International Journal of Advanced Robotic Syst

semantic

Efficient Sematic Segmentation Network for Mobile Robot Path Planning

Journal:	*International Journal of Advanced Robotic Systems*
Manuscript ID	ARX-24-0203
Manuscript Type:	Research Article
Keywords:	Semantic segmentation, convolution neural network, path planning, binary classification, mobile robots
ARX Research Topics:	COMPUTER VISION < VISION SYSTEMS, SEGMENTATION AND GROUPING < VISION SYSTEMS, VISION-BASED NAVIGATION AND SLAM < VISION SYSTEMS, COLLISION AVOIDANCE AND MULTI-VEHICLE SYSTEMS < ROBOT MANIPULATION AND CONTROL, PATH PLANNING AND NAVIGATION < MOBILE ROBOTS AND MULTI-ROBOT SYSTEMS, MOBILITY AND MOTION PLANNING < SERVICE ROBOTICS, VISION SENSORS < ROBOT SENSORS AND SENSOR NETWORKS
Abstract:	One of the key challenges in semantic segmentation for path planning of mobile robots (MR) lies in deploying segmentation models on edge devices or embedded systems with limited processing capabilities. This research applies a Binary Neural Network (BNN)-based semantic segmentation model to images captured by a monocular camera, with the model's backbone being Resnet18 and incorporating a decoder based on the PSP-Net architecture. By reducing processing parameter magnitudes and optimizing network propagation, this approach enhances training efficiency, speeds up inference, and reduces hardware requirements. The outcomes of this methodology outperform those of conventional CNN architectures, particularly with a notable 3.11-fold increase in inference velocity while maintaining an accuracy of 6.21%. Furthermore, BNNs enable the deployment of semantic segmentation models on hardware with restricted physical or resource capacities. The proposed method achieved IoU and Dice scores of 83.15% and 88.2% respectively, with throughput improvements of 1.6 to 2.5 times and latency reduced to 0.2 times the average of other models evaluated. In conclusion, this architectural framework shows promise for integration into intelligent mechatronic systems, especially in advancing knowledge systems for robotic navigation in diverse environments.

SCHOLARONE™
Manuscripts

Article

Efficient Sematic Segmentation Network for Mobile Robot Path Planning

Abstract: One of the key challenges in semantic segmentation for path planning of mobile robots (MR) lies in deploying segmentation models on edge devices or embedded systems with limited processing capabilities. This research applies a Binary Neural Network (BNN)-based semantic segmentation model to images captured by a monocular camera, with the model's backbone being Resnet18 and incorporating a decoder based on the PSP-Net architecture. By reducing processing parameter magnitudes and optimizing network propagation, this approach enhances training efficiency, speeds up inference, and reduces hardware requirements. The outcomes of this methodology outperform those of conventional CNN architectures, particularly with a notable 3.11-fold increase in inference velocity while maintaining an accuracy of 6.21%. Furthermore, BNNs enable the deployment of semantic segmentation models on hardware with restricted physical or resource capacities. The proposed method achieved IoU and Dice scores of 83.15% and 88.2% respectively, with throughput improvements of 1.6 to 2.5 times and latency reduced to 0.2 times the average of other models evaluated. In conclusion, this architectural framework shows promise for integration into intelligent mechatronic systems, especially in advancing knowledge systems for robotic navigation in diverse environments.

Keywords: Semantic segmentation, convolution neural network, path planning, binary classification, mobile robots.

For Peer Review

1. Introduction

In recent decades, researchers from academia and industry have shown considerable interest in mobile robots (MRs) due to their wide range of applications in various fields such as industry, agriculture, and the military [1]. The focus of path planning lies in guiding MRs through dynamic and intricate environments while balancing safety requirements with the need for efficiency [2]. The advancement of robust path planning algorithms is crucial for the continuous innovation necessary to meet the evolving demands of practical MR implementations. MRs rely on data captured by cameras, particularly through semantic segmentation, which poses a significant challenge in vision-based applications when utilizing deep learning (DL) [3].

Semantic segmentation (SS) is expected to enhance object and area classification through sensor and camera systems as imaging technology progresses [4]. However, achieving high-quality segmented images from high-resolution cameras necessitates expertise in both computer science and remote sensing. Various object detection models, such as semi-surveillance and three-dimensional models using Lidar or sensor systems, are employed, each with its own limitations [5]. Challenges exist in segmenting images from monocular cameras, despite their advantages in accessibility and cost-effectiveness [6]. Deploying convolutional networks in intelligent embedded devices presents obstacles related to computational resources, overfitting, and overall system efficiency. Further development is required for neural networks to meet standards of optimal architecture, training efficiency, inference speed, performance accuracy, and ease of implementation on peripheral devices [7]. Addressing these challenges involves implementing a binary neural network (BNN) architecture that utilizes triggers and weights to perform binary operations on outputs, converting conventional convolutional networks (CNNs) to BNNs [8]. A decoder-encoder architecture is utilized as a lightweight model for semantic segmentation, allowing effective comprehension and representation of image data to generate accurate semantic segmentation maps. Encoders extract attributes from input images using filters to create characteristic maps, while decoders resize encoded feature maps back to their original dimensions. This process results in the production of a semantic segmentation map by predicting semantic labels for each pixel [9].

The paper proposes a novel lightweight network architecture founded on the Pyramid Scene Parsing network (PSP) integrated with binary convolutional (binary conv) layers [10]. To elaborate, the initial stratum comprises the convolutional (conv) backbone classes that leverage Resnet18 [11] to extract the input data attributes. Subsequent to traversing a pyramid aggregation module, the features are segregated into distinct dimensions, thereby enhancing the model's capacity to acquire data of varying scales. The adoption of binary conv layers instead of traditional 2D convolutional layers within the decoder accelerates both training and inference speeds [12]. A visual representation illustrates the binary convolutional operation of the output, which is then merged with binary conv and passed to the final layer. Furthermore, adjustments will be made to the activation functions to incorporate the binary conv layer. In essence, the prediction layer utilizes synthesis to generate the final segment map. Consequently, MRs can employ local search algorithms from a frontal perspective to detect obstacles and navigate a global path, thus facilitating real-time movement.

The primary contributions can be outlined as follows:

- Binary-SegNet: Fusion of binary conv FCN decoder with PSPNet model's encoder utilizing Resnet18 to achieve efficient lightweight semantic segmentation.

-The amalgamation of Resnet18 architecture with PSP model and Binary Neural Network (BNN) enhances accuracy, reduces model dimensions, and boosts training and inference speeds.

- The incorporation of the Adam optimizer further enhances performance and computational efficiency. Moreover, data preprocessing is refined through Gaussian filters.

- Based on Binary-Segnet model, optimal global path planning (GPP) is enhanced safety in local path planning (LPP) using adaptive DWA [13]. The optimized MR's navigation strategy is maintained stabe with the steering angle variation less than 0.2 rad.

- Experimental results conducted on Cityscape, Pascal VOC, a self-collected TQB-, and HaUI datasets demonstrate the proposed model's superiority over existing methods for MR's navigation [13] using monocular cameras.

The ensuing segments of the research paper are organized in the following manner: Section 2 elucidates the present state-of-the-art methodologies. The proposed framework of the Binary-SegNet alongside the corresponding network training procedures is delineated in Section 3. Section 4 comprehensively details both simulated and empirical experiments. Finally, Section 5 encapsulates the paper and outlines prospective avenues for future research advancements.

2. Related Works

Significant advancements have been achieved in the field of computer vision recently, encompassing both the broader scope and the specific domain of semantic segmentation [14]. Various architectural alternatives exhibit unique merits and demerits, prompting a detailed comparison of methodologies. This serves to highlight the compatibility and effectiveness of the proposed approach, notable for its capacity to leverage GPU computational power without compromising parameter count, thereby enhancing operational efficiency. Models like U-Net, FCN, and PSP-Net are typical examples constructed on CNNs featuring an encoder-decoder configuration [15]. The encoder comprises multiple convolution and max pooling layers to decrease image dimensions and increase feature characterization (e.g., VGG-16 [2,3,16], ResNet18 [11], etc.), while the decoder combines convolutional layers with upsampling to augment image size and provide semantic predictions for individual pixels.

For Peer Review

For Peer Review

2.1. Fully Convolutional Networks

Fully Convolutional Networks (FCNs) [17] undertake pixel classification using a softmax layer, omitting the final composite layer to preserve input image resolution integrity. Due to their computational efficiency and simplicity, FCNs effectively process large images in their entirety. The encoder of FCNs captures source image attributes and diminishes output size, with permutation convolution across multiple classes generating decoders for precise segmentation mapping. However, the image resolution may fall short compared to more recent models like U-Net [18].

2.2. Unet

The distinctive segmentation capability of U-Net stems from its encoder-decoder structure, facilitating high-resolution processing of input images [18]. Feature maps are produced through convolutional layers that decrease in size but increase in channel count. The original image is reconstructed in the transposed convolutional layer, forming the characteristic "U" shape of U-Net as encoder and decoder components interconnect. Nonetheless, processing images with large dimensions poses challenges for U-Net, necessitating balanced and substantial training data [19].

2.3. Pyramid Pyramid Sence Parsing

PSP-Net employs the pyramid pooling model to extract multi-level information from input images, combining data from different levels to classify each pixel accurately. Empirical studies and practical applications have validated the effectiveness of this model, with networks like ResNet [20] or VGG [2,16] commonly serving as foundational components for characteristic extraction tasks. The decoder design incorporates multiple pyramid pooling blocks to extract varied attributes from input images, leading to precise segment map generation. However, the complexity of PSP-Net's architecture [21] compared to alternative networks demands robust training datasets for optimal performance.

In summary, the evaluated models offer unique advantages tailored to specific semantic segmentation needs. To serve as a foundational guideline for MRs, it was crucial to develop models that strike a balance between computational efficiency, lightweight design, and exceptional precision and performance.

3. Proposed Binary-Segnet

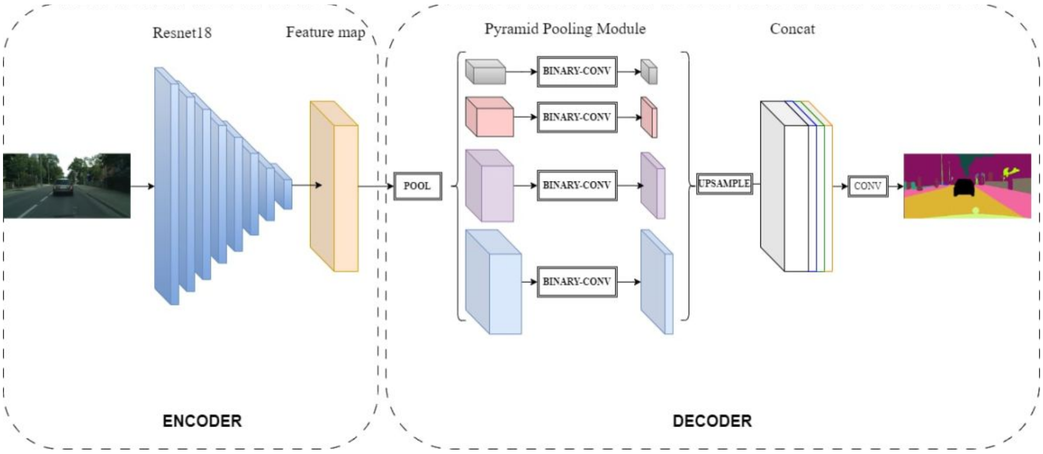


Figure 1. The Binary-SegNet architecture.

Building upon the foundational principles of Binary Neural Networks (BNNs), emphasizing binary convolutional layers, this paper presents the Binary-Segnet model as an advanced version of the Pyramid Scene Parsing Network. The PSPNet, following an encoder-decoder architecture, efficiently handles segmentation tasks. Encoders, which are Convolutional Neural Networks (CNNs), specialize in extracting input data, with commonly used networks such as ResNet [20], VGG-16 [2,16], among others. The decoder in PSP-Net acquires and comprehends information across different contexts by amalgamating dilated pyramids [10,21]. The model's remarkable precision has been demonstrated in various training and inference procedures across diverse datasets. However, the model's size and its ability to utilize PSP-Net on devices with limited configurations present challenges. Through the optimization of the model towards parametric binary, this limitation can be mitigated. Specifically, the approach involves replacing the weighted two-dimensional convolutional layers with binary convolutions, activated by real-type values in combination with binary convolutions classes and layers. The architecture of Binary-SegNet is depicted in Figure 1.

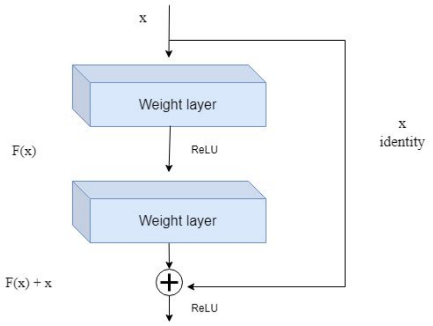


Figure 2. A residual block in the Resnet network [20].

The proposed model is constructed employing the ResNet18 architecture as the encoder component. It represents a sophisticated neural network framework predicated on the principles of residual learning. Each residual unit encompasses two convolutional layers interconnected via a skip connection. A depiction of a residual unit is presented in Figure 2.

By incorporating an input value x into the output of the layer, this configuration somewhat mitigates the degradation of contextual information and enhances the continuity among layers. Concurrently, the inclusion of the residual value inhibits the activation function from converging to zero, thereby averting the propagation of the vanishing gradient issue. The subsequent layers facilitate the model's capacity to efficiently capture features by enabling the system to assimilate the transformations pertinent to the input data rather than necessitating a learning process from the foundational level. MaxPooling layers are integrated to progressively diminish the size of the input while augmenting the dimensionality of the feature representation, thereby generating the foundational feature map. Crucial information is accentuated for optimal extraction. Initially conceived as an efficient network, ResNet18 employs a markedly reduced number of parameters compared to alternative backbone architectures when processing two-dimensional images.

The juxtaposition of the parameter counts is depicted as presented in Table 1. It is evident that, alongside the validated competitive performance demonstrated through various projects and research endeavors, the optimal computational efficiency of resnet18 is also significant. This observation is pragmatically relevant for its deployment as a foundational architecture in systems with limited resources in a cost-effective manner.

Table 1. The comparison of the proposed Binary-SegNet with the different model based on Cityscape dataset [22].

Model	Number of Parameters (million)
Resnet-18 [11,23]	11.24
Resnet-50 [24]	23.901

For Peer Review

Resnet-101 [25]	42.820
Resnet-152 [26]	58.24
VGG-16 [2,3,16]	134.7
DenseNet-169 [27]	12.8

The decoder block is constructed on the principles of a Pyramid Pooling module. Specifically, the features derived from the high-level extracted layers undergo processing and synthesis across multiple scales. The data originating from the encoder is segmented into decoding regions of varying dimensions. Four distinct partitions are designated to handle the decoding of characteristic information. The dimensions of these blocks are calibrated in accordance with the tensor size received from the encoder. In par-ticular, the model suggests employing convolutional kernels of dimensions 1x1, 2x2, 3x3, and 6x6 respectively, with characteristic scales that increment systematically. Average Pooling layers are utilized to mitigate noise while concentrating on the features intended for extraction. Subsequently, the acquired information undergoes dimensional homogenization via an Upsampling layer. The Concat block assumes the responsibility of synthesizing this information to generate the ultimate prediction map.

The model's remarkable accuracy has been substantiated through its implementation in training and inference operations across a diverse array of datasets. The magnitude of the model and its ability to leverage the PSP-Net on devices with configuration limitations present significant challenges. By optimizing the model towards parametric binary configurations, the aforementioned challenges can be addressed. In particular, we undertake the substitution of the weighted two-dimensional convolutional layers, activating them with real-valued parameters in conjunction with binary convolution classes and layers. Consistent with the PSP-Net methodology, authors have employed the encoder alongside Resnet18 to produce and extract feature maps. The decoder blocks serve the ongoing purpose of replicating the efficacy of pyramid blocks. One significant application of binary convolutions is to improve the efficiency of computational devices and processing speed. Subsequent to implementing the pyramid structures, feature maps are created and then fed through the final convolutional layer. Given the unique attributes of binary weights, there is a need to reassess activation functions. The initial activation function in the network shifts from ReLU to the TANH function [28] shown in Equation (1).

$$\mathrm{TANH}(z)=\frac{e^{z}-e^{-z}}{e^{z}+e^{-z}} \tag{1}$$

This occurs when the weighted trigger functions of standard 2D convolutional classes are combined, and the binary convolution is unique in that it employs weights that have been bifurcated via the Sign function. Differentiable approximations of binary step functions, also known as sign functions, are utilised in binary conv. An exposition of the model posited in this article is provided above. The fundamental objective of this architecture is to preserve and integrate the merits of the two constituent methods: the nimbleness of BNNs and the high performance of PSP-Net are fused into a single framework.

3.1. Binary Neural Network

In general, the efficiency of convolutional neural networks (CNNs) in both training and inference phases is remarkably high. Various methodologies have been introduced to enhance the setup of com-putational devices, such as model pruning, low-rank decomposition, and knowledge distillation, all aimed at boosting efficiency. A notable approach is the advancement of a binary convolutional network, show-casing a promising direction. The parameters of the model, comprising weights and activations, are transformed into binary values of +1 or -1 through the utilization of the Sign function. This transformation process is executed as follows:

For Peer Review

$$x^b = \text{Sign}(x) = \begin{cases} +1 & \text{if } x \geq 0 \\ -1 & \text{otherwise,} \end{cases} \quad (2)$$

The authors deviate from the conventional framework by introducing binary convolutional layers instead of 2D convolutional layers. The primary goal of this research is to reduce the memory footprint necessary for parameter calculations and storage, consequently amplifying processing speed when com-pared to traditional CNNs. This approach also serves to mitigate the computational load on the processor.

3.2. Convolutional 2D network

2D convolutional layers serve as the fundamental building blocks within Convolutional Neural Networks (CNN) for the computation and analysis of data characteristics, where parameters are maintained as real values to derive a set of output values. Subsequently, the convolution output undergoes processing through an activation function (e.g., ReLU) to introduce nonlinear properties. The convolution process can be denoted in Equation (3):

$$Y(i,j) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} X(i+m, j+n) \times W(m,n) + b, \quad (3)$$

where $Y(i,j)$ is the value at position (i,j); $X(i+m,j+n)$ is the input value at the position $(i+m,j+n)$; $W(m,n)$ is the weight at the position (m,n); b is the term bias. Furthermore, weights and bias are classified as parameters that undergo backpropagation for the purpose of learning and updating. Using filters, network training seeks to identify critical distillations. Indeed, the intricacy associated with parameter calculation and updating results in the CNN being overfitted [20].

3.3. Binary convolutional network

In order to overcome the limitation associated with 2D convolutional layers, a proposal is made to substitute them with binary convolutional layers. The conversion of parameters to binary form is achieved through the application of the Sign function. In contrast to their predecessors, the binary convolutional layers operate with parameters ranging from -1 to +1. This transition is anticipated to expedite the computation and transmission of parameters, thus reducing the burden on the computational resources of mobile robots. The function of the convolutional operation is expressed as Equation (4):

$$Y(i,j) = \text{Sign} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} X(i+m, j+n).W(m,n), \quad (4)$$

where M and N are the sizes of the binary filter, respectively. The use of binary conv classes has the following main contributions:

- Reduction in memory usage and computational efficiency by constraining model parameters to -1 and +1.
- Distinct from CNN, binary convolution utilizes simpler operations, enhancing computational performance.
- Proposed model with binary parameters require less memory storage compared to those with real values, consequently shrinking the model size.

Therefore, binary convolutional layers present a promising framework for devising a semantic segmentation model tailored for MR's navigation, given the aforementioned advantages.

4. Experimental Results and Discussion

In this section, the generated dataset utilized for training and testing the model is presented. Subsequently, the outcomes of ablation and comparative experiments are individually examined.

4.1. Data Generation

For Peer Review

Three datasets were utilised to train the proposed model including such as follows: CityScapes (5000 images) [22], Pascal_VOC (11500 images) [29], HaUI and TQB-datasets [6, 13] comprising 1200 images obtained from the Ta Quang Buu library, in Figure 3.

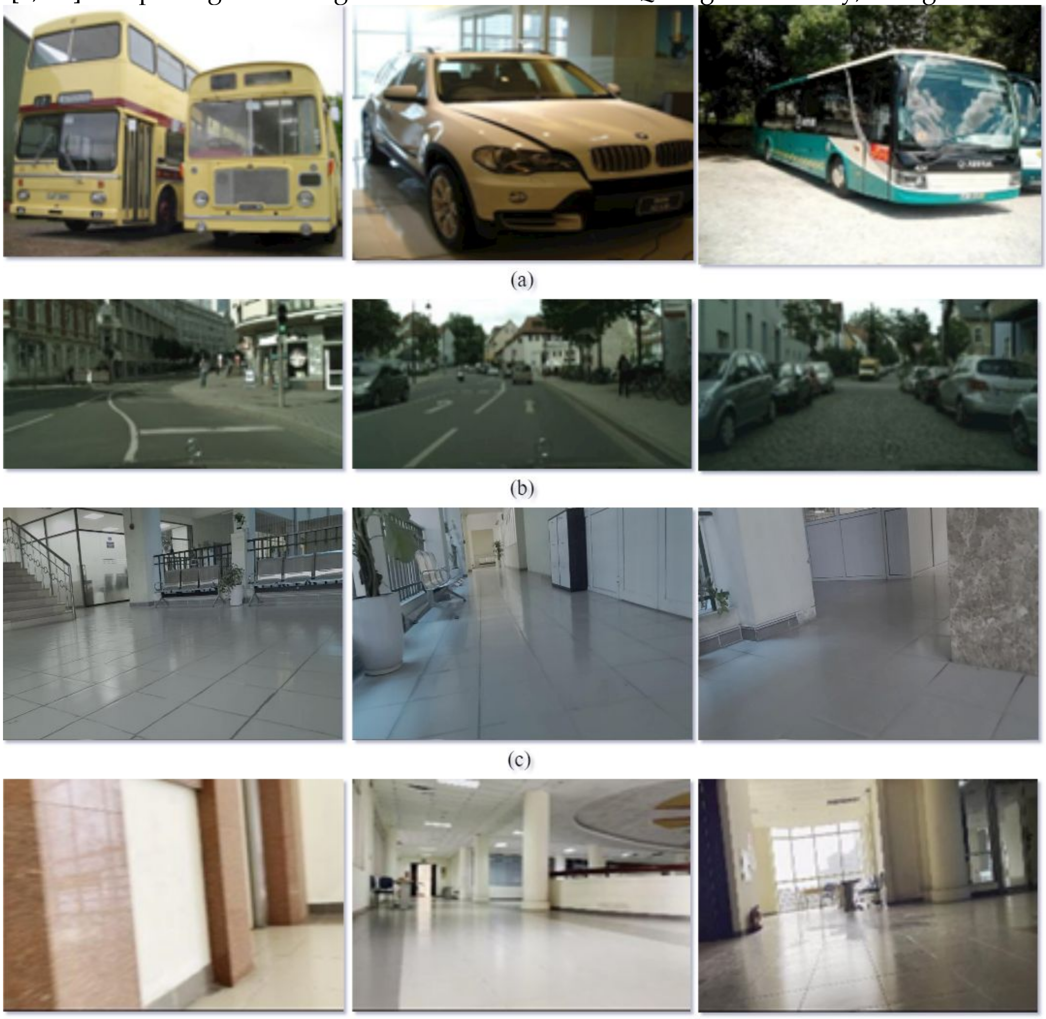


Figure 3. The image datasets for training Binary-SegNet model including (a): Pascal_VOC, (b): Cityscape, (c): HaUI-dataset and (d): TQB_dataset.

Firstly, Pascal VOC extensively collects utilised medicinal materials in computer vision comprising 11500 labelled images representing 20 distinct objects, including aircraft, bicycles, automobiles, individuals, and etc (see Figure 3a). A range of operations were executed on this dataset to improve its quality. These operations encompassed resizing the image to 256x256 pixels, randomly inverting the image horizontally, arbitrarily altering the RGB values, normalising the image by calculating the mean and standard deviation of each colour channel, and transforming the image from a numpy array to a torch tensor.

Secondly, Cityscape particularly serves as a resource in the domain of computer vision with regard to duties involving the comprehension of street imagery (see Figure 3b). The dataset comprises over five thousand images, each of which has a resolution of 1024×2048 pixels. Resizing images to 256x256 pixels, randomly inverting images horizontally, altering RGB values at random, normalising images using the mean and standard deviation of each colour channel, and converting images from numpy arrays to torch tensors are all reinforcements on this set.

Next, self-collected HaUI-dataset consists of 1000 images gathered from the landscape at Hanoi University of Industry Vietnam (see Figure 3c). The output consists of the following operations: resizing the images to 224x224 pixels, flip along the axis of symmetry, rotate the photo at an angle of 15 degrees, normalising them using the mean and standard deviation of each colour channel, and transforming them from numpy arrays to torch tensors.

Finally, self-collected TQB-dataset consists of 1200 images of HUST's Ta Quang Buu library, Vietnam (see Figure 3d).The output consists of the following operations: resizing the images to 224x224 pixels, normalising them using the mean and standard deviation of each colour channel, and transforming them from numpy arrays to torch tensors.

In addition, the learning pace for each parameter was modified using the Adam Optimizer [30], enabling it to function efficiently with a wide variety of models. More precisely, the learning rate is modified to 0.001 during the training phase. Furthermore, the training data file is subjected to Gaussian blur [31] in order to generate images of varying quality. The objective is to increase the segmentation efficiency of models and generate datasets with greater generalizability. Utilising the Gaussian function is as follows:

$$G^{(x,y)} = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}, \tag{5}$$

For Peer Review

4.2. Model Training

The computational framework is comprised of an Intel(R) Core(TM) i9-12900H central processing unit alongside an NVIDIA GeForce RTX 4080 graphical processing unit, endowed with 64GB of video memory. The computational software environment encompasses Windows 11, Python version 3.10, CUDA version 12.1.68, and PyTorch version 1.13.1. The datasets employed for the training of the model consist of 1200 images. The ratio of the training to validation dataset is established at 9:1, comprising 1080 training images and 120 test images. Evaluation metrics are formulated based on the performance exhibited during the training and prediction phases. The efficacy of inference and the accuracy of label matching are substantiated through metrics including Loss, Accuracy, Intersection over Union (IoU), and Dice coefficients. The efficiency of model processing and its operational speed are ascertained through measurements of Latency and Throughput. In order to provide a comparative analysis with alternative methodologies, U-Net, Fully Convolutional Networks (FCN), and IRDC models utilizing VGG and MobileNet architectures are scrutinized within an identical training framework. The training process is conducted over a total of 60 epochs, with downsampling factors set at 8 and 16. The Adam optimizer [30] is employed for the calibration of training parameters.

A reduced train loss indicates that the model is effectively learning the characteristics of the input. The cross-entropy loss function is shown as follows:

$$J(w) = \frac{1}{N}\sum_{n=1}^{N} H(p_n, q_n) = -\frac{1}{N}\sum_{1}^{N}\left[y_n \log \hat{y}_n + (1-y_n)\times\log(1-\hat{y}_n)\right], \tag{6}$$

Model training performance is evaluated based on the following technical specifications:

Mean Intersection over Union (mIoU): is an indicator used to evaluate the accuracy of object detection and image segmentation. The calculation formula and description are depicted belows:

$$\mathrm{IoU} = \frac{O}{U}, \tag{7}$$

where O is the area of overlap and U is the area of union, respectively. The IoU index has a value from 0 to 1, the closer the value to 1 indicates the more accurate the segment. During training, the point model demonstrates an increased mIoU value, showing improved segmentation accuracy with each epoch. Therefore, the calculation formula and description are depicted belows:

$$\text{meanIoU} = \frac{1}{n}\sum_{1}^{n}\text{IoU}_i, \tag{8}$$

where n is the number of classes and IoU_i is the IoU for the ith class.

Dice Metric: is utilised to assess the effectiveness of image segmentation models. An analogous approach to IoU, it predicts the regions accurately labelled on the pixcels by utilising the degree of overlap between regions. Dice metric is mathematically defined as follows:

$$\text{DSC}(A,B) := \frac{2|A \cap B|}{|A|+|B|}, \tag{9}$$

344

where B represents the predicted pixels, while A represents the set of genuine pixels. In contrast to Dice's IoU, the oscillation observed when the accuracy of the contiguous regions between the model's prediction and the correct label of the images changes distinguishes this approach. Therefore, the mean Dice (mDice) represents the average dice coefficient scores as follows:

$$\text{meanDice} = \frac{1}{n}\sum_{1}^{n}\text{Dice}_i, \tag{10}$$

where n is the number of classes in the data set and Dice_i is the IoU for the ith class.

Accuracy: evaluates the performance of a trained model. The correct predictions are proportional to the total number of samples.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}, \tag{11}$$

where TP: True positive; FP: False positive; TN: True negative; and FN: False negative, respectively.

Latency and Throughput Latency: represent the delay measured during the inference of the trained model. Throughput measures the processing speed of the model, calculated based on the number of images processed per second.

4.3. Training Results

To evaluate the functionality and effectiveness of the model, an assessment is conducted. This study offers guidance on both the proposed model and comparable models like PSPNet, Unet, and FCN. Evaluation metrics were derived from datasets such as Pascal_VOC, Cityscape, self-collected TQB and HaUI datasets. The selection of the most appropriate evaluation metric by the author is based on the characteristics of each trained dataset to highlight the advantages in segmentation. Segmented results were produced by Binary-SegNet during training on Pascal_VOC, Cityscape, TQB, and HaUI datasets depicted in Figures 4, 5, 6, and 7 respectively.

For Peer Review

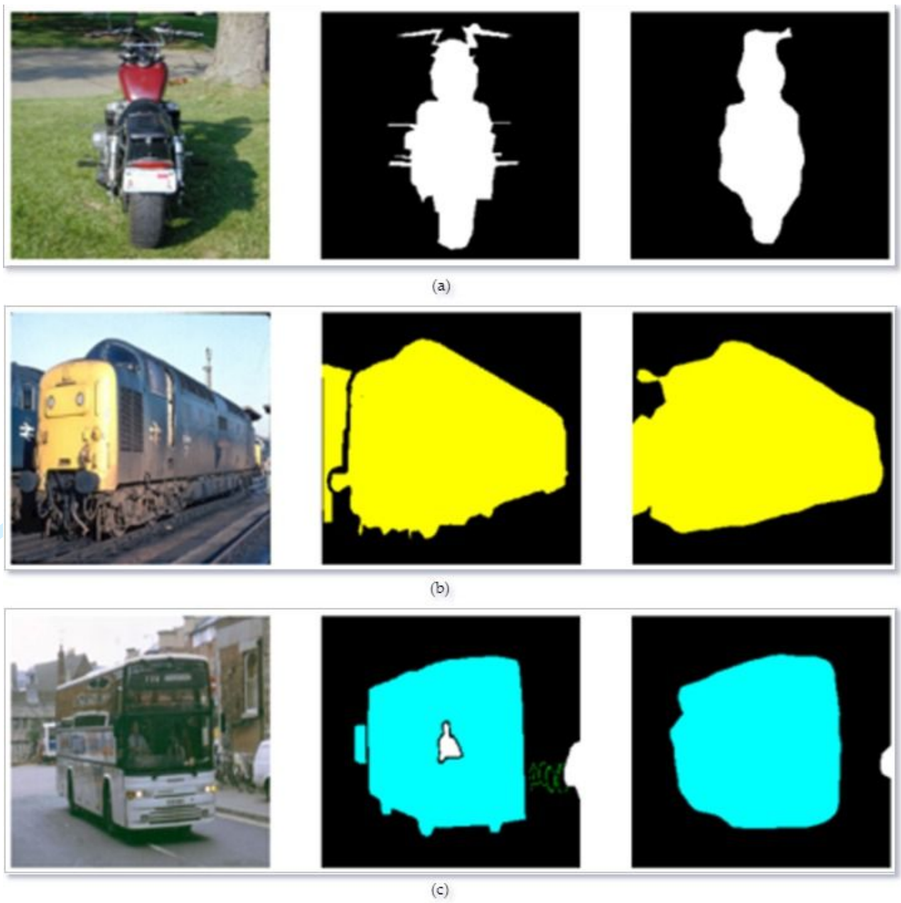


Figure 4. Image prediction using Pascal_VOC dataset [29] from left to right with input image, manually typed label, model pre-diction label, respectively

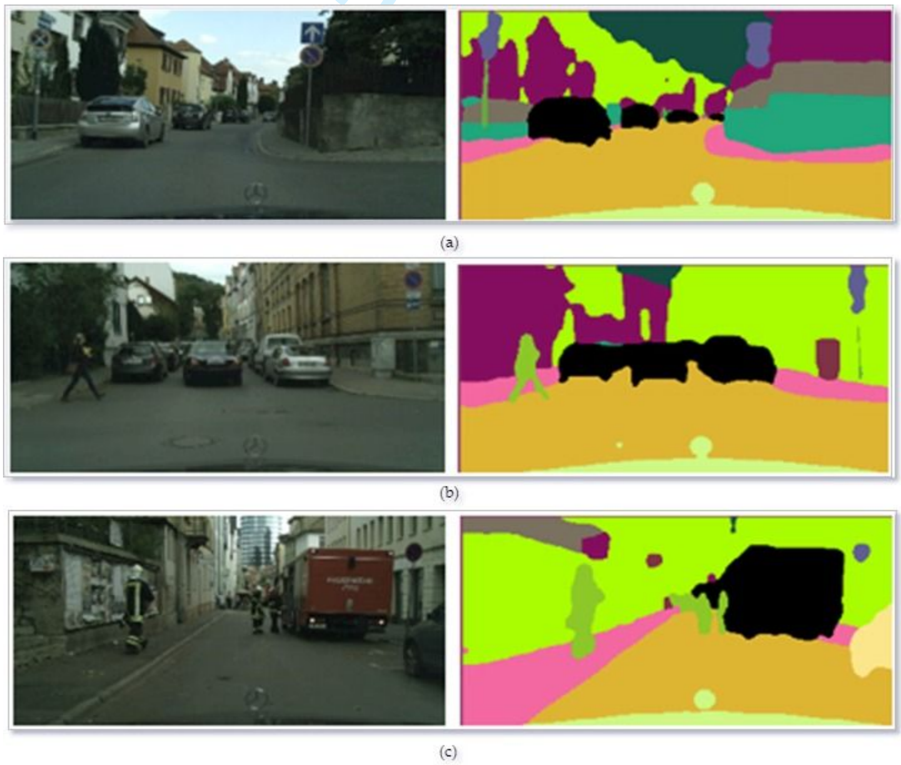


Figure 5. Inference results on Cityscape dataset [22] from the left to the right with the image inputs and segmented results, repectively.

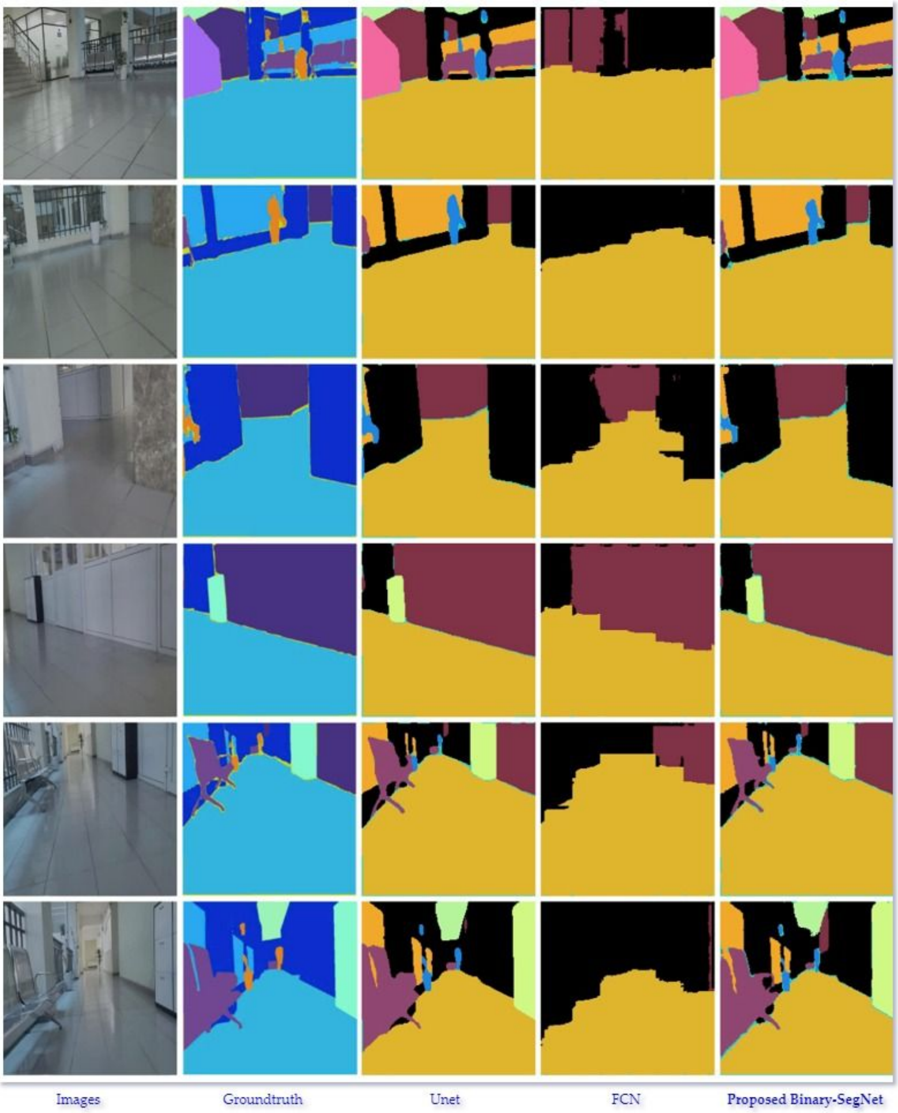


Figure 6. The comparison of the inference results with the stages of the art method is based on the HaUI dataset.

As illustrated, the proposed method produces inference results that are highly similar to the ground truth. Overlapping regions are specifically segmented. Accuracy is ensured, especially for landscape objects in the MR operating environment such as obstacles. This demonstrates the notable segmentation efficiency of Binary-SegNet. Comparisons with other models of Unet and FCN in Figure 6 show the strong competitiveness of the proposed method. The thorough exploitation of the effectiveness from the Pyramid modules combined with the binary network yields accuracy equivalent to a complex model like Unet.

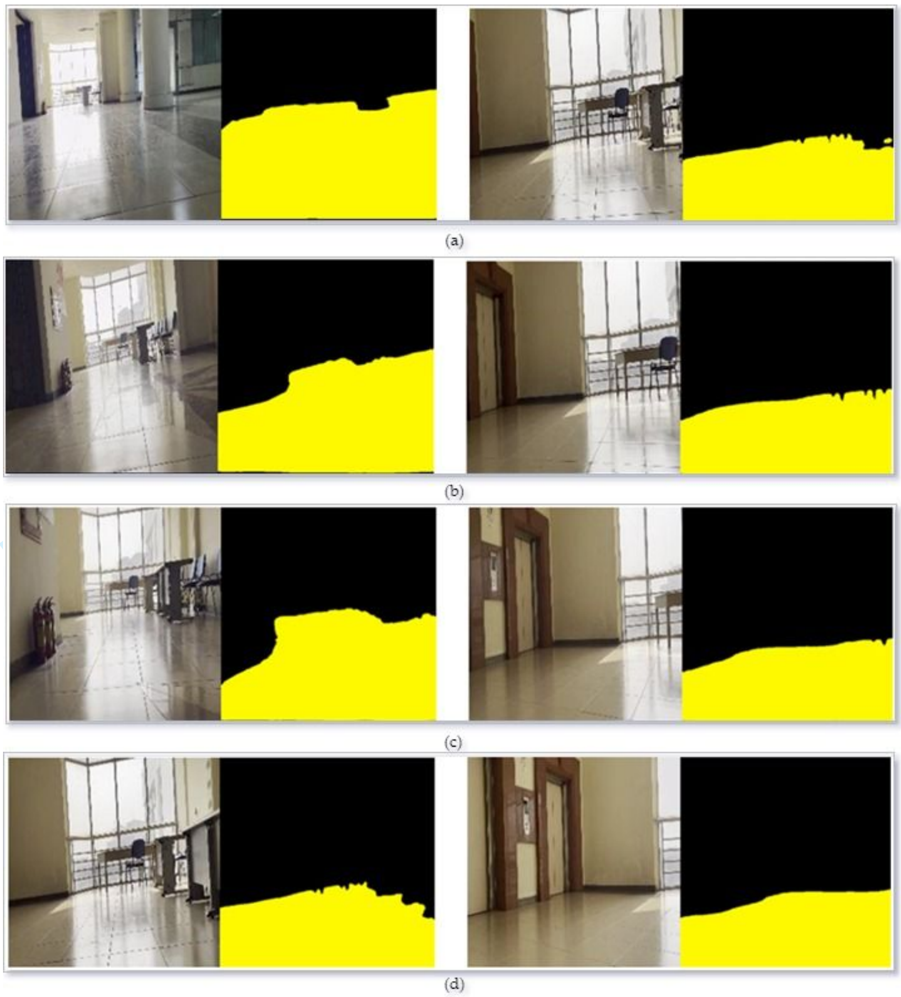


Figure 7. Inferences on TQB_dataset dataset [6,13] with the yellow part corresponds to the road, the black part corresponds to the object.

Meanwhile, the method’s performance significantly surpasses FCN in capturing object features and making accurate predictions. Based on the above training data sets, the model quality parameter sets are analyzed and their advantages are reinforced through index Tables from 3 to 5. We conducted a comparison between our proposed model and existing segmentation models of Resnet 152-PSP-Net, FCN-VGG16, and Unet-VGG19 based on the TQB-dataset (see Figure 7). Therefore, the simultaneous use of IoU and Dice metrics provides a comprehensive view of both the performance and competitiveness of the compared models. A high IoU value demonstrates the effectiveness of the prediction between overlapping regions, indicating that semantic information is efficiently analyzed and synthesized based on the proposed architecture. The Dice metric inherits the characteristics of IoU, and additionally, Dice is more sensitive to changes in predicted regions with differences. The obtained results demonstrate the flexibility of the proposed architecture. Furthermore, Figure 8 illustrates that the robust segmentation efficacy of the proposed methodology remains assured. Furthermore, the comparative analysis with these baseline models substantiates the advantageous performance of the proposed technique. Owing to the augmentation of our datasets with more demanding images and the comprehensive utilization of training data, our enhanced segmentation model employing the Adam optimizer is capable of obtaining a more precise depiction of the environment and facilitating a more expedited training process. Consequently, these accomplishments will significantly bolster the construction of the MR’s frontal view. In particular, in the dynamic operating environment of MR, where the context is constantly changing, the proposed model demonstrates the ability to adapt quickly and effectively.

Fo

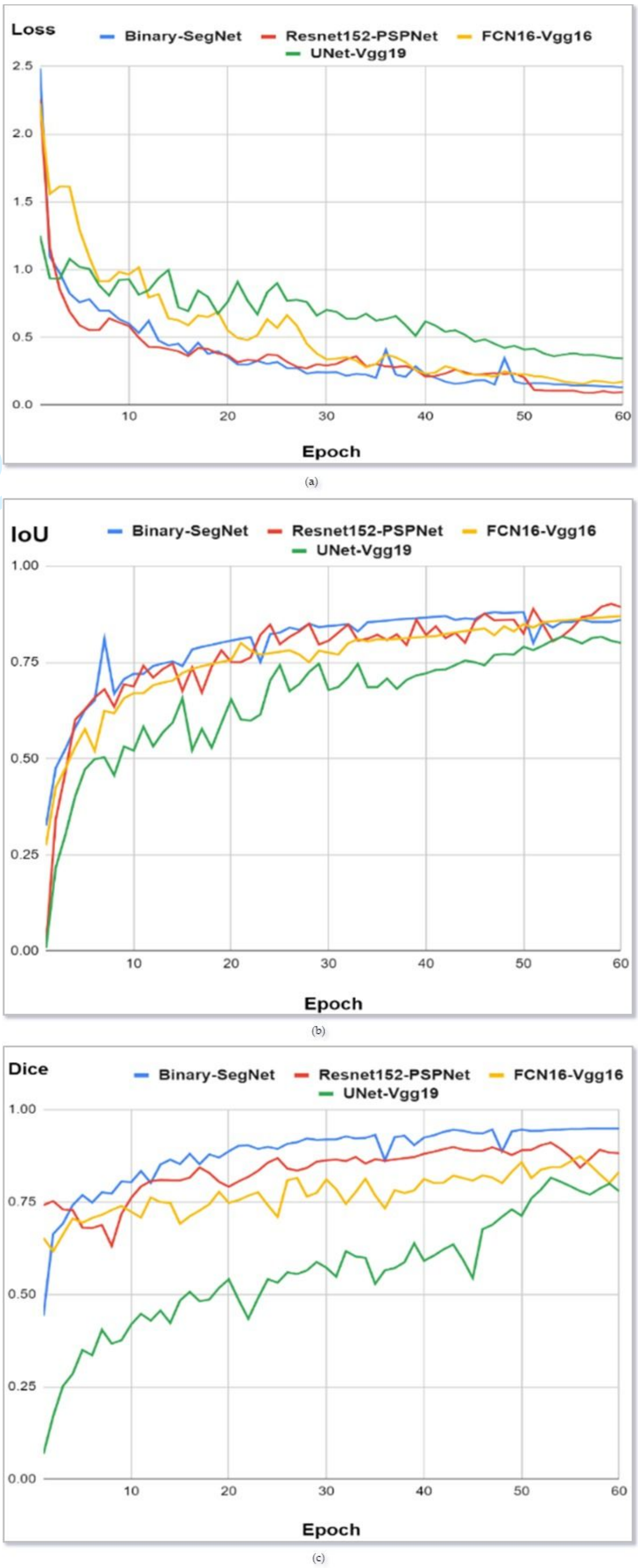


Figure 8. The diagrams of Train Loss, IoU, and Dice of Binary-SegNet training based on TQB dataset [6, 13].

For Peer Review

Moreover, the proposed method exhibits exceptional efficacy in the precise delineation of strata and their boundaries. The identifiers pertain specifically to vehicles and streetscape objects (e.g., automobiles, buses, pedestrians, motorcycles, and etc), in Table 2. A limited subset of layers exhibits ordinary performance as a result of elusive attributes or unbalanced datasets. The extent to which the MRs are adopted in their operational environment is not significantly affected. The outcomes serve as evidence of the model's capability to handle intricate training datasets containing a significant number of classes in order to solve the semantic segmentation problem. Furthermore, our model achieves precise boundary segmentation of various objects despite fluctuating illumination conditions, blurred input images, and noise. The model demonstrated remarkable performance when presented with inputs comprising a variety of objects against a common background. The outcome is comparable to that of the methods against which it was compared. The efficacy of segmentation is enhanced when examining data collected from a monocular camera under various conditions. This enhances the robot's perception of its environment and improves the navigation algorithm's performance in terms of both efficiency and accuracy.

Table 2. The results of evaluating the performance of the model on the classes of the Pascal_VOC dataset [29].

Class	IoU (%)	Dice (%)
Car	95.00	98.15
Motorbike	80.36	82.54
Bus	70.58	81.00
Person	75.36	88.45
Airplane	90.98	91.99
Bicycle	90.65	93.07
Bird	98.52	98.11
Boat	45.51	48.69
Bottle	90.31	94.92
Chair	90.68	90.64
Cow	85.16	85.61
Table	85.08	86.98
Dog	80.38	82.62
Horse	90.61	90.87
Potted plant	60.24	71.58
Sheep	90.53	92.51
Sofa	60.84	65.23
Train	89.15	90.14
TV/monitor	73.58	75.90
Tree	87.77	92.16
Unlabeled	88.37	88.69

Next, Table 3 demonstrates that the proposed Binary-SegNet for customs training is effective. The model's accuracy, as measured by two metrics with IoU (less than 7%) and Dice (less than 1%) is merely subpar in comparison to PSP-Net, a model renowned for its intricate architecture, substantial parameter count, memory demands, and high computational demands. The simultaneous use of IoU and Dice for experiments demonstrates sensitivity to changes in overlapping regions. Here, it refers to the predicted values and ground truth. No impact is observed on the implementation of the paragraph model on the mobile automaton due to the insignificant deviation. The remaining models all performed worse on the same dataset. At the same time, the convergence of values during training is not as fast and gives high results as the proposed method.

Table 3. The comparison of proposed Binary-SegNet model with the different methods though IoU and Dice.

Model	IoU (%)	Dice (%)
PSP-Net [10,21]	89,02	90.54
Unet [18,19]	82.15	81.91
FCN-VGG19 [28]	83.09	87.04
Our Binary- SegNet	**83, 14**	**88,21**

Then, the performance and computation speed of the parameters in the network were collected and compared. As shown and confirmed in Table 4, our proposed model provides superior evaluation parameters in comparison to all other models. The method in comparison had much greater latency than our method, PSPNet [10,21] was 4.5 times slower; Unet [18,19] is 5.4 times slower; and 3.3 times for FCN-VGG19 [28]. The throughput of SegNet-Binary is 1,7 times greater than that of PSPNet [10,21]; 2.4 times Unet [18,19] and 1.6 times that of FCN-VGG19 [28]. The observed throughput and latency exhibit promise when implemented in the context of navigating congested robots. Overall, the proposed method offers high competitiveness compared to the models used for comparison in terms of accuracy and reliability.

Table 4. The comparison of proposed Binary-SegNet model with the different methods though Latency and Throughput.

Model	Latency	Throughput
PSP-Net [10,21]	0.0955	84.369
Unet [18,19]	0.11577	60.8796
FCN-VGG19 [28]	0.0711	93.884
Our Binary- SegNet	**0.021496**	**150.32**

Finally, the proposed model is comprehensively evaluated against other approaches. The metrics and parameters are illustrated as shown in Table 5. In comparison to competing models, the proposed method generated remarkable outcomes across all performance metrics (including accuracy, mIoU, latency, and throughput). Our model's compatibility with computational equipment has been demonstrated. The authors deduce, based on the aforementioned findings, that Binary-SegNet is compatible with the MR's limited resources and maintains high inference speed and accuracy for the segmentation task.

Table 5. Comparison of the proposed Binary-SegNet model with all metrics based on the Cityscape dataset.

Model	Acc	mIoU	Latency	Throughput
Binary FCN-VGG 16 [2,3]	0.94272	0.69863	0.05249	50.120
IRDC-Net [6]	0.97798	0.7226	0.03925	60.024
Our Binary- SegNet	**0.97716**	**0.74365**	**0.025783**	**101.90**

In summary, Binary Seg-Net model ensures three key factors: high accuracy, fast computation and inference speed, and efficient utilization of system resources.

4.4. MR's strategy of navigation

To begin, the image plane will be approximated from the image coordinates using the focal length of the camera. Following this, the intrinsic camera matrix is displayed in the second transformation, which converts the image plane to the pixel plane. Then, the homography transformation [32] is set up the pixel plane for MR's path planning based on the perspective projection, in Figure 7.

For Peer Review

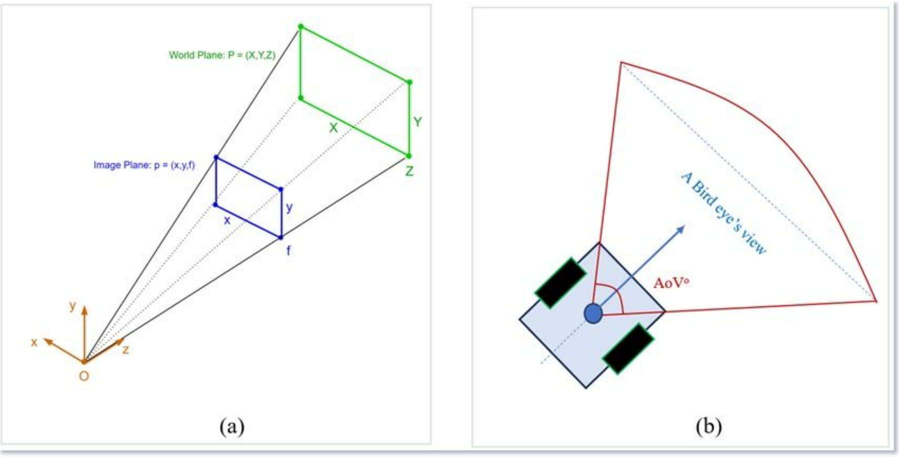


Figure 9. A perspective projection with (a): homography transformation and (b): MR's bird-eye view.

For Peer Review

The homography matrix 3 x 3 at the ground surface signifies the transformation between four points of the image plane and four points of world plane (see Figure 9a), as observed from a bird's-eye view in Figure 9b. Therefore, the transformation described by Equation (12) such as follows:

$$p = M_{int} \times M_{ext} \times W \tag{12}$$

where 3 x 4 intrinsic parameters, and 4 x 4 extrinsic parameters. Because the camera poses are fixed to form MR's bird's eye view.

In the ground surface (Z=0), the homography transformation matrix H is expressed as follows in following Equation (13):

$$H = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}; \; x = \frac{h_{11}X + h_{12}Y + h_{13}}{h_{31}X + h_{32}Y + h_{33} + 1}; \text{ and } y = \frac{h_{21}X + h_{22}Y + h_{23}}{h_{31}X + h_{32}Y + h_{33} + 1}. \tag{13}$$

Finally, using the checkerboard, the construction of MR's perception is determined by the image plane observed from above (see Figure 10).

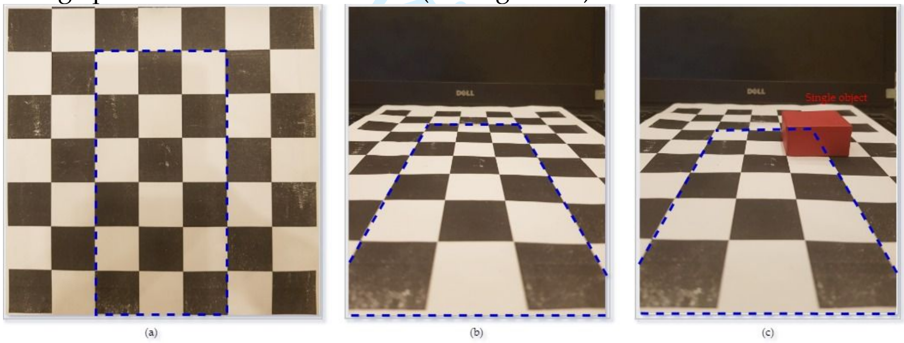


Figure 10. The homography transformation using a checkerboard (a) with (b) image plane and (c) having one object in the bird's eye view.

When comparing this analysis to the state of the art methods [2,3,6] the optimal MR's strategy was maintained. In contrast, the segmented images obtained in this study were processed using the proposed Binary-SegNet model, which enabled the construction of the grid-map, in Figure 11.

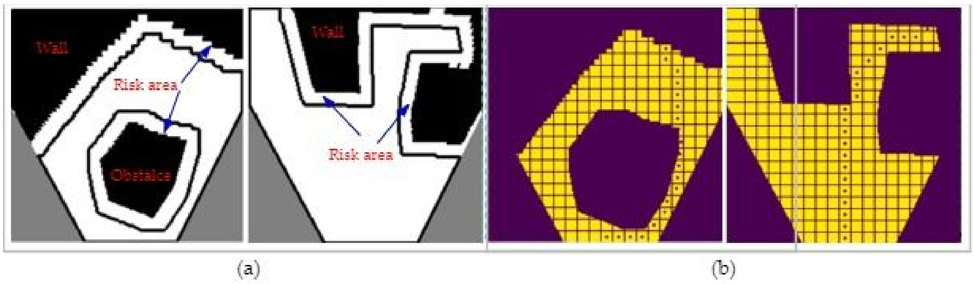


Figure 11. MR’s moving environment with (a) additional risk areas around obstacles based on image plane and (b) based on the grid-map method.

Then, the authors conduct an evaluation of the transformation outcomes and affirm the integral significance of binary semantic segmentation in the construction of the floor's frontal perspective. Subsequently, the MR’s optimal path planning [13] will be adeptly devised by authors, in Figure 12. The empirical findings substantiate the methodology for the detection of collision-free zones. Moreover, the researchers the efficacy of the optimal GPP is assessed in a consistent manner within Scenario 1, as illustrated in Figure 11. The ensuing 200 x 200 grid environments are conducted in an uninterrupted sequence. Upon standardization of the frontal view, a comprehensive framework for navigation and obstacle evasion will be developed. In Figure 13, traditional A* path planning with safety cost in the heuristic function ensures the obstacle avoidance ability (blue dot-lines). However, the implemented MR’s path planning [13] will adjust the path points on the original A* path to the improved smoothed A* path (green dotted line). Moreover, the trajectory created after smoothing still ensures the tasks such as finding the shortest path, and the change in steering angle is minimal, helping the MR move stably during the trajectory tracking process, and the image quality of the monocular camera is also guaranteed to reduce the impact of vibration.

In order to assess the effectiveness of our proposed semantic segmentation in conjunction with the navigation of three-wheeled MR (see Figure 14), the subsequent experiments were executed within Scenario 2 of the 3 m × 1.5 m environment depicted in Figure 15, which encompasses four distinct obstacles. In addition, as illustrated in Figure 12, a specialised local search algorithm was developed to enhance the obstacle avoidance safety of the MR when it effectively follows the global path. Upon comparison of the novel results with those obtained in [13], it is possible to conclude that semantic segmentation is an essential component in the construction of the frontal perspective of the ground. This facilitates the development of an optimal trajectory for MR. The objective of the practical experiments performed in this research was to improve techniques for identifying collision-free regions in local search zones while following a specified global path, in Figure 15 (snapshots (a)-(h)). The MR endeavors to follow a comprehensive trajectory from the initial position S to the terminal objective G, in accordance with the MR navigation framework delineated in Figure 12. In particular, MR has turned to avoid the top corner of obstacle 3 (see Figures 15b and c), while still maintaining GPP. Continuing when obstacle 2 with increasing length, with the lower corner interfering with GPP, MR through the data processing from the Seg-Net model for the images collected from the camera has been processed in time to ensure successful arrival at the destination (see Figures 15f and g).

Due to the fact that the MR’s camera pose, the proposed Seg-Net results would be compromised by a seamless trajectory. That is to say, the outcomes produced by our proposed model would be superior to those of the previous binary segmentation FCN-VGG 16 [2,3]. Furthermore, the Seg-Net’s processing speed is incredibly enhanced and improved in comparison to the light weight IRDC-Net [6]. As a consequence, the stability and deviation of the MR's steering angle (less than 0.2 rad) have been enhanced during trajectory tracking, in Figure 16.

For Peer Review

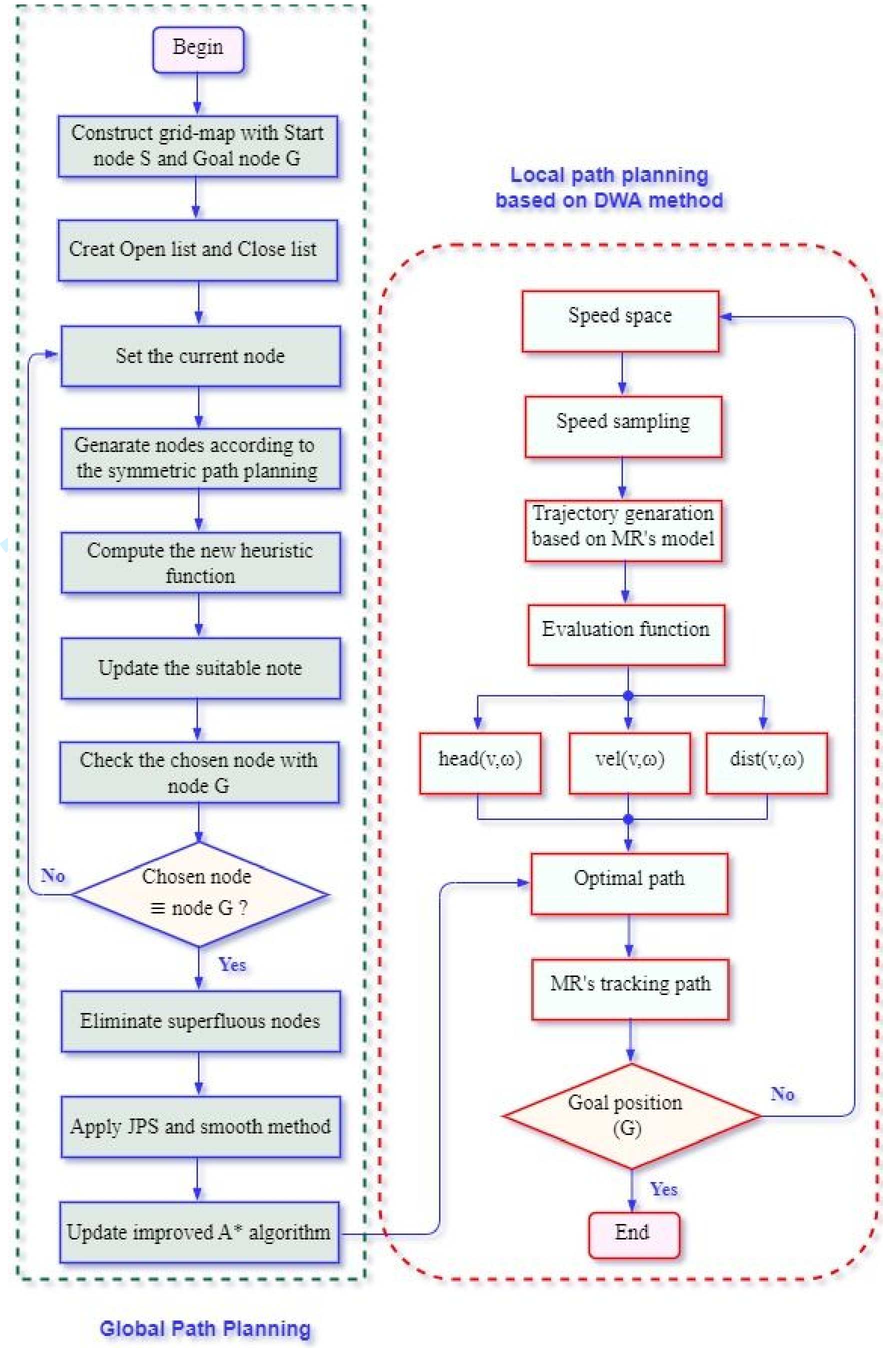

Begin
Construct grid-map with Start node S and Goal node G
Creat Open list and Close list
Set the current node
Genarate nodes according to the symmetric path planning
Compute the new heuristic function
Update the suitable note
Check the chosen node with node G
Chosen node ≡ node G ?
No
Yes
Eliminate superfluous nodes
Apply JPS and smooth method
Update improved A* algorithm
Global Path Planning
Local path planning based on DWA method
Speed space
Speed sampling
Trajectory genaration based on MR's model
Evaluation function
head(v,ω)
vel(v,ω)
dist(v,ω)
Optimal path
MR's tracking path
Goal position (G)
No
Yes
End

Figure 12. The optimization trajectory workflow [13].

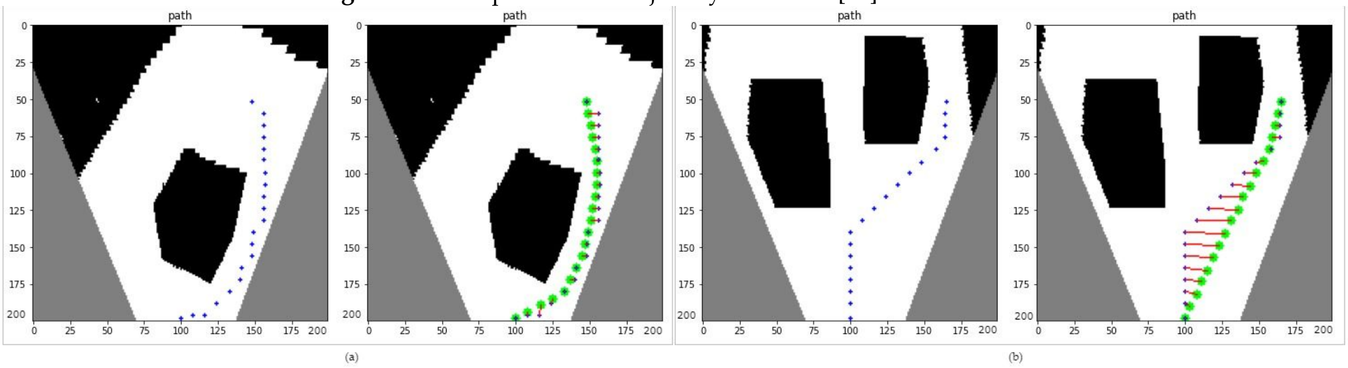


Figure 13. The optimal path planning based on improve A* algorithm with traditional A* path points (blue points) and improved A* path points with smooth and JPS algorithm [13] (green points) with (a) single object and (b) two objects in the grid-map.

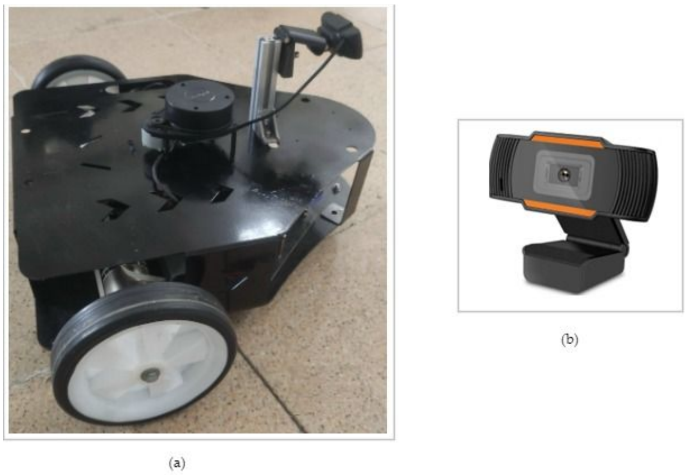


Figure 14. The experimental three-wheeled MR: (a) fully equiped MR, and (b) camera PCW01-QC with 1280x720p resolution.

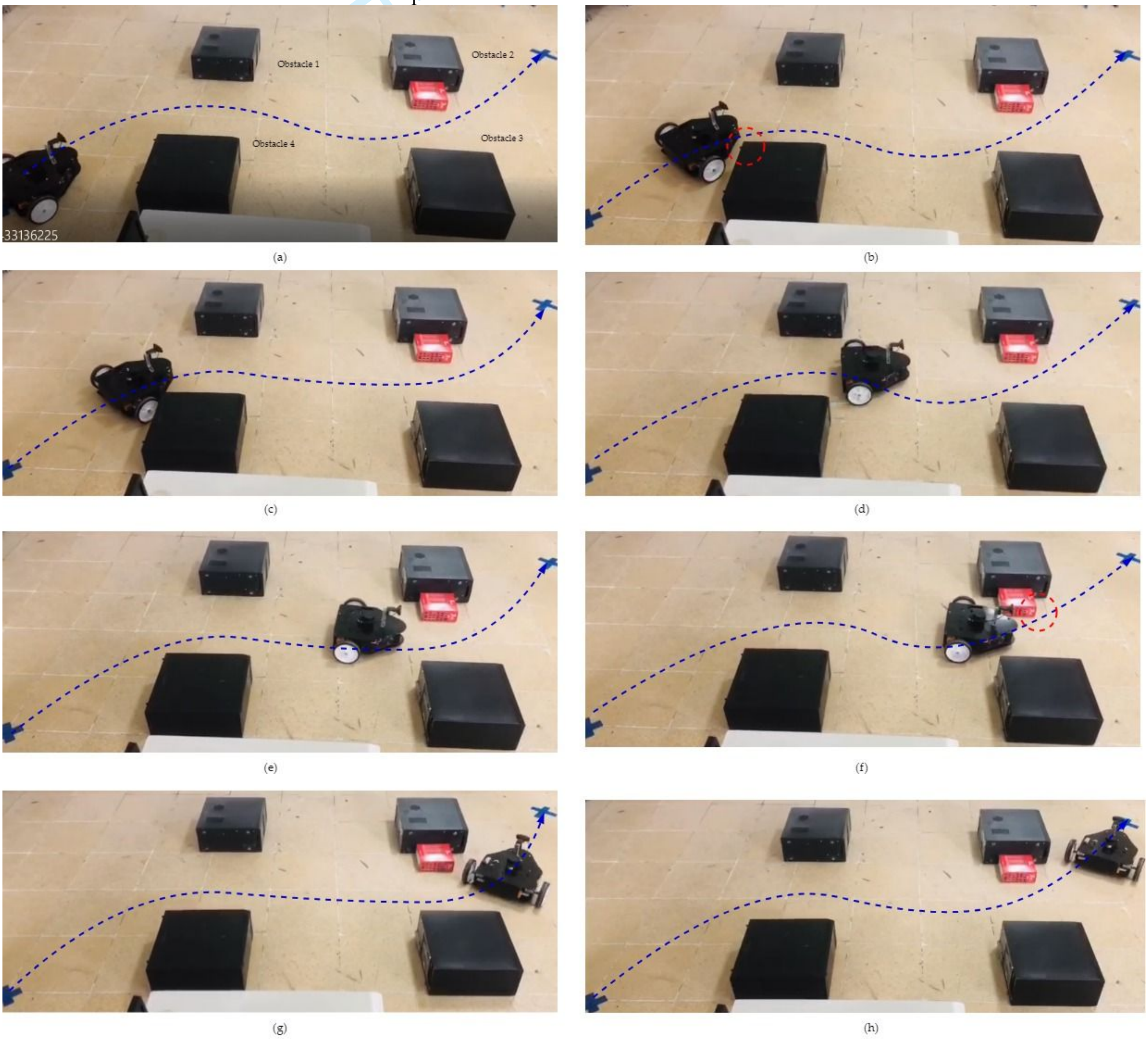
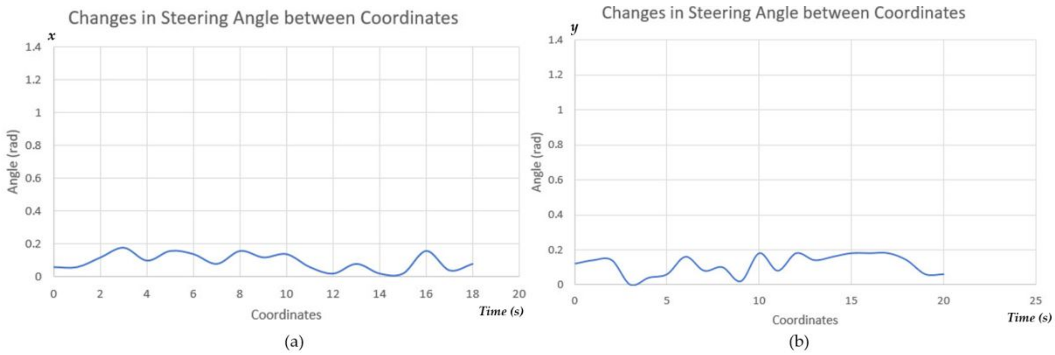


Figure 15. MR tracks the GPP based on the proposed Binary-SegNet model.



(a)

(b)

Figure 16. Changes in Steering Angle while mobile robot’s tracking optimal trajectory: (a) steering angle between coordinates in x axis, (b) steering angle between coordinates in y axis.

For Peer Review

With the successful application of the optimized mobile robot navigation strategy, particularly the process of smoothing the path while maintaining the mobile robot's safety, the tracking trajectory controlling is robust with steering angle variations of less than 0.2 rad in Figure 16.

5. Conclusions

The paper presents a real-time solution for MR’s navigation that extracts corridor scenes from a single image. In particular, this paper presents a lightweight segmentation model incorporating binary conv layers into the architecture. In order to effectively reduce the model’s dimensions and raise up the training and inference speed, binary conv layers are utilized in place of the customary 2D convolutional layers within the decoder. Furthermore, Binary-SegNet incorporates with a Adam optimizer to reduce the computational cost while increasing segmentation accuracy. By comparing the evaluation results to those of more the state of the art methods such as FCN-VGG16 [2,3], Unet [18,19], FCN-VGG19 [28] and IRDC-Net [6,13], the practicability of the proposed method is illustrated. Verifiable data is obtained from completed testing datasets of Pascal_VOC, Cityscape, HaUI, and TQB datasets. Performance is enhanced while the duration of training is optimised. This situation provides an opportunity to optimise the utilisation of system resources or to employ pentatonic segmentation models on devices that have hardware capacity constraints. The practical outcome demonstrates that the mobile robot's trajectory was effectively tracked with a reduced steering angle variation of 0.2 rad. In fact, our efficient binary Seg-Net model updated from binary to multi-class classification to accurately identify obstacles. Additionally, the segmented results will aid the continous local safety areas according to MRs global path planning [13]. In uncharted environments, the MRs’ avoidance capabilities are ultimately improved in the face of both static and dynamic obstacles.

References

1. Liu Y et al. A Review of Sensing Technologies for Indoor Autonomous Mobile Robots. *Sensors* 2024; 24(4): 1222. https://doi.org/10.3390/s24041222
2. Dang TV, Bui NT. Obstacle Avoidance Strategy for Mobile Robot Based on Monocular Camera. *Electronics* 2023; 12(8): 1932. https://doi.org/10.3390/electronics12081932
3. Dang, TV, Bui, NT. Multi-Scale Fully Convolutional Network-Based Semantic Segmentation for Mobile Robot Navigation. *Electronics* 2023; 12(3): 533. https://doi.org/10.3390/electronics12030533
4. Sohail A et al. A Systematic Literature Review on Machine Learning and Deep Learning Methods for Semantic Segmentation. *IEEE Access* 2022; 10: 134557-134570. https://doi.org/10.1109/ACCESS.2022.3230983
5. Kim JY, Ha JE. Foreground Object Detection in Visual Surveillance With Spatio-Temporal Fusion Network. *IEEE Access* 2022; 10:122857-122869. https://doi.org/10.1109/ACCESS.2022.3224063

6. Dang, TV, Tran, DMC, Tan PX. IRDC-Net: Lightweight Semantic Segmentation Network Based on Monocular Camera for Mobile Robot Navigation. Sensors 2023; 23(15):6907. https://doi.org/10.3390/s23156907
7. Paisitkriangkrai S, Sherrah J, Janney P, Hengel VD. Effective semantic pixel labelling with convolutional networks and conditional random fields. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015: IEEE. https://doi.org/10.1109/CVPRW.2015.7301381
8. Hirtzlin T et al. Stochastic Computing for Hardware Implementation of Binarized Neural Networks. *IEEE Access* 2019; 7: 76394-76403. https://doi.org/10.1109/ACCESS.2019.2921104
9. Wei X et al. Semantic pixel labelling in remote sensing images using a deep convolutional encoder-decoder model. *Remote Sensing Letters* 2018; 9(3):199-208. https://doi.org/10.1080/2150704X.2017.1410291
10. Zhao H, Shi J, Qi X, Wang Q, Jia J. Pyramid scene parsing network. In: *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, 2017: IEEE. https://doi.org/10.1109/CVPR.2017.660
11. Yuzhi C. Application of Resnet18-Unet in separating tumors from brain MRI images. *Journal of Physics Conference Series* 2023; 2580(1):012057. https://doi.org/10.1088/1742-6596/2580/1/012057
12. Bjekić M et al. Wall segmentation in 2D images using convolutional neural networks. *PeerJ Computer Science* 2023; 9(12):e1565. https://doi.org/10.7717/peerj-cs.1565
13. Dang, TV, Tan, PX. Hybrid Mobile Robot Path Planning Using Safe JBS-A*B Algorithm and Improved DWA Based on Monocular Camera. *Journal of Intelligent & Robotic Systems* 2024; 110(151): 1-21. https://doi.org/ 10.1007/s10846-024-02179-z
14. Xu W, Yang L, Cao S. A Review of Semantic Segmentation Based on Context Information. In: *Proceedings of the 2018 IEEE 4th Information Technology and Mechatronics Engineering Conference* (ITOEC 2018), Chongqing, China, 14-16 December 2018: IEEE. https://doi.org/10.1109/ITOEC.2018.8740714
15. Wang Y, Sun Z, Zhao W. Encoder- and Decoder-Based Networks Using Multiscale Feature Fusion and Nonlocal Block for Remote Sensing Image Semantic Segmentation. *IEEE Geoscience and Remote Sensing Letters* 2020; 18(7): 1159-1163. https://doi.org/10.1109/LGRS.2020.2998680
16. Agus, EM, Bagas YS, Yuda M, Hanung, AN, Zaidah I. Convolutional Neural Network featuring VGG-16 Model for Glioma Classification. *International Journal on Informatics Visualization* 2022; 6: 660–666. https://doi.org/10.30630/joiv.6.3.1230
17. Shelhamer V., Long J., Darrell T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2016; 39(4): 640-651. https://doi.org/10.1109/TPAMI.2016.2572683
18. Giang TL, Dang KB, Le QT, Nguyen VG, Tong SS, Pham VM. U-Net convolutional networks for mining land cover classification based on high-resolution UAV imagery. *IEEE Access* 2020; 8: 186257-73. https://doi.org/10.1109/ACCESS.2020.3030112
19. Alfarhan M, Deriche M, Maalej A. Robust Concurrent Detection of Salt Domes and Faults in Seismic Surveys Using an Improved UNet Architecture. *IEEE Access* 2022; 10: 39424- 39435. https://doi.org/10.1109/ACCESS.2020.3043973
20. Gao L, Huang Y, Zhang X, Liu Q, Chen Z. Prediction of Prospecting Target Based on ResNet Convolutional Neural Network. *Applied Sciences* 2022; 12(22): 11433. https://doi.org/10.3390/app122211433
21. Oršic′ M, Šegvic S. Efficient semantic segmentation with pyramidal fusion. *Pattern Recognition* 2021; 110: 107611. https://doi.org/10.1016/j.patcog.2020.107611
22. Cordts, M et al. The Cityscapes Dataset for Semantic Urban Scene Understanding. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 27–30 June 2016; pp. 3213-3223.
23. Zhang J et al. AFC-ResNet18: A Novel Real-Time Image Semantic Segmentation Network for Orchard Scene Understanding. *Journal of the ASABE* 2024; 67(2): 493-500. https://doi.org/10.13031/ja.15682
24. Yang H, Liu Y, Xia T. Defect Detection Scheme of Pins for Aviation Connectors Based on Image Segmentation, and Improved Resnet-50. *International Journal of Image and Graphics* 2022; 24(01): 2450011. https://doi.org/10.1142/S0219467824500116
25. Vaishali S, Neetu S. Enhanced copy-move forgery detection using deep convolutional neural network (DCNN) employing the ResNet-101 transfer learning model. *Multimedia Tools and Applications* 2023; 83(4): 1-25. https://doi.org/10.1007/s11042-023-15724-z
26. Kim T et al. Development of ResNet152 UNet++-Based Segmentation Algorithm for the Tympanic Membrane and Affected Areas. *IEEE Access* 2023; 11:56225-56234. https://doi.org/10.1109/ACCESS.2023.3281693
27. Adhinata FD, Ramadhan GN. Real Time Fire Detection using Color Probability Segmentation and DenseNet Model for Classifier. *International Journal of Advanced Computer Science and Applications* 2022; 13(9): 300-305. https://doi.org/10.14569/IJACSA.2022.0130935
28. Thiruvenkadam K, Padmapriya ST, Karuppanagounder S, Praveenkumar S. E-Tanh: a novel activation function for image processing neural network models. *Neural Computing and Applications* 2022; 34: 16563-16575. https://doi.org/10.1007/s00521-022-07245-x
29. Agrawal A. Exploiting CNNs for Semantic Segmentation with Pascal VOC. arXiv:2304.13216. https://doi.org/10.48550/arXiv.2304.13216
30. Liu M, Yao D, Liu Z, Guo J, Chen J. An Improved Adam Optimization Algorithm Combining Adaptive Coefficients and Composite Gradients Based on Randomized Block Coordinate Descent. *Computational Intelligence and Neuroscience* 2023; 5: 4765891. https://doi.org/10.1155/2023/4765891

For Peer Review

31. Kostková J, Flusser J, Lébl M, Pedone M. Handling Gaussian Blur without Deconvolution. Pattern Recognition *2020*; 103: 107264. https://doi.org/10.1016/j.patcog.2020.107264

32. Zhao L et al. Multi-Source Fusion Image Semantic Segmentation Model of Generative Adversarial Networks Based on FCN. *IEEE Access* 2021; 9: 101985-101993. https://doi.org/10.1109/ACCESS.2021.3097054

33. Hartley, R.; Xisserman, A. Multiple View Geometry in Computer Vision; Cambridge University Press: Cambridge, UK, 2000.

For Peer Review