Karbala International Journal of Modern Science

Manuscript 3339

YOLOv8-CAB: Improved YOLOv8 for Real-time object detection

Moahaimen Talib

Ahmed H. Y. Al-Noori

Jameelah Suad

YOLOv8-CAB: Improved YOLOv8 for Real-time object detection

Abstract

This study presents a groundbreaking approach to enhance the accuracy of the YOLOv8 model in object detection, focusing mainly on addressing the limitations of detecting objects in varied image types, particularly for small objects. The proposed strategy of this work incorporates the Context Attention Block (CAB) to effectively locate and identify small objects in images. Furthermore, the proposed work improves the feature extraction capability without increasing model complexity by increasing the thickness of the Coarse-to-Fine(C2F) block. In addition, Spatial Attention (SA) has been modified to accelerate detection performance. The enhanced YOLOv8 model (Namely YOLOv8-CAB) strongly emphasizes the performance of detecting smaller objects by leveraging the CAB block to exploit multi-scale feature maps and iterative feedback, thereby optimizing object detection mechanisms. As a result, the innovative design facilitates superior feature extraction, “especially the weak features,” contextual information preservation, and efficient feature fusion. Rigorous testing on the Common Objects in Context (COCO) dataset was performed to demonstrate the efficacy of the proposed technique. It is resulting in a remarkable improvement over standard YOLO models. The YOLOv8-CAB model achieved a mean average precision of 97% of detecting rate, indicating a 1% increase compared to conventional models. This study highlights the capabilities of our improved YOLOv8 method in detecting objects, representing a breakthrough that sets the stage for advancements in real-time object detection techniques.

Keywords

Artificial Intelligence; Deep Learning; Computer Vision; Object Detection; You Only Look Once.

Creative Commons License

RESEARCH PAPER

YOLOv8-CAB: Improved YOLOv8 for Real-time Object Detection

Moahaimen Talib [a,*], Ahmed H.Y. Al-Noori [b], Jameelah Suad [a]

[a] Department of Computer Science, College of Science, Mustansiriyah University, Baghdad, Iraq
[b] School of Science, Engineering and Environment, University of Salford, Salford, United Kingdom

Abstract

This study presents a groundbreaking approach to enhance the accuracy of the YOLOv8 model in object detection, focusing mainly on addressing the limitations of detecting objects in varied image types, particularly for small objects. The proposed strategy of this work incorporates the Context Attention Block (CAB) to effectively locate and identify small objects in images. Furthermore, the proposed work improves the feature extraction capability without increasing model complexity by increasing the thickness of the Coarse-to-Fine(C2F) block. In addition, Spatial Attention (SA) has been modified to accelerate detection performance. The enhanced YOLOv8 model (Namely YOLOv8-CAB) strongly emphasizes the performance of detecting smaller objects by leveraging the CAB block to exploit multi-scale feature maps and iterative feedback, thereby optimizing object detection mechanisms. As a result, the innovative design facilitates superior feature extraction, "especially the weak features," contextual information preservation, and efficient feature fusion. Rigorous testing on the Common Objects in Context (COCO) dataset was performed to demonstrate the efficacy of the proposed technique. It is resulting in a remarkable improvement over standard YOLO models. The YOLOv8-CAB model achieved a mean average precision of 97 % of detecting rate, indicating a 1 % increase compared to conventional models. This study highlights the capabilities of our improved YOLOv8 method in detecting objects, representing a breakthrough that sets the stage for advancements in real-time object detection techniques.

Keywords: Artificial intelligence, Deep learning, Computer vision, Object detection, You only look once

1. Introduction

Object recognition for small objects in images is a critical and indispensable task in the field of computer vision, finding applications in various domains, such as identifying pathological cells [1], crime prediction [2], plant classification [3], Epidemic prevention [4], human age recognition [5], and navigation assistance [6]. Despite the advancements in object detection models, accurately detecting small and irregularly shaped objects remains challenging. This difficulty arises because most models primarily concentrate on medium or large objects, often ignoring the intricacies associated with smaller objects.

Recent efforts have focused on creating network structures that are both efficient and accurate for real-time applications. One of the notable examples of these structures include MobileNet [7—9], ShuffleNet [10,11], ResNet [12], and DarkNet [13], specifically designed for performance on hardware platforms, such as CPUs and GPUs, as proposed by researchers.

During training, convolutional neural network (CNN) models learn directly from the original pixel data. That allows them to discover data features and express complex contextual information effectively. Certain CNNs have exhibited substantial enhancements in accuracy and generalizability [14], successfully addressing various image analysis tasks, such as image categorization [15], image region segmentation [16], and image quality enhancement. Object detection algorithms can be broadly classified into two categories: Two-stage algorithms,

Received 22 August 2023; revised 8 December 2023; accepted 10 December 2023.
Available online 23 January 2024

* Corresponding author.
E-mail addresses: moahaimen@gmail.com (M. Talib), a.h.y.al-noori@edu.salford.ac (A.H.Y. Al-Noori), dr.jameelahharbi@gmail.com (J. Suad).

https://doi.org/10.33640/2405-609X.3339

including Fast R–CNN, Faster R–CNN, and Mask R–CNN [17–19], and one-stage algorithms, such as the well-known **You Only Look Once** (YOLO) series algorithms [13,20–24] and single shot multi-box detector (SSD) algorithms [25,26], among others.

The YOLO algorithms have undergone substantial development and are widely recognized as some of the most effective algorithms in the field. Notably, the YOLOv8 algorithm [27], introduced in 2023, has achieved exceptional accuracy, surpassing previous iterations. The YOLO algorithm is primarily designed to identify and categorize objects that occupy the entire image. However, its performance for detecting smaller-scale objects may be comparatively less than certain contemporary algorithms when configured to operate in a unique environment with specific dimensions [28,29].

CNN robustness becomes evident when evaluating its models' performance on visible images. The ability to capture profound input characteristics has led to intensive research in the challenging area of discovering frail and tiny objects in videos using CNNs.

Recent advancements in object detection algorithms have provided robust solutions for identifying medium to large objects in various contexts. However, detecting small and geometric objects still represents an essential challenge, especially for objects whose detection is critical for applications like micro-organism classification or precision agriculture. While the YOLOv8 [27] algorithm introduced an innovative approach, it falls short in environments where object scale and clarity are compromised. The work presents YOLOv8-CAB, an evolution of YOLOv8, specifically engineered to enhance small object detection. Integrating the Context Attention Block (CAB) within the model's architecture addresses the intricacies associated with detecting fine-scale objects without compromising the real-time processing capabilities. This is a significant step forward from the conventional Coarse-to-Fine models.

Empirical evaluations on the COCO dataset demonstrate a 2.1 % increase in mAP for objects under a certain size threshold compared to the baseline YOLOv8, outperforming contemporary models like the NanoDet model in speed and precision. These improvements are not just incremental; they enable the application of YOLOv8-CAB in scenarios where rapid and precise detection of small objects can be life-saving, such as in medical diagnostics or disaster response scenarios.

These advantages substantially contribute to enhancing object detection precision in images. Subsequently, the presented work enhances the existing methodology and proposes an innovative detection framework, YOLOv8-CAB, explicitly designed for identifying diminutive and feeble entities within visual representations. The model demonstrates a high level of reliability and efficiency in the task of object detection within images.

The network prioritizes shallow information and optimizes feature extraction by replacing the Coarse-to-Fine (C2F) [30] module with CAB in its backbone. Moreover, the iterative utilization of the feature extraction module allows for extracting detailed information along with profound features. The spatial attention [31] module has been integrated and improved within the residual blocks, facilitating the adjustment of feature weights and integrating features across the channel dimension. The detection stage incorporates the enhancement of multi-scale feature detection to enhance the detection capabilities for small and low-intensity objects, implemented through four-scale feature maps. The primary goal is to improve the accuracy of object detection. A scientific investigation was conducted on the Common Objects in Context (COCO) [32] dataset to assess the specific influence of each element in the proposed model's network. Empirical results prove the proposed model has better precision and real-time detection capabilities when applied to video data.

Furthermore, as seen later, the proposed method has achieved higher accuracy than other state-of-the-art techniques (such as YOLOv5 and YOLOv8).

The primary contribution of this study can be briefly outlined as follows:

a) This study presents the YOLOv8-CAB approach for detecting small and geometric objects by examining their distinctive characteristics. The method builds upon the YOLOv8 framework and involves an analysis of the network structure, channel compression, parameter optimization, and other relevant factors. Substantial advancements have been implemented to enhance the design of this novel network, YOLOv8-CAB. Specifically, the feature extraction network has been meticulously engineered to fully exploit shallow characteristics while adding four layers to the detection head network to prioritize detecting small and fragile objects. The proposed models exhibit enhanced speed and accuracy compared with current image object detection algorithms.
b) The developed feature extraction network expands upon and iterates the shallow C2F module, replacing it with the Context Attention Block (CAB), explicitly designed to capture local and

global context effectively and efficiently, allowing the network to detect small objects with better performance.

c) In the head, the C2F has been modified by increasing the thickness of the C2F module, which allocates more layers and filters in the convolution operations. This enhancement may improve the model's performance on certain tasks due to the added capacity for feature extraction.

d) Improving spatial attention by adding Selective Kernel (SK) attention [33] to spatial attention, the proposed contribution enhances the spatial attention module by incorporating modifications inspired by the SK attention mechanism. These modifications include a split operation for multiscale spatial modeling, separate fuse and scale steps for flexible feature weighting, and integrating local and global context. These improvements enhance small object detection and selectively highlight important spatial regions and channels. The modified spatial attention module shows potential for surpassing the performance of the original module on various computer vision tasks.

The rest of the paper is divided into the following: Section 2 focuses mainly on related work concerned with object detection using different types of CNN models, the Proposed method including Module. Architecture, the suggested C2F modification, and Spatial Attention Module improvement in section 3. Experimental analysis and results in section 4. Finally, the conclusion and the suggestions for future works are in section 5.

2. Related work

In 2014, R. Girshick, J. Donahue, T. Darrell, and J. Malik introduced the groundbreaking R−CNN [42] for object detection. The presented object detection system divided the process into generating proposals and predicting objects [14]. However, this approach was computationally expensive. Faster R−CNN was proposed in 2015 [17,18], aiming to improve both speed and precision by introducing the region proposal network (RPN) and Region of Interest (ROI) pooling. Various network models have been derived from Faster R−CNN to address different problem domains. However, these models inherit some limitations from R−CNN and Faster R−CNN. The two-stage design of these models makes inference slower due to separate proposal generation and detection steps. Fixed proposal scales may lead to missing small objects, and ROI pooling on sparse proposals can result in losing spatial details. The separate training of the RPN and detection network components also hinders optimization. Moreover, these models lack inherent feature enhancement mechanisms for handling visually similar classes. In contrast, YOLOv8-CAB overcomes these limitations using a single-stage YOLOv8-CAB enhanced detection approach with multi-scale feature processing, enabling faster and more accurate detection, especially for small and visually similar objects.

In 2018, A. Wong and M. Javad developed Tiny SSD [34], a concise deep neural network structure for real-time embedded object detection. This method utilizes Fire modules inspired by Squeeze Net and auxiliary convolutions based on the single shot detector architecture to reduce model size while maintaining accuracy. However, Tiny SSD has some limitations. The model's accuracy, measured by Mean Average Precision (mAP), still falls behind that of detection models due to a trade-off between model compression and maintaining accuracy. Furthermore, the evaluation of the model is constrained in terms of diversity, as it has been scrutinized exclusively utilizing 20 categories extracted from the PASCAL VOC dataset.

In comparison, YOLOv8-CAB emerges as a superior model. It strikes a better balance between model size and accuracy, making it suitable for diverse and challenging datasets. Moreover, the YOLOv8-CAB method does not entail a strict tradeoff between accuracy and model compression, resulting in a more versatile and efficient model for object detection.

In 2022, A. Mishra and H. Aljasmi developed a real-time vision-based laboratory safety monitoring system [35]. They utilized YOLOv5 and YOLOv7 object detection models trained on a unique dataset featuring students wearing four types of personal protective equipment in lab settings. Although their models, especially YOLOv5 versions, performed high accuracy on this modest four-class dataset, they faced limitations in detecting small objects due to the limited samples in this dataset. Furthermore, the proposed model evaluation lacked variety since it was examined using only 20 categories from the PASCAL VOC dataset.

In comparison, YOLOv8 offers clear advantages. The versatility and robustness of YOLOv8-CAB enable it to perform efficiently, even with small objects, addressing one of the primary limitations of Mishra and his teamwork. Additionally, YOLOv8-CAB is more effective in diverse environments than limited to a specific context, such as a single lab.

Therefore, YOLOv8 is a superior choice for real-time monitoring in various settings.

In a study conducted in 2022, F. Sun, J. Gu, L. Deng, and H. Liu proposed an algorithm called SF YOLOv5 [36] for detecting objects. This algorithm builds upon the YOLOv5 algorithm to achieve results. The primary objective of their work was to improve object detection by utilizing their algorithm. They successfully improved detection accuracy and reduced model parameters and computational requirements, outperforming methods, such as YOLOv5s, YOLOv5n, YOLOv3, YOLOv7, and ResNeXt CSP on datasets. However, the study primarily focused on detecting objects, and further exploration is needed to assess its performance with small and full-scale targets. In comparison with YOLOv8-CAB, it surpasses SF YOLOv5 through advanced context modeling via CAB blocks, robust multi-scale architecture, optimized end-to-end training, state-of-the-art performance on small and full objects, and disambiguation capabilities, providing more efficient and well-rounded improvements.

In their 2022 publication, Chang Xu, Jinwang Wang, Wen Yang, Huai Yu, Lei Yu, and Gui-Song Xia introduced a new approach for detecting tiny objects in aerial images. They proposed the Normalized Wasserstein Distance (NWD) and a Ranking-Based Assignment (RKA) strategy [37]. The authors modeled bounding boxes as Gaussian distributions and measured similarity using the Wasserstein distance, which NWD then normalized to a 0−1 range. The RKA strategy assigns positive labels based on ranking proposals rather than simple thresholding. Despite their advances, their method has certain limitations; it is highly dependent on modeling bounding boxes as Gaussian distributions, which might not precisely capture all objects. Performance improvements are primarily attributed to the ranking assignment rather than feature discrimination enhancement, limiting the approach's generalizability. Additionally, more in-depth failure analysis can strengthen their methodology.

In their 2023 study, the authors Kim, M., Kim, H., Sung, J., Park, C., & Paik, J. developed a novel method for small object detection using a high-resolution processing module [38] and a sigmoid fusion module. This method improves small object detection by increasing mAP, using only 57 % of the model parameters and 71 % of the computational power in Giga Floating Point Operations Per Second (Gflops) compared with existing models, such as YOLOX. Although some enhancements have been observed, the model still has challenges. These challenges include the need for more power when dealing with high-resolution images, its ability to perform well across different object detection tasks, and its resilience in handling diverse image conditions and noise. By contrast, YOLOv8-CAB surpasses the limitations of previous work through its efficient CAB modules for context-aware feature enhancement, robust multi-scale architecture to detect diverse object sizes, optimized end-to-end joint training for optimization, and built-in noise and variation resilience from CAB's focus on discriminative features.

YOLOv8 offers a substantial improvement over this approach. It does not rely on specific modeling of bounding boxes and focuses on feature discrimination, allowing for a more generalized and effective application. In addition, YOLOv8-CAB's strategy does not require the intensive analysis needed by Xu and his colleges approach [38], making it a more robust and efficient solution for tiny object detection in aerial images.

In their 2022 study, authors Jiang, B., Chen, S., Wang, B., & Luo, B. developed a novel method for enhanced multi-graph data representation in semi-supervised classification [40], showing promise in leveraging multiple graph structures for enhanced data representation. However, MGLNN's dependency on multiple graph integration may limit its efficacy in singular, complex graph scenarios. In contrast, YOLOv8-CAB excels in object detection, particularly in accurately identifying small and geometric objects. While MGLNN demonstrates effectiveness in multi-graph learning with notable performance in semi-supervised tasks, YOLOv8-CAB achieves superior results in diverse and complex object detection tasks, where MGLNN's approach may not be directly applicable. Thus, YOLOv8-CAB emerges as a more versatile and robust solution, capable of addressing a wider range of real-world detection challenges, surpassing MGLNN in both adaptability and practicality.

In their 2023 study, the authors A.M. Roy J. Bhaduri developed a novel method for YOLO-based object detection models, the approach integrating DenseNet blocks with YOLOv5 [41]. In this work, the proposed model excels in accuracy and speed, particularly in real-time applications, achieving a mean average precision of 85.25 %. However, its primary focus on road damage detection may limit its applicability in detecting smaller, more diverse objects. In contrast, in this paper, the YOLOv8-CAB model, specifically designed for detecting small and geometric objects, demonstrates versatility in various domains. Integrating the CAB and modifications in C2F and Spatial Attention in YOLOv8-CAB address finer nuances of small object detection and contribute to higher accuracy. While YOLOv8-CAB shows potential superiority in broader detection scenarios, this claim would benefit from further empirical evidence demonstrating its enhanced performance. Furthermore, acknowledging DenseSPH-YOLOv5's strengths in its specialized application area provides a balanced perspective. Finally, exploring the potential challenges of YOLOv8-CAB, such as in highly complex environments, would offer a more comprehensive comparison between the two models.

YOLOv8-CAB surpasses previous studies through its versatile context-aware feature enhancement, robust multi-scale architecture, and built-in noise resilience, providing more efficient and generalized tiny object detection capabilities. This is achieved via its CAB modules, optimized end-to-end training, and emphasis on discriminative features.

3. Proposed method

3.1. Module architecture

YOLOv8 is an advanced deep-learning model that further enhances the capabilities of YOLOv5 in the object detection field. It incorporates a range of network structures and utilizes C2F modules to enhance YOLOv5. The CSP modules effectively reduce parameters and FLOPs while maintaining accuracy. Additionally, YOLOv8 introduces a new anchor box generator that considers the distribution of object sizes in the dataset. In this study, YOLOv8-CAB, a method based on YOLOv8, has been proposed. This method is specifically designed to increase the detection accuracy of small and medium-sized objects among multiple objects.

The architecture of YOLOv8-CAB comprises three key components: A network designed to extract features with minimal computational resources within the backbone network, a path aggregation feature pyramid network architecture in the neck network, incorporating multi-scale detection head, and cross-stage feature fusion Fig. 1 shows the entire Yolov8-CAB Network Architecture.

When processing an input color image, YOLOv8-CAB initiates a focus operation that reduces the image size while increasing the number of channels. This aids in enhancing the model's accuracy by focusing on the most crucial image parts. The architecture includes 53 layers using convolution with 22.8 million parameters and operates on images sized 640 × 640 pixels. The image sample is initially reduced to 320 × 320 pixels using max pooling. Subsequently, convolutional layers extract distinctive features, resulting in a 10 × 10 × 1024 tensor known as the “bottleneck input.”

In the second phase, convolution with 1024 filters of dimensions 3 × 3 further reduces the parameters, generating a 1 × 1 × 1024 tensor called the “bottleneck output.” This, combined with the output of the previous convolutional layers, is passed to the next concatenation layer. Phase three involves applying convolutional layers to extract intricate features, and fully connected layers enable object-bound box and class predictions. The resulting feature map, known as scale 2, is obtained through concatenation and additional convolutional operations. The detection layer merges the combined feature map from the phase to produce a list of bounding boxes and class probabilities. To prevent overlapping detections, we use a technique called maximum suppression.

The model predicts three bounding boxes per feature map at every location, providing object locations by bounding boxes and object scores. Fig. 2 shows the CAB Block Architecture.

3.2. C2F modification

The integration and exchange of scale visual features in both directions are improved by expanding the C2F module, which leads to enhanced feature learning. Conceptually, it augments the conventional bottom-up feature pyramid network backbone with an additional top-down refinement pathway. Specifically, the C2F block leverages a top-down convolutional projection sequence to up-sample semantically strong responses from preceding coarse resolution layers. This results in enhanced feature maps with amplified spatial extent and acuity. The up-projected features are concatenated with bottom-up feature maps originating from finer pyramid scales. The C2F block, as shown in Fig. 3, synergizes pyramidal features across scales through this composite integration, enriching fine-grained representations with augmented contextual focus.

Consequently, the hierarchical blending of semantically enriched and spatially precise features demonstrates substantial empirical gains.

Fig. 4 shows the head structure with the extended shallow C2F. Quantitative experiments exhibit state-of-the-art trade-offs between accuracy and efficiency on competitive object detection benchmarks. Ablation studies verify the contribution of bidirectional C2F enhancement to the localization and recognition of small objects. The C2F block is an effective architectural innovation for multi-scale visual modeling. The YOLOv8 model efficiently utilizes the characteristics found in the layers of the CNN by making these adjustments. This careful balance of deep and shallow

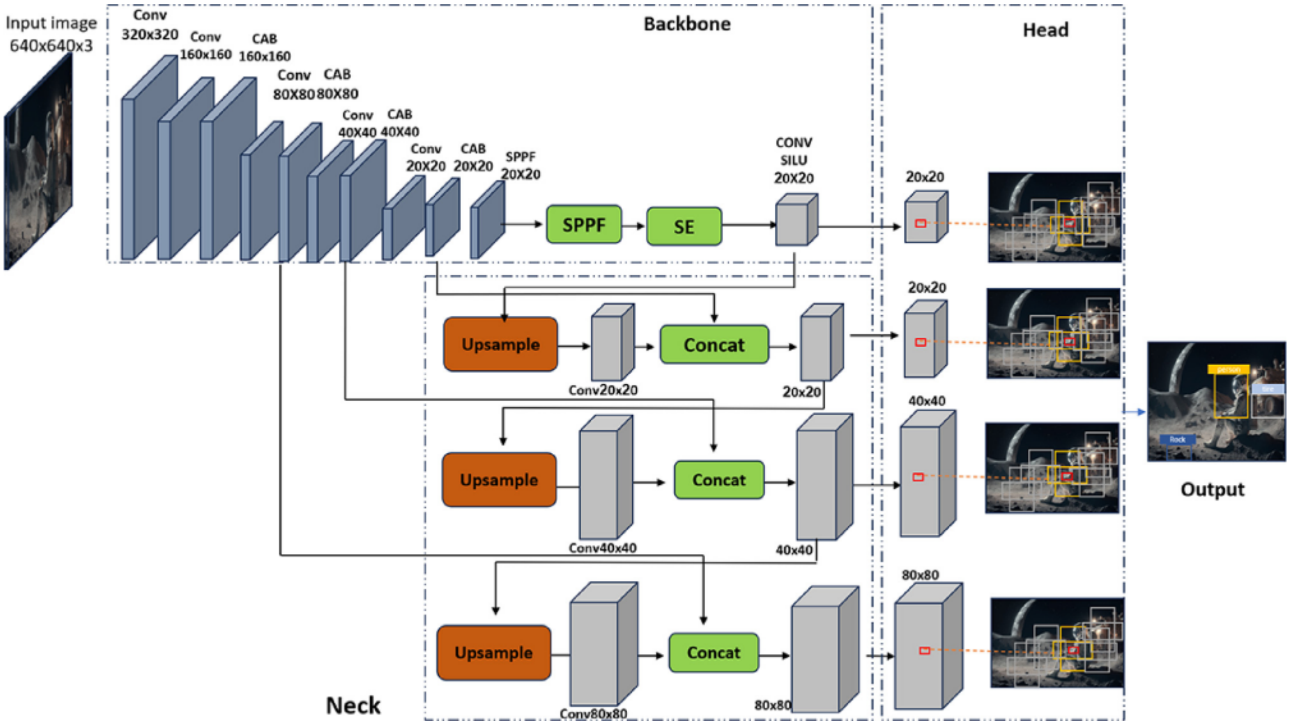

Fig. 1. Full Yolov8-CAB network architecture.

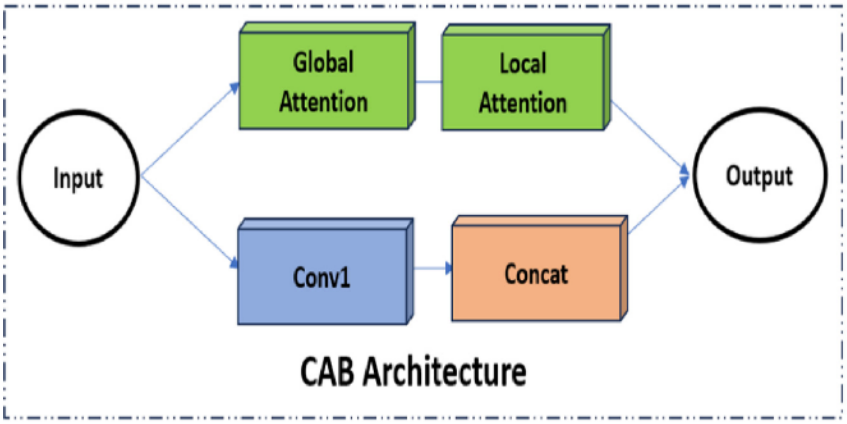

Fig. 2. CAB block architecture.

feature extraction allows the model to accurately identify and locate diminutive and feeble entities within images while minimizing the computational burden.

3.3. Improvement of the spatial attention module

Integrating the Spatial Attention Module with the YOLOv8 model represents a significant contribution with immense potential to enhance small object detection capabilities. The Selective Kernel (SK) Attention Module brings in an approach to adjust the importance of the feature maps, enabling the model to pay attention to significant regions in the input data. This selective attention mechanism empowers YOLOv8-CAB to prioritize features while downplaying ones, enhancing object detection accuracy.

In the context of small object detection, this innovation becomes particularly advantageous. By leveraging the SK Attention Module, YOLOv8-CAB becomes more adept at capturing these nuances, thus enhancing YOLOv8-CAB's ability to identify small objects accurately.

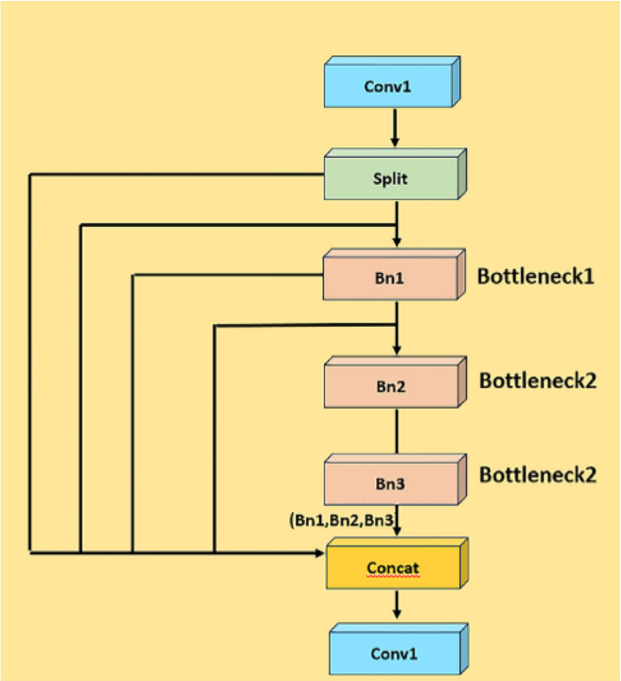

Fig. 3. Extended C2f Module.

On the other hand, attention mechanisms in deep learning models directly affect the decision-making process. Attention allows the model to focus on specific input parts more relevant to the task. By dynamically weighing different parts of the input, attention helps the model to make more informed decisions, especially in cases where certain input features are more crucial than others. In essence, attention allows the model to allocate its computational resources more efficiently, leading to potentially better performance and improved generalization in many tasks. Fig. 5 shows the improved spatial attention.

3.4. YOLOv8-CAB algorithm

To harness the detailed architecture of YOLOv8-CAB and its components, we introduce the following algorithm to effectively detect objects, especially focusing on small and medium-sized entities:

Input: video data, classes (small objects).
Output: bounding boxes with class labels.

1. Split video into frames at 30 fps and resize each frame to 320 × 320.
2. Preprocess each frame for enhanced image clarity.
3. Detect ROIs using the YOLOv8-CAB model:
   - Grid frames and determine object origins using a CAB-integrated Backbone with 53 convolutional layers and 22.8 million parameters.
   - Predict bounding boxes and class probabilities in the Head containing 11 layers, three blocks, and 177 K parameters.
4. Consolidate bounding boxes using intersection over union (IOU) and non_max_suppression.
5. Retain boxes with the top IOU scores, discarding those below a threshold (e.g., 0.5).
6. Continue the process until all boxes are either selected or discarded, with the YOLOv8s head able to predict up to 128 bounding boxes and class probabilities.

End

4. Experiment analysis

4.1. Experimental dataset

The COCO dataset is widely employed in computer vision for object detection tasks. It provides annotations specifically customized for object

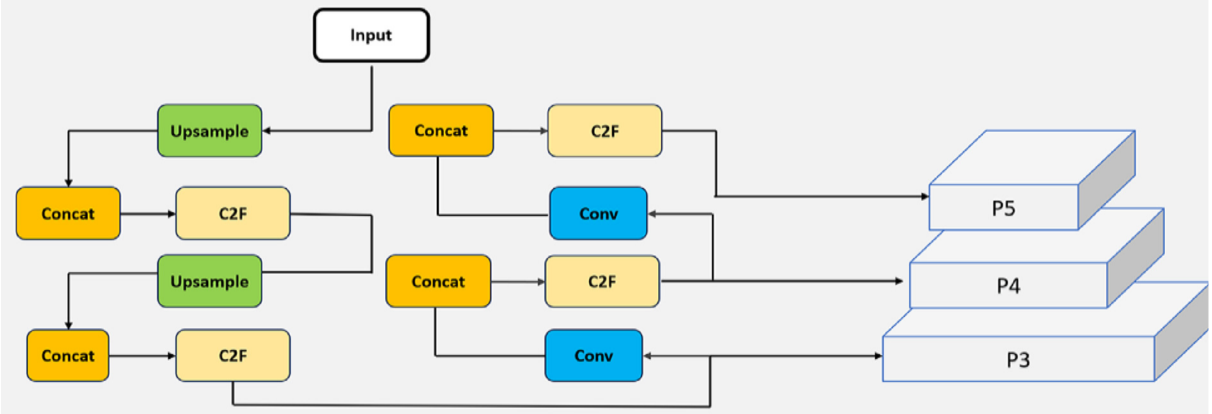

Fig. 4. The head structure with the extended shallow C2F.

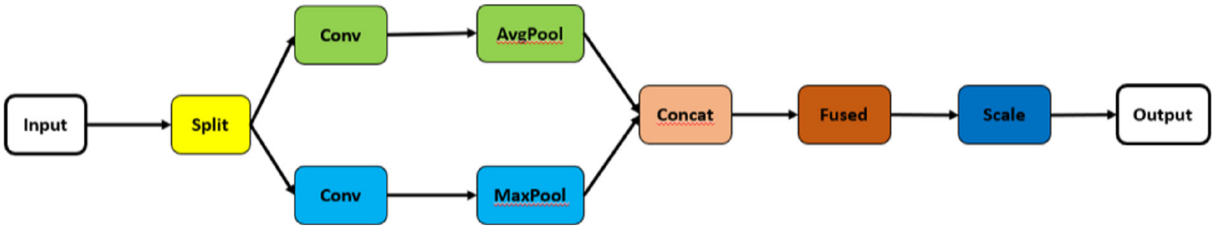

Fig. 5. Improved spatial attention.

detection, with each image accompanied by bounding boxes outlining the objects. The dataset contains over 330,000 images, each meticulously annotated with 80 distinct object categories and five descriptive captions depicting the scene. Approximately 15,000 images in this dataset predominantly have a resolution of 100 × 600 pixels. However, the COCO dataset still has some limitations that dramatically affect some object detection. Among these limitations:

1. The limit of object categories contained only 80 object categories. This number represents very limited compared with countless other categories in the real world. 2. Like any other large-scale dataset, COCO is vulnerable to annotation errors. Although this error is relatively small, it sometimes affects both training and evaluation.3. Some classes in the COCO dataset have many more instances than others. This imbalance in class distributions represents a challenge for models to perform equally across all categories. However, the COCO dataset is the best choice compared to the other datasets.

To ensure an unbiased representation, the original COCO dataset has been selected as the source for both positive and negative samples without any modifications. A total of 17,498 consecutive image samples were intercepted for training and testing the large-scale original data. The experimental findings demonstrate that the provided data are adequate for effectively training and evaluating the model, resulting in a notable level of accuracy in detecting the desired outcomes.

4.2. Experimental setup

The novel YOLOv8-CAB algorithm uses multiple evaluation metrics and hyperparameters, as seen in the following:

1. The input image dimensions are set to 640 × 640. The main reason for choosing this dimension is to maintain the balance between computational efficiency and image details for detection. In other words, high-quality images would improve the detection accuracy but at a substantial computational cost. On the other hand, using low-quality images may cause the loss of vital features in each image. The specified resolution has been experimentally found to produce significant results in prior YOLO versions.
2. The training process lasts for 300 epochs since training on this number of epochs allows the model to learn essential features without overfitting the training data.
3. In every cycle, a group of 32 samples has been used. This batch size is widely accepted in deep learning practices since it provides a good balance between model update frequency and computational feasibility. Furthermore, this batch size ensures adequate GPU memory utilization without causing memory overflow.
4. Initially, the learning rate is set to 0.001 (representing a standard choice that allows the model to cover at moderate speed without potentially skipping over optimal solutions or wasting computational resources). After every five cycles, it decreases by a factor of 0.01. The intersection over the union threshold is set at 0.20. Furthermore, the specified values for momentum and weight decay are set at 0.937 and 0.0005, respectively. The experiments were conducted using the Google Colab PyTorch platform and executed on a GPU GeForce V100 18 GB. The time consumption for training samples is about five hours, and the inference speed is 200 fps.
5. The YOLOv8-CAB algorithm strikes an impressive balance between accuracy and computational efficiency in object detection. It incorporates a powerful backbone network enhanced with the SK attention mechanism, significantly sharpening the focus on critical features, which may improve detection accuracy while adding to the algorithm's complexity.
6. Computational Requirements: Operating at 33.8 GFLOPs, the model provides quick and precise object detection capabilities.
7. Memory Usage: Consuming 57 GB of memory reflects the model's robust data handling and precision.

8. Training Time: Efficiently condensed to just over two minutes per epoch, the complete training duration is approximately eleven hours, exemplifying the model's efficient learning process.
9. Inference Time: An impressive inference speed of 50 ms is achieved, enabling rapid object detection crucial for real-time application scenarios.

As mentioned before, the novel algorithm was trained and tested using the COCO dataset to enhance its performance at each stage. A comparative analysis was conducted between the new algorithm and the baseline YOLOv8. A verification process was necessary to ascertain the potential enhancement of this algorithm in detecting small-size targets while maintaining accuracy across other scales.

A series of comparative experiments were conducted on the COCO dataset to evaluate the effectiveness of the proposed targets. A selection of intricate scene images in various scenarios was used to compare the detection capabilities between the YOLOv8-CAB algorithm and the YOLOv8 algorithm in real-life settings.

4.3. Valuation index

For evaluation, various metrics were integrated, including mAP, recall (R), average precision (AP), and precision (P). Precision and Recall (R) are represented by Equations (1) and (2), respectively.

$$P = \frac{TP}{TP + FP} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

where TP represents the number of accurately predicted bounding boxes, FP represents the number of positive samples, and FN represents the number of false positives.

Average precision (AP) evaluates the accuracy exhibited by a given model, and mAP is the average of AP values across different categories. The variable k is used to denote the number of categories. Equations (3) and (4) represent the formulas for calculating AP and mAP, respectively.

$$AP = \int_{1}^{0} p(r)dr \tag{3}$$

$$mAP = \frac{1}{C}\sum_{K=1}^{N} P(K)\Delta R(K) \tag{4}$$

4.4. Experimental result analysis

A series of ablation experiments were performed using the COCO dataset to evaluate the efficacy of the proposed enhanced approach for detecting small-sized targets. A comparative analysis was then conducted between the outcomes derived from the implementation of YOLOv8 and the results of this study. The COCO dataset exhibits extensive coverage, including diverse scenes, weather conditions, and lighting variations, making it suitable for testing small-sized targets in intricate environments. The dataset contains various attributes, such as scene visibility, object classification, and occlusion, making it comprehensive and authoritative for this experiment. The dataset was used for control experiments to ensure the credibility of the results.

To effectively demonstrate the experiment's credibility, the mAP0.5 and mAP0.5:0.9 evaluation indices were used in this investigation. Also, the (Average Precision for Small Objects (AP^{small}) measures a model's accuracy in detecting small objects, focusing on precision for smaller-scale items in an image dataset. While the average Recall for Small Objects (AR^{small}). Table 1 displays the test results.

The findings in Table 1 indicate a discernible enhancement in the performance of the enhanced algorithm at each stage, particularly in detecting small-scale targets within intricate scenes. Furthermore, a noteworthy enhancement of 2.1 % is observed in the recall rate, indicating a substantial scope for further improvement. The three improved methods in this experiment demonstrate the potential for performance enhancement, especially when replacing C2f in the backbone with the CAB block, efficiently capturing both local and global contexts. Increasing the number of layers in C2f resulted in enhanced feature extraction capabilities. Moreover, incorporating SK to enhance spatial attention has improved the algorithm's performance.

The adjustments, including multi-scale spatial modeling and feature weighting (fuse and scale steps), offer enhanced flexibility and selective

Table 1. Comparison of metrics at each modification.

Dataset	Result	YOLO5	YOLO7	YOLO8	Yolo-CAB
COCO	mAP0.5 %	46.2	43.1	46.2	47.1
	mAP0.5:0.95 %	27.2	21.9	27.2	28.2
	AP^{small}%	16	22.9	48.0	49.0
	AR^{small}%	19.3	24.3	65.2	66.5
VOC	mAP0.5 %	40.3	32.8	41.2	41.5
	mAP0.5:0.95 %	23.4	20.3	25.8	26.7
	AP^{small}%	64.8	70.6	70.5	71.3
	AR^{small}%	67.2	73.1	73.4	74.2

emphasis on substantial spatial regions and channels. The empirical findings show that enhancing the algorithm at each iteration improves learning capacity. To evaluate the detection performance of the YOLOv8-CAB algorithm, we measured the mAP for ten different object categories within the COCO dataset, as shown in Fig. 6.

The results indicate that the YOLOv8-CAB algorithm outperforms the baseline YOLOv8 in detecting various object categories. Notably, four distinct categories are observed where the recognition accuracy surpasses the overall average of the dataset.

The modified algorithm consistently enhances the detection capabilities. It shows notable advancements in detecting larger objects, such as automobiles, and smaller objects, including cups, bowls, televisions, felines, remote controls, rodents, oranges, and similar items.

Table 2 shows comprehensively compares performance metrics across various versions of the YOLO algorithm.

An analysis was conducted to examine the factors contributing to the superior performance of the YOLOv8-CAB algorithm. The main finding of this experiment can be summarized in the following points:

a) One significant challenge conventional methods (such as FPN + PAN) face is the difficulty in layer-by-layer feature extraction. Often, targets with reduced dimensions can be confused with objects of average dimensions, resulting in substantial information loss. In contrast, YOLOv8-CAB's feature fusion technique seamlessly integrates shallow information into the final output. This approach reduces information loss in the shallower layers.
b) During the feature extraction process, YOLOv8-CAB disregards irrelevant information. It retains features extracted from pixels of small-sized targets, leading to enhanced accuracy.
c) The C2f modification in the YOLOv8-CAB model augments its depth, facilitating deeper knowledge acquisition and improved feature extraction capabilities.
d) Incorporating SK attention into the YOLOv8-CAB model enhances spatial attention. As a result, the model demonstrates increased efficiency in capturing features of smaller objects.
e) In some cases, the COCO cannot provide enough samples for some small objects, representing the main challenges to detection performance. Therefore, some strategies have been employed, such as data augmentation and Feature Pyramid Network (FPN). In this work, YOLOv8-CAB has adopted the same augmentation techniques used in YOLOv8 and YOLOv5 architecture. The augmentation technique, known as Mosaic augmentation, proves its performance in the model's generalization and makes it robust to various real-world conditions.
f) Although trained in high-quality platform hardware, YOLOv8-CAB is optimized for diverse hardware, ensuring broad compatibility. Considerations for real-time applications are addressed without compromising speed or accuracy. However, it is recommended to use high hardware requirements to prevent lag detection for small objects in real-time.
g) YOLOv8-CAB shows a significant increase in detection performance compared with conventional YOLOv8. Furthermore, when compared with other state-of-the-art techniques, YOLOv8-CAP shows considerable improvement, too. Table 3 demonstrates the mAP of YOLOv8-CAP with Nanodet techniques [39]. It is obvious that YOLOv8-CAP has high-accuracy detection with

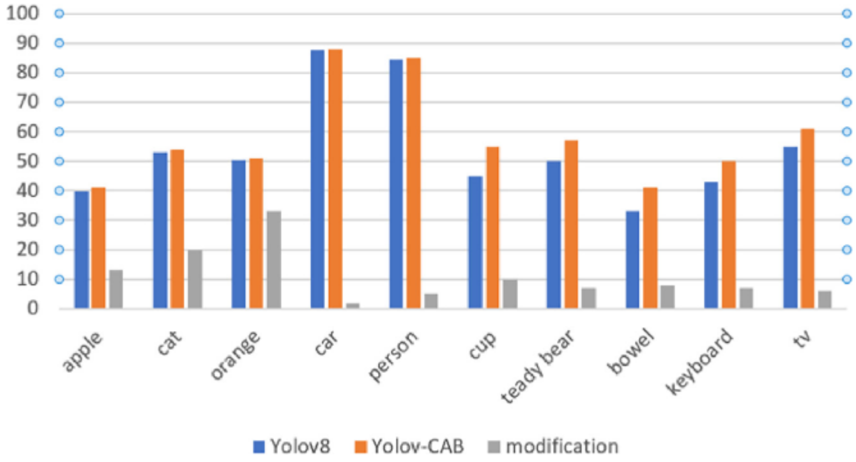

Fig. 6. A comparison of 10 small object classes between Yolov8 and Yolov8-CAB.

Table 2. Performance of different YOLO versions.

Algorithm	Module			Metrics Results				AP^{small}%	AR^{small}%	F1%
	CAB	SA	Modified-C2f	mAP:0.5 %	mAP:0.5:0.95 %	P%	R%			
Yolov8	X			46.2	27.2	88.4	63.2	48.0	65.2	73.82
Yolov8-CAB		X		46.7	27.6	88.9	64	48.5	66.7	74.41
Yolov8-CAB			X	47.4	28	89.3	64.7	49.0	66.5	75.07

Table 3. Comparison between YOLOv8-CAP with nanodet.

Algorithm	Metrics Results				AP^{small}%	AR^{small}%	F1 Score%
	mAP:0.5 %	mAP:0.5:0.95 %	P%	R%			
Nanodet	39.5	Never used	66	61	37.5	59.5	71
Yolov8	47.4	28	89.3	64.7	45	62.5	75.07

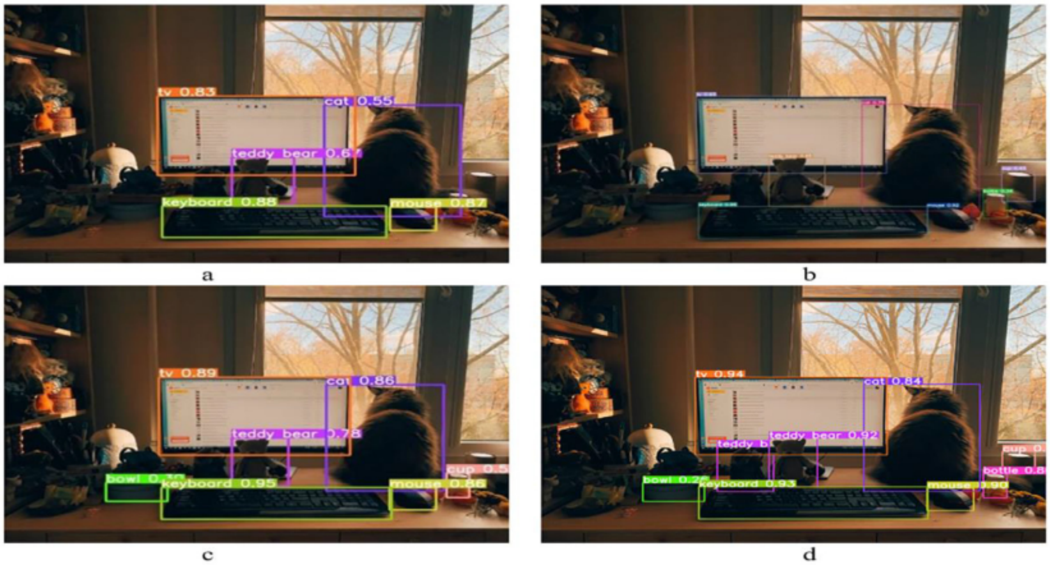

Fig. 7. (a) Illustrates that YOLOv5 exhibited limitations in detecting small objects in the image, such as the bottle and the cup on the right side, and YOLOv5 did not detect the bowl due to the darkness on the left of the image. (b) shows that YOLOv7 also missed the bowl for the same reason as YOLOv5. (c) shows that YOLOv8 has better detection confidence and detects more objects, but it missed the cup on the right side of the image. By contrast, YOLOv8-CAB, as shown in (d), successfully detects all objects, achieving higher accuracy and much better detection confidence, demonstrating its outstanding performance in challenging scenarios.

0.5 mAP compared with only 66 for the nano det technique.

h) The selection process for the comparison experiments involved choosing images from the COCO dataset that exhibit complex scenes, higher levels of interference, and substantial overlap. Fig. 7 shows the detection result comparison of various YOLO versions.

5. Conclusion

In this study, YOLOv8-CAB, an optimized detection technique tailored explicitly for small objects, has been introduced. Based on the foundation of YOLOv8, this method enhances feature utilization by iterating the feature extraction network's shallow C2F module and adds a better attention module to the remaining blocks. These adjustments compel the network to learn robust and distinctive features, proving highly effective in detecting small objects in images.

The proposed model shows a unique ability to identify diminutive entities within images, irrespective of their diminished resolution and indistinct characteristics. The results show that using CNN image preprocessing of different images can further enhance the accuracy of detecting small objects.

Although our focus has been primarily on still and complex images, the outstanding performance of YOLOv8-CAB sets a promising premise for future exploration in video object detection, given the strong correlation between video sequence frames. We anticipate even better detection performance in such scenarios. In summary, the YOLOv8-CAB model signifies a notable advancement in the domain of small object detection, exhibiting considerable potential for forthcoming applications.

Conflict of interest

There is no conflict of interest.

Acknowledgments

I want to thank everyone who contributed to building the future and worked hard to develop it.

References

[1] M. Khalaf, B.N. Dhannoon, Skin lesion segmentation based on U-shaped network, Karbala Int J Mod Sci 8 (2022) 493–502, https://doi.org/10.33640/2405-609X.3248.

[2] R.M. Aziz, A. Hussain, P. Sharma, P. Kumar, Machine learning-based soft computing regression analysis approach for crime data prediction, Karbala Int J Mod Sci 8 (2022) 1–19, https://doi.org/10.33640/2405-609X.3197.

[3] S.K. Behera, A.K. Rath, P.K. Sethy, Fruit recognition using support vector machine based on deep features, Karbala Int J Mod Sci 6 (2020) 235–245, https://doi.org/10.33640/2405-609X.1675.

[4] E.A. Abbood, T.A. Al-Assadi, GLCMs based multi-inputs 1D CNN deep learning neural network for COVID-19 texture feature extraction and classification, Karbala Int J Mod Sci 8 (2022) 28–39, https://doi.org/10.33640/2405-609X.3201.

[5] A.M. Bilal, M.B. Kurdy, Age-invariant face recognition using trigonometric central features, Karbala Int J Mod Sci 5 (2019) 7, https://doi.org/10.33640/2405-609X.1209.

[6] W. Li, Y. Zhu, D. Zhao, Missile guidance with assisted deep reinforcement learning for head-on interception of maneuvering target, Compl Intell Syst 8 (2021) 1205–1216, https://doi.org/10.1007/s40747-021-00577-6.

[7] X. Xu, M. Du, H. Guo, J. Chang, X. Zhao, Lightweight FaceNet based on MobileNet, Int J Intell Sci 11 (2021) 1–16, https://doi.org/10.4236/ijis.2021.111001.

[8] J. Liu, X. Wang, Early recognition of tomato gray leaf spot disease based on MobileNetv2-YOLOv3 model, Plant Methods 16 (2020) 7, https://doi.org/10.1186/s13007-020-00624-2.

[9] M. Abd Elaziz, A. Dahou, N.A. Alsaleh, A.H. Elsheikh, A.I. Saba, M. Ahmadein, Boosting covid-19 image classification using mobilenetv3 and aquila optimizer algorithm, Entropy 23 (2021) 1383, https://doi.org/10.3390/e23111383.

[10] N. Ullah, J.A. Khan, S. El-Sappagh, N. El-Rashidy, M.S. Khan, A holistic approach to identify and classify COVID-19 from chest radiographs, ECG, and CT-scan images using ShuffleNet convolutional neural network, Diagnostics 13 (2023) 162, https://doi.org/10.3390/diagnostics13010162.

[11] R. Yang, X. Lu, J. Huang, J. Zhou, J. Jiao, Y. Liu, F. Liu, B. Su, P. Gu, A multi-source data fusion decision-making method for disease and pest detection of grape foliage based on ShuffleNet V2, Remote Sens (Basel). 13 (2021) 5102, https://doi.org/10.3390/rs13245102.

[12] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas. (2016), pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.

[13] M. Yuan, Q. Zhang, Y. Li, Y. Yan, Y. Zhu, A suspicious multi-object detection and recognition method for millimeter wave SAR security inspection images based on multi-path extraction network, Rem Sens 13 (2021) 4978, https://doi.org/10.3390/rs13244978.

[14] A.R. Pathak, M. Pandey, S. Rautaray, Application of deep learning for object detection, Procedia Comput Sci 132 (2018) 1706–1717, https://doi.org/10.1016/j.procs.2018.05.144.

[15] M. Shafiq, Z. Gu, Deep residual learning for image recognition: a survey, Appl Sci 12 (2022) 8972, https://doi.org/10.3390/app12188972.

[16] S. Dong, P. Wang, K. Abbas, A survey on deep learning and its applications, Comput Sci Rev 40 (2021) 100379, https://doi.org/10.1016/j.cosrev.2021.100379.

[17] X. Xu, M. Zhao, P. Shi, R. Ren, X. He, X. Wei, H. Yang, Crack detection and comparison study based on faster R-CNN and Mask R-CNN, Sensors 22 (2022) 1215, https://doi.org/10.3390/s22031215.

[18] J. Li, D. Zhang, J. Zhang, J. Zhang, T. Li, Y. Xia, Q. Yan, L. Xun, Facial expression recognition with faster R-CNN, Proc Comput Sci 107 (2017) 135–140, https://doi.org/10.1016/j.procs.2017.03.069.

[19] H. Nguyen, Improving faster R-CNN framework for Fast vehicle detection, Math Probl Eng (2019) 11, https://doi.org/10.1155/2019/3808064, 2019.

[20] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only Look once: unified, real-time object detection, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR),IEEE, Las Vegas, 2016, pp. 779–788, https://doi.org/10.1109/CVPR.2016.91.

[21] J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Honolulu. (2017), pp. 7263–7271, https://doi.org/10.48550/arXiv.1612.08242.

[22] C. Zhang, F. Kang, Y. Wang, An improved apple object detection method based on lightweight YOLOv4 in complex backgrounds, Rem Sens 14 (2022) 4150, https://doi.org/10.3390/rs14174150.

[23] G. Jocher, Yolov5 in PyTorch, 2020, https://doi.org/10.5281/zenodo.3908559. https://github.com/ultralytics/yolov5. (accessed August 21 2023).

[24] C.-Y. Wang, A. Bochkovskiy, H.-Y.M. Liao, C.-Y. Wang, A. Bochkovskiy, H.-Y.M. Liao, YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVF, Vancouver. (2023), pp. 7464–7475, https://doi.org/10.1109/CVPR52729.2023.00721.

[25] B. Leibe, J. Matas, N. Sebe, M. Welling, Computer Vision – ECCV, first ed., Springer International Publishing, Cham. (2016) https://doi.org/10.1007/978-3-319-46448-0.

[26] M. Maktab, M. Razaak, P. Remagnino, Enhanced single shot small object detector for aerial imagery using super-resolution feature fusion and deconvolution, Sensors 22 (2022) 4339, https://doi.org/10.3390/s22124339.

[27] G. Jocher, A. Chaurasia, J. Qiu, YOLO by Ultralytics, 2023, https://doi.org/10.5281/zenodo.3908559. https://github.com/ultralytics/ultralytics. (accessed August 21 2023).

[28] C. Xu, X. Wang, Y. Yang, Selective multi-scale feature learning by discriminative local representation, IEEE Access 7 (2019) 127327–127338, https://doi.org/10.1109/ACCESS.2019.2939716.

[29] L. Deng, H. Li, H. Liu, J. Gu, A lightweight YOLOv3 algorithm used for safety helmet detection, Sci Rep 12 (2022) 10981, https://doi.org/10.1038/s41598-022-15272-w.

[30] K.-Y. Jeng, Y.-C. Liu, Z.Y. Liu, J.-W. Wang, Y.-L. Chang, H.-T. Su, W.H. Hsu, A coarse-to-fine (C2F) representation for end-to-end 6-DoF grasp detection, in: Conference on Robot Learning, Rob, Lear. Foun. Inc. London. 2020 224814237, https://doi.org/10.48550/arXiv.2010.10695.

[31] J. Hu, L. Shen, G. Sun, Squeeze-and-Excitation networks, in: IEEE/CVF Conference on Computer Vision and Pattern

Recognition, IEEE, Salt Lake City. 2018, pp. 7132—7141, https://doi.org/10.1109/CVPR.2018.00745.

[32] S. Jain, S. Dash, R. Deorari, Kavita, object detection using coco dataset, in: International Conference on Cyber Resilience (ICCR), IEEE, Dubai. 2022, pp. 1—4, https://doi.org/10.1109/ICCR56254.2022.9995808.

[33] X. Li, W. Wang, X. Hu, J. Yang, Selective Kernel networks, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Long Beach, 2019, pp. 510—519, https://doi.org/10.1109/CVPR.2019.00060.

[34] A. Wong, M.J. Shafiee, F. Li, B. Chwyl, Tiny SSD: a tiny single-shot detection deep convolutional neural network for real-time embedded object detection, in: 2018 15th Conference on Computer and Robot Vision (CRV), IEEE, Toronto. 2018, pp. 95—101, https://doi.org/10.1109/CRV.2018.00023.

[35] L. Ali, F. Alnajjar, M. Parambil, M. Younes, Z. Abdelhalim, H. Aljassmi, Development of YOLOv5-based real-time smart monitoring system for increasing lab safety awareness in educational institutions, Sensors 22 (2022) 8820, https://doi.org/10.3390/s22228820.

[36] H. Liu, F. Sun, J. Gu, L. Deng, SF-YOLOv5, A lightweight small object detection algorithm based on improved feature fusion mode, Sensors 22 (2022) 5817, https://doi.org/10.3390/s22155817.

[37] C. Xu, J. Wang, W. Yang, H. Yu, L. Yu, G. Xia, Detecting tiny objects in aerial images: a normalized Wasserstein distance and a new benchmark, ISPRS J Photogramm Rem Sens 190 (2022) 79—93, https://doi.org/10.1016/j.isprsjprs.2022.06.002.

[38] M. Kim, H. Kim, J. Sung, C. Park, J. Paik, High-resolution processing and sigmoid fusion modules for efficient detection of small objects in an embedded system, Sci Rep 13 (2023) 244, https://doi.org/10.1038/s41598-022-27189-5.

[39] P. Yong, S. Li, K. Wang, Y. Zhu, A real-time detection algorithm based on Nanodet for pavement cracks by incorporating attention mechanism, in: 8th International Conference on Hydraulic and Civil Engineering, Inst. Electr. Electron. Eng. Inc. Xi'an, 2022, pp. 1245—1250, https://doi.org/10.1109/ICHCE57331.2022.10042517.

[40] B. Jiang, S. Chen, B. Wang, B. Luo, MGLNN: semi-supervised learning via multiple graph cooperative learning neural networks, Neural Network 153 (2022) 204—214, https://doi.org/10.1016/j.neunet.2022.05.024.

[41] A.M. Roy, J. Bhaduri, DenseSPH-YOLOv5: An automated damage detection model based on DenseNet and Swin-Transformer prediction head-enabled YOLOv5 with attention mechanism, Adv Eng Inf 56 (2023) 102007, https://doi.org/10.1016/j.aei.2023.

[42] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, Columbus, 2014, pp. 580—587, https://doi.org/10.1109/CVPR.2014.81.