 doi.org/10.24191/jeesr.v19i1.018

Performance Evaluation of Semantic Segmentation using Efficient Neural Network (ENet) on Various Traffic Scene Conditions

Miza Fatini Shamsul Azmi, Fadhlan Hafizhelmi Kamaru Zaman, Husna Zainol Abidin

Abstract— **Object recognition, object detection, and semantic segmentation are fundamental components of the intelligent vehicle. Recently, there have been various methods proposed to create a reliable and accurate model to provide intelligent assistance to drivers. However, a reliable and accurate model in adverse conditions such as snow, rain, and fog remain a problem for advance driving assistance systems. The methods proposed only effectively solve the problem in a specific condition. Therefore, in this work, we focus on performing semantic segmentation in normal, rainy, foggy, and low light conditions using Efficient Neural Network (ENet) and ResNet18 and subsequently evaluating the trained model's performance in these conditions. In the experiment, we used a daytime dataset from CamVid and synthetically transformed the daytime dataset into rainy, foggy, and low light conditions. To verify the accuracy of the proposed method, the Intersection over Union (IoU) is used, and the result is elaborated in the section result and discussion. This approach only performs accurately during daylight. From the experiments, both methods do suffer from various conditions, but the ENet method performs better in certain conditions compared to ResNet18.**

Index Terms— **Semantic segmentation, object detection, deep learning, ENet, vehicle intelligence**

I. INTRODUCTION

OBJECT detection and semantic segmentation are two important techniques used by researchers to create a reliableautonomous vehicle. In [1], the results obtained show that when object detection and semantic segmentation were applied together, they improved the model efficiency compared totraining separately. A semantic task in an automated vehicle needsto be accurate, robust, and real-time as discussed in [2]. Scene understanding is mainly about semantic segmentation, object detection, object recognition, and the relationship between objects [3]. The relationship between an object in a driving scene includes the understanding and prediction of pedestrians [4], vehicles [3], traffic lights [5], and scenes in various conditions [6]. This is why vehicle-related scene understanding plays an important role in autonomous vehicles. Based on [7], labeling surface in scene understanding can be summarized as the following: (1) Labelling visible surfaces and objects: apply the classifier to label pixels into two groups, which are visible foreground and background, using an image labeling algorithm and a pre-trained object detector; (2) label hidden surfaces; and (3) region overlay as a scene and shape.

However, with the development of deep learning, researchers started to adopt end-to-end models because the use of deep learning has increased their performance in complex conditions. One well-known way to implement deep learning is through the Convolution Neural Network (CNN). Based on the approach, it has been applied to traffic scenes [5] , pedestrian detection [8], lane detection [9], and weather detection [10]. Deep learning understands the scene by separating it into several key aspects. First, it determines the classification of each pixel through deep learning. Then the deep neural network will be able to recognize regions in the scene with boundaries. As the focus of learning goes deeper, the deep learning neural network can classify the objects in the scene [11]. To improve the efficacy of the semantic segmentation algorithm, the application of CNN to semantic image segmentation is proposed in [12].

The structure of CNN consists of two layers, which are the extraction layer and the feature map. There are many examples of CNN architecture that are well known, such as LeNet [13], AlexNet [14], GoogleNet [15], Visual Geometry Group (VGGNet) [16], and Residual neural network (ResNet) [17]. The ResNet architecture contains a skip connection which can improve thelearning ability of a network. It could also train up to hundredsor thousands of layers and still perform well. Based on [18], Efficient Neural Network (ENet) has adopted ResNet strategies where the ResNet architecture contains skip connection, which can improve the learning ability of a network. It could also train up to hundreds or thousands of layers and still

This manuscript is submitted on 10th February 2021 and accepted on 2nd September 2021. This work is supported by the Ministry of Education (MOE) Malaysia under the grant 600-IRMI/FRGS 5/3/ (081/2019).

Miza Fatini Shamsul Azmi, Fadhlan Hafizhelmi Kamaru Zaman, and Husna Zainol Abidin are with the Vehicle Intelligence and Telematics Lab, School of Electrical Engineering, Universiti Teknologi MARA, 40450 Shah Alam, Selangor (email: fadhlan@uitm.edu.my).

perform well. ENet was then tested on three different datasets of road scenes and indoor scenes. The average accuracyand intersection over union (IoU) metrics show that ENet is faster than SegNet.

Semantic segmentation is essential for scene understanding.It classifies each pixel in the image into a specific group. In [19], the model performed object detection by labeling pixels belonging to the same group. The method is used to detect roads, understand traffic signs, and detect theidentification of vehicles and pedestrians. The detector includes the road lane as a reference to detect objects which might affect the driving situation, and then it excludes theobjects that might affect the driving condition from the recognition task. The accuracy of detection and recognition of these objects will be the guideline to ensure the safety of autonomous vehicles. At the same time, image information can be applied to understand scenes in bad weather conditions, such as fog, snow, and sand storms.

The rest of this paper is arranged as follows. Section 2 outlines some related work on object detection, semantic segmentation, and ENet. Section 3 describes and elaborates on the details of the methodology and ENet architecture as well as the dataset used in the evaluation. Section 4 analyses and discusses the results obtained from the experiment conducted. Finally, section 5 outlines the conclusion and future work.

II. RELATED WORKS

This section reviews the most relevant works to thispaper, including object detection, deep learning, semantic segmentation networks, and ENet.

A. Object detection

Object detection is the task of recognizing and localizing multiple objects in a scene with bounding boxes. Based on [20], object detection methods are mainly divided into two-stage and single-stage methods. Extracting Regions of Interest (ROI) or Region Proposals (RP) from a scene of several groups of objects is the first step of a two-stage method. Then it is verified, classified, and separated. The Region-based Convolutional Network (RCNN) is one of the state-of-the-art detection methods and a classic object detection model as discussed in [21]. As for the single-stage method, it does not have an RP module, but instead, the aim is to map features directly to the bounding box. For example, based on [22], the YOLO model processes images in real-time at 45 frames per second and YOLO has twice the mean average precision of other real-time systems. However, YOLO still lags in terms of accuracy.

B. Deep learning

There are different types of deep learning techniques. For example, the Fully Connected Neural Network (FCNNs), CNN, recurrent neural network, generative adversarial network, deepreinforcement learning, and autoencoder. Deep learning is a method where the computer mimics the human brain as it makes use of a deep neural network. Its algorithm consists of an input layer, an output layer, and a hidden layer. Recently, deep neuralnetworks have been applied in various sectors, for example, autonomous vehicles. In [23], deep learning is used to achieve automatic detection of pedestrians. A method in use is DeepLabv3, which contains various semantic information from the encoder, which allows extracting features at arbitrary resolution. This work focuses on two classes, which are pedestrians and background. Based on its accuracy, theapproach shows that the pedestrian detection error is still huge. Based on [24], a traditional method of detection was divided into 3 phases which are selecting the region, extracting features, and classifying with a classifier. The problem with these methods is that the region of selection strategy is not specific and the hand design does not react well to change in conditions. Therefore, with deep learning, detection and classification have been more efficient than with the traditional method.

C. Semantic segmentation

The objective of semantic segmentation is to label each pixel with semantics (pixel-level) or by simultaneously detecting objects and doing per-instance pixel labeling (instance-level). The Fully Convolution Network (FCN) is the first to adopt a full convolution network for semantic segmentation. In [25], fully connected layers in a CNN classifier for predicting classification are replaced with convolutional layersto produce output maps. These maps are then be up-sampled to dense pixel labels by deconvolution.

One semantic segmentation application in bad weather is shown in [26], where the authors used images of fog conditions and applied a specific training so that the networks learn to focus on the undisturbed sensor and ignore unknown noise. The evaluation of the model was applied in good weather and with unknown disturbance conditions. The work has also proposed using early and late fusion to increase the accuracy in such conditions. Although the model has created a solution for foggy conditions, the same method could not be applied to other weather or different light conditions because the method used only focuses on a very specific problem.

In this work, ENet semantic segmentation is applied to several conditions, such as rain [27], fog [26], low light [28], and day [29]. In previous methods, bright and clear surroundings images were favorable for being used to test a new method. However, the method proposed is only applicable to that trained condition. There are 3 types of methods to evaluate the accuracy of semantic segmentation prediction, which are based on region accuracy, contour-based score, and measuring per image. The examples of region-based accuracies are overall pixel, per-class (PC), Jaccard Index, and trimap. As for object detection, mAP (mean Average Precision) is popular for measuring the accuracy of object detectors. The metrics are usually evaluated in comparison to ground truth data. For object detection, the ground truth includes the image, classes of objects, and bounding boxes for each object in the image [30]. Once the bounding box of prediction and the corresponding ground truth are detected, intersection over union (IoU) can be applied. IoU is the ratio between the intersection and union of the prediction box and

the ground truth box. It is also known as the Jaccard index [31]. The IoU metric is very simple. The target and prediction masks are divided by the total number of pixels present in both masks. The score is usually calculated for each class. Besides IoU, there is another method for evaluating semantic segmentation accuracy by using pixel accuracy metrics. Accuracy is obtained by taking the ratio of correctly classified pixels (Accuracy = [TP+TN]/[TP+TN+FP+FN]). Many datasets have been published for semantic segmentation research, such as CamVid [32], Cityscape [33], KITTI [34], Toronto City [35], and ApolloScape [36]. Most of the reviewed datasets provide ground-truth labels for 2D object detection and semantic segmentation tasks. Based on [32] and [37], there are 32 semantic classes. The classes include animals, pedestrians, children, rolling carts/luggage, bicyclists, motorcycles, cars, trucks, trains, roads (drivable surface), shoulder roads, drivable lane marking, non-drivable, sky, tunnel, archway, building, wall, tree, vegetation, fence, sidewalk, parking block, pole, traffic cone, bridge, sign, text, and traffic light.

D. Semantic segmentation

The state-of-the-art CNN architecture network is going deeper by day. AlexNet had only 5 convolutional layers, while the VGG network and GoogleNet had 19 and 22 layers respectively. However, increasing network depth does not work by stacking up layers together. The main problem faced by the deep network is that the gradient back propagates to its earlier layer. To overcome this problem, auxiliary loss in the middle of the layer is added as extra supervision, but this problem was not solved until ResNet was introduced. The core idea of ResNet is that it introduces a shortcut connection that skips one or more layers in the architecture. ResNet was not the first method of using shortcut connection. It was also found in Long Term Short Memory (LSTM) and highway networks. ResNet is gaining more popularity in the research community and its architecture is being studied frequently. The entire architecture of ENet islargely based on ResNets. There are 2 main blocks in the architecture, which are the initial block and the bottleneck module. The initial block consists of 16 x 256 x 256 after concatenation of the convolution (13 filters) and MaxPooling (2 x 2). All bottlenecks have the same structure and each branch consistsof three convolutional layers. The first projection reduces the dimensionality, while the latter projection expands the dimensionality. In that layer, there is a convolution. These are examples of the ENet model used in semantic segmentation [38], intelligent vehicles [39], road lane marking [40], and mapping of road lanes [41]. While in [39], the research shows the comparison of ENet CRF Lidar and its performance in efficiency and precision. The experiment results indicate that ENet-CRF-Lidar can provide reliable multi-scale object recognition performance.

III. METHODOLOGY

In this section, the proposed ENet architecture is described in subsection A, which is based on ENet, a deep neural network architecture for real-time semantic segmentation [18]. Then implementation of training and testing is discussed subsequently, as well as the dataset. To evaluate the performance of ENet, the network is trained and tested using the collected dataset which contains traffic scenes.

A. Network Architecture

Based on [18], the ENet model is split into the initial block as depicted in Figure.1 and the bottleneck module is shown in Figure 2. The ENet initial block consists of MaxPooling performed with non-overlapping 2×2 windows, and the convolution has 13 filters, which sums up to 16 feature maps after concatenation. All bottlenecks have the same structure. Each branch consists of three convolutional layers. The "first" 1 x 1 projectionreduces the dimensionality, while the latter 1 x 1 projection expands the dimensionality. In between these convolutions, a regular, dilated or full convolution (no annotation) takes place.Batch normalization and PReLU are placed between all convolutions. To regularize the bottleneck, Spatial Dropout is used. MaxPooling on the initial block is added when the bottleneck is downsampled. The first branch is replaced by a non-overlapping 2 x 2 convolution and the activations are zero-padded to equal the number of feature maps. In the decoder, MaxPooling is replaced by MaxUnpooling.

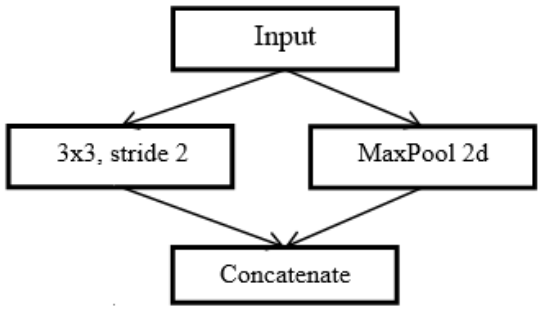


Figure 1 . Initial Block of the model

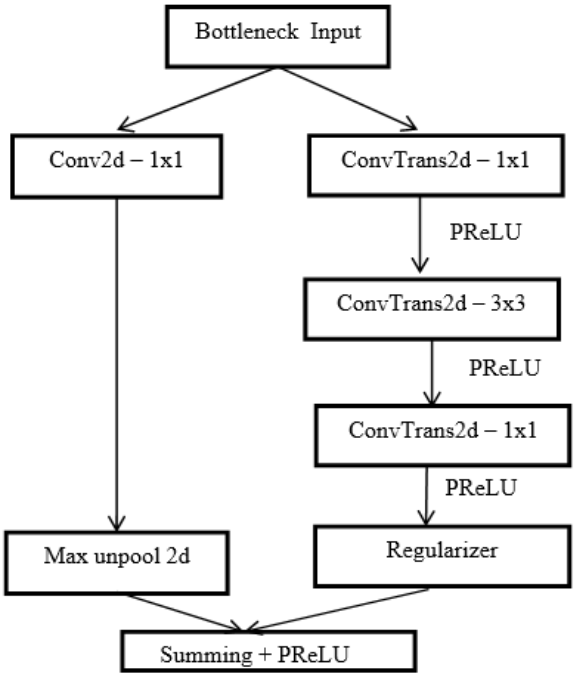


Figure 2 . ENet bottleneck module.

B. *Dataset*

In this work, the dataset was obtained from the CamVid dataset [40]. The dataset contains traffic scenes such as pedestrians, traffic lights, and vehicles as shown in Figure 3. The CamVid dataset only provides an image with a bright light condition, as shown in (a) and (b), while (c) and (d) are datasets synthetically transformed into rain conditions, where (e) and (f) aretransformed into fog conditions. (h) and (i) are transformed intolow light conditions. These photos are transformed with an online photo converter. The number of images in CamVid is 701. About 367 images are transformed for each condition, which makes the total number of images in the dataset 1468. Besides the dataset, the label ID images are also obtained from CamVid.

Figure 3. Example of a dataset for test and training.

C. *Training and testing*

In this paper, the training stage has been carried out under several conditions, such as normal, fog, rain, and low light conditions. The experiment is divided into three types:(1) Experiment A – images of the normal condition are used for training and testing. 367 images were used for training and 100 for testing. (2) Experiment B – images with various conditions are used for training and it is tested on images with various conditions such as fog, rain, and low light conditions. (3) Experiment C – images with normal conditions are used for training while images with various conditions are used for testing. About 367 images are used for training and 100 images are used for testing.

D. *Performance Metrics*

To evaluate the performance of semantic segmentation, most previous works have used the intersection over union metric or pixel accuracy. The Intersection-over-union (IoU) is also known as the Jaccard index and is very effective for compare pixel accuracy or Dice Coefficient (F1 Score). IoU is the intersection of the pixels or overlapping area found in both theprediction mask and the ground truth mask over all pixels found in either the prediction or target mask, known as the area of union. Therefore, the mean IoU score is used to calculate our semantic segmentation prediction and the concept as shown in Figure 4. The number of images used for training and testingis casually choosen based on the capacity of the CPU and GPU.

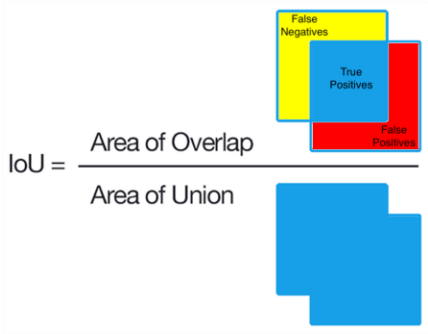


Figure 4. Intersection-over-union (IoU).

IV. RESULTS AND DISCUSSIONS

In this section, the results of experiments A, B, and C will be discussed in detail. For these experiments, the CamVid dataset provides traffic scene images and it consists of 32 semantic classes. The ground truth and predicted segmentation image. Each object has a specific color to represent the pixel group. Therefore, Figure 5 shows the list of RGB color codes for each semantic class.

Figure 5. Color code for segmentation

Experiment A: Normal condition is used for training and tested in normal condition.

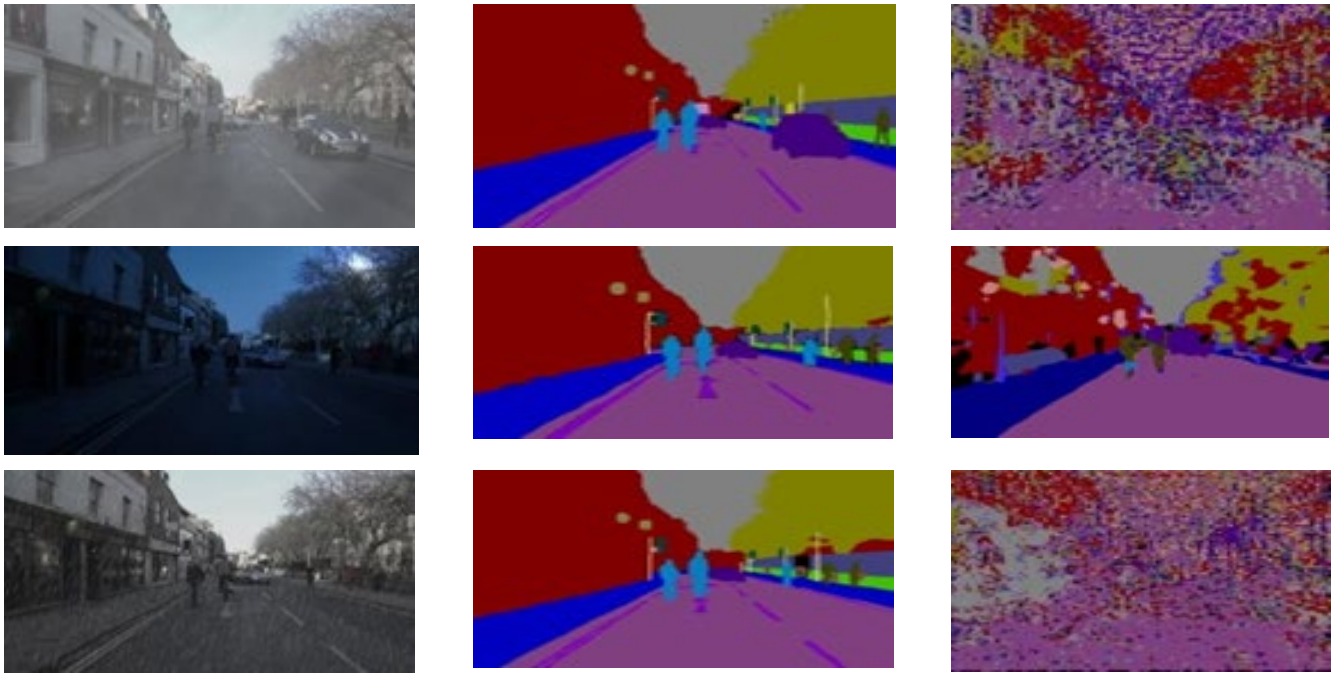
In this experiment, the dataset used to train and test the ENet semantic segmentation model was taken under natural daylight, and the sample result of the segmentation is shown in Figure 6. Based on the visual comparison between ground truth images and ENet segmentation images, the model was able to segment the imageinto distinct classes such as sky, buildings, poles, road pavement, trees, fences, cars, and pedestrians based on the color assigned by ENet. Generally, it can be seen that most of the objects were segmented correctly. However, a small object such as a traffic sign was not detected accurately. Besides, road marking were also not detected. However, road making can be dealt with effectively with specific training and template features. The percentage of accuracy measured in IoU for Experiment A is 94.3%.

Input Image Ground Truth ENet Segmentation

Figure 6. The test result of ENet semantic segmentation for Experiment A

Experiment B: Various condition is used for training and tested in various condition

For experiment B, the conditions of images used to train and test the ENet model are foggy, low-light, and rainy. Three ENet models are trained separately for each condition and tested on images containing similar conditions. The semantic segmentation result in Figure 7 shows that the ENet model can segment correctly only images under low light conditions. While for images with fog and rain, it could only recognize parts of buildings and roads correctly. Other objects have poor detection accuracy. The IoU for fog, low light, and rain conditions are 47.7%, 88.9%, and 43.1% respectively.



Input Image Ground Truth ENet Segmentation

Figure 7. The test result of ENet semantic segmentation for Experiment B. The first row shows the result for foggy condition, the middle row shows the result for low-light condition and the bottom row shows the result for rainy condition

Experiment C: Normal condition is used for training and tested in various conditions.

For experiment C, images of normal conditions are used to train the ENet model and images of various conditions are used for testing. The result in Figure 8 shows that the model failed to correctly segment objects in the image. Only part of the road and building are correctly segmented. Other than that, it was not recognized and detected accurately at all. Basedon results in Experiment A and B, ENet performance highly depends on the variations of images used to train the model. Compared to the results of Experiment A and C, even though both models were trained on the normal condition, the results shows that the model could not generalize well to unseen conditions. The model works well when tested under normal conditions, but fails when tested under different conditions. The percentage of IoU for fog, low light, and rain conditions are 33.1%, 37.6%, and 38.7%. This is tabulated in Table I.

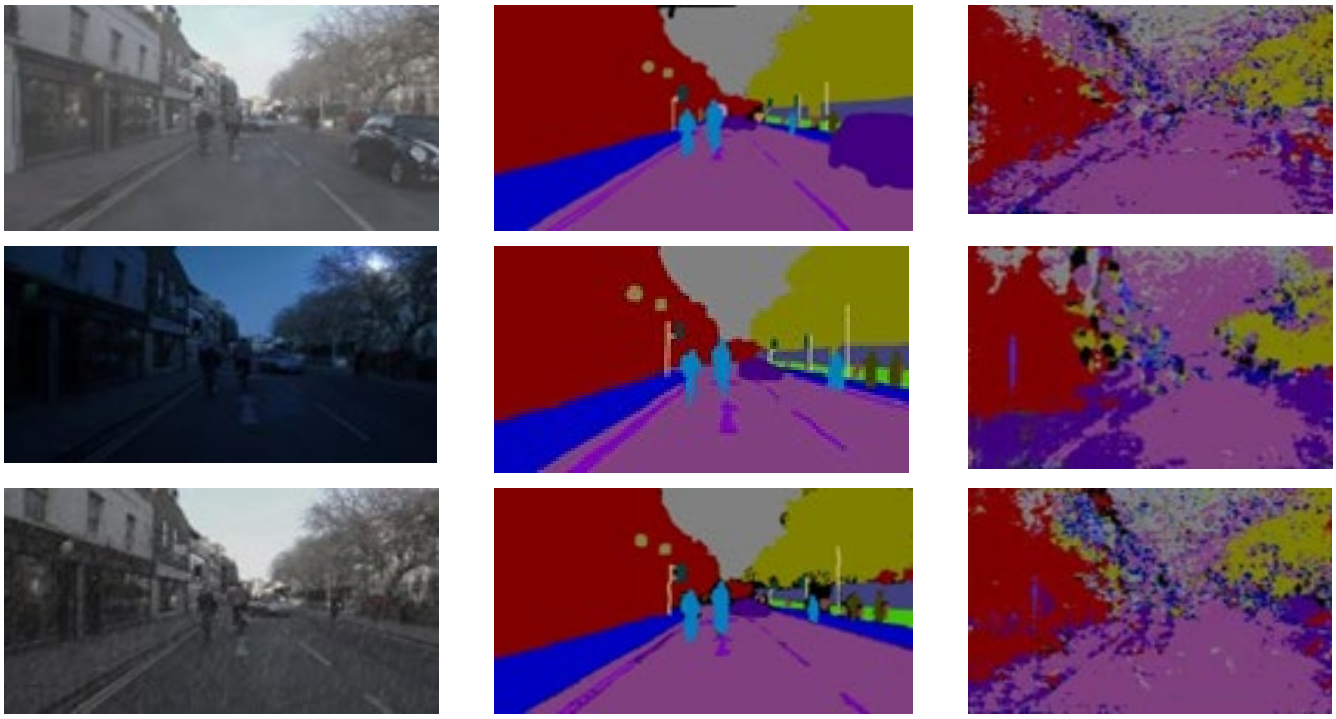


Figure 8. The test result of ENet semantic segmentation for Experiment C. First row shows the result for foggy condition, the middle row shows the result for low-light condition, the bottom row shows the result for rainy condition.

TABLE I
Summary of Intersection Over Union (IoU) performance from experiment

Experiment A, B and C	Condition	ENet IoU	ResNet18 IoU
A	Train normal, test normal	94.3%	62.0%
B	Train fog, test fog	47.7%	53.6%
	Train low light, test lowlight	88.9%	49.3%
	Train rain, test rain	43.1%	42.0%
C	Train normal, test fog	33.1%	56.1%
	Train normal, test low-light	37.6%	30.3%
	Train normal, test rain	38.7%	30.5%

Based on the results of ENet IoU in Table 1, the highest IoU is obtained when test and training are using the same dataset at the normal condition as shown in experiment A. In experiment B, the low-light condition produces acceptable IoU at 88.9%, while foggy and rainy conditions produce poor IoU at 47.7% and 43.1% respectively, even when trained and tested using the same conditions. However, the result drastically changes when the normal dataset is used for training and the various conditions are used for testing. Not more than 40% of accuracy was detected when testing in rain, fog, and low light conditions. As for ResNet18, the highest percentage of overall performance is 62%, which is while using the normal condition dataset for test and training. The result shows that the foggy dataset in experiments B and C scores 53% and 56%, low light scores 49.3% and 30.3%, and the rain dataset scores 42% and 30.5% in experiments B and C.

After performing semantic segmentation in normal, rainy, foggy, and low light conditions, the results show that the accuracy of semantics can be affected by the conditions. This proves that most existing models proposed only work accurately

when datasets are in bright light or normal conditions, instead of robust weather such as foggy, rainy, and low light. The accuracy of semantics is drastically affected, as shown in experiments B and C on both methods. The result also shows that ENet was a better method than ResNet18 in terms of normal and low light conditions. However, in the condition of rain, ENet performs slightly better than ResNet18. While in foggy conditions, ResNet accuracy is better than ENet. Overall, both methods are affected by the various conditions, but the ENet method performs better in certain conditions compared to ResNet18.

V. CONCLUSIONS

In this paper, the problem of semantic image segmentation under various conditions, such as fog, rain, and low light conditions, was presented and so was the evaluation of ENet and existing method performance. The findings from the simulation work have shown that the Enet method can accurately segment the image under normal conditions with an IoU of 94.3%. However, theperformance is poor when it is tested using images under rain, fog, and low light intensity conditions, even when the model is trained using similar conditions. ENet performs slightly better compared to ResNet18 on normal, rain, and low intensity. As for foggy conditions, ResNet is more accurate compared to ENet. This supports the idea that only certain methods will work accurately in certain conditions. Therefore, to improve the accuracy during rain, fog and low light, it can apply enhancement and use a variety of traffic scenes as a dataset to improve training.

ACKNOWLEDGMENT

The author would like to thank Ministry of Education for the FRGS grant (600-IRMI/FRGS 5/3/ (081/2019)) and Faculty of Electrical Engineering, Universiti Teknologi MARA for the support.

REFERENCES

[1] R. Mottaghi *et al.*, "The role of context for object detection and semantic segmentation in the wild," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, no. June, p. 9, 2013, doi: 10.1109/CVPR.2014.119.

[2] D. Feng *et al.*, "Deep Multi-Modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges," *IEEE Trans. Intell. Transp. Syst.*, no. Feb, pp. 1–27, 2020, doi: 10.1109/tits.2020.2972974.

[3] Y. Konishi, Y. Hanzawa, M. Kawade, and M. Hashimoto, "Fast 6D Pose Estimation Using Hierarchical Pose Trees," in *European Conference on Computer Vision (Eccv)*, 2016, no. October, pp. 1–17, doi: 10.1007/978-3-319-46448-0.

[4] S. Yang, W. Wang, C. Liu, and W. Deng, "Scene understanding in deep learning-based end-to-end controllers for autonomous vehicles," *IEEE Trans. Syst. Man, Cybern. Syst.*, no. October, pp. 1–12, 2018, doi: 10.1109/TSMC.2018.2868372.

[5] X. Pan, J. Shi, P. Luo, X. Wang, and X. Tang, "Spatial As Deep : Spatial CNN for Traffic Scene Understanding," *Assoc. Adv. Artif. Intell.*, no. Dec, pp. 1–8, 2017.

[6] S. Di *et al.*, "Rainy Night Scene Understanding With Near Scene Semantic Adaptation," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1–9, 2020.

[7] R. Guo and D. Hoiem, "Labeling Complete Surfaces in Scene Understanding," *Int. J. Comput. Vis.*, no. October, pp. 1–16, 2015, DOI: 10.1007/s11263-014-0776-7.

[8] Y. Y. Chen, G. Y. Li, S. Y. Jhong, P. H. Chen, C. C. Tsai, and P. H. Chen, "Nighttime pedestrian detection based on thermal imaging and convolutional neural networks," *Sensors Mater.*, vol. 32, no. 10, pp. 1–11, 2020, doi: 10.18494/SAM.2020.2838.

[9] W. Wang, H. Lin, and J. Wang, "CNN based lane detection with instance segmentation in edge-cloud computing," *J. Cloud Comput. Syst. Appl.*, vol. 9, no. 1, pp. 1–10, 2020, doi: 10.1186/s13677-020-00172-z.

[10] M. Toğaçar, B. Ergen, and Z. Cömert, "Detection of weather images by using spiking neural networks of deep learning models," *Neural Comput. Appl.*, no. October, pp. 1–13, 2020, doi: 10.1007/s00521-020-05388-3.

[11] D. Bau, J.-Y. Zhu, H. Strobelt, A. Lapedriza, B. Zhou, and A. Torralba, "Understanding the role of individual units in a deep neural network," *Proc. Natl. Acad. Sci.*, vol. 2, no. Sep, pp. 1–8, 2020, doi: 10.1073/pnas.1907375117.

[12] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 1–14, 2016, doi: 10.1109/TPAMI.2017.2699184.

[13] C. Zhang, M. Yang, and H. Zeng, "Pedestrian detection based on improved LeNet-5 convolutional neural network," J*ournal Algorithms Comput. Technol.*, vol. 13, no. January, pp. 1–10, 2019, doi: 10.1177/1748302619873601.

[14] A. Krizhevsky and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Neural Inf. Process. Syst.*, vol. 25, pp. 1–9, 2012.

[15] R. Di, "Image Classification Using Network Inception- Architecture & Applications," *Int. J. Innov. Res. Sci. Eng. Technol. ·*, vol. 10, no. 1, pp. 1–7, 2021, doi: 10.15680/IJIRSET.2021.1001055.

[16] B. B. Shabarinath and P. Muralidhar, "Convolutional neural network based traffic-sign classifier optimized for edge inference," *IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON*, no. November, pp. 1–6, 2020, doi: 10.1109/TENCON50793.2020.9293767.

[17] J. Xia, D. Xuan, L. Tan, and L. Xing, "ResNet15 : Weather Recognition on Traffic Road with Deep Convolutional Neural Network," *Hindawi Adv. Meteorol.*, vol. 2020, pp. 1–11, 2020, [Online]. Available: https://doi.org/10.1155/2020/6972826.

[18] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation," *Towar. Data Sci.*, no. April, pp. 1–10, 2016, [Online]. Available: http://arxiv.org/abs/1606.02147.

[19] C. Science *et al.*, "Roadway Image Preprocessing for Deep Learning-Based Driving Scene Understanding," *2019 IEEE Int. Conf. Big Data Smart Comput.*, pp. 1–4, 2019, [Online]. Available: https://doi.org/10.1109/BIGCOMP.2019.8679168.

[20] J. Peng, Z. Nan, L. Xu, J. Xin, and N. Zheng, "A Deep Model for Joint Object Detection and Semantic Segmentation in Traffic Scenes," in *2020 International Joint Conference on Neural Networks (IJCNN).*, 2020, pp. 1–8, doi: 10.1109/ijcnn48605.2020.9206883.

[21] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 1–8, 2014, doi: 10.1109/CVPR.2014.81.

[22] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, no. Decemember, pp. 1–10, 2016, doi: 10.1109/CVPR.2016.91.

[23] R. V. Carneiro *et al.*, "Mapping Road Lanes Using Laser Remission and Deep Neural Networks," *Proc. Int. Jt. Conf. Neural Networks*, no. July, pp. 1–8, 2018, doi: 10.1109/IJCNN.2018.8489363.

[24] P. R. Chen, S. Y. Lo, H. M. Hang, S. W. Chan, and J. J. Lin, "Efficient road lane marking detection with deep learning," *2018 IEEE 23rd Int. Conf. Digit. Signal Process. (DSP).*, pp. 1–5, 2018.

[25] E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 1–10, 2017, doi: 10.1109/TPAMI.2016.2572683.

[26] A. Pfeuffer and K. Dietmayer, "Robust Semantic Segmentation in Adverse Weather Conditions by means of Sensor Data Fusion," *FUSION 2019 - 22nd Int. Conf. Inf. Fusion*, pp. 1–8, 2019.

[27] S. Halder, J. F. Lalonde, and R. De Charette, "Physics-based rendering for improving robustness to rain," *Proc. IEEE Int. Conf. Comput. Vis.*, no. 2019-Octob, pp. 1–10, 2019, doi: 10.1109/ICCV.2019.01030.

[28] L. Sun, K. Wang, K. Yang, and K. Xiang, "See clearer at night: towards robust nighttime semantic segmentation through day-night image conversion," in *Security + Defence (2019).*, 2019, p. 8, doi: 10.1117/12.2532477.

[29] C. Li, W. Xia, Y. Yan, B. Luo, and J. Tang, "Segmenting Objects in Day and Night:Edge-Conditioned CNN for Thermal Image Semantic Segmentation," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 1, pp. 1–12, 2019, [Online]. Available: http://arxiv.org/abs/1907.10303.

[30] G. Csurka, D. Larlus, and F. Perronnin, "What is a good evaluation measure for semantic segmentation?," *BMVC 2013 - Electron. Proc. Br. Mach. Vis. Conf. 2013*, pp. 1–11, 2013, doi: 10.5244/C.27.32.

[31] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, "Acquisition of localization confidence for accurate object detection," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11218 LNCS, pp. 1–16, 2018, doi: 10.1007/978-3-030-01264-9_48.

[32] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognit. Lett.*, vol. 30, no. 2, pp. 1–10, 2009, doi: 10.1016/j.patrec.2008.04.005.

[33] M. Cordts *et al.*, "The Cityscapes Dataset for Semantic Urban Scene Understanding," *2016 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR).*, pp. 1–11, 2016, [Online]. Available: https://doi.org/10.1109/CVPR.2016.350.

[34] H. Martins, S. Bruno, and E. Luna, "LIFT-SLAM: a deep-learning feature-based monocular visual SLAM method," *Neurocomputing*, vol. 2, no. Jun, p. 2021, 2021, [Online]. Available: https://arxiv.org/pdf/2104.00099.pdf.

[35] S. Wang *et al.*, "TorontoCity : Seeing the World with a Million Eyes," *2017 IEEE Int. Conf. Comput. Vis. (ICCV).*, pp. 1–9, 2012.

[36] X. Huang *et al.*, "The apolloscape dataset for autonomous driving," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, no. June, pp. 1–7, 2018, doi: 10.1109/CVPRW.2018.00141.

[37] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in *ECCV (2008).*, 2008, pp. 1–16, doi: 10.1007/978-3-540-88682-2_5.

[38] L. I. N. Jinzhao, L. I. Guoquan, and W. Huiqian, "Vehicle Vehicle type type detection detection based based on on deep deep learning learning in in traffic traffic scene scene," *Procedia Comput. Sci.*, vol. 131, pp. 1–9, 2018, doi: 10.1016/j.procs.2018.04.281.

[39] M. A. Mikhalkova, V. O. Yachnaya, E. N. Yablokov, and V. R. Lutsiv, "Automatic detection of pedestrians in traffic scene images," *2020 Wave Electron. its Appl. Inf. Telecommun. Syst. WECONF 2020*, pp. 1–6, 2020, doi: 10.1109/WECONF48837.2020.9131458.

[40] D. N. Spoorthi and B. P. Ashwini, "Efficient Neural Network for Real Semantic Segmentation," in *Proceedings of the 3rd International Conference on Electronics and Communication and Aerospace Technology, ICECA 2019*, 2019, pp. 1–4, doi: 10.1109/ICECA.2019.8822228.

[41] Q. Deng, X. Li, P. Ni, H. Li, and Z. Zheng, "Enet-CRF-Lidar: Lidar and Camera Fusion for Multi-Scale Object Recognition," *IEEE Access*, vol. 7, pp. 1–10, 2019, doi: 10.1109/ACCESS.2019.2956492.

Miza Fatini Shamsul Azmi received the bachelor‟s degree in electrical and elecronic (Honos.) from Universiti Teknologi MARA (UiTM) in 2019. She is currently pursuing thee master‟s of science in electrical engineering from UiTM. Her current research is centered on performance evaluation of semantic segmentation using deep learning approach for autonomous driving

Fadhlan Hafizhelmi Kamaru Zaman received the B.Sc (Hons.) and P.hD. degrees from International Islamic University Malaysia in 2008 and 2015, respectively. He is currently a Senior Lecturer at Department of Computer Engineering, University of Technology MARA, Malaysia. His research interests are in surveillance system, pattern recognition, signal and image processing, artificial intelligence and computer vision. He is also the Head of Vehicle Intelligence and Telematics Lab. Fadhlan is also a professional Engineer with Board of Engineers Malaysia, and a Chartered Engineer from the Institution of Engineering and Technology, UK

Assoc. Prof. Ir. Ts. Dr Husna Zainol Abidin currently is an associate professor at Universiti Teknologi MARA, Shah Alam. She has published more than 60 research papers. She completed her Bachelor of Engineering in Electrical in 2001 from the University of Wollongong, Australia. She obtained her Master of Engineering in Electrical and PhD in Engineering from the Universiti Tenaga Nasional in 2006 and 2015 respectively. Her research interest areas includes Wireless Networking and Optimization.