Sensors and Materials, Vol. 37, No. 8 (2025) 3707–3722
MYU Tokyo

S & M 4147

Small-traffic-sign Detection Model Based on Improved YOLOv7

Hsin-Chun Lin,[1] Yung-Yao Chen,[1] Sin-Ye Jhong,[2,3] Cong-Cheng Zhang,[4]
Kai-Lung Hua,[4] Sheng-Tao Chenm,[5] and Chih-Hsien Hsia[6,7]*

[1]Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology,
No. 43, Sec. 4, Keelung Rd., Taipei City 106, Taiwan
[2]Smart Electric Vehicle Center, National Taiwan University of Science and Technology,
No. 43, Sec. 4, Keelung Rd., Taipei City 106, Taiwan
[3]Department of Engineering Science, National Cheng Kung University,
No. 1, University Rd., East District, Tainan City 701, Taiwan
[4]Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, No. 43, Sec. 4, Keelung Rd., Taipei City 106, Taiwan
[5]Department of Avionics Engineering, Republic of China Air Force Academy,
No. Sisou 1, Jieshou W. Rd., Kaohsiung City 820, Taiwan
[6]Department of Computer Science and Information Engineering, National Ilan University,
No. 1, Sec. 1, Shennong Rd., Yilan City, Yilan County 260, Taiwan
[7]Department of Business Administration, Chaoyang University of Technology, Taichung 413, Taiwan

(Received August 27, 2024; accepted November 26, 2024)

Keywords: small-object detection, traffic sign detection, attention mechanism, space-to-depth convolution, YOLOv7

With the increasing popularity of self-driving cars, road-condition detection systems have become a significant research focus. Traffic sign detection, which is a crucial component of these systems, directly affects the safety of both drivers and pedestrians. Owing to the urgent requirement for efficient traffic sign detection for autonomous driving applications, achieving high-performance and rapid responses for long-distance sign detection is crucial. You only look once v7 (YOLOv7) is a one-stage object detection model that offers excellent detection speed but faces challenges in long-range detections owing to the inherent loss of small-object features in its convolutional and maxpooling layers. To address these challenges, we propose enhancements for YOLOv7 by integrating a space-to-depth convolution module to better preserve small-object features and an attention mechanism to help it focus more effectively on relevant objects. We further enhanced it by adding extra detection heads specifically designed to extract small-object features and incorporated Gaussian noise to enhance its robustness. The improved model was evaluated on the National Taiwan University of Science and Technology Taiwan traffic sign dataset, which comprises 29 types of traffic sign. The results demonstrated the effectiveness of these enhancements, improving the mAP50 of YOLOv7 from 59.5 to 84.7% and offering a significantly better traffic sign detection performance.

*Corresponding author: e-mail: hsiach@niu.edu.tw
https://doi.org/10.18494/SAM5347

ISSN 0914-4935 © MYU K.K.
https://myukk.org/

1. Introduction

With continuous hardware improvements and advancements in computer-vision algorithms, autonomous-driving systems have become a popular research area in recent years. The core technologies necessary for a fully autonomous-driving vehicle include pedestrian,(1,2) vehicle detection,(3,4) lane detection,(5) and traffic sign detection.(6,7) Traffic signs are crucial for warning, prohibiting, and guiding vehicles and pedestrians to ensure road safety and prevent accidents. Therefore, the effective detection of these signs is essential for autonomous vehicles to protect both drivers and pedestrians. Additionally, it is crucial for these vehicles to identify traffic signs from a sufficient distance to enable ample reaction times. However, long-range detection is challenging owing to limited RGB camera resolutions and pixel information, which directly affect the quality of the information acquired. Even without compression, camera resolution significantly affects the information that can be obtained from a given scene.

Figure 1 illustrates the challenge we address in this study. Accurate traffic sign detection is vital for driver safety and autonomous driving systems, as misinterpreting or failing to recognize signs can lead to dangerous situations. Detecting signs from a distance is crucial, allowing drivers and automated systems time to process and react appropriately. However, this introduces challenges. Signs appear smaller from afar, reducing their pixel representation in images. This decreased resolution can distort or obscure features, especially in poor lighting or weather. In

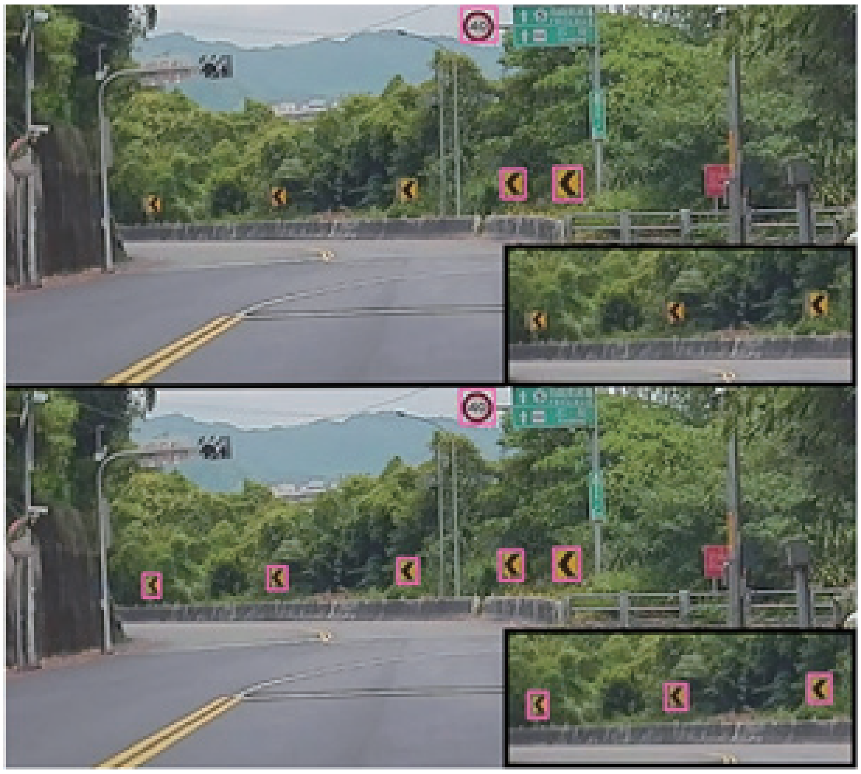


Fig. 1. (Color online) Comparison of the inferences of the original YOLOv7 (top) and the modified version proposed in this study. (Bottom) Inferences for a mountainous scene.

extreme cases, signs may become unrecognizable, compromising detection model reliability. Complex environments, such as urban areas with dense traffic and varied backgrounds, exacerbate this issue. Therefore, developing models that can accurately detect small, distant traffic signs under these conditions is critical for ensuring robust, long-distance detection and enhancing road safety. In this work, we collected 4970 photos from roads in Taiwan encompassing prohibition (seven categories), restricted (two categories), mandatory (three categories), warning (10 categories), indicative (five categories), and auxiliary (two categories) signs, obtaining a total of 29 sign classes. These photos were captured at various locations, including urban neighborhoods, highways, and mountain paths. This dataset, which is referred to as the National Taiwan University of Science and Technology (NTUST) Taiwan traffic dataset, was used in the experiments.

In summary, in this paper, we present significant contributions to the field of object detection, particularly in traffic sign recognition. The findings of this study are crucial for enhancing road safety and improving object recognition models' performance. (1) To address the issue of small-object features gradually disappearing during feature extraction in the backbone, we implemented the space-to-depth (SPD) convolution module. This approach better preserves detailed features. (2) To enhance small-object detection, we added a detection head to the underlying backbone. We also incorporated Gaussian noise during training to improve model robustness. Additionally, we integrated a Simam attention mechanism into the head, helping the model focus on objects of various sizes in complex scenes. (3) To validate the model's performance on Taiwan traffic signs, we created the NTUST Taiwan traffic dataset by collecting diverse traffic scenes in Taiwan. We then used this dataset to improve the model's performance on Taiwan-specific traffic signs.

2. Related Work

2.1 Object detection

AlexNet,(8) a graphics processing-unit-based convolutional neural network (CNN) introduced in 2012, demonstrated groundbreaking image classification performance, ushering in revolutionary developments in the field of computer vision. This advancement spurred interest in exploring the potential of computer vision for more complex tasks such as object detection and segmentation, making deep learning (DL) a popular research area. Object detection methods can be broadly classified as one- and two-stage approaches, with the latter exhibiting better performance. The two-stage approach involves generating candidate frames and then feeding them into a pretrained classification model,(9–11) or the classic machine learning (ML) method.(12–13) This intuitive and effective approach has laid the foundation for object detection methods, leading to the development of models such as faster region-based CNNs (R-CNNs),(14) and feature pyramid networks (FPNs) that offer(15) high detection accuracies. Additionally, although two-stage methods are slower than one-stage methods, they offer superior accuracy. In recent years, one-stage object-detection models, such as you-look-only-once (YOLO) models,(16) have achieved outstanding performance and efficiency, leading to their widespread adoption.

2.2 Traffic sign detection

Traffic sign recognition is a long-standing task and early studies relied on traditional computer vision techniques to locate and identify traffic signs.(12) In such techniques, the color distribution of a sign (e.g., red for warning and blue for compliance) is typically analyzed and then the sign area is segmented using feature descriptors such as a histogram of oriented gradients (HOG).(13) Subsequently, classic ML techniques such as support vector machine (SVM),(17) *k*-nearest neighbor classification (KNN), and random forest techniques were employed to classify the signs. Owing to significant technological advancements in recent years, CNNs have been increasingly employed for traffic sign recognition. These DL methods offer better sign recognition performance by leveraging features beyond simple color and HOG analysis. Recently, many researchers have focused on optimizing the YOLO model for traffic sign recognition by adjusting its internal structure to improve its accuracy and efficiency.(18–20)

2.3 Small-object detection

Small-object recognition is a branch of object recognition tasks that emerged after the maturation of object recognition technology. In recent years, with the development of applications such as unmanned aerial vehicles, the identification of small objects has become an important area of research. One of the biggest challenges in small-object recognition is that only a limited number of features can be used to locate and identify objects. Taking YOLO as an example, although we will not discuss its feature pyramid in-depth, generally, the input image data usually extract features through convolutional and maxpooling layers. However, after going through these two modules, some small features are gradually lost, especially when we use multiple convolution modules and maxpooling layers to extract features. This problem is particularly important.

The most direct way to improve the model's performance in recognizing small objects is to strengthen its feature extraction capabilities.(15) This can be achieved by directly introducing detection heads for small objects or optimizing the model.(21) On the other hand, integrating the attention mechanism can help improve the model's performance in detecting small objects. Through the introduction of contextual information,(22) the model can perform more accurate detection based on the information surrounding an object. The super-resolution method,(23) which can repair the features of objects with insufficient information and improve their recognizability, is another effective me. Additionally, by generating more samples of smaller objects,(24) we can expand the training data in a relatively simple and intuitive manner, allowing the model to learn more completely during the training process. This strategy helps the model obtain more complete information about small objects and improve its performance in practical applications.

2.4 Attention mechanism

In 2017, researchers such as Wang *et al.* reported their study,(25) in which they used attention modules stacked layer by layer, with each module focusing on a single feature of information.

However, continuing to stack these modules can have detrimental effects on the model. Eventually, they introduced the concept of residual networks, which gave rise to the first model of the attention mechanism.

Then, Hu *et al.* came out with squeeze-and-excitation network (SEnet)(26) and integrated the attention module they developed into the common CNN model at that time, which greatly improved the accuracy of the model. However, unlike SEnet, the convolutional block attention module (CBAM)(27) model focuses not only on the channels of the feature map but also on the spatial features of the set of channels via the two-layer attention mechanism. This allows CBAM to obtain better results than SEnet. The main goal of integrating the attention mechanism into the model is to enable the model to focus more on object features while reducing attention to irrelevant objects. This attention mechanism is also widely used in multiple tasks,(28–30) highlighting the importance of such a mechanism.

2.5 Feature map perturbation

Under normal circumstances, potential noises are hidden in the photos taken. Although humans cannot directly detect these noises, they are potential interferences that can negatively impact detection or classification models. To solve this problem, before training a model, Gaussian noise or salt and pepper noises are usually introduced into an image to augment the data, thereby improving the robustness of the model. It is also one of the simplest practices in early ML.

However, this approach does not fully consider the impact of noise on the model. Therefore, in a previous paper,(31) it was proposed to directly introduce random information into a feature map. In this way, the neural network can be perturbed more directly, thereby improving the robustness of the model. This method is used to solve both model overfitting(32) and insufficient data problems.(33) More importantly, this approach imposes no additional computational burden during inference and can improve model accuracy with existing resources.

3. Proposed Methodology

YOLOv7(34) introduces several key improvements that enhance both the speed and accuracy of object detection. For instance, it leverages the efficient layer aggregation network (ELAN) structure to boost the model's feature learning capacity while maintaining computational efficiency. Additionally, it employs dynamic label assignment (DLA) to improve the detection of objects of varying sizes. The model also incorporates the cross stage partial (CSP) structure to reduce computational costs and uses the FPN and path aggregation network (PAN) for multiscale feature fusion, further enhancing its capability to detect objects of different scales.

We propose an enhanced version of YOLOv7, which features an improved backbone and head and offers better performance in detecting small symbols; its structure is shown in Fig. 2. First, we introduce an SPD convolution module to replace the original convolution and maxpooling layers to mitigate the loss of small features. Second, we integrate the simple parameter-free attention mechanism (SimAM),(35) which is located before spatial pyramid

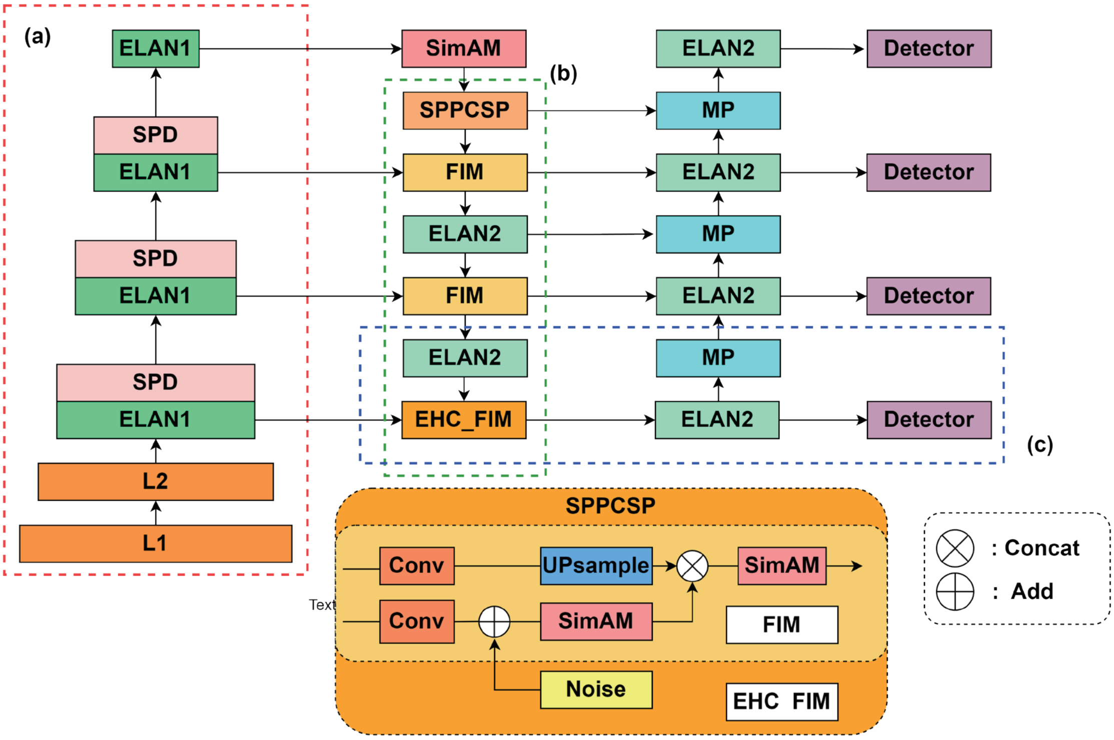


Fig. 2. (Color online) Framework of the improved YOLOv7 proposed in this study. (a) The SPD module replaces the upsampling and maxpooling layers in the original backbone, significantly enhancing the preservation of small-object characteristics. (b) SimAM is integrated into FIM, enabling the model to focus more effectively on object features. (c) Additional object detection heads are added to enhance the recognition performance for small objects. The FIM of the small-object detection head is also improved to increase the model robustness.

pooling cross stage partial (SPPCSP),(34) and input the head of the feature map; after the feature integration module (FIM), we apply SimAM during the downward propagation of the feature map from the backbone. This mechanism enables the model to primarily focus on the important parts of an object. Third, we introduce additional detection heads to extract small features to further boost the small-object detection capability of YOLOv7 and enhance the head FIM (EHC FIM). By perturbing the feature map during training, we improved the inference performance without increasing the computational complexity. These improvements were aimed at enhancing the model's efficiency in identifying small signs.

3.1 Simple attention module

In this section, we discuss in detail the integration and operating principles of SimAM,(35) in YOLOv7. Traffic signs typically occupy only a small part of an image with considerable background information. The primary purpose of AM is to help the model focus on traffic signs to enhance its detection accuracy for these signs. The main advantage of SimAM is that it is parameter-free. Therefore, unlike traditional AMs, SimAM does not involve numerous parameters or computational elements, such as convolutional layers, normalization, and activation functions, and it instead implements AM through mathematical functions. Thus, it simplifies the model structure, reduces unnecessary computational loads, and improves computation efficiency.

The traffic sign detection process involves the following steps: first, the input image passes through the backbone network that extracts object features from them layer-by-layer and then passes these features to higher levels through up-sampling to detect larger objects until they reach the final output layer. After processing the feature pyramid, the feature map in the head is combined with the backbone feature map at the corresponding level. Note that SimAM is employed before the feature map is input into the head and during its downward transmission to enhance the object attention at each level of the feature pyramid.

The AM model uses neurons from different areas to enhance the information intensity of certain neurons, thereby improving the detection and recognition performance of the model. AM is defined as

$$e_t\left(w_t,b_t,y,x_i\right)=\left(y_t-\hat{t}\right)+\frac{1}{M-1}\sum_{i=1}^{M}\left(y_o-\hat{x}_i\right)^2, \tag{1}$$

where t and x_i denote the target and other neurons, respectively, and $\hat{t} = w_t t + b_t$ and $\hat{x}_i = w_t x_i + b_t$. When $\hat{t} = y_t$ and all other $\hat{x}_i$ values are y_o, the equation generates the minimum value. Additionally, M is the number of feature maps on a single channel and is calculated by multiplying the height of the feature map by its width, and λ is the hyperparameter. By minimizing this equation, we can determine the linear separability between t and all other neurons in the same channel, thereby simplifying y_t and y_o to binary labels (1 and −1) and employing a regularizer in the equation. The final function is expressed as

$$e_t\left(w_t,b_t,y,x_i\right)=\frac{1}{M-1}\sum_{i=1}^{M-1}\left(-1-\left(w_t x_i+b_t\right)\right)^2+\left(1-\left(w_t x_i+b_t\right)\right)^2+w_t^2, \tag{2}$$

where w_t and b_t are computed as

$$w_t=-\frac{2\left(t-\mu_t\right)}{\left(t-\mu_t\right)^2+2\sigma_t^2+2\lambda}, \tag{3}$$

$$b_t=-\frac{1}{2}\left(t+\mu_t\right)w_t. \tag{4}$$

The mean (μ_t) and the variance (σ_t^2), excluding t, can be calculated using Eqs. (2) and (3), and these parameters are defined as

$$\mu_t = \frac{1}{M-1}\sum_{i=1}^{M-1} x_i, \tag{5}$$

$$\sigma_t^2 = \frac{1}{M-1}\sum_{i=1}^{M-1}\left(x_i - \mu_t\right)^2. \tag{6}$$

By computing w_t and b_t as well as the mean and variance of each channel, we can effectively reduce the additional computational burden resulting from the iteration. The attention is expressed as

$$e_t^* = \frac{4\left(\hat{\sigma}^2 + \lambda\right)}{\left(t - \hat{\mu}\right)^2 + 2\hat{\sigma}^2 + 2\lambda}. \tag{7}$$

Finally, a sigmoid function is applied to all e_t^* values to prevent them from becoming excessively large, and the result is denoted as E, which is then applied to the input feature map X to obtain $\tilde{X}$ as

$$\tilde{X} = sigmoid\left(\frac{1}{E}\right) \odot X. \tag{8}$$

3.2 SPD module

In this section, we explain the integration of the SPD layers(36) in YOLOv7. Conventional CNNs typically offer limited detection performance for small objects. Sunkara and Luo(36) focused on classical convolution operations and noted that they can ignore small pieces of information, thereby affecting the model's capability to learn excellent features for small objects or minute details. In YOLOv7, the size of the feature map is halved at each higher level of the feature pyramid to allow the high-level pyramid to learn the features of the large objects. However, the features of the small objects are lost during this process. Additionally, it offers suboptimal traffic sign recognition performance because the system must remotely recognize vehicles and signs. The traffic signs are generally small in photographs captured from sufficient distances. As shown in Fig. 3, we replaced the staggered convolution and pooling layers with SPD layers to ensure that the features of small objects can be preserved to the maximum extent.

3.3 Tiny layer enhancement

3.3.1 Normal tiny layer

The successful detection of road signs from a distance requires the processing of small objects. YOLOv7 was developed using the Common Objects in Context (COCO) dataset comprising images with dimensions of 640 × 640 pixels, and any object with a side length of <32

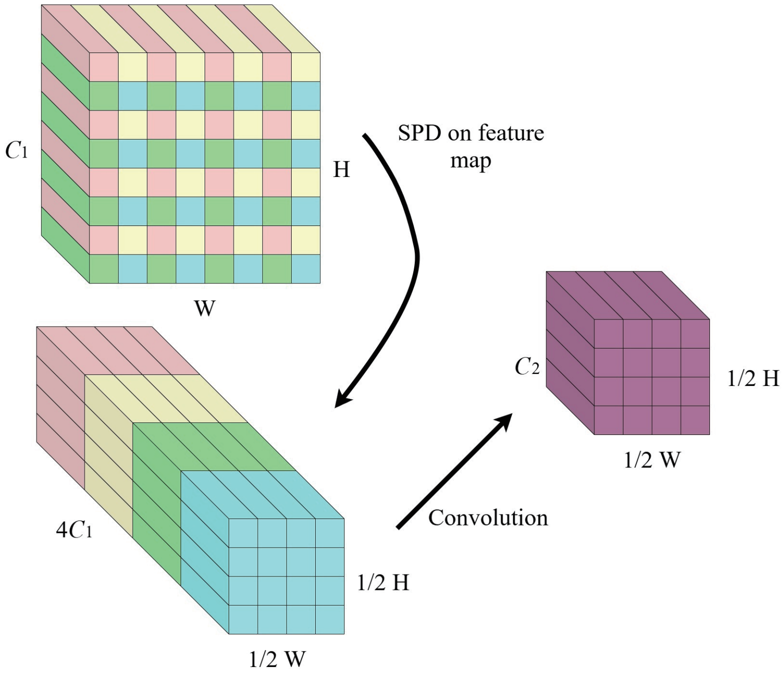


Fig. 3. (Color online) Illustration of the SPD module operation.

pixels was considered a tiny object. The object sizes typically ranged from 30 to 60 pixels. Additionally, the NTUST Taiwan traffic sign dataset comprised images of 1280 × 712 pixels; however, we resized them before inputting them into the model, with the expected object sizes ranging from 30 to 60 pixels. Therefore, we focused on capturing the features of smaller objects by adding more object detection heads. Moreover, the model can obtain more detailed information on small objects at the bottom of the feature pyramid, leading to better performance for the traffic dataset.

3.3.2 Enhanced tiny layer

With the additional sensing heads, the model can better detect tiny objects. Compared with the original detection layer of YOLOv7, our improved detection head was closer to the lower layers. However, the convolutional layers are not very deep at these lower levels. Because our data were extracted directly from a camera without preprocessing, the images may have contained noise, which can affect the capability of the sensing head to extract features from the objects. Therefore, we introduced Gaussian noise into the detector head for small objects to help it adapt to various environments, thereby improving the model's robustness.

3.4 Experimental results

In this subsection, we describe the experiments conducted using the NTUST Taiwan traffic sign dataset to evaluate the performance of the modified YOLOv7 model. Furthermore, ablation experiments were conducted to evaluate the effect of each model component.

3.4.1 Datasets

The NTUST Taiwan traffic sign dataset included road images from city centers, highways, and mountain roads. We installed a camera inside a car directly under the rearview mirror; the camera specifications are listed in Table 1. The images included various types of traffic sign, including prohibition, mandatory, restricted, warning, instruction, and auxiliary signs, with a total of 29 categories and more than 4970 images of 1280 × 712 pixels. As shown in Fig. 4, datasets comprising outdoor images often suffer from data imbalances, and the dataset employed in this study is no exception. It featured significant quantity differences among the different categories, with small objects posing additional challenges.

Figure 5 illustrates some data samples from the NTUST Taiwan traffic sign dataset, which were collected outdoors. Some sign categories are unique to certain mountainous road sections, which made their collection particularly challenging. Additionally, urban settings pose greater complexities than suburban settings, with occlusions owing to high traffic volumes being the most prevalent issue. Furthermore, business district signs are often misidentified because their colors and symbols are similar to those of advertising signs. These challenges are compounded by the difficulty of detecting small objects, particularly for images captured from a distance.

3.4.2 Implementation details

The proposed YOLOv7-based model resizes the input images to 640 × 640 pixels. The weights from the official pretrained model were used as the initial training weights, and most

Table 1
Camera specifications.

Chip	1/1.8" CMOS chip
Color	Multicolor
Max resolution	2048 × 1536 pixels
Pixel size	3.45
Frame rate	55 fps

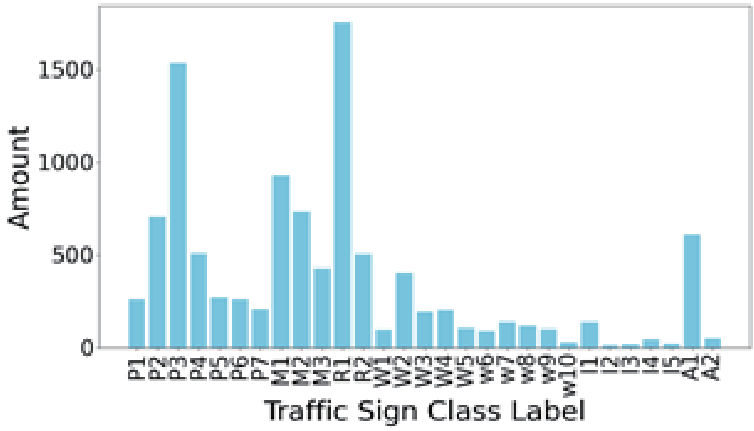


Fig. 4. (Color online) Statistical distribution of each traffic sign class in the dataset.

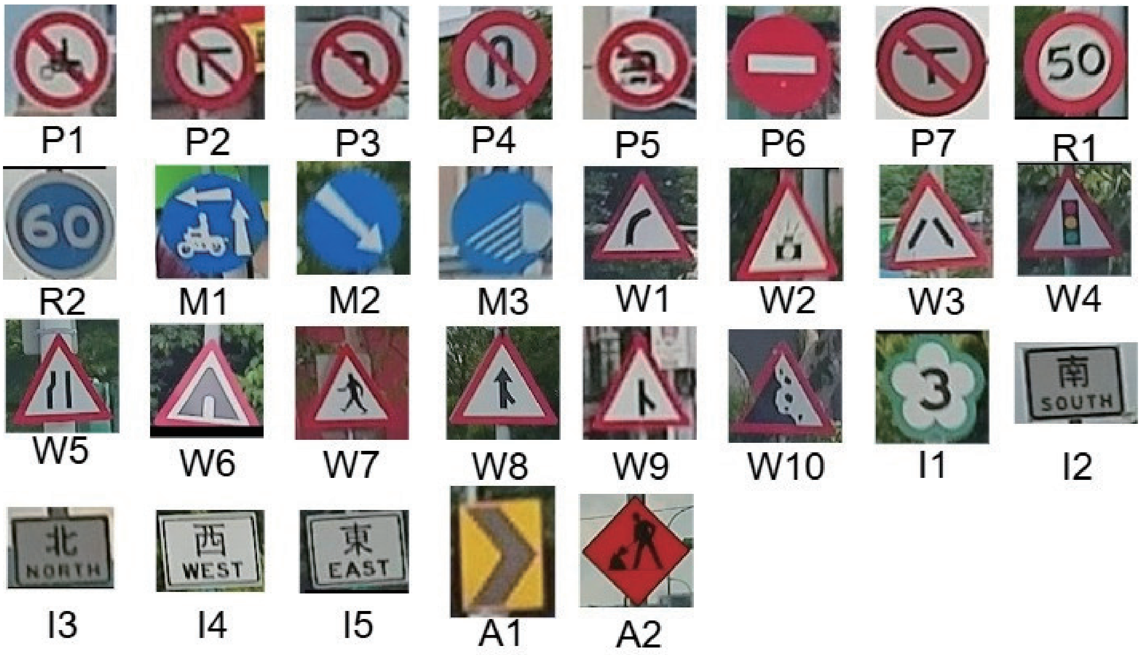


Fig. 5. (Color online) Sign types and corresponding labels included in the NTUST Taiwan traffic sign dataset.

training settings were adopted from conventional YOLOv7 configurations. During our experiments, we disabled the left- and right-flipping in the official configuration file. The model exhibited good convergence performance after 200 epochs on our dataset. During inference, we used a batch size of 32 and a confidence threshold of 0.3. Additionally, the experiments were conducted on a system comprising an NVIDIA 2080ti 11G GPU.

4. Results and Discussion

4.1 Main experiment

We successfully improved the accuracy of YOLOv7 on the Taiwan traffic sign dataset by integrating SimAM and the SPD, and tiny-layer enhancement modules. As shown in Table 2, the proposed YOLOv7 model showed a significantly improved accuracy compared with the conventional YOLOv7 model without substantially increasing the computation time. Specifically, the SPD convolution module performed exceptionally well because our dataset primarily comprised small objects, making it well-suited for this task. Additionally, ideal results were obtained by introducing disturbances into the feature maps of the small-object detection heads. However, this approach increases the computational complexity owing to the large dimensions of the underlying feature maps. Therefore, our task entailed a higher computational effort owing to the considerable number of small objects in our dataset. The results are listed in Table 2.

Table 2
Results of the conventional and proposed YOLOv7 models for the NTUST Taiwan traffic sign dataset.

Method	Precision (%)	Recall (%)	mAP50 (%)	mAP50:95 (%)
YOLOv7	71.3	61.4	59.5	45.5
This work	**90.0**	**81.8**	**84.7**	**64.7**

4.2 Ablation studies

We analyzed how various component combinations affect the performance of YOLOv7 on the NTUST Taiwan traffic sign dataset. We first evaluated the performance of each module individually and then combined the most effective results. In the following sections, we will discuss these results in detail, demonstrating that each component contributed positively to the overall performance.

4.2.1 SimAM attention module

The YOLOv7 model was divided into three main areas (L1, L2, and L3) and SimAM was employed in each area. In L1, it was incorporated from the trunk to the head, whereas in L2, it was integrated into the head and propagated through the feature map from the top to the bottom. This enabled the model to focus more on objects of various sizes, thereby enhancing its capability to handle and detect small objects effectively. Finally, in L3, SimAM was integrated into the bottom-up propagation process of the underlying features of the model. This allowed the passing of the object features detected at lower levels to higher levels, thereby enhancing the capability of the model to capture large objects. As indicated by the results listed in Table 3, this design enabled the model to adapt better to objects of different sizes at different levels, thereby improving its detection performance.

4.2.2 SPD module

We used the SPD convolution module to replace the bottom-up upsampling modules in the backbone and head to reduce the loss of small-object features during upsampling. Table 4 presents the crossover results of the experiments for the NTUST Taiwan traffic sign dataset. Surprisingly, in the first experiment, wherein all maxpooling modules were replaced with SPD-Conv blocks, we did not achieve the best results. We believe that this may be due to the potential noise disrupting the underlying layer, resulting in significant differences in image output. However, not all results were unsatisfactory. Satisfactory results were obtained in the second experiment, wherein the SPD module was added to the backbone of YOLOv7, indicating that the SPD module can be effectively applied to each layer to retain more detailed features.

4.2.3 Enhanced tiny layer

We added more sensing heads to the original YOLOv7 model to enhance its detection performance for tiny objects. Accuracy can be improved by obtaining feature information

Table 3
Ablation experiment results for SimAM.

L1	L2	L3	Precision (%)	Recall (%)	mAP50 (%)	mAP50:95 (%)	Time (ms)
✓			77.5	64.0	64.0	48.2	**4.4**
✓	✓		**81.4**	**74.6**	**73.7**	**56.2**	4.5
✓	✓	✓	75.9	74.4	71.3	54.0	4.8

Table 4
Ablation experiment results for the SPD convolution module.

Layer	Precision (%)	Recall (%)	mAP50 (%)	mAP50:95 (%)	Time (ms)
1-7	**82.0**	72.7	72.6	54.9	6.3
1-5	79.7	80.7	79.5	60.5	6.1
1-3	78.1	77.2	75.6	58.7	6.3
3-5	80.4	**81.3**	**80.0**	**61.0**	**4.3**

Table 5
Ablation experiment results for the enhanced tiny layer.

γ	Precision (%)	Recall (%)	mAP50 (%)	mAP50:95 (%)	Time (ms)
0	82.4	78.7	78.4	59.7	**5.5**
0.005	**85.4**	78.7	79.5	61.0	**5.5**
0.01	84.2	**83.6**	**82.4**	**63.6**	**5.5**
0.02	82.7	81.2	80.4	61.7	**5.5**

directly from the lowest layer. YOLOv3[37] and its future versions primarily use feature pyramids to obtain object features of different sizes, perform feature extraction, and pass the extracted features to the next layer. During this feature transfer process, the size of the entire feature map decreases gradually, leading to a gradual loss of the characteristics of small objects. Using the detection head directly at the bottom layer is a simple and effective method for improving recognition performance for small objects. Although this approach enhances accuracy, it requires feature size reduction, especially for feature maps close to the bottom layer, which can affect the inference and training time. Additionally, Gaussian noise was introduced in the detection head to enhance the model resilience. We added noise with a gamma standard deviation at various levels and identified the appropriate values for our dataset and the small-object detection head, which improved the small-object detection performance (see Table 5).

5. Conclusions

In this study, we effectively enhanced the YOLOv7 model for small-traffic-sign detection by integrating three additional modules and testing the model on the Taiwan NTUST traffic sign dataset. Although we successfully improved the model's mAP50 from 59.5 to 84.7%, there are potential limitations to consider. The added complexities, such as tiny detector heads, Gaussian noise, and SPD convolution, could increase the computational load, potentially affecting real-time performance, especially in resource-limited environments. Additionally, although the dataset primarily focused on long-distance traffic sign detection, it may not fully represent diverse real-world conditions, such as varying weather and lighting scenarios. Future work should aim to optimize the model's computational efficiency and expand the dataset to include more challenging environments, thereby enhancing both the speed and robustness of the model in practical applications.

Acknowledgments

The authors would like to thank the anonymous reviewers of the paper for the many helpful suggestions. This work was partially supported by the National Science and Technology Council (NSTC), Taiwan, under grant number 113-2222-E-013-001, and the Aerospace Technology Research and Development Center (ATRDC) of ROCAF for research grants in 2024.

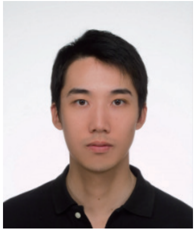
References

1 Y.-Y. Chen, G.-Y. Li, S.-Y. Jhong, P.-H. Chen, C.-C. Tsai, and P.-H. Chen: Sens. Mater. **32** (2020) 3157. https://doi.org/10.18494/SAM.2020.2838
2 S.-Y. Jhong, Y.-Q. Wang, W.-J. Cheng, H.-W. Hwang, and Y.-Y. Chen: 2022 Int. Conf. System Science and Engineering (ICSSE) (IEEE, 2022) 7. https://doi.org/10.1109/ICSSE55923.2022.9948235
3 C.-H. Hsia, S.-C. Yen, and J.-H. Jang: Sens. Mater. **31** (2019) 1803. https://doi.org/10.18494/SAM.2019.2351
4 Y.-Q. Wang, P.-H. Chen, S.-Y. Jhong, K.-M. Yen, and Y.-Y. Chen: 2021 IEEE Int. Conf. Consumer Electronics-Taiwan (ICCE-TW) (IEEE, 2021) 1. https://doi.org/10.1109/ICCE-TW52618.2021.9602987
5 S.-Y. Jhong, C.-H. Ko, Y.-F. Su, K.-L. Hua, and Y.-Y. Chen: 2023 Int. Conf. Advanced Robotics and Intelligent Systems (ARIS) (IEEE, 2023) 1. https://doi.org/10.1109/ARIS59192.2023.10268521
6 S. Chen, Z. Zhang, L. Zhang, R. He, Z. Li, M. Xu, and H. Ma: IEEE Internet Things J. **11** (2024) 19500. https://doi.org/10.1109/JIOT.2024.3367899
7 J. Wang, Y. Chen, X. Ji, Z. Dong, M. Gao, and C. S. Lai: IEEE Trans. Intell. Trans. Syst. **25** (2024) 710. https://doi.org/10.1109/TITS.2023.3309644
8 A. Krizhevsky, I. Sutskever, and G. E. Hinton: Advances in Neural Information Processing Systems 25, F. Pereira, C. J. Burges, L. Bottou, and K. Q. Weinberger, Eds. (Curran Associates, Inc., New York, 2012) pp. 1097–1105. https://doi.org/10.1145/3065386
9 L. Wu, H. Li, J. He, and X. Chen: J. Phys.: Conf. Ser. **1176** (2019) 032045. https://doi.org/10.1088/1742-6596/1176/3/032045
10 Z. Zuo, K. Yu, Q. Zhou, X. Wang, and T. Li: Proc. 2017 IEEE 37th Int. Conf. Distributed Computing Systems Workshops (ICDCSW) (IEEE, Atlanta, 2017) 286–288. https://doi.org/10.1109/ICDCSW.2017.34
11 C. Han, G. Gao, and Y. Zhang: Multimed. Tools Appl. **78** (2019) 13263. https://doi.org/10.1007/s11042-018-6428-0
12 A. Sugiharto and A. Harjoko: 2016 3rd Int. Conf. Information Technology, Computer, and Electrical Engineering (ICITACEE) (IEEE, 2016) 317. https://doi.org/10.1109/ICITACEE.2016.7892463
13 C. Rahmad, I. F. Rahmah, R. A. Asmara, and S. Adhisuwignjo: 2018 Int. Conf. Information and Communications Technology (ICOIACT) (IEEE, 2018) 50. https://doi.org/10.1109/ICOIACT.2018.8350804
14 S. Ren, K. He, R. Girshick, and J. Sun: Adv. Neural Inf. Process. Syst. 28, eds. C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Curran Associates, Inc., 2015). https://doi.org/10.1109/TPAMI.2016.2577031
15 T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie: 2017 IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (IEEE, 2017) 2117. https://doi.org/10.1109/CVPR.2017.106
16 C.-H. Hsia, H.-C. Peng, and H.-T. Chan: Electronics **12** (2023) 1. https://doi.org/10.3390/electronics12102312
17 Y. Xie, L.-F. Liu, C.-H. Li, and Y.-Y. Qu: 2009 IEEE Intelligent Vehicles Symp. (IEEE, 2009) 24–29. https://doi.org/10.1109/IVS.2009.5164247
18 M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf: IEEE Intell. Syst. Appl. **13** (1998) 18. https://doi.org/10.1109/5254.708428
19 K. Taunk, S. De, S. Verma, and A. Swetapadma: 2019 Int. Conf. Intelligent Computing and Control Systems (ICCS) (IEEE, 2019) 1255. https://doi.org/10.1109/ICCS45141.2019.9065747
20 T. K. Ho: Proc. 3rd Int. Conf. Document Analysis and Recognition (ICDAR) (IEEE, 1995) 278–282. https://doi.org/10.1109/ICDAR.1995.598994
21 J. Chu, C. Zhang, M. Yan, H. Zhang, and T. Ge: Sensors **23** (2023) 8. https://doi.org/10.3390/s23083871
22 G. Qi, Y. Zhang, K. Wang, N. Mazur, Y. Liu, and D. Malaviya: Remote Sens. **14** (2022) 2. https://doi.org/10.3390/rs14020420

23 J.-S. Lim, M. Astrid, H.-J. Yoon, and S.-I. Lee: Proc. 2021 Int. Conf. Artificial Intelligence in Information and Communication (ICAIIC) (IEEE, 2021) 181. https://doi.org/10.1109/ICAIIC51459.2021.9415217
24 C. Deng, M. Wang, L. Liu, Y. Liu, and Y. Jiang: IEEE Trans. Multimedia **24** (2022) 1968. https://doi.org/10.1109/TMM.2021.3074273
25 F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang: Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (IEEE, 2017) 3156. https://doi.org/10.1109/CVPR.2017.683
26 J. Hu, L. Shen, and G. Sun: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (IEEE, 2018) 7132. https://doi.org/10.1109/CVPR.2018.00745
27 S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon: Proc. European Conf. Computer Vision (ECCV) (Springer, 2018) 3.
28 Z. Li, Y. Sun, L. Zhang, and J. Tang: IEEE Trans. Pattern Anal. Mach. Intell. **44** (2022) 9904. https://doi.org/10.1109/TPAMI.2021.3132068
29 H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain: Proc. 2020 IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (IEEE, 2020) 5781. https://doi.org/10.1109/cvpr42600.2020.00582
30 Z. Li, M. Duan, B. Xiao, and S. Yang: IEEE Trans. Ind. Inf. **19** (2023) 7278. https://doi.org/10.1109/TII.2022.3231923
31 F. Juefei-Xu, V. N. Boddeti, and M. Savvides: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (IEEE, 2018) 3310. https://doi.org/10.1109/CVPR.2018.00349
32 H. Noh, T. You, J. Mun, and B. Han: arXiv preprint (2017). https://doi.org/10.48550/arXiv.1710.05179
33 Z. Liu, Y. Zhou, Y. Xu, and Z. Wang: Proc. 2023 IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (IEEE, 2023) 20402. https://doi.org/10.1109/CVPR52729.2023.01954
34 C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao: Proc. 2023 IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (IEEE, 2023) 7464. https://doi.org/10.1109/CVPR52729.2023.00721
35 L. Yang, R.-Y. Zhang, L. Li, and X. Xie: Proc. 38th Int. Conf. Machine Learning (PMLR, 2021) 11863. https://proceedings.mlr.press/v139/yang21o.html
36 R. Sunkara and T. Luo: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, M. R. Amini, S. Canu, A. Fischer, T. Guns, P. K. Novak, and G. Tsoumakas, Eds. (Springer, Cham, 2022) 443–459. https://doi.org/10.1007/978-3-031-26409-2_27
37 J. Redmon and A. Farhadi: arXiv preprint (2018). https://doi.org/10.48550/arXiv.1804.02767

About the Authors

Hsin-Chun Lin received his B.S. degree from the Department of Electronic Engineering, National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan, in 2020, and his M.S. degree from the Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan, in 2022. He is currently pursuing his Ph.D. degree at the Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan. His current research interests include visual odometry and AI applications.
(d11102001@gapps.ntust.edu.tw)

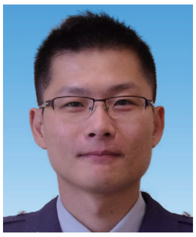


Yung-Yao Chen received his B.S. degree in electrical and control engineering from National Chiao Tung University, Hsinchu, Taiwan, in 2004, and his M.S. degree from the same institution in 2006. He earned his Ph.D. degree in electrical engineering from Purdue University, USA, in 2013. He is currently an associate professor in the Department of Electronic and Computer Engineering and a co-director of the Taiwan Tech Smart Electric Vehicle Research Center, National Taiwan University of Science and Technology, Taipei, Taiwan. His current research interests include vision-based automation, smart manufacturing, autonomous driving, and human–computer interaction. (yungyaochen@gapps.ntust.edu.tw)

Sin-Ye Jhong received his M.E. degree from the Graduate Institute of Automation Technology, National Taipei University of Technology, Taipei, Taiwan, in 2019. He is currently pursuing his Ph.D. degree at the Department of Engineering Science, National Cheng Kung University, Tainan, Taiwan. Since 2023, he has been a researcher at the Smart Electric Vehicle Center, National Taiwan University of Science and Technology, Taiwan. His research interests include digital image and video processing, computer vision, and deep learning. (n98081034@ncku.edu.tw)

Cong-Cheng Zhang received his B.S. degree from the Department of Computer Science and Information Engineering, Fu Jen Catholic University, New Taipei, Taiwan, in 2021 and his M.S. degree from the Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan, in 2024.
(m11015q10@mail.ntust.edu.tw)

Kai-Lung Hua received his B.S. degree in electrical engineering from National Tsing Hua University, Hsinchu, Taiwan, in 2000, his M.S. degree in communication engineering from National Chiao Tung University, Hsinchu, Taiwan, in 2002, and his Ph.D. degree from the School of Electrical and Computer Engineering, Purdue University, USA, in 2010. Since 2010, he has been with the National Taiwan University of Science and Technology, where he is currently a professor at the Department of Computer Science and Information Engineering. His research interests include digital image and video processing, computer vision, and machine learning.
(hua@mail.ntust.edu.tw)



Sheng-Tao Chen is an assistant professor at the Republic of China Air Force Academy, Taiwan. He received his M.S. degree from the Department of Information Management, the School of Defense Science, Management College, National Defense University Taipei, Taiwan, in 2013 and his Ph.D. degree from the Department of Electrical and Electronic Engineering of Chung Cheng Institute of Technology, National Defense University, Taoyuan, Taiwan, in 2020. His research interests include AIoT, wireless sensor networks, and computer vision. (iiccanffly@gmail.com)

Chih-Hsien Hsia received his Ph.D. degree in electrical and computer engineering from Tamkang University and his second Ph.D. degree from National Cheng Kung University, Taiwan. He is currently a distinguished professor and a chairperson at the Department of Computer Science and Information Engineering, NIU. His research interests include DSP IC design, GenAI/AI in multimedia, and cognitive engineering. (hsiach@niu.edu.tw)