

An Efficient End-to-End Obstacle Recognition Framework for Assistive Navigation Based on Enhanced EfficientNetV2-B0 and Feature Pyramid Network

Hadi Santoso ^{1,*}, Vemas A. Pratama ¹, Mudrik Alaydrus ², Misni ¹, and Isack Farady ³

¹ Faculty of Computer Science, Universitas Mercu Buana, Jakarta, Indonesia

² Faculty of Engineering, Universitas Mercu Buana, Jakarta, Indonesia

³ Department of Electrical Engineering, Yuan Ze University, Taiwan

Email: hadi.santoso@mercubuana.ac.id (H.S.); pratamavemas7@gmail.com (V.A.P.);
mudrikalaydrus@mercubuana.ac.id (M.A.); misni.muhammad@mercubuana.ac.id (M.);
isackf@saturn.yzu.edu.tw (I.F.)

*Corresponding author

Abstract—Visual impairment severely restricts independent mobility in crowded urban environments, where unpredictable obstacles such as pedestrians, vehicles, and street furniture pose constant safety risks. The white cane, the most common assistive tool, detects objects only on contact and provides no contextual information. Although computer vision-based obstacle detection offers a promising solution, real-time deployment on resource-constrained Internet of Things (IoT) and mobile devices remains unresolved. Existing approaches are limited by heavy architectures unsuitable for edge deployment and by applying backbone networks, feature fusion, attention mechanisms, and refinement modules independently rather than as a unified system. This study proposes a lightweight end-to-end obstacle recognition framework integrating EfficientNetV2-B0, an adaptive Feature Pyramid Network (FPN) with learnable fusion weights, a Squeeze-and-Excitation (SE) attention module, and a residual refinement block into a single architecture. The key novelty lies in the adaptive scale-aware fusion strategy that dynamically balances multi-scale feature contributions via SoftMax-normalized importance weights—unlike conventional fixed elementwise FPN operations. Evaluated on ten obstacle classes (3133 training, 432 validation images), the proposed model achieves 98.84% accuracy with weighted precision, recall, and F1-Score of 98.92%, 98.84%, and 98.85%, respectively—an improvement of 75.51% over a baseline EfficientNetB0+SE model (23.33%), confirming the effectiveness of integrated architecture for real-time assistive navigation.

Keywords—obstacle detection, visually impaired navigation, EfficientNetV2-B0, Feature Pyramid Network (FPN), lightweight deep learning

I. INTRODUCTION

Independent mobility is a common problem for visually impaired people. The problem becomes more serious in crowded urban areas. In these places, obstacles move randomly and are difficult to predict. This situation often causes collisions. Such conditions affect safety and also limit daily activities like traveling, working, and social interaction. A white cane is widely used because it is simple and practical. However, it works only when the cane touches an object and the detection range is very limited. It also cannot provide information about object type or surrounding space. Because of this limitation, visually impaired users often receive incomplete information about their environment, particularly in complex urban locations.

In recent years, computer vision and deep learning techniques have been used to support navigation for visually impaired people. Many studies apply Convolutional Neural Networks (CNNs) to extract visual information and detect obstacles. Many studies have tried different backbone models to improve obstacle detection. EfficientNet-based architectures have demonstrated strong performance across various image classification tasks, including medical imaging and environmental recognition [1]. Some studies use a Feature Pyramid Network (FPN) to handle objects with different sizes. Other studies add attention blocks such as Squeeze-and-Excitation (SE). These blocks help the model focus on useful features. In addition, refinement methods are used to reduce errors in detection results. However, most of these approaches treat each component separately, which may reduce overall effectiveness when the system is used in real-time conditions.

Another issue is that many deep learning-based obstacle detection models require high computational resources. This makes them difficult to deploy on Internet of Things (IoT) or edge devices with limited hardware capability. For assistive navigation systems, it is important that the model can work in real time while still providing acceptable detection accuracy. Lightweight models are therefore needed, but combining lightweight networks with multi-scale feature extraction and attention mechanisms is still not widely explored.

This research aims to develop an integrated obstacle detection framework that differs from previous works by combining multiple architectural components within a single lightweight model. In addition, this study proposes an adaptive feature fusion strategy within the FPN to dynamically balance multi-scale representations. Unlike existing studies that evaluate FPN, attention mechanisms, or refinement modules as separate enhancements, this study unifies EfficientNetV2-B0, multi-scale feature extraction, channel-wise attention, and feature refinement into one end-to-end architecture. This integration is designed to improve detection robustness in dynamic urban environments while maintaining low computational cost. By focusing on architectural synergy rather than isolated improvements, this research addresses the gap between detection accuracy and real-time deployment requirements for assistive navigation systems on resource-constrained devices.

In addition to architectural integration, this study introduces an adaptive feature fusion mechanism within the FPN, where learnable weighting parameters are employed to dynamically balance spatial and semantic information across pyramid levels. This differs from conventional FPN designs that rely on fixed element-wise operations, enabling more flexible and discriminative multi-scale feature representation. The objectives of this study are:

- (1) To develop a lightweight end-to-end obstacle recognition framework based on EfficientNetV2-B0 for real-time deployment on IoT and mobile devices;
- (2) To integrate adaptive FPN, SE attention, and feature refinement into a unified architecture;
- (3) To introduce an adaptive scale-aware feature fusion mechanism with learnable weighting parameters;
- (4) To validate the framework through real-world data evaluation and a mobile-based case study.

The rest of the paper is organized as follows: Section II reviews related literature on assistive navigation and deep learning-based obstacle detection. Section III describes the materials and methods. Section IV presents results and discussion. Section V presents the conclusions.

II. LITERATURE REVIEW

Visual impairment affects a large number of people and creates many daily limitations. One of the main problems faced by visually impaired people is independent mobility, especially in urban environments [2]. According to reports from the World Health Organization, a large number of

people experience visual impairment and many of them face difficulties during daily navigation [3]. In crowded places such as city centers, public transportation areas, and commercial zones, visually impaired people often encounter moving obstacles like pedestrians, vehicles, and street facilities. These conditions increase the risk of accidents and make safe movement more difficult [4–6].

A white cane is the most widely used assistive tool for visually impaired people. The white cane gives basic tactile feedback when it touches an object. The information it provides is limited and depends only on direct contact. It cannot identify the type of obstacle, estimate distance, or provide information about the surrounding environment in real time [7, 8]. This makes navigation difficult, particularly when the user is in a new or changing environment. This situation encourages researchers to develop intelligent assistive systems that can provide additional environmental information [9].

Deep learning has been used in several assistive systems for visually impaired navigation [10, 11]. Many studies apply CNNs for obstacle detection tasks [4, 7, 8, 12]. CNN-based methods show better performance than traditional vision techniques in several cases. Maurya and Verma [7] reported improved detection results using a CNN-based approach.

Other studies focus on real-time assistive systems. Jouini *et al.* [8] developed a CNN-based assistive device for obstacle detection and user feedback. While the system works in real time, it relies on a relatively heavy model. Similar issues are found in several IoT-based assistive systems [5, 13, 14]. These systems often require higher computational resources and complex processing pipelines, which makes them less suitable for low-power or edge devices.

Efficient models are often required when computational resources are limited. For this reason, several studies use lightweight neural networks instead of large architectures. EfficientNetV2 is one of the models used in such cases. The network adjusts its structure by changing depth, width, and input size [15–17]. This allows the model to reduce computation while keeping acceptable accuracy. EfficientNetV2 has been applied in different tasks, including object detection and environment recognition, and has been reported to work well in real-time and edge-based systems [18–20].

EfficientNetV2 has also been compared with other backbone networks such as MobileNet and Cross Stage Partial (CSP)-DarkNet. In these comparisons, the model shows a reasonable balance between accuracy and processing speed [18, 19]. This balance is important for systems that need fast response, such as assistive navigation applications. For assistive navigation systems, such characteristics are important because the system must provide timely feedback to the user.

In real-world obstacle detection, objects appear at different sizes and distances. FPN is commonly used to handle this problem [21, 22]. FPNs enhance detection performance by combining high-level semantic information with low-level spatial details, enabling improved recognition of small and distant objects.

Several studies further modify the basic FPN structure. Zhu [21] proposed a dynamic FPN that changes feature fusion across scales. Other studies introduce weighted or bidirectional FPN structures to improve detection results in complex environments such as aerial or marine scenes [22]. Although FPN improves multi-scale detection, many assistive systems apply it without additional mechanisms to highlight important features. As a result, detection performance may decrease in cluttered or crowded environments.

Attention mechanisms have been introduced to improve feature representation by emphasizing informative channels and suppressing irrelevant noise. The SE block is a widely used channel attention mechanism that performs adaptive feature recalibration through global context modeling [23–25]. SE integration has demonstrated consistent performance improvements across various tasks, including plant disease classification, medical image analysis, and action recognition [23, 25, 26].

In EfficientNet-based models, SE blocks are often used to improve feature learning. Several studies report that SE blocks can enhance feature discrimination without adding much computational cost [25, 27]. However, in many assistive navigation studies, SE blocks are usually applied alone and not combined with multi-scale feature fusion methods. This limits the benefit that can be obtained from using SE mechanisms together with other architectural components.

Apart from backbone and attention modules, feature refinement is also used to improve detection results. Feature refinement methods focus on correcting errors and improving spatial details before the final output [28, 29]. Some refinement approaches are based on residual connections so that the original feature information is not lost. These methods help improve feature quality in a gradual manner.

Zhou *et al.* [28] showed that feature refinement can improve visual restoration performance by adjusting feature representations. Similar ideas are also applied in other tasks such as action detection and video analysis, where refinement helps improve accuracy and stability [29]. However, in assistive navigation systems

for visually impaired users, refinement modules are still not commonly used. In particular, refinement is rarely combined with lightweight backbone networks and multi-scale feature extraction methods.

From the existing studies, several limitations can be observed. Many obstacle detection systems use heavy models that are difficult to run in real time. Other approaches rely on limited feature representations, which affects detection quality. In addition, backbone networks, FPN, SE blocks, and refinement modules are often used separately rather than as a single integrated model. As a result, the full potential of these components is not fully utilized.

This work focuses on combining these components into one framework. The aim is to use an efficient backbone together with multi-scale feature fusion, channel attention, and feature refinement. By integrating these elements, the model is expected to achieve reliable obstacle detection while remaining suitable for real-time operation on devices with limited.

III. MATERIALS AND METHODS

This study uses an experimental approach to develop an obstacle recognition model for assistive navigation. The main objective of the proposed method is to detect obstacles accurately while keeping the model lightweight enough for real-time use. For this reason, deep learning techniques are applied with attention to both detection performance and computational efficiency.

This study is designed as experimental research employing deep learning techniques to develop an obstacle recognition model intended for assistive systems for visually impaired individuals. The primary focus of the proposed framework is not only to achieve high classification accuracy but also to ensure architectural efficiency suitable for real-time deployment scenarios. Therefore, the methodology emphasizes the integration of lightweight feature extraction, multi-scale representation learning, attention-based enhancement, and feature refinement within a single end-to-end model.

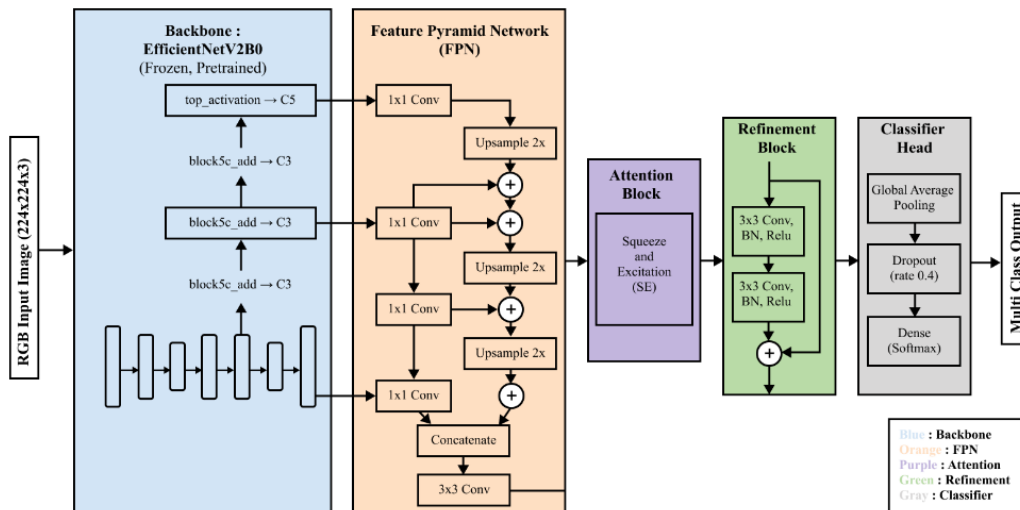


Fig. 1. Framework of proposed method.

The proposed system follows a sequential workflow that begins with dataset acquisition and preprocessing, followed by feature extraction using a pre-trained backbone network. The overall framework of the proposed method is illustrated in Fig. 1. Subsequently, multi-scale feature fusion is performed to address variations in obstacle size and spatial distribution. Channel-wise attention and refinement mechanisms are then applied to enhance discriminative capability before final classification. This structured design ensures that each component contributes explicitly to the overall performance of the system [30].

Mathematically, the model learns a mapping function.

$$f_{\theta}: \mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{R}^C \quad (1)$$

where H and W denote image height and width, C represents the number of obstacle classes, and θ corresponds to the trainable parameters.

A. Materials

1) Dataset

The dataset used in this research consists of two complementary components: secondary data obtained from publicly available sources and primary data collected independently by the researcher. The secondary dataset provides a broad range of labeled obstacle images captured under relatively controlled conditions, serving as a foundation for initial model learning.

In contrast, the primary dataset is collected using real-world cameras in natural environments, capturing obstacles under varying illumination conditions, backgrounds, and viewing angles. This dataset reflects realistic scenarios encountered by visually impaired individuals during daily navigation and is therefore crucial for evaluating the practical applicability of the proposed model.

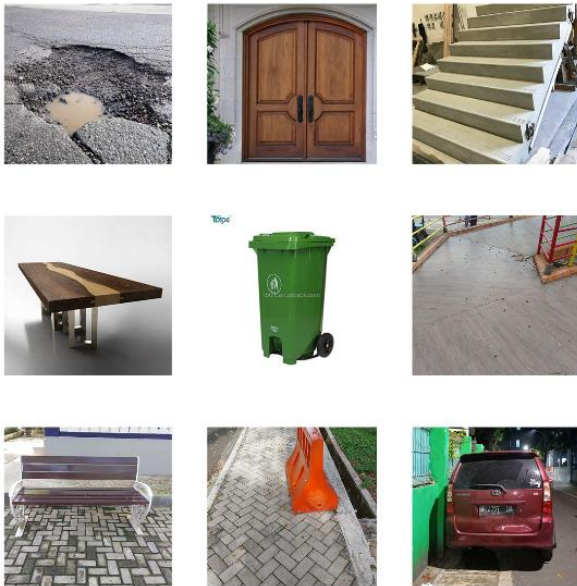


Fig. 2. Obstacle dataset.

All images are standardized to a resolution of 224×224 pixels and represented in Red, Green, Blue (RGB) color space. The dataset is divided into training, validation, and testing subsets. To prevent data leakage and overly optimistic performance estimates, the primary dataset is exclusively used during validation and testing phases, allowing the assessment of model generalization on unseen real-world data.

The dataset consists of ten obstacle classes relevant to urban navigation scenarios. Fig. 2 presents representative samples from the obstacle classes in the dataset, including chairs, doors, fences, garbage bins, obstacles, potholes, stairs, tables, and vehicles. Table I summarizes the dataset composition, including the number of samples per class across training and validation subsets.

TABLE I. DATASET COMPOSITION

Class	Training	Validation	Total
Chair	280	24	304
Door	460	51	511
Fence	110	12	122
Garbage Bin	155	17	172
Obstacle	820	91	911
Plant	215	24	239
Pothole	270	30	300
Stairs	540	60	600
Table	300	33	333
Vehicle	810	90	900
Total	3133	432	3565

2) Data preprocessing

Prior to model training, all images undergo preprocessing to ensure compatibility with the backbone network. Image resizing is performed to achieve uniform input dimensions of 224×224 pixels, while pixel values are normalized implicitly through the pre-trained backbone's internal preprocessing pipeline [31]. The data pipeline is further optimized using caching, shuffling, and prefetching techniques to reduce input latency and improve training efficiency on hardware accelerators.

3) Feature extraction

EfficientNetV2-B0 is employed as the backbone network due to its proven balance between computational efficiency and representational power [32]. The backbone is initialized with ImageNet pre-trained weights, enabling the model to leverage generic low- and mid-level visual features learned from large-scale data.

The classification layers of the backbone are removed, and all backbone parameters are frozen during training to stabilize optimization and prevent catastrophic forgetting. Given an input image I , the backbone produces hierarchical feature maps at different semantic depths:

$$\{C_3, C_4, C_5\} = f_{EffNetV2}(I) \quad (2)$$

where deeper feature maps encode higher-level semantic information and shallower feature maps preserve fine-grained spatial details [33]. These hierarchical representations form the basis for subsequent multi-scale feature fusion, allowing the model to capture obstacles of varying sizes and visual characteristics.

4) Data augmentation

To enhance generalization capability and mitigate overfitting, data augmentation techniques are applied dynamically during training [34]. Given an original image I , an augmented version I' is generated as Eq. (3):

$$I' = T(I) \quad (3)$$

where T denotes a stochastic transformation function composed of horizontal flipping, rotation, and zooming operations. These augmentation strategies encourage the model to learn invariant and robust features without altering the semantic meaning of the image.

B. Methods

The following subsections describe the foundational methods that form the basis of the proposed architecture design.

1) Multi-scale feature fusion using a FPN

To address the challenge of scale variation among obstacles, an FPN is integrated on top of the backbone network [35]. The FPN employs a top-down architecture with lateral connections, enabling high-level semantic features to be propagated to lower-resolution feature maps [36].

Unlike the conventional FPN that applies direct element-wise addition, the proposed method introduces an adaptive feature fusion mechanism to dynamically balance spatial and semantic information across pyramid levels. At each pyramid level, feature fusion is defined as Eq. (4):

$$P_l = \alpha_l \cdot \text{Conv}_{1 \times 1}(C_1) + (1 - \alpha_l) \cdot \text{Up}(P_{l+1}) \quad (4)$$

where $\text{Up}(\cdot)$ denotes spatial up-sampling and α_l is a learnable weight generated using a sigmoid activation function. This adaptive weighting mechanism allows the network to control the contribution of low-level spatial features and high-level semantic features according to the input characteristics.

Furthermore, instead of performing direct concatenation, a scale-aware fusion strategy is introduced to assign importance weights to each pyramid level before aggregation. The unified multi-scale feature representation is formulated as Eq. (5):

$$F = \text{Conv}_{3 \times 3}([\gamma_3 P_3, \gamma_4 \text{Up}(P_4), \gamma_5 \text{Up}(P_5)]) \quad (5)$$

where γ_i represents normalized importance weights obtained using a SoftMax function. This mechanism enables the model to emphasize more informative feature maps while suppressing less relevant ones across different scales.

This enhanced fusion strategy allows the model to simultaneously exploit spatial precision from shallow layers and semantic richness from deeper layers in a more adaptive and discriminative manner, thereby improving robustness in detecting obstacles with varying sizes and visual characteristics.

2) Channel-wise attention and feature refinement

To further enhance feature discrimination, an SE block is applied to the fused feature map [37, 38]. The SE mechanism performs global average pooling to capture global contextual information:

$$z_c = \frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W F_c(i, j) \quad (6)$$

The resulting channel descriptor is transformed using a bottleneck structure and sigmoid activation to produce adaptive channel weights.

Following attention-based recalibration, a refinement block based on residual learning is employed to improve feature quality and stabilize training [39]. The refinement process is formulated as Eq. (7):

$$F_{ref} = F_{in} + g(F_{in}) \quad (7)$$

where $g(\cdot)$ represents a sequence of convolutional and batch normalization layers. This residual formulation improves gradient flow and allows the network to refine features without introducing excessive complexity.

3) Classification, training strategy, and evaluation metrics

The refined feature map is passed through a global average pooling layer to reduce spatial dimensionality while preserving semantic information. A dropout layer is applied to reduce overfitting, followed by a fully connected layer with SoftMax activation to produce class probability outputs:

$$\hat{y}_k = \frac{e^{v_k}}{\sum_{j=1}^K e^{v_j}} \quad (8)$$

The model is trained using the Adam optimizer with a learning rate of 1×10^{-4} . Sparse categorical cross-entropy is employed as the loss function:

$$L = - \sum_{k=1}^K y_k \times \log(\hat{y}_k) \quad (9)$$

Early stopping and model checkpointing strategies are applied to prevent overfitting and preserve the best performing model. Model performance is evaluated primarily using classification accuracy:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^N I(\hat{y}_i = y_i) \quad (10)$$

With particular emphasis on validation and testing results obtained from the primary dataset to assess robustness in real-world scenarios.

C. Proposed Methodology

1) Problem definition

Visually impaired individuals face significant safety risks when navigating urban environments due to the presence of diverse and unpredictable obstacles. Although deep learning-based obstacle detection has shown promising results, most existing approaches either rely on

heavy architectures unsuitable for edge deployment or apply feature extraction, multi-scale fusion, and attention mechanisms as isolated components rather than as an integrated system.

Formally, the problem is defined as a multi-class image classification task. Given an input image $I \in R^{(H \times W \times 3)}$, the goal is to learn a mapping function f_θ that produces a predicted class label $\hat{y} \in \{1, \dots, C\}$ where $C = 10$, while satisfying the real-time inference constraint on resource-constrained devices.

2) System model

The proposed system is modeled as an end-to-end lightweight obstacle recognition pipeline consisting of four sequential stages: (1) input preprocessing, (2) hierarchical feature extraction, (3) adaptive multi-scale feature fusion with attention, and (4) classification output. The overall system model can be expressed as Eq. (11):

$$\hat{y} = f_{cls}(f_{ref}(f_{se}(f_{fpn}(f_{enc}(I)))))) \quad (11)$$

This sequential formulation ensures that each component contributes explicitly to the final prediction while maintaining a unified and differentiable end-to-end training objective.

3) Architecture

The proposed architecture consists of four main components integrated into a single end-to-end framework, as illustrated in Fig. 3.

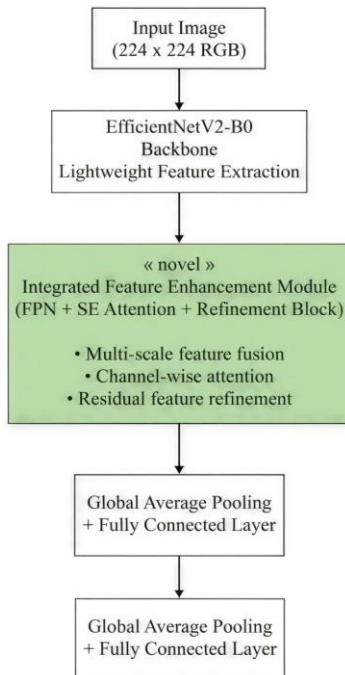


Fig. 3. General architecture.

EfficientNetV2-B0 serves as the lightweight backbone initialized with ImageNet pre-trained weights, producing three hierarchical feature maps C3, C4, and C5 at different semantic depths. These maps are fed into the adaptive FPN, which introduces learnable weighting parameters α_l at each pyramid level as defined in Eq. (4), and scale-aware

SoftMax weights across levels as defined in Eq. (5). The fused feature map F is then recalibrated channel-wise through the SE block as defined in Eq. (6), and stabilized through the residual refinement block as defined in Eq. (7). Finally, the classification head applies global average pooling, dropout, and SoftMax activation as defined in Eqs. (8)–(10) to produce final class probabilities. The model is trained using the Adam optimizer with a learning rate of 1×10^{-4} and sparse categorical cross-entropy loss.

4) Working model

The system processes an input RGB image of size $224 \times 224 \times 3$ through the frozen EfficientNetV2-B0 backbone to extract hierarchical feature maps C3, C4, and C5. The adaptive FPN then constructs a unified multi-scale feature representation by dynamically balancing spatial and semantic information using learnable weights per pyramid level and SoftMax-normalized importance weights across levels. The SE block subsequently recalibrates the fused features channel-wise to emphasize discriminative representations. The residual refinement block then improves feature stability before the classification head produces a probability distribution over the ten obstacle classes. The predicted class with the highest probability is returned as the final output. In real-time deployment, this prediction is sent directly to the mobile application interface to inform the user about the detected obstacle.

IV. RESULT AND DISCUSSION

This section presents the experimental results of the proposed obstacle recognition model based on EfficientNetV2-B0 integrated with FPN, SE attention, and a refinement mechanism. The analysis covers training performance, quantitative results, and architectural insights, with emphasis on real-time mobile deployment for assistive navigation. All experiments were conducted on an AMD Ryzen 7 5700U system with 16GB RAM (Axioo Hype 7), with training accelerated using Google Colab GPU. The model was implemented using TensorFlow and Keras frameworks. Table II summarizes the experimental configuration used in this study.

TABLE II. EXPERIMENTAL CONFIGURATION

Parameter	Value
Framework	TensorFlow/Keras
Backbone	EfficientNetV2-B0
Input Size	$224 \times 224 \times 3$
Optimizer	Adam
Learning Rate	1×10^{-4}
Loss Function	Sparse Categorical Cross-Entropy
Epochs	15
Training Images	3133
Validation Images	432
Number of Classes	10
Hardware	AMD Ryzen 7 5700U, 16GB RAM + Google Colab GPU

A. Training and Validation Results

A total of 3133 images were used for training, while 432 images were used for validation. The dataset consists of

ten obstacle classes. The training process was conducted for 15 epochs. Transfer learning was applied by using a backbone model with pre-trained weights. Training accuracy and validation accuracy were observed during each epoch.

In the first epoch, the training accuracy was 38.75%. The validation accuracy at this epoch was 84.26%. In the second epoch, the training accuracy became 81.99%. The validation accuracy increased to 95.83%.

After several epochs, the accuracy values continued to increase. At the sixth epoch, the training accuracy reached 96.10%. The validation accuracy also reached its highest value at this epoch, which was 98.84%. The validation loss at this point was recorded as 0.0476. Afterward, the performance remained stable, and the validation accuracy stayed above 97% until the final epoch.

The comparison between training and validation accuracy across epochs is shown in Fig. 4. The training accuracy curve and validation accuracy curve show a similar trend during the training process. No large difference between the two curves was observed.



Fig. 4. Training vs test accuracy graph.

B. Classification Performance

The final trained model was further evaluated using precision, recall, and F1-Score for each obstacle class. Table III summarizes the classification performance across all 10 classes.

TABLE III. CLASSIFICATION REPORT

Class	Precision	Recall	F1-Score	Support
Chair	1.000	1.000	1.000	24
Door	1.000	0.9804	0.9901	51
Fence	1.000	0.9167	0.9565	12
Garbage bin	0.8947	1.000	0.9444	17
Obstacle	1.000	1.000	1.000	91
Plant	1.000	1.000	1.000	24
Pothole	1.000	0.9667	0.9831	30
Stairs	0.9524	1.000	0.9756	60
Table	1.000	0.9697	0.9846	33
Vehicle	1.000	0.9889	0.9944	90
Accuracy	-	-	0.9884	432
Macro Avg.	0.9847	0.9822	0.9829	432
Weighted Avg.	0.9892	0.9884	0.9885	432

The model achieved an overall classification accuracy of 98.84% on the validation dataset. The weighted average precision, recall, and F1-Score reached 98.92%, 98.84%, and 98.85%, respectively, indicating highly consistent performance across classes.

Most obstacle categories, including chair, obstacle, plant, and vehicle, achieved near-perfect precision and recall values. Slight performance variation was observed in classes such as fence and garbage bin, which can be attributed to visual similarity with surrounding background objects. Nevertheless, the F1-Scores for these classes remained above 94%, demonstrating robust recognition capability.

Among the ten classes, chair, obstacle, and plant achieved perfect scores across all metrics, reflecting their high visual distinctiveness and sufficient sample representation within the dataset. Door, pothole, and vehicle also achieved F1-Scores above 98%, confirming strong recognition capability for these commonly encountered urban obstacles.

Fence achieved a recall of 0.9167, indicating that one validation sample was misclassified. This is likely due to visual similarity between fence structures and background elements such as walls or railings, particularly under varying viewpoint and lighting conditions. Garbage Bin achieved a precision of 0.8947, suggesting that samples from other classes were occasionally misclassified as Garbage Bin. This confusion can be attributed to the irregular shape and variable appearance of garbage bins across different environments. Despite these minor variations, both classes maintained F1-Scores above 94%, confirming that the proposed model handles inter-class confusion well overall.

C. Confusion Matrix Analysis

After presenting numerical metrics, the confusion matrix is used to analyze misclassification patterns in detail.

The confusion matrix is defined as Eq. (12):

$$M \in R^{C \times C} \quad (12)$$

where C is the number of classes and each element $M_{i,j}$ represents the number of samples from ground-truth class i predicted as class j .

In this study, a 10×10 confusion matrix is constructed based on the validation dataset. The visualization is shown in Fig. 5.

Most predictions are concentrated along the main diagonal, indicating a high number of correct classifications. Minor confusion occurs between visually similar classes, such as *fence* and *stairs*, but these errors are limited and do not significantly affect overall performance.

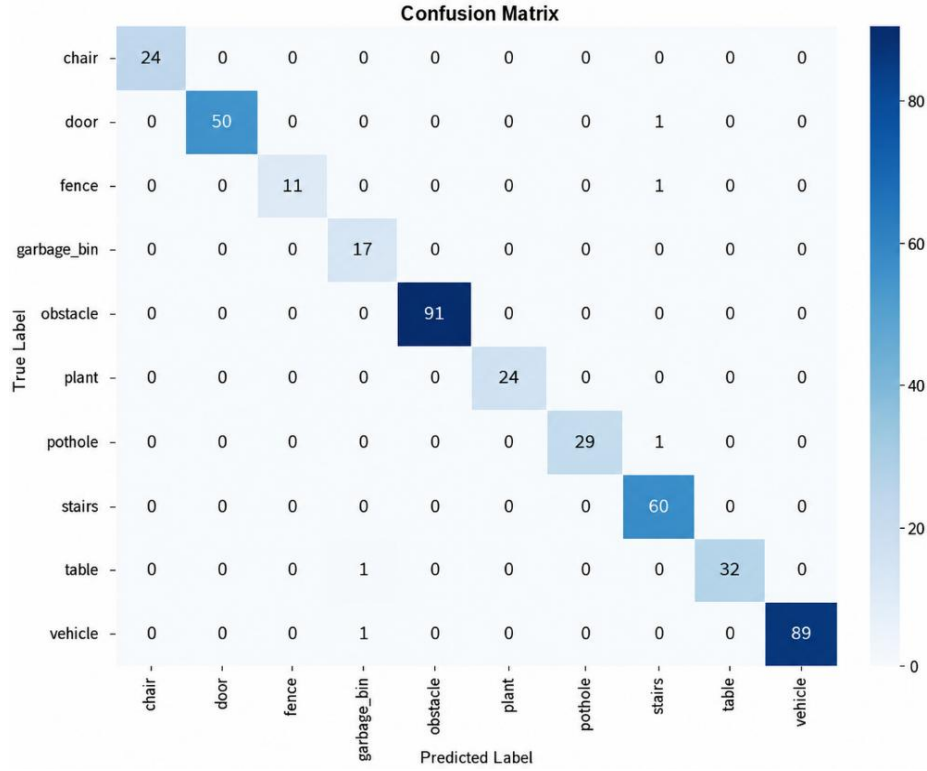


Fig. 5. Confusion matrix.

D. Real-Time Deployment and Case Study

To evaluate the practical applicability of the proposed system, a real-time deployment was conducted using a mobile-based implementation. The trained model was integrated into an Android application that utilizes live camera input for continuous obstacle detection and classification, following the system pipeline illustrated in Fig. 6.

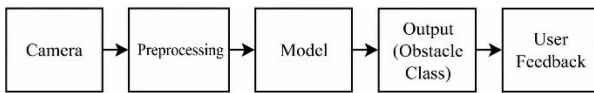


Fig. 6. Proposed system pipeline.

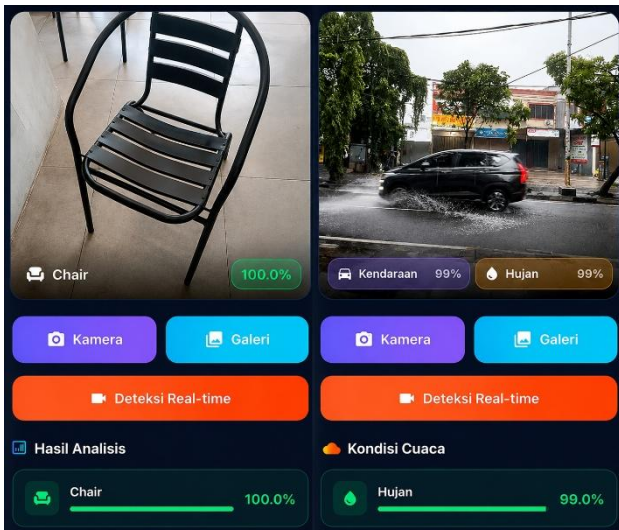


Fig. 7. Real-time obstacle detection on mobile.

The system processes each frame captured by the device camera and performs inference using the proposed deep learning architecture. The output is then used to inform the user about nearby obstacles, enabling real-time navigation assistance.

The mobile application interface, as shown in Fig. 7, demonstrates the system's ability to provide real-time obstacle classification in a practical setting.

Three case study scenarios were designed to simulate visually impaired navigation conditions across different environments. In each scenario, a participant with vision obstructed by cloth navigated a designated path relying solely on the system output, mimicking real-world usage conditions. All scenarios were conducted on the same mobile device (MediaTek Helio G96, 8 GB RAM) with a consistent inference speed of approximately 1 second per frame (~1 Frame Per Second (FPS)). Table IV summarizes the scenario configurations and results.

TABLE IV. CASE STUDY RESULTS

No.	Environment & Lighting	Target Obstacles	Result	Accuracy (%)
1	Indoor, normal lighting	Table, Chair, Garbage Bin	10/10	100
2	Outdoor, natural daylight	Plant	10/10	100
3	Indoor, dim lighting	Chair, Table	9/10	90

In Scenario 1, the user navigated a 10-meter indoor path under normal lighting containing tables, chairs, and garbage bins. The system successfully detected all obstacles with 100% accuracy and the user avoided all objects without collision. In Scenario 2, the user navigated

a 5 m outdoor path under natural daylight directed toward a plant obstacle, also achieving 100% detection accuracy, demonstrating robustness under bright ambient illumination. In Scenario 3, the user navigated a 5 m indoor path under dim lighting, directed toward chairs and tables. Nine out of ten trials were correctly classified, yielding 90% accuracy. The single misclassification is attributed to insufficient ambient lighting, which reduced the visual contrast of obstacle edges.

Overall, the system achieved an average case study accuracy of 96.67% across all three scenarios, confirming reliable performance across varied environmental conditions. However, several limitations must be acknowledged: the evaluation involved sighted participants with obstructed vision rather than actual visually impaired individuals, the test environments were relatively controlled, and the inference speed of ~ 1 FPS may be insufficient for highly dynamic environments. A more comprehensive evaluation involving actual visually impaired participants across diverse environments is planned as future work.

E. Discussion

The classification results obtained in this study are closely related to the integrated design of the proposed architecture. Unlike conventional approaches that apply backbone networks, feature fusion, and attention mechanisms independently, this study emphasizes the interaction between multiple components within a unified lightweight framework.

EfficientNetV2-B0 serves as the backbone network to ensure computational efficiency while maintaining strong feature extraction capability. The integration of the FPN enables effective multi-scale feature representation,

allowing the model to handle obstacles with varying sizes and spatial characteristics. The SE mechanism further improves the model by adaptively recalibrating channel-wise features, enabling the network to focus on more informative representations. In addition, the refinement block contributes to improving feature stability and representation quality through residual learning.

In addition to architectural integration, this study introduces an adaptive feature fusion mechanism within the FPN. Unlike conventional FPN implementations that rely on fixed element-wise operations, the proposed approach employs learnable weighting parameters to dynamically balance the contribution of spatial and semantic features across pyramid levels. Furthermore, scale-aware weighting is applied during multi-scale aggregation to emphasize more informative feature maps while suppressing less relevant ones. This adaptive strategy enhances the flexibility of feature representation and improves robustness in complex environments.

To further validate the effectiveness of the proposed approach, a baseline model using EfficientNetB0 combined with SE attention was evaluated. The baseline achieved a validation accuracy of 23.33%, which is significantly lower than the performance obtained by the proposed model. This result indicates that the backbone and attention mechanism alone are insufficient, and that the integration of multi-scale feature fusion and refinement plays a critical role in improving classification performance.

Table V summarizes the comparative performance between the baseline and the proposed model.

The 75.51% improvement confirms that architectural integration and adaptive feature fusion are critical factors in achieving reliable obstacle recognition performance.

TABLE V. COMPARATIVE PERFORMANCE

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
EfficientNetB0 + SE (Baseline)	23.33	-	-	-
Proposed (EfficientNetV2-B0 + Adaptive FPN + SE + Refinement)	98.84	98.92	98.84	98.85

Overall, the proposed framework demonstrates that combining lightweight architecture design with adaptive multi-scale feature fusion can effectively address challenges related to scale variation, feature redundancy, and representation stability. Although the achieved accuracy is high, the primary contribution of this study lies not only in performance improvement but also in the proposed integration strategy and adaptive feature fusion mechanism. These findings highlight the potential of the proposed approach for deployment in real-time assistive navigation systems on resource-constrained devices.

V. CONCLUSION

Obstacle recognition is an important task for assistive systems used by visually impaired people. This study focuses on classifying obstacles using a deep learning-based approach. The proposed model is built using EfficientNetV2-B0 as the backbone and combined with FPN, SE attention, and a refinement block. These components help the model extract visual features from

obstacles of different sizes. At the same time, the overall model design remains lightweight and suitable for practical use.

The model was tested using a dataset with ten different obstacle classes. Based on the experiments, a validation accuracy of 98.84% was obtained. Model performance was measured using accuracy, precision, recall, and F1-Score. The evaluation shows that most obstacle categories can be classified correctly. Multi-scale feature extraction supports the recognition of objects with varying sizes, while the attention module assists the model in focusing on important features.

The experiments were carried out using both public datasets and real-world images. This shows that the proposed approach can work on data obtained from different sources. However, there is still scope for improvement. Future work will explore several directions to further improve the proposed system. First, the framework will be extended to support object detection with bounding box regression for more precise obstacle localization. Second, the dataset will be expanded to cover

more obstacle types and diverse environmental conditions including nighttime and crowded scenes. Third, model optimization techniques such as quantization and TensorFlow Lite conversion will be applied to improve inference speed beyond the current ~1 FPS on mobile devices. Fourth, integration with audio feedback or haptic notification will be developed to enable hands-free navigation assistance. Finally, evaluation involving actual visually impaired participants across uncontrolled real-world environments is planned to provide a more rigorous assessment of practical usability.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

H.S. provided the idea, conceptualized the research problem, designed the overall research framework, provided ongoing supervision throughout the research, provided the mathematical concept of the research, and acquired the financial support for the project leading to this publication. V.A.P. performed dataset collection and pre-processing, developed and implemented the proposed obstacle recognition model using EfficientNetV2-B0 integrated with FPN and SE mechanism, conducted the experiments, and analyzed the results. In addition, V.A.P. designed the training strategy and network regularization scheme by combining dropout layers, batch normalization, and max pooling to reduce overfitting and improve model generalization. V.A.P. prepared the methodology and experimental results. H.S. prepared and reviewed the original manuscript, editing and refining the manuscript to improve technical accuracy, clarity of experimental results, and scientific rigor, specifically critical review, commentary or revision—including pre- or post-publication stages. M.A. verified, either as part of the work or separately, the replication/reproducibility of the results/experiments and other research outputs as a whole. M. supervised the planning and implementation of research activities. I.F. provided comments, criticism, and suggestions for this research. All authors read and approved the final version of the manuscript.

FUNDING

This research was funded by the Institute for Research and Community Service, Mercu Buana University, Jakarta. This research is an Internal Lecturer Grant under the Faculty of Computer Science International Cooperation Scheme. This research is a collaboration between the Faculty of Mercu Buana University and Yuan Ze University.

ACKNOWLEDGMENT

The researchers would like to thank the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia, the Institute for Research and Community Service, Mercu Buana University, for their generous support in carrying out this research, and the

Faculty of Computer Science, Mercu Buana University, Jakarta.

REFERENCES

- [1] M. Hasanudin *et al.*, “Deep learning-based autism detection using facial images and EfficientNet-B3,” *J. Tek. Inform. Jutif*, vol. 7, no. 1, pp. 81–92, Feb. 2026. doi: 10.52436/1.jutif.2026.7.1.4574
- [2] J. Wang, S. Wang, and Y. Zhang, “Artificial intelligence for visually impaired,” *Displays*, vol. 77, 102391, Apr. 2023. doi: 10.1016/j.displa.2023.102391
- [3] World Health Organization. (2025). Vision impairment and blindness. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment>
- [4] E. Casanova, D. Guffanti, and L. Hidalgo, “Technological advancements in human navigation for the visually impaired: A systematic review,” *Sensors*, vol. 25, no. 7, 2213, Apr. 2025. doi: 10.3390/s25072213
- [5] I. Patel, M. Kulkarni, and N. Mehendale, “Review of sensor-driven assistive device technologies for enhancing navigation for the visually impaired,” *Multimed. Tools Appl.*, vol. 83, no. 17, pp. 52171–52195, Nov. 2023. doi: 10.1007/s11042-023-17552-7
- [6] D. H. Hareva, A. Sebastian, A. R. Mitra, I. A. Lazarusli, and C. A. Haryani, “Mobile surveillance siren against moving object as a support system for blind people,” *J. Image Graph.*, pp. 170–177, Jun. 2023. doi: 10.18178/joig.11.2.170-177
- [7] A. Maurya and P. Verma, “Optimized deep CNN based obstacle detection for aiding visually impaired persons,” *Investigación Operacional*, vol. 45, no. 1, pp. 71–79, 2024.
- [8] K. Jouini, M. H. Maaloul, and O. Korbaa, “Real-time, CNN-based assistive device for visually impaired people,” in *Proc. 2021 14th Int. Congr. Image Signal Process., BioMed. Eng. Informat. (CISP-BMEI)*, 2021, pp. 1–6. doi: 10.1109/CISP-BMEI53629.2021.9624387
- [9] F. Ashiq *et al.*, “CNN-based object recognition and tracking system to assist visually impaired people,” *IEEE Access*, vol. 10, pp. 14819–14834, 2022. doi: 10.1109/ACCESS.2022.3148036
- [10] A. Singh and V. Vishwakarma, “CNN based currency reader for visually impaired,” *International Journal of Software & Hardware Research in Engineering (IJSHRE)*, vol. 11, no. 10, 2023. <https://journals.in/wp-content/uploads/2023/10/1.IJSHRE-111002-Aviraj.pdf>
- [11] H. Alhichri, Y. Bazi, N. Alajlan, and B. Bin Jdara, “Helping the visually impaired see via image multi-labeling based on SqueezeNet CNN,” *Appl. Sci.*, vol. 9, no. 21, 4656, Nov. 2019. doi: 10.3390/app9214656
- [12] M. Rifki, A. Bastian, and A. Mardiana, “Development of CNN-LSTM-based image captioning dataset to enhance visual accessibility for disabilities,” *JITK J. Ilmu. Pengetah. Dan Teknol. Komput.*, vol. 10, no. 4, pp. 980–992, Jun. 2025. doi: 10.33480/jitk.v10i4.6657
- [13] M. Al Duhayyim, “Internet of things enabled indoor activity monitoring for visually impaired people with hybrid deep learning and optimized algorithms for enhanced safety,” *Sci. Rep.*, vol. 15, no. 1, 35438, Oct. 2025. doi: 10.1038/s41598-025-19030-6
- [14] I. H. Hsieh, H. C. Cheng, H. H. Ke, H. C. Chen, and W. J. Wang, “A CNN-based wearable assistive system for visually impaired people walking outdoors,” *Appl. Sci.*, vol. 11, no. 21, 10026, Oct. 2021. doi: 10.3390/app112110026
- [15] A. H. Nurfauzi, Y. Azhar, and D. R. Chandranegara, “Penerapan model EfficientNetV2-B0 pada benchmark IP102 dataset untuk Menyelesaikan Masalah Klasifikasi Hama Serangga,” *Jurnal Repositor*, vol. 5, no. 3, 2023.
- [16] N. McMahon, E. M. Grua, and C. Eising, “Leveraging EfficientNetV2 for litter detection,” in *Proc. IET Conference Proceedings CP887*, 2024, vol. 2024, no. 10, pp. 315–318. doi: 10.1049/icp.2024.3322
- [17] Z. Zhao, E. B. A. Bakar, N. B. A. Razak, and M. N. Akhtar, “Progressive optimization of EfficientNetV2 for classification based on images of corroded objects,” *Comput. Appl. Math.*, vol. 44, no. 8, 419, Dec. 2025. doi: 10.1007/s40314-025-03382-w
- [18] R. Kridalukmana, D. Eridani, R. Septiana, and I. P. Windasari, “YoloV8_s, EfficientNetv2_s, and CSP_Darknet_l comparison as recognition model’s backbone for drone-captured images,” *JOIV*:

- International Journal on Informatics Visualization*, vol. 9, no. 2, pp. 683–689, 2025.
- [19] T. Njoroge, R. Kibuku, and K. Mugoye, “Comparative and edge-hybrid modeling of EfficientNetV2 and MobileNetV2 for multi-class crop disease classification with statistical validation,” *J. Edge Comput.*, vol. 4, no. 2, pp. 234–262, Nov. 2025. doi: 10.55056/jec.905
 - [20] J. Yin, Y. Lyu, K. Suo, and X. Huang, “PCB defect detection algorithm based on EfficientNetV2,” *Journal of Computer-Aided Design & Computer Graphics*, vol. 37, no. 7, pp. 1260–1269, Jul. 2025. doi: 10.3724/SP.J.1089.2023-00551
 - [21] M. Zhu, K. Han, C. Yu, and Y. Wang, “Dynamic feature pyramid networks for object detection,” in *Proc. Fifteenth International Conference on Signal Processing Systems (ICSPS 2023)*, 2024, vol. 13091, pp. 503–511.
 - [22] J. Gao, Y. Zhang, and X. Geng, “Augmented weighted bidirectional feature pyramid network for marine object detection,” *Expert Systems with Applications*, vol. 237, 121688, 2024.
 - [23] S. Karthikeyan, R. Charan, S. Narayanan, and L. Jani Anbarasi, “Enhanced plant disease classification with attention-based convolutional neural network using squeeze and excitation mechanism,” *Front. Artif. Intell.*, vol. 8, 1640549, Aug. 2025. doi: 10.3389/frai.2025.1640549
 - [24] A. Y. Ashurov *et al.*, “Enhancing plant disease detection through deep learning: A depthwise CNN with squeeze and excitation integration and residual skip connections,” *Front. Plant Sci.*, vol. 15, 1505857, Jan. 2025. doi: 10.3389/fpls.2024.1505857
 - [25] I. Pacal, O. Celik, B. Bayram, and A. Cunha, “Enhancing EfficientNetv2 with global and efficient channel attention mechanisms for accurate MRI-based brain tumor classification,” *Clust. Comput.*, vol. 27, no. 8, pp. 11187–11212, Nov. 2024. doi: 10.1007/s10586-024-04532-1
 - [26] B. V. Mone, B. Purnama, and B. Erfianto, “Classification badminton poses class intra-variation using MoveNet and ConvLSTM with improve SE block for player action analysis,” *Int. J. Intell. Eng. Syst.*, vol. 18, no. 3, pp. 273–289, Apr. 2025. doi: 10.22266/ijies2025.0430.19
 - [27] N. Sengodan, “Breast cancer histopathology classification using CBAM-EfficientNetV2 with transfer learning,” arXiv Preprint arXiv:2410.22392, 2025. <https://arxiv.org/abs/2410.22392>
 - [28] S. Zhou, D. Chen, J. Pan, J. Shi, and J. Yang, “Adapt or perish: Adaptive sparse transformer with attentive feature refinement for image restoration,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2024, pp. 2952–2963. doi: 10.1109/CVPR52733.2024.00285
 - [29] Z. Zhao, D. Wang, and X. Zhao, “M3Net: Movement enhancement with multi-relation toward multi-scale video representation for temporal action detection,” *Pattern Recognit.*, vol. 155, 110702, Nov. 2024. doi: 10.1016/j.patcog.2024.110702
 - [30] J. Li, M. Zheng, D. Dong, and X. Xie, “PC-CS-YOLO: High-precision obstacle detection for visually impaired safety,” *Sensors*, vol. 25, no. 2, 534, Jan. 2025. doi: 10.3390/s25020534
 - [31] M. Tan and Q. V. Le, “EfficientNetV2: Smaller models and faster training,” in *Proc. International Conference on Machine Learning*, 2021, pp. 10096–10106.
 - [32] G. Karthiha and S. Allwin, “Transfer learning approaches for EfficientNetV2 B0 and ImageNet skin cancer classification in a convolutional neural network,” *PeerJ Comput. Sci.*, vol. 11, e3362, Dec. 2025. doi: 10.7717/peerj-cs.3362
 - [33] J. Wang, P. Yin, W. Yang, Y. Wang, and S. Wang, “Exploiting multi-scale hierarchical feature representation for visual tracking,” *Complex Intell. Syst.*, vol. 10, no. 3, pp. 3617–3632, Jun. 2024. doi: 10.1007/s40747-024-01345-y
 - [34] T. An, R. Zhou, F. Qin, Y. Dai, Y. Gong, and P. Chen, “Comparative study of the parameter acquisition methods for the Cauer thermal network model of an IGBT module,” *Electronics*, vol. 12, no. 7, 1650, Mar. 2023. doi: 10.3390/electronics12071650
 - [35] H. Wang and T. Wang, “Multi-scale residual aggregation feature pyramid network for object detection,” *Electronics*, vol. 12, no. 1, 93, Dec. 2022. doi: 10.3390/electronics12010093
 - [36] C. Wang and C. Zhong, “Adaptive feature pyramid networks for object detection,” *IEEE Access*, vol. 9, pp. 107024–107032, 2021. doi: 10.1109/ACCESS.2021.3100369
 - [37] H. Cao *et al.*, “Enhanced specific emitter identification using multi-sampling convolutional attention network,” in *Proc. 2024 16th Int. Conf. Wireless Commun. Signal Process (WCSP)*, 2024, pp. 602–607. doi: 10.1109/WCSP62071.2024.10827072
 - [38] M. Hayat, “Squeeze & excitation joint with combined channel and spatial attention for pathology image super-resolution,” *Frankl. Open*, vol. 8, 100170, Sep. 2024. doi: 10.1016/j.fraope.2024.100170
 - [39] S. Zhang, T. Ma, and Y. Wen, “RC-block: Refinement coefficient for rectifying deformation field,” in *Proc. 2024 IEEE Int. Conf. Multimedia Expo (ICME)*, 2024, pp. 1–6. doi: 10.1109/ICME57554.2024.10688036

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).