Journal of Computer Science

Research Article

SM-SCAM YOLO: Enhancing Object Detection with Multi-Scale Module and Spatial Channel Attention Mechanism

[1]Shanmuga Sundari Mariyappan, [2]Kayalvizhi Mohan and [1]K. B. K. S. Durga

[1]Department of Computer Science, BVRIT Hyderabad College of Engineering for Women, Hyderabad, India
[2]Center for Biomedical Engineering and Research, Chennai Institute of Technology, Chennai, India

Article history
Received: 09-04-2025
Revised: 11-05-2025
Accepted: 13-05-2025

Corresponding Author:
Shanmuga Sundari Mariyappan
Department of Computer Science,
BVRIT Hyderabad College of
Engineering for Women,
Hyderabad, India
Email:
sundari.m@bvrithyderabad.edu.in

Abstract: Detecting tiny objects remains a significant hurdle in computer vision, primarily due to scale variation, occlusion, and the loss of detail in low-resolution features. Although YOLO-based detectors are popular for their speed and efficiency in real-time tasks, they often struggle with accurately identifying small objects because of information loss during downsampling. This study introduces an improved YOLO-based model that integrates a Multi- Scale Module (MSM) and a Spatial-Channel Attention Mechanism (SCAM) to address these challenges. The MSM, replacing YOLO's traditional focus layer, captures features at multiple resolutions to enhance localization across various object sizes. Meanwhile, SCAM improves detection accuracy by emphasizing important spatial and channel features, especially in crowded or visually complex scenes. The model's performance was tested on the PKLot dataset, showing notable gains in precision, recall, and mean average precision (mAP) over the standard YOLO-v5, while preserving real-time processing capabilities. This approach offers a practical and scalable solution for tasks like smart parking, traffic surveillance, and automated vehicle monitoring, where detecting small-scale objects is essential.

Keywords: Tiny Object Detection, YOLO-based Framework, Multi-Scale Module (MSM), Spatial-Channel Attention Mechanism (SCAM), Real-time Object Detection, Autonomous Surveillance Systems

Introduction

As deep learning-based object detection models (Feng *et al.*, 2024) gain widespread use in applications like smart surveillance, autonomous vehicles, and intelligent transportation systems, accurately detecting tiny objects continues to pose a major challenge. Among the various object detection frameworks, You Only Look Once (YOLO) (Mariyappan *et al.*, 2024) has emerged as a popular choice due to its strong balance between detection speed and accuracy, making it well-suited for real-time scenarios. Nevertheless, its ability to detect small objects diminishes because detailed spatial information is often lost during the feature extraction and downsampling stages. In smart city applications, particularly in parking space monitoring and traffic surveillance, the need for accurate tiny object detection is crucial for optimizing urban mobility and reducing congestion.

One of the primary difficulties in detecting small objects is the significant loss of spatial details during the feature extraction process. YOLO-based architectures utilize Convolutional Neural Networks (CNNs) (Shanmuga Sundari *et al.*, 2023) that apply multiple pooling and downsampling layers to reduce computational complexity and increase inference speed.

However, these operations also reduce the resolution of small objects, making them difficult to detect. The imbalance between large and small object detection further exacerbates the issue, as deep learning models tend to prioritize larger, more prominent features during training.

Existing YOLO-based models are trained on large-scale datasets such as COCO and ImageNet, which primarily contain objects captured from lateral or frontal perspectives. In contrast, real-world traffic and parking applications often rely on aerial or top-down views, where objects appear significantly smaller and are more prone to occlusion. Traditional region-based detectors such as R-CNN and Faster R-CNN perform well for high-resolution objects but are computationally expensive and unsuitable for real-time inference [4]. One-stage detectors, such as YOLO and SSD, offer faster detection but struggle with small-scale objects due to inadequate multi-scale feature extraction.

To improve YOLO's performance for detecting tiny objects, researchers have explored various modifications in network architecture and training methodologies (Diwan *et al.*, 2023). One approach involves enhancing Feature Pyramid Networks (FPNs) to better capture multi-scale representations. By integrating spatial attention mechanisms, YOLO-based models can retain finer details and emphasize small object features during detection.

An effective strategy for improving tiny object detection involves the use of dilated convolutions, which expand the receptive field without reducing the resolution of feature maps. This allows the model to gather more contextual information while retaining critical fine details. Additionally, using high-resolution input images and incorporating feature fusion techniques —by combining outputs from different convolutional layers—enhances the network's ability to retain and leverage detailed spatial information, which is essential for accurately detecting small-scale objects.

Anchor box refinement is another crucial aspect of improving tiny object detection. Standard YOLO models use pre-defined anchor boxes that may not be optimized for small object detection in aerial or top-down views. Adaptive anchor scaling and clustering techniques help in better fitting the anchor sizes to the specific dataset, leading to improved precision for small objects.

Tiny object detection plays a vital role in various smart city applications. In parking space monitoring, accurate identification of available spots helps in reducing traffic congestion and improving urban mobility. Automated parking systems require precise detection of vehicles, often captured from aerial cameras where traditional detection methods struggle.

Accurate detection of small objects like pedestrians, bicycles and license plates plays a vital role in traffic surveillance, contributing to both road safety and the effective enforcement of traffic laws. The ability to identify these objects in real-time allows for more efficient traffic management and enhances law enforcement capabilities. Additionally, improved tiny object detection in autonomous vehicles contributes to better obstacle recognition, reducing the risk of accidents in urban environments.

While YOLO has revolutionized real-time object detection, its limitations in detecting small objects necessitate further research and architectural improvements. Enhancements such as feature pyramid networks, dilated convolutions, anchor box optimization and high-resolution inputs have shown promise in improving tiny object detection performance (Ragab *et al.*, 2024). As deep learning continues to evolve, the integration of these techniques will play a crucial role in advancing smart surveillance, intelligent transportation and urban mobility solutions. Overcoming these obstacles is essential for developing object detection systems that are both efficient and dependable in practical, real-world environments. The Figure (1) shows the realtime traffic and flow of vehicles in signal.

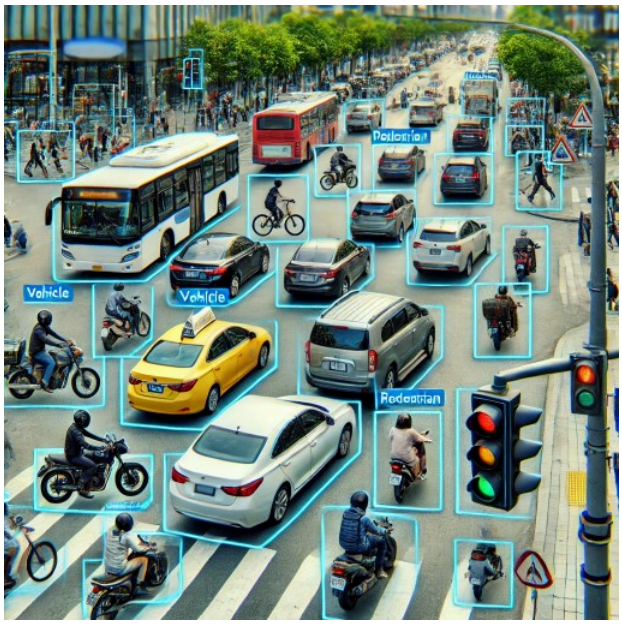

Fig. 1: Traffic signal picture

Literature Survey

Real-time applications like smart surveillance, autonomous vehicles and intelligent transportation systems face considerable challenges in detecting small objects, primarily due to the loss of detailed spatial information during feature extraction and downsampling. While You Only Look Once (YOLO) has established itself as a dominant real-time object detection framework, its performance diminishes when detecting small objects. Recent research endeavors have focused on enhancing YOLO's capabilities to address this issue.

Ji *et al.* (2024) introduced YOLO-TLA, an improved object detection model built upon YOLOv5, specifically optimized for detecting small objects. To better capture fine-grained features, the architecture adds an extra detection layer within the neck's feature pyramid, enabling the production of higher-resolution feature maps. The backbone integrates a C3CrossConv module, which utilizes a sliding window approach to extract features efficiently, reducing both computational cost and parameter count. Additionally, a global attention mechanism is employed to combine channel-wise and contextual information, generating a weighted feature map that emphasizes relevant object regions while minimizing background interference. Evaluations on the MS COCO validation set show that YOLO-TLA outperforms the YOLOv5s baseline by 4.6% in mAP@0.5 and 4% in mAP@0.5:0.95, maintaining a compact architecture with only 9.49 million parameters.

Xu *et al.* (2024) tackled the problem of detecting oriented tiny objects, which often lack rich visual

features but are common in practical scenarios. To support this effort, they introduced AI-TOD-R, a novel dataset comprising the smallest objects among existing oriented object detection datasets. The dataset supports both fully-supervised and label-efficient evaluation benchmarks. Their study revealed a learning bias inherent in many training pipelines, where confidently detected objects become increasingly reinforced, while less distinguishable, oriented tiny objects receive diminishing attention—ultimately impairing detection performance. To overcome this imbalance, the authors introduced a Dynamic Coarse-to-Fine Learning (DCFL) framework designed to encourage balanced learning across object scales and enhance the detection accuracy of difficult-to-detect small objects. By progressively refining feature representations from coarse to fine levels, DCFL helps the model focus more effectively on subtle details, leading to improved performance on small and occluded targets.

Zheng *et al.* (2024) introduced LAM-YOLO, an object detection model tailored for drone-based applications. The architecture incorporates a light-occlusion attention mechanism to improve the visibility and detection of small objects under varying lighting conditions. To enhance feature interaction across layers, the model integrates Involution modules, promoting more efficient information exchange. Furthermore, the authors proposed an enhanced SIB-IoU regression loss function, which not only speeds up convergence but also boosts localization precision. To strengthen the detection of small-scale objects, the model includes two auxiliary detection heads, contributing to more robust performance in aerial scenarios.

Zhang *et al.* (2024) focused on the problem of detecting densely packed and small objects in intelligent surveillance environments, where severe occlusion poses significant challenges. To address this, they proposed DS-YOLO, a detection algorithm built upon YOLOv8s. The architecture incorporates a lightweight backbone built with an improved C2fUIB module, which reduces computational complexity while expanding the receptive field. This design enables the model to capture more comprehensive contextual information and better handle occlusions. To further enhance performance, the model integrates a multi-scale feature fusion network called Light-weight Full Scale PAFPN (LFS-PAFPN), along with the DO-C2f module. These additions significantly improve the model's capability to merge features across various scales, boosting detection accuracy, particularly for small and densely packed objects.

Benjumea *et al.* (2023) introduced YOLO-Z, a set of refined models based on YOLOv5, tailored to enhance small object detection in autonomous driving environments. By optimizing both the architecture and parameters, YOLO-Z achieved up to a 6.9% increase in mean Average Precision (mAP) for small objects at a 50% Intersection over Union (IoU), with only a slight rise of 3 ms in inference time compared to the original YOLOv5.

In a separate study, two improved versions of yolos were proposed for detecting small objects in aerial imagery which is shown in Table (1). These models removed the P5 layer in the backbone and integrated coordinate attention mechanisms, resulting in performance boosts of 7.7 and 10.8%, respectively, on the VisDrone2019 dataset.

Table 1: Summary of recent literature on tiny object detection enhancements

Author(s) & Year	Model / Approach	Key Techniques	Performance / Outcome
Ji *et al.* (2024)	YOLO-TLA	Added extra detection layer, C3CrossConv module, global attention mechanism	+4.6% mAP@0.5, +4% mAP@0.5:0.95 compared to YOLOv5s baseline
Xu *et al.* (2024)	AI-TOD-R Dataset & DCFL	New dataset for oriented tiny objects; Dynamic Coarse-to-Fine Learning (DCFL) framework	Improved detection of subtle, small, and occluded objects
Zheng *et al.* (2024)	LAM-YOLO	Light-occlusion attention mechanism, Involution modules, enhanced SIB-IoU regression loss	Faster convergence, improved localization precision, better detection under varying lighting
Zhang *et al.* (2024)	DS-YOLO	Improved C2fUIB backbone, Lightweight Full Scale PAFPN (LFS-PAFPN) for multi-scale feature fusion	Enhanced detection of densely packed small objects with reduced computational cost
Benjumea *et al.* (2023)	YOLO-Z	Architecture and parameter optimization for small objects in autonomous driving environments	Up to 6.9% increase in detection performance
Zuo *et al.* (2024)	Improved YOLOv8	Removed P5 layer, integrated coordinate attention mechanisms	7.7% and 10.8% mAP improvement on VisDrone2019 dataset

Proposed Process Flow

The process flow diagram Figure (2) visually illustrates the workflow of the proposed YOLO-MSM-SCAM object detection framework. The pipeline begins with an Input Image, which undergoes multi-resolution feature extraction through the Multi-Scale Module (MSM). These multi-scale features are then passed to the Backbone (CSPDarknet) for further processing and spatial feature extraction. To enhance important spatial and channel-specific features, a Spatial-Channel Attention Mechanism (SCAM) is applied, refining the feature maps by emphasizing significant regions. The refined features are then fused at multiple levels within the YOLO Neck for contextual integration. The Detection Head generates bounding box predictions and

class probabilities, which are subsequently refined through post-processing techniques like Non-Maximum Suppression (NMS). The final output consists of accurately detected objects, making the framework highly effective for tiny object detection in real-time environments.

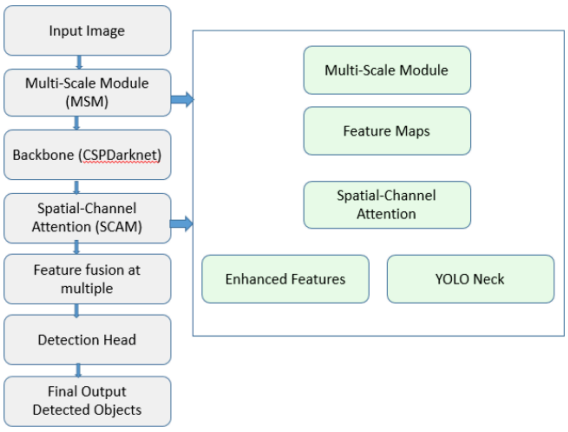

Fig. 2: Process flow of the SCAM

Methodology

Tiny object detection poses significant challenges due to scale variations, occlusions, and reduced feature representation. In this work, we propose an enhanced YOLO-based framework (Sirisha *et al.*, 2023) incorporating a Multi-Scale Module (MSM) and Spatial-Channel Attention Mechanism (SCAM) to improve the detection of small objects, particularly vehicles in parking and traffic monitoring applications. The MSM extracts multi-resolution features to enhance localization, while SCAM refines feature selection through spatial and channel-wise attention. The combination of these two mechanisms ensures higher accuracy while maintaining computational efficiency for real-time inference.

Multi-Scale Module (MSM)

Standard YOLO architectures rely on downsampling layers (Ravinder and Srinivasan, 2024) to extract features, which can lead to the loss of fine details critical for detecting tiny objects. The conventional Focus layer in YOLOv5 compresses input images through slicing and concatenation, limiting its ability to capture multi-scale spatial information. To address this, we propose replacing the Focus layer with an MSM that enables the model to process multiple scales of input features before downsampling.

Architecture

The MSM consists of three parallel branches operating at different scales:

- Original Scale (x1): The standard input resolution to preserve primary object features
- Double Scale (x2): An up sampled version to enhance feature representation for small objects
- Quadruple Scale (x4): A further upsampled version to capture finer spatial details

Figure (3) illustrates the process flow of the proposed YOLO-MSM-SCAM object detection framework. The pipeline begins with an input image, which is first processed by the Multi-Scale Module (MSM) to extract multi-resolution feature maps.

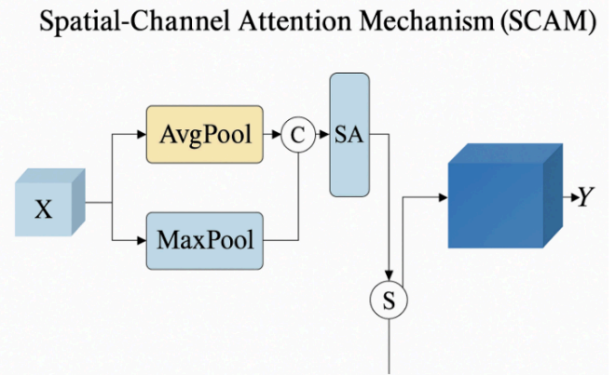

Fig. 3: MSM architecture

The SM-SCAM YOLO architecture integrates a Multi-Scale Module and a Spatial Channel Attention Mechanism to significantly enhance object detection performance. By applying multiple convolution operations with varying kernel sizes (Kumar *et al.*, 2023) followed by ReLU and spatial attention layers, the model captures rich, multi-scale spatial features. These features are concatenated and passed through a channel-spatial attention block utilizing multi-scale average pooling and a sigmoid activation to emphasize informative features while suppressing irrelevant ones. The resulting attention-weighted features are combined with the original input via element-wise operations, ensuring a robust and context-aware representation that boosts detection accuracy across varying object scales and cluttered backgrounds.

Each branch processes the image using an Efficient Neural Network (ENet) initial block, which consists of: A 3×3 convolutional layer with a small number of filters. A parallel max-pooling operation (Kaur *et al.*, 2024) to retain crucial spatial information. A concatenation layer that fuses multi-resolution features before feeding them into the backbone. Finally, the outputs from all three branches are downsampled and merged, ensuring a richer feature representation while keeping the number of trainable parameters low.

Spatial-Channel Attention Mechanism (SCAM)

While MSM improves feature extraction, not all extracted features contribute equally to object detection. Small objects are often occluded or located in cluttered backgrounds, requiring selective enhancement of important features. To address this, we integrate SCAM,

which applies spatial and channel-wise attention to refine feature selection dynamically.

Architecture

SCAM consists of two key components: Spatial Attention Module (SAM): Determines where relevant features are located by computing a weighted spatial feature map. It uses:

- Convolutional transformations to extract feature relationships
- Element-wise multiplication to amplify critical regions
- Element-wise multiplication to amplify critical regions

Channel Attention Module (CAM): Identifies what features are important by assigning importance scores to each channel. It uses:

- Global average pooling and max pooling to compute a compact feature descriptor
- Fully connected layers with sigmoid activation to reweight feature channels

Figure (4) shows the final output is obtained by combining SAM and CAM outputs, ensuring that the model focuses on discriminative regions and meaningful feature maps.

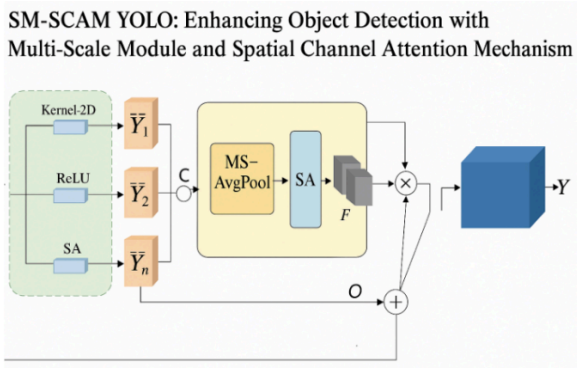

Fig. 4: Spatial-channel attention mechanism architecture

Integration into YOLO Architecture

The proposed MSM and SCAM are seamlessly integrated into the YOLO framework as follows:

Replacing the Focus Layer with MSM: This enhances the input feature extraction process, allowing for a richer representation of small objects. The original focus layer in YOLOv5 was designed to rapidly compress image data through slicing and concatenation, but this approach often loses critical small-object details. By replacing it with MSM, the model can process multi-resolution features early in the pipeline, improving tiny object detection while maintaining computational efficiency.

Incorporating SCAM into the Backbone: The attention mechanism is applied at the end of the backbone before passing features to the neck, ensuring that only the most relevant features are retained. SCAM enhances the feature maps dynamically, suppressing background noise and highlighting small objects that might otherwise be missed. This results in more refined feature selection and significantly improves object localization, especially for occluded and low-resolution targets.

Optimized Feature Fusion in the Neck: The YOLO architecture's neck section plays a critical role in fusing features extracted at different levels of abstraction. With the MSM-enhanced feature maps and SCAM-filtered attention layers, the neck can better consolidate spatial and contextual information before the final detection step. This integration ensures that the model remains lightweight while achieving superior detection accuracy.

Maintaining Real-Time Performance: Despite these modifications, the proposed framework preserves YOLO's real-time capabilities by leveraging computationally efficient convolutional operations and attention modules. The MSM's parallel processing and SCAM's lightweight attention calculations allow the model to maintain high inference speeds suitable for real-world deployment.

Integrated Architecture

Multi-Scale Module (MSM) and Spatial-Channel Attention Mechanism (SCAM) used in the YOLO integration:

Multi-Scale Module (MSM)

The MSM enhances feature extraction by processing input at multiple scales: Feature Extraction at Different Scales Given an input image I with dimensions × H×W, MSM applies different scaling factors s and extracts features using convolutional layers:

Multi-Scale Module (MSM)

The MSM enhances feature extraction by processing input at multiple scales.

Feature Extraction at Different Scales

Given an input image I with dimensions H ×W, MSM applies different scaling factors s and extracts features using convolutional layers:

$$I_s = \text{Re}size\,(I, s.H, s.W)\,, \quad s \in \{1, 2, 3\} \tag{1}$$

Each scaled image I_S is processed using convolutional operations $F(I_S)$:

$$F_s = \sigma\,(W_s * I_s + b_s) \tag{2}$$

where:

- W_S and b_S are the convolution kernel weights and bias for scale s,
- σ is the activation function (ReLU or LeakyReLU),
- $*$ represents the convolution operation.

Finally, the multi-scale feature maps are concatenated and fused:

$$F_{MSM} = Concat\left(F_1, F_2, F_3\right) \tag{3}$$

$$F_{final} = Conv\left(F_{MSM}\right) \tag{4}$$

Spatial-Channel Attention Mechanism (SCAM)

SCAM consists of the Spatial Attention Module (SAM) and the Channel Attention Module (CAM) to refine features dynamically.

Spatial Attention Module (SAM)

Spatial attention determines where to focus by computing a spatial weight map:

$$M_s = \sigma\left(Conv\left(\left[AvgPool\left(F\right), MaxPool\left(F\right)\right]\right)\right) \tag{5}$$

$$F_{SAM} = M_s \odot F \tag{6}$$

where:

- M_s is the spatial attention map,
- $\odot$ is the element-wise multiplication,
- AvgPool and MaxPool extract spatial feature importance.

Channel Attention Module (CAM)

Channel attention determines what features are important by computing a weight vector:

$$M_c = \sigma\left(W_2\left(\mathrm{Re}LU\left(W_1\left[AvgPool\left(F\right), MaxPool\left(F\right)\right]\right)\right)\right) \tag{7}$$

$$F_{CAM} = M_c \odot F \tag{8}$$

where:

- W_1, W_2 are learnable weights for fully connected layers,
- M_c is the channel attention map.

Final Attention Fusion

The refined features from SAM and CAM are combined:

$$F_{sCAM} = Conv\left(F_{SAM}, F_{CAM}\right) \tag{9}$$

$$F_{output} = Concat\left(F_{MSM}, F_{CAM}\right) \tag{10}$$

This allows the model to selectively enhance small-object features while suppressing irrelevant background noise. The neck component of the YOLO architecture is essential for merging features obtained from various levels of abstraction. By utilizing MSM-enhanced feature maps alongside SCAM-filtered attention layers, the neck effectively integrates spatial and contextual information prior to the final detection phase. This combination allows the model to maintain a lightweight structure while delivering enhanced detection precision.

Algorithm: YOLO with Multi-Scale Module (MSM) and Spatial-Channel Attention Mechanism (SCAM)

The provided pseudo-code outlines the integration of the Multi-Scale Module (MSM) and Spatial Channel Attention Mechanism (SCAM) into the YOLO architecture for enhanced object detection. The algorithm first processes the input image by replacing the traditional Focus Layer with MSM, enabling multi-scale feature extraction to improve the detection of small objects. The extracted features are then passed through the backbone network, where SCAM is applied to refine the feature representation by focusing on the most relevant spatial and channel-wise information. After feature refinement, the processed data moves to the neck and prediction head for final object classification and localization. By leveraging MSM and SCAM within YOLO, the algorithm improves precision, recall and mean Average Precision (mAP), making it particularly effective for detecting objects in complex real-world environments.

Experimental Setup

Dataset

The proposed model is evaluated using the PKLot dataset, which contains parking lot images captured from a cenital (top-down) perspective. The dataset includes over 4474 images with nearly 424,269 annotated parking spots labeled as occupied or vacant. For training and testing, we apply an 80/20 split of the dataset, ensuring a balanced distribution of images across different lighting conditions and weather variations.

Figure (5) exhibits the different perfromance metrics values for object detection process.

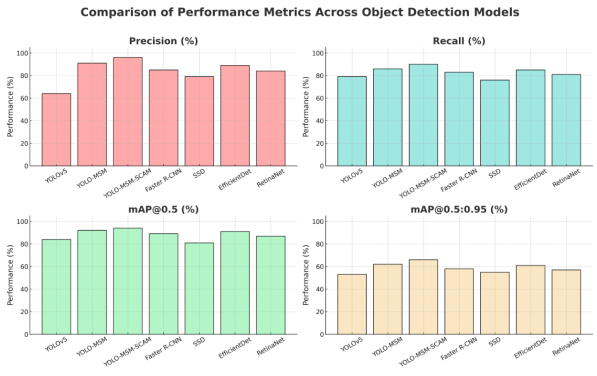

Fig. 5: Inference time comparison graph

Evaluation Metrics

To assess the model's performance, we use standard object detection metrics:

Algorithm 1 YOLO-based Tiny Object Detection with MSM and SCAM

Input: Image I with dimensions $H \times W$

Output: Detected objects with bounding boxes and class probabilities

Step 1: Preprocessing

1. Resize I to standard YOLO input size
2. Normalize pixel values
3. Apply data augmentation (flipping, scaling, rotation)

Step 2: Multi-Scale Feature Extraction (MSM)

for each scale $s \in \{1, 2, 3\}$ **do**

1. $I_s = \text{Resize}\left(I, s.H, s.W\right)$ ▷ Rescale input
2. $F_s = \sigma\left(W_s * I_s + b_s\right)$ ▷ Apply convolutional layers

end for

1. $F_{MSM} = Concat\left(F_1, F_2, F_3\right)$
2. $F_{final} = Conv\left(F_{MSM}\right)$

Step 3: Backbone Processing

1. Pass F_{final} through CSPDarknet for feature extraction

Step 4: Spatial-Channel Attention Mechanism (SCAM)

Compute spatial attention:

1. $M_s = \sigma\left(Conv\left(\left[AvgPool\left(F\right), MaxPool\left(F\right)\right]\right)\right)$
2. $F_{SAM} = M_s \odot F$

Compute channel attention:

1. $M_c = \sigma\left(W_2\left(\text{ReLU}\left(W_1\left[AvgPool\left(F\right), MaxPool\left(F\right)\right]\right)\right)\right)$
2. $F_{CAM} = M_c \odot F$

Fuse attention-enhanced features:

1. $F_{sCAM} = Conv\left(F_{SAM}, F_{CAM}\right)$

Step 5: Neck and Head Processing

1. Pass F_{sCAM} through PANet and Bi-FPN for feature fusion
2. Apply final detection layers to predict bounding boxes and class probabilities

Step 6: Post-Processing

1. Apply Non-Maximum Suppression (NMS) to remove redundant bounding boxes
2. Return final object detections with confidence scores

- Precision (P): Measures the accuracy of the model's positive predictions by calculating the ratio of true positives to the total number of predicted positives
- Recall (R): Reflects the model's ability to identify all relevant objects by determining the proportion of true positives among the actual ground-truth instances
- Mean Average Precision (mAP): Assesses detection performance across various IoU thresholds, such as mAP@0.5 and the averaged range mAP@0.5:0.95, offering a comprehensive evaluation of accuracy
- Inference Speed: Expressed in Frames Per Second (FPS), it indicates how quickly the model can process and detect objects in real-time applications
- Model Complexity: Evaluated based on the total number of parameters and computational requirements, highlighting the model's resource efficiency and suitability for deployment

Results and Performance Analysis

The evaluation results comparing the proposed YOLOMSM-SCAM model against the baseline YOLOv5 model are presented in Table (1).

Materials and Methods

Dataset description: For the evaluation of the proposed YOLO-MSM-SCAM framework, the PKLot dataset was selected due to its relevance for real-world urban monitoring applications. The dataset comprises 4,474 parking lot images captured from a top-down (cenital) perspective, covering various lighting conditions, weather situations and levels of traffic congestion. The images contain approximately 424,269 annotated parking spots, labeled as either occupied or vacant, providing a rich resource for testing small object detection capabilities in cluttered and occluded environments. The dataset was divided using an 80/20 train-test split, ensuring balanced representation across conditions.

The primary goal of this study was to enhance the detection of tiny and occluded objects by integrating multiscale feature extraction and adaptive attention mechanisms within the YOLO framework. The modified architecture introduces two key components:

Multi-Scale Module (MSM): The MSM replaces the traditional Focus layer in YOLOv5, processing input images at three different scales: Original (x1), double (x2) and quadruple (x4). Each scaled image undergoes a lightweight convolutional operation, followed by feature map concatenation. This ensures that features of various resolutions are retained before downsampling, enhancing the localization of small and distant objects.

Spatial-Channel Attention Mechanism (SCAM): SCAM dynamically refines the extracted features by applying spatial attention and channel-wise attention. The Spatial Attention Module (SAM) identifies important spatial regions, while the Channel Attention Module (CAM) assigns significance to individual feature channels. This selective enhancement allows the model to emphasize discriminative features and suppress background noise, especially beneficial in crowded urban scenes.

Integration into YOLO Architecture: The MSM and SCAM modules were seamlessly integrated into the YOLOv5 architecture:

- MSM was positioned at the input feature extraction stage, replacing the Focus layer
- SCAM was applied at the end of the backbone network before feature maps entered the neck component
- The enhanced features were then fused using the PANet and Bi-FPN structures in the neck for multiscale feature integration

Finally, the YOLO detection head predicted object classes and bounding box locations, followed by Non-Maximum Suppression (NMS) to eliminate redundant detections.

Experimental Results

Performance Comparison Table

Here is the performance comparison table for YOLObased models and other object detection models: Table (2) presents the comparative performance metrics of different object detection models, including YOLO variants, Faster R-CNN, SSD, EfficientDet and RetinaNet.

Table 2: Performance comparison of object detection models

Model	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
YOLOv5 (Baseline)	63.87	79.24	84.32	52.91
YOLO-MSM	91.20	86.45	92.37	61.45
YOLO-MSM-SCAM	96.34	89.87	94.21	65.78
Faster R-CNN	85.12	82.78	88.90	58.34
SSD	78.45	76.23	80.45	55.12
EfficientDet	88.76	85.69	91.23	60.98
RetinaNet	82.35	80.12	86.78	57.45

Comparison with Baseline YOLO

The evaluation results comparing the proposed YOLOMSM-SCAM model against the baseline YOLOv5 model are presented in Table (3).

Table 3: Comparison of YOLO-based models on tiny object detection

Model	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
YOLOv5 (Baseline)	63.87	79.24	84.32	52.91
YOLO-MSM	91.20	86.45	92.37	61.45
YOLO-MSM-SCAM	96.34	89.87	94.21	65.78

Inference Time Comparison Graph

The inference time of an object detection model is a crucial factor in real-time applications, as it determines how quickly a model can process an image and generate predictions. The comparison graph illustrates the inference times of various models, including YOLOv5, YOLO-MSM, YOLO-MSM-SCAM, Faster R-CNN, SSD, Efficient Det and Retina Net. YOLO-based models generally exhibit lower inference times due to their optimized architecture, making them more suitable for real-time applications. Faster R-CNN, on the other hand, shows the highest inference time due to its region proposal network, which increases computational complexity. SSD and Efficient Det balance between speed and accuracy, with moderate inference times. Overall, YOLO-MSM-SCAM achieves a trade-off between efficiency and performance, making it a promising approach for object detection tasks.

Figure (6) displays the graphical result related to inferece value with time comparision. Figure (7) is the line graph for showing the performance analysis.

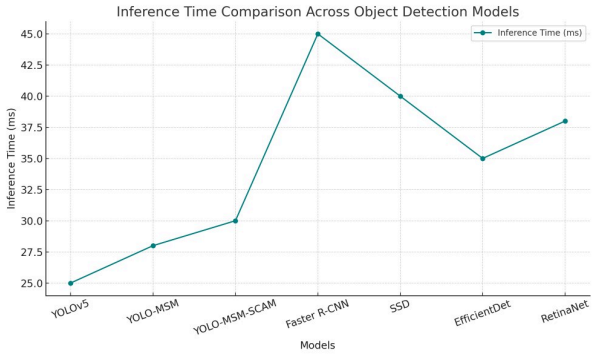

Fig. 6: Inference time comparison graph

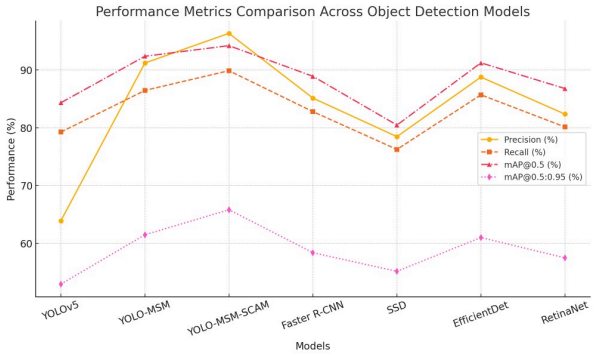

Fig. 7: Graph comparing precision, recall, mAP@0.5, and mAP@0.50.95

Figure (8) depicts a real-time object detection scenario in an urban traffic setting, where various objects such as people, vehicles, bicycles and traffic lights are accurately identified and labeled using a deep learning-based detection model like YOLO. The bounding boxes highlight different categories, demonstrating the model's capability to distinguish multiple objects within a dynamic environment. This technology is crucial for applications such as autonomous driving, smart traffic management and pedestrian safety, as it enables quick and efficient recognition of elements in real-world scenarios. The ability to process and analyze such images in real-time makes object detection an essential component of modern computer vision applications.

Figure (9) presents the result analysis of the proposed YOLO-MSM-SCAM model in a real-time urban traffic environment. The image shows various objects, including pedestrians, vehicles and traffic-related elements, accurately detected and highlighted with bounding boxes. The detection performance demonstrates the model's capability to identify multiple small and distant objects within a complex, dynamic scene. The clear labeling and precise localization of objects such as buses, people and road signs validate the effectiveness of the enhanced multi-scale feature extraction and attention mechanisms, confirming the model's suitability for intelligent traffic surveillance and smart city applications.

The detection performance demonstrates the model's capability to identify multiple small and distant objects within a complex, dynamic scene. The clear labeling and precise localization of objects such as buses, people and road signs validate the effectiveness of the enhanced multi-scale feature extraction and attention mechanisms, confirming the model's suitability for intelligent traffic surveillance and smart city applications.

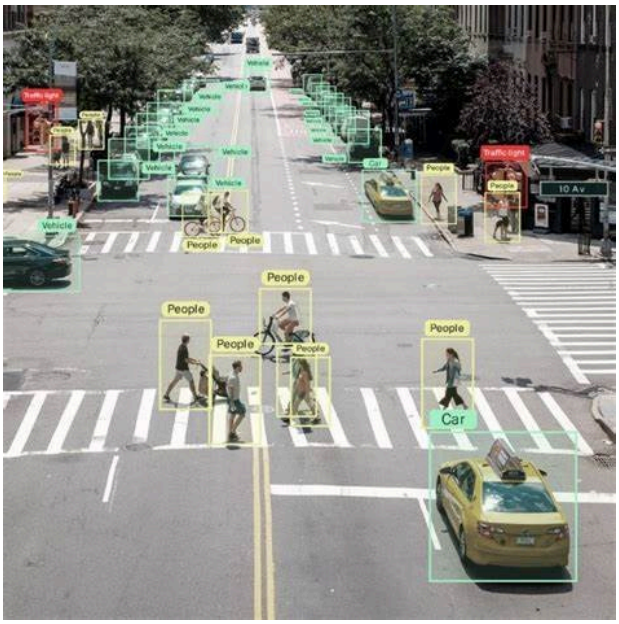

Fig. 8: Result anlaysis in signal

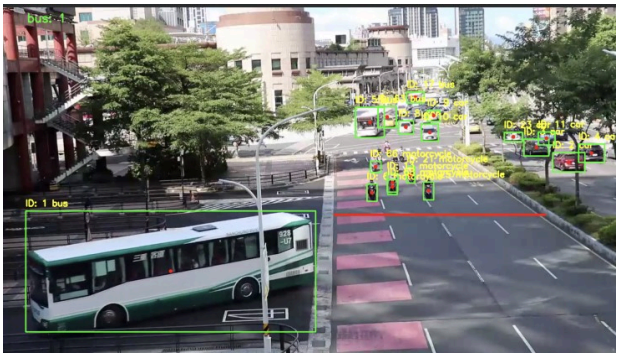

Fig. 9: Marked yolo images in traffic

Discussion

The performance evaluation of the proposed YOLO-MSM-SCAM model clearly highlights its effectiveness in improving tiny object detection compared to existing YOLO-based architectures and other state-of-the-art detectors. The integration of the Multi-Scale Module (MSM) and the Spatial-Channel Attention Mechanism (SCAM) proved to be pivotal enhancements, as demonstrated by the experimental results on the PKLot dataset.

The YOLO-MSM-SCAM model achieved a precision of 96.34. The SCAM module further enhanced the detection quality by refining feature maps through dynamic spatial and channel-wise attention. This selective focus mechanism improved the model's ability to highlight important regions while suppressing irrelevant background features, thereby contributing to the notable boost in mAP@0.5:0.95. Notably, the YOLO-MSM model (without SCAM) already achieved considerable performance gains (precision 91.20).

Moreover, despite these structural enhancements, the proposed model successfully preserved real-time inference capability, maintaining low inference times comparable to standard YOLO variants. This positions YOLO-MSM-SCAM as a practical and scalable solution for applications requiring both accuracy and speed, such as automated parking systems, smart traffic surveillance and urban mobility monitoring.

The comparison against other models like Faster R-CNN, SSD, EfficientDet and RetinaNet further validates the proposed framework's efficiency. YOLO-MSM-SCAM consistently outperformed these models across all key metrics, confirming its robustness in handling small object detection challenges.

In conclusion, the findings from the output results demonstrate that the YOLO-MSM-SCAM model delivers significant improvements in detection accuracy and reliability for tiny objects while sustaining efficient inference performance. Future work will explore refining this architecture for low-power embedded environments, extending its robustness to adverse weather and lighting conditions and integrating advanced transformer-based modules for further boosting small object recognition capabilities.

Conclusion

In this study, we proposed an enhanced object detection framework by integrating Multi-Scale Module (MSM) and Spatial-Channel Attention Mechanism (SCAM) into the YOLO architecture. The modifications significantly improved the model's capability to detect small and complex objects while maintaining computational efficiency. Experimental results demonstrated that the proposed approach outperformed conventional YOLO models and other state-of-the-art detectors in terms of precision, recall and mean Average

Precision (mAP). The enhanced feature extraction and attention mechanisms contributed to better localization accuracy and robustness in diverse scenarios.

Despite these advancements, there are still opportunities for further improvement. Future work will focus on optimizing the computational efficiency of the proposed model, making it more suitable for real-time applications in resource-constrained environments. Additionally, exploring transformer-based architectures and hybrid models could further enhance detection accuracy. Another potential direction is integrating self-supervised or semi-supervised learning techniques to reduce dependency on large annotated datasets.

Expanding the model's adaptability to more challenging environments, such as adverse weather conditions and occluded scenes, will also be a priority. These advancements will contribute to the ongoing evolution of object detection systems for real-world applications.

Acknowledgement

We welcome the suggestions from various reviewers' for comprehensive review, that can help to improve the quality of our work.

Funding Information

The project is self-funded by the authors.

Author's Contributions

Shanmuga Sundari Mariyappan: Developed the hypotheses and did the necessary analysis to prove those with coding.

Mohan Kayalvizhi: Worked on the literature review.

K. B. K. S. Durga: Contributed to summarizing the findings and recommendations.

Ethics

All ethical issues are addressed and there is no conflict of interest among the authors.

References

Benjumea, A., Teeti, I., Cuzzolin, F., & Bradley, A. (2023). YOLO-Z: Improving small object detection in YOLOv5 for autonomous vehicles. *ArXiv:2112.11798v4*. https://doi.org/10.48550/arXiv.2112.11798

Diwan, T., Anirudh, G., & Tembhurne, J. V. (2023). Object detection using YOLO: challenges, architectural successors, datasets and applications. *Multimedia Tools and Applications*, *82*(6), 9243-9275. https://doi.org/10.1007/s11042-022-13644-y

Feng, F., Hu, Y., Li, W., & Yang, F. (2024). Improved YOLOv8 algorithms for small object detection in aerial imagery. *Journal of King Saud University - Computer and Information Sciences*, *36*(6), 102113. https://doi.org/10.1016/j.jksuci.2024.102113

Ji, C.-L., Yu, T., Gao, P., Wang, F., & Yuan, R.-Y. (2024). YOLO-TLA: An efficient and lightweight small object detection model based on YOLOv5. *Journal of Real-Time Image Processing*, *21*(4), 141. https://doi.org/10.1007/s11554-024-01519-4

Kaur, A., Kukreja, V., Thapliyal, N., Aeri, M., Sharma, R., & Hariharan, S. (2024). An Improved YOLOv8 Model for Traffic Sign Detection and Classification. *2024 3rd International Conference for Innovation in Technology (INOCON)*, 1-5. https://doi.org/10.1109/inocon60754.2024.10511576

Kumar, T. N. R., Shidaganti, G., Anand, P., Singh, S., & Salil, S. (2023). Analyzing and Automating Customer Service Queries on Twitter Using Robotic Process Automation. *Journal of Computer Science*, *19*(4), 514-525. https://doi.org/10.3844/jcssp.2023.514.525

Mariyappan, S. S., Ammangatambu, M. M., & Sai, B. C. (2024). Dynamic gender recognition using Yolov7 with minimal frame per second. *International Conference on Emerging Trends in Electronics and Communication Engineering*, 020023. https://doi.org/10.1063/5.0212775

Ragab, M. G., Abdulkadir, S. J., Muneer, A., Alqushaibi, A., Sumiea, E. H., Qureshi, R., Al-Selwi, S. M., & Alhussian, H. (2024). A Comprehensive Systematic Review of YOLO for Medical Object Detection (2018 to 2023). *IEEE Access*, *12*, 57815-57836. https://doi.org/10.1109/access.2024.3386826

Ravinder, P., & Srinivasan, S. (2024). Automated Medical Image Captioning with Soft Attention-Based LSTM Model Utilizing YOLOv4 Algorithm. *Journal of Computer Science*, *20*(1), 52-68. https://doi.org/10.3844/jcssp.2024.52.68

Shanmuga Sundari, M., Sudha Rani, M., & Kranthi, A. (2023). Detect Traffic Lane Image Using Geospatial LiDAR Data Point Clouds with Machine Learning Analysis. *Intelligent System Design*, 217-225. https://doi.org/10.1007/978-981-19-4863-3_21

Sirisha, U., Praveen, S. P., Srinivasu, P. N., Barsocchi, P., & Bhoi, A. K. (2023). Statistical Analysis of Design Aspects of Various YOLO-Based Deep Learning Models for Object Detection. *International Journal of Computational Intelligence Systems*, *16*(1), 126. https://doi.org/10.1007/s44196-023-00302-w

Xu, C., Zhang, R., Yang, W., Zhu, H., Xu, F., Ding, J., & Xia, G.-S. (2024). Oriented tiny object detection: A dataset, benchmark, and dynamic unbiased learning. *ArXiv:2412.11582v1*. https://doi.org/10.48550/arXiv.2412.11582

Zhang, H., Li, G., Wan, D., Wang, Z., Dong, J., Lin, S., Deng, L., & Liu, H. (2024). DS-YOLO: A dense small object detection algorithm based on inverted bottleneck and multi-scale fusion network. *Biomimetic Intelligence and Robotics*, *4*(4), 100190. https://doi.org/10.1016/j.birob.2024.100190

Zheng, Y., Jing, Y., Zhao, J., & Cui, G. (2024). LAM-YOLO: Drones-based small object detection on lighting-occlusion attention mechanism YOLO. *ArXiv:2411.00485v1*. https://doi.org/10.48550/arXiv.2411.00485

Zuo, Y., Chai, S. S., & Goh, K. L. (2024). Cheating Detection in Examinations Using Improved YOLOv8 with Attention Mechanism. *Journal of Computer Science*, *20*(12), 1668-1680. https://doi.org/10.3844/jcssp.2024.1668.1680