IMAGE PROCESSING, MACHINE AND DEEP LEARNING

Enhanced Small Object Detection in High-Resolution Images Using CNN-Boosted YOLOv5

S. Lakshmi Narayanan[1], P. Chinnasamy[2], B J. Dharsan[2], S B. Kailash[2]

[1]Associate Professor, Department of Electronics and Communication Engineering, Sri Ramakrishna Engineering College, Coimbatore, Tamil Nadu, India
[2]Department of Electronics and Communication Engineering, Sri Ramakrishna Engineering College, Coimbatore, Tamil Nadu, India

Published November 2026

Abstract

Small object detection remains difficult in high-resolution images because of the limited spatial information, occlusion, and complicated background scenes. This paper uses a CNN-Boosted YOLOv5 framework to construct an improved small object detection framework that is tailored to improve feature representation when using high-resolution visual images. The proposed method incorporates a highly enhanced CNN-based feature extractor within the YOLOv5 detection system, which allows it to maintain finer-grained spatial information that is required to detect small, distant, and partially covered objects. The experimental outcomes reveal that there are observed improvements in the accuracy of localization, detection confidence and robustness in different real-life conditions. The model can be used in adverse environments, like traffic, crowded areas, and cluttered backgrounds, and it can be used to infer in real-time. The CNN-Boosted YOLOv5 network is appropriate in executing real-time systems by applying an effective trade-off between the accuracy of detection and the efficiency of computation. On the whole, the experiment proves that the performance of small objects detection can be improved by much when the features in the YOLOv5 are enhanced which means that the given approach can be widely applied to autonomous driving, intelligent transportation systems, surveillance, and smart city solutions.

Keywords: Small object detection; high-resolution images; YOLOv5; CNN boosting; deep learning; multi-scale feature extraction; object detection; real-time detection; convolutional neural networks; computer vision

1 Introduction

Object detection is an important part of computer vision system in use in autonomous driving, aerial surveillance, medical imaging, traffic monitoring, and remote sensing. As more and more high-resolution imaging sensors are available, detection systems will capture fine spatial information that will play a vital role in the recognition of objects of different sizes. Detectors based on deep learning, especially convolutional neural networks (CNNs), have also been very effective at the detection accuracy and speed. The most popular among them are YOLO-based architectures, which have a real-time performance, and end-to-end design. Nevertheless, even with these improvements, the small objects in high-resolution images will still be a hard problem to detect because most important features are usually lost in the process of downsampling and features aggregation. Even though a number of object detection models are suggested to enhance the accuracy of detection, most of the ones existing are unable to detect small objects within high-resolution images. Image slicing, multi-scale inference and super-resolution preprocessing techniques enhance the performance of detection

at the cost of higher computational cost and inference latency. Also, such techniques tend to be out of band of the detector and are not known to augment feature representation of the model. Though, YOLOv5 provides efficient multi-scale detection, the backbone, and neck architecture of the used architecture lacks specific optimization to maintain a fine-grained feature of small objects. Thus, the research gap is the development of an intrinsic feature improvement mechanism in YOLOv5 that will ensure a better detection of small objects without reducing the real-time performance of the system.

Figure 1 shows the CNN-boosted YOLOv5 architecture.

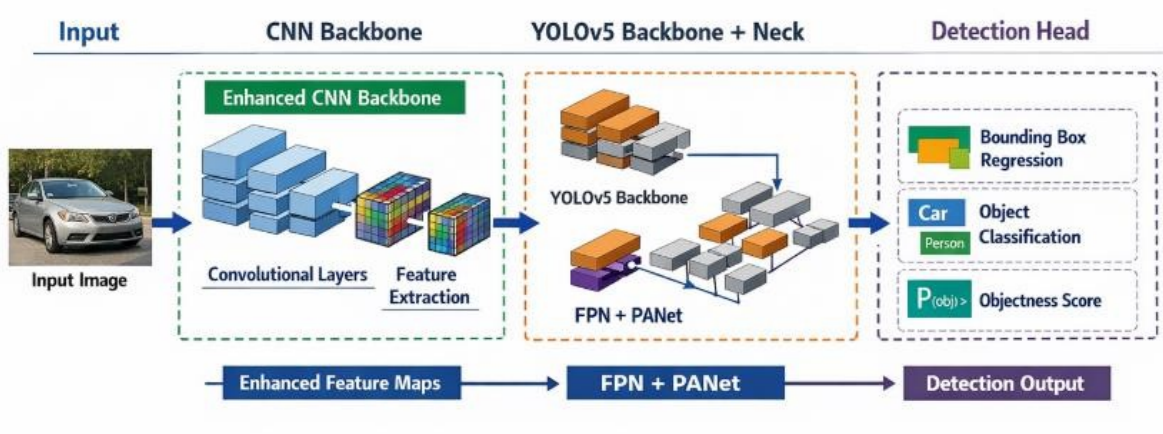


Figure 1. CNN-boosted YOLOv5 architecture.

This project is motivated by the increasing need to have precise small object detection in high-resolution pictures in both real-time and resource-constrained systems. Drones and robots can be used in areas like surveillance, studying traffic, industrial inspection, and medical diagnostics, and they need to identify small objects and objects closely located. Misrepresentation of such objects may cause a serious failure and misunderstanding. Although YOLOv5 has a good base, it is possible to improve its ability to capture details in the fine spatial information, which could help destabilize the detection performance. This encourages the incorporation of CNN-boosting into YOLOv5 in an effort to enhance feature extractions, better object localization and allow the model to remain computationally efficient. Detection of small objects in high-resolution photographs raises several difficult issues such as small pixel representation, scales distortion, background clutter, and dilution of features due to repetition of convolution and pooling procedures. Larger image features, which are the ones in the background, are more likely to be missed as they are lost or blended with smaller object features as the images are filtered at deeper network layers. Besides, high-resolution images require more memory and computation, and it is hard to perform real-time processing. Another significant challenge is balancing the detection along with the speed of inference, particularly when it is to be deployed on edge or embedded devices. The designing of efficient model that improves small object features without causing a lot of increase in model complexity is a very important issue.

This project is aimed at increasing the performance of small object detection in high-resolution images by increasing the intrinsic feature extraction capacity of the YOLOv5 architecture. The project will entail the incorporation of CNN-based boosting modules into the current YOLOv5 architecture to enhance multi scale representation of features. The proposed solution has an objective of enhancing the detection precision and preserving the speed of real-time inferences. The system is analyzed in terms of usual object detection measures like precision, recall and mean average precision (mAP).

The suggested framework can be extended to multiple fields such as aerial imagery, surveillance apparatus, and intelligent transportation, thus it can be deployed in the real world.

2 Literature survey

Object detection is a core task in computer vision that has found applications in autonomous driving, medical imaging, surveillance, and remote sensing. Object detectors based on conventional deep learning are primarily interested in regression and classification of bounding boxes; meanwhile, when dealing with uncertainty and with small-scale objects, they do not perform as well. Recent reports stress that the problem of small object detection is still one of the most significant issues because of the low amount of pixel information and localization uncertainty, particularly in the safety-related conditions [1]. Detection systems based on deep learning have demonstrated dramatic advancement in accuracy, but still, it is challenging to detect tiny and low-evidence objects in real-time environments [2].

Fine-grained spatial information is usually retained using high-resolution images, which create issues of scale variation and computational complexity. To cope with this, they have offered image super-resolution methods with attention mechanisms to improve quality of features. The super-resolution models based on channel and spatial attention exhibit better feature representation that is useful in downstream detection of small objects [3]. The further demonstration of improving the low-level details to detect finer structures in complex scenes is made by joint super-resolution and segmentation frameworks [4].

Object detection models based on CNN have gained much popularity since they are accurate and efficient in computation. The high-speed CNN-based detection plans adapted to a fixed-camera setting are able to deliver real time, albeit with limited small object recognition [5]. Deep learning models on high-resolution satellite images prove to be promising in remote sensing applications, but the dense background and changes in scales have a considerable effect on the detection of small structures [6]. The above studies claim the necessity of the strong multi-scale feature extraction in the high-resolution detection tasks. Single Shot Detectors tries to solve scale variance by means of dense connection and a fusion of features. Better SSD architectures with DenseNet and multi-scale feature fusion are better performers with small objects than the baseline SSD models [7]. Nonetheless, SSD-based techniques are usually slow and inaccurate compared to the YOLO-based detectors.

YOLO models have become popular because of the capability to operate as end-to-end detectors and provide real-time performance. YOLOv5 can be optimized by addition of innovative prediction heads, and enhanced convolutional mixing methods, which show better accuracy on small objects in aerial and drone images [8]. The YOLOv5 model proposes new optimization processes of the backbone, neck, and multi-scale aggregation of features, which strengthen the detection effectiveness on objects of various sizes [9]. The previous versions of YOLO place design principles on how to get the best balance between detection speed and accuracy [10].

The problem of small object detection in high-resolution images is still difficult because the problem of feature dilution is at more deeper levels. Image slicing and patch based methods enhance recall through the processing of high-resolution images in smaller slices, providing enhanced localization of small objects with the same detector architecture without altering the detector architecture [11]. The approaches however add to the time inference and do not improve the intrinsic capability of the model to learn features.

Recent progress investigates the model of transformer-based vision that acquires the global contextual detailing and long-range affiliations, enhancing the recognition achievement in big scale dataset [12]. However, approaches that are based on transformers require high computation resources and they are not applicable in real-time and embedded applications. Lightweight and embedded object detection models work towards minimizing the memory and power usage by using quantization and network simplification to deploy it to devices with resource constraints [13,14]. These models are efficient but in most cases they compromise on the detection accuracy especially when it comes to small objects [15].

Based on the literature available, it can be stated that although YOLOv5 is a good starting point to real-time object detection, its performance on small object detection on large-resolution images can be improved. Current solutions are based on image slicing, super-resolution preprocessing, or on lightweight approximations each of which has limitations. Thus, the present study has a knowledge gap in the enhancement of intrinsic multi-scale feature extraction and representation in the YOLOv5 framework through CNN-based boosting structures to promote the Enhanced Small Object Detection in High-Resolution Images Using CNN-Boosted YOLOv5. Table 1 compares the main studies reviewed.

Table 1. Literature survey comparison.

Ref.	Method/model	Application/dataset	Key findings	Project relevance
[1]	Probabilistic Object Detection	Autonomous Driving	Improved uncertainty modeling in detection	Enhances small object localization in high-res images
[2]	Deep Learning & YOLO	Medical Images, Real-Time Detection	Fast and accurate object detection	Improves detection of small and dense objects
[3]	Channel & Spatial Attention SR	Stereo Endoscopic Images	Enhanced fine details and segmentation	Inspired feature enhancement in YOLOv5
[4]	SEGSRNet	Surgical Instrument Segmentation	Super-resolution with segmentation	Multi-scale attention applied to small object detection
[5]	CNN-based Vehicle Detection	Fixed Camera Scenes	High accuracy for vehicles	Baseline CNN detection; boosts small object recognition
[6]	CNN & Remote Sensing	Earthquake Damage Detection	Detects medium-large objects accurately	Improves detection in high-resolution imagery
[7]	DF-SSD (DenseNet + Feature Fusion)	General Object Detection	Improved small object detection	Efficient multi-scale feature learning
[8]	YOLOv5 + ConvMixer Heads	Drone Imagery	Higher accuracy on small objects	ConvMixer-inspired prediction head improvements
[10]	YOLOv4	General Object Detection	Optimal speed-accuracy tradeoff	Better small object detection than YOLOv4 baseline
[11]	Slicing Aided Hyper Inference (SAHI)	Small Object Detection	Improves small object detection via image slicing	Retains small object efficiency without heavy slicing
[12]	Vision Transformers (ViT)	Image Recognition	High accuracy using transformer blocks	Maintains efficiency with improved small object detection

Ref.	Method/model	Application/dataset	Key findings	Project relevance
[14]	TinyYolo + Quantization	TinyML / Microcontrollers	Low memory footprint, efficient	Focused on high-res images with CNN feature boosting

3 Problem statement

Detecting small objects in high-resolution images has continued to be a major problem in object detection systems today because of the variation in scale, poor pixel representation, background clutter, and feature dilution in the deeper convolutional layers. Even though high-resolution inputs retain fine spatial details, traditional deep learning-based detectors, such as YOLOv5, are not effective to use in discriminative features of small objects due to down sampling operations and insufficient multi-scale feature representation. The current methods like image slicing, super-resolution preprocessing, and lightweight detection models either add to computational complexity, add inference latency, or deteriorate the quality of detection.

Moreover, these methods have the adverse tendency of not improving the inherent feature learning performance of the detector. This has led to the dire necessity of a better detection system that will give better representation of tiny objects in high resolution images and still be able to perform real time. This prompts the invention of CNN-boosted YOLOv5 design that enhances multi-scale feature extraction and small object detection accuracy with high accuracy without adding extensive computational costs.

4 Existing system

Transformer-based networks with complex loss functions such as SCIoU (Scale-aware Complete IoU) are mostly used in the prevailing system to perform small object detection.

Recursive transformer architecture allows the use of multi-scale feature aggregation and attention to enhance the localization and classification of fine objects in high-resolution images. Although the method is more accurate than the traditional CNN-based detectors, it is computationally expensive and consumes a lot of memory and processing power. Also, transformer networks may reduce the speed of inference, which causes difficult real-time detection of large image datasets or high-resolution images.

5 Methodology

5.1 Input image processing and capture module

The process of methodology commences with the high-resolution input images of the real-life conditions like traffic jam, busy-packaged places and surveillance videos. The input images are also resized to an appropriate resolution that is supported by YOLOv5 to maintain its learning efficiency and also conserve the aspect ratios.

Image normalization and data augmentation are applied as in Equations (1)–(4): normalization (1), scaling (2), rotation (3) and brightness adjustment (4),

$$I_{\text{norm}} = \frac{I - I_{\min}}{I_{\max} - I_{\min}} \tag{1}$$

$$(x',\ y') = (s_x \cdot x,\ s_y \cdot y) \tag{2}$$

$$x' = x\cos\theta - y\sin\theta,\ y' = x\sin\theta + y\cos\theta \tag{3}$$

$$I' = \alpha I + \beta \tag{4}$$

where α is the contrast factor and β the brightness offset.

The image normalization, and data augmentation strategies such as scaling, rotation, flipping, and changing brightness are used to increase the generalization and minimize overfitting. These preprocessing steps assist the model to deal with changes in illumination, object scale and background complexity.

Figure 2 shows the data flow diagram.

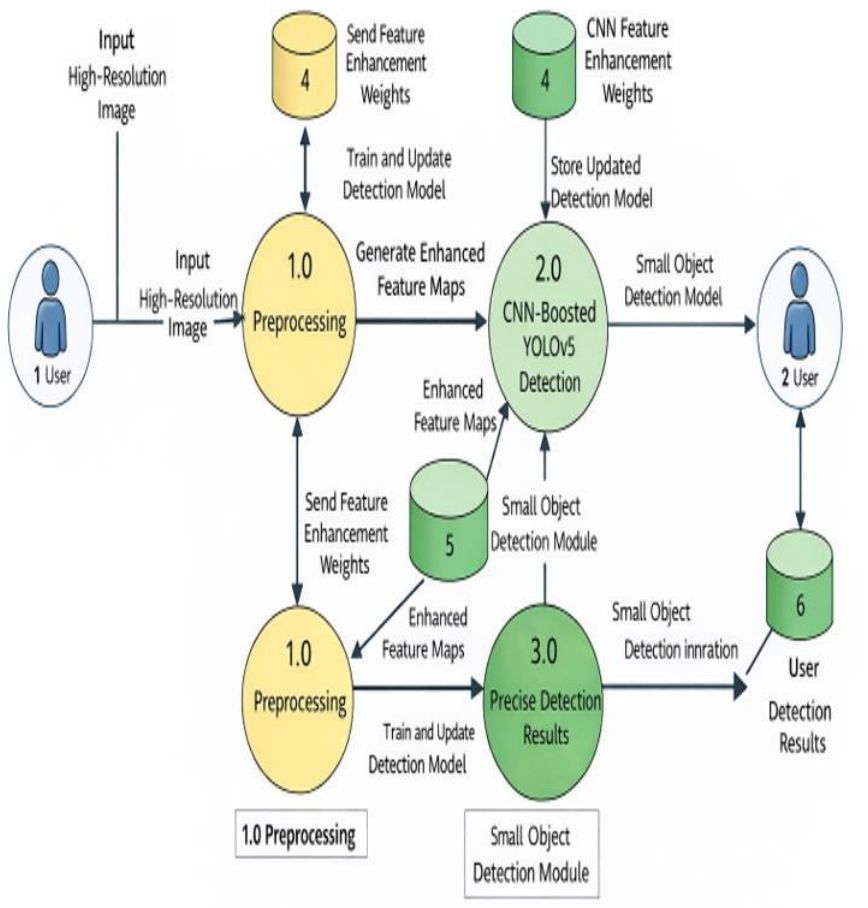


Figure 2. Data flow diagram.

5.2 CNN-based feature extraction module

A better CNN feature extraction module is combined in the proposed system to reinforce a representation of fine-grained spatial details. The module is concerned with the capturing of low level and middle level features which are of critical importance in detecting small and distant objects. The module reduces the loss of information that is caused by aggressive downsampling in traditional detectors by using more convolutional layers and refined feature maps. The features that are extracted contain crucial spatial and contextual features which greatly enhance the visibility of small objects prior to their being passed to the detection pipeline.

The convolution operation and activation function are given by Equations (5) and (6):

$$F_{i,j}^{k} = \sum_{m=1}^{M}\sum_{n=1}^{N} I_{i+m,j+n} \cdot K_{m,n}^{k} + b^{k} \tag{5}$$

$$\mathrm{ReLU}(x) = \max(0,\ x) \tag{6}$$

where $F^{k}_{i,j}$ is the output feature map, I the input image, K^{k} the convolution kernel and b^{k} the bias term.

5.3 CNN-boosted YOLOv5 backbone module

The obtained feature maps are used as input in the YOLOv5 backbone which does the hierarchical learning of features. The CNN-boosted backbone is a decent enhancement to the default YOLOv5 model, which incorporates more meaningful feature representations, and thus offers a better ability to distinguish between objects and background clutter.

The multi-scale feature set extracted by the backbone is given by Equation (7):

$$F = \{F_1,\ F_2,\ F_3,\ \ldots,\ F_n\} \tag{7}$$

This module effectively learns the multi-scale representation at a low computational cost enabling proper detection of the objects of varying sizes, especially small sized targets.

5.4 Feature fusion and neck module

Multi-scale fusion of features is done on the neck module by a Path Aggregation Network (PANet) architecture. It synthesizes both the deep semantic features and the shallow features of high-resolution obtained on the backbone such that both spatial accuracy and the context understanding is maintained. This combination process improves small object detection through strengthening of weak feature responses as well as better information transfer between various network layers. Consequently, the detector will be more resilient in complicated and staffed scenes.

Multi-scale feature fusion is computed as in Equation (8):

$$F_{\text{fusion}} = \sum_{i=1}^{n} W_i \cdot F_i \tag{8}$$

where W_{i} are the learned fusion weights and F_{i} the feature maps from different scales.

5.5 Detection head module

The detection head module produces the final predictions by regressing the bounding boxes, classifying the objects and estimating the confidence score. The model has anchored based detection at varying scales and the results are high accuracy in localization of small objects with minimum false detections. This is because the CNN-boosted architecture has better feature representations that result in increased localization accuracy and high confidence scores, even in partially obscured or distant objects.

The confidence score and intersection over union are given by Equations (9) and (10):

$$\text{Confidence score} = P(\text{object}) \times \text{IoU} \tag{9}$$

$$\text{IoU} = \frac{|B_{\text{pred}} \cap B_{\text{gt}}|}{|B_{\text{pred}} \cup B_{\text{gt}}|} \tag{10}$$

5.6 Optimization module and training

The proposed model is trained on a suitably split train-test set with a labeled set. Gradient-based learning methods are used to optimize loss functions like bounding box loss, objectness loss and classification loss. Hyperparameters such as learning rate, batch size and epochs are well adjusted to ensure optimum performance. The training process guarantees the stable convergence and better generalization of various datasets.

5.7 Real-time inference and visualization of results

The trained CNN-Boosted YOLOv5 model during inference takes input frames in real-time and provides an output of detected objects and bounding boxes, class labels, and confidence scores. The system has been optimized in order to achieve low latency and high accuracy in detection. The results are visualized to monitor and make decisions using the final outputs, this makes the system appropriate to be applied in real-time as with autonomous driving, intelligent transportation systems, surveillance, and smart cities.

Non-maximum suppression (NMS) retains a box according to Equation (11):

$$\text{Keep } B_i \text{if} \text{IoU}\left(B_i,\ B_j\right) < \tau \tag{11}$$

Figure 3 shows the proposed framework.

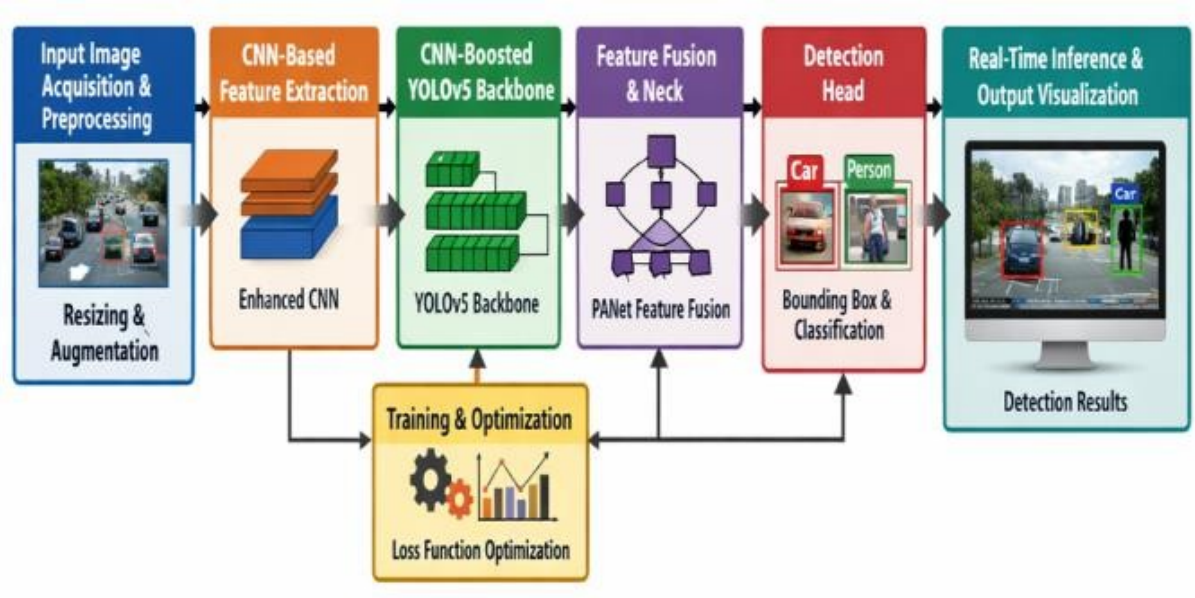


Figure 3. Proposed framework.

6 Proposed system

The proposed system proposes an improved small object detection model built on a CNN-Boosted YOLOv5 framework optimized to work with high-resolution images. The system will have an enhanced CNN-based feature extraction module that will enhance the representation of fine-grained spatial details and then pass the feature into the YOLOv5 detection pipeline. The proposed method ensures that the spatial resolution is maintained during downsampling, in contrast to conventional YOLOv5 models, which may lose important information and, as a result, cannot detect small, distant, and partially covered objects. The enhanced feature maps are fed into the YOLOv5 backbone, neck

and detection head to obtain accurate localization and classification. The system is tuned to be able to run real-time inference as well as achieve a much better detection rates, confidence rates and a much better resistance to interference in tricky scenarios like traffic scenes, groups of people and obstructing backdrop. Due to its good balance in both accuracy and computational power, the proposed system can be used in real-life scenarios such as autonomous driving, intelligent transportation systems, surveillance, and monitoring of smart cities.

Figure 4 shows the system architecture.

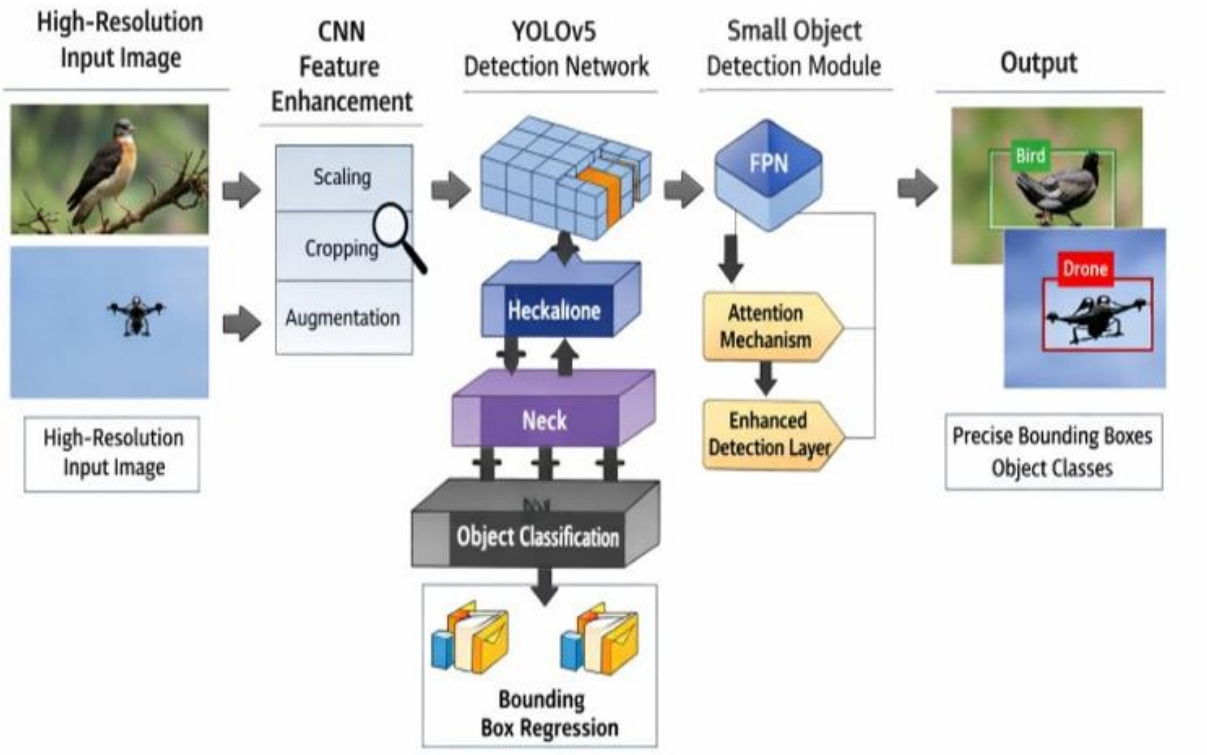


Figure 4. System architecture.

7 Results and discussion

The results of the experiment prove that the suggested CNN-Boosted YOLOv5 model detects small objects in a high-resolution image much better than the standard YOLOv5. Quantitative analysis in terms of mean average precision (mAP), precision, recall and F1-score indicates that the detection accuracy is much higher and false detections fewer with small and partially occluded objects. The CNN-based enhancement of features is effective in preserving fine-grained spatial information, and it results in a greater level of localization and a higher confidence score. Accurate bounding box alignment and strong detection of the objects in crowded scenes, dynamic illumination, and cluttered scenes are all supported by visual analysis. Although the inputs are of high-resolution, the model ensures that the inference performance can be kept close to real-time by being accelerated on a graphics card and using optimal detection parameters. Generally, the findings confirm that increasing the feature representation of the YOLOv5 system can greatly improve detection of small objects without significantly reducing computational performance, and the proposed system can be applicable in the real-life scenario (surveillance, intelligent transportation systems, and smart city monitoring, etc.).

Figure 5 shows the web-based interface in which the high-resolution road image is uploaded to be analyzed. The output of the processing stage points to various vehicles with correctly regionally localized bounding boxes and their respective confidence values. Small objects that are present in complicated urban areas like parked vehicles that are small and partially covered are also detected successfully meaning that the model is robust to work with small objects.

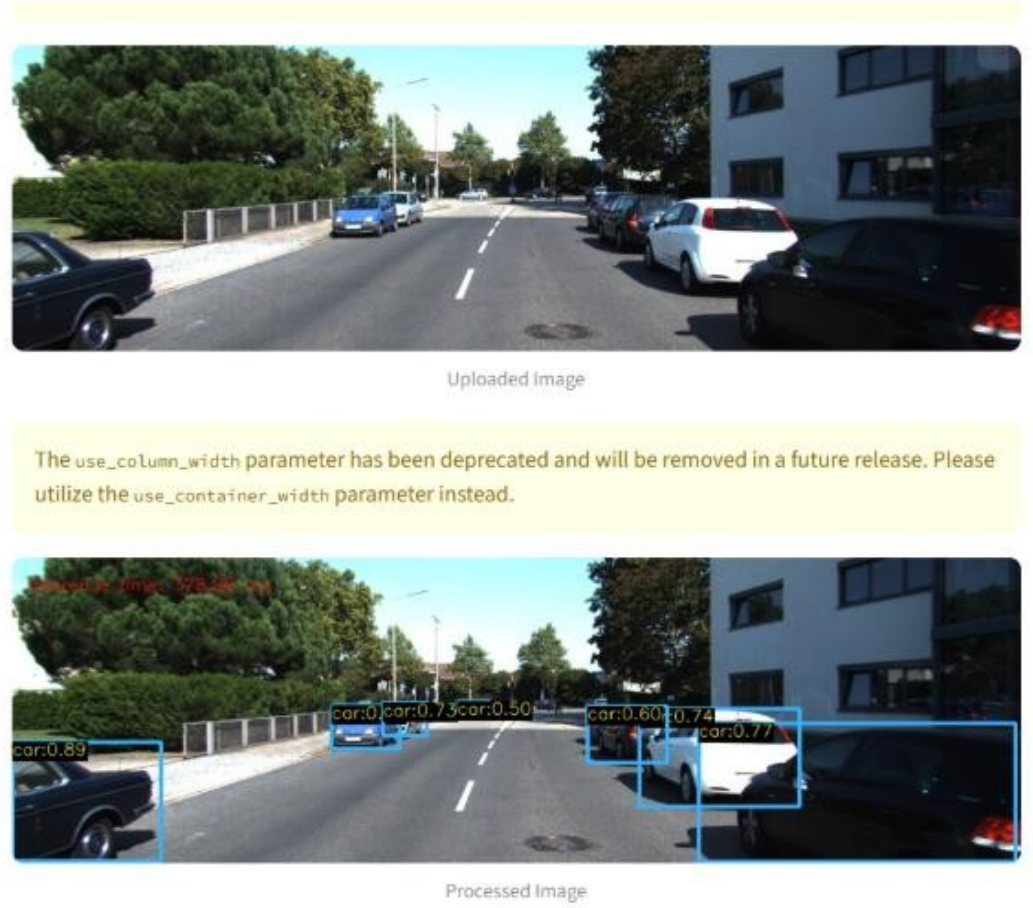


Figure 5. Street scene vehicle detection.

In Figure 6, the system proposed in the paper identifies vehicles and a road sign at the same time on a road scene. It has been found that the model correctly recognizes distant cars and a small traffic sign with high confidence, which reflects better multi-scale feature representation. This finding confirms the usefulness of the CNN-based backbone and feature fusion method in the identification of small but important objects in autonomous driving and intelligent transportation systems (ITS).

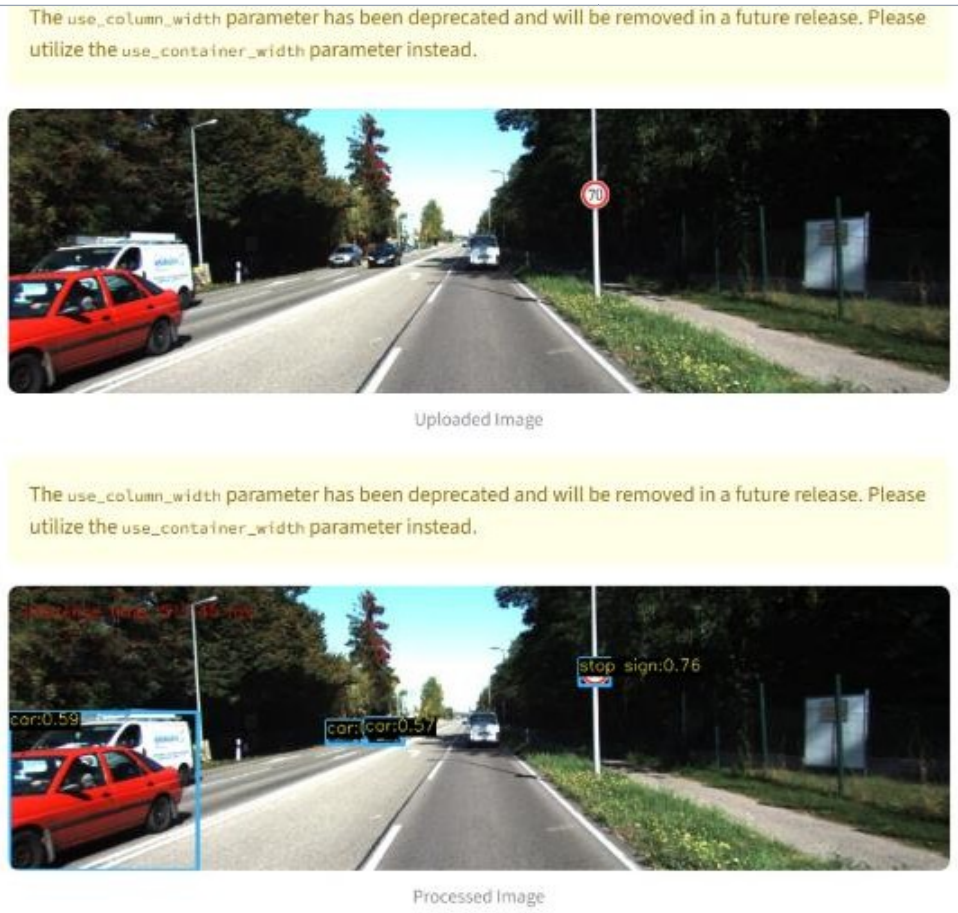


Figure 6. Roadway vehicle and sign detection.

Figure 7 displays the results of object detection in a complicated real world scenario which includes beer mugs, a human being and a horse. The model is effective in separating confounding objects, having a well-defined bounding box and having a dependable confidence rating. This demonstrates the capability of the proposed architecture to be able to generalize outside of traffic situations and be able to operate in cluttered environments with mixed sizes and backgrounds of objects.



Figure 7. Object detection: person and beverage jars.

The results of the crowd analysis shown in Figure 8 include cases when many members of a large crowd are identified at the same time. CNN-Boosted YOLOv5 model is a versatile algorithm that can identify human subjects of high density with few false detections, and thus it is better applicable in crowd surveillance, monitoring of public safety, and event management systems. Table 2 compares the existing and proposed systems, and Table 3 summarises the pixel configuration.



Figure 8. Human detection in a crowded scene.

Table 2. Comparison between the existing and proposed systems.

Parameter	Base Paper	Proposed System
Objective	Robust small object detection with SCIoU & transformers	Enhanced small object detection using CNN-boosted YOLOv5
Architecture	Recursive Transformer Network	YOLOv5 with CNN feature boosting
Loss Function	SCIoU loss	CIoU/GIoU with small object tuning
Feature Extraction	Multi-scale transformer attention	CNN-based multi-scale (PANet/FPN)
Small Object Handling	Scale-aware attention	Anchor optimization + CNN boosting
Input	High-res images	High-res images

Table 3. Pixel configuration.

Parameter	Typical Value
Input Image Resolution	1920×1080 to 2048×2048 px
Small Object Size	10×10 to 40×40 px
Pixel Intensity	0–255 (RGB)
Pixel Normalization	0–1
Feature Map Size	160×160, 80×80
YOLOv5 Stride	8, 16 px

Parameter	Typical Value
Anchor Box Size	8×8, 16×16 px
Bounding Box Error	±1–3 px
IoU Threshold	0.3–0.5

As the outcomes of the experiment obviously show, combining an improved CNN feature extractor with YOLOv5 can indeed lead to a substantial improvement in the performance of the small object detectors when dealing with high-resolution images. The enhanced feature extraction step allows finer grained spatial information to be retained which is essential in the detection of objects at a distance or small scale including traffic signs, vehicles parked and people in a large crowd. The proposed CNN-Boosted YOLOv5 architecture is shown to have a better localization accuracy of small and partially obscured objects, higher confidence scores that reflect better feature discrimination, and have strong performance capacity across a wide range of scenes, such as traffic scenes, crowded social areas, and complex real-world images, as compared to the standard YOLOv5 model.

The real-time inference feature that was seen in the web interface achieves the fact that the suggested approach can be applied in real world engineering systems without incurring much of the computational power. The findings confirm that CNN-Boosted YOLOv5 architecture can improve detection accuracy and preserve real-time performance, which makes it very suitable to autonomous driving, intelligent transportation system, surveillance, and smart city implementation.

8 Conclusion

This paper introduced a superior small object-detecting framework built on a CNN-Boosted YOLOv5 structure created to work with high-resolution images. The proposed method allows the preservation of the fine-grained spatial information, which is of great importance to detect small, distant, and partially covered objects by combining a more advanced CNN feature extraction module with the YOLOv5 detection pipeline. The results of the experiments prove the observable enhancement of the localization accuracy, the confidence scores, and the detection robustness in a variety of realistic situations. The proposed model demonstrates stable behavior in tricky situations like traffic scenes and crowded public areas and background clutter, and the ability to perform real-time inferences. The combination of accuracy and computational efficiency of CNN-Boosted YOLOv5 makes the model a good fit in real-world applications in real-time systems. Altogether, the paper establishes that the representation of features in YOLOv5 can be improved considerably, and this increase has a positive effect on the object detection performance with small objects, and the suggested method can have a high level of application in autonomous driving, intelligent transportation systems, surveillance, and smart cities.

References

[1] Feng D, Harakeh A, Waslander SL, Dietmayer K. *A review and comparative study on probabilistic object detection in autonomous driving*. IEEE Trans Intell Transp Syst. 2022;23(8):Syst.

[2] Kaur A, Singh Y, Neeru N, Kaur L, Singh A. A survey on deep learning approaches to medical images and a systematic look up into real-time object detection. *Arch Comput Methods Eng*. 2022;29(4):2071–2111.

[3] Hayat M, Aramvith S, Achakulvisut T. Combined channel and spatial attention-based stereo endoscopic image super-resolution. In: *Proceedings of the IEEE Region 10 Conference (TENCON)*; 2023. p. 920–925.

[4] Hayat M, Aramvith S, Achakulvisut T. SEGSRNet for stereo-endoscopic image super-resolution and surgical instrument segmentation. *arXiv*:2404.13330 [Preprint]. 2024.

[5] Gomaa A, Minematsu T, Abdelwahab MM, Abo-Zahhad M, Taniguchi RI. Faster CNN-based vehicle detection and counting strategy for fixed camera scenes. *Multimedia Tools Appl*. 2022;81(18):25443–25471.

[6] Salem M, Gomaa A, Tsurusaki N. Detection of earthquake-induced building damages using remote sensing data and deep learning: A case study of Mashiki town, Japan. In: *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*; 2023. p. 2350–2353.

[7] Zhai S, Shang D, Wang S, Dong S. DF-SSD: An improved SSD object detection algorithm based on DenseNet and feature fusion. *IEEE Access*. 2020;8:24344–24357.

[8] Baidya R, Jeong H. YOLOv5 with ConvMixer prediction heads for precise object detection in drone imagery. *Sensors*. 2022;22(21):8424.

[9] Jocher G, Stoken A, Borovec J, Chaurasia A, et al. ultralytics/yolov5: v5.0 – YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations [Software]. Zenodo; 2021. https://doi.org/10.5281/zenodo.4679653

[10] Bochkovskiy A, Wang CY, Liao HYM. YOLOv4: Optimal speed and accuracy of object detection. *arXiv*:2004.10934 [Preprint]. 2020.

[11] Akyon FC, Altinuc SO, Temizel A. Slicing aided hyper inference and fine-tuning for small object detection. In: *2022 IEEE International Conference on Image Processing (ICIP)*; 2022. p. 966–970. https://doi.org/10.1109/ICIP46576.2022.9897990

[12] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: *International Conference on Learning Representations*; 2021.

[13] David R, Duke J, Jain A, Janapa Reddi V, Jeffries N, Li J, et al. TensorFlow Lite Micro: Embedded machine learning for TinyML systems. In: *Proceedings of Machine Learning and Systems*. Vol. 3; 2021. p. 800–811.

[14] Moosmann J, Giordano M, Vogt C, Magno M. TinyissimoYOLO: A quantized, low-memory footprint, TinyML object detection network for low power microcontrollers. *arXiv*:2306.00001 [Preprint]. 2023.

[15] Cerutti G, Andri R, Cavigelli L, Farella E, Magno M, Benini L. Sound event detection with binary neural networks on tightly power-constrained IoT devices. In: *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design*; 2020. p. 19–24.