I. J. Engineering and Manufacturing, **2023, 6, 53-61**
Published Online on December 8, 2023 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijem.2023.06.05

Advancing Road Scene Semantic Segmentation with UNet-EfficientNetb7

Anagha K J*
Centre for Artificial Intelligence, TKM College of Engineering, Kollam, Kerala, India
Email: anaghakj1999@gmail.com
ORCID ID: http://orcid.org/0009-0004-7214-8810
*Corresponding Author

Sabeena Beevi K
Department of Electrical and Electronics Engineering, TKM College of Engineering, Kollam, Kerala, India
Email: sabeena3000@tkmce.ac.in
ORCID ID: http://orcid.org/0000-0002-5129-9386

Received: 03 June, 2023; Revised: 03 August, 2023; Accepted: 29 August, 2023; Published: 08 December, 2023

Abstract: Semantic segmentation is an essential tool for autonomous vehicles to comprehend their surroundings. Due to the need for both effectiveness and efficiency, semantic segmentation for autonomous driving is a difficult task. Present-day models' appealing performances typically come at the cost of extensive computations, which are unacceptable for self-driving vehicles. Deep learning has recently demonstrated significant performance improvements in terms of accuracy. Hence, this work compares U-Net architectures such as UNet-VGG19, UNet-ResNet101, and UNet-EfficientNetb7, combining the effectiveness of compound-scaled VGG19, ResNet101, and EfficientNetb7 as the encoders for feature extraction. And, U-Net decoder is used for regenerating the fine-grained segmentation map. Combining both low-level spatial information and high-level feature information allows for precise segmentation. Our research involves extensive experimentation on diverse datasets, including the CamVid (Cambridge-driving Labeled Video Database) and Cityscapes (a comprehensive road scene understanding dataset). By implementing the UNet-EfficientNetb7 architecture, we achieved notable mean Intersection over Union (mIoU) values of 0.8128 and 0.8659 for the CamVid and Cityscapes datasets, respectively. These results outshine alternative contemporary techniques, underscoring the superior precision and effectiveness of the UNet-EfficientNetb7 model. This study contributes to the field by addressing the crucial challenge of efficient yet accurate semantic segmentation for autonomous driving, offering insights into a model that effectively balances performance and computational demands.

Index Terms: Semantic segmentation, Autonomous driving, UNet, road scenes, VGG19, ResNet101, EfficientNetb7

1. Introduction

Semantic segmentation, sometimes referred to as pixel level classification, is the process of giving each pixel of an image a pre-determined label or class. It has several applications, including robotics, self-driving cars, and medical imaging. Semantic segmentation is an essential component of intelligent vehicles since they need to understand and be in context with their surroundings to be safely integrated on today's roadways. Research towards creating autonomous vehicles began in 1989, but it was hampered by the constraints of standard neural networks and hardware resources [1].

The advancement of intelligent vehicles has been hastened by the progress in convolutional neural networks (CNN) and GPU technology. Despite ongoing research efforts, the subject remains challenging due to significant geographical disparities. However, there has been recent progress in semantic scene segmentation and it has become a widely studied area within deep learning.

Before the development of CNN-based systems, semantic image segmentation methods relied on manually created features and conventional classifiers. It is difficult for machines to comprehend complicated real-world scenes and function at a human level [2]. However, because CNN's have demonstrated their efficacy in image classification, feature extraction in a semantic segmentation framework uses them as the foundation. The general layout of the CNN-based encoder-decoder type semantic segmentation framework is shown in Fig.1. Convolutional neural networks gradually decrease the input image resolution to create the high-level feature map representing the original image. Such a tiny feature map is excellent for classifying images in which there is just one dominant object, and CNN's have

outperformed humans in this area of image classification. However, CNN's performance delays regarding segmentation since the tiny feature maps lose the spatial information necessary for analyzing complicated scenes in the image. Thus, this work compares various models to combine the effectiveness of VGG [3], ResNet [4], and EfficientNet [5] as an encoder for high-level feature extraction, along with the UNet [6] decoder for generating fine segmentation maps. Although U-Net models are commonly used for medical imaging, they are not frequently utilized for road scene segmentation.

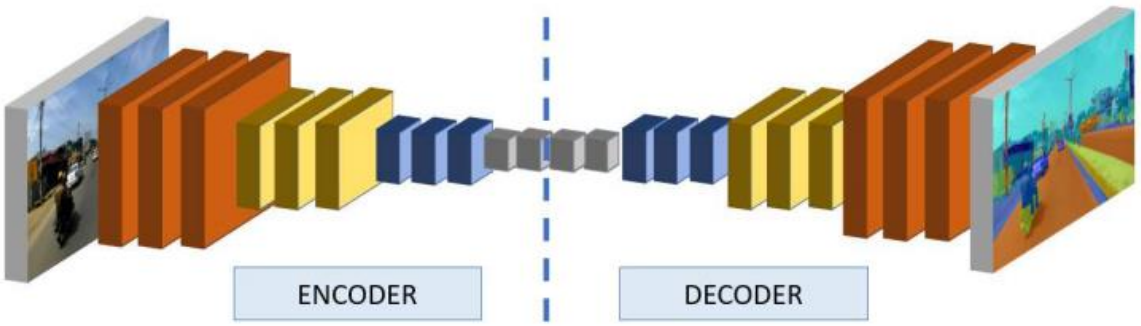

Fig. 1. Encoder-Decoder type semantic segmentation architecture

Cityscapes [7], KITTI [8], and CamVid [9] are examples of typical datasets for assessing semantic segmentation architectures. These datasets are primarily concerned with scene understanding in metropolitan street scenes, such as those found in the USA or Europe, where the road system and surrounding environment are well organised. They have well-marked lanes, a low participant-to-traffic density, fewer alterations in the objects and background, and tight adherence to traffic regulations. This work mainly works on the CamVid and Cityscapes dataset, which focuses primarily on scene understanding in a structured environment.

The proposed method integrates spatial information at a low level and feature information at a high level for segmentation and surpasses current state-of-the-art techniques in road scene semantic segmentation. The U-Net decoder is employed to generate a detailed segmentation map, while compound-scaled VGG19, ResNet101, and EfficientNetb7 serve as the encoders for feature extraction. Although U-Net models are widely used in medical imaging, they are not as common in road scene segmentation. However, when utilized in this study, they lead to improved segmentation maps to aid self-driving cars.

In this paper, we aim to achieve an innovative approach that strikes a balance between efficiency and accuracy in semantic segmentation for autonomous driving scenarios. By integrating state-of-the-art encoder architectures with the UNet decoder, the goal is to improve the precision of road scene segmentation and advance the capabilities of self-driving vehicles in comprehending their environments. This research also aims to shed light on the viability of U-Net models, which have found success in medical imaging, within the domain of road scene segmentation, thus contributing to the enhancement of segmentation techniques for complex real-world scenarios.

The upcoming sections of the paper are arranged in the subsequent manner: Section 2 offers an overview of pertinent literature, Section 3 explains the methodology utilized, Section 4 illustrates the findings, and Section 5 summarizes the conclusions and future prospects.

2. Related Works

The first publication of CNN in the area of semantic segmentation was FCN [10]. For denser predictions, the complete image was given into the FCN-based approach. Deconvolutional networks and unpooling layers were first suggested in [11]. Badrinarayanan et al. suggested the encoder-decoder type module SegNet [12], which uses VGG19 as the feature extractor, to address the issue with coarse segmentation outputs. Since 2012, many designs based on convolutional neural networks, including VGG19 [3], ResNet [4], and the recently released EfficientNet [5], have been developing and setting standards for image categorization.

The EfficientNet, as it was described in [5], is made up of a compound coefficient that investigates model scaling and modifies the network's depth, width, and resolution for improved performance. The usage of these CNNs has greatly facilitated recent advancements in the field of semantic segmentation as feature extractors. The VGG19 [3] based architecture outperformed conventional approaches significantly; however, the coarse output pixel map limited the pixel accuracy. In [13], a multiscale semantic segmentation technique was proposed in light of the emergence of robust CNN architectures. The feature maps from various resolutions were combined using a skip-net architecture with end-to-end learning. Since CNNs gradually reduce the initial image resolution, the performance is hampered by the loss of spatial information of small and thin objects. The idea of dilated convolution was first proposed in [11,13] to improve the feature map's resolution while maintaining the neuron's receptive field. Dilated residual networks were suggested by Yu et al., which solved the issue of gridding artifacts [14].

DeepLabV1 [15] and DeepLabV2 [16] employed fully connected Conditional Random fields (CRF), dilated convolutions, and state-of-the-art CNNs as feature extractors. DeepLabV3 [17] introduced the idea of Atros Spatial Pyramid Pooling (ASPP). With the introduction of an efficient decoder module in DeepLabV3+ [18], the outcomes of DeepLabV3 were considerably enhanced. ERFNet (Efficient Residual Factorized Network), which makes use of factorized convolution with residual connections, was proposed by Romera et al. [19]. ParseNet, which combines L2 normalization and global average pooling, is introduced in [20]. Zhao et al. proposed PSPNet, which employed a pyramid pooling module on the feature map of the last layer [21]. In real-time applications, segmentation models like ENet [22] and ICNet [23] are helpful. Ronneberger et al. [6] suggested a U-shaped, fully convolutional network architecture in which features maps from various encoder and decoder layers were upsampled and concatenated.

This paper focuses on the UNet-EfficientNetb7 model that combines the U-Net as the decoder and EfficientNetb7 as the encoder. This model yields improved segmentation maps despite geographical differences. The proposed model exhibits higher accuracy on the CamVid and Cityscapes dataset compared to recent studies.

3. Methodology

3.1 Techniques used

Semantic image segmentation divides an image into different regions or categories, each containing pixels with comparable characteristics and assigned to a particular category. UNET is a popular convolutional neural network variation for segmenting medical images. Olag Ronneberger et al. first proposed UNET in 2015 [6]. As shown in Fig.2., the basic UNET consists of two paths: a contracting path (encoder) and an expansive path (decoder). The input, convolutional, pooling, and fully connected layers make up the general CNN encoder component. The number of features increases by half with each downsampling when features are extracted from the images. To enable localization, the decoder component employs transposed convolution. It entails upsampling the feature map, resulting in a reduction in the number of feature channels. The convolutional layer was then used to map the channels to the appropriate classes after performing the concatenation of the respective feature maps of the encoder and decoder.

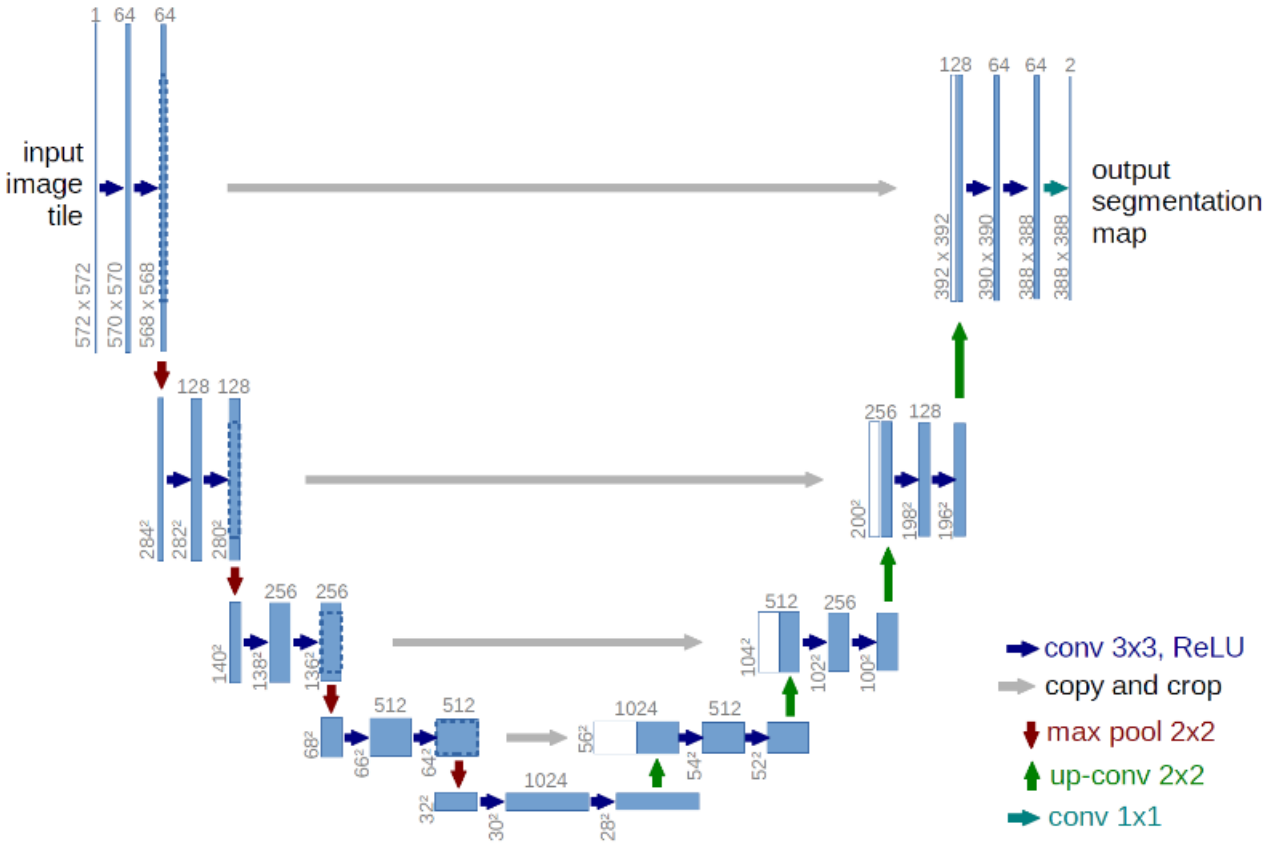

Fig. 2. UNet architecture

3.1.1 UNet-VGG19

The encoder and decoder are the basic building blocks of the framework, just like the standard UNET [6]. The decoder section uses the output of the encoder VGG19 to segment the image by upsampling the feature maps, while the encoder part employs standard VGG19 [3] to extract features from the images. The most often used CNN has 19 layers, called VGG19. 16 of the 19 layers are convolutional, 5 are MaxPool layers, three are entirely linked, and 1 SoftMax layer. The six-step procedure is followed in the construction of VGG19.

- The architecture receives the image as its first input, typically an image with the resolution (224, 224, 3)
- Then, a kernel of size (3, 3) was used to find the image's underlying patterns.
- Padding was utilised to maintain the image's resolution.
- To reduce the size of the image, pooling was used.
- Typically, layers produce linear output; consequently, a fully linked layer was used to convert linear output to non-linear output.
- To forecast the probability distribution of the numerous classes, the SoftMax layer is use

3.1.2 UNet-ResNet101

Encoder and decoder are the basic building blocks of the framework, just like the standard UNET [6]. The decoder section uses the output of the encoder ResNet101 to segment the image by upsampling the feature maps, while the encoder part employs standard ResNet101 to extract features from the images.

The first section, which comprises the down-sampling path, uses Resnet101 as its primary structural support. Each stage consists of a 2×2 max-pooling layer applied after two 3×3 convolutions with the batch norm. A 2×2 up-convolution follows two 3×3 convolutions to form the horizontal bottleneck. The decoder is shown as having two 3×3 convolutional layers in 4 stages, followed by 2×2 up-sampling in the up-sampling pipeline. At each stage, the feature maps are cut in half. The model skips links between up-sampling and down-sampling paths to give local and global information. The segmented output is finally provided by the output 1×1 convolutional layer, where the number of feature mappings is comparable to the number of desired segments.

3.1.3 UNet-EfficientNetb7

The encoder and decoder are the basic building blocks of the framework, just like the standard UNET [6]. The decoder section uses the output of the encoder EfficientNetb7 to segment the image by upsampling the feature maps, while the encoder part employs standard EfficientNetb7 to extract features from the images.

The development of CNN architectures depends on the available resources, and when those resources rise, scaling takes place to get better performance. The model has historically been scaled by arbitrarily increasing the CNN width, depth, or input image resolution. Despite the time-consuming manual tuning required, this approach occasionally produces subpar performance. To increase performance, Tan et al. introduced a novel compound scaling strategy that uniformly adjusts the network's depth, width, and resolution using a predetermined set of scaling factors [5]. EfficientNetB0, a new baseline architecture, was initially created, and then it was scaled up to produce the EfficientNet family using the compound scaling method. Eight EfficientNets versions, numbered EfficientNetB0 to EfficientNetB7, are supported by this methodology. By balancing the architecture's compound width, depth, and image resolution coefficients, scaling the network incrementally enhances model performance.

The mobile inverted bottleneck convolution (MBConv) [24] with a squeeze and excitation optimization is the fundamental component of the EfficientNet design. The number of these MBConv blocks varies depending on the EfficientNet network family. Depth, width, resolution, and model size all increase from EfficientNetB0 to EfficientNetB7, and accuracy improves [5]. On ImageNet, the best-performing model, EfficientNetB7, beats earlier state-of-the-art CNNs in terms of accuracy while being 8.4 times smaller and 6.1 times faster [5]. According to the filter size, striding, and number of channels, it can be separated into seven blocks. This work combined the EfficientNetB7 encoder with a UNet decoder to get the best performance.

3.2 Datasets used

3.2.1 CamVid

CamVid (Cambridge-driving Labeled Video Database) [9] offers 701 driving scenario photos divided into 367, 101, and 233 images for training, validation, and testing. There are 32 annotated categories, and the image resolution is 960 x 720. The predefined 32 classes are void, building, wall, tree, vegetation, fence, sidewalk, parking block, column/pole, traffic cone, bridge, sign, miscellaneous text, traffic light, sky, tunnel, archway, road, road shoulder, lane markings (driving), lane markings (non-driving), animal, pedestrian, child, cart luggage, train, car, motorcycle, truck/bus, bicyclist, pickup/SUV, and other moving objects.

3.2.2 Cityscapes

In the fields of semantic segmentation and autonomous driving, Cityscapes [7] is one of the most widely used datasets. It was originally recorded as a video, therefore the pictures are specifically chosen frames that were taken in 50 various cities. The decision was made based on the requirement for numerous items, a wide range of situations, and a wide range of backgrounds. 30 distinct classes are offered in total, divided into 8 categories. Around 5000 photos in the Cityscapes collection have fine annotation, which is divided into 2975, 500, and 1525 images for training, validation, and testing. The included urban street sceneries were taken during numerous spring, summer, and fall months during the daytime and in suitable weather.

3.3 Proposed framework

Fig.3. displays a detailed encoder-decoder block diagram for semantic segmentation. The encoder-decoder module typically consists of a CNN-based encoder that extracts the features from the original image.

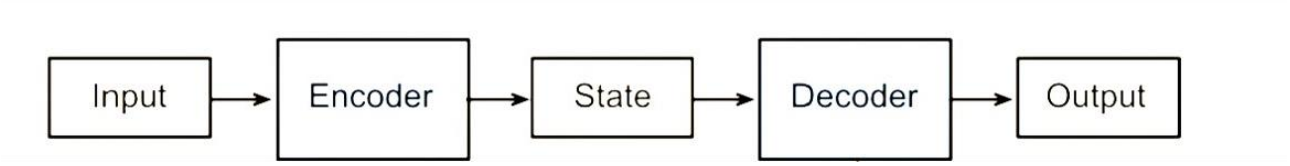

Fig.3. Proposed Framework

To extract high-level characteristics from the input image, the encoder is built from a succession of convolutional layers. In order to provide a more compact representation of the input image, these layers gradually increase the number of channels while decreasing the spatial resolution of the feature maps. The decoder, on the other hand, is in charge of restoring the spatial information lost during encoding and upsampling the compressed feature maps to the original resolution of the input image. Transpose convolutional layers (sometimes called upsampling layers or deconvolutional layers) are commonly used to do this. The network is programmed to make a classification prediction for each input image pixel. To do this, we add a classification head to the decoder, where each semantic class is represented by a convolutional layer that generates a softmax activation over the output channels. Dice loss function is used to optimise the network during training; these functions penalise the model if it makes a wrong prediction for a given pixel. To create a model that reliably predicts semantic labels for fresh input images, the optimisation process revises the network's parameters to minimise this loss function. This work uses VGG19, ResNet101, and EfficientNetb7 as the encoders for feature extraction, with a UNet decoder for regenerating the fine-grained segmentation map

3.4 Experimentation Setup

The experiments conducted in this study were centered on the CamVid and Cityscapes datasets. Three different models, namely UNet-VGG19, UNet-ResNet101, and UNet-EfficientNetb7, were employed for these experiments. The hardware utilized for these experiments consisted of an HPC (High-Performance Computing) machine running on Ubuntu 78.04.05 LTS. The computational power was enhanced with an NVIDIA Tesla V100 16G Passive GPU, providing the necessary processing capabilities.

The technical framework employed for implementing the models was Keras, which operated on the TensorFlow platform within a Jupyter notebook environment. To optimize the model training process, the RMSprop optimizer was chosen, paired with the dice loss function, which is particularly suited for segmentation tasks. This combination of optimizer and loss function is commonly used in semantic segmentation experiments.

The primary performance metric for evaluating the models' effectiveness was the mean Intersection over Union (mIoU). By analyzing the mIoU scores generated by each model, the study aimed to identify the most suitable and effective model for the specific segmentation task at hand.

4. Results

The proposed semantic segmentation architecture is developed in Tensorflow 2.0. The network is trained on an image size of 128x128, with a batch size of 4 and 100 epochs. Training is carried out using the RMSprop optimizer with a learning rate of 0.0001. The results are submitted to the online server and evaluated in terms of mean Intersection over Union (mIoU). MIoU (Mean Intersection over Union) is a common evaluation metric used in semantic segmentation to measure the performance of the segmentation model. It calculates the mean of the Intersection over Union (IoU) scores for all classes in the image. IoU is the ratio of the intersection of predicted and ground truth segments to their union. The higher the MIoU score, the better the performance of the segmentation model. Intersection over Union (IoU), also called as Jaccard index for one class, is computed as in (1):

$$IoU = \frac{TP}{TP+FP+FN} \quad (1)$$

where TP, FP, and FN mean the number of True Positive, False Positive, and False Negative pixels, respectively. From the results of mIoU, as shown in Table 1., it can be seen that the UNet-EfficientNetb7 model performs best on the CamVid dataset when compared to other models. The UNet-EfficientNetb7 model achieved a training mIoU of 90.88%, validation mIoU of 83.32%, and a testing mIoU of 81.28%. Table 2. shows the semantic segmentation quantitative results on Cityscapes dataset, the UNet-EfficientNetb7 achieved a training mIoU of 93.13%, validation mIoU of 87.48%, and a testing mIoU of 86.59%. From the above two tables, it is evident that the UNet-EfficientNetb7 model works better than the current best available models in semantic segmentation.

Table 1. mIoU on CamVid Dataset

Model	Training mIoU (%)	Validation mIoU (%)	Test mIoU (%)
UNet-VGG19	90.25	82.24	79.54
UNet-ResNet101	92.11	82.86	79.35
UNet-EfficientNetb7	92.11	83.32	81.28

Table 2. mIoU on Cityscapes Dataset

Model	Training mIoU (%)	Validation mIoU (%)	Test mIoU (%)
UNet-VGG19	92.95	86.34	85.49
UNet-ResNet101	93.51	86.86	84.23
UNet-EfficientNetb7	93.13	87.48	86.59

The loss graphs of the three models – UNet-VGG19, UNet-ResNet101, and UNet-EfficientNetb7 are shown in Fig.4. The blue line indicates training loss and the red line indicates test loss. After each epoch, both training and test loss are seen to decrease for all three models. Fig.5. shows the semantic segmentation results of UNet-VGG19, UNet-ResNet101, and UNet-EfficientNetb7 on CamVid respectively. The first column displays the sample input images from the CamVid dataset representing various scenarios. The predicted segmentation map of the UNet-VGG19, UNet-ResNet101, and UNet-EfficientNetb7 models are displayed in the second, third, and fourth columns, respectively.

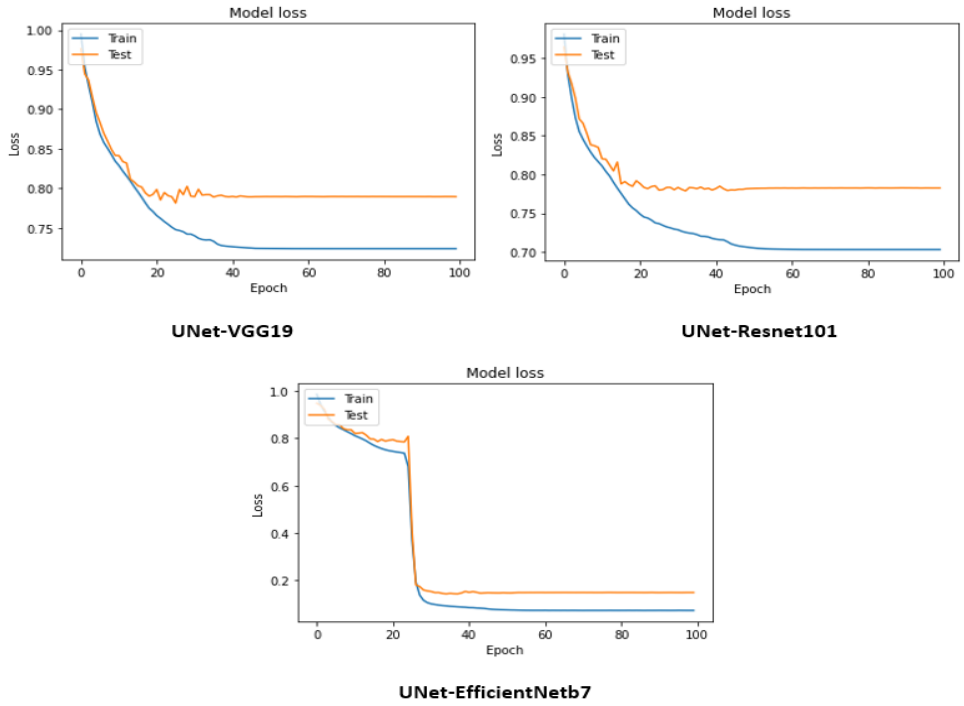

Fig.4. Loss graph of various UNet-Model

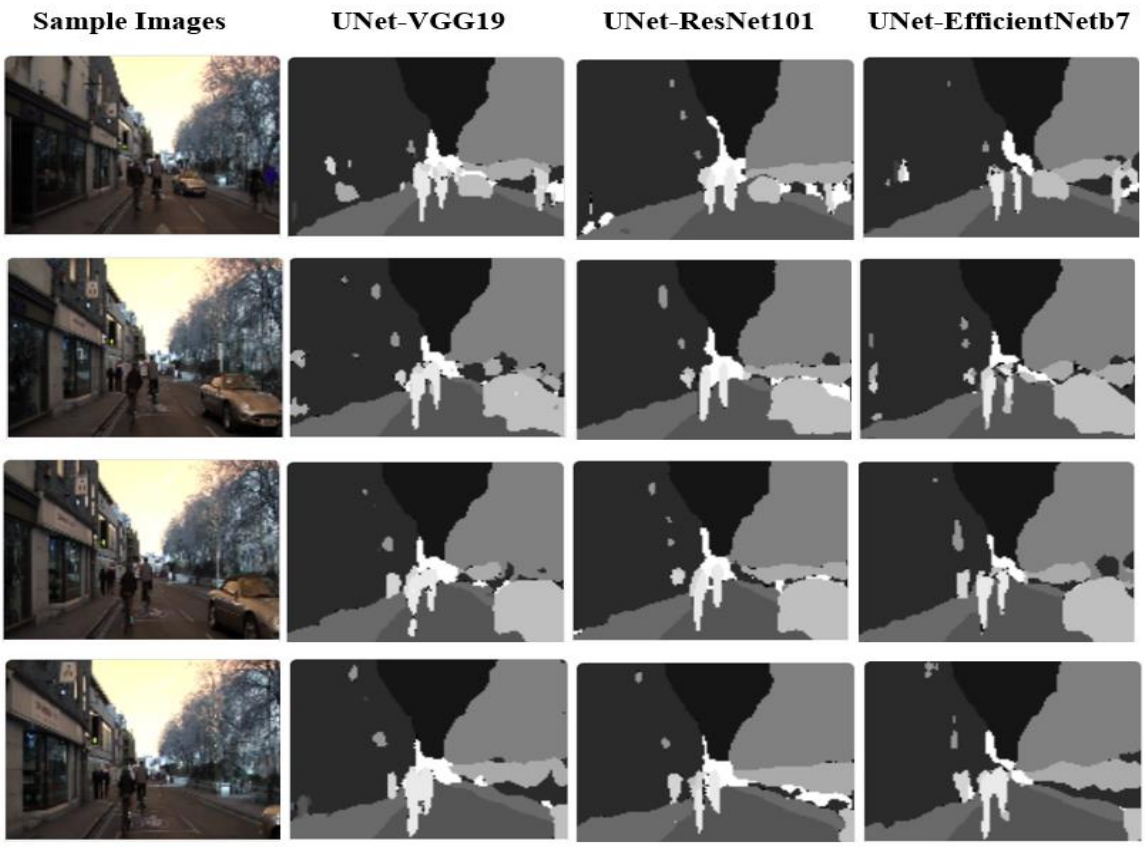

Fig.5. Results of semantic segmentation models on CamVid dataset

The results of the semantic segmentation using UNet-VGG19, UNet-ResNet101, and UNet-EfficientNetb7 on the Cityscapes dataset are shown in turn in Fig.6. Images from the Cityscapes dataset representing various settings are included in the first column as sample input images. The second, third, and fourth columns, respectively, show the segmentation images predicted by the UNet-VGG19, UNet-ResNet101, and UNet-EfficientNetb7 models.

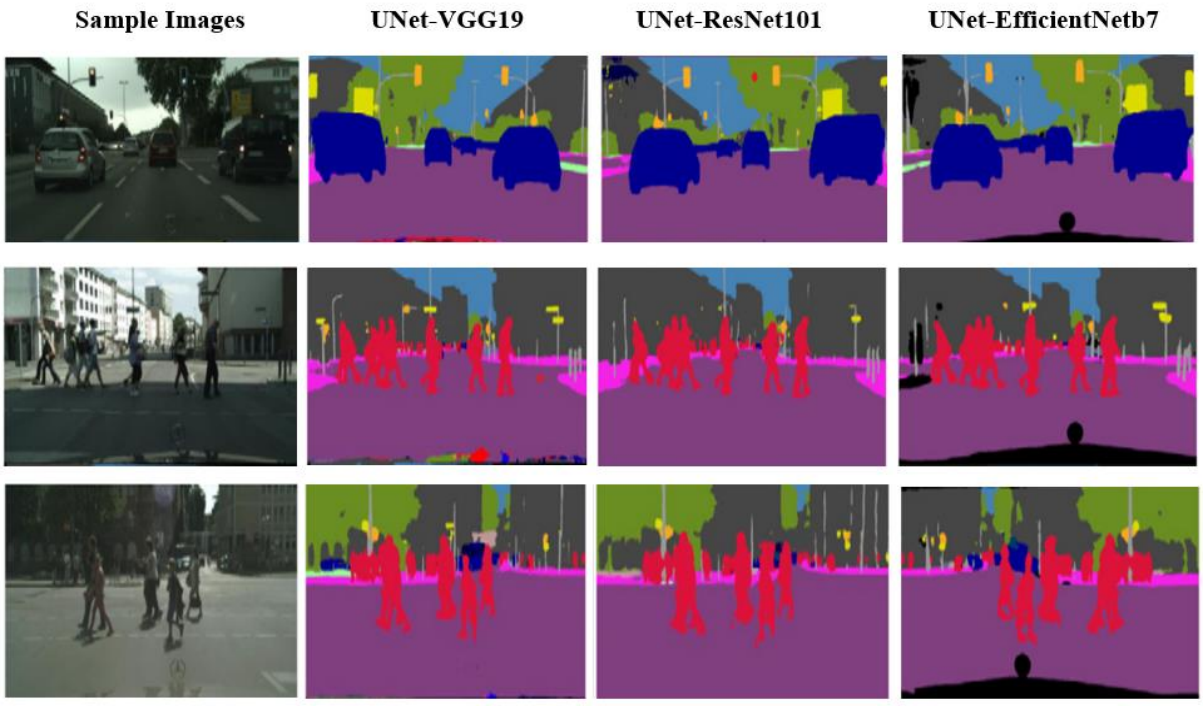

Fig.6. Results of semantic segmentation models on Cityscapes dataset

The results indicate that our approach effectively captures both spatial and temporal contextual information, resulting in improved accuracy without incurring a significant increase in computational cost. This outperforms the current state-of-the-art techniques with lower latency.

5. Conclusion and Future Work

Developing an advanced semantic image segmentation solution to achieve comprehensive scene understanding in intelligent transportation systems presents a complex challenge. This study focuses on the comparison of diverse models for pixel-level segmentation, specifically on the CamVid and Cityscapes datasets. The research employs VGG19, ResNet101, and EfficientNetb7 as encoders, combined with a UNet decoder that integrates both low-level spatial information and high-level features to achieve precise segmentation.

The CamVid dataset serves as a compact driving/road scene understanding dataset, while the Cityscapes dataset offers a large-scale collection of diverse street scenes. The models developed in this study achieve notable mIoU scores—reaching 83% on the CamVid dataset and 86% on the Cityscapes dataset for the UNet-EfficientNetb7 model. These results distinctly demonstrate the superiority of the UNet-EfficientNetb7 model, surpassing existing state-of-the-art models in terms of both effectiveness and accuracy.

As part of future research, these UNet-based models will be subjected to evaluation on larger and more diverse datasets. This evaluation aims to determine whether the models maintain or even enhance their performance on datasets featuring high-resolution images. By extending the assessment to broader datasets, the study seeks to ensure the reliability and adaptability of the proposed models for a wider range of real-world scenarios and challenges in semantic segmentation.

References

[1] Baheti, B., Innani, S., Gajre, S., Talbar, S. (2020). Eff-unet: A novel architecture for semantic segmentation in unstructured environment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 358-359).

[2] Baheti, B., Gajre, S., Talbar, S. (2019, October). Semantic scene understanding in unstructured environment with deep convolutional neural network. In TENCON 2019-2019 IEEE Region 10 Conference (TENCON) (pp. 790-795). IEEE.

[3] Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556..

[4] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[5] Tan, M., Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114). PMLR.

[6] Ronneberger, O., Fischer, P., Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234- 241). Springer, Cham.

[7] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., ... Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3213-3223).

[8] Menze, M., Geiger, A. (2015). Object scene flow for autonomous vehicles. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3061-3070).

[9] Brostow, G. J., Fauqueur, J., Cipolla, R. (2009). Semantic object classes in video: A high-definition ground truth database. Pattern Recognition Letters, 30(2), 88-97.

[10] Long, J., Shelhamer, E., Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

[11] Noh, H., Hong, S., Han, B. (2015). Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE international conference on computer vision (pp. 1520- 1528).

[12] Badrinarayanan, V., Kendall, A., Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence, 39(12), 2481-2495.

[13] Yu, F., Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122.

[14] Yu, F., Koltun, V., Funkhouser, T. (2017). Dilated residual networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 472-480).

[15] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L. (2014). Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062.

[16] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4), 834-848.

[17] Chen, L. C., Papandreou, G., Schroff, F., Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.

[18] Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV) (pp. 801- 818

[19] Romera, E., Alvarez, J. M., Bergasa, L. M., Arroyo, R. (2017). Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. IEEE Transactions on Intelligent Transportation Systems, 19(1), 263-272.

[20] Liu, W., Rabinovich, A., Berg, A. C. (2015). Parsenet: Looking wider to see better. arXiv preprint arXiv:1506.04579.

[21] Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J. (2017). Pyramid scene parsing network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2881-2890).
[22] Paszke, A., Chaurasia, A., Kim, S., Culurciello, E. (2016). Enet: A deep neural network architecture for real-time semantic segmentation. arXiv preprint arXiv:1606.02147.
[23] Zhao, H., Qi, X., Shen, X., Shi, J., Jia, J. (2018). Icnet for real-time semantic segmentation on high-resolution images. In Proceedings of the European conference on computer vision (ECCV) (pp. 405-420).
[24] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4510-4520).

Authors' Profiles

Anagha K J is a fresh postgraduate student in Artificial Intelligence, at TKM College of Engineering, Kollam, India. Her research interest is in machine learning, neural networks, natural language processing, computer vision, robotics, and data science.

Dr. Sabeena Beevi K is the Head of the Department of Electrical and Electronics Engineering, T K M College of Engineering, Kollam, Kerala, India. She received the Ph.D. degree from the University of Kerala, in 2018 in Electrical and Electronics Engineering and her M.Tech. Degree in computer science with specialization in Digital Image Computing in 2009 from the Computer Science Department of Kerala University. In June 1998, she joined the Electrical & Electronics Engineering Department at T K M College of Engineering, where she is currently working as an Associate Professor. Her research interests include pattern recognition, machine learning, medical image analysis and AI applications in Electrical engineering. She bagged the Best Thesis Award 2019 from the IEEE Communication Society through the 'GATE'7' contest. In 2021 she received Guidence Award for the Best Project from 5th National Level IEEE Project Competition. She serves as Reviewer of IEEE Journal of Biomedical and Health Informatics, IEEE EMBS, IEEE Access, Reviewer of SPICES'17, NetACT'19 Springer Conference, ComNet'20 Springer Conference, RAICS 2020 and SPICES 2022 IEEE Conferences. She has published several papers in international journals and conferences. She is a senior member of IEEE, IE (I), IEEE Engineering in Medicine & Biology Society and Computer Society of India (CSI).

How to cite this paper: Anagha K J, Sabeena Beevi K, "Advancing Road Scene Semantic Segmentation with UNet-EfficientNetb7", International Journal of Engineering and Manufacturing (IJEM), Vol.13, No.6, pp. 53-61, 2023. DOI:10.5815/ijem.2023.06.05