RESEARCH ARTICLE

RSP Science Hub

International Research Journal on Advanced Science Hub
2582-4376
www.rspsciencehub.com
Vol. 07, Issue 12 December

Check for updates

http://dx.doi.org/10.47392/IRJASH.2025.132

Interactive Selective Spatial Feature Extractor for Small Object Detection in Challenging Scenes

Shylaja D N[1], Leesabanu[2], Shweta S[3], M Kalpana[4]

[1,2,3,4] Assistant Professor, Computer Science and Engineering, Bangalore Technological Institute, Bangalore, India.

Email ID: *shylajadn222@gmail.com[1], leesasubhan@gmail.com[2], vanshika072023@gmail.com[3], mailingkalpana@gmail.com[4]*

Article history

Received: 29 September 2025
Accepted: 27 October 2025
Published: 26 December 2025

Keywords:

C2f-Darknet ; DOTA; Interactive Selective Spatial Feature Extractor ;Small Object Detection; VisDrone;

Abstract

Small object detection in computer vision has made great strides in accuracy and robustness. However, practical applications are still hampered by two main issues: inaccurate detection of small objects and the difficulty of deploying these models on resource- constrained devices due to their extensive parameters and high computational demands. To address these limitations, we propose the Interactive Selective Spatial Feature Extraction method. Which is designed for efficient small object detection, utilizing a modified C2f-Darknet backbone for robust feature extraction and an attention mechanism to focus on crucial spatial areas. The core innovation of this technique lies in its ability to extract and combine multi-scale information to capture fine-grained details and border content, and an internal interaction mechanism that allows communication and information sharing to selectively attend to the most relevant features while suppressing noise. Proposed a state-of-art system that is evaluated on a publicly available DOTA dataset and VisDrone dataset. The proposed methodology detects the accurate region.

1. Introduction

Small object detection is a challenging subfield of computer vision that focuses on identifying and localizing objects in images or videos that are significantly smaller than the overall image size. Unlike traditional object detection tasks, which often deal with objects that occupy a substantial portion of the image, small object detection aims to accurately detect objects that are barely visible, such as distant pedestrians in aerial imagery or tiny defects in industrial inspection. Small object detection finds applications in various fields, including remote sensing for vehicle and ship detection, medical image analysis for cancer detection, autonomous vehicles for pedestrian and obstacle detection, security surveillance for face recognition and object detection, industrial inspection for defect detection and quality control, and robotics for object picking and placement, as well as navigation and obstacle avoidance. By leveraging deep neural networks, inspired by the human brain's architecture, computer vision researchers have made significant strides in object

detection. These networks excel at pattern recognition, enabling precise object identification within images. This technological advancement holds the potential to revolutionize various fields, including military applications. Small object detection (SOD) is a cutting-edge technology that can significantly enhance defence and wartime capabilities. By leveraging computer vision algorithms, SOD enables the identification and tracking of small objects within images or video feeds, even in challenging conditions. This technology can be applied to various defence and wartime scenarios, including surveillance, reconnaissance, border security, maritime security, target acquisition, intelligence gathering, situational awareness, counter-IED, and search and rescue. SOD can help detect and track small objects like hidden weapons, explosives, drones, boats, vehicles, mines, underwater drones, debris, enemy soldiers, vehicles, equipment, and survivors in disaster zones. However, challenges such as algorithm development, computational resources, data acquisition, and real-time processing need to be considered the potential of SOD. By overcoming these challenges, SOD can revolutionize defence and wartime operations, providing a significant advantage in various mission-critical scenarios. For soldiers operating in low-light or nighttime conditions, this technology could serve as a high-tech nightlight, enhancing situational awareness by detecting people and objects in the dark.

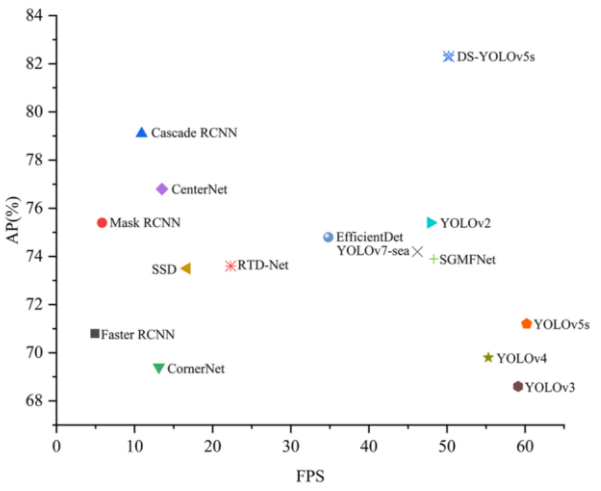

Figure 1 FPS and AP Comparison of Object Detection Methods On the Visdrone

By providing soldiers with the ability to perceive their surroundings more accurately, even in challenging visibility conditions, this technology can significantly improve their safety and operational effectiveness. Datasets are essential for assessing the performance of small object sppoting and tracking algorithms. Videos gathered by satellites or UAVs present unique challenges for these algorithms due to their larger dimensions, smaller objects, and higher object density. The increased altitude of target objects in satellite videos poses an additional layer of difficulty. Several publicly available datasets are widely used for research and development in this area. Figure 1 FPS and AP Comparison of Object Detection Methods On the Visdrone

2. Method

The proposed system is the interactive special significant map generator which is focused mainly on region of interest which is the official future extractor, which concentrates mainly on the needed part of the given dense image as well as the challenging scenarios. The overall architecture is divided into three parts. The first part is backbone. described in Figure 4 The backbone is modified C2f-Darknet is responsible for fusing visual features followed by convolutional layers that through input it extract relevant features. Which generates grid cells by dividing the image into equal-sized cells using multiscale image features. To extract features, proposed architecture employs convolutional layers that progressively learn complex patterns from image data. which is feature extractor used to extract the multi scale feature map from the given input images. This feature maps forwarded into the proposed method Interactive Selective Spatial Feature Extractor which accepts the feature maps and launch the weights for each special locations are letting the most relevant regions for the small object detections which are important the next layer. The interactive special significant map generator typically begins by generating two distinct spatial representations of the input feature map: an average-pooled map capturing the mean activation across channels, and a max-pooled map capturing the maximum activation across channels. These representations are then pipelined along the channel dimension, creating a combined feature map. Subsequently, a small-kernel convolution is applied to this concatenated map to learn spatial dependencies. The output of this convolution is passed through a sigmoid activation function, generating a spatial attention map where values

range from 0 to 1. This attention map highlights the most important spatial regions within the input. Finally, the original input feature map is multiplied elementwise with this attention map, effectively amplifying activations in the attended regions while suppressing activations in less important areas. Each feature node represents a specific spatial-scale location within the feature pyramid. These nodes receive inputs from other nodes across different levels and scales, embodying the interactive information flow inherent in our architecture. Operations performed by these nodes typically include weighted sums of input features with learnable weights to dynamically adjust feature contributions, non-linear activations to introduce non-linearity, and spatial transformations such as convolutions or depthwise separable convolutions to capture spatial relationships within the features. we integrated the interactive mechanism which eliminates the collections by simplifying the network and reducing the redundancy this connection that flow both top down and bottom up to extract the richer information exchange between the different level of the feature pyramid which is a weighted fusion. When fusing multiple features of input with different resolutions, first process is to resize them to the same resolution and then bind them up. Different input features are at different resolutions, they usually contribute to the output features unequally by adding an additional weight for each input during feature fusion, making the network to educate the significance of each input features interactive module gets the fused features . Which combine features from different sources allowing network to learn the importance of each input features. Which improves the information flow and robustness of future representation for detecting the small objects. Which is cross scale connection after this process we refine the future map by emphasizing the most important special locations of the small objects which improves the quality of fused features this is the complete operation in the neck of the architecture the last part is the add part the detection add where the bounding boxes and class probabilities are being predicted. Figure 2 shows Image Processing in Proposed System

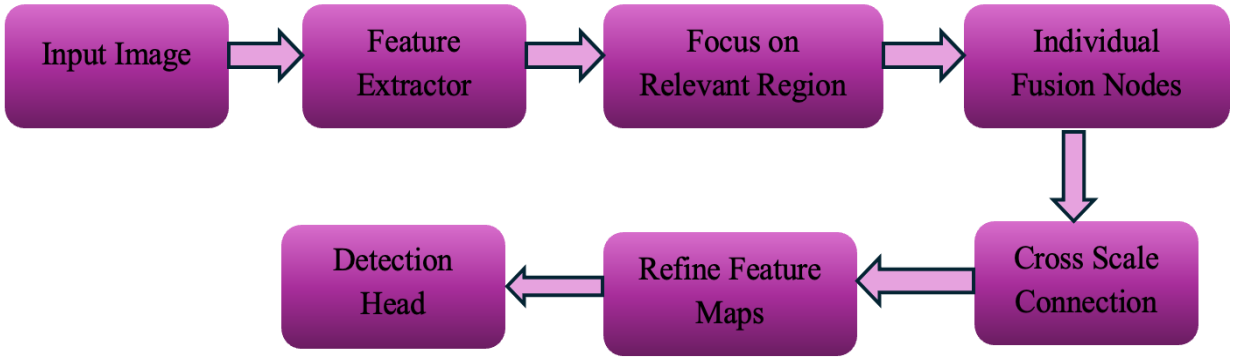

Figure 2 Image Processing in Proposed System

The interactive special significant map generator typically begins by generating two distinct spatial representations of the input feature map: an average-pooled map capturing the mean activation across channels, and a max-pooled map capturing the maximum activation across channels. These two representations are then appended along the channel dimension, creating a combined feature map. Subsequently, a small-kernel convolution is applied to this concatenated map to learn spatial dependencies. The output of this convolution is passed through a sigmoid activation function, generating a spatial map attention where values range from 0 to 1. This attention map highlights the most important spatial regions within the input. Finally, the original input feature map is accumulated elementwise with this output-attention map, effectively amplifying activations in the attended regions while suppressing activations in less important areas.

3. Implementation

Give a brief description of the small object detector design and methodology in this section. Afterwards, discussed the preprocessing procedures that were carried out to incorporating the input into our model.

3.1. Algorithmic Steps for Small Object Detection

The algorithm steps described below outline the process for small object detection using the proposed architecture.

Algorithm 1: Interactive Selective Spatial Feature Extraction

Step 1: Backbone Feature Extraction (using modified C2f-Darknet)

feature_map = backbone(input_image)

Step 2: Generate multi-scale feature maps

multi_scale_feature_maps= generate_multi_scale_feature_maps(feature_map)

Step 3: Interactive Selective Spatial Feature Extractor (ISSFE)

Generate two spatial representations of the feature map:

- Average pooled map (captures the mean activation across channels)
- Max pooled map (captures the maximum activation across channels)

average_pooled_map = average_pooling(feature_map)

max_pooled_map = max_pooling(feature_map)

map=concat_maps(average_pooled_map, max_pooled_map)

Step 4: Apply small-kernel convolution to capture spatial dependencies

convolution_result= small_kernel_conv(concat_map)

Step 5: Apply Sigmoid activation to generate spatial attention map

attention_map = sigmoid_activation(conv_result)

Step 6: Amplify or suppress features based on the attention map

refined_feature_map = apply_attention_map(feature_map, attention_map)

Step 7: Fusing multi-scale feature maps with attention mechanism

fused_feature_map= fuse_feature_maps_with_attention(multi_scale_feature_maps, attention_map)

Step 8: Final feature refinement based on small object detection needs

refined_feature_map= refine_feature_map(fused_feature_map)

Step 9: Prediction of bounding boxes and class probabilities

detection_results = detection_head(refined_feature_map) return detection_results

3.2. Stepwise Description of Feature Processing

- Backbone Feature Extraction: The backbone (e.g., modified C2f-Darknet) processes the input image to generate an initial feature map.
- Pooling Operations: We generate two distinct feature maps: an average-pooled map and a max-pooled map.
- Attention Map Creation: These two pooled maps are concatenated, followed by small-kernel convolution to learn spatial dependencies, and the sigmoid activation creates the attention map.
- Feature Map Amplification: The original feature map is multiplied with the attention map to emphasize the most relevant spatial regions.
- Fusion of Multi-Scale Features: Multiple feature maps from different scales are resized and fused, weighted by the attention map.
- Refinement: The fused feature map is refined for better detection of small objects.
- Prediction Head: The last predictions (bounding boxes and class probabilities) made sfrom the refined feature map.

Used to detect a small object from the existing image by bounding boxes based on the precision for each epoch and also provides the confidence score and class probability. Then it is passed into the vectorization and decoding stage to find the

proper co-ordinations with the set of nodes given in the description as input by using confidence score. Applying Non-Max Suppression (NMS) is a preprocessing step that is primarily used to remove the bounding boxes that are anchors for a single object using the probability score. Performs to find the best candidate for the detected boxes and completes object detection. Overall, an algorithm or model is implemented to predict the object specified in the description. Additionally, techniques like Bounding boxes that are unnecessary or that may overlap considerably were eliminated using non-maximum suppression (NMS). This approach attempts to automatically create high-quality labeled data by first generating potential labels (region-text pairs) and then refining them using a model to ensure relevance and accuracy. This method could be beneficial for tasks that require large amounts of labeled data.

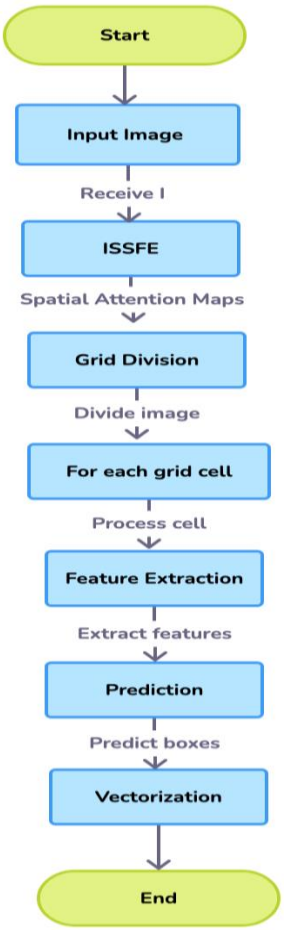

Figure 3 Data Flow Diagram

4. Results and Discussion

This project implements object detection in Python on Google Colab. The code leverages TensorFlow 3.12.13 and runs on a machine possessing an Intel Core i3(2.0 GHz) -6006U CPU and the operating system Windows 10 Pro (64-bit) and Mac. While the project utilizes Google Colab GPU for faster computations, the local machine is specifications are also relevant. The project trains on a VisDrone dataset comprises 10,109 images with a resolution of 2000x1500 pixels, containing 542,000 instances of 10 targets categories commonly found in traffic senario. In addition to TensorFlow, the project also uses other Python libraries like NumPy, Matplotlib, Keras, and Pandas. These libraries facilitate data manipulation, visualization, deep learning model building, and file handling, respectively. It is likely that this project will involve researching existing object detection algorithms through various research papers to inform the model's design. To pre-trained for 50 epochs on Google Colab 4GPUs with total batch size of 16. As shown in Table I, Hyper parameter setting details used in the experiment and Table II contains the configuration setting details. Table 1 shows Hyper Parameter Setting Table 2 shows Experimental Configuration

Table 1 Hyper Parameter Setting

Name of Hyper Parameter	Number
Number of epoch	50
Batch size	16
Input size	640

Table 2 Experimental Configuration

Platform	Name
The Operating System	Windows 10 Pro of 64-bit
CPU	Intel Core i3 of 2.0GHz-6006U CPU
GPU	Google Colab
Python	2.7
PyTorch	0.12.0
S Software	Pycharm 2022

Table 3 Comparison of Different Models on VisDrone-Test-Dev Set

Model	Precision	Recall	mAP @.5%	mAP @.5:.95%
SSD512	9.9	39.4	22.8	-
FPN	26.2	38.6	28.1	-
RetinaNet	12.7	28.8	20.1	-
YOLOx-s	23.5	43.5	32.7	19.1
YOLOx-l	34.3	43.3	36	20
YOLOv3-spp	48.3	32.6	31.3	17
YOLOv5-s	42.7	33.2	30.6	16.5
YOLOv5-l	30.3	45.1	32.8	18.1
ISSFE(ours)	**44.9**	**34.3**	**39.5**	**22.8**

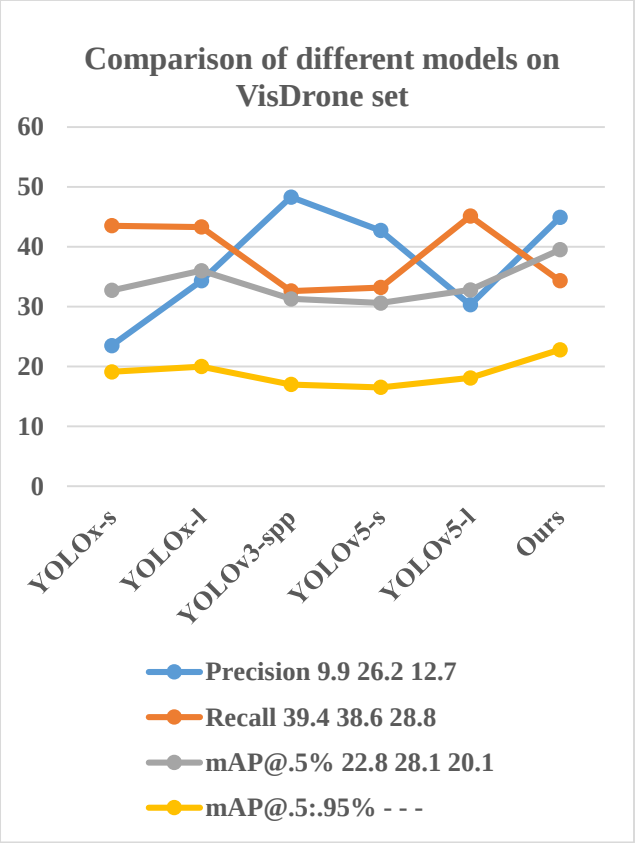

Figure 4 State-of-Art Comparison

Conclusion

The Interactive Selective Spatial Feature Extraction method offers an efficient solution for detecting small objects in crowded images. It uses a modified C2f-Darknet backbone for feature extraction and an attention mechanism to aim on important areas. This approach upgrade the spotting of small objects by highlighting relevant features, reducing noise, and combining multi-scale information, making it more effective in complex and varied image situations. The future scope of this method includes enhancing real-time performance for applications like autonomous driving and surveillance, integrating multimodal data for improved detection in diverse environments, and refining the attention mechanism to adapt dynamically to varying object characteristics and scene complexities. The future scope of this method includes enhancing real-time performance for applications like autonomous driving and surveillance, integrating multimodal data for improved detection in diverse environments, and refining the attention mechanism to adapt dynamically to varying object characteristics and scene complexities.

References

[1]. Ding J, Xue N, Xia G,S Bai X ,Yang W ,Yang M, Object Detection in Aerial Images: A Large-Scale Benchmark and Challenges. IEEE Trans. Pattern Anal. Mach. Intell. 2022, 44, 7778–7796.

[2]. Zhu P, Wen L, Du D, Bian X, Fan H, Hu Q, Ling H, Detection and Tracking Meet Drones Challenge. IEEE Trans. Pattern Anal. Mach. Intell. 2022, 44, 7380–7399.

[3]. Boshra Khalili, Andrew W.Smyth “SOD-YOLOv8 - Enhancing YOLOv8 for Small Object Detection in Traffic Scenes”, arXiv:2408.04786v1 [cs.CV] 8 Aug 2024.

[4]. Sumit Kushwahaa , M Dhanalakshmi et.al,” Efficient Liver Disease Diagnosis Using Infrared Image Processing For Enhanced Detection And Monitoring, Journal of Environmental Protection and Ecology 25, No 4, 1266–1278 (2024) Public health

[5]. Liu, M., Wang, X., Zhou, A., Fu, X., Ma, Y., and Piao, C. (2020). UAV- YOLO: Small object detection on unmanned aerial vehicle perspective. Sensors, 20(8), 2238.

[6]. M Dhanalakshmi, Vyshali Rao et. al. Design of Automatic Greener Healthy System for

Data Center using deep learning",15th International Conference Computing Communication Networking Technologies(ICCCNT),2024

[7]. Wang, G., Chen, Y., An, P., Hong, H., Hu, J., and Huang, T. (2023). UAV-YOLOv8: a small-object-detection model based on improved YOLOv8 for UAV aerial photography scenarios. Sensors, 23(16), 7190.

[8]. Shasi,M.Dhanalakshmi,,S.Saravanan,G.Sujatha, K Tamilarasi, sampath Bhoopathi, "Fog Computing-Based Framework and Solutions for Intelligent Systems: Enabling Autonomy in Vehicles", "Computational Intelligence for Green Cloud Computing and Digital Waste Management",IGI Global Publishing, Feb 2024

[9]. Law, H. & Deng, J. Cornernet: Detecting objects as paired keypoints. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 734–750 (2018).

[10]. Yoshihashi, R.; Trinh, T.T.; Kawakami, R.; You, S.; Iida, M.; Naemura, T. Differentiating Objects by Motion: Joint Detection and Tracking of Small Flying Objects. arXiv 2017, arXiv:1709.04666.

[11]. Duan, K. et al. Centernet: Keypoint triplets for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 6569–6578 (2019).

[12]. P. Hosanna Princye, M. Dhanalakshmi," Expression Analysis Of Death Receptor Ligands Using Biomedical Images", Journal of Environmental Protection and Ecology 25, No 4, 1279–1290 (2024) Public health.

[13]. Kim, H. M., Kim, J. H., Park, K. R. & Moon, Y. S. Small object detection using prediction head and attention. In: 2022 International Conference on Electronics, Information, and Communication (ICEIC), IEEE, pp 1–4 (2022).

[14]. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10778–10787.

[15]. Bochkovskiy, A.; Wang, C.; Liao, H.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv 2020, arXiv:2004.10934.