Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.060605

ARTICLE

KD-SegNet: Efficient Semantic Segmentation Network with Knowledge Distillation Based on Monocular Camera

Thai-Viet Dang[1,*], Nhu-Nghia Bui[1] and Phan Xuan Tan[2,*]

[1]School of Mechanical Engieering, Hanoi University of Science and Technology, Hanoi, 10000, Vietnam
[2]College of Engineering, Shibaura Institute of Technology, Tokyo, 135-8548, Japan

*Corresponding Authors: Thai-Viet Dang. Email: viet.dangthai@hust.edu.vn; Phan Xuan Tan. Email: tanpx@shibaura-it.ac.jp

Received: 05 November 2024; Accepted: 10 January 2025; Published: 17 February 2025

ABSTRACT: Due to the necessity for lightweight and efficient network models, deploying semantic segmentation models on mobile robots (MRs) is a formidable task. The fundamental limitation of the problem lies in the training performance, the ability to effectively exploit the dataset, and the ability to adapt to complex environments when deploying the model. By utilizing the knowledge distillation techniques, the article strives to overcome the above challenges with the inheritance of the advantages of both the teacher model and the student model. More precisely, the ResNet152-PSP-Net model's characteristics are utilized to train the ResNet18-PSP-Net model. Pyramid pooling blocks are utilized to decode multi-scale feature maps, creating a complete semantic map inference. The student model not only preserves the strong segmentation performance from the teacher model but also improves the inference speed of the prediction results. The proposed method exhibits a clear advantage over conventional convolutional neural network (CNN) models, as evident from the conducted experiments. Furthermore, the proposed model also shows remarkable improvement in processing speed when compared with light-weight models such as MobileNetV2 and EfficientNet based on latency and throughput parameters. The proposed KD-SegNet model obtains an accuracy of 96.3% and a mIoU (mean Intersection over Union) of 77%, outperforming the performance of existing models by more than 15% on the same training dataset. The suggested method has an average training time that is only 0.51 times less than same field models, while still achieving comparable segmentation performance. Hence, the semantic segmentation frames are collected, forming the motion trajectory for the system in the environment. Overall, this architecture shows great promise for the development of knowledge-based systems for MR's navigation.

KEYWORDS: Mobile robot navigation; semantic segmentation; knowledge distillation; pyramid scene parsing; fully convolutional networks

1 Introduction

Predominantly, autonomous navigation seeks to circumvent obstacles with high precision and efficiency requirements [1]. Due to the significant computational demands of complex environments, navigation systems are prohibitively expensive. Notwithstanding their diminutive size, low cost, and significant computational requirements, cameras deliver comprehensive scene information. Fundamentally, path planning entails guiding MRs through environments that are dynamic and complex, all the while balancing the primary requirement for safety with the shortest route possible [2]. By means of semantic segmentation, MRs acquire copious amounts of data from the camera. Semantic segmentation (SS) is anticipated to enhance the ability of sensor and camera systems to differentiate between area and object classifications as imaging technology advances [3].

Recently, numerous vision-based applications have encountered a substantial challenge in the implementation of deep learning (DL) for SS [4]. Nevertheless, SS's remarkable development is still impeded by several obstacles. First, there is a dearth of information essential for SS operations, particularly in industrial sectors. Furthermore, the data collection system is expensive and labor-intensive, which constitutes the second issue. Further essential criteria for the camera-based SS system include resource optimization, processing speed acceleration, and output quality assurance [5]. Recent segmentation models that require immense amounts of computation have achieved remarkable performance, including DenseASPP [6], DeepLab V3 [7], fully convolutional networks (FCNs) [8], PSP-Net [9], DANet [10], CCNet [11], and OCR-Net [12]. Nonetheless, real-time mobile devices and online systems with constrained processing resources are incompatible with these intricate models. ENet [13], ICNet [14], DFANet [15], ESP-Net [16], and BiSeNet [17] are shown as specialised lightweight architectures. Subsequently, other researchers endeavoured to substitute significant backbones with networks that were either more efficient or shallower, including EfficientNet [18], MobileNet [19], and ShuffleNet [20]. Nonetheless, it is essential to acknowledge that these methods have specific inherent limitations. Primarily owing to the cost-effectiveness and availability of the technique, image segmentation utilising monocular cameras.

The paper proposes the SS architecture, which integrates the pyramid scene parsing (PSP) model with the knowledge distillation (KD) process. In the proposed KD, the teacher model employs ResNet152 for general extraction tasks, while the student model utilises the smaller ResNet18 network. Both advocate the utilisation of pyramid pooling modules as distinct analytical components at varying scales. Subsequently, the data is synthesised to produce the final prediction map. The authors instruct the student model with data derived from the features of the teacher model. Moreover, the suggested network preserves the teacher model's superior accuracy while ensuring the student model's swift reasoning capability.

Main contributions are as follows:

- KD-SegNet: Combining pyramid pooling blocks with a ResNet network as an encoder block into an efficient semantic segmentation model.
- Apply KD to train student model with backbone as ResNet18 network under support from ResNet152-PSPNet as teacher model.
- By implementing the Adam optimizer, performance and computational efficiency can be further enhanced. Data preprocessing is further enhanced through the implementation of Gaussian filters.
- Based on fours datasets of Cityscapes, Kitti, Cifar10 and self collected TQB-dataset, the proposed model has shown superiority over the state-of-the-art SS based on monocular camera.
- Empirical findings confirm the feasibility of MR's navigation based on KD-SegNet in both simulated and real-world scenarios.

Continue as follows with the following sections of the article. Related works are shown in Section 2. Section 3 presents the architecture of the proposed KD-SegNet. Section 4 contains experimental results and comparisons. Section 5 presents the final conclusions.

2 Related Works

Most of the current research focuses on the development of lightweight networks to increase performance in practical situations. In addition, compact segmentation models are also utilized in the domain of model compression, which can be broadly categorized as follows: pruning, quantization, and KD [21]. Shelhamer et al. introduced original FCN employing convolution and upsampling to classify each pixel within an image [8]. However, the efficacy of FCN is significantly limited due to its fundamental architecture.

Distillation of knowledge has been implemented extensively to compress semantic segmentation models. The "dark knowledge" is transferred from a teacher to a student model.

2.1 Semantic Segmentation

Significant progress has been made in recent years in computer vision as a whole and in SS in particular [2,3]. The FCN framework has witnessed recent developments such as FCN [8], PSP-Net [9], ESP-Net [22], DeepLabv3 [23], RefineNet [24], and LRR [25]. In contrast to the current CNN-based semantic segmentation frameworks, our approach incorporates the subsequent modifications to enhance its performance. As the backbone network, Chollet employed Xception to adjust the stride size and channel counts to attain an optimal equilibrium between efficiency and effectiveness [26]. Furthermore, low-level features provide boundary information to employ suitable skip connections. In contrast to U-Net [27], Rajamani et al. implemented up-convolution by combining features from multiple layers, resulting in improved accuracy. Drawing inspiration from the DenseNet [28], the precision of the present block was augmented by utilizing the output from the preceding one. Side prediction referred to the segmented results whose features are derived from multiple blocks. Moreover, this cross-module structure allows transmitting information between contiguous modules.

Prominent attributes encompass the model's aptitude for employing the Graphics Processing Unit (GPU) while simultaneously optimizing its operation and avoiding parameter reductions. U-Net [27], FCN [8], and PSP-Net [9] are illustrative CNN-based models constructed using an encoder-decoder architecture. For the purpose of reducing image size and increasing characterization size, the encoder is composed of numerous convolution layers and max pooling layers (e.g., VGG16 [2], ResNet18 [29], etc.). In conjunction with upsampling, the decoder employs convolutional layers to enlarge the image and provide semantic prediction for every pixel.

2.2 Knowledge Distillation

Recent research has made extensive use of KD-based techniques to enhance the precision of lightweight SS networks [30]. In 2015, Lu et al. presented KD's concept [31]. To improve the performance of a compact model, KD framework would transfer knowledge from complex teacher model. In [32], Liu et al. introduced the KD technique for training small SS based on the structured prediction. Then, two structural KD schemes (SKD) were constructed as follows: firstly, pair-wise distillation found similarities; secondly, holistic distillation was carried out by using GAN. Hence, the student model learned more structured information to operate as effectively as the teacher model. SSTKD [33] added statistical texture knowledge to efficiently transfer the inter-class distance from the teacher model. In contrast to SKD, Wang et al. proposed the intra-class feature variation KD method (IFVD) based on intra-class feature (IF) [34]. While still maintaining the passed-through features of the teacher model, robust IF variation enhanced the correctness of the student model. To provide more information to teacher model, Shu et al. suggested employing a channel-wise distillation technique (CWD) to acquire the soft probability mapping by normalizing each channel [35]. By minimizing the Kullback-Leibler divergences between the teacher and student models, the KD process can concentrate more significantly on each channel. In combination with incremental learning to improve accuracy, Arnaudo et al. utilized contrastive regularization knowledge distillation (CRKD), in aerial image processing [36]. A cross-image relational KD (CIRKD) was developed for student segmentation network of urban road [37] by examining the pixel relationships in global pictures. In light of the widespread use of transformers in the visual MR perception, Sampath et al. introduced the Transformer-based KD architecture (TransKD) [38]. Simultaneously on the teacher and student model, TransKD performed the transformer-to-transformer strategy of feature maps and patch embeddings.

In contrast to the above approaches, our paper places greater emphasis on the rapid inference speed of the student model and the high accuracy of the teacher model. Teacher models demonstrate enhanced inference capabilities when compared to student models due to the more complex data representation and more profound architecture of the former.

2.3 Limitations

For SegNet, limitations arise from the inherent characteristics of the Encoder-Decoder architecture. The generalization ability of this architecture can vary across complex training scenarios. Notably, imbalanced and noisy training data can pose challenges. High-resolution images also lead to increased computational time and resource costs. Older models like U-Net and DeepLabv3 have been outperformed by newer segmentation architectures. Additionally, complex architecture presents challenges in terms of interpretability and accessibility for SegNet.

Applying knowledge distillation to SegNet also faces related issues. Notably, training time per epoch increases, impacting the overall learning process. Accuracy and compression performance depend on the model architecture. However, KD may not be universally suitable for all architectures and datasets. Careful control of parameters and guidance during teacher-student model training minimizes rigidity in the development and deployment process.

3 Proposed Method

3.1 SegNet Model

SegNet is a CNN-based architecture designed for semantic image segmentation [39]. It has been used for a variety of tasks and has achieved competitive results. SegNet consists of two main components: an encoder-decoder. The encoder extracts feature from the input image. It consists of multiple convolutional and max-pooling layers stacked on top of each other. Common networks chosen for this task are VGG16 [2], ResNet [6], MobileNet [19]. The decoder predicts a label for each pixel in the image. It consists of multiple convolutional layers that increase the size and up-sampling layers. The combination of these two blocks forms a characteristic bottleneck architecture. Typical variants of SegNet include: DeepLabv3 [23], U-Net [27], FCN [8], and PSPNet [9]. The basic description of the architecture is shown in Fig. 1.

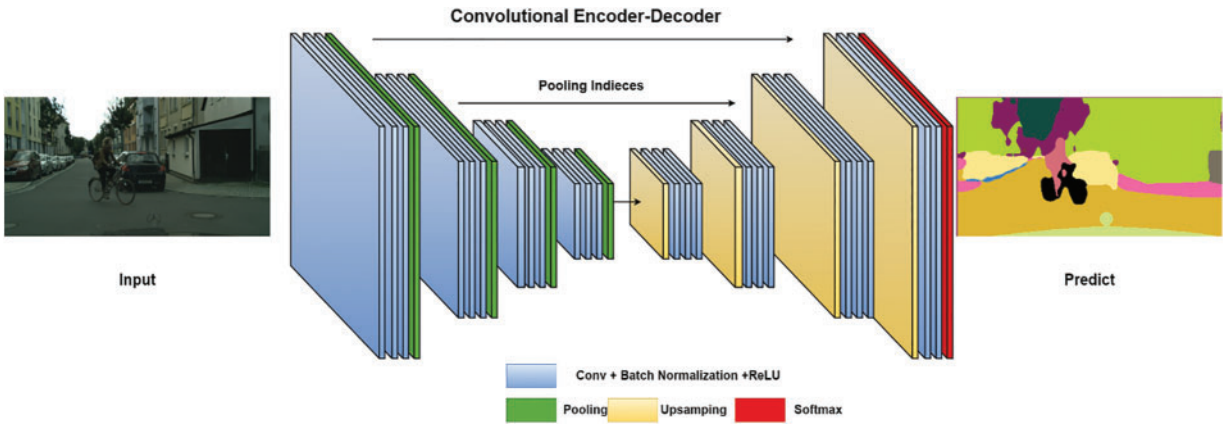

Figure 1: The SegNet architecture with the Encoder-Decoder block

3.2 Proposed KD Scheme

The proposed KD model aims to preserve two factors: the high accuracy of the teacher model and the fast inference speed of the student model. In general, teacher models exhibit superior inference performance in comparison to student models because of their more intricate data representation and deeper architecture. Experiments and real-world deployments have provided evidence of this across a vast array of evaluation metrics, such as mIoU, loss, accuracy, and mDice. The disparity in performance between student and teacher models can be primarily ascribed to two elements: the network architecture's ability to generalize and the degree to which semantic information is preserved throughout gradient descent, in Fig. 2.

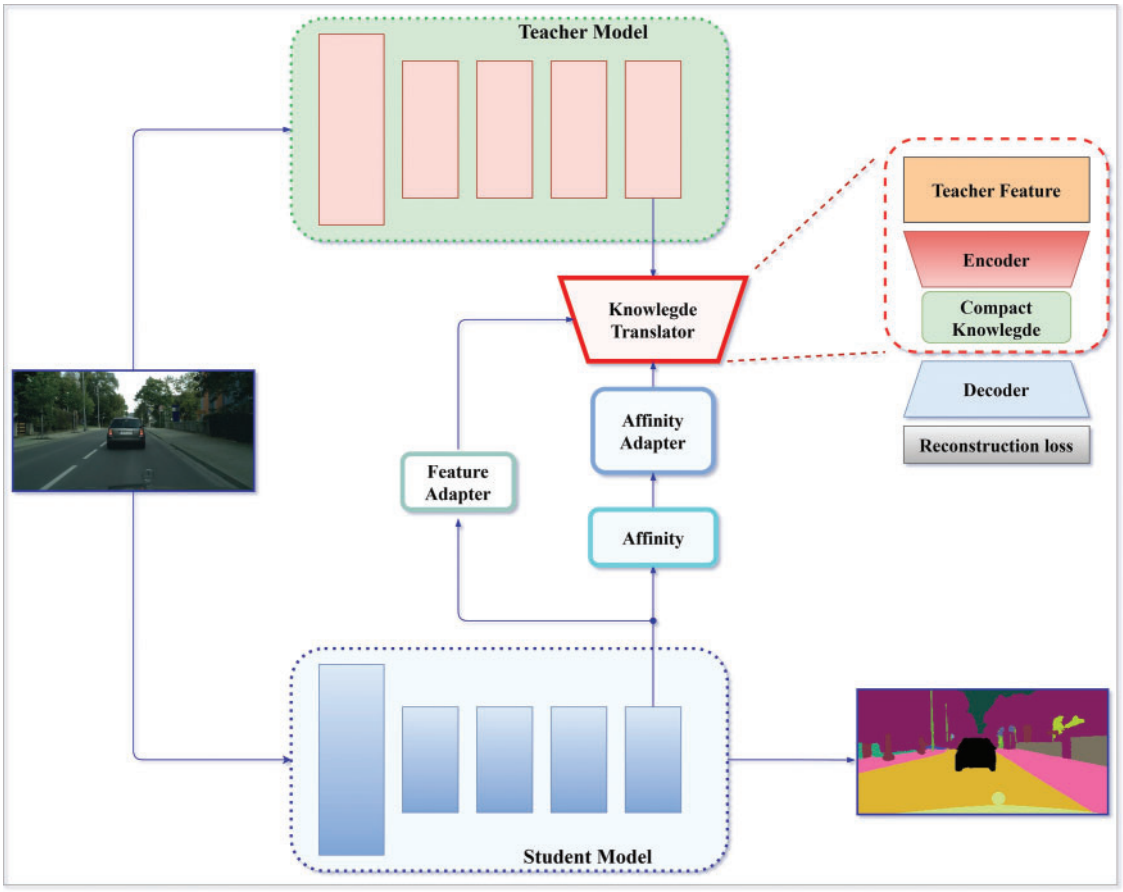

Figure 2: KD's description

In segmentation tasks, the depth and complexity of teacher models may, nevertheless, present obstacles with regard to training duration and inference efficacy. On the other hand, simplified student models provide notable benefits regarding network data propagation and computational performance. Student models are ideally adapted for implementation on resource-constrained systems, including embedded devices and edge devices, owing to their reduced size and increased speed. By capitalizing on the merits and demerits of both the teacher and student models, it is possible to train a lightweight model using KD for retaining the teacher model's swift inference speed while acquiring enhanced accuracy. In addition, training time can be drastically reduced, and a much smaller number of training epochs are required to attain performance comparable to that of conventional models when employing this method.

3.3 Proposed KD-SegNet

Like the presented SS models, the proposed model is built on the common encoder-decoder structure. Specifically, the teacher and student model architectures are designed according to the PSP-Net [9]. The ResNet network [6] with two distinct depths was chosen as the encoder block to extract the features of the input image. Decoders are pyramid pooling blocks that receive and process information from the encoder to generate a segmentation map. The specific architectural proposal is presented in Fig. 3.

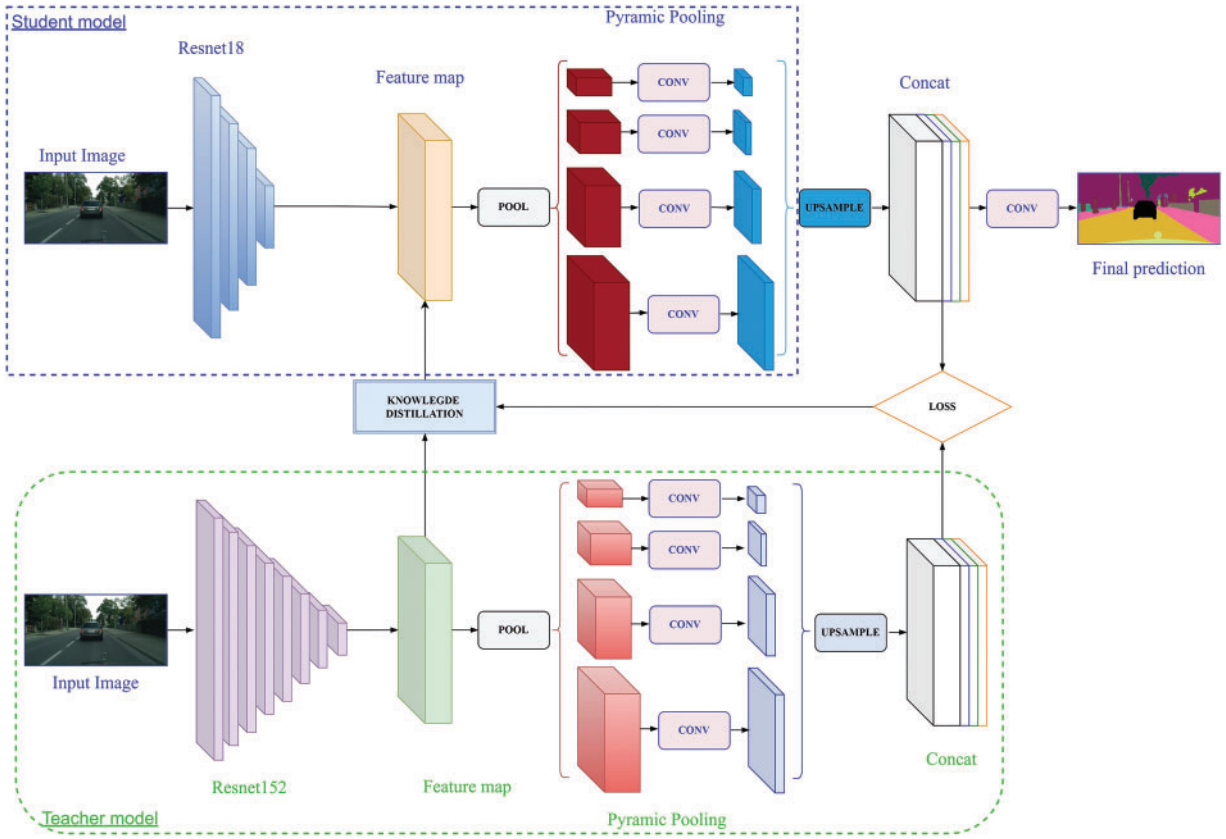

Figure 3: Proposed KD-SegNet architecture

3.3.1 Proposed Teacher Model

Teacher model is the PSP-Net model [9] with the ResNet152 network [40] chosen as the encoder block. ResNet152 consists of 152 layers. It is a deep neural network architecture built on residual blocks. Each residual block consists of two convolutional layers connected through a skip connection, in Fig. 4. The curved arrow originating from the beginning and ending at the end of the residual block represents the addition of Input X to the output of the layer. Adding this value will counteract the derivative being 0, thereby preventing the vanishing gradient phenomenon. With H(x) being the predicted value and F(x) being the true value, the goal during model training is to make H(x) equal or approximate F(x).

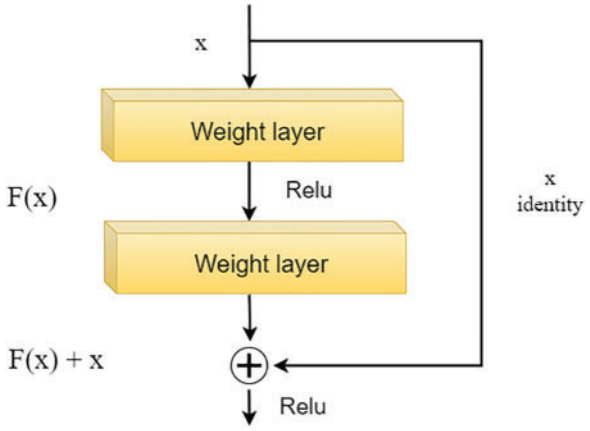

Figure 4: A residual block in the ResNet network

Furthermore, residual blocks allow the model to capture features effectively by allowing the model to learn transformations relative to the input data rather than learning from scratch. The layers in ResNet152 use max pooling to gradually reduce the input size and increase the size of the features, thereby creating the initial feature map. This feature map is used to train the student network on a similar scale. Despite its excellent accuracy in learning the features of the training data, ResNet152 still has limitations. The model has a large depth and volume. In addition, there are a large number of parameters that need to be calculated and propagated complexly. Overfitting is likely to occur on devices with limited computing capabilities. This leads to a loss of training and inference time for the model.

The decoder block is designed based on pyramid pooling layers. These blocks are used to extract features at different scales. This information is aggregated and used to create a complete prediction map. Data from the encoder is divided into decoding regions of different sizes. The size of these blocks is adapted to the size of the tensor received from the encoder. Each decoding region uses average pooling to reduce noise and focus on the features to be extracted. The extracted features are brought to a uniform size through deconvolution layers. Then, concatenation is used to combine information at different scales. The final convolution layer is used to aggregate information and make predictions. In summary, PSP with a pyramid pooling block architecture achieves effective segmentation results. This is especially true for adapting to the details and context of the training data. In combination, authors have the PSP-ResNet152 teacher model with superior accuracy in the SS task.

3.3.2 Proposed Student Model

The student model closely resembles the teacher model, as the authors have created a comparable structure for both models during the training process. In the student model, the change is the substitution of ResNet152 [40] with ResNet18 [29] as the backbone. ResNet18 utilizes residual blocks as a fundamental component. ResNet18 is comprised of a fundamental structure that has a total of 18 layers. Depth in this context refers to the number of layers in the network, including both the activated and linear composite layers. Batch normalization and synthesis classes are not included. The primary distinction is in the dimensions and lightweight design of the model. The number of parameters required for computation and propagation in this model is much lower than that of its teacher model (see Table 1).

Table 1: Comparison between the teacher model: ResNet152 and the student model: ResNet18 based on the number of parameters

Model	Number of parameters (million)
ResNet-152 [40]	58.24
ResNet-18 [29]	11.24
ResNet-50 [41]	23.901
ResNet-101 [42]	42.820
VGG16 [2,3]	134.7
DenseNet-169 [43]	12.8

Due to its simple architecture, ResNet18 has lower performance compared to other models [29]. This is addressed by distilling knowledge from the weights of the teacher model ResNet152 to improve training efficiency without increasing the weight and size of the network, which would reduce the model's learning and inference speed [40]. Knowledge is transferred from the teacher-to-student model in combination with the mitigation of the loss function. The class predicted by the teacher model is distributed to the classes of the student model. During distillation, the probability of each object class [31] is calculated as follows:

$$p_i = \frac{\exp\left(\frac{z_i}{T}\right)}{\sum_j \exp\left(\frac{z_j}{T}\right)} \tag{1}$$

where p_i is the probability of the *i*th class calculated from z, T is the temperature parameter. T = 1 to define the standard SoftMax function. As T grows up, the probability distribution generated by the SoftMax function becomes softer, showing that the prediction is close to the correct value. The teacher model and the student model both use the same T-value to calculate SoftMax during distillation. The overall loss function, incorporating both distillation and student losses, is calculated as:

$$L(x; W) = \alpha \times H(y, \sigma(z_s; T = 1)) + \beta \times H(\sigma(z_t; T = \tau), \sigma(z_s; T = \tau)) \tag{2}$$

Importantly, in order to maintain consistency between the teacher and student models, the decoder block is left unchanged in the proposed student model structure. The pyramid pooling module incorporates features from four different scales. It divides the feature map into different subregions and performs average pooling on the sub-regions. The four subregion sizes used are 1 × 1, 2 × 2, 3 × 3, and 6 × 6. The original size of the feature map is obtained through bilinear interpolation. Finally, the features extracted at different scales are aggregated into the final pyramid pooled global features. This pre-serves the multi-scale contextual information. This allows for utilizing the high performance of the pyramid pooling blocks while also improving the KD process. Inference to generate the image feature map is performed only on the student model to take advantage of its smaller size and easier deployment on devices with diverse configurations, especially on systems with limited resources and computing capabilities.

4 Experiment 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

Four datasets were utilized to train the proposed model: Cityscapes (5000 images) [44], Kitti (80,000 images) [45], CIFAR10 (60,000 images) [46], and TaQuangBuu's (TQB) dataset comprising 1200 images obtained from the Ta Quang Buu library [47] (https://github.com/buinghia3101/TQB-Dataset.git) (accessed on 09 January 2025).

4.1.1 Cityscapes

Cityscapes [44] particularly serves as a resource in the domain of vision-based duties involving the comprehension of street imagery. The dataset comprises over five thousand images, each of which has a resolution of 1024 × 2048 pixels. The total number of photos used was 1000, of which 85% were for training and 15% for testing. Augmentation including randomly inverting images horizontally, altering RGB values at random, normalizing images using the mean and standard deviation of each color channel, and converting images from Numpy arrays to torch tensors are all reinforcements on this set.

4.1.2 Kitti

Kitti [45] includes more than 80,000 images taken from MR-mounted cameras. The images are high resolution (2048 × 1024 pixels) and taken at 10 Hz. Augmentation including randomly inverting images horizontally, altering RGB values at random, normalizing images using the mean and standard deviation of each color channel, and converting images from Numpy arrays to torch tensors are all reinforcements on this set.

4.1.3 Cifar10

Cifar10 [46] comprises 60,000 color images of 32 × 32 pixels, divided into 10 classes. The dataset has split 50,000 photos for training and 10,000 test photos as default. The dataset undergoes image resizing to 256 × 256 pixels, followed by extracting a 224 × 224 pixel square from the center of the resized image. The data is then converted from Numpy array format to PyTorch tensor format. Finally, the image data is normalized by subtracting the mean value for each color channel and dividing by the corresponding standard deviation value.

4.1.4 TQB Dataset

TQB dataset [47] consists of 1200 images gathered from HUST's Ta Quang Buu library (see Fig. 5). The total number of photos used was 1200, of which 85% were for training and 15% for testing. The output consists of the following operations: resizing the images to 224 × 224 pixels, normalizing them using the mean and standard deviation of each color channel, and transforming them from Numpy arrays to Torch tensors.

Moreover, the learning rate of the model is adjusted through the Adam optimizer [48], adapting to the specific characteristics and size of the training data. When the gradient changes non-uniformly, accelerating the parameter convergence process. Thus, it brings efficiency in various diverse training scenarios. This involves fine tuning the learning rate to prevent information loss during gradient descent. Additionally, the training dataset is preprocessed using the Gaussian blur function to enhance generalization. The model adapts better to cases where the training data is of low quality. The generalization and performance of the model are ensured. The quality of the images is modified and augmented into the training process. The

Gaussian function [49] is applied as follows:

$$G^{(x,y)} = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{3}$$

where x, y are distance on the two axes from the original coordinates on the image. σ is the standard deviation of the distribution.

Figure 5: TQB dataset [47] with input images of specific scene with number (a): 70, (b): 210, (c): 630, (d): 100, (e): 955, (f): 1015, (g): 1140, and (h): 1870

4.2 Evaluation Metric

4.2.1 Train Loss

Train loss demonstrates the level of efficiency during model training. A loss value closer to zero demonstrates that the model captures the characteristics to be extracted. The loss value is calculated based on the cross-entropy function [47] as follows:

$$J(w) = \frac{1}{N}\sum_{n=1}^{N} H(p_n, q_n) = -\frac{1}{N}\sum_{n=1}^{N} [y_n \log \hat{y}_n + (1 - y_n) \log (1 - \hat{y}_n)] \tag{4}$$

4.2.2 Accuracy

Accuracy evaluates the performance of predictions [40]. The value is calculated based on an average of the correct predictions of the model. The closer the accuracy is to 1, the more accurate the model proves the prediction.

$$Acc = \frac{TP}{TP + FP} \tag{5}$$

4.2.3 Full-Width Accuracy (Fwa_Acc)

Fwa_Acc calculates accuracy for all pixels in an image, including pixels in the positive class and negative class [50]. The calculation formula is as follows:

$$\text{Fwc\_Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{6}$$

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

4.2.4 Accuracy Class (Acc_cls)

Acc_cls is calculated as the percentage of correctly predicted classes [50]. The calculation formula is as follows:

$$\text{Acc\_cls} = \frac{\sum_{i}^{n} \text{TP}_i}{\sum_{i}^{n} \text{N}_i} \tag{7}$$

where TP is the number of pixels correctly classified into class i, N is the total number of pixels belonging to class i in the ground truth dataset.

4.2.5 Mean Intersection over Union (mIoU)

First, is described by following Eq. (8), in [50]:

$$\text{IoU} = \frac{\text{O}}{\text{U}}. \tag{8}$$

where O is the overlapping area and U is the consolidated area, respectively. A review value closer to 1 indicates the more accurate the segment. Then, during training, the mean mIoU value increases, indicating improved segmentation accuracy in (9).

$$\text{meanIoU} = \frac{1}{\text{n}} \sum_{\text{i=1}}^{\text{n}} \text{IoU}_\text{i}. \tag{9}$$

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

4.3 Training Results

The study presents inference outcomes and comparisons with correspondingly structured models (U-Net, PSP-Net, FCN) to assess the functionality and efficacy of the model. Cifar10, Cityscape, Kitti, and TQB datasets are harvested to derive the evaluation metrics. Initial distillation performance on the Cifar10 dataset is represented by the loss value indices. To exemplify the benefits of the proposed method, suitable evaluation metrics are furnished in accordance with the characteristics of each trained dataset. The Cityscapes and TQB datasets are utilized to gather segmentation outcomes. Proposed KD-SegNet's segmentation performance and rapid inference speed are ultimately demonstrated through a comprehensive comparison with contemporary methods.

Utilizing the 60,000-image Cifar10 dataset, KD's efficacy was evaluated. As shown in Fig. 6, the accuracy attained 98.91%, with student loss and distillation loss measuring 0.13 and 0.12, respectively. The classes in the training dataset are appropriately captured and predicted by the student model of KD-ResNet18. The rate of training is considerably quicker in comparison to the conventional approach used to train the

ResNet18 student network. Typically, the student model will have a smaller size and number of parameters than the teacher model. In practice, the proposed KD-SegNet model uses distillation with quantization and distillation combined with model pruning [51]. The coordination of the above techniques results in models with optimal size and increased computational speed and parameter propagation.

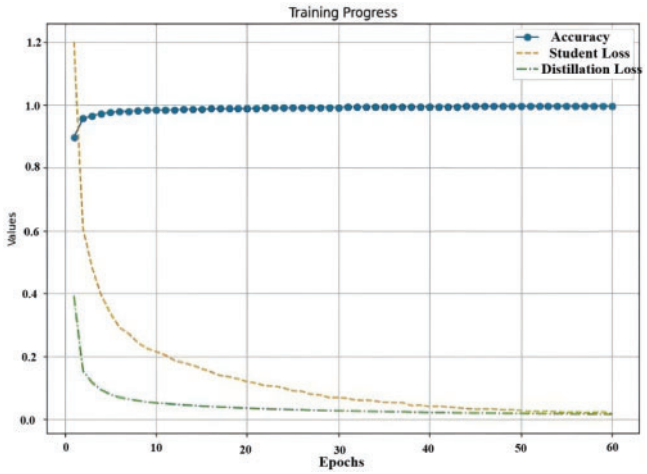

Figure 6: Experimental KD's results from ResNet152 for ResNet18 [29] training

Furthermore, Fig. 7 illustrates the evaluation metrics including as follows: Loss, Acc, fwa_Acc, Acc_cls and mIoU based on Cityscapes dataset, to compare the training efficiency of the proposed method to that of the student model trained using the conventional method. KD-SegNet easily surpasses ResNet18-PSPNet across all performance metrics. The method under consideration attains a Loss value that is 0.14 units lower in magnitude than the model being compared. Moreover, segmentation performance is enhanced as Acc increases by over 0.08, fwa_Acc increases by over 0.11, Acc_cls increases by over 0.19, and mIoU increases by over 0.23.

A comparative analysis continuously between the acceleration of prediction accuracy and the rate of convergence of parameters, in Fig. 8. The proposed model generates a comparatively accurate prediction map in just 30 epochs, with a mIoU of 65% after 30 epochs. On the contrary, the conventional model produces a prediction map with a mIoU of merely 61% after 120 epochs. The superior efficacy of the proposed KD-SegNet in terms of training speed and segmentation is substantiated by these two comparisons. MR navigation facilitates the rapid acquisition of semantic information pertaining to objects.

A summary of the conclusions drawn by the proposed method is presented in Fig. 9. The model captures and deduces scene elements such as people, vehicles, and streetscapes with remarkable accuracy. Furthermore, data obtained from the pre-existing Kitti dataset is incorporated into the distillation and inference procedure on the Cityscapes dataset utilized by the student model. The experimental results illustrate that the performance of inference is enhanced through the transfer and integration of supplementary semantic information into the model. This results in the SS task becoming more generalized. When coupled with the nimble characteristics of the student model, the suggested approach possesses the capacity to be implemented on autonomous systems that provide guidance to resource-constrained MRs. Specifically in navigation tasks, objects such as roads and obstacles are prioritized for capture. As illustrated, roads, humans, and vehicles are accurately segmented in corresponding dark yellow, pink, and black colors, respectively. The Mio values for all classes are above 90%.

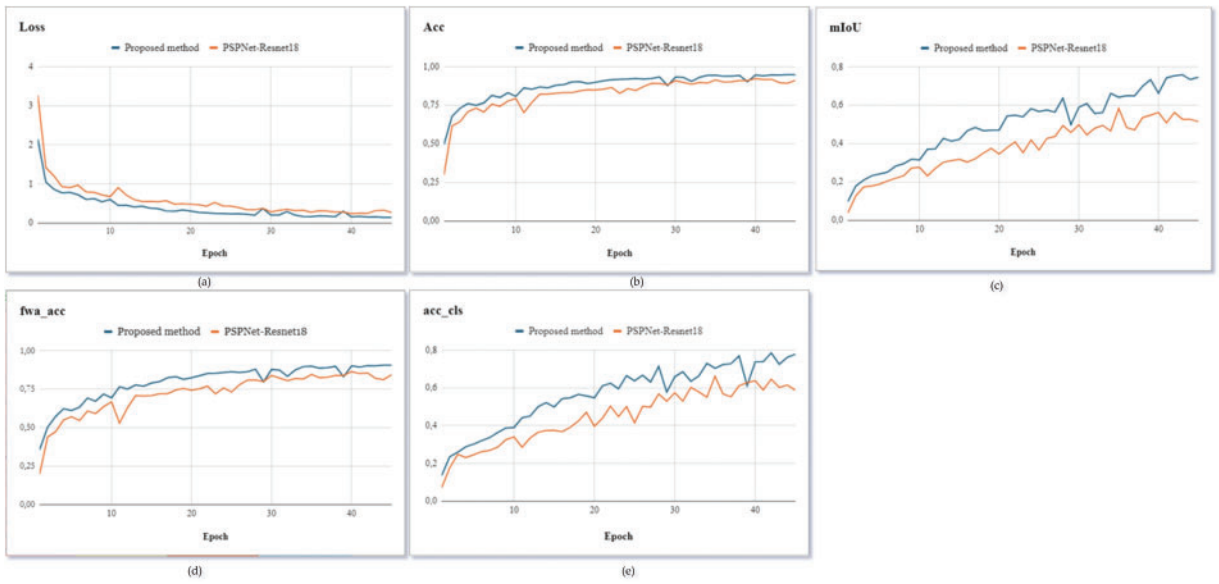

Figure 7: Experimental KD's results from ResNet152 for ResNet18 [29] training

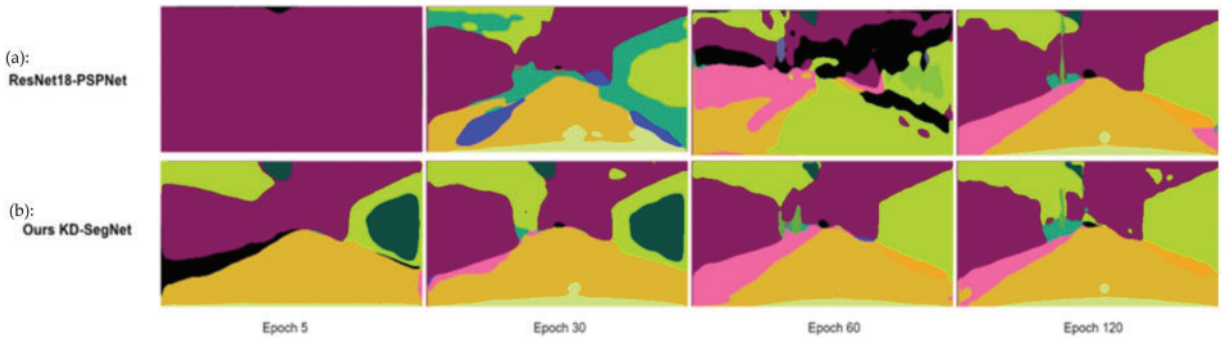

Figure 8: Inference results of the proposed KD-Segnet with the conventional PSPNet-ResNet18 model

Figure 9: Image prediction using Pascal_VOC dataset for each pair of input image and model prediction, respectively

The last network performance test was also run on Fig. 10's background image, which features several intersections and barriers based on the TQB dataset. This benefit enhances the computation of steering angle and speed for MRs. Safety and cost effectiveness are guaranteed.

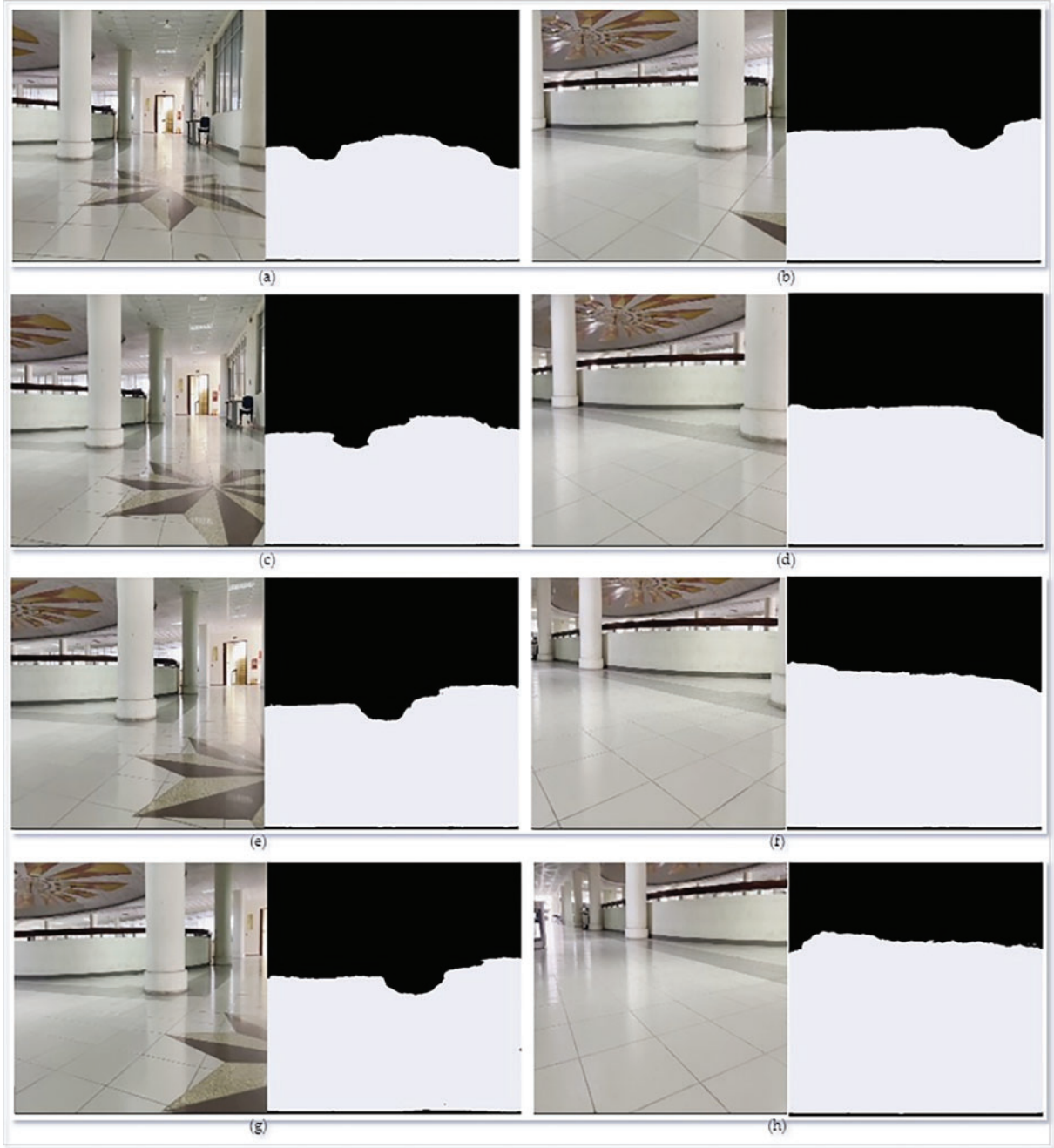

Figure 10: KD-SegNet's results consist of obstacles, corners, and intersections based on TQB dataset, from (a) to (h), taken from the MR's bird's-eye view when moving from the starting point and preparing to make a left turn

For the SS assignment, the authors present a comparison with other contemporary approaches. Table 2 presents a comparison with other models based on accuracy and mIoU. Not only compared with conventional models, the proposed model also shows superiority when compared with other light-weight models of MobileNetV2 and EfficientNet. The disparity in prediction performance is insignificant, measuring at less than 1%. However, upon comparison with the alternative methods, the proposed method exhibits a substantial enhancement.

Table 2: Comparison of the proposed KD-SegNet model based on synthesis of metrics collected on the Cityscape dataset

Model	Accuracy	mIoU
IRDC [47]	91.265	0.7263
Binary-SegNet [39,52]	93.584	0.6882
PSPNet-ResNet18 [9,29]	90.151	0.7081
PSPNet-ResNet50 [9,41]	90.525	0.6907
PSPNet-ResNet152 [9,40]	96.405	0.7804
FCN-VGG19 [52]	94.909	0.6610
Unet-VGG19 [52,53]	89.477	0.6598
PSPNet-MobileNetV2 [9,47]	93.005	0.6130
FCN-MobileNet [47,53]	88.151	0.5723
Efficientnet-Unet [18,54]	90.082	0.5138
KD-SegNet (Our)	**96.319**	**0.7716**

Then, Table 3 points out the improved performance of the student model while maintaining a rapid training rate. Data is collected and contrasted in the illustration in terms of the number of epochs and training time. In contrast to conventional models, proposed KD-SegNet attains equivalent inference performance in 51.26% less time and 25% fewer epochs on average during training. The time to inference averaged between 0.095 and 0.120 per image for the experiment. Meanwhile, the method is compared to the same time consumption but with inferior accuracy.

Table 3: Comparison of the proposed KD-SegNet model based on the number of epochs and training duration guarantees an mIoU score of 65%

Model	Number of epochs	Time (Hours)
Binary-SegNet [39,52]	102	0.81
PSPNet-ResNet18 [9,29]	95	0.96
PSPNet-ResNet50 [9,41]	71	1.058
PSPNet-ResNet152 [9,40]	66	1.606
FCN-VGG16 [2,3]	81	0.904
KD-SegNet (Our)	**20**	**0.518**

Moreover, the suggested framework demonstrates significant enhancement in processing velocity when juxtaposed with less resource-intensive architectures such as MobileNetV2 and EfficientNet, as evaluated through latency and throughput metrics (see Table 4). With the use of the KD model in training the student model, in addition to the accuracy inherited from the teacher model's weights, the fast inference speed of the deployed model is also prioritized for evaluation. Along with the high accuracy demonstrated in the previous sections, the measurement results of the wave transmission process and sample calculation in the proposed model have all achieved excellent results. For the proposed method, the obtained latency value is only 0.005 s while the throughput reaches 88.14 samples per second. The computation and parameter propagation during inference leverage the lightweight structure of the student model. The flexible characteristics of the student model provide superior sample processing speed compared to the methods being compared.

Table 4: Comparison of the proposed KD-SegNet model based on latency and throughput

Model	Latency (second)	Throughput (second)
PSPNet-ResNet152 [9,40]	0.026	65.12
PSPNet-ResNet50 [9,41]	0.008	75.91
FCN-VGG19 [53]	0.011	55.69
FCN-MobileNetV2 [47,53]	0.009	40.17
Unet-ResNet152 [40,54]	0.038	30.10
Unet-VGG19 [55]	0.011	55.63
KD-SegNet (Our)	**0.005**	**88.14**

4.4 MR's Path Planning Based on KD-SegNet

The initial step involves approximating the image plane by utilizing the focal length of the camera and the image coordinates. The intrinsic camera matrix is subsequently depicted during the pixel plane-to-image plane conversion in the second transformation. Subsequently, the pixel plane is homogenized for MR's path planning in accordance with the perspective projection, in Fig. 11. Furthermore, the homography matrix 3×3 at the ground surface signifies the transformation between four points (x, y) of the image plane and four points $(x', y')$of world plane (see Fig. 11a), as observed from a bird's-eye view (see Fig. 11b). Therefore, the transformation described by Eq. (10) such as follows:

$$p = M_{int} \times M_{ext} \times W \tag{10}$$

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

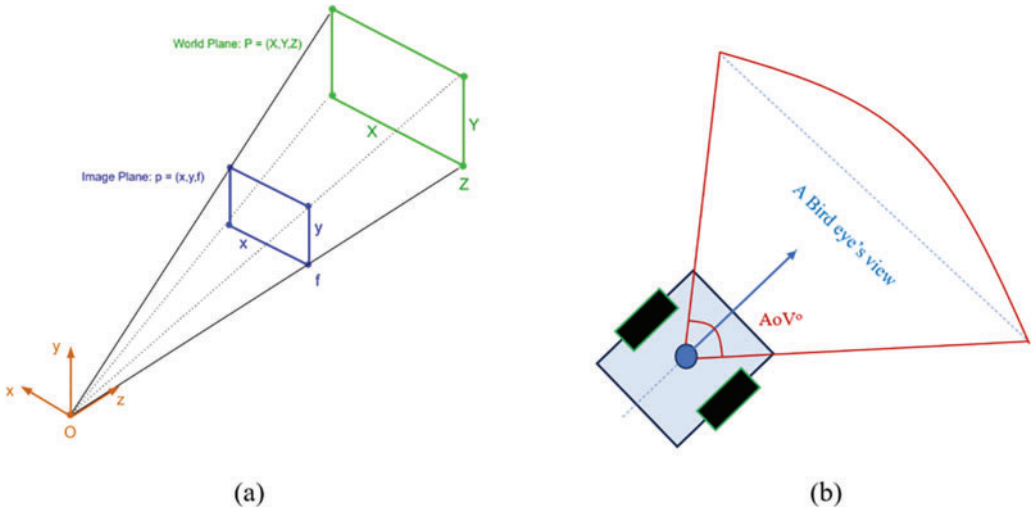

Figure 11: A perspective projection with (a): homography transformation and (b): MR's bird-eye view

In the ground surface (Z = 0), the homography transformation matrix H is expressed as follows in following Eq. (11):

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

Finally, using the checkerboard (see Fig. 12), the construction of MR's perception is determined by the pixel plane observed from above. Confirmed empirically, the homography transformation from the image plane to the pixel plane in bird's eye view.

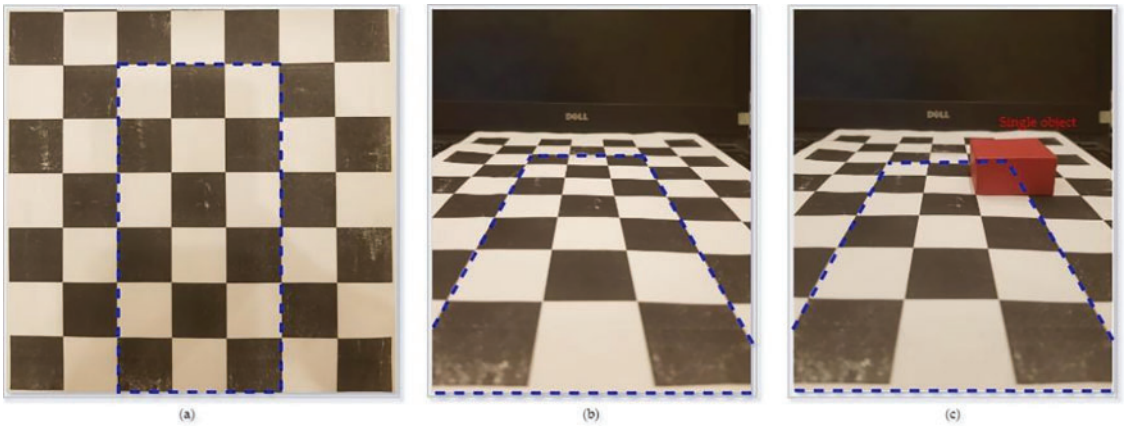

Figure 12: The homography transformation using a checkerboard (a) with (b) image plane and (c) having one object in the bird's eye view

Furthermore, MR's environment is constructed by grid-map with cells representing the state of allowing movement or being occupied by obstacles, in Fig. 13. The size of the calculated cell is chosen to match the MR's size and the moving speed in the known grid-map.

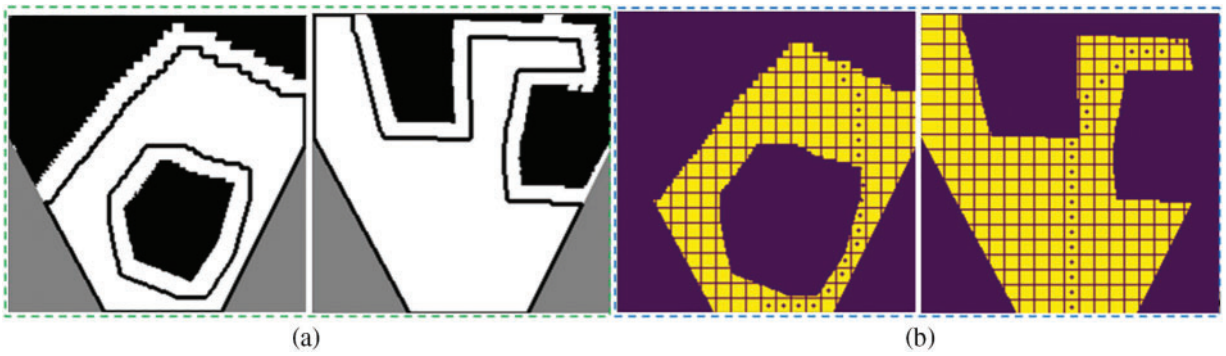

Figure 13: MR's environment with (a) additional risk zones surrounding obstacles and (b) grid-map-based obstacles with a risk zone in the MR's environment

In comparing the output performance of the proposed SS network, the author applies the same optimal navigation algorithm [2,3,47]. On the contrary, the segmented images acquired for this research were processed utilizing the proposed KD-SegNet that was proposed, thereby facilitating the generation of the

grid-map. Furthermore, a specialized local search algorithm was created to improve the obstacle avoidance safety of the MR while it effectively follows the global A* path [2,3,56]. Based on a comparison between the novel findings and those reported in [2,3,56], it is feasible to deduce that SS plays a critical role in generating the frontal perspective of the ground. This process aids in the formulation of an ideal trajectory for MR.

Based on the proposed theoretical framework, the authors executed mixed reality navigation experiments within a controlled simulation environment. The spatial and topographical configurations are depicted in Figs. 14 and 15. The experimental area encompasses approximately four hundred square meters, incorporating obstacles such as individuals, storage facilities, and structural pillars. The MR will start at different locations S (see Figs. 14 and 15) and plan the path to the destination G. To test the environment perception capability through the segmented image output data of the KD-Segnet network, in both environments there are mobile objects, human 1 and human 2, which will move from top to bottom on the left and from bottom to top on the right, respectively. With Scenario 1, in Fig. 14, four snapshots from Fig. 14a to b illustrate the ability to establish a successful path strategy for MR. Furthermore, at Fig. 14c, when MR detects that human 1 intersects and causes a collision with its moving trajectory, MR adjusts its direction and velocity to ensure safety and successfully reaches the destination G (see Fig. 14d). Based on the bird-eyes view of the monocular fixed on the MR, four snapshots of from Fig. 14e–h illustrate the live 2D input image and corresponding to them are four snapshots of from Fig. 14i–l represent segmented images.

Further to reinforce the correctness of the proposed method, twelve snapshots from Fig. 15a to m, demonstrate that the navigation plan is feasible when MR follows the moving trajectory from the starting point S to the destination point G, calculating and avoiding the mobile obstacles human 1 and human 2, as well as the static obstacles in the environment of Scenario 2. By leveraging the superior accuracy and rapid processing capabilities of the knowledge distillation model, the segmentation process is both expedient and dependable. The mean Intersection over Union (mIoU) metric recorded during the training phase with the data acquired from the environment attains a value of 89%. Consequently, the mobile trajectory of the mixed reality system is continuously computed to correspond with the conditions posed by obstacles. Similar to the segmentation illustration, the white region is interpreted as the area where the navigation trajectory can be established. The remaining objects are systematically prioritized by the algorithm for evasion. As evidenced in the two scenarios presented, the mixed reality system efficiently determines the optimal path trajectory from the initial point to the target point, while simultaneously avoiding collisions with static obstacles present in the environment. Utilizing the adaptability and accelerated inference capabilities of KD-SegNet, the system persistently recalculates the optimal path every three frames. This results in the mixed reality system partially adjusting to variations in correlation with mobile obstacle entities (individuals). Such functionality significantly enhances the efficacy of mixed reality orientation planning. As illustrated in both Figs. 14 and 15, initially, the MR scans the entire surrounding context to determine a preliminary trajectory. The MR approaches and adheres to the initially established trajectory. Upon encountering an obstacle in a new context area, the steering angle is recalculated to minimize the likelihood of a collision, thereby adjusting the movement trajectory based on obtained MR's path planning.

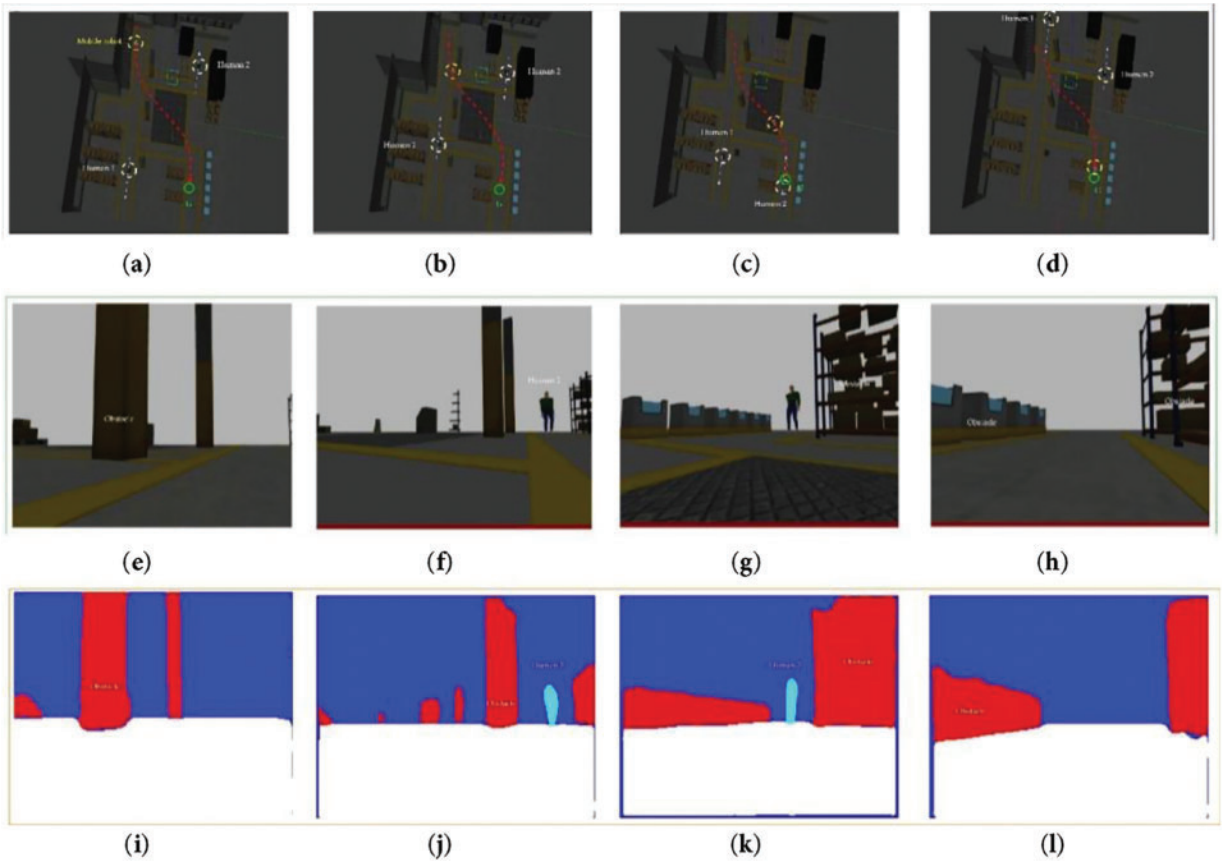

Figure 14: Simulation results of MR navigation in the Gazebo environment. In senarios 1: including global position (yellow dotted-line circle) and trajectory (red-dotted line), as well as frames collected and segmented during the corresponding movement with from (a) to (d): MR's Gazebo environment, from (e) to (h): 2D input images, and from (i) to (l): segmented images

To demonstrate the pratical application, the real MR is stationed with a Jetson nano and monocular camera, in Fig. 16. The aim of the practical experiments conducted in this study was to enhance methods for detecting collision-free areas in local search zones while adhering to a predetermined global trajectory, in Fig. 17. A seamless trajectory would compromise the proposed KD-SegNet outcomes because of the MR camera pose [57]. In other words, the results generated by the model we have put forth would surpass those achieved by the prior binary segmentation FCN-VGG [2,3]. In addition, the processing speed of the KD-SegNet is significantly accelerated and refined when compared to the lightweight IRDC-Net that is built upon FCN-MobileNetV2 [47,58].

Consequently, the deviation and stability of the MR's steering angle according to both x-axis and y-axis are less than 0.05 rad. Then, the MR's stability and robustness has been improved with smaller angle steering changes throughout trajectory tracking when compared to [58], in Fig. 18.

Finally, the authors the substantial correlation between the precision of MR's trajectory tracking and the performance of KD-SegNet's output. The MR maintains a seamless tracking trajectory with minimal steering angle variation of less than 0.05 rad by generating the MR's frontal view and moving region. Moreover, due to the stable mounting of the monocular camera on the MR, even minor adjustments in steering angle enhance stability and improve the overall quality of segmentation outcomes (see Table 5).

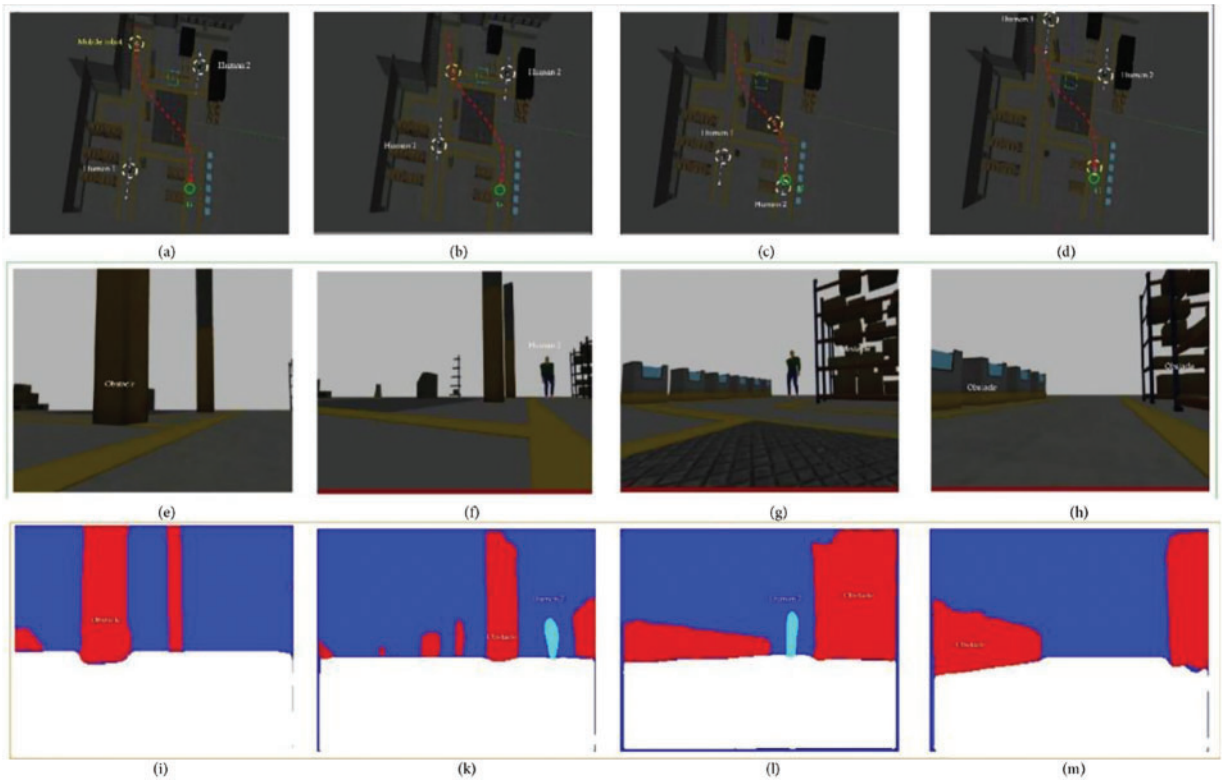

Figure 15: Simulation results of MR navigation in the Gazebo environment. In senarios 2: includes global position (yellow dotted-line circle) and trajectory (red-dotted line), as well as frames collected and segmented during the corresponding movement with from (a) to (d): MR's Gazebo environment, from (e) to (h): 2D input images, and from (i) to (m): segmented images

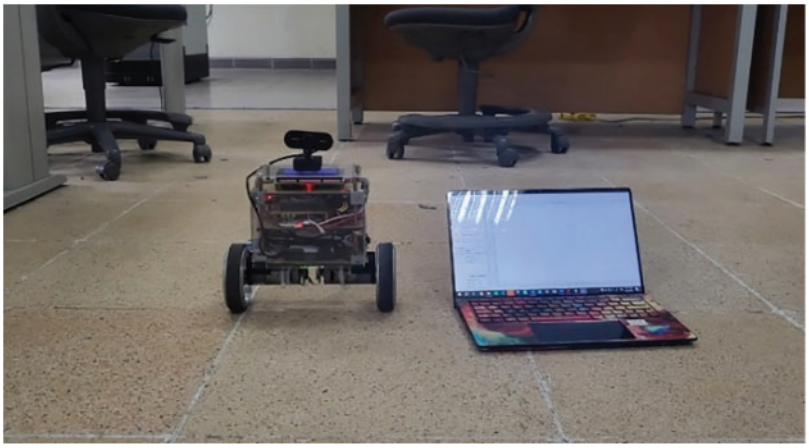

Figure 16: Practical MR

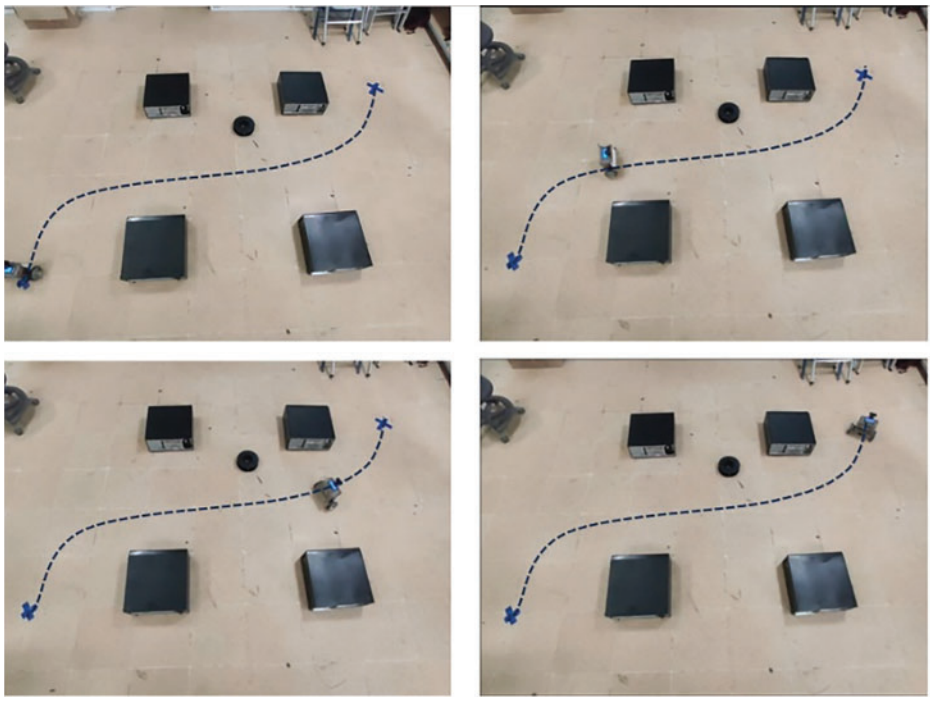

Figure 17: MR tracks the path based on the MR's bird view from proposed KD-SegNet

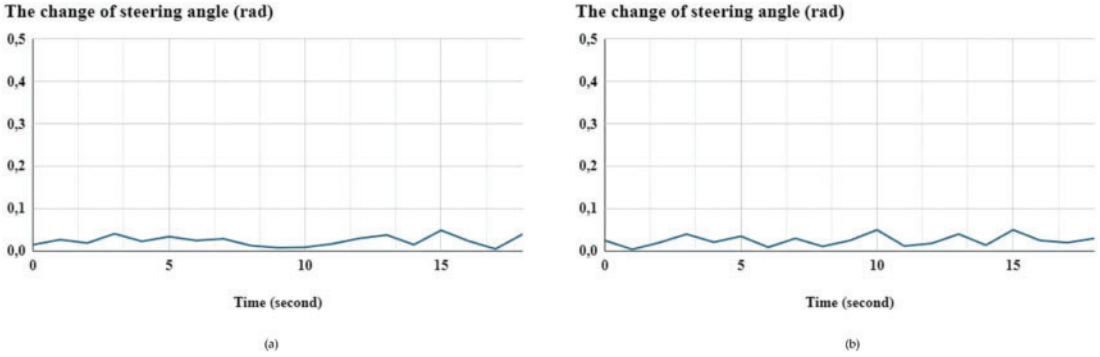

Figure 18: The steering angle changing between coordinates with (a): rotation around x-axis and (b) rotation around y-axis

Table 5: The influence of steering angle changing on the proposed KD-SegNet

MR's frontal view	Steering angle change (rad)	Accuracy (%)
When rotating around x-axis	0.00	100
	0.01	98.6
	0.02	98.2
	0.03	97.4
	0.04	96.7

(Continued)

Table 5 (continued)

MR's frontal view	Steering angle change (rad)	Accuracy (%)
	0.05	96.3
	0.1	86.5
	0.15	83.1
When rotating around *y*-axis	0.00	100
	0.01	99
	0.02	98.5
	0.03	98.1
	0.04	97.3
	0.05	96.4
	0.1	85.8
	0.15	82.1

5 Conclusions

The paper proposes a real-time MR navigation solution that relies on extracting features from monocular camera images. Specifically, the SS model integrating with KD enhances the efficiency of training and the speed of inference. Both the student and teacher models retain their outstanding performance. Hence, the proposed KD-SegNet becomes a compact model that is efficient in segmenting, and capable of training and inferring quickly. Furthermore, combining the Adam optimizer with Gaussian preprocessing further improves the accuracy of segmentation and ensures environmental compatibility. Comparative outcomes with contemporary methods including Binary-SegNet, PSP-Net, FCN, U-Net, and IRDC-Net serve to illustrate the practicability of the proposed method. Experiments are performed on four datasets of Cityscapes, Kitti, Cifar10, and TQB datasets yielding excellent evaluation outcomes. This opens up the opportunity to optimally exploit system resources or deploy SS tasks on systems with limited computational capabilities. When implemented in MR navigation, a 0.05 rad reduction in steering angle change enhances the trajectory of the vehicle. Furthermore, it is evident that the proposed method effectively captures rapid modifications in the operational MR's environment. By minimizing Kullback-Leibler divergence, the authors intend to extract the channel-wise probability map between the teacher and student models as a means of expanding the scope of their investigation. Future experiments will combine model weight distillation with other model compression methods to optimize performance for semantic segmentation tasks. Based on the obtained semantic segmented outputs, real-time MR's path planning is designed. Overall, the proposed KD-SegNet model shows great promise for the development of MR's perception systems.

Acknowledgement: Lab 821M-C7: Computer Vision and Autonomous Mobile Robot (CVMR), SME, HUST; iRobot Lab, SMAE, HaUI, Vietnam; and Mobile Multimedia Communications Laboratory (MMC), SIT, Japan is gratefully acknowledged for providing work location, guidance, and expertise.

Funding Statement: This research is funded by Hanoi University of Science and Technology (HUST) under project number T2023-PC-008.

Author Contributions: Conceptualization, Nhu-Nghia Bui and Thai-Viet Dang; Methodology, Phan Xuan Tan and Thai-Viet Dang; Software, Nhu-Nghia Bui; Formal analysis, Thai-Viet Dang; Investigation, Nhu-Nghia Bui; Resources, Thai-Viet Dang; Data curation, Nhu-Nghia Bui; Writing—original draft, Nhu-Nghia Bui and Thai-Viet Dang; Writing—review & editing, Phan Xuan Tan and Thai-Viet Dang; Visualization, Nhu-Nghia Bui; Funding, Phan Xuan Tan and

Thai-Viet Dang; Supervision, Phan Xuan Tan and Thai-Viet Dang; Project administration, Thai-Viet Dang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available on request from the corresponding author, Thai-Viet Dang with TQB dataset: https://github.com/buinghia3101/TQB-Dataset.git and MR's navigation on Gazebo environment: https://youtu.be/di5n4Em62vE?feature=shared, accessed on 25 December 2024.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

Glossary

CNN	Convolutional Neural Network
CIRKD	Cross-Image Relational KD
CRKD	Contrastive Regularization Knowledge Distillation
CWD	Channel-Wise Distillation
GPU	Graphics Processing Unit
DL	Deep learning
FCNs	Fully Convolutional Networks
IFVD	Intra-class Feature Variation KD
KD	Knowledge Distillation
mIoU	mean Intersection over Union
MRs	Mobile Robots
PSP	Pyramid Scene Parsing
SS	Semantic Segmentation
SKD	Structural KD
TransKD	Transformer-based KD

References

1. Liu Y, Wang S, Xie Y, Xiong T, Wu M. A review of sensing technologies for indoor autonomous mobile robots. Sensors. 2024;24(4):1222. doi:10.3390/s24041222.
2. Dang TV, Bui NT. Multi-scale fully convolutional network-based semantic segmentation for mobile robot navigation. Electronics. 2023;12(3):533. doi:10.3390/electronics12030533.
3. Dang TV, Bui NT. Obstacle avoidance strategy for mobile robot based on monocular camera. Electronics. 2023;12(8):1932. doi:10.3390/electronics12081932.
4. Sohail A, Nawaz NA, Ali Shah A, Rasheed S, Ilyas S, Ehsan MK. A systematic literature review on machine learning and deep learning methods for semantic segmentation. IEEE Access. 2022;10(1):134557–70. doi:10.1109/ACCESS.2022.3230983.
5. Zhao S, Yue X, Zhang S, Li B, Zhao H, Wu B, et al. A review of single-source deep unsupervised visual domain adaptation. IEEE Trans Neural Netw Learning Syst. 2022;33(2):473–93. doi:10.1109/TNNLS.2020.3028503.
6. Wang B, Shi W. Automatic pterygopalatine fossa segmentation and localisation based on DenseASPP. Robot Comput Surg. 2024;20(2):e2633. doi:10.1002/rcs.2633.
7. Kuo TC, Cheng TW, Lin CK, Chang MC, Cheng KY, Cheng YC. Using DeepLab v3 + -based semantic segmentation to evaluate platelet activation. Med Biol Eng Comput. 2022;60(6):1775–85. doi:10.1007/s11517-022-02575-3.
8. Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Mach Intell. 2017;39(4):640–51. doi:10.1109/TPAMI.2016.2572683.
9. Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA: IEEE. p. 6230–9. doi:10.1109/CVPR.2017.660.

10. Zhao S, Feng Z, Chen L, Li G. DANet: a semantic segmentation network for remote sensing of roads based on dual-ASPP structure. Electronics. 2023;12(15):3243. doi:10.3390/electronics12153243.
11. Huang Z, Wang X, Wei Y, Huang L, Shi H, Liu W, et al. CCNet: criss-cross attention for semantic segmentation. IEEE Trans Pattern Anal Mach Intell. 2019;45(6):6896–908. doi:10.1109/ICCV43118.2019.
12. Zhou T, Wang W. Cross-image pixel contrasting for semantic segmentation. IEEE Trans Pattern Anal Mach Intell. 2024;46(8):5398–412. doi:10.1109/TPAMI.2024.3367952.
13. Dong C. Image semantic segmentation using improved ENet network. J Inf Process Syst. 2021;17(5):892–904. doi:10.3745/JIPS.02.0164.
14. Zhao S, Hao G, Zhang Y, Wang S. A real-time semantic segmentation method of sheep carcass images based on ICNet. J Robotics. 2021;2021(2):1–12. doi:10.1155/2021/8847984.
15. Li H, Xiong P, Fan H, Sun J. DFANet: deep feature aggregation for real-time semantic segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA: IEEE. doi:10.1109/cvpr.2019.00975.
16. Mehta S, Rastegari M, Caspi A, Shapiro L, Hajishirzi H. ESPNet: efficient spatial pyramid of dilated convolu-tions for semantic segmentation. In: Computer Vision—ECCV 2018: 15th European Conference; 2018; Cham: Springer International Publishing. p. 561–80. doi:10.1007/978-3-030-01249-6_34.
17. Yu C, Gao C, Wang J, Yu G, Shen C, Sang N. BiSeNet V2: bilateral network with guided aggregation for re-al-time semantic segmentation. Int J Comput Vis. 2021;129(11):3051–68. doi:10.1007/s11263-021-01515-2.
18. Abdelrahman A, Viriri S. EfficientNet family U-Net models for deep learning semantic segmentation of kidney tumors on CT images. Front Comput Sci. 2023;5:1235622. doi:10.3389/fcomp.2023.1235622.
19. Rybczak M, Kozakiewicz K. Deep machine learning of MobileNet, efficient, and inception models. Algorithms. 2024;17(3):96. doi:10.3390/a17030096.
20. Zhang X, Zhou X, Lin M, Sun J. ShuffleNet: an extremely efficient convolutional neural network for mobile devices. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA: IEEE. p. 6848–56. doi:10.1109/CVPR.2018.00716.
21. Amirkhani A, Khosravian A, Masih-Tehrani M, Kashiani H. Robust semantic segmentation with multi-teacher knowledge distillation. IEEE Access. 2021;9:119049–66. doi:10.1109/ACCESS.2021.3107841.
22. Jin R, Yu T, Han X, Liu Y. The segmentation of road scenes based on improved ESPNet model. Secur Commun Netw. 2021;2021:1681952. doi:10.1155/2021/1681952.
23. Wang Y, Yang L, Liu X, Yan P. An improved semantic segmentation algorithm for high-resolution remote sensing images based on DeepLabv3. Sci Rep. 2024;14(1):9716. doi:10.1038/s41598-024-60375-1.
24. Lin G, Liu F, Milan A, Shen C, Reid I. RefineNet: multi-path refinement networks for dense prediction. IEEE Trans Pattern Anal Mach Intell. 2019:42(5):1228–42. doi:10.1109/TPAMI.2019.2893630.
25. Ghiasi G, Fowlkes CC. Laplacian pyramid reconstruction and refinement for semantic segmentation. In: Computer Vision—ECCV 2016: 14th European Conference; 2016; Cham: Springer International Publishing. p. 519–34. doi:10.1007/978-3-319-46487-9_32.
26. Chollet F. Xception: deep learning with depthwise separable convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA: IEEE. p. 1800–7. doi:10.1109/CVPR.2017.195.
27. Rajamani KT, Rani P, Siebert H, ElagiriRamalingam R, Heinrich MP. Attention-augmented U-Net (AA-U-Net) for semantic segmentation. Signal Image Video Process. 2023;17(4):981–9. doi:10.1007/s11760-022-02302-3.
28. Brahimi S, Ben Aoun N, Benoit A, Lambert P, Ben Amar C. Semantic segmentation using reinforced fully convolutional densenet with multiscale kernel. Multime Tools Appl. 2019;78(15):22077–98. doi:10.1007/s11042-019-7430-x.
29. Zhang J, Yang J, An T, Wu P, Ma C, Zhang C, et al. AFC-ResNet18: a novel real-time image semantic seg-mentation network for orchard scene understanding. J ASABE. 2024;67(2):493–500. doi:10.13031/ja.15682.
30. Wang C, Zhong J, Dai Q, Qi Y, Shi F, Fang B, et al. Multi-view knowledge distillation for efficient semantic segmentation. J Real Time Image Process. 2023;20(2):39. doi:10.1007/s11554-023-01296-6.

31. Lu H, Liu Z, Zhang M. Distilling the knowledge in object detection with adaptive balance. In: 2022 16th IEEE International Conference on Signal Processing (ICSP); 2022 Oct 21–24; Beijing, China: IEEE. p. 272–5. doi:10.1109/ICSP56322.2022.9965214.
32. Liu Y, Chen K, Liu C, Qin Z, Luo Z, Wang J. Structured knowledge distillation for semantic segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA: IEEE. doi:10.1109/cvpr.2019.00271.
33. Ji D, Wang H, Tao M, Huang J, Hua X, Lu H. Structural and statistical texture knowledge distillation for se-mantic segmentation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA: IEEE; p. 16855–64. doi:10.1109/CVPR52688.2022.01637.
34. Wang Y, Zhou W, Jiang T, Bai X, Xu Y. Intra-class feature variation distillation for semantic segmentation. In: Computer Vision–ECCV 2020: 16th European Conference; 2020 Aug 23–28; Glasgow, UK: Springer International Publishing. p. 346–62.
35. Shu C, Liu Y, Gao J, Yan Z, Shen C. Channel-wise knowledge distillation for dense prediction. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada: IEEE. p. 5311–20. doi:10.1109/ICCV48922.2021.00526.
36. Arnaudo E, Cermelli F, Tavera A, Rossi C, Caputo B. A contrastive distillation approach for incremental se-mantic segmentation in aerial images. In: Image Analysis and Processing—ICIAP 2022; 2022; Cham: Springer International Publishing. p. 742–54. doi:10.1007/978-3-031-06430-2_62.
37. Yang C, Zhou H, An Z, Jiang X, Xu Y, Zhang Q. Cross-image relational knowledge distillation for semantic segmentation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA: IEEE. p. 12309–18. doi:10.1109/CVPR52688.2022.01200.
38. Sampath V, Maurtua I, Aguilar Martín JJ, Iriondo A, Lluvia I, Rivera A. Vision Transformer based knowledge distillation for fasteners defect detection. In: 2022 International Conference on Electrical, Computer and Energy Technologies (ICECET); 2022 Jul 20–22; Prague, Czech Republic: IEEE. p. 1–6. doi:10.1109/ICECET55527.2022.9872566.
39. Zhang X, Xu G, Wu X, Liao W, Xiao L, Jiang Y, et al. Fast-SegNet: fast semantic segmentation network for small objects. Multimed Tools Appl. 2024;83(34):81039–55. doi:10.1007/s11042-024-18829-1.
40. Kim T, Oh K, Kim J, Lee Y, Choi J. Development of ResNet152 UNet++-based segmentation algorithm for the tympanic membrane and affected areas. IEEE Access. 2023;11:56225–34. doi:10.1109/ACCESS.2023.3281693.
41. Yang H, Liu Y, Xia T. Defect detection scheme of pins for aviation connectors based on image segmentation and improved RESNET-50. Int J Image Grap. 2024;24(1):2450011. doi:10.1142/S0219467824500116.
42. Vaishali S, Neetu S. Enhanced copy-move forgery detection using deep convolutional neural network (DCNN) employing the ResNet-101 transfer learning model. Multimed Tools Appl. 2024;83(4):10839–63. doi:10.1007/s11042-023-15724-z.
43. Adhinata FD, Ramadhan NG. Real time fire detection using color probability segmentation and DenseNet model for classifier. Int J Adv Comput Sci Applications. 2022;13(9). doi:10.14569/issn.2156-5570.
44. Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, et al. The cityscapes dataset for semantic urban scene understanding. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA: IEEE. p. 3213–23. doi:10.1109/CVPR.2016.350.
45. Abu Alhaija H, Mustikovela SK, Mescheder L, Geiger A, Rother C. Augmented reality meets computer vision: efficient data generation for urban driving scenes. Int J Comput Vis. 2018;126(9):961–72. doi:10.1007/s11263-018-1070-x.
46. Ayi M, El-Sharkawy M. RMNv2: Reduced Mobilenet V2 for CIFAR10. In: 2020 10th Annual Computing and Communication Workshop and Conference (CCWC); 2020 Jan 06–08; Las Vegas, NV, USA: IEEE. doi:10.1109/CCWC47524.2020.9031131.
47. Dang TV, Tran DM, Tan PX. IRDC-Net: lightweight semantic segmentation network based on monocular camera for mobile robot navigation. Sensors. 2023;23(15):6907. doi:10.3390/s23156907.

48. 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. Comput Intell Neurosci. 2023;2023(1):4765891. doi:10.1155/2023/4765891.
49. Kostková J, Flusser J, Lébl M, Pedone M. Handling Gaussian blur without deconvolution. Pattern Recognit. 2020;103(2):107264. doi:10.1016/j.patcog.2020.107264.
50. Asgari Taghanaki S, Abhishek K, Cohen JP, Cohen-Adad J, Hamarneh G. Deep semantic segmentation of nat-ural and medical images: a review. Artif Intell Rev. 2021;54(1):137–78. doi:10.1007/s10462-020-09854-1.
51. Polino A, Pascanu R, Alistarh D. Model compression via distillation and quantization. arXiv:1802.05668. 2018.
52. Dang TV, Bui NN, Bui NT, Bui. Binary-SegNet: efficient convolutional architecture for semantic segmenta-tion based on monocular camera. In: From smart city to smart factory for sustainable future: conceptual framework, scenarios, and multidiscipline perspectives; Cham: Springer Nature Switzerland; 2024. p. 275–85.
53. Zhao L, Wang Y, Duan Z, Chen D, Liu S. Multi-source fusion on image semantic segmentation model of genera-tive adversarial networks based on FCN. IEEE Access. 2021;9:101985–93. doi:10.1109/ACCESS.2021.3097054.
54. Alfarhan M, Deriche M, Maalej A. Robust concurrent detection of salt domes and faults in seismic surveys using an improved UNet architecture. IEEE Access. 2020;10:39424–35. doi:10.1109/ACCESS.2020.3043973.
55. Raju Y, Narayana M. Satellite image segmentation using UNet++ with VGG19 deep learning model. Edelweiss Appl Sci Technol. 2024;8(6):5707–22. doi:10.55214/25768484.v8i6.3242.
56. Dang TV, Nguyen DS. Optimal navigation based on improved A* algorithm for mobile robot. In: Intelligent systems and networks. Singapore: Springer Nature Singapore; 2023. p. 574–80. doi:10.1007/978-981-99-4725-6_68.
57. Nguyen VT, Do CD, Dang TV, Bui TL, Tan PX. A comprehensive RGB-D dataset for 6D pose estimation for industrial robots pick and place: creation and real-world validation. Results Eng. 2024;24(4):103459. doi:10.1016/j.rineng.2024.103459.
58. Dang TV, Tan PX. Hybrid mobile robot path planning using safe JBS-A*B algorithm and improved DWA based on monocular camera. J Intell Robot Syst. 2024;110(4):151. doi:10.1007/s10846-024-02179-z.