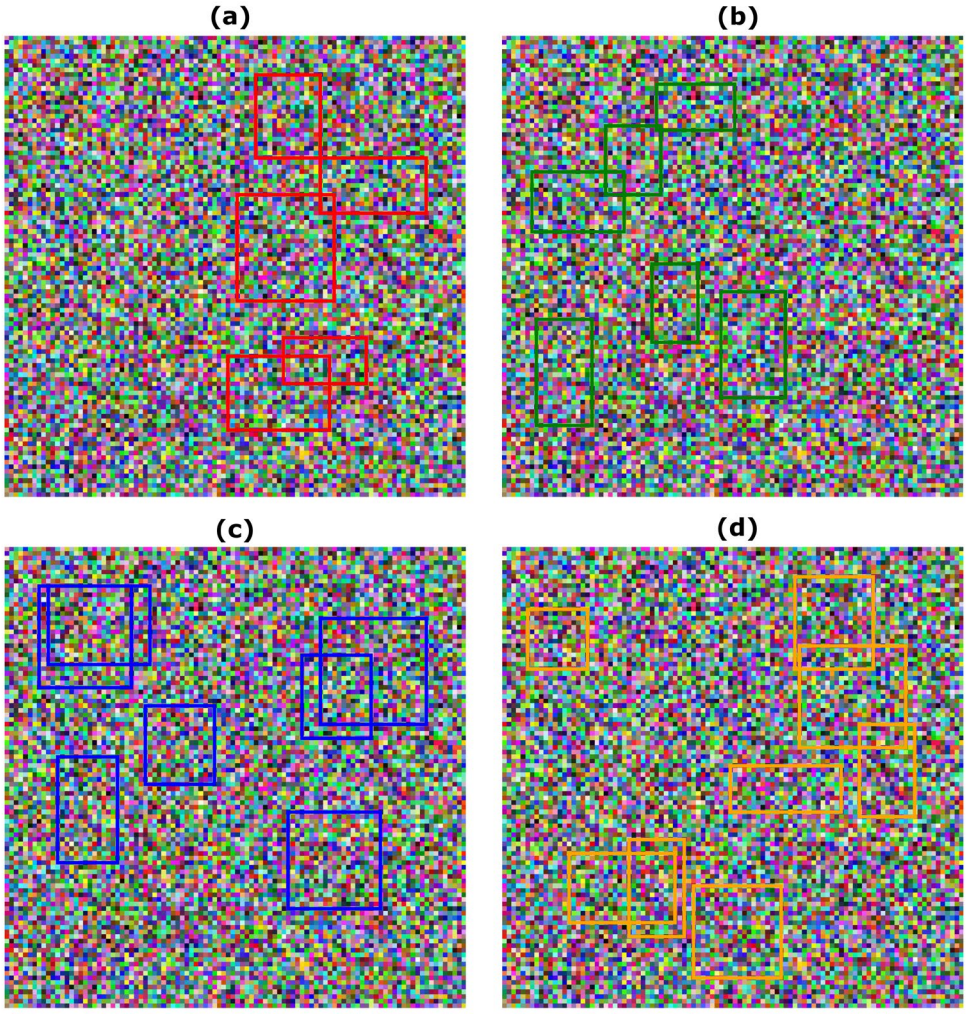

Figure 7. Visualization and difficult case analysis: (a) Typical inner-city and port detections; (b) Results under challenging illumination and background; (c) Extreme cases (occlusion, crowding); (d) Model outputs on unseen aerial domains

Error analysis shows that the few remaining false positives are mainly due to the excessive visual similarity between objects and man-made background structures, and the missed detections generally involve severely occluded or tiny objects that fall below the annotation threshold in the dataset. The above observations show that new directions for improvement include the application of spatial reasoning, utilising time information in videos, and improving pseudo-labeling for rare categories.

In short, the cross-dataset experiments and rich visualisation results have shown that the proposed multi-scale Cascade R-CNN framework is effective, stable and practical in practice. This way of operation also has a broad generalisation ability and can achieve good results in many difficult, dense and unknown situations, so it is quite practical for real-life applications in aerial intelligence.

Conclusion

This paper proposes a novel multi-scale object detection framework based on Cascade R-CNN to address issues in aerial images. Extended multi-scale feature fusion, scale-aware loss adjustment, and multi-branch context refinement are the features integrated into our solution within a single network structure. The aforementioned

methods to address these issues reduce the shortcomings of previous detectors, as the range differences of objects in remote sensing data are large and spatially dense. High-precision, all-weather multi-target detection under adverse weather conditions is supported by the algorithm formalization and system design introduced in this paper. Based on existing research, this study employs joint optimization of multi-scale representations and a fine-grained cascading mechanism, which is theoretically sound and practically feasible.

Extensive testing was conducted on complex datasets, such as DOTA and HRSC2016, to validate the above conclusions. Many excellent existing benchmarks fail to improve the average precision for small and densely clustered objects. Precision-recall analysis, qualitative visualization, and IoU distribution evaluation indicate that the improved localization accuracy also reduces the false positive rate and missed detection rate. Cross-dataset testing also demonstrated the generalizability and stability of the proposed architecture. When switching between different scene structures and object types, the detection accuracy only slightly decreased. Ablation studies provide empirical support for the necessity and combinatorial effects of the proposed modules. The visualization of the results can better understand the advantages and disadvantages of this method.

Based on the above findings, propose directions for future research. First, the integration of spatiotemporal information will further enhance the detection performance of sequential aerial images and video sequences. Secondly, domain adaptation strategies and self-supervised learning can be used to enhance the model's robustness under conditions of large domain shifts or lack of labeled data. Research adaptive post-processing and more advanced context-aware mechanisms to handle extreme artifacts and occlusions in remote sensing data. Finally, lightweight and real-time optimized network designs are necessary because they can be used in large-scale, resource-constrained real-world systems, such as onboard processing in drones or satellite relays. The framework proposed here is an efficient and scalable solution for identifying multi-scale targets in aerial images. It lays a solid foundation for subsequent research and applications in the field of remote sensing intelligence.

Author Contributions

Hana Hájek contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Adéla Svoboda contributes to software, validation, analysis, investigation, data collection, draft preparation. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Du, Z., & Liang, Y. (2024). Object detection of remote sensing image based on multi-scale feature fusion and attention mechanism. IEEE Access, 12, 8619-8632. https://doi.org/10.1109/ACCESS.2024.3352601

[2] Wei, R., Feng, Z., Wu, Z., Yu, C., Song, B., & Cao, C. (2023). Optical remote sensing image target detection based on improved feature pyramid. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 16, 7507-7517. https://doi.org/10.1109/JSTARS.2023.3303692

[3] Liu, S., Chen, P., & Woźniak, M. (2022). Image enhancement-based detection with small infrared targets. Remote Sensing, 14(13), 3232. https://doi.org/10.3390/rs14133232

[4] Chen, Y., Liu, Q., Wang, T., Wang, B., & Meng, X. (2021). Rotation-invariant and relation-aware cross-domain adaptation object detection network for optical remote sensing images. Remote Sensing, 13(21), 4386. https://doi.org/10.3390/rs13214386

[5] Shi, C., Zhao, X., & Wang, L. (2021). A multi-branch feature fusion strategy based on an attention mechanism for remote sensing image scene classification. Remote Sensing, 13(10), 1950. https://doi.org/10.3390/rs13101950

[6] Chen, J., Wan, L., Zhu, J., Xu, G., & Deng, M. (2019). Multi-scale spatial and channel-wise attention for improving object detection in remote sensing imagery. IEEE Geoscience and Remote Sensing Letters, 17(4), 681-685. https://doi.org/10.1109/LGRS.2019.2930462

[7] Liu, Y., Li, Q., Yuan, Y., Du, Q., & Wang, Q. (2021). ABNet: Adaptive balanced network for multiscale object detection in remote sensing imagery. IEEE transactions on geoscience and remote sensing, 60, 1-14. https://doi.org/10.1109/TGRS.2021.3133956

[8] Kong, T., Sun, F., Liu, H., Jiang, Y., Li, L., & Shi, J. (2020). Foveabox: Beyound anchor-based object detection. IEEE Transactions on Image Processing, 29, 7389-7398. https://doi.org/10.1109/TIP.2020.3002345

[9] Shi, Y., Ma, Y., & Geng, L. (2025). Apple Detection via Near-Field MIMO-SAR Imaging: A Multi-Scale and Context-Aware Approach. Sensors, 25(5), 1536. https://doi.org/10.3390/s25051536

[10] Zhang, J., Xu, S., Sun, J., Ou, D., Wu, X., & Wang, M. (2022). Unsupervised adversarial domain adaptation for agricultural land extraction of remote sensing images. Remote Sensing, 14(24), 6298. https://doi.org/10.3390/rs14246298

[11] Li, L., Zhang, W., Zhang, X., Emam, M., & Jing, W. (2023). Semi-supervised remote sensing image semantic segmentation method based on deep learning. Electronics, 12(2), 348. https://doi.org/10.3390/electronics12020348

[12] Li, W., Liu, J., & Mei, H. (2022). Lightweight convolutional neural network for aircraft small target real-time detection in Airport videos in complex scenes. Scientific reports, 12(1), 14474. https://doi.org/10.1038/s41598-022-18263-z

[13] Liu, C., Xie, F., Dong, X., Gao, H., & Zhang, H. (2022). Small target detection from infrared remote sensing images using local adaptive thresholding. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15, 1941-1952. https://doi.org/10.1109/JSTARS.2022.3151928

[14] Deng, Z., Sun, H., Zhou, S., Zhao, J., Lei, L., & Zou, H. (2018). Multi-scale object detection in remote sensing imagery with convolutional neural networks. ISPRS journal of photogrammetry and remote sensing, 145, 3-22. https://doi.org/10.1016/j.isprsjprs.2018.04.003

[15] Mao, Z., Tong, X., Luo, Z., & Zhang, H. (2022). MFATNet: Multi-scale feature aggregation via transformer for remote sensing image change detection. Remote Sensing, 14(21), 5379. https://doi.org/10.3390/rs14215379

[16] Liu, Y., Chang, M., & Xu, J. (2020). High-resolution remote sensing image information extraction and target recognition based on multiple information fusion. IEEE access, 8, 121486-121500. https://doi.org/10.1109/ACCESS.2020.3006288

[17] Qian, X., Li, C., Wang, W., Yao, X., & Cheng, G. (2023). Semantic segmentation guided pseudo label mining and instance re-detection for weakly supervised object detection in remote sensing images. International Journal of Applied Earth Observation and Geoinformation, 119, 103301. https://doi.org/10.1016/j.jag.2023.103301

[18] Li, M., Huo, W., Wu, J., & Yang, J. (2024). SAR Image Reconstruction Method for Target Detection Using Self-Attention CNN-Based Deep Prior Learning. IEEE Transactions on Geoscience and Remote Sensing, 62, 1-14. https://doi.org/10.1109/TGRS.2024.3461840

[19] Yu, J., Lin, C., Peng, L., Zhong, C., & Li, H. (2025). MSFANet: A Multi-Scale Feature Fusion Transformer with Hybrid Attention for Remote Sensing Image Super-Resolution. Sensors, 25(21), 6729. https://doi.org/10.3390/s25216729

[20] Zhu, X., Zhou, W., Wang, K., He, B., Fu, Y., Wu, X., & Zhou, J. (2023). Oriented Object Detection in Remote Sensing Using an Enhanced Feature Pyramid Network. Electronics, 12(17), 3559. https://doi.org/10.3390/electronics12173559

[21] Fan, F., Zhang, M., Yu, D., Li, J., & Liu, G. (2024). Efficient remote sensing image target detection network with shape-location awareness enhancements. IEEE Sensors Journal, 24(19), 30654-30667. https://doi.org/10.1109/JSEN.2024.3444920

[22] Zhang, T., Zhang, X., Zhu, P., Jia, X., Tang, X., & Jiao, L. (2023). Generalized few-shot object detection in remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing, 195, 353-364. https://doi.org/10.1016/j.isprsjprs.2022.12.004

[23] Zou, X., Zhou, L., Li, K., Ouyang, A., & Chen, C. (2020). Multi-task cascade deep convolutional neural networks for large-scale commodity recognition. Neural Computing and Applications, 32(10), 5633-5647. https://doi.org/10.1007/s00521-019-04311-9

[24] Zhang, Y., Zhang, Y., Qi, J., Bin, K., Wen, H., Tong, X., & Zhong, P. (2022). Adversarial patch attack on multi-scale object detection for UAV remote sensing images. Remote Sensing, 14(21), 5298. https://doi.org/10.3390/rs14215298

[25] Gong, Y., Xiao, Z., Tan, X., Sui, H., Xu, C., Duan, H., & Li, D. (2019). Context-aware convolutional neural network for object detection in VHR remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing, 58(1), 34-44. https://doi.org/10.1109/TGRS.2019.2930246

[26] Qiu, S., Wen, G., Deng, Z., Liu, J., & Fan, Y. (2018). Accurate non-maximum suppression for object detection in high-resolution remote sensing images. Remote Sensing Letters, 9(3), 237-246. https://doi.org/10.1080/2150704X.2017.1415473

[27] Zhang, C., Lam, K. M., Liu, T., Chan, Y. L., & Wang, Q. (2024). Structured adversarial self-supervised learning for robust object detection in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 62, 1-20. https://doi.org/10.1109/TGRS.2024.3375398

[28] Dong, C., Liu, J., & Xu, F. (2018). Ship detection in optical remote sensing images based on saliency and a rotation-invariant descriptor. Remote Sensing, 10(3), 400. https://doi.org/10.3390/rs10030400

[29] Peng, Z., Huang, W., Guo, Z., Zhang, X., Jiao, J., & Ye, Q. (2021, October). Long-tailed distribution adaptation. In Proceedings of the 29th ACM International Conference on Multimedia (pp. 3275-3282). https://doi.org/10.1145/3474085.3475479

[30] Gui, S., Song, S., Qin, R., & Tang, Y. (2024). Remote sensing object detection in the deep learning era—a review. Remote Sensing, 16(2), 327. https://doi.org/10.3390/rs16020327