

Efficient lightweight residual network for real-time road semantic segmentation

Amine Kherraki¹, Muaz Maqbool², Rajae El Ouazzani¹

¹IMAGE Laboratory, School of Technology, Moulay Ismail University of Meknes, Meknes, Morocco

²Head of AI Department, OMNO AI, Lahore, Pakistan

Article Info

Article history:

Received Jan 31, 2022

Revised Sep 2, 2022

Accepted Oct 1, 2022

Keywords:

Computer vision

Convolution neural network

Deep learning

Semantic segmentation

Tonomous driving

ABSTRACT

Intelligent transportation system (ITS) is currently one of the most discussed topics in scientific research. Actually, ITS offers advanced monitoring systems that include vehicle counting, pedestrian detection. Lately, convolutional neural networks (CNNs) are extensively used in computer vision tasks, including segmentation, classification, and detection. In fact, image semantic segmentation is a critical issue in computer vision applications. For example, self-driving vehicles require high accuracy with lower parameter requirements to segment the road scene objects in real-time. However, most related work focus on one side, accuracy or parameter requirements, which make CNN models difficult to use in real-time applications. In order to resolve this issue, we propose the efficient lightweight residual network (ELRNet), a novel CNN model, which is an asymmetrical encoder-decoder architecture. Indeed, in this network, we compare four varieties of the proposed factorized block, and three loss functions to get the best combination. In addition, the proposed model is trained from scratch using only 0.61M parameters. All experiments are evaluated on the popular public the cambridge-driving labeled video database (CamVid) road scene dataset and reached results show that ELRNet can achieve better performance in terms of parameters requirements and precision compared to related work.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Amine Kherraki

IMAGE Laboratory, School of Technology, Moulay Ismail University of Meknes

Marjane, Meknes, Morocco

Email: amine.kherraki.9@gmail.com

1. INTRODUCTION

Currently, intelligent transportation system (ITS) design has become a challenging topic in scientific research [1]. In fact, ITS applications can organize the traffic flow by performing certain tasks like road monitoring, and traffic light management, in order to ensure the pedestrians safety [2]. More recently, road scene semantic segmentation becomes one of the main issues treated by computer vision, which can aid to implement an efficient ITS. Semantic segmentation aims to label classes and segment the entire image into different parts [3]. Figure 1 shows some samples from the cambridge-driving labeled video database (CamVid) dataset and their corresponding semantic segmentation. Lately, convolutional neural networks (CNNs) are widely used for classification, and they were adapted for segmentation tasks, such as semantic segmentation in autonomous vehicles, medical field [3]–[5]. Luckily, CNN has demonstrated encouraging results in solving semantic segmentation challenges and they exceeded traditional methods in terms of accuracy and computation time. However, the combination between high precision and computational resources still remains a major challenge to improve in semantic segmentation.



Figure 1. Samples from CamVid dataset [6]. The input images are at the top and their corresponding segmentation red green blue (RGB) masks are at the bottom

In literature, several studies have been carried out on road scene semantic segmentation. In 2016, a deep neural network called efficient neural network (ENet) was proposed for real-time applications due to its few requirements in terms of parameters number [7]. However, it does not have satisfactory results in terms of precision. After that, a powerful neural network called deep labelling (DeepLab) [8] was proposed, and it achieved good results in terms of precision. Yet, DeepLab is resource-intensive while it requires 54.6 million parameters. Thus, this model takes a long time to process a high-resolution image on a powerful graphics processing unit (GPU), so it is unsuitable for real-time applications.

To solve the resource problem while maintaining high precision, and to make semantic segmentation more practical, great efforts are made to propose a new network based on new blocks. In this regard, we propose a new neural network inspired by the residual networks (ResNet) architecture [9], which can combine high precision and low resources, therefore, it is more suitable for practical cases and real-time use. In our proposed efficient lightweight residual network (ELRNet), we perform a new block with four varieties that can be installed and deployed in the intelligent vehicle's sensors. Indeed, intelligent sensors like video and rear cameras can understand the road scenes and perceive surrounding objects effectively. The experiments of our ELRNet on the CamVid database, which contains images captured from cameras installed in a vehicle, show that the proposed block significantly reduces the cost of computation while keeping a good precision. In comparison with related models, the good results of ELRNet are due to the strength of the proposed blocks, which have been tested using several criteria and hyper-parameters, such as dilation rates and precise entry channel. Actually, our ELRNet may be deployed in every intelligent vehicle that requires a road scene analysis using mounted cameras. In this paper, our main contributions are: i) a novel network for road scene semantic segmentation based on new factorized blocks, which does not require many parameters; ii) the proposed network reached good results in terms of mean intersection over union (mIoU) compared to the state of the art; and iii) the proposed network achieves good results in terms of speed, indeed, it provides real-time inference.

The rest of this paper is organized: in section 2, we review some works on traffic scene semantic segmentation. After that, a new network called ELRNet is proposed in section 3. Section 4 presents the experimental results on the challenging CamVid dataset. Finally, the conclusion and future work are provided in section 5.

2. RELATED WORK

In this section, we introduce related work on semantic segmentation, including real-time segmentation and offline segmentation. The real-time segmentation requires a model that combines speed inference and precision, and this constraint is an important and challenging task. Recently, and according to [10], the authors have proposed an efficient fast CNN for semantic segmentation called end-to-end speech processing toolkit (ESPNet). This network does not require much memory and parameters, however, it does not achieve a good precision compared to related work. Thereafter, a new model based on asymmetrical depth-wise asymmetric bottleneck network named (DABNet) has been proposed in [11]. The latter achieves good results in terms of precision and parameters requirement. Later in [12], the authors rethink semantic segmentation. Thus, they proposed a new context-guided block to learn the local feature and the surrounding context which makes encouraging results. More recently, Daliparthi [13] proposed a new CNN named IkshanaNet, which is inspired by the human brain. However, this work did not achieve good results in terms of precision and parameters requirements. Lately, a novel semantic segmentation neural network named efficient dense modules with asymmetric convolution (EDANet) has been proposed [14]. The EDANet is based on dense modules as it achieves encouraging results. Afterward, Chen *et al.* [15] presented a new neural network for road lane detection called lane marking detector (LMDNet). This network has been evaluated on the CamVid dataset, and it achieved an encouraging result in terms of mIoU. However, the

authors did not give more information about the parameter requirements. Subsequently, Visin *et al.* [16] proposed a new neural network called ReSeg, which is based on vgg16 layers. Yet, reached results are not interesting compared to other models in literature. So far, the objective of all the networks that have been already mentioned is to make a compromise between inference speed and precision. However, more improvements should be made to increase the precision and reduce the resource requirements.

The segmentation task can be carried out on online and offline applications. The offline segmentation does not care about time; thus, it is slow in time processing. In this paragraph, we will review some recent work on offline segmentation. Many segmentation networks are based on the fully convolutional network (FCN), which uses the visual geometry group network (VGGNet) as a backbone to replace the fully connected (FC) layers with convolution layers [3]. According to [8], the authors designed a new convolution module with dilation rates, to improve the DeepLab v3 network. This modification made a great change, however, it still requires an improvement in parameters and resources. Recently, U-Net [16], which has a U-shaped network architecture, has been chosen as one of the most used network architectures, especially in medical field. The authors have managed to overtake many states of the art CNN models. However, it is a very expensive network in terms of resources and parameters. After, bilateral segmentation network (BiSeNet) [17] is a CNN model that is based on the ResNet and Xception backbone. Further, the authors have succeeded to merge spatial and context information to get encouraging results. Mostly, in terms of precision, these models can achieve great results. But the parameters constraint is always the big challenge. Therefore, these models cannot be used in some edge devices, such as unmanned aerial vehicle (UAV) [18] and internet of things (IoT) systems [19].

3. RESEARCH METHOD

3.1. Dataset and metric

We use the most popular CamVid road scene semantic segmentation dataset [6], which does not require great machine performances. For reliability, we follow the related work methodology to split the whole dataset images into 367 training images, 233 test images, and 101 validation images. Besides, we use 11 classes for semantic segmentation.

In addition, we report our results using the standard metric for semantic segmentation which is mIoU, and it is characterized [20]:

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

Where true positives (TP), false positives (FP), and false negatives (FN) are the number of true pixel-level positives, false positives, and false negatives respectively, and it can be calculated for each semantic class.

3.2. Implementation Setup

All the experiments are carried out using the PyTorch framework with the compute unified device architecture (CUDA) backends. The Adam [21] optimization algorithm is used to train our proposed network. Indeed, the proposed ELRNet is trained from scratch, without a pre-trained weight. We use a learning rate of 0.045. We mention that we use the Tesla K80 GPU with 12G memory for training. Reached results analysis show that the proposed network does not have a high requirement in terms of memory, also it has a low inference time.

3.3. The proposed approach

3.3.1. Factorized residual blocks

In this subsection, we will give details about the proposed factorized block. To reduce the computing time and memory costs, the convolution factorization block separates standard convolution layers into several steps. Thus, this method has become widely exploited in lightweight neural networks, such as ENet [7] and EDANet [14], which have been already cited in the previous section. In this context, we have created our own factorized blocks as shown in Figure 2. The Figure 2(a) is block contains 3×3 conv layer, the latter is divided into four layers 3×1 conv, 3×1 dilated-conv, 1×3 conv, and 1×3 dilated-conv. We have used dilated layers to get a large receptive field with fewer parameters. Then, we add a layer of convolution 1×1 conv in an attempt to reduce the calculation cost. In parallel, we apply batch normalization (BN) with rectified linear unit (ReLU) between each layer level in order to make our CNN faster and more stable.

Afterward, we have made several varieties of the proposed block to find out the best combination. As shown in Figure 2(b), we integrated the "shuffle channel" which has been already used in [22] to make a successful exchange of information between the different groups of channels. In Figure 2(c), we applied the "max element wise", in common element-wise operations like addition. The information from every channel is averaged out. Whereas, element-wise max makes the best use of all channels as it only keeps higher values, which in essence is

the most dominant part of that channel. Finally, in Figure 2(d), we applied "max element wise x2", and the experiments show that it increased the accuracy of the mIoU compared to "max element wise".

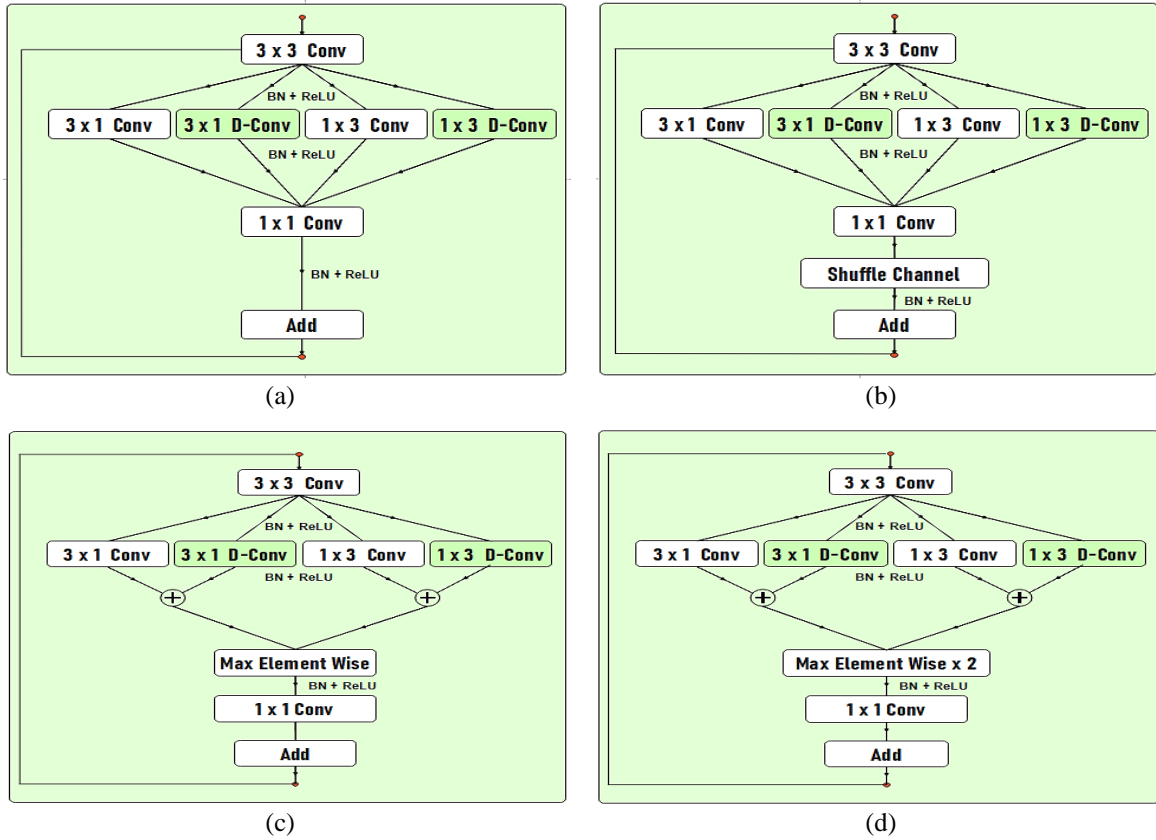


Figure 2. The proposed varieties of factorized blocks (a) Simple residual layers, (b) Residual layers with shuffle channel, (c) Residual layers with max element wise multiplied by two, and (d) Residual layers with max element wise

3.3.2. The proposed network: ELRNet

In this subsection, we introduce our proposed network as illustrated in Table 1 and Figure 3. The ELRNet is an asymmetric network, which is inspired by residual blocks that combine high precision and few computational resources. Therefore, it uses few convolutional layers with different hyperparameters, and this made the proposed model lightweight. The ELRNet consists of five blocks, the first block contains the initial stage. In the second block, we have a downsampling block, and the proposed factorized block is repeated five times with 64 input channels, and a dilation rate of $r=1$. The third one contains the proposed factorized block repeated four times, with 128 input channels, and different dilation rates $r=\{2, 4, 8, 16\}$, intending to have a larger field of view. The fourth and fifth blocks make Upsampling with three factorized blocks, and a dilation rate of $r=1$. The fourth block uses 64 input channels, and the fifth one uses 16 input channels, with the convtranspose2d output convolutional layer. We implement our proposed network with the different blocks that are mentioned in the previous section.

Table 1. The detailed ELRNet architecture

Stage	Block	Block type	Number of channels
Encoder	Block 1	Initial block	16
	Block 2	Downsampling block	64
		Factorized block $\times 5$ ($r=1$)	64
Decoder	Block 3	Downsampling block	128
		Factorized block $\times 4$ ($r=\{2, 4, 8, 16\}$)	128
	Block 4	Upsampling block	64
		Factorized block $\times 3$ ($r=1$)	64
		Upsampling block	16
	Block 5	Factorized block $\times 3$ ($r=1$)	16
		ConvTranspose2d	16

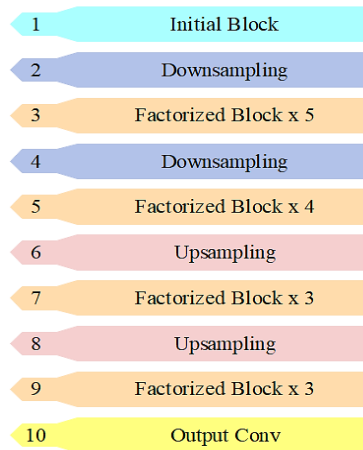


Figure 3. The overall architecture of the proposed network

4. RESULTS AND DISCUSSION

We examine our ELRNet on CamVid dataset, and we set the batch size to 8. In the training phase, we used the image size 360×480 like related work. In addition, we achieved 106 frames per second (FPS), therefore, we exceed several existing models. First, we carried out the proposed ELRNet with different varieties of factorized blocks to find out the appropriate one in road scene semantic segmentation. As we can see in Table 2, we have studied all the proposed blocks on the whole classes of the CamVid dataset. We note that block (b) is the one that gets the highest mIoU with 66.36%, as it succeeds to get the highest IoU in six classes out of eleven, including building, pole, sign symbol, fence, car, and pedestrian. It is followed by block (a) in the second place, where it achieves an average of 65.87% and can outperform the other blocks in five out of eleven classes, including sky, road, pavement, tree, and bicyclist. Block (c) came in the third place with a score of 65.10%, but it does not outperform the other blocks in terms of individual IoU, but with that, the mIoU score is close to the results of the other blocks. As for the last rank, block (d) achieves a mIoU score of 63.25% and it does not excel in any classes in terms of mIoU. However, the results achieved by this block have exceeded several models in the literature.

Table 2. Comparative results of the proposed factorized block on CamVid test set

Factorized blocks	Sky	Building	Pole	Road	Pavement	Tree	Sign symbol	Fence	Car	Pedestrian	Bicyclist	mIoU (%)
(a)	91.77	80.31	34.47	94.40	80.00	73.89	42.82	35.32	80.85	53.48	57.31	65.87
(b)	91.58	81.28	35.94	93.66	78.01	73.13	45.68	37.02	82.01	56.63	54.98	66.36
(c)	91.63	80.04	35.48	92.98	76.40	72.85	44.20	31.43	81.18	54.30	55.55	65.10
(d)	91.17	79.18	34.57	93.16	77.36	71.79	35.93	30.84	80.58	50.82	50.39	63.25

After a careful analysis of the proposed blocks, we found out that ELRNet using block (b) gets the highest mIoU. Then, we made other intensive studies by experimenting three loss functions, in particular, focal loss [23], cross-entropy [24], and lovasz-softmax [25]. Table 3 shows the results related to these implementations.

Firstly, the lovasz-softmax loss function achieves a mIoU of 66.59%, and the latter has been able to outperform other loss functions in seven of the eleven classes, which are; sky, building, pole, road, pavement, pedestrian, and bicyclist. Second, the cross-entropy loss function achieves a mIoU of 66.36%, and it outperforms the other loss functions in two out of eleven classes, in particular, sign symbol, and car. Finally, the focal loss function reaches a mIoU of 65.66%, and it outperforms the other loss functions in two out of eleven classes, notably, tree, and fence. In general, the results were close among all the used loss functions.

Table 4 shows a comparison of the most recent models on the CamVid dataset. The analysis of the obtained results shows that our ELRNet works perfectly in terms of mIoU, number of parameters, and FPS. Thus, we note that the proposed network is more adequate for practical cases and real-time applications compared to its associated works. Regarding the mIoU metric, we observe that our proposed model exceeds most models in the literature, as there is a big difference with some models. In terms of parameters, and as summarized in Table 4, we notice that our model outperformed most of the related work. We can also note that the previous works that have few parameters do not have good precision, which makes them not usable,

especially in real-time applications. In addition, we found out that the weight, size, and FPS are directly related to the number of used layers. Although the majority of related work has not evaluated these metrics, we have outperformed most of them. In addition, some models are already pre-trained, however, our ELRNet is trained from scratch.

Figure 4 shows some test images of the implemented models. Hence, the results show that our proposed ELRNet with its different block varieties, in particular, (a), (b), (c), and (d), have modicum noise and can distinguish different classes. Besides, we can see that the output of the predicted images is pure, and look like the ground truth. Despite all the efforts made in the related work, there are still problems in the semantic segmentation concerning small classes such as tree, sign symbol, and pedestrian, and this is reflected in the pictures and also the tables of results. In this way, it is necessary to work on the segmentation of small classes to ensure the safety of pedestrians, as well as animals that can be harmed. Oppositely, big classes like sky, car, and road, are very easily and accurately segmented.

Table 3. Detailed analysis of ELRNet using the best factorized block and achieved results on CamVid dataset with different loss functions

Loss functions	Sky	Building	Pole	Road	Pavement	Tree	Sign symbol	Fence	Car	Pedestrian	Bicyclist	mIoU (%)
Cross entropy	91.58	81.28	35.94	93.66	78.01	73.13	45.68	37.02	82.01	56.63	54.98	66.36
Focal	91.50	80.32	34.88	93.92	78.42	74.46	38.89	39.16	81.73	55.16	53.80	65.66
Lovasz softmax	91.65	81.30	36.87	93.96	79.49	73.63	43.17	32.40	81.04	62.10	56.85	66.59

Table 4. Performance comparison of ELRNet with related work on CamVid test set

Models	mIoU (%)	Parameters (M)	Size (MB)	FPS	Pretrained
SegNet [26]	55.6	29.5	56.2	-	Yes
FCN-8s [3]	57	134	-	39	Yes
DeconvNet [27]	48.9	252	-	26	Yes
SegNet-basic [26]	46.3	1.4	-	70	No
DABNet [11]	66.4	0.84	-	117	No
ENet [7]	51.3	0.37	0.7	149	No
CGNet [12]	65.6	0.5	3.34	-	No
LMDNet [15]	63.5	-	66	34.4	No
FC-DenseNet56 [28]	58.9	1.5	-	-	No
EDANet [14]	66.4	0.68	-	-	No
Dilation8 [29]	65.3	140.8	-	-	No
DFANet [30]	64.7	-	-	120	No
ReSeg [16]	58.8	-	-	-	No
ELRNet (our)	66.59	0.64	2.75	106.41	No

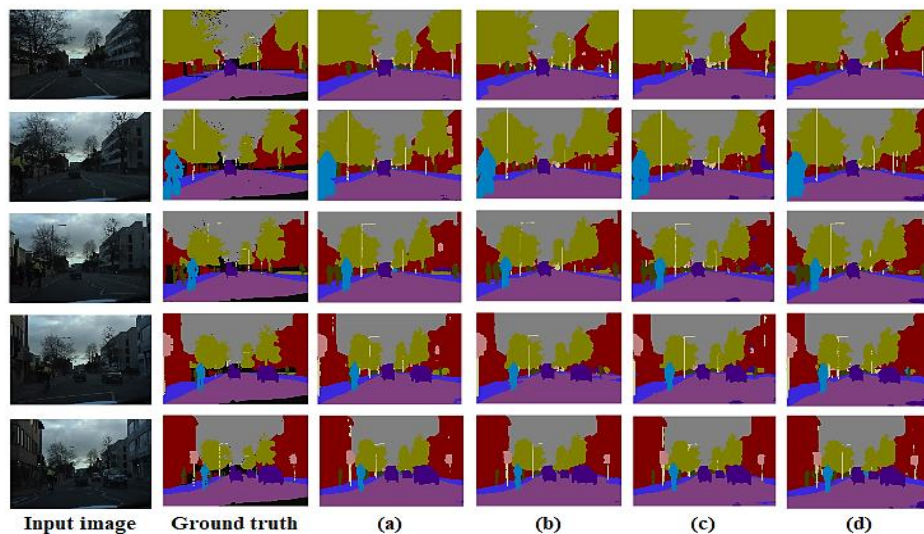


Figure 4. Results of semantic segmentation on CamVid test set

5. CONCLUSION

In this paper, we present a new and efficient lightweight residual neural network named ELRNet, which designs an asymmetrical encoder-decoder architecture for real-time road semantic segmentation. We also propose a factorized block with four varieties. Besides, we do extensive experimentation on these blocks in order to find the best combination. Our proposed ELRNet is evaluated on the most popular urban scene CamVid dataset, which demonstrates the segmentation performance of the proposed network. Generally, our network shows major amelioration in terms of parameters requirement, inference speed, and precision compared to related work. We mention that the ELRNet is trained from scratch and it achieves a mIoU of 66.59% with only 0.64M parameters. In addition, the proposed network gets a good real-time inference ability of 106.41 FPS. The extensive experiments demonstrate the efficiency of the proposed network using different varieties of the factorized block. Furthermore, the number of parameters has been importantly decreased. As a perspective, we are looking forward to developing other modules and blocks, which can improve the precision and further reduce the parameter requirements.




REFERENCES

- [1] M. Boukabous and M. Azizi, "A comparative study of deep learning based language representation learning models," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 2, pp. 1032–1040, May 2021, doi: 10.11591/ijeecs.v22.i2.pp1032-1040.
- [2] A. Kherraki and R. El Ouazzani, "Deep convolutional neural networks architecture for an efficient emergency vehicle classification in real-time traffic monitoring," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 1, pp. 110–120, Mar. 2022, doi: 10.11591/ijai.v11.i1.pp110-120.
- [3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 3431–3440. doi: 10.1109/CVPR.2015.7298965.
- [4] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998, doi: 10.1109/5.726791.
- [5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, vol. 25. [Online]. Available: <https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html>
- [6] 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. 88–97, Jan. 2009, doi: 10.1016/j.patrec.2008.04.005.
- [7] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," Jun. 2016, [Online]. Available: <http://arxiv.org/abs/1606.02147>
- [8] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, Apr. 2018, doi: 10.1109/TPAMI.2017.2699184.
- [9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.
- [10] 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)*, 2018, pp. 552–568. [Online]. Available: https://openaccess.thecvf.com/content_ECCV_2018/html/Sachin_Mehta_ESPNet_Efficient_Spatial_ECCV_2018_paper.html
- [11] G. Li, S. Jiang, I. Yun, J. Kim, and J. Kim, "Depth-wise asymmetric bottleneck with point-wise aggregation decoder for real-time semantic segmentation in urban scenes," *IEEE Access*, vol. 8, pp. 27495–27506, 2020, doi: 10.1109/ACCESS.2020.2971760.
- [12] T. Wu, S. Tang, R. Zhang, J. Cao, and Y. Zhang, "CGNet: A light-weight context guided network for semantic segmentation," *IEEE Transactions on Image Processing*, vol. 30, pp. 1169–1179, 2021, doi: 10.1109/TIP.2020.3042065.
- [13] V. S. S. A. Daliparthi, "lkshana: A theory of human scene understanding mechanism," *ArXiv*, Jan. 2021, [Online]. Available: <http://arxiv.org/abs/2101.10837>
- [14] S.-Y. Lo, H.-M. Hang, S.-W. Chan, and J.-J. Lin, "Efficient dense modules of asymmetric convolution for real-time semantic segmentation," in *Proceedings of the ACM Multimedia Asia*, Dec. 2019, pp. 1–6. doi: 10.1145/3338533.3366558.
- [15] P.-R. Chen, S.-Y. Lo, H.-M. Hang, S.-W. Chan, and J.-J. Lin, "Efficient road lane marking detection with deep learning," in *2018 IEEE 23rd International Conference on Digital Signal Processing (DSP)*, Nov. 2018, pp. 1–5. doi: 10.1109/ICDSP.2018.8631673.
- [16] F. Visin *et al.*, "ReSeg: A recurrent neural network-based model for semantic segmentation," in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2016, pp. 426–433. doi: 10.1109/CVPRW.2016.60.
- [17] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, pp. 334–349. doi: 10.1007/978-3-030-01261-8_20.
- [18] I. Idrissi, M. Azizi, and O. Moussaoui, "A lightweight optimized deep learning-based host-intrusion detection system deployed on the edge for IoT," *International Journal of Computing and Digital Systems*, vol. 11, no. 1, pp. 209–216, Jan. 2022, doi: 10.12785/ijcds/110117.
- [19] I. Idrissi, M. Azizi, and O. Moussaoui, "IoT security with deep learning-based intrusion detection systems: A systematic literature review," in *2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS)*, Oct. 2020, pp. 1–10. doi: 10.1109/ICDS50568.2020.9268713.
- [20] R. Padilla, S. L. Netto, and E. A. B. da Silva, "A survey on performance metrics for object-detection algorithms," in *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, Jul. 2020, pp. 237–242. doi: 10.1109/IWSSIP48289.2020.9145130.
- [21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Dec. 2014, [Online]. Available: <http://arxiv.org/abs/1412.6980>
- [22] Y. Wang *et al.*, "Lednet: A lightweight encoder-decoder network for real-time semantic segmentation," in *2019 IEEE International Conference on Image Processing (ICIP)*, Sep. 2019, pp. 1860–1864. doi: 10.1109/ICIP.2019.8803154.
- [23] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, Feb. 2020, doi: 10.1109/TPAMI.2018.2858826.
- [24] M. Yi-de, Q. Liu, and Q. Zhi-bai, "Automated image segmentation using improved PCNN model based on cross-entropy," in *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004.*, 2004, pp. 743–746. doi:




- 10.1109/ISIMP.2004.1434171.
- [25] M. Berman, A. R. Triki, and M. B. Blaschko, "The lovasz-softmax loss: a tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 4413–4421. doi: 10.1109/CVPR.2018.00464.
 - [26] 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. 2481–2495, Dec. 2017, doi: 10.1109/TPAMI.2016.2644615.
 - [27] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 1520–1528. doi: 10.1109/ICCV.2015.178.
 - [28] S. Jégou, M. Drozdal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jul. 2017, pp. 1175–1183. doi: 10.1109/CVPRW.2017.156.
 - [29] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *CoRR*, vol. abs/1511.0, Nov. 2015, [Online]. Available: <http://arxiv.org/abs/1511.07122>
 - [30] H. Li, P. Xiong, H. Fan, and J. Sun, "DFANet: Deep feature aggregation for real-time semantic segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9514–9523. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2019/html/Li_DFANet_Deep_Feature_Aggregation_for_Real-Time_Semantic_Segmentation_CVPR_2019_paper.html

BIOGRAPHIES OF AUTHORS






Amine Kherraki    was born in Meknes, Morocco, in 1994. He received the B.S degree in School of Technology from Hassan First University at Berrechid, Morocco, in 2017. He received the M.S degree in Computer Science at the National School of Applied Science, Sidi Mohamed Ben Abdellah University, Fez, Morocco, in 2019. Currently, He is Ph.D. candidate at the High School of Technology of Meknes, Moulay Ismail University in Morocco. His research interests include Deep Learning, Computer Vision, Pattern Recognition, and Intelligent Transportation Systems. He can be contacted at email: amine.kherraki.9@gmail.com.



Muaz Maqbool    was born in Sahiwal, Pakistan, in 1997. He received his B.S degree in Computer Science from National University of Computer & Emerging Sciences, Lahore, Pakistan, in 2019. Currently, he is the CTO of OMNO AI, and he has been advising industry-academic projects for one year. His research interests include Computer Vision enabled Sports, Traffic, Retail and Automotive Analytics. He can be contacted at email: muazmaqbool65@gmail.com.



Rajae El Ouazzani    received her Master's degree in Computer Science and Telecommunication by the Mohammed V University of Rabat (Morocco) in 2006 and the PhD in Image and Video Processing by the High National School of Computer Science and Systems Analysis (Morocco) in 2010. From 2011, she is a Professor in the High School of Technology of Meknes, Moulay Ismail University in Morocco. Since 2007, she is an author of several papers in international journals and conferences. Her domains of interest include multimedia data processing and telecommunications. She can be contacted at email: elouazzanirajae@gmail.com.