

SMALL OBJECT DETECTION APPROACH BASED ON ENHANCED SINGLE-SHOT DETECTOR FOR DETECTION AND RECOGNITION OF INDONESIAN TRAFFIC SIGNS

Phie Chyan^{1*}, Norbertus Tri Suswanto Saptadi², Jeremias Mathias Leda³

^{1,2} Department of Informatics, Faculty of Information Technology, Universitas Atma Jaya Makassar

³ Department of Electrical Engineering, Faculty of Engineering, Universitas Atma Jaya Makassar
Jln. Tanjung Alang, No 23, Makassar, 90134, Indonesia

Corresponding author's e-mail: * phiechyan@gmail.com

ABSTRACT

Article History:

Received: 24th, May 2024

Revised: 10th, July 2024

Accepted: 15th, August 2024

Published: 14th, October 2024

Keywords:

Small Object Detection;

Enhanced SSD;

Traffic Signs.

The detection and recognition of traffic signs are crucial components of advanced driving assistance systems (ADAS) that enhance road safety. Current traffic sign detection and recognition model technology is proficient in identifying and interpreting traffic signs. However, for accurate detection and recognition, the traffic sign in the image must be of a certain minimum pixel size or distance from the driver's sight line for proper detection. The ADAS system should be capable of detecting and recognizing road traffic signs from a considerable distance as they come into the driver's line of vision. The higher the vehicle speed, the greater the distance required for the sign to be detected and recognized, allowing the driver sufficient time to react according to the sign's meaning. Addressing these challenges, this research proposes an enhanced version of the single shot detector (SSD) algorithm, commonly used in object detection, to improve the algorithm's ability to detect small objects. The proposed method involves adding an auxiliary layer module to the original SSD architecture to increase the feature map resolution and expand the conventional layer's receptive space. With the Enhanced SSD algorithm, the detection capability of the SSD can be significantly enhanced in terms of accuracy. The limitations of this study are related to the influence of occlusion and clutter, which might affect the performance of object detection, especially for small objects, which are more susceptible to being influenced by various factors. The research results demonstrate that Enhanced SSD improves object detection accuracy compared to the original SSD, with a mean average precision (mAP) of 97.87 compared to 95.35 for detecting 21 traffic signs in Indonesia.



This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-sa/4.0/).

How to cite this article:

P. Chyan, N. T. S. Saptadi and J. M. Leda., "SMALL OBJECT DETECTION APPROACH BASED ON ENHANCED SINGLE-SHOT DETECTOR FOR DETECTION AND RECOGNITION OF INDONESIAN TRAFFIC SIGNS," *BAREKENG: J. Math. & App.*, vol. 18, iss. 4, pp. 2653-2662, December, 2024.

Copyright © 2024 Author(s)

Journal homepage: <https://ojs3.unpatti.ac.id/index.php/barekeng/>

Journal e-mail: barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id

[Research Article](#) · [Open Access](#)

1. INTRODUCTION

Traffic signs are a type of road equipment that informs road users about the rules and instructions needed to achieve order and security on the road [1]. Identifying and recognizing road traffic signs are crucial components incorporated into the advanced driving assistance system (ADAS) module in contemporary vehicles. This ensures it functions effectively and optimally to aid drivers while navigating roads [2].

Traffic signs, especially in Indonesia, are broadly divided into three categories: prohibitions, instructions, and warnings [3]. These signs are generally designed using specific shapes like circles, triangles, and squares. They are given particular colors that state the type of sign, such as red for prohibition, yellow for warning, and blue for instructions [4].

Identifying and acknowledging road traffic signs are crucial aspects integrated into the advanced driving assistance system (ADAS) module in modern vehicles. This is essential for its practical and optimal performance in supporting drivers while navigating roads. Researchers from various countries, including Indonesia, have conducted studies on this topic, resulting in the evolution of accurate detection and recognition models for traffic signs [1], [2], [4]–[7]. However, the minimum pixel size needed for accurately detecting and recognizing image traffic signs remains unresolved. Currently, existing models that originated from previous research, despite their high accuracy, can only identify signs when they reach a certain size in the image. This poses a problem when the signs are at a distance from the driver, as it may result in delayed or unsafe reactions. ADAS systems must detect and recognize traffic signs from a considerable distance, allowing drivers ample time to react appropriately based on the sign's meaning, especially at higher speeds.

This research aims to create a small object detection model for traffic sign detection and recognition. This model is designed to identify objects with tiny dimensions in images. The developed model is an improved version of the single shot detector (Enhanced-SSD) algorithm, known for its high accuracy and speed in object detection. Auxiliary modules have been integrated to enhance its perception of tiny objects and improve the algorithm's ability to detect small objects accurately. Implementing this method will enhance the performance of traffic sign detection and recognition and allow for signs to be detected from a greater distance as they enter the driver's field of view.

2. RESEARCH METHODS

Object detection algorithms based on convolutional neural networks can be divided into two main types. The first type is a single-stage algorithm, which uses convolution networks like YOLO and SSD (single shot detector) to directly predict the categories and positions of objects in the input image. The second type is a region proposal-based algorithm, also known as a two-stage algorithm. This algorithm starts by generating a region proposal through a region proposal network and then proceeds to classify the objects based on the region proposal [8]. Object detection results are obtained after these two stages. Examples of two-stage algorithms include Faster RCNN (faster region-based convolutional neural network). While two-stage algorithms are usually more flexible and accurate, single stage algorithms like SSD have the advantage of faster processing speed, up to 7 times faster than two-stage algorithms like Faster RCNN [9]. This characteristic is particularly suitable for real-time applications such as ADAS systems requiring fast processing [10]. To detect and recognize small objects, such as traffic signs, we utilize a modified version of SSD called Enhanced-SSD to increase the accuracy of the SSD algorithm. According to the literature, a small object can be defined as an object with a maximum resolution dimension of one-tenth of the image resolution dimension, or an alternative definition, as an object whose pixel dimensions are less than 32 x 32. Detecting small objects is challenging due to their characteristics, including being unrecognizable, low resolution, set against a complex background, and having limited context information [11]. By using Enhanced-SSD to detect small objects, we achieved a better trade-off between accuracy and faster processing than the original SSD and two-stage algorithm. The following sub-section will discuss the details of the Enhanced-SSD model.

2.1 SSD Model Architecture

The SSD method employs a feed-forward convolutional network to produce a series of bounding boxes and determine the presence of an object class within each box [12]. Afterward, a non-maximum suppression process is applied to generate the ultimate detection. The layers of the original SSD consist of a network layer, which is the basic architecture for processing high-quality images, called the base layer. In addition to the base layer, an extra layer structure supports object detection, known as the feature layer for multi-scale feature maps. This layer is a convolutional feature layer added at the end of the base layer. These layers progressively decrease in size, enabling detection and prediction at various scales. Each additional feature layer can produce a series of fixed predictions using convolutional filters, as demonstrated in the original SSD architecture in **Figure 1**. The prediction of parameters for potential detection in an $m \times n$ feature layer with x channels is based on a $3 \times 3 \times p$ kernel, which generates scores for categories or offset shapes about the coordinates of the bounding boxes. As convolutional layers progressively decrease in spatial dimension, the feature map resolution is also reduced. SSD utilizes a low-resolution layer to detect large objects and, conversely, uses a high-resolution layer to detect small-scale objects. For instance, a 4×4 feature map is utilized to detect large objects, and an 8×8 feature map is utilized to detect smaller objects, as shown in **Figure 2**. SSD adds 6 convolutional layers after VGG16, with 5 used for object detection. In total, the SSD can make 8732 predictions using these 6 layers.

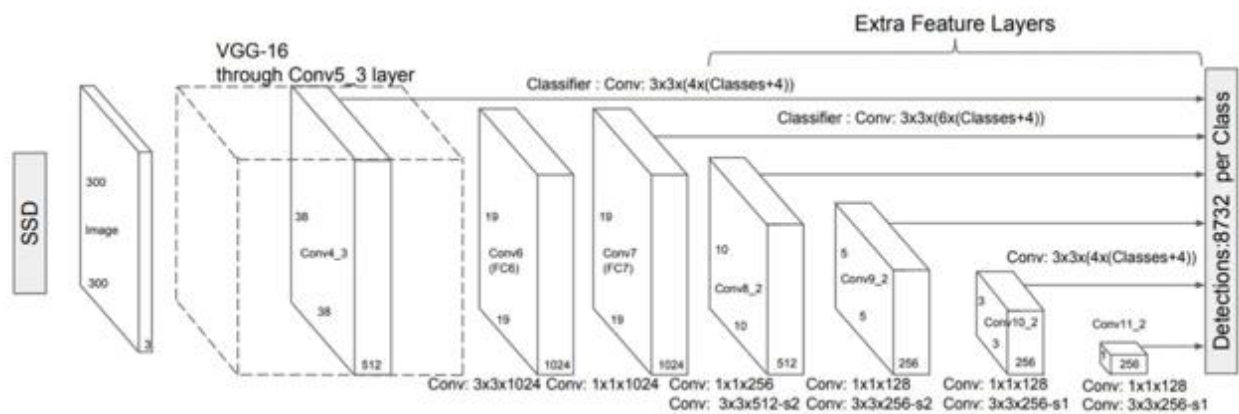


Figure 1. The SSD Model Architecture

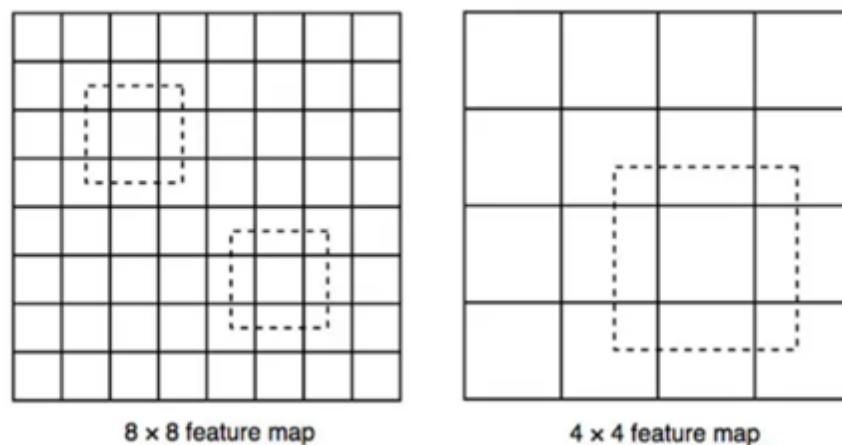


Figure 2. Higher resolution feature map (left) and lower resolution feature map (right)

2.2 SSD Enhanced Module

In the original SSD architecture, as discussed in section 2.1, the multi-scale feature map is utilized directly to detect objects regardless of the scale of the detected objects [13]. The weakness of this method

causes SSD to have shortcomings in capturing detailed local and global semantic features. This causes the SSD's ability to detect objects of various sizes unequal, especially for small objects. To be able to detect small objects well, the detector must be able to combine context information and detailed features. For this reason, a slight change was made to the architecture so that the SSD does not directly use the feature map from the prediction layer to detect objects but waits until the context layer fusion results are obtained using a particular auxiliary module, as illustrated in **Figure 3**.

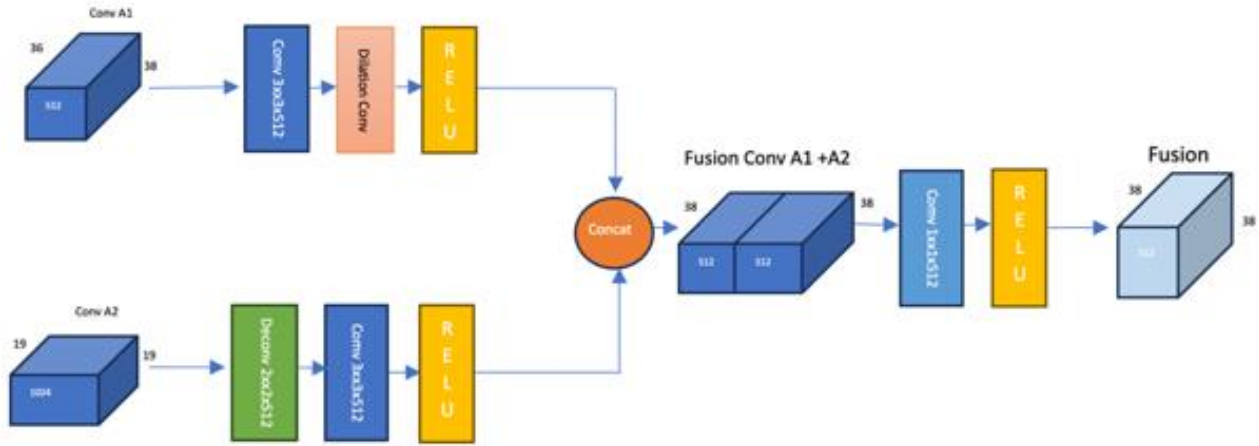


Figure 3. Auxiliary module for Enhanced SSD

The auxiliary module aims to expand the SSD module by adding additional layers. Layer that will increase the feature map resolution and develop the receptive space of conventional VGG. The features from the two auxiliary modules are then concatenated to replace those from the Conv 5_3 layer. Assuming $x_i, i \in P$ is a multi-scale feature for detection, the fusion module feature can be expressed as following **Equation (1)** and **Equation (2)**:

$$x_f = \mathcal{C} \{ \mathcal{D}(x_{Conf5_3}), \mathcal{T}(x_{fc7}) \} \quad (1)$$

$$loc = class = \phi_{c,l}(x_p) \quad p \in \mathcal{P} \quad (2)$$

Where \mathcal{D} is the first aux module for the x_{Conf5_3} feature of the conf5_3 layer and \mathcal{T} is the function of the second aux module which carries out the deconvolution process. The next is \mathcal{C} , which states the concatenation function $\phi_{c,l}$ and is a method for predicting objects from feature maps. The first aux module \mathcal{D} which performs the convolution process, is usually used in image segmentation [14], [15] to obtain context information from various image dimensions. \mathcal{T} is used to integrate semantic information for small object detection in shallow layers, so it uses feature maps from higher layers and combines them into shallower layers. However, because the dimension of the feature map is different, an upsample map is needed for each layer to ensure the size of each feature map is the same.

2.3 Training

The main difference between detection models based on single-stage detectors (including enhanced SSD) and two-stage detector models that use region proposals such as Faster-RCNN is that ground-truth information must be provided to an output in a series of output detectors [16]. After the provisioning process is carried out, the loss and backpropagation functions are then implemented at both input and output points. Training also requires choosing a series of standard boxes and scales for detection and utilizing the data augmentation method.

During training, it is necessary to determine the default box corresponding to the annotated object label and then train according to this configuration. For each ground truth, default boxes will be selected, and these will vary in size, location, and aspect ratio. The matching strategy will be carried out by matching default

boxes with each label annotated with a certain threshold. This strategy will simplify the learning problem and allow the network to predict multiple overlapping default boxes.

SSD's Training Objective is a derivative of the MultiBox objective but with the addition of supporting multiple object categories [17]. For example, $x_{ij}^p = \{1,0\}$ represents whether the default box i matches the ground truth box from category p . In the matching approach described above, it is possible to have $\sum_i x_{ij}^p > 1$. The combined loss function aims to minimize both the localization loss (loc) and the confidence loss ($conf$) through weighted summation, which is expressed in Equation (3)

$$L(x, c, l, g) = \frac{1}{N} (L_{conf}(x, c) + \alpha L_{loc}(x, l, g)) \quad (3)$$










Where N is the count of default matching boxes. If $N = 0$, then the loss is set to 0. Localization loss refers to the discrepancy between the parameters of the predicted bounding box (l) and the actual ground truth (g). Confidence loss, on the other hand, is the loss calculated using the softmax function across various classes (c), with the weight factor α being determined as 1 through the process of cross-validation.





3. RESULTS AND DISCUSSION

3.1 Dataset

The dataset used in this research is the Indonesian Traffic Sign Dataset [3], which consists of applicable traffic signs in Indonesia. These traffic signs are divided into four groups: prohibition signs, warning signs, command signs, guidance signs, and traffic lights. There are a total of 21 types of signs in the dataset, with details provided in Table 1 below:

Table 1. Indonesian Traffic Sign Dataset

No	Group Category	Road Sign	Sample Count	Sample Image
1	Prohibition sign	No parking sign	100	
		No stopping sign	100	
		Do not enter sign	100	
		No U-turn sign	100	
		No right turn sign	100	
		No left turn sign	100	
		No through road sign	100	
2	Warning sign	Pedestrian crossing warning sign	100	
		Traffic light ahead sign	100	

No	Group Category	Road Sign	Sample Count	Sample Image
		Railroad crossing warning sign	100	
		Left T intersection warning sign	100	
		Lane ends ahead warning sign	100	
3	Mandatory sign	Keep left sign	100	
		Lane choice control sign	100	
4	Direction sign	U-turn location sign	100	
		Pedestrian crossing sign	100	
		Bus stop sign	100	
		Parking area sign	100	
5	Traffic light	Red light	100	
		Yellow light	100	
		Green light	100	

The dataset comprises 2100 images, divided into 21 categories, with 100 images for each. The sign dataset was split into a 70:30 ratio for training and validation, resulting in 1470 images for training and 630 images for validation. The dataset used has been annotated, and the annotation results are stored in a text file named according to the image file. The data in the text file includes the index number of the sign type, the coordinates of the bounding box where the sign object is located, and the size of the bounding box.

3.2 Model Training

The training machine runs on the Windows 10 Operating System with an Intel Core i7 11800H processor, 32GB of RAM, and an NVIDIA RTX 3070Ti GPU with 12GB of VRAM. The training procedures are based on the original SSD with adjustments to support new additional layers for enhanced SSD models. A batch size of 16 and an input size of 300 x 300 are used. The training begins with a learning rate of 10^{-3} for the first 20,000 iterations, then gradually reduces to 10^{-4} and 10^{-5} at the 40,000 and 60,000 iterations using a stochastic gradient descent (SGD) optimizer with weight_decay of 0.0005. Stochastic Gradient Descent (SGD) is a popular optimization algorithm used in training deep learning models. It is a variant of the traditional gradient descent algorithm designed to optimize the model's weights by minimizing the loss function. To evaluate the performance increase from the Enhanced SSD, we compare the original SSD algorithm by detecting the accuracy of the two algorithms for each sign class. The performance evaluation metric is Average Precision (AP), which refers to the area enclosed by the Precision-Recall curve. Precision measures how many predicted objects are relevant (correctly identified). It is calculated as the ratio of true

positive detections to the total number of predicted detections. Meanwhile, recall measures how many of the actual objects were correctly predicted. It is calculated as the ratio of true positive detections to the total number of ground truth objects. In practical applications, AP is not calculated directly but is based on a smoothed accuracy value obtained from a precision-recall curve. The calculation formula follows **Equation (4)** and **Equation (5)**. Mean Average Precision (mAP) is the average of AP across all classes in the dataset, and it provides an overall measure of the accuracy and robustness of an object detection model across different object categories. Higher mAP values indicate better overall performance of the model.

$$AP = \int_0^1 P_{smooth}(r) dr \quad (4)$$

$$P_{Smooth}(r) = \max_{r' \geq r} P(r') \quad (5)$$

3.3 Experiment Result

The experiments conducted with enhanced SSD yielded a Mean Average Precision (mAP) of 97.87% for all sign classes, surpassing the original SSD algorithm, which achieved a mAP of 95.35%. Generally, both algorithms showed similar accuracy for classes with large objects; however, enhanced SSD outperformed the original SSD in detecting classes with small objects, such as identifying the color of traffic lights from a distance. **Table 2** compares the average accuracy for each sign class using enhanced SSD and original SSD.

Table 2. Original SSD vs Enhanced SSD performance result

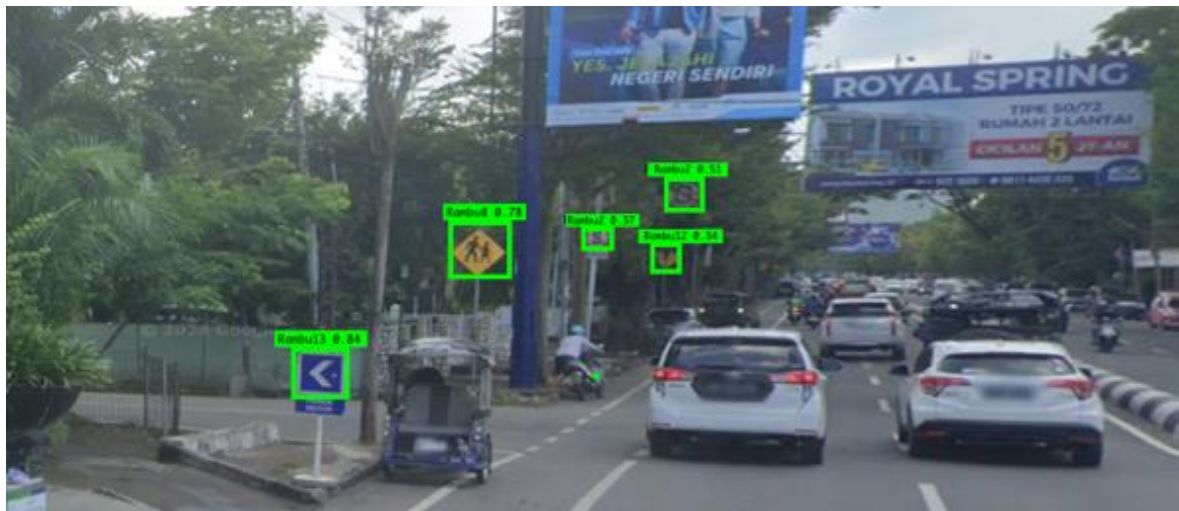
No	Road Sign	AP(%)	
		Original SSD	Enhanced SSD
1	No parking sign	88.56	95.56
2	No stopping sign	96.67	97.05
3	Do not enter sign	100	100
4	No U-turn sign	96.45	95.32
5	No right turn sign	92.66	96.18
6	No left turn sign	98.16	97.74
7	No through road sign	94.03	93.39
8	Pedestrian crossing warning sign	100	100
9	Traffic light ahead sign	100	100
10	Railroad crossing warning sign	100	100
11	Left T intersection warning sign	100	100
12	Lane ends ahead warning sign	100	100
13	Keep left sign	100	98.86
14	Lane choice control sign	96.67	97.08
15	U-turn location sign	96.18	97.33
16	Pedestrian crossing sign	100	100
17	Bus stop sign	100	100
18	Parking area sign	100	100
19	Red light	76.26	93.65
20	Yellow light	73.78	94.18
21	Green light	93.85	98.83
mAP(%)		95,35	97,87

3.4 Model Implementation Result

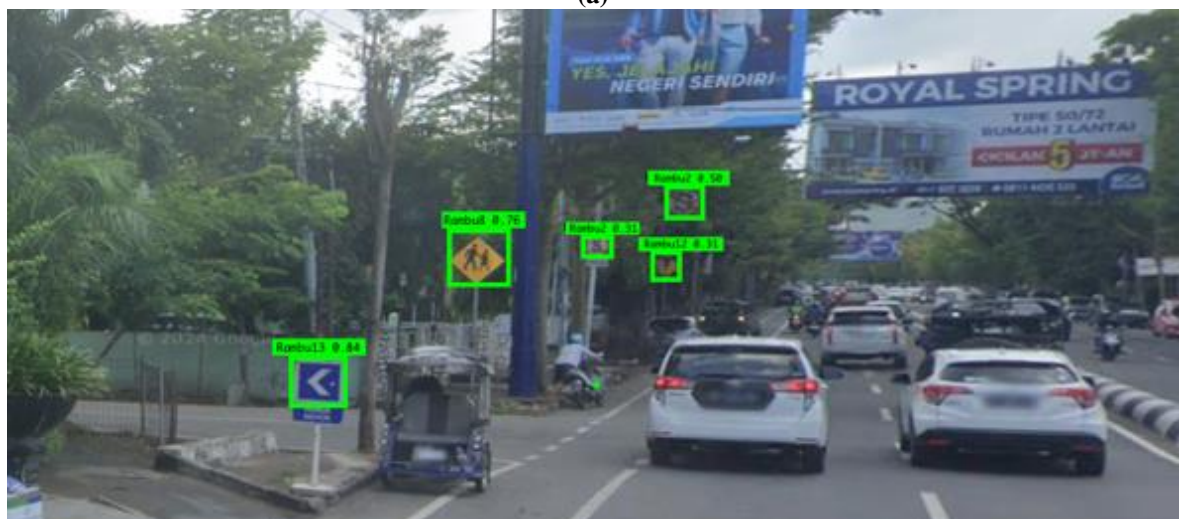
The results of direct testing of images using the original SSD and enhanced SSD algorithms are displayed in **Figure 4**. The comparison revealed that the improved SSD algorithm effectively detects small objects with high accuracy and confidence, outperforming the original SSD. This means that detection errors or failures will be further reduced with the enhanced SSD.

The Enhanced SSD architecture, an improvement over the original SSD, has successfully increased the detection accuracy of the SSD algorithm, particularly for small objects. However, further research is needed to enhance the accuracy of this detection by making additional modifications to the network layer

components of the SSD architecture. This will optimize feature extraction from small objects based on their characteristics, leading to developing a more optimal network model.



(a)



(b)

Figure 4. Comparison results of traffic sign detection with SSD: (a) results of traffic sign detection using enhanced SSD; (b) traffic sign detection results using original SSD

4. CONCLUSIONS

The results from our small object detection for traffic sign recognition show that our modified SSD algorithm significantly improves detection and recognition, achieving a mAP of 97.87% compared to 95.35% for the original SSD algorithm. This enhancement can enhance the capability of the original SSD algorithm in detecting small objects, particularly in supporting ADAS systems that require precise detection of traffic signs from long distances. The limitations of this study are related to the influence of occlusion and clutter, which might affect the performance of object detection, especially for small objects, which are more susceptible to being influenced by various factors. Future research will address these limitations, including integrating these object detection algorithms into vehicle ADAS simulations.

ACKNOWLEDGMENT

We sincerely thank LPPM Atma Jaya Makassar University for their invaluable support of our research, facilitated through the 2024 UAJM lecturer research grants.

REFERENCES

- [1] J. Xu, Y. Huang, and D. Ying, "Traffic Sign Detection and Recognition Using Multi-Frame Embedding of Video-Log Images," *Remote Sens.*, vol. 15, no. 12, 2023, doi: 10.3390/rs15122959.
- [2] R. K. Megalingam, K. Thanigundala, S. R. Musani, H. Nidamanuru, and L. Gadde, "Indian traffic sign detection and recognition using deep learning," *Int. J. Transp. Sci. Technol.*, vol. 12, no. 3, pp. 683–699, 2023, doi: 10.1016/j.ijtst.2022.06.002.
- [3] A. W. Sudjana and H. Supeno, "Implementasi Deep Learning untuk Object Detection Menggunakan Algoritma YOLO (You Only Look Once) pada Rambu Lalu Lintas di Indonesia," *Univ. Pas.*, pp. 1–8, 2021.
- [4] Y. Zhu and W. Q. Yan, "Traffic sign recognition based on deep learning," *Multimed. Tools Appl.*, vol. 81, no. 13, pp. 17779–17791, 2022, doi: 10.1007/s11042-022-12163-0.
- [5] A. Vennelakanti, S. Shreya, R. Rajendran, D. Sarkar, D. Muddegowda, and P. Hanagal, "Traffic Sign Detection and Recognition using a CNN Ensemble," in *2019 IEEE International Conference on Consumer Electronics (ICCE)*, 2019, pp. 1–4. doi: 10.1109/ICCE.2019.8662019.
- [6] M. Flores-Calero *et al.*, "Traffic Sign Detection and Recognition Using YOLO Object Detection Algorithm: A Systematic Review," *Mathematics*, vol. 12, no. 2. 2024. doi: 10.3390/math12020297.
- [7] M. Akbar, A. S. Purnomo, and S. Supatman, "Multi-Scale Convolutional Networks untuk Pengenalan Rambu Lalu Lintas di Indonesia," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 11, no. 3, pp. 310–315, 2022, doi: 10.32736/sisfokom.v11i3.1452.
- [8] C. Peng, M. Zhu, H. Ren, and M. Emam, "Small Object Detection Method Based on Weighted Feature Fusion and CSMA Attention Module," *Electron.*, vol. 11, no. 16, 2022, doi: 10.3390/electronics11162546.
- [9] W. Liu *et al.*, "SSD: Single shot multibox detector," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9905 LNCS, pp. 21–37, 2016, doi: 10.1007/978-3-319-46448-0_2.
- [10] M. Weber, T. Weiss, F. Gechter, and R. Kriesten, "Approach for improved development of advanced driver assistance systems for future smart mobility concepts," *Auton. Intell. Syst.*, vol. 3, no. 1, p. 2, 2023, doi: 10.1007/s43684-023-00047-5.
- [11] X. N. Huynh, G. B. Jung, and J. K. Suhr, "One-Stage Small Object Detection Using Super-Resolved Feature Map for Edge Devices," *Electronics*, vol. 13, no. 2. 2024. doi: 10.3390/electronics13020409.
- [12] E. S. L. Roa, J. S. U. López, and I. D. C. Silva, "Real time face mask detection with SSD," in *2021 IEEE 2nd International Congress of Biomedical Engineering and Bioengineering (CI-IB&BI)*, 2021, pp. 1–4. doi: 10.1109/CI-IBBI54220.2021.9626095.
- [13] Hao Zhang, X. gong Hong, and L. Zhu, "Detecting Small Objects in Thermal Images Using Single-Shot Detector," *Autom. Control Comput. Sci.*, vol. 55, no. 2, pp. 202–211, 2021, doi: 10.3103/S0146411621020097.
- [14] S. Zhang, Z. Ma, G. Zhang, T. Lei, R. Zhang, and Y. Cui, "Semantic Image Segmentation with Deep Convolutional Neural Networks and Quick Shift," *Symmetry*, vol. 12, no. 3. 2020. doi: 10.3390/sym12030427.
- [15] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, in NIPS'16. Red Hook, NY, USA: Curran Associates Inc., 2016, pp. 4905–4913.
- [16] T. Diwan, G. Anirudh, and J. V. Tembhurne, "Object detection using YOLO: challenges, architectural successors, datasets and applications," *Multimed. Tools Appl.*, vol. 82, no. 6, pp. 9243–9275, 2023, doi: 10.1007/s11042-022-13644-y.
- [17] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable Object Detection Using Deep Neural Networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2155–2162. doi: 10.1109/CVPR.2014.276.

