

Residual balanced attention network for real-time traffic scene semantic segmentation

Amine Kherraki¹, Shahzaib Saqib Warraich², Muaz Maqbool², Rajae El Ouazzani¹

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

²AI Department, OMNO AI, Lahore, Pakistan

Article Info

Article history:

Received Jul 26, 2022

Revised Sep 3, 2022

Accepted Sep 11, 2022

Keywords:

Computer vision

Convolution neural network

Deep learning

Self-driving

Traffic scene semantic

segmentation

ABSTRACT

Intelligent transportation systems (ITS) are among the most focused research in this century. Actually, autonomous driving provides very advanced tasks in terms of road safety monitoring which include identifying dangers on the road and protecting pedestrians. In the last few years, deep learning (DL) approaches and especially convolutional neural networks (CNNs) have been extensively used to solve ITS problems such as traffic scene semantic segmentation and traffic signs classification. Semantic segmentation is an important task that has been addressed in computer vision (CV). Indeed, traffic scene semantic segmentation using CNNs requires high precision with few computational resources to perceive and segment the scene in real-time. However, we often find related work focusing only on one aspect, the precision, or the number of computational parameters. In this regard, we propose RBANet, a robust and lightweight CNN which uses a new proposed balanced attention module, and a new proposed residual module. Afterward, we have simulated our proposed RBANet using three loss functions to get the best combination using only 0.74M parameters. The RBANet has been evaluated on CamVid, the most used dataset in semantic segmentation, and it has performed well in terms of parameters' requirements and precision compared to related work.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Amine Kherraki

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

Meknes, Morocco

Email: amine.kherraki.9@gmail.com

1. INTRODUCTION

Intelligent transportation systems (ITS) have been the focus of researchers and scientists due to the critical role they play, in particular, road safety, and traffic scene monitoring. In fact, intelligent systems can perform many tasks such as traffic flow tracking, determining the speed of vehicles, and traffic signs management [1], [2]. Lately, semantic segmentation has become one of the most important tasks that have been highlighted, due to its importance in identifying objects and image understanding [3]. Figure 1 presents samples of traffic scene images from the CamVid dataset and their corresponding semantic segmentation. In fact, deep learning (DL) algorithms, especially convolutional neural network (CNN), were able to obtain very advanced results in traffic scene semantic segmentation, compared to traditional methods [4], [5]. Before, CNN were used only for tasks' identification and classification, but in the past few years, and they have been adapted to make semantic segmentation in several areas, such as magnetic resonance images (MRI) [6], satellite images [7], traffic scenes [8]. But the biggest barrier to real-time semantic segmentation is always the combination of high precision and low computational resources.



Figure 1. Samples from CamVid dataset [9]. The input images are at the bottom and their corresponding segmentation red green blue (RGB) masks are at the top

In the literature, diverse research papers on traffic scene semantic segmentation have been conducted. In Badrinarayanan *et al.* [10] proposed a new semantic segmentation encoder-decoder network named SegNet. The latter is based on VGG16, and it requires a large amount of parameters. According to [11], a deep CNN named ENet has been proposed. This network has satisfactory results in terms of parameter requirement; however, it does not have good accuracy. In [12], a new deep neural network called DeconvNet has been proposed. The DeconvNet uses 252M parameters, therefore, it requires a large computation cost, which makes it undesirable for real-time applications. To overcome the resource problem while keeping high accuracy, great efforts are needed to propose a new CNN model. In this regard, we propose a new residual and robust network called RBANet, and it is based on a new balanced attention module (BAM) and a new residual module. Our RBANet can gather between the parameters requirement and precision, which makes it practical for real-time applications. Our major contributions to this work are as: i) a new robust network for traffic scene semantic segmentation based on a new residual module, as well as a new BAM; ii) the proposed network achieved good results in terms of mean intersection over union (mIoU) and the required number of computational parameters.

The remainder of this paper is structured as. We will look at some research on traffic scene semantic segmentation in section 2. Then, in section 3, we explain the proposed network RBANet by giving details on the proposed residual module and the proposed BAM. Section 4 shows the experimental results on the CamVid dataset. Finally, the conclusion and future directions are presented in section 5.

2. RELATED WORK

This section addresses recent traffic scene semantic segmentation research work, such as real-time and offline segmentation. Real-time traffic scene semantic segmentation necessitates a model that combines speed and precision, which is a substantial and tough problem. Mehta *et al.* [13] proposed a new fast and efficient CNN for traffic scene semantic segmentation called ESPNet. The latter does not need great parameters' number; however, it does not attain excellent precision when compared to similar work. Subsequently Wu *et al.* [14] reconsider traffic scene semantic segmentation. As a consequence, they presented a novel context-guided block to learn the surrounding, local features, and context. This work achieves encouraging results. According to [15], a CNN inspired by the human brain called IkshanaNet, has been proposed. However, this network did not yield satisfactory results in terms of precision and parameter requirements. Later, the semantic segmentation neural network called EDANet has been suggested in [16]. The EDANet is structured and based on dense modules, which obtain impressive results. Thereafter, Li *et al.* [17] have introduced a novel model based on asymmetrical Depth-Wise Bottleneck named DABNet. The latter delivers good results in terms of precision, mIoU and parameters' requirements. According to [18], Visin *et al.* have developed a CNN named ReSeg based on VGG16 backbone. However, ReSeg did not produce satisfactory results in terms of precision, and the authors did not mention the number of parameters. Later, a new deep neural network based on feature aggregation called DFANet has been proposed in [19]. The latter achieved an acceptable result in terms of precision; however, the authors did not mention the parameters' number. Until now, all of the networks that have been designed are to achieve a compromise between inference speed and precision. However, further progress is required to improve precision while lowering resource costs.

Semantic segmentation tasks may be used in both online and offline applications. Offline segmentation is slow to process since it is indifferent to time. In this part, we will look at some recent work on offline traffic scene semantic segmentation. Chen *et al.* [20] have developed a CNN module using dilation to

improve the DeepLabv3 model. This adjustment made a significant difference in terms of precision, but it still has to be improved in terms of computation cost. After, a deep CNN model named BiSeNet has been proposed in [21], and it is based on the ResNet and Xception backbone. BiSeNet has achieved encouraging results in terms of precision, however, it is very expensive in terms of parameters' requirements, which is not suitable for real-time applications. In general, the models that are expensive in terms of computation resources may produce excellent gains in accuracy, however, the parameters constraint is usually the most difficult barrier. As a result, the heavy models are incompatible with various edge devices such as Raspberry Pi, Arduino, and field-programmable gate array (FPGA) [22].

3. METHOD

3.1. Dataset and metric

We have used CamVid, the most widely used traffic scene semantic segmentation dataset [9], which does not necessitate strong machine performance. Indeed, we have used the relevant work strategy to divide the entire image dataset into rain, test and validation sets that contain 367 images, 233 images, and 101 images respectively. In addition, there exist 32 classes, but only 11 of them are used for semantic segmentation in the CamVid dataset [23]. Furthermore, we provide our outcomes using the standard measure for semantic segmentation, mIoU, which is defined as (1) [24]:

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

with TP , FP , and FN are the number of true pixel-level positives, false positives, and false negatives, and may have been computed for each semantic class.

3.2. Implementation setup

The PyTorch framework with CUDA backends is used for all experiments. Our proposed RBANet is trained from scratch, without pre-trained weight. We mention that we have used Adam [25] as an optimization algorithm, and we have chosen a learning rate of 0.045. We have used Google Collaboratory with Tesla K80 GPU and 12 GB RAM to train our proposed model. The outcomes of our simulation prove that the suggested network does not require a great parameters' number.

3.3. The proposed approach

3.3.1. Residual module

In this part, we will go through the proposed residual module in depth. Our proposed module is based on convolution factorization, which divides typical convolution layers into many stages to minimize processing time and memory expenses [26]. Therefore, this approach is frequently used in various lightweight CNN, such as ENet [11], EDANet [16], DABNet [17], ESNet [27], and LEDNet [28]. In this regard, we have created a novel residual module as shown in Figure 2.

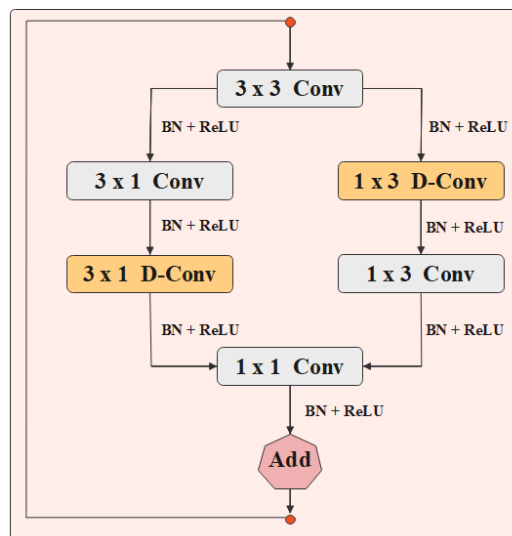


Figure 2. The proposed residual module

The proposed module contains 3×3 Conv layer, the latter is divided into two layers, in particular, 3×1 Conv, and 1×3 Dilated-Conv. After that, in the next level, we have linked the layer 3×1 Conv with another layer 3×1 Dilated-Conv, as well as we have linked the layer 1×3 Dilated-Conv to the layer 1×3 Conv. In parallel, and to get a larger field of view using fewer parameters, we applied dilated layers. Then, in an attempt to lower the computation cost, we have added a layer of 1×1 Conv. In addition, and in order to make our CNN stable and quicker, we have used batch normalization (BN) [29] with rectified linear unit (ReLU) [30] in each level of the convolution layer.

3.3.2. Balanced attention module

In this subsection, we will discuss the proposed balanced attention module (BAM), which discusses the potential of multi-layered attention convolutional blocks that are relayed in [31], [32]. The architecture of the proposed BAM involves the application of channel and spatial attention mechanisms to feature maps from preceding convolutional down sampler blocks using a balanced feature sharing mechanism. The channel attention block is used to extract useful information from the input image, while the spatial attention block further identifies the most useful information within the output received from the channel attention as illustrated in Figure 3(a).

In Figure 3(b), the channel attention module is depicted where the output features from preceding convolutional layers are refined. Firstly, the input features are max pooled and average pooled simultaneously. Average pooling is used to aggregate spatial information to induce a smoothing effect while max pooling is used to induce a sharpening effect by preserving contextual information in terms of object edges. The output features generated by the two pooling layers are simultaneously passed to the multi-layer perceptron (MLP) layers whose output vectors are then concatenated element-wise. The final resultant vector is then sent to the rectified linear unit (ReLU) activation function, which generates the important feature maps for the spatial attention module.

In Figure 3(c), the spatial attention module is depicted where the output feature maps from the preceding channel attention module are further refined. Firstly, the input feature maps are max pooled and average pooled simultaneously for similar reasons as in the channel attention module. The output vectors from the pooling layers are then concatenated element-wise before being passed as input to a convolutional layer to produce a single channel feature map. Subsequently, the final feature map is sent to the ReLU function, which produces a spatial mask for the identification of important features. The spatial mask is applied to all significant feature maps from the preceding channel attention module to identify essential features by using element-wise multiplication. Finally, the output features from spatial attention are added with the input features using residual connections, to generate more refined and highlighted feature maps.

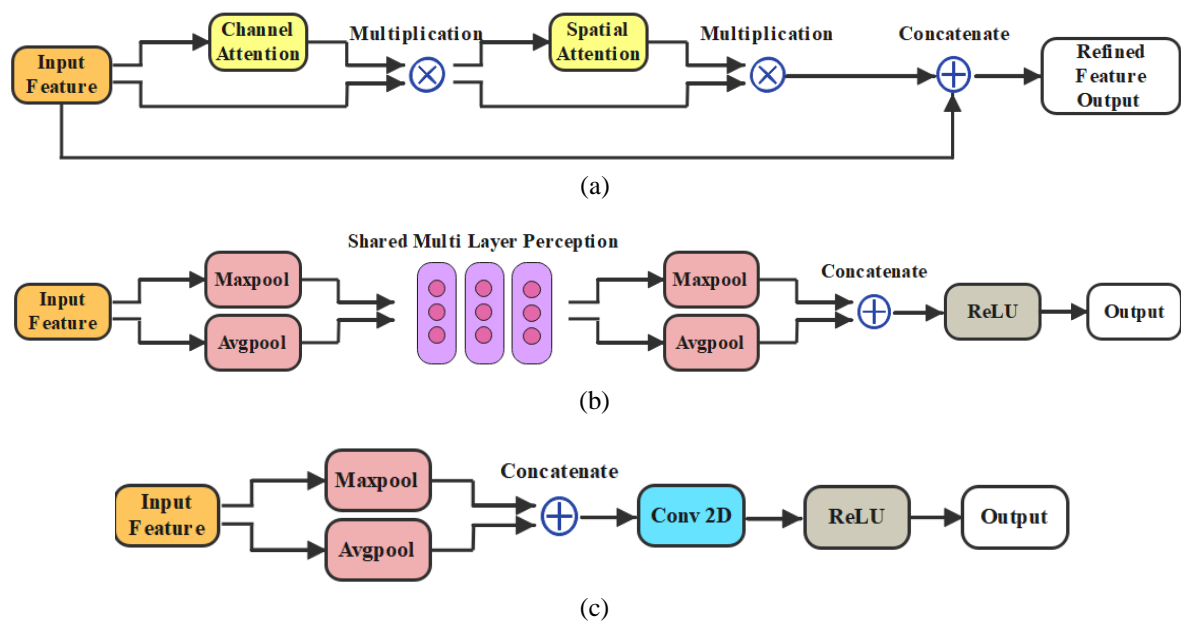


Figure 3. The proposed module (a) the overall BAM, (b) channel attention block, and (c) spatial attention block

3.3.2. The proposed RBANet

In this subsection, we will describe our intended architecture network, which is depicted in Table 1 and Figure 4. The RBANet is a robust network inspired by residual connections, which combine great accuracy with a limited amount of parameters [33]. As a result, it employs few convolutional layers with varying hyperparameters, therefore, it makes the proposed model lightweight. The proposed network is made up of five blocks, where the first one involves the initial stage. The second, third, and fourth blocks are composed of a downsampling Block, the proposed attention module, and the proposed residual module. The residual module is repeated four times in the second block with 16 input channels, and different dilation rates of $r=\{2, 2, 4, 4\}$. In the third block, the residual module is repeated four times with 64 input channels, and different dilation rates of $r=\{8, 8, 16, 16\}$. The fourth block consists of the same element mentioned for the precedent blocks, except that the residual module is repeated five times with 128 input channels, and different dilation rates of $r=\{32, 32, 64, 32, 32\}$. The fifth and sixth blocks use upsampling with the residual module repeated four times for each block. The fifth block uses 64 input channels and a dilation rate of $r=\{16, 16, 8, 8\}$, whereas, the sixth one uses 16 input channels, with a dilation rate of $r=\{4, 4, 2, 2\}$. Finally, we have used the ConvTranspose2d output convolutional layer.

Table 1. The detailed RBANet architecture

| Stage | Block | Block Type | Number of Channels |
|---------|---------|---|--------------------|
| Encoder | Block 1 | Initial Block | 16 |
| | | Downsampling Block | 16 |
| | Block 2 | The proposed BAM | 16 |
| | | The proposed Residual Module $\times 4(r=\{2, 2, 4, 4\})$ | 16 |
| | Block 3 | Downsampling Block | 64 |
| | | The proposed BAM | 64 |
| | | The proposed Residual Module $\times 4(r=\{8, 8, 16, 16\})$ | 64 |
| | | Downsampling Block | 128 |
| | Block 4 | The proposed BAM | 128 |
| | | The proposed Residual Module $\times 5(r=\{32, 32, 64, 32, 32\})$ | 128 |
| Decoder | Block 5 | Upsampling Block | 64 |
| | | The proposed Residual Module $\times 4(r=\{16, 16, 8, 8\})$ | 64 |
| | Block 6 | Upsampling Block | 16 |
| | | The proposed Residual Module $\times 4(r=\{4, 4, 2, 2\})$ | 16 |
| | | ConvTranspose2d | 16 |

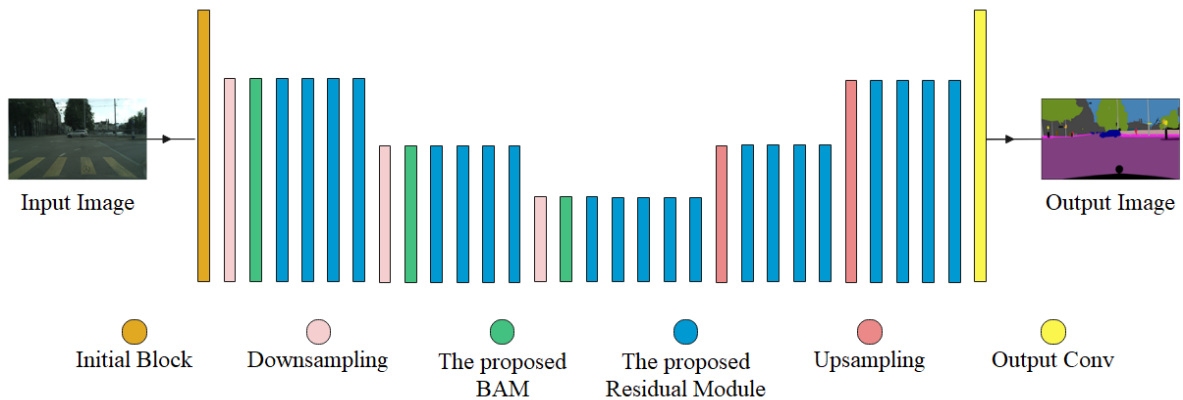


Figure 4. The overall architecture of the proposed RBANet

4. RESULTS AND DISCUSSION

We perform our RBANet on the CamVid dataset with a batch size of 8. During the training stage, we chose an image with a resolution of 360×480 as the relevant literature to keep the reliability. Furthermore, we have reached 66.82% in terms of mIoU using only 0.74M parameters, and 106 frames per second (fps), thus, we have outperformed many related work models. As we can see in Table 2, we have studied the proposed RBANet on the whole classes of the CamVid dataset. Afterward, we have made a careful study by experimenting three loss functions, in particular, Focal Loss [34], Cross-Entropy [35], and LovaszSoftmax [36]. To start, the LovaszSoftmax loss function has achieved 66.82% in terms of mIoU, and it has been eligible

to exceed the other loss functions in seven classes of the eleven, in particular, bicyclist, pedestrian, fence, sign symbol, pavement, pole, and building. Thus, we can say that our proposed RBANet using the LovaszSoftmax managed to obtain high precision in the most difficult and small classes. Afterward, with a mIoU of 65.64%, the cross-entropy loss function surpasses the other loss functions in three classes of the eleven, namely car, road, and sky. Moreover, the proposed RBANet using the cross-entropy has been able to get high precision in large classes, which are easy to recognize by the model. Finally, the Focal loss function achieves a mIoU of 64.58 % and surpasses the other loss functions in one class, which is Tree. In overall, the outcomes were nigh between all the loss functions.

Table 2. Detailed analysis of RBANet results on CamVid dataset using different loss functions

| Loss functions | Sky | Building | Pole | Road | Pavement | Tree | Sign Symbol | Fence | Car | Pedestrian | Bicyclist | mIoU (%) |
|----------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| Cross Entropy | 91.39 | 80.22 | 34.11 | 94.99 | 81.35 | 73.21 | 42.97 | 29.44 | 82.29 | 55.12 | 56.94 | 65.64 |
| Focal | 91.22 | 80.03 | 32.71 | 94.17 | 80.26 | 73.75 | 40.24 | 33.55 | 76.75 | 53.31 | 54.42 | 64.58 |
| Lovasz Softmax | 91.21 | 80.95 | 37.85 | 94.74 | 82.04 | 72.87 | 44.65 | 32.83 | 79.04 | 60.12 | 57.84 | 66.82 |

Table 3 compares several latest models on the CamVid traffic scene dataset. According to the reached results, our RBANet performs fluently in terms of mIoU, number of parameters and fps. Therefore, we observe that the proposed network is better suitable for practical uses and real-time applications than its competitors [14], [37]. Concerning the mIoU metric, we notice that our proposed RBANet outperforms the majority in the state of the art [17], [19], also there is a wide difference with some models [10], [12], [38]. In terms of the parameters' number, our RBANet outperformed the bulk of similar research as seen in Table 3. In addition, we can observe that some works with restricted parameters are inaccurate such as SegNet-Basic [10], and ENet [11], which makes them unsuitable for real-time applications. Furthermore, we discovered that the number of layers used in our proposed model has a direct relationship with the model weight size and fps. Despite the fact that most of the relevant work has not assessed these metrics, we have exceeded the majority of them, in particular, [16], [18], [39]. Besides that, whereas some models are formally pre-trained like, [10], [12], our RBANet is trained from scratch without using a pre-trained weight.

Table 3. Performance comparison of ELRNet with related work on CamVid test set

| Models | mIoU (%) | Parameters (M) | Size (MB) | Fps | Pre-trained |
|-------------------|--------------|----------------|------------|------------|-------------|
| SegNet [10] | 55.6 | 29.5 | 56.2 | - | Yes |
| FCN-8s [38] | 57 | 134 | - | 39 | Yes |
| DeconvNet [12] | 48.9 | 252 | - | 26 | Yes |
| SegNet-Basic [10] | 46.3 | 1.4 | - | 70 | No |
| DABNet [17] | 66.4 | 0.84 | - | 117 | No |
| DFANet [19] | 64.7 | - | - | 120 | No |
| ENet [11] | 51.3 | 0.37 | 0.7 | 149 | No |
| CGNet [14] | 65.6 | 0.5 | 3.34 | - | No |
| LMDNet [37] | 63.5 | - | 66 | 34.4 | No |
| Dilation8 [39] | 65.3 | 140.8 | - | - | No |
| EDANet [16] | 66.4 | 0.68 | - | - | No |
| ReSeg [18] | 58.8 | - | - | - | No |
| RBANet (our) | 66.82 | 0.74 | 3.31 | 115 | No |

Figure 4 depicts a few test images using the proposed RBANet. Consequently, the outcomes prove that our proposed model can differentiate between distinct classes despite some little noise. Furthermore, we can observe that the predicted image output is clean and appears to be the ground truth. Considering all the achievements that are made in the related work, there are still issues in semantic segmentation of small classes such as tree, sign symbol, and pedestrian, as seen in Figure 5 and Table 3. As a result, it is vital to work on the segmentation of smaller classes in order to save the life of pedestrians and animals that may be damaged. On contrary, large classes such as road, car, and sky may be simply and precisely segmented.

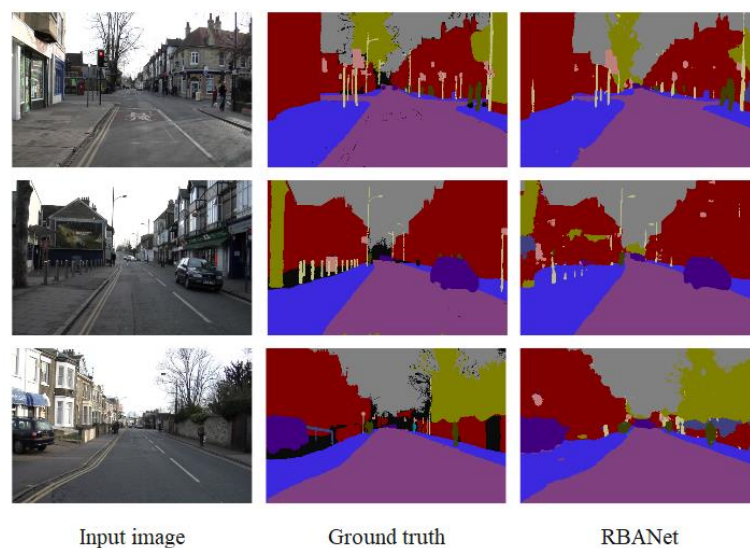


Figure 5. RBANet evaluation result on CamVid dataset

5. CONCLUSION

In this paper, we present a new robust and lightweight neural network called RBANet for real-time traffic scene semantic segmentation. Besides, we also propose a new attention module that uses spatial attention and channel attention. At the same time, we have proposed a new residual module. The proposed RBANet has been evaluated on the CamVid dataset, and this latter proves the segmentation rendering of the proposed model. In general, the RBANet displays a great improvement in terms of mIoU, and parameters' requirements compared to the state of the art. Our proposed model is trained from scratch, and it achieves a mIoU of 66.82% using only 0.74M parameters. The wide experiments demonstrate the effectiveness of the proposed RBANet using different loss functions. Moreover, the parameters' requirement has been importantly relieved. In future work, we will mainly look forward to evolving new attention modules to reduce the computational cost, and improve precision.




REFERENCES

- [1] 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.
- [2] M. Boukabous and M. Azizi, "Review of learning-based techniques of sentiment analysis for security purposes," in *The Proceedings of the Third International Conference on Smart City Applications*, 2021, pp. 96–109.
- [3] M. Berrahal and M. Azizi, "Augmented binary multi-labeled CNN for practical facial attribute classification," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 23, no. 2, pp. 973–979, Aug. 2021, doi: 10.11591/ijeecs.v23.i2.pp973-979.
- [4] I. Idrissi, M. Boukabous, M. Azizi, O. Moussaoui, and H. El Fadili, "Toward a deep learning-based intrusion detection system for IoT against botnet attacks," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 10, no. 1, pp. 110–120, Mar. 2021, doi: 10.11591/ijai.v10.i1.pp110-120.
- [5] 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.
- [6] S. Kumar, A. Negi, J. N. Singh, and H. Verma, "A deep learning for brain tumor MRI images semantic segmentation using FCN," in *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, Dec. 2018, pp. 1–4, doi: 10.1109/CCAA.2018.8777675.
- [7] B. Neupane, T. Horanont, and J. Aryal, "Deep learning-based semantic segmentation of urban features in satellite images: A review and meta-analysis," *Remote Sensing*, vol. 13, no. 4, Feb. 2021, doi: 10.3390/rs13040808.
- [8] A. Kherraki, M. Maqbool, and R. El Ouazzani, "Traffic scene semantic segmentation by using several deep convolutional neural networks," in *2021 3rd IEEE Middle East and North Africa COMMUNICATIONS Conference (MENACOMM)*, Dec. 2021, pp. 1–6, doi: 10.1109/MENACOMM50742.2021.9678270.
- [9] 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.
- [10] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: a deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, Nov. 2015.
- [11] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: a deep neural network architecture for real-time semantic segmentation," *arXiv:1606.02147*, pp. 1–10, Jun. 2016.
- [12] 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.




- [13] 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. 561–580.
- [14] 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.
- [15] V. S. S. A. Daliparthi, "Ikshana: a theory of human scene understanding mechanism," *Prepr. arXiv2101.10837v4*, 2021.
- [16] 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.
- [17] 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.
- [18] 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.
- [19] 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.
- [20] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *Prepr. arXiv1706.05587*, Jun. 2017.
- [21] 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.
- [22] S. Sarkar, R. Wankar, S. N. Srirama, and N. K. Suryadevara, "Serverless management of sensing systems for fog computing framework," *IEEE Sensors Journal*, vol. 20, no. 3, pp. 1564–1572, Feb. 2020, doi: 10.1109/JSEN.2019.2939182.
- [23] L. Wang, D. Li, H. Liu, J. Peng, L. Tian, and Y. Shan, "Cross-dataset collaborative learning for semantic segmentation in autonomous driving," in *The Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22)*, Mar. 2021, pp. 2487–2494.
- [24] 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.
- [25] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," *arXiv Prepr. arXiv1412.6980*, Dec. 2014.
- [26] X. Zhang, Z. Chen, Q. M. J. Wu, L. Cai, D. Lu, and X. Li, "Fast semantic segmentation for scene perception," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 2, pp. 1183–1192, Feb. 2019, doi: 10.1109/TII.2018.2849348.
- [27] H. Lyu, H. Fu, X. Hu, and L. Liu, "Esnet: Edge-based segmentation network for real-time semantic segmentation in traffic scenes," in *2019 IEEE International Conference on Image Processing (ICIP)*, Sep. 2019, pp. 1855–1859, doi: 10.1109/ICIP.2019.8803132.
- [28] 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.
- [29] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate Shift," in *International conference on machine learning*, 2015, pp. 448–456.
- [30] A. F. Agarap, "Deep Learning using rectified linear units (ReLU)," *arXiv preprint arXiv:1803.08375*, Mar. 2018.
- [31] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.
- [32] L. Chen *et al.*, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 6298–6306, doi: 10.1109/CVPR.2017.667.
- [33] 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.
- [34] 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.
- [35] Yi-de Ma, Qing Liu, and Zhi-bai Quan, "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.
- [36] 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.
- [37] 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.
- [38] T. Hu, Y. Deng, Y. Deng, and A. Ge, "Fully convolutional network variations and method on small dataset," in *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, Jan. 2021, pp. 40–46, doi: 10.1109/ICCECE51280.2021.9342059.
- [39] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *4th International Conference on Learning Representations, ICLR 2016-Conference Track Proceedings*, pp. 1–13, Nov. 2015.

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 Moulay Ismail University, Meknes, 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.






Shahzaib Saqib Warraich    was born in Sargodha, Pakistan, in 1998. He received his Bachelor's degree in Electrical and Electronics Engineering from the National University of Sciences and Technology (NUST), Islamabad, in 2021. Currently, he is an AI Engineer and researcher at OMNO AI and Adlytic AI, Pakistan. His research interests include deep learning, computer vision, and pattern recognition. He can be contacted at email: shahzaibsaqibwarraich1@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 and 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 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 Ph.D 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.