Proceedings of the 7th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2025.24679

UAV Image Object Detection Based on Attention Mechanism and Dilated Convolution

Shijie Lyu

Georgia Institute of Technology, Atlanta, USA
slyu41@gatech.edu

Abstract. Existing algorithms for unmanned aerial vehicle (UAV) image object detection often face challenges such as low detection accuracy for small objects and missed detections of multi-scale objects. To address these issues, this paper proposes a UAV image object detection algorithm that integrates a channel attention mechanism with parallel-structured dilated convolution feature fusion. To enhance the algorithm's feature representation capabilities in terms of channel attention and receptive field, the ResNet50 backbone is redesigned by incorporating the Squeeze-and-Excitation Network (SENet) and a Parallel-Structured Dilated Convolution Feature Fusion Network (PSDCFFN). Additionally, Region of Interest (ROI) Align is employed, and the Region Proposal Network (RPN) anchor sizes are optimized using K-Means clustering to minimize coordinate deviations during object regression. Experimental results demonstrate that the proposed algorithm significantly improves object detection accuracy in UAV images. On the RSOD-Dataset and a custom UAV image dataset, the mean Average Precision (mAP) reaches 92.52% and 98.07%, respectively.

Keywords: UAV Image, Faster R-CNN, Attention Mechanism, Dilated Convolution, Feature Fusion, Object Detection

1. Introduction

Unmanned aerial vehicles (UAVs) equipped with imaging devices have become critical tools in various applications, including environmental monitoring, urban planning, and military reconnaissance, due to their flexibility and ability to capture high-resolution aerial images [1]. However, UAV images present unique challenges for object detection, such as wide variations in object scales, small object sizes, complex and cluttered backgrounds, and a high number of objects [2]. These characteristics often lead to low detection accuracy and missed detections, particularly for small and multi-scale objects, in traditional object detection algorithms [3, 4].

In recent years, convolutional neural networks (CNNs) have significantly advanced object detection tasks, offering superior speed and accuracy compared to traditional methods [5]. Deep learning-based object detection approaches are broadly categorized into two types: region proposal-based methods, such as R-CNN [6], Fast R-CNN [7], and Faster R-CNN [8], and regression-based methods, such as You Only Look Once (YOLO) [8] and Single Shot MultiBox Detector (SSD) [9].

DOI: 10.54254/2755-2721/2025.24679

Region proposal-based methods generally outperform regression-based methods in terms of detection accuracy, making them suitable for complex UAV image scenarios.

To address the specific challenges of UAV image object detection, researchers have proposed various strategies. For instance, feature fusion mechanisms have been introduced to combine low-level visual features with high-level semantic features to enhance multi-scale feature representation. However, such approaches often increase model complexity and computational cost, slowing detection speed. Other studies have modified YOLOv2 to fuse features extracted from input images of different scales, improving detection accuracy for vehicle targets in UAV images, but at the cost of increased computational complexity. For small object detection, methods such as enhancing low-level features and increasing feature map resolution have been explored [10]. For example, replacing VGG16 with a lightweight network in SSD reduced model parameters but struggled with objects exhibiting wide scale variations [10]. Similarly, enhancements to Faster R-CNN with Flat-FPN and soft-NMS improved small object detection but introduced significant computational overhead and information loss due to multiple downsampling operations. Multi-scale pooling and deconvolution have also been employed to improve small object detection, though they increase the number of region proposals. To tackle complex backgrounds, attention mechanisms have been incorporated to leverage inter-object correlations, yet their effectiveness remains limited for multi-scale objects [5].

To overcome these limitations, this paper proposes a novel UAV image object detection algorithm based on Faster R-CNN, integrating a channel attention mechanism and parallel-structured dilated convolution feature fusion. The proposed method enhances feature extraction by incorporating SENet [11] and a custom-designed Parallel-Structured Dilated Convolution Feature Fusion Network (PSDCFFN) into the ResNet50 backbone. Additionally, ROI Align is used to reduce localization errors, and RPN anchor sizes are optimized via K-Means clustering to better adapt to UAV image characteristics. Experimental results validate the effectiveness of the proposed algorithm in improving detection accuracy for multi-scale and small objects in UAV images.

2. Preliminary knowledge

2.1. Faster R-CNN

Faster R-CNN [8] is a two-stage object detection framework that integrates a Region Proposal Network (RPN) with a detection network. It employs VGG16 as the default feature extraction backbone and uses a multi-task loss function for RPN, combining classification and regression losses. The loss function is defined as follows:

$$L\left(\left\{p_i\right\},\left\{t_i\right\}\right)=\frac{1}{N_{cls}}\sum_i L_{cls}\left(p_i,p_i^*\right)+\lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}\left(t_i,t_i^*\right)$$

where i denotes the index of an anchor, p_i is the predicted probability of the i-th anchor being an object, p_i^* is the ground-truth label, t_i represents the predicted bounding box offsets, t_i^* is the ground-truth offsets, N_{cls} and N_{reg} are the number of classification and regression samples, respectively, and λ is a balancing parameter. The classification loss L_{cls} is a log-loss function, and the regression loss L_{reg} uses the smooth L_1 loss, defined as:

DOI: 10.54254/2755-2721/2025.24679

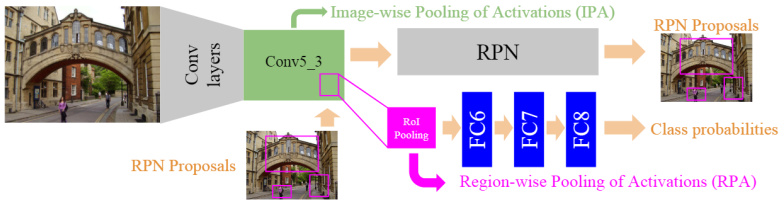

Figure 1: Image- and region-wise descriptor pooling from the Faster R-CNN architecture

2.2. Attention mechanisms

The attention mechanism enables models to focus on relevant features by assigning importance weights. The Squeeze-and-Excitation Network (SENet) [11] is a channel attention mechanism that explicitly models interdependencies between feature channels. SENet employs a "feature recalibration" strategy, learning the importance of each channel to enhance useful features and suppress irrelevant ones [12]. The SENet architecture consists of three main operations, as illustrated in Figure 2 of the original paper.

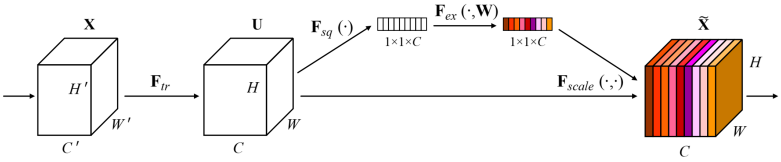

Figure 2: Diagram of a Squeeze-and-Excitation building block [11]

3. Camdc-Faster RCNN algorithm

The proposed CAMDC-Faster RCNN algorithm, builds upon the Faster R-CNN framework by integrating the Squeeze-and-Excitation Network (SENet) channel attention mechanism and the characteristics of dilated convolution. A novel feature extraction network, AMDC-ResNet50, is designed to enhance the capability to extract features from multi-scale and small objects in unmanned aerial vehicle (UAV) images. Additionally, Region of Interest (ROI) Align is employed to replace ROI Pooling, reducing positional errors during object regression. The Region Proposal Network (RPN) anchor sizes are redesigned based on K-Means clustering tailored to the characteristics of UAV image targets.

3.1. Feature extraction network: AMDC-ResNet50

The Faster R-CNN algorithm traditionally uses VGG16 as its feature extraction backbone, which suffers from high parameter counts and limited capability to detect multi-scale and small objects. Furthermore, standard convolution operations sum the results across all channels without considering inter-channel relationships [8]. To address these limitations, ResNet50, proposed by He et al. [12], is adopted as the baseline due to its deeper architecture, fewer parameters compared to VGG16, and shortcut connections that mitigate the vanishing gradient problem in deep networks.

The proposed AMDC-ResNet50, enhances ResNet50 by incorporating the SENet channel attention mechanism and dilated convolution. SENet is integrated into the initial layers of ResNet50 to recalibrate feature channels, allocating computational resources to the most informative channels [11]. After the Conv4 layer, a custom-designed Parallel-Structured Dilated Convolution Feature

DOI: 10.54254/2755-2721/2025.24679

Fusion Network (PSDCFFN) is introduced to improve the network's ability to represent multi-scale and small objects by expanding the receptive field and capturing diverse contextual information.

3.2. Parallel-structured dilated convolution feature fusion network: PSDCFFN

UAV image targets exhibit multi-scale characteristics and include small objects, making the size of the convolutional kernel critical, as it determines the local receptive field. A receptive field that is too small may fail to capture complete semantic information, while an overly large receptive field may include excessive background noise, hindering small object detection [12]. Dilated convolution addresses this by introducing gaps in the convolutional kernel, expanding the receptive field without increasing the number of parameters, thus capturing multi-scale contextual information beneficial for detecting multi-scale and small objects.

Inspired by TridentNet [12], which uses multi-branch dilated convolution but lacks feature fusion across branches, this paper proposes the PSDCFFN. Integrated into ResNet50, PSDCFFN employs three parallel paths to process features. First, Batch Normalization ensures data follows a normal distribution, facilitating network convergence. Then, each path applies dilated convolution with distinct dilation rates (R=1,2,5) to extract features at different scales. Finally, a hybrid feature fusion strategy combines these features at both pixel and channel levels through element-wise addition and channel concatenation, enhancing the network's ability to represent multi-scale contextual information. To mitigate the gridding effect in dilated convolution, the hybrid dilated convolution (HDC) structure [13] is adopted, ensuring effective coverage of multi-scale receptive fields.

3.3. Improved localization and anchor optimization

In Faster R-CNN, the ROI Pooling process introduces quantization errors by discretizing region boundaries and feature map bins, leading to positional inaccuracies in object regression [8]. To address this, ROI Align is employed, which avoids quantization by using bilinear interpolation to compute precise feature values, thereby reducing localization errors.

Additionally, the default RPN anchor sizes in Faster R-CNN are not optimized for UAV images, which contain objects with diverse scales and aspect ratios. To adapt to these characteristics, K-Means clustering is applied to analyze the size distribution of objects in the UAV image dataset. The clustering results, with k=9, yield anchor scales of 32×32, 64×64, 128×128, and 256×256, and aspect ratios of 1:2, 3:2, and 2:1, with a base stride of 8, improving the alignment of anchors with UAV image targets.

4. Experiments

4.1. Datasets

Two datasets were used: the RSOD-Dataset and a custom UAV image dataset. The RSOD-Dataset, a public aerial image dataset from Wuhan University, contains 976 images with 6,950 object instances across four categories: aircraft (4,993 instances), oil tanks (1,586), overpasses (180), and playgrounds (191). It features challenging characteristics such as severe background interference, variable object scales, and small object sizes. The dataset was split into 780 training images and 196 testing images (8:2 ratio). The UAV image dataset, collected from the internet and UAV flights, comprises 1,458 images captured at a height of 578 meters, including ground targets such as pedestrians, motor vehicles (primarily cars), and non-motorized vehicles (e.g., motorcycles, electric

DOI: 10.54254/2755-2721/2025.24679

scooters, bicycles). It was divided into 1,166 training images and 292 testing images (8:2 ratio). Sample images from the UAV dataset.

4.2. Model training

To prevent overfitting due to the limited dataset size, transfer learning was employed. The model was initialized with weights pre-trained on the VOC2007 dataset and fine-tuned on the RSOD-Dataset and UAV image dataset. Training spanned 100 epochs, with a learning rate of 0.0001 for the first 5 epochs and 0.00001 for the remaining 95 epochs, and a weight decay of 0.0005.

4.3. Experimental results and analysis

The performance of the proposed CAMDC-Faster RCNN algorithm was evaluated using metrics including Average Precision (AP), mean Average Precision (mAP), F1 score, precision (P), and recall (R), calculated as follows:

$$mAP = \frac{1}{k}\sum_{i=1}^{k} AP_i$$

$$P_{\text{precision}} = \frac{T_P}{T_P+F_P} \times 100\%$$

$$R_{\text{recall}} = \frac{T_P}{T_P+F_N} \times 100\%$$

$$F_1 = \frac{2\times P_{\text{precision}} \times R_{\text{recall}}}{P_{\text{precision}}+R_{\text{recall}}}$$

where T_P, F_P and F_N denote true positives, false positives, and false negatives, respectively.

The proposed CAMDC-Faster RCNN was compared against several baseline algorithms on both datasets. Results on the RSOD-Dataset show that CAMDC-Faster RCNN achieved an mAP of 92.52%, surpassing Faster R-CNN (88.96%), Faster RCNN + ResNet50 (90.34%), and other variants. Notably, for the aircraft category, which includes multi-scale and small objects, the AP improved by 4.95 percentage points compared to Faster R-CNN. On the UAV image dataset the mAP reached 98.07%, with a 9.99 percentage point improvement in AP for the pedestrian category compared to Faster R-CNN. The parameter count of CAMDC-Faster RCNN (167.67M) is also significantly lower than that of Faster R-CNN (521.68M), indicating improved efficiency.

To assess the effectiveness of SENet, experiments were conducted by adding SENet to different layers of the ResNet50 feature extraction network in Faster RCNN + ResNet50 on the RSOD-Dataset. Adding SENet to Conv2 layers improved the mAP to 91.04%, with marginal gains in F1 score and recall, demonstrating that channel attention enhances feature representation. However, adding SENet to all layers (Conv2–Conv4) slightly reduced the mAP to 90.03%, suggesting that excessive attention mechanisms may introduce noise or overfitting.

DOI: 10.54254/2755-2721/2025.24679

Table 1: The impact of adding the PSDCFFN at different positions on algorithm performance

Conv 2	Conv 3	Conv 4	AP/%				mAP/%	F1	Recall Rate/%	Precision Rate/%	Parameter Count/MB
			1	2	3	4					
			72.96	95.87	95.31	99.99	91.04	0.893	91.78	86.95	108.31
√			75.53	95.33	95.14	99.99	91.50	0.880	92.41	83.93	112.05
	√		74.46	95.26	93.84	98.60	90.54	0.883	92.28	84.61	126.94
		√	76.78	96.53	95.66	99.52	92.12	0.893	92.89	86.02	167.67
√	√		66.14	88.49	89.45	99.05	85.78	0.819	87.46	76.92	130.68
√		√	71.98	93.37	94.44	99.11	89.72	0.861	91.08	81.67	171.42
	√	√	72.22	90.64	93.40	99.52	88.95	0.892	90.39	87.96	182.54
√	√	√	57.49	89.60	93.27	91.84	81.05	0.802	81.53	78.92	186.30

5. Conclusion

This paper addresses the challenges of UAV image object detection, including small object sizes, wide scale variations, and complex backgrounds, by proposing the CAMDC-Faster RCNN algorithm. By integrating the SENet channel attention mechanism and a novel PSDCFFN into the ResNet50 backbone, the algorithm enhances feature representation for multi-scale and small objects. The adoption of ROI Align reduces localization errors, and K-Means-based anchor optimization improves adaptability to UAV image characteristics. Experimental results on the RSOD-Dataset and a custom UAV image dataset demonstrate significant improvements, with mAP values of 92.52% and 98.07%, respectively, outperforming baseline methods. The algorithm effectively handles multi-scale and small objects, though challenges remain in detecting heavily occluded or indistinct targets. Future work will focus on enhancing robustness to occlusions and further optimizing computational efficiency for real-time UAV applications.

References

[1] Su A, Sun X, Zhang Y, et al.Efficient rotation-invariant histogram of oriented gradient descriptors for car detection in satellite images [J].IET Computer Vision, 2017, 10: 634-640.
[2] Bay H, Ess A, Tuytelaars T, et al.Speeded-up robust features(SURF) [J].Computer Vision and Image Understanding, 2008, 110(3): 346-359.
[3] Girshick R, Donahue J, Darrell T, et al.Rich feature hierarchies for accurate object detection and semantic segmentation [C]//2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014: 580-587.
[4] Girshick R.Fast R-CNN [C]//2015 IEEE International Conference on Computer Vision, 2015: 1440-1448.
[5] Ren S, He K, Girshick R, et al.Faster R-CNN: Towards real-time object detection with region proposal networks [C] //28th International Conference on Neural Information Processing Systems, 2015: 91-99.
[6] Redmon J, Divvala S, Girshick R, et al.You only look once: Unified, real-time object detection [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779-788.

DOI: 10.54254/2755-2721/2025.24679

[7] Liu W, Anguelov D, Erhan D, et al.SSD: Single shot MultiBox detector [C]//European Conference on Computer Vision, 2016: 21-37.
[8] Hu J, Shen L, Sun G.Squeeze-and-excitation networks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: 7132-7141.
[9] He K, Zhang X, Ren S, et al.Deep residual learning for image recognition [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.
[10] Li Y, Chen Y, Wang N, et al.Scale-aware trident networks for object detection [C]//2019 IEEE/CVF International Conference on Computer Vision, 2019: 6054-6063.
[11] Fang Y, Li Y, Tu X, et al.Face completion with Hybrid Dilated Convolution [J].Signal Processing: Image Communication, 2020, 80: 115664.
[12] Long Y, Gong Y, Xiao Z, et al.Accurate object localization in remote sensing images based on convolutional neural networks [J].IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(5): 2486-2498.
[13] Xiao Z, Liu Q, Tang G, et al.Elliptic Fourier transformation-based histograms of oriented gradients for rotationally invariant object detection in remote-sensing images [J].International Journal of Remote Sensing, 2015, 36(2): 618-644.