# Enhanced U-Net Approach: Semantic Segmentation for Self-Driving Cars Applications

Wijden Bouzidi
*Lab. Electronics & Microelectronics*
*Faculty of Sciences of Monastir*
*National Engineering School of Monastir*
*University of Monastir, Monastir, Tunisia*
bouzidi.wijden@enim.u-monastir.tn

Soulef Bouaafia
*Lab. Electronics & Microelectronics*
*Faculty of Sciences of Monastir*
*University of Monastir, Monastir, Tunisia*
soulef.bouaafia@fsm.rnu.tn

Mohamed Ali Hajjaji
[1]*Lab. Electronics & Microelectronics*
*Faculty of Sciences of Monastir*
*University of Monastir, Monastir, Tunisia*
[2]*Higher Institute of Applied Science and Technology of Sousse*
*University of Sousse, Sousse, Tunisia*
mohamedali.hajjaji.issats@gmail.com

Luis M. Bergasa
*Electronics Department*
*University of Alcalá (UAH)*
Madrid, Spain
luism.bergasa@uah.es

***Abstract*—Semantic segmentation is a crucial task for the development of autonomous vehicles. Autonomous Driving (AD) stack is mainly classified into three subsystems, which are respectively control, planning, and perception. This work deals with the perception subsystem of AD systems. Indeed, on-road semantic segmentation for the AD systems is addressed using deep-learning algorithms. This paper proposes a deep learning approach based on an enhanced U-Net model that exploits squeeze and excitation (SE) networks. In order to demonstrate the efficiency of the added SE blocks, we opted to use U-Net as our baseline. In these experiments, the two models are trained from scratch on the CamVid dataset. Compared with the U-Net baseline, the mean class accuracy (mCA) and the mean intersection over union (mIoU) of the proposed model are increased by 5.54% and 5.87%, respectively. It achieves a mIoU of 59.88 % and a mCA of 84.15% on the CamVid dataset. The achieved results reveal that the improved U-Net reaches a better compromise between accuracy and computational complexity than previous studies.**

***Index Terms*—semantic segmentation, autonomous vehicles, Enhanced U-Net, CamVid dataset, Squeeze and excitation.**

## I. Introduction

The ability of autonomous vehicles to operate effectively and safely has been a widespread study issue in the past few years, and several manufacturers and academic institutions have been working to develop the first fully functional model. It is an exciting sector with many potential advantages, including enhanced safety, lower prices, more comfortable travel, more mobility, and a smaller environmental impact. However, developing self-driving vehicles with the ultimate level of autonomy to the point that human intervention is not required in any situation remains unresolved. The modular architecture of self-driving vehicle is typically divided into three major components: perception, planning, and control subsystems.

Accurate real time environment perception is a fundamental component for any autonomous vehicle. The perception subsystem handles a variety of tasks, including object localization, semantic segmentation, and object recognition. Semantic segmentation, called also scene parsing, particularly seeks to categorize each pixel of the input image [1], being a classification task at the pixel level. For scenario comprehension and eventual adoption of this novel technology, semantic segmentation is extremely crucial. Therefore, it is employed in wide range of sectors, including robots, medical imaging, and autonomous vehicles. For a thorough understanding of the operating condition, autonomous vehicles rely on the data collected by sensors of the surroundings [2]. Semantic segmentation is essential for interpreting scenes since the visual signals are so rich within that type of information. The more quickly and accurately we can perform the semantic segmentation task, the more the autonomous vehicle will be able to comprehend its environment and, as a result, make the best decisions in the appropriate instant. In spite of this, semantic segmentation is a challenging task because of the complex relationships that exist between the pixels in each frame of an image as well as between frames.

Although the rapid advancement of novel technologies like deep learning [3], that have really enhanced the goal of semantic segmentation, efficient real-time semantic segmentation remains a huge issue in current research. In this paper, we take advantages of deep learning techniques, Particularly

convolutional neural networks(CNNs), to develop an accurate semantic segmentation approach. The key contributions of this paper are:

- Utilizing the benefits of convolutional neural networks to build a precise U-Net-based model with additional squeeze and excitation blocks to perform an efficient semantic segmentation task for autonomous driving.
- The developed approach is trained and evaluated using the CamVid dataset,focused on semantic segmentation for self-driving missions.

The remainder of this paper is structured as follows: Section II presents the recent achievements in semantic segmentation. We systematically cover a brief overview of the existing methods. The proposed approach is described on Section III. Then, Section IV and V present the evaluation metrics and the experimental results on the CamVid dataset, respectively. Finally, the conclusions are presented in Section VI.

## II. Related Work

The rapid developments in deep-learning research have led to tremendous advancements in computer vision tasks [4] [5]. The major contribution to this accomplishment was the invention of CNNs, which significantly improved accuracy and speed for perception tasks such as object detection and recognition.

Autonomous vehicles ought to be fully aware and should have a thorough understanding of their surroundings. Using a set of item categories, such as pedestrians, cars, buildings, etc., semantic segmentation categorizes all images at the pixel level. For instance, by examining the findings of semantic segmentation, vehicles will be able to identify the navigable area and the surrounding obstacles. Additionally to CNNs, Auto-Encoders were employed to create semantic segmentation algorithms which are significantly more effective than previous models. Several approaches for semantic segmentation techniques have been introduced in recent years. These methods may be classified based on their key contributions. Fully convolutional networks (FCN) [6], encoder-decoder-based approaches such as Segnet [7], ERFNet [20] and U-Net [8], and ESPnetv2 [9] are a few examples of these types. Since images contain a range of semantic information, it is crucial to develop simple segmentation models without sacrificing accuracy since a huge number of trainable parameters are required to fully represent the complexity of potential images. Recent studies have focused on convolutional auto-encoders(CAEs), which are auto-encoders that contain convolutional and deconvolutional layers as their encoder and decoder components. The backbone of CAEs conceived for semantic segmentation were thus based on CNNs that were originally established for object detection and recognition. Semantic segmentation was performed through FCN [6] using a fully convolutional framework with a huge number of parameters. It was a pioneering attempt to discard fully connected layers.

The VGG architecture served as the backbone for both SegNet [7] and SegNet-Basic [10] architectures. For the up-sampling process in the decoder, it exploited the pooling indexes of the encoder. To improve segmentation accuracy, certain other systems, like U-Net [8], employed some sort of skip connections between the encoder and the decoder as well as other approaches, such as data augmentation. Real-time semantic segmentation remains a growing field, especially since some sectors, like autonomous driving and robotics, involve extremely reliable semantic segmentation with a minimum amount of processing time. The models listed above and some additional models, such Hyperseg [11], Dilated [12], and DeepLab [13], increased the efficiency of state-of-the-art techniques. Some models, like FPN, were designed with lower computational complexity. The encoder architecture employed in the original FPN model has a structure that is similar to ResNets [14], which might pose issues when it is generalized to operate in real-time scenarios even though it is effective in the semantic segmentation missions. With fewer parameters, super-lighter models like ESCNet [15], ApesNet [16], Enet [17], ERFNet [20] and ESPNet [9] attempted to provide real-time semantic segmentation. These models provide workable solutions to meet the real-time need, but critical applications like road scene interpretation in autonomous cars require substantially higher segmentation accuracy.

## III. Proposed Architecture

In order to develop an efficient semantic segmentation model for self-driving vehicles, two factors must be taken into consideration. It must first be able to run under real time situation. Second, it should be accurate enough for the autonomous car to rely on the findings in order to interpret its surroundings. Fig 1 shows the backbone of our proposed network based on the U-Net architecture.

Encoder and decoder sub-networks constitute its two sub-networks. The decoder sub-network up-samples the features through using transposed convolution corresponding to each down-sample Stage in the encoder sub-net. The obtained feature maps from the encoder that have the same resolution are fused with the up-sampled features. After each series of consecutive $3 \times 3$ convolution layers with Leaky Rectified Linear Unit activation functions, the encoder sub-network performs a maximum pooling operation. The size of the filters used in each layer distinguishes each layer in the sub-network. Fig 1 depicts each convolution unit and the layer it is connected with. The number of filters employed in the next layer is doubled with each down-sampling step. Four feature maps with different sizes are produced as a result of repeating this cycle. The number of filters employed in each level of the decoder sub-network is reduced by half before being scaled up by a $2 \times 2$ transposed convolution. Then, again a sequence of two $3 \times 3$ convolution followed by Leaky ReLU operations is performed. The final segmentation mask is

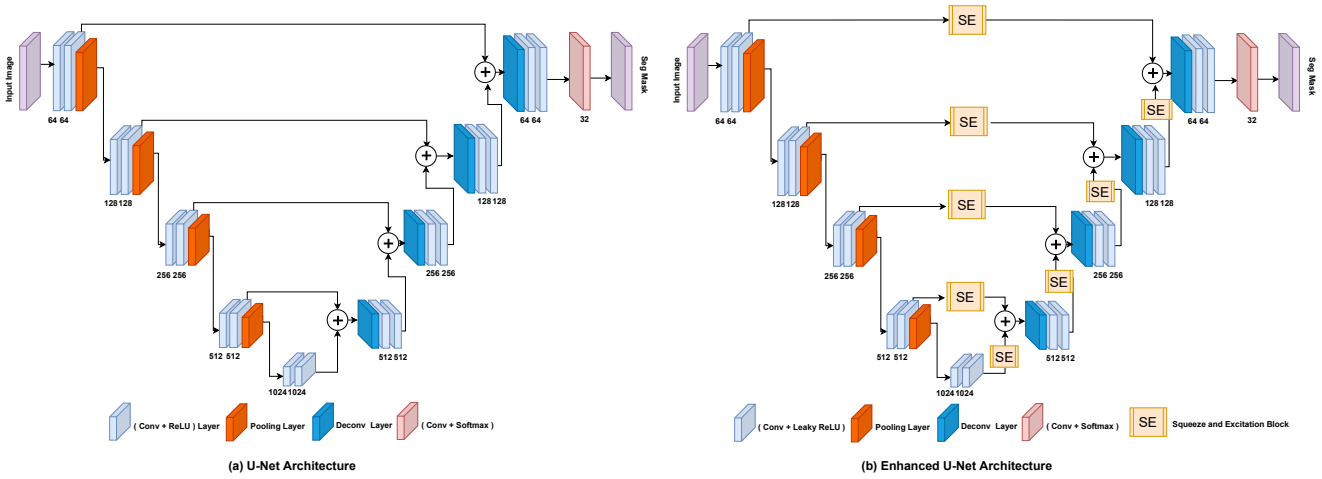



Fig. 1: Proposed Architecture based on U-NET Model

produced using a $1 \times 1$ convolution operation with a Softmax layer after this process has been repeated four times. In order to tackle the shortcomings of U-Net in terms of accuracy, we suggest a reformed skip-connected architecture that employ squeeze and excitation blocks [18].The utilized module present an architectural component that enables dynamic channel-wise feature recalibration and is introduced to improve the representational power of a network.

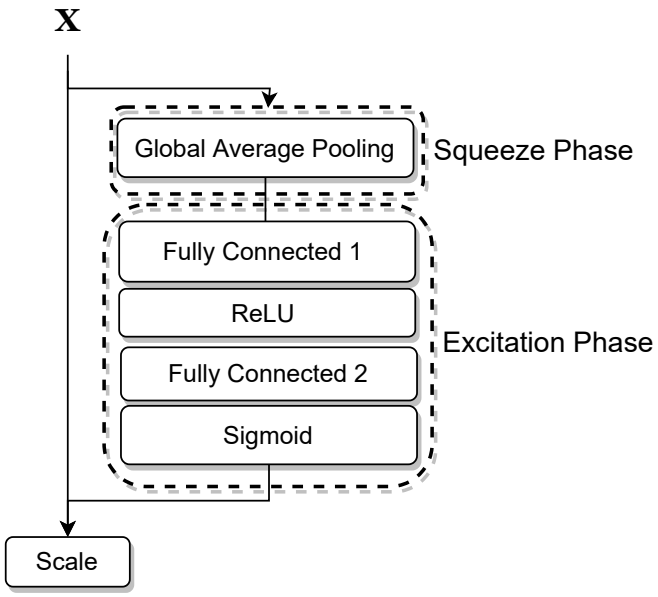



Fig. 2: Illustration of the squeeze and excitation Module

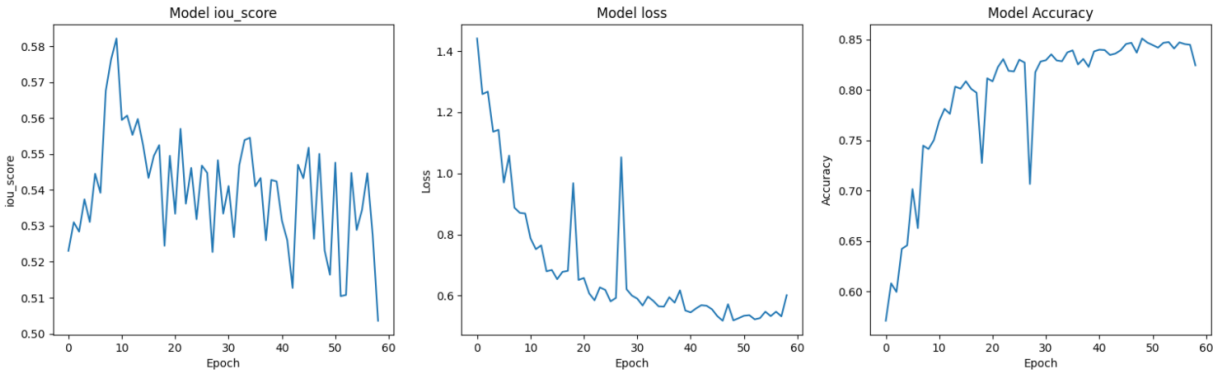
This block may be placed into a convolutional neural network to enhance channel inter-dependencies across various feature channels. Squeeze-and-excitation blocks incorporate a type of self-attention on channels and accurately represent channel connections and inter-dependencies. These additional modules are based on two key principles:

- Channel-inter-dependencies modeling inside modules explicitly.
- Feature recalibration: Suppress ineffective features and precisely accentuate those that are useful.

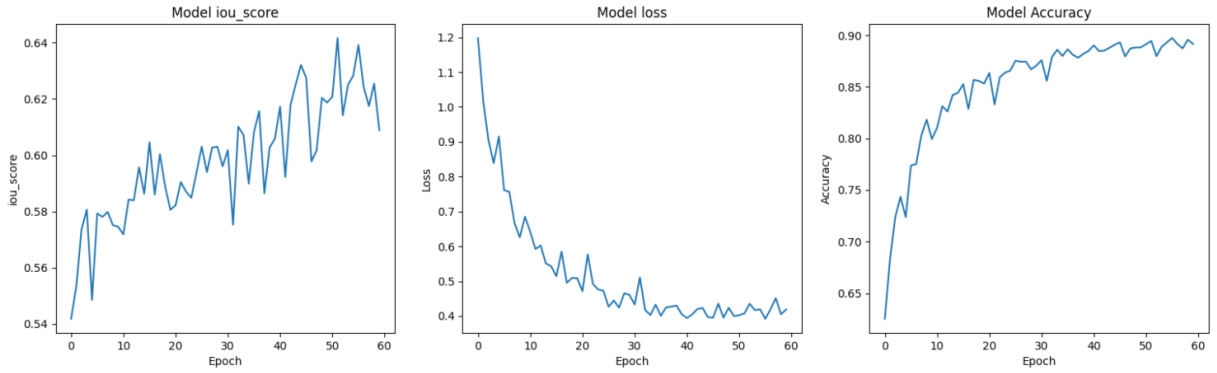
As illustrated in Fig 2, the squeeze phase in the SE block consists on squeezing global spatial features into a channel descriptor. In order to create channel-wise statistics, this phase involves a global average pooling over the spatial dimensions. Channel-wise dependencies are fully captured by the excitation process. In order to rescale the feature maps, the excitation step transforms the squeeze operation's output into a vector of activations.

## IV. Dataset and Evaluation Metrics

Cambridge-driving Labeled Video Database (CamVid) [19] is an image dataset comprising road scenes. It was first captured as a video with five different scenes. It consists of images with a resolution of $960 \times 720$, which are organized into 32 sets. Some of the most important categories to understand Street View are children, signs, motorcycles, lanes, traffic lights and traffic cones, buses, cars and street markings. The CamVid dataset was used to train the model, and this was followed by data augmentation. The model was trained using randomly generated $512 \times 512$ image cropping. With a batch size of 32 and an initial learning rate of $1e^{-4}$, the model is trained over 60 epochs using the Adam as the optimizer. The CamVid test dataset is utilized to test the given model, and it is then evaluated utilizing variety of evaluation metrics. We must consider the issue of class imbalance while selecting the evaluation metrics. Therefore, even while class weighting during the training process helps to minimize the effects of class imbalance, the developed model often has a tendency




(a) Evaluation Results of U-Net on the Camvid dataset




(b) Evaluation Results of Enhanced U-Net on the Camvid dataset

Fig. 3: Evaluation Results

to perform better on classes with higher frequency. As a consequence, we use metrics that perform some sort of average calculation between the findings of the evaluation of each class independently. Additionally, a different metric for estimating model complexity is employed. For the current work, the following metrics are employed :

- Mean class accuracy (mCA): The prediction accuracy in semantic segmentation is defined as the proportion of properly identified pixels to all pixels. We adhere to the accepted practice of taking the average accuracy amongst all estimated accuracies of the defined classes in order to prevent class imbalance from producing false accuracy results. The obtained value is called as mCA.
- Mean intersection over union (mIoU): This criteria is commonly utilized for classification tasks. In tasks involving semantic segmentation, the intersection over union ratio measures the proportion of pixels with labels from both the prediction and the ground truth frames to the proportion of pixels with labels in either frame separately. The mIoU is calculated by using the average of all categories after the IoU for each segment class has been individually established.

## V. Experimental Results

### A. Experimental Condition

For the implementation of the chosen approach, we use the deep learning framework Tensorflow and the Nvidia acceleration libraries to perform the learning process and the testing phase on a development environment equipped with an Intel i7 CPU and Nvidia GeForce RTX 2070. We utilized the U-Net model as a baseline to assess the effectiveness of our suggested network and to demonstrate the functionality of the additional squeeze and excitation modules. Following the common settings, we train both of U-Net and our proposed model on the CamVid dataset which was randomly divided into three main partitions for train, test and validation purposes. The model takes 40 hours for a number of 60 epochs. The CamVid testing dataset is used as input after training, and the model is then assessed using evaluation metrics.

### B. Results and Discussion

As presented in Fig 3, the proposed model proves significant enhancement in terms of class accuracy, loss and intersection over union compared to the original U-Net.

As shown in Table I, the mCA and the mIoU are increased by 5.54% and 5.87% respectively. The Comparison with other

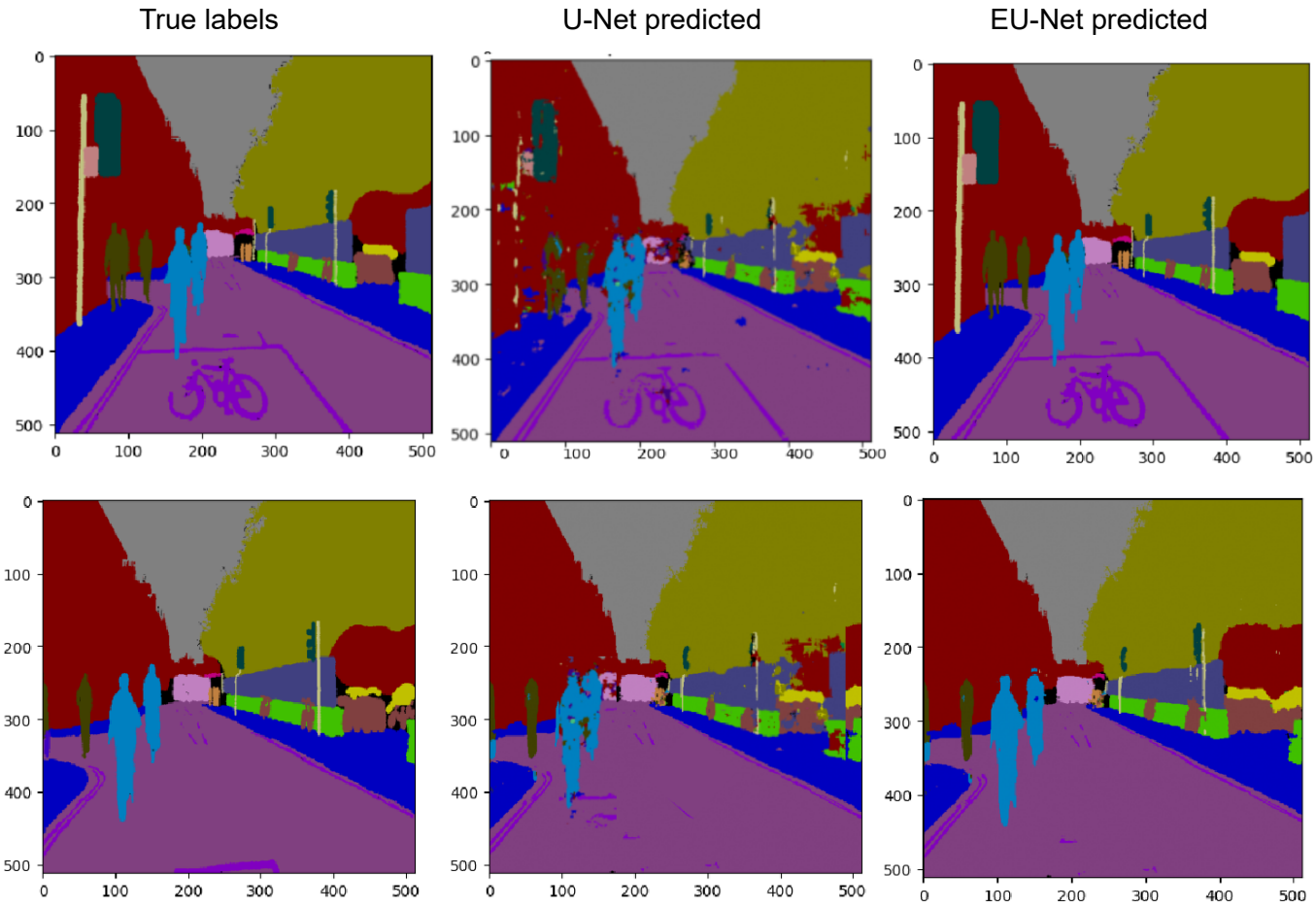



Fig. 4: Qualitative results for Testing Images : Tested images from the CamVid dataset with U-Net vs Enhanced U-Net.

TABLE I: Comparison with state of the art models

| Model | mCA (%) | IoU (%) |
|---|---|---|
| ApesNet [16] | 69.3 | 48 |
| ENet [17] | 68.3 | 51.3 |
| **U-Net** [8] | **78.61** | **54.01** |
| SegNet [7] | 65.2 | 55.6 |
| ESCNet [15] | 70.9 | 56.1 |
| ESPNet [9] | 68.3 | 67.7 |
| **Enhanced U-NET** | **84.15** | **59.88** |

state of the art models is then reported in Table I. Among the baselines mentioned above, the proposed model achieves a better mean class accuracy of 84.15% and a mean intersection over union of 59.88%. Fig 4 depicts segmentation results of the original U-Net and the suggested model on the CamVid dataset. The left column shows the True labels, the central and the right columns represents the predicted mask with U-Net and Enhanced U-Net respectively. The level of complexity is critically important since our system will perform in an outdoor environment and is going to be employed for real-time perception. Therefore as result, the more complicated a model is, the longer it will take to calculate the output. This makes its use in real-time applications less convenient.

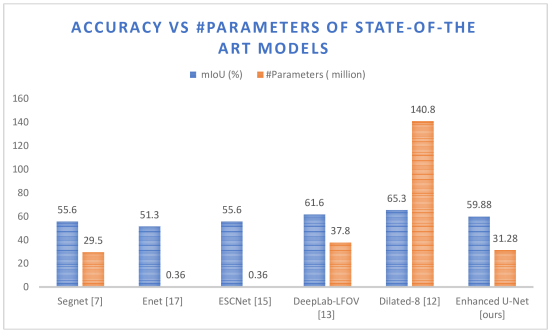



Fig. 5: Accuracy vs #parameters of State of the art algorithms for semantic segmentation.

Fig 5 compares the accuracy and complexity of our devel-

TABLE II: Comparison of number of parameters: U-Net vs Enhanced U-Net

| Model | Total params |
|---|---|
| U-NEt [16] | 31.197.600 |
| Enhanced U-Net | 31.286.680 |

oped model with those of other state-of-the-art algorithms. Enhanced U-Net outperforms the lightweight approaches such as ENet [17] and ESPNet [9] (0.36 million parameters). Although certain models, such as DeepLab-LFOV [13] and Dilated-8 [12], would reach a greater mIoU or mCA, these are less suitable for real-time use due to the large number of parameters. As we can clearly see, our proposed approach provides promising performance with a relatively low set of parameters. In addition, it achieves a better accuracy then U-Net, without a significant increase in the computational cost as presented in Table II.

## VI. Conclusion

Semantic segmentation is a fundamental task for the perception component of self-driving vehicles. It is crucial to understand how the ego vehicle will operate in a road scenario. As deep learning techniques have advanced over the recent years, a growing amount of research is increasingly relying on how to use deep learning to enhance perception, among other autonomous driving modules. In order to achieve the accuracy of real-time considerations, we proposed a deep learning method that takes advantages of both U-Net and squeeze and excitation architectures. The CamVid dataset was used to train the suggested model, which outperformed other cutting-edge algorithms in terms of accuracy and efficiency. Nevertheless, there is still an enormous gap to achieve a better trade-off between accuracy and inference speed while preserving small size to allow these models to be applied in real-time systems.
Although the domain is very active, there is still much room for further development.

## References


[1] Janai, J.; Güney, F.; Behl, A.; Geiger, A. Computer Vision for Autonomous Vehicles: Problems, Datasets and State of the Art. Found. Trends Comput. Graph. Vis. 2020, 12, 85.

[2] W. Zhou, S. Lv, Q. Jiang, and L. Yu, Deep road scene understanding, IEEE Signal Processing Letters, vol. 26, no. 4, pp. 587591, 2019.

[3] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT press, Cambridge, MA, USA, 2016.

[4] Yang, Chen, et al. UL-CNN: An Ultra-Lightweight Convolutional Neural Network Aiming at Flash-Based Computing-In-Memory Architecture for Pedestrian Recognition, Journal of Circuits, Systems and Computers 30.02, 2021.

[5] Takos, G. A Survey on Deep Learning Methods for Semantic Image Segmentation in Real-Time. arXiv 2020, arXiv:abs/2009.12942.

[6] ] J. Long, E. Shelhamer, and T. Darrell, Fully convolutional networks for semantic segmentation, in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 34313440, Boston, MA, USA, 2015.

[7] ] V. Badrinarayanan, A. Kendall, and R. Cipolla, SegNet: a deep convolutional encoder-decoder architecture for image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 12, pp. 24812495, 2017.

[8] SIDDIQUE, Nahian, PAHEDING, Sidike, ELKIN, Colin P., et al. U-net and its variants for medical image segmentation: A review of theory and applications. IEEE Access, 2021, vol. 9, p. 82031-82057.

[9] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, Espnet: efficient spatial pyramid of dilated convolutions for semantic segmentation, in Proceedings of the European Conference on Computer Vision (ECCV), pp. 552568, Munich, Germany, 2018.

[10] V. Badrinarayanan, A. Handa, and R. Cipolla, Segnet: a deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling, 2015, https://arxiv.org/abs/1505.07293.

[11] Nirkin, Y.; Wolf, L.; Hassner, T. HyperSeg: Patch-wise Hypernetwork for Real-time Semantic Segmentation. arXiv 2020, arXiv:abs/2012.11582.

[12] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, Pyramid scene parsing network, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 28812890, Honolulu, HI, USA, 2017.

[13] Liu, C.; Chen, L.C.; Schroff, F.; Adam, H.; Hua, W.; Yuille, A.; Li, F.-F. Auto-DeepLab: Hierarchical Neural Architecture Search for Semantic Image Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 1620 June 2019; pp. 8292.

[14] HE, Kaiming, ZHANG, Xiangyu, REN, Shaoqing, et al. Deep residual learning for image recognition. In : Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 770-778.

[15] J. Kim and Y. S. Heo, Efficient semantic segmentation using spatio-channel dilated convolutions, IEEE Access, vol. 7, pp. 154239154252, 2019.

[16] C. Wu, H. P. Cheng, S. Li, H. Li, and Y. Chen, ApesNet: a pixel-wise efficient segmentation network for embedded devices, IET Cyber-Physical Systems: Theory & Applications, vol. 1, no. 1, pp. 7885, 2016.

[17] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, Enet: a deep neural network architecture for real-time semantic segmentation, 2016, https://arxiv.org/abs/1606.02147.

[18] HU, Jie, SHEN, Li, et SUN, Gang. Squeeze-and-excitation networks. In : Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 7132-7141.

[19] G. J. Brostow, J. Fauqueur, and R. Cipolla, Semantic object classes in video: a high-definition ground truth database, Pattern Recognition Letters, vol. 30, no. 2, pp. 8897, 2009.

[20] E. Romera, J.M. Alvarez, L.M. Bergasa and R. Arroyo, "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation", IEEE Transactions on Intelligent Transportation Systems, vol. 19, no. 1, pp. 263-272, 2017.