International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

U-In-Effnet: Semantic Segmentation with the Effect of Magnifying Glasss

Shubha Rao A[1] and Dr. Mahantesh K [2]

Submitted:21/03/2023 **Revised:**25/05/2023 **Accepted:**10/06/2023

Abstract. Each of the growing advancement in the field of Semantic segmentation has taken us a step ahead towards effectively achieving Artificial Intelligence. CNN has again and again proved its semantic competency in extracting rich features at various level of abstraction. Semantic segmentation of the image for better understanding of the objects and its context is essential for wide range of applications ranging from scene classification to automatic driving vehicles. An encoder- decoder inspired U-net architecture using Efficient Net as backbone is been proposed. To replicate the effect of magnifying glass for diverse feature rich extraction inception blocks with different kernel size is added into decoder. The proposed method is tested on the most unstructured, delineated Indian Driving Dataset (IDD) and the popular benchmark Pascal VOC 2012 dataset. The architecture with its well defined segmentation map outperforms the previous benchmark results by attaining mean Intersection over Union (mIoU) of 0.78 and 0.63 on Pascal VOC and IDDLite dataset respectively.

Keywords. *Unet, EfficientNet, Semantic Segmentation, CNN.*

1. Introduction

Convolution Neural Network (CNN) has brought a change in the way images are processed, understood and recognized [1]. It had lead us from the coarse image classification task to fine level object detection, bounding box prediction, style transfer. Sematic Segmentation is finer pixel level image processing which involves prediction of each pixel with its identifying class. Semantic segmentation has been widely used for scene classification, localization and navigation related tasks [2]. Most of the old school methods involved manual feature extraction, could process images only of low resolution which ate up all the finer level details. Along with CNN, advancements in GPU have led the researchers into building innovative approaches for segmentation tasks like FCN [3], U-Net [4], PSP-Net [5], Deep Lab [6] to Mask-RCNN [7]. In the era, heading towards self-driving vehicles which works mainly on the provided context information, proper semantic segmentation with perfect blend of fine-low level features and coarse-semantic level feature is evident [8].

U-Net is most often popularly used architecture without any complex pre and post processing as requirement, yet efficient enough in fusing the segmentation maps at various level of the pyramid [9]. Unet merges the information of what the images contains to where exactly, by effectively merging the feature maps in hierarchy from fine to coarse level. The general architecture of the Unet is shown in Figure 1. Unet consists an encoder - which follows a contracting path using max pooling, a bridge and a decoder – which is an expanding process using upsampling for image reconstruction. The key element responsible for faster convergence lies in the architectures skip connection which aids in retaining the losses incurred along the downhill.

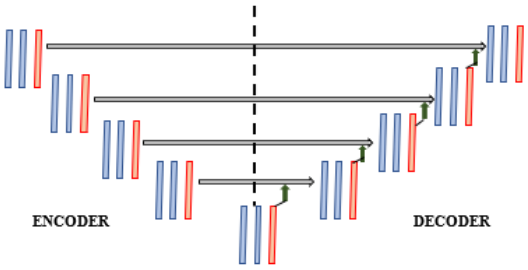

Fig 1. The general architecture of U-Net.

2. Related Work

An automatic liver segmentation in CT images called EAR-U-Net is proposed, using EffiicientNet B4 as framework for U-Net. The method uses an added Attention gate at skip connection to focus on the features which are appropriate for the task and residual block to tackle with the vanishing gradient problem. The EAR-U-Net has proved its efficiency in MICCIA-LiTS17 challenge with the dice score of 0.952 [10]. A Mask R-CNN inspired model for Efficient Panoptic segmentation with customized head which integrates the combination of rich features. The method has proved its accuracy on KITTI, Military Vistas, Cityscapes and Indian Driving Dataset [11]. U-Net based architecture with EfficientB7

[1] *Research Scholar, Department of ECE, SJB Institute of Technology, Bangalore, India*
[2]*Associate Professor, Department of ECE, SJB Institute of Technology, Bangalore, India*
mail2shugar@gmail.com[1]

as backbone along with simple convolution layers using combination of jacquard and binary cross entropy is employed for segmentation of IDD-Lite dataset with test mIoU of 0.6276 [12].

Half U-Net with full scale fusion of features with the aid of ghost modules is proposed for mammography and lung nodule segmentation. This simplified architecture reduces the total number of operations, parameters while achieving similar accuracy as the original U-Net and its variants [13]. A multi-scale based attention merged with Unet at the skip connection called the MSA-UNet is proposed for liver segmentation. The model has achieved a dice score of 98% and mIoU of 96% on publically available datasets [14]. To address the issue of class imbalance which is often faced in segmentation, another variant with batch normalization, reduced filters and dropout layers called class balanced Unet is proposed by the authors. The method shows an improvement of 0.01% in liver and 0.11% in tumor, but failed to recognize oddly shaped tumors [15]. The W-MSA structure present in Swin is merged with convolution and the general architecture of Unet for faster convergence of the model while preserving the locality information from the patches for medical CT image segmentation [16].

3. U-in-EffNet: Proposed Methodology

Many variants of U-Nets are been proposed by the researchers along the history. The usage of transfer learning based pre-trained model as encoder for U-net is a popular culture. Encoder with pre-trained weights of a base model works as ideal feature extractor. In the proposed methodology, the state-of-the art EfficientNet model is used as encoder, instead of using traditional convolutional layer as decoder, a magnifier which increases the field of view, capable of capturing diverse features is designed. Architecture details of encoder and decoder is elaborated in the further sections.

3.1 Encoder: EfficientNet-B7

The general thumb rule followed while working with the CNNs states if efficiency is not good enough - increase the complexity of the model either by adding more layers, increase the filters or by using higher resolution. Well, the unsaid ugly truth here is the increased complexity is not always directly proportional to efficiency. It saturates, more gets worse additionally by eating up the memory and making the model too complex. A novel technique EfiicientNet was proposed which does the scaling up task but in a smarter way called compound scaling [17]. The architecture is made up of several blocks, each block is made up of several sub-blocks called MBConv block which was originally proposed by the authors of MobileNetV2 [18]. MBConv Block has series of Expand, DepthwiseConv, Squeeze, Reshape, Expand, Excite layer. With the aid of inverted residuals EfficientNet family of models justifies the name given with smart compound scaling and fewer number of parameters. For the proposed UNet based semantic segmentation model, EfficientNet-B7 is used as the backbone along with the pre-trained weights on ImageNet dataset.

3.2 Decoder: Magnifier using Inception Block

As opposed to the conventional decoder, the proposed Unet has a decoder which is asymmetrical in nature with skip connection from its corresponding layers of the encoder. Transpose convolutions are used to upsample the data to increase the resolution as required for concatenation from its relative skip connection. Addition of skip connection ensures better localization of information while retaining the global spatial relationships. The concatenated features are passed onto to the inception block which works as magnifying glass. The inception block employs multiple size of kernels to capture the diverse sparse features present in the image which is concatenated before passing to the next layer, resulting in efficient prediction of feature rich segmentation maps [19]. The architectural detail of the proposed methodology is shown in Figure 3.

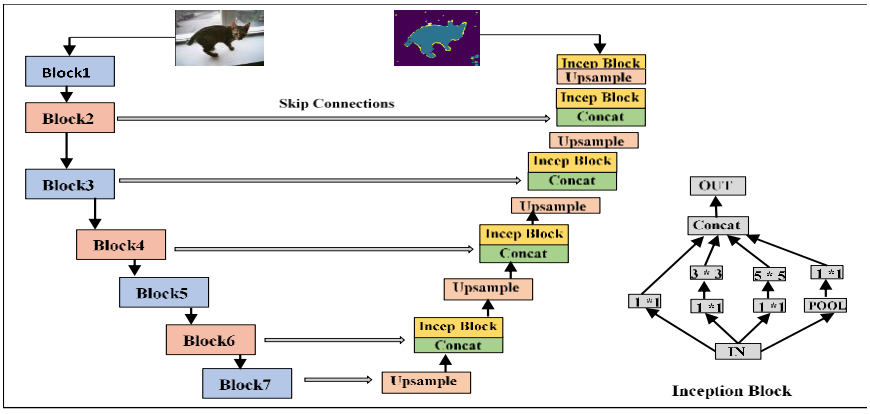

Fig 2. Architecture of the proposed U-in-EffNet.

3.3 Datasets

Pascal VOC2012: It is the most popular publically available benchmark image datasets used for Object detection, Segmentation and Action classification. The dataset contains visual image object from the real life environment with the ground truth of bounding box, object class labels and reference spatial points [20]. It has around 11,530 images, with object categories belonging to aeroplane, bicycle, boat, bus, car, motorbike, train, bottle, chair, dining table, potted plant, sofa, TV/monitor, bird, cat, cow, dog, horse, sheep, and person. Sample images from the dataset can be seen in Figure 4 (a).

Fig 3. Sample images from datasets (a) Pascal VOC2012 (b) IDD-Lite

Indian Driving Dataset (IDD): According to the statistics provided by the World Health Organization (WHO) road accidents stands at the 8th position for world leading death causes [21]. During the time where autonomous navigation system with ADAS features are gaining major consumer preferences and manufacturers attention, techniques with precise knowledge of surrounding environment is apparent basic requirement. A much diverse, unstructured, complex dataset with road sequence images along with ground truth is made available under four different hierarchies of classes [22]. For the evaluation of proposed method level-1 with 7 classes called IDD Lite version is used with 1404 to train, 204 for validation and 408 test images. Some of the sample images from the dataset are shown in Figure 4 (b).

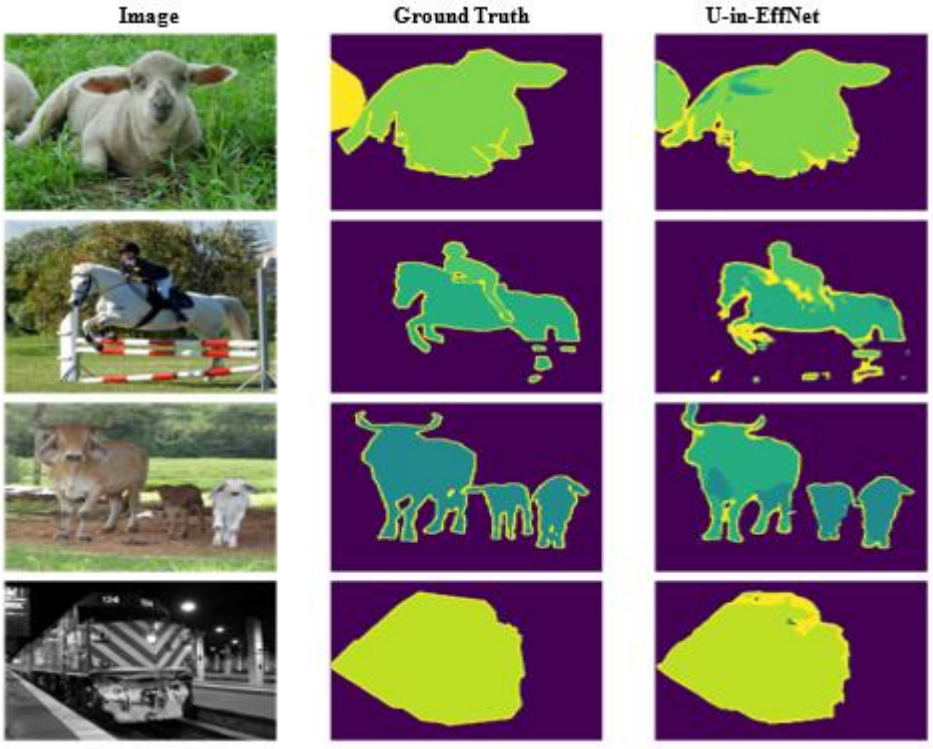

Fig 5. Segmentation maps predicted by U-in-EffNet for PascalVOC2012

3.4 Result & Discussion

The proposed architecture was tested under experimental set up by using TensorFlow 2.0 framework as backbone via Python3 programming language with 52GB RAM and T4 GPU. The input was normalized, augmented before passing it to the model, later trained using Adam as an optimizer with learning rate of 0.0001 and Focal Loss as loss function. The results are evaluated with best saved model using mean Intersection over Union which is a standard metric commonly used for segmentation tasks. For the Pascal VOC2012 dataset the model was trained for 60 epochs with batch size of 20, the segmentation results are shown in Figure 4. For IDD Lite dataset the model was trained for 10 epochs with the batch size of 4, Figure 5. shows few of the segmentation results along with its ground truth. A comparative analysis of the proposed methodology with previous work for Pascal VOC2012 and IDD Lite dataset is been tabulated in Table 1. and Table 2. respectively.

$$Adam = w_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \quad (1)$$

$$Focal\ Loss = -\alpha_t (1 - p_t)^{\gamma} log (p_t) \quad (2)$$

$$Mean\ IoU = \frac{1}{n} \sum \frac{TP}{TP_n + FP_n + FN_n} \quad (3)$$

In comparison to the earlier works which are based on FCN, ERFNet, ResNet, Unet, the proposed U-EfficientNet with induced inception block proves its greater discriminating ability between the classes. The model shows 0.7% increased efficiency in mIoU for PascalVOC 2012 and 0.1% increased efficiency in IDDLite level-1 dataset in comparison to earlier work. The models performance is poor when it comes to level-2 minority classes like curb, wall, billboard, traffic sign, traffic light, pole, fence which occupies a very small percentage of the overall image. The graphical representation of the results can be seen in Figure 6.

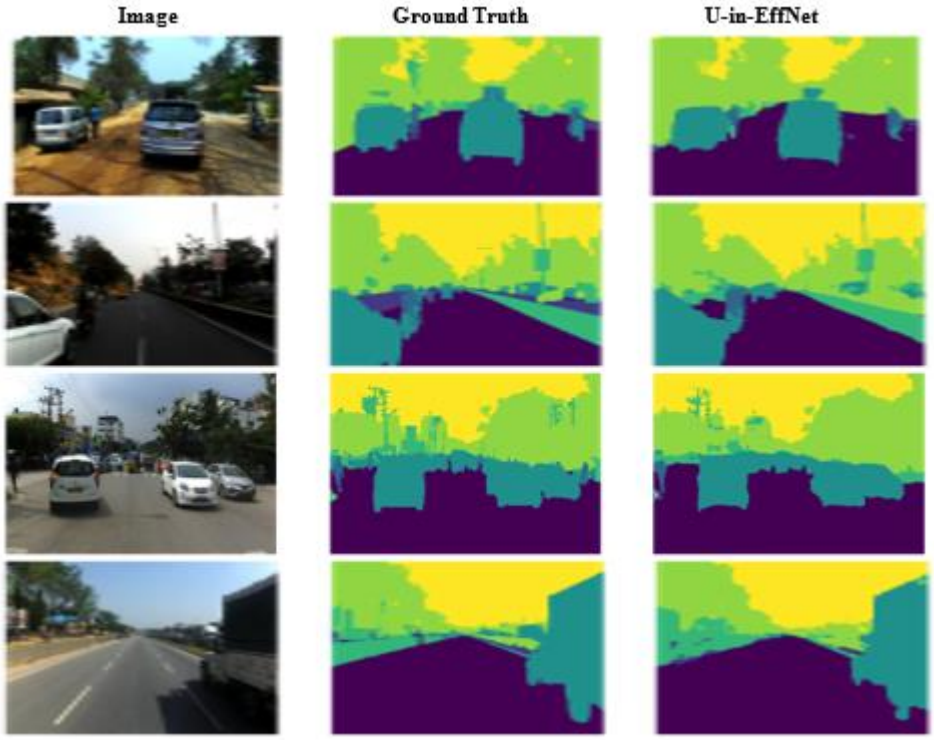

Fig 5. Segmentation maps predicted by U-in-EffNet for IDD-Lite

Table 1. Comparative analysis of Proposed U-in-EffNet on Pascal VOC2012 dataset result with previous work

Methodology	Test mIoU
R-CNN [23]	47.9
SDS [24]	51.6
FCN-8s [3]	62.2
DiCENet [25]	66.5

ESPNetV2[26]	68.0
ParseNet [27]	69.8
DeepLab-MSc-CRF-LargeFOV [28]	71.6
U-in-EffNet (Proposed)	**78.5**

Table 2. Comparative analysis of Proposed Proposed U-in-EffNet on IDD dataset result with previous work

Methodology	Validation mIoU	Test mIoU
Dilated ResNet18[29]	55.03	-
ERFNet [30]	66.13	-
DeeplabV3+ with ResNet18 [31]	63.03	56.14
U-Net with ResNet 50 [31]	68.59	60.76
U-Net with EffNetB7 [31]	73.76	62.76
U-in-EffNet (Proposed)	**74.25**	**63.35**

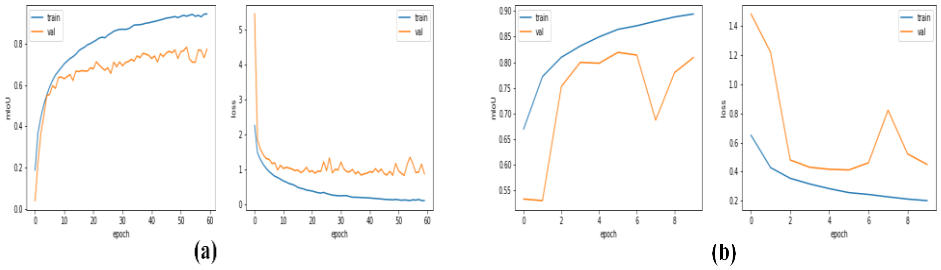

Figure 6. Graphical representation of the result (a) Pascal VOC2012 (b) IDD-Lite

4. Conclusion

Pixel level classification of images from real life scenarios are often unstructured contains many minority class which is very challenging task for any state of the art technique to achieve higher mIoU. Since CNN learns to map the data with its related identifying class only by seeing enough samples, it struggles with inadequate and similar samples. The proposed Unet architecture with EfficientNet as backbone benefits from its smart scaling capability. With the additional inception block in the decoder enables it to learn complex semantic level features while preserving the spatial relationships among the pixels. The architecture achieves higher mIoU for both visual and traffic images.

In future, the method needs to be evaluated on various challenging dataset belonging to different domains. Meantime, we can think about benefitting from knowledge gained from segmentation - towards its applications for automated annotation of images.

References

[1] T. Deng, "A Survey of Convolutional Neural Networks for Image Classification: Models and Datasets," 2022 International Conference on Big Data, Information and Computer Network (BDICN), 2022, pp. 746-749, doi: 10.1109/BDICN55575.2022.00145.

[2] Yujian Mo, Yan Wu, Xinneng Yang, Feilin Liu, Yujun Liao. "Review the state-of-the-art technologies of semantic segmentation based on deep learning", *Neurocomputing*, Vol. 493, 2022, pp 626-646, ISSN 0925-2312, doi: https://doi.org/10.1016/j.neucom.2022.01.005.

[3] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(4):640–651, April 2017.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. Unet: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, pages 234–241, Cham, 2015. Springer International Publishing.

[5] H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia, "Pyramid Scene Parsing Network," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6230-6239, doi: 10.1109/CVPR.2017.660.

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. 2018 Europian Conference on Computer Vision (ECCV), 2018.

[7] K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980-2988, doi: 10.1109/ICCV.2017.322.

[8] Parekh D, Poddar N, Rajpurkar A, Chahal M, Kumar N, Joshi GP, Cho W. A Review on Autonomous Vehicles: Progress, Methods and Challenges. *Electronics*. 2022; 11(14):2162. https://doi.org/10.3390/electronics11142162.

[9] Kugelman, J., Allman, J., Read, S.A. *et al.* A comparison of deep learning U-Net architectures for posterior segment OCT retinal layer segmentation. *Sci Rep* 12, 14888. 2022. doi: https://doi.org/10.1038/s41598-022-18646-2.

[10] Wang, Jinke et al. Automatic Liver Segmentation Using EfficientNet and Attention-Based Residual U-Net in CT. Journal of Digital Imaging. 2022. 35. 10.1007/s10278-022-00668-x.

[11] Mohan, R., Valada, A. EfficientPS: Efficient Panoptic Segmentation. Int J Comput Vis 129, 1551–1579 (2021). https://doi.org/10.1007/s11263-021-01445-z.

[12] B. Baheti, S. Innani, S. Gajre and S. Talbar, "Eff-UNet: A Novel Architecture for Semantic Segmentation in Unstructured Environment," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 1473-1481, doi: 10.1109/CVPRW50498.2020.00187.

[13] Lu Haoran, She Yifei, Tie Jun, Xu Shengzhou. "Half-UNet: A Simplified U-Net Architecture for Medical Image Segmentation". Frontiers in Neuroinformatics, Vol. 16, 2022, doi: 10.3389/fninf.2022.911679.

[14] Wu, J., Zhou, S., Zuo, S. *et al.* U-Net combined with multi-scale attention mechanism for liver segmentation in CT images. *BMC Med Inform Decis Mak.* 21, 283 (2021). https://doi.org/10.1186/s12911-021-01649-w.

[15] Ayalew, Y.A., Fante, K.A. Mohammed, M. "Modified U-Net for liver cancer segmentation from computed tomography images with a new class balancing method". *BMC biomed eng* **3**, 4 (2021). https://doi.org/10.1186/s42490-021-00050-y.

[16] Xiaomeng Feng, Taiping Wang, Xiaohang Yang, Minfei Zhang, Wanpeng Guo, Weina Wang. ConvWin-UNet: UNet-like hierarchical vision Transformer combined with convolution for medical image segmentation[J]. Mathematical Biosciences and Engineering, 2023, 20(1): 128-144. doi: 10.3934/mbe.2023007.

[17] Mingxing Tan, Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. Proceedings of the 36th International Conference on Machine Learning (ICML), 2019. http://arxiv.org/abs/1905.11946.

[18] Sandler, Mark, Howard, Andrew, Zhu, Menglong, Zhmoginov, Andrey, Chen, Liang-Chieh. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. 4510-4520. 10.1109/CVPR.2018.00474.

[19] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, Alexander A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI'17, page 4278–4284. AAAI Press, 2017.

[20] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman. The PASCAL Visual Object Classes Challenge 2011 (VOC 2011). Results. http://www.pascalnetwork. org/challenges/VOC/voc2011/workshop/index.html.

[21] The World Health Organization. "Global Status Report on Road Safety". 2018. Accessed on: Jan. 15, 2020.

[22] Varma, Girish, Subramanian, C.V.*et al.* "IDD: A Dataset for Exploring Problems of Autonomous Navigation in Unconstrained Environments". 1743-1751. 2019. 10.1109/WACV.2019.00190.

[23] S. Gupta, P. Arbelaez, and J. Malik. Perceptual organization and recognition of indoor scenes from RGB-D images. In CVPR, 2013.

[24] Perez-Siguas, R. ., Matta-Solis, H. ., Matta-Solis, E. ., Matta-Perez, H. ., Cruzata-Martinez, A. ., & Meneses-Claudio, B. . (2023). Management of an Automatic System to Generate Reports on the Attendance Control of Teachers in a Educational Center. International Journal on Recent and Innovation Trends in Computing and Communication, 11(2), 20–26. https://doi.org/10.17762/ijritcc.v11i2.6106

[25] B. Hariharan, P. Arbel´aez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In European Conference on Computer Vision (ECCV), 2014.

[26] S. Mehta, H. Hajishirzi and M. Rastegari, "DiCENet: Dimension-Wise Convolutions for Efficient Networks" in IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 44, no. 05, pp. 2416-2425, 2022. doi: 10.1109/TPAMI.2020.3041871.

[27] Mehta, Sachin & Rastegari, Mohammad & Shapiro, Linda & Hajishirzi, Hannaneh. (2019). ESPNetv2: A Light-Weight, Power Efficient, and General Purpose Convolutional Neural Network. 9182-9192. 10.1109/CVPR.2019.00941.

[28] Liu, Wei, Rabinovich, Andrew & Berg, Alexander. (2015). ParseNet: Looking Wider to See Better.CoRR. http://arxiv.org/abs/1506.04579.

[29] 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," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834-848, 1 April 2018, doi: 10.1109/TPAMI.2017.2699184.

[30] Fisher Yu, Vladlen Koltun, and Thomas A. Funkhouser. Dilated residual networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 636– 644, 2017.

[31] Eduardo Romera, Jose M. Alvarez, Luis Miguel Bergasa, and Roberto Arroyo. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. IEEE Transactions on Intelligent Transportation Systems, 19:263–272, 2018.

[32] Wang Wei, Natural Language Processing Techniques for Sentiment Analysis in Social Media , Machine Learning Applications Conference Proceedings, Vol 1 2021.

[33] B. Baheti, S. Innani, S. Gajre and S. Talbar, "Eff-UNet: A Novel Architecture for Semantic Segmentation in Unstructured Environment," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 1473-1481, doi: 10.1109/CVPRW50498.2020.00187.