Eff-UNet: A Novel Architecture for Semantic Segmentation in Unstructured Environment

Bhakti Baheti* Shubham Innani* Suhas Gajre Sanjay Talbar
Center of Excellence in Signal and Image Processing, SGGSIET, Nanded, India
{bahetibhakti, 2016bec035, ssgajre, sntalbar}@sggs.ac.in

Abstract

Since the last few decades, the number of road causalities has seen continuous growth across the globe. Nowadays intelligent transportation systems are being developed to enable safe and relaxed driving and scene understanding of the surrounding environment is an integral part of it. While several approaches are being developed for semantic scene segmentation based on deep learning and Convolutional Neural Network (CNN), these approaches assume well structured road infrastructure and driving environment. We focus our work on recent India Driving Lite Dataset (IDD), which contains data from unstructured driving environment and was hosted as an online challenge in NCVPRIPG 2019. We propose a novel architecture named as Eff-UNet which combines the effectiveness of compound scaled EfficientNet as the encoder for feature extraction with UNet decoder for reconstructing the fine-grained segmentation map. High level feature information as well as low level spatial information useful for precise segmentation are combined. The proposed architecture achieved 0.7376 and 0.6276 mean Intersection over Union (mIoU) on validation and test dataset respectively and won first prize in IDD lite segmentation challenge outperforming other approaches in the literature.

1. Introduction

Since the last few decades, the number of road fatalities has seen continuous growth across the globe. As stated by the World Health Organization (WHO) survey [30], road traffic accidents are the eighth leading cause of death with 1.3 million deaths each year and cause approximately 20-50 million injuries. The report also states that 74% of the total road traffic deaths are from middle income countries [30]. The statistics show that the roads are becoming more deadly in developing countries [29] having unstructured driving environment like ambiguous road boundaries, unmarked or incompletely delineated lanes, wear and tear of road infrastructure, high within class diversity, less adherence to traffic rules, etc.

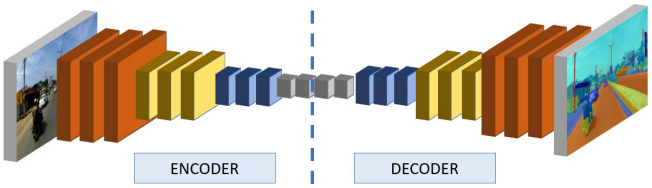

Figure 1: Design of Encoder-Decoder type semantic segmentation architecture based on CNN

Nowadays, rapid research is happening towards development of intelligent vehicles for safe and relaxed driving. Advanced Driver Assistance Systems (ADAS) are being developed with several active safety features that are based on semantic segmentation, pedestrian and vehicle detection, lane detection, etc. in the surrounding environment. Semantic segmentation is the process of assigning pre-defined label or class to each pixel of an image which is also known as pixel level classification. Various applications of it include medical imaging, robotics, self driving car etc. As intelligent vehicles must have an understanding and context of the surrounding environment for their safe integration on the current roads, semantic segmentation is a crucial element in the intelligent vehicles. Actually, the research towards building autonomous vehicles started way back in 1989, but the limitations of conventional neural networks and hardware resources restricted their progress. With the advancements in convolutional neural networks and GPU technologies, the development of intelligent vehicles has been accelerated. Though a lot of research is happening in this era, it remains a challenging problem due to wide variations in geographies.

Semantic scene segmentation has witnessed tremendous progress recently and has become a trending research domain with deep learning. Figure 1 shows the general design

*BB and SI have contributed equally towards this research work

2160-7516/20/$31.00 ©2020 IEEE
DOI 10.1109/CVPRW50498.2020.00187

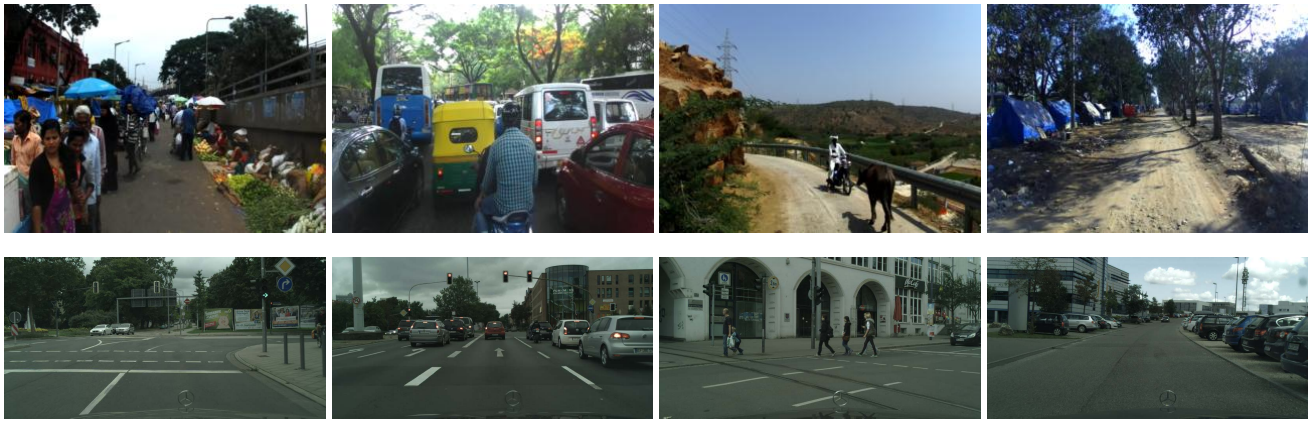

Figure 2: Few example road traffic images in different scenarios. Sample images from the IDD dataset [31] in unstructured environment are shown in first row and sample images from the Cityscapes dataset [10] having structured environment are shown in the second row

of CNN based encoder decoder type semantic segmentation framework. Despite of the several advancements in the last few years, scene understanding in complex real-world scenarios yet remains a challenging task compared to human level performance [2]. Prior to rise of CNN based approaches, semantic image segmentation algorithms were based on hand-crafted features and classical classifiers. But since the CNNs have proved effectiveness in image classification, they are being used as the backbone for feature extraction in semantic segmentation framework. Convolutional Neural Networks progressively reduce the input image resolution by factor of 32 to obtain the high level feature map representing the original image. Such small feature map is suitable for image classification where only one dominant object is present in the image and CNNs have surpassed human level performance in this task of image classification. But the performance of CNN shatters for segmentation task as the spatial information useful for analysis of complex scenes in the image is lost in the tiny feature maps. To prevail over this problem, we propose to combine the effectiveness of EfficientNet [28] as an encoder for high level feature extraction along with UNet [20] decoder for generating fine segmentation maps.

Various popular datasets in literature used for evaluating performance of semantic segmentation assume well structured environment like in Europe and North America. But such environment doesn't exist in most of the other parts of the world. India Driving Dataset (IDD) [31] is the world's first dataset of unstructured driving scenarios. IDD captures unstructured driving environments with higher uncertainties and ambiguities. Some sample images from the Cityscapes [10] and IDD dataset are shown in Figure 2. IDD Lite [16] semantic segmentation challenge was launched as a part of Seventh National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG 2019) in India. (https://cvit.iiit.ac.in/ncvpripg19/idd-challenge/). Performance on unknown test data was evaluated by submitting results to their online server. Our proposed approach achieved benchmark performance in the leader board winning first prize in the challenge.

The rest of the paper is structured as follows. Recent and relevant literature in the domain of semantic segmentation along with the typical datasets is reviewed in Section 2. The IDD Lite dataset used in this research is described in Section 3. The proposed approach is illustrated in Section 4. Results are discussed in Section 5 and Section 6 concludes the paper with future scope.

2. Related Work

The literature in the field of semantic segmentation can be categorized into two categories. In general, the approaches before the evolution of deep learning are called classical methods or traditional methods. With these conventional methods, semantic scene segmentation was a very difficult task. They mainly focused on handcrafted features like SIFT, SURF, HoG, and classifiers like random forest classifiers [23][4], conditional random fields [26] etc. These methods explicitly clustered or grouped the pixels while classifying the patch category. However, these methods were dependent on handcrafted features rather than knowing the structure of the data to perform pixel-level classification. On the other hand, the conditional random field for a structured prediction problem was the right approach. In [26] [13], a CRF based solution was provided for the segmentation task. Classification of pixels based on another method known as boosting was also presented in [13][24]. But it was observed that the performance of segmentation

with these classical methods was limited.

Since 2012, various Convolutional Neural Network based architectures like VGG16 [25], ResNet [11], MobileNet [12], Xception [9], and recently introduced EfficientNet [28] have been evolving and have established benchmarks in image classification. The EfficientNet, as proposed in [28], consists of the compound coefficient which studied model scaling and adjusted the depth, width, and the resolution of the network for better performance. The field of semantic segmentation also witnessed very significant progress recently by the use of these CNNs as feature extractor. One of the initial efforts for semantic segmentation using CNN was based on Fully Convolutional Neural Network (FCN) [22]. This VGG16 [25] based architecture achieved significant improvement over classical methods, but pixel accuracy was bounded because of coarse output pixel map. FCN was the first work that introduced CNN in the field of semantic segmentation. This FCN based method was fed with full image for denser predictions. The idea of deconvolutional network and unpooling layer was introduced in [17]. To overcome the drawback of coarse segmentation outputs, Badrinarayanan *et al*. proposed an encoder-decoder type module known as SegNet [1], which consists of VGG16 as the feature extractor. This decoder module uses maxpooling indices from the encoder while the upsampling the feature maps in the decoder. With the powerful CNN architectures emerging, multiscale semantic segmentation algorithm was proposed in [32][17]. With end-to-end learning, the feature maps from different resolutions were merged using skip-net architecture. Since the CNNs progressively downsample the original image resolution, the loss of spatial information of small and thin objects hampers the performance. To overcome this problem, the concept of dilated convolution was introduced in [32][17] to increase the resolution of the feature map while maintaining the receptive field of the neuron. Yu *et al*. proposed dilated residual networks, which removed the problem of gridding artifacts [33].

DeepLabV1 [6] and DeepLabV2 [5] used the state-of-art CNNs as feature extractor along with dilated convolutions and fully connected Conditional Random Field (CRF). The concept of Atros Spatial Pyramid Pooling (ASPP) was introduced in DeepLabV3 [7]. An effective decoder module was introduced in DeepLabV3+ [8] which further improved the results of DeepLabV3. Romera *et al*. proposed ERFNet (Efficient Residual Factorized Network) that used factorized convolution with residual connections [19]. In [14], ParseNet is proposed which combined global average pooling and L2 normalization. PSPNet is proposed by Zhao *et al*. which used a pyramid pooling module on the last layer feature map [35]. Segmentation models like ENet [18], ICNet [34] are useful in real-time applications. Ronneberger *et al*. [20] proposed a U-shaped fully convolutional network architecture where upsampling and concatenation of features maps from a different encoder layers and decoder layers occurred.

There are various typical datasets for evaluating semantic segmentation architectures like Cityscapes [10], KITTI [15], CamVid [3]. These datasets mainly focus on scene understanding in the urban street scenes like in Europe or the USA where the road infrastructure and driving environment are well structured. They have well-defined lanes, low density of traffic participants, fewer variations in the objects and background and traffic rules are strictly followed. But in many other parts of the world like Asia and Africa, there is minimal possibility of such a structured environment. We mainly work on the recent IDD lite dataset [16] prepared from IDD [31], which focuses primarily on scene understanding in an unstructured environment. This dataset consists of high within-class diversity, no strict adherence to traffic rules, new classes such as drivable areas besides roads, a new type of vehicle like trucks, auto-rickshaws, high density of traffic etc. The development of intelligent vehicles in such an unstructured environment is a highly challenging and demanding task.

3. Dataset Description

We evaluate the results of the proposed architecture on India Driving Dataset (IDD), which is the only dataset to best of our knowledge focusing on an unstructured environment having less developed road infrastructure. The dataset has a mixture of urban and rural scenarios. The IDD Lite semantic segmentation online challenge was hosted as a part of the 7^{th} National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG) in December 2019 at Hubli, India. The challenge used IDD-Lite, a subsampled version of IDD for resource constrained training, deployment and architecture search with similar label statistics as IDD and less number of labels. The motivation behind IDD Lite is to create a small scale segmentation dataset which is useful in training image segmentation models in resource constrained environment like lack of memory resources and high end GPU. Such small scale dataset gives descent quantitative and qualitative results enabling fast prototyping on low resource hardware. The labels and scenes in IDD dataset are quite different from Cityscapes, CamVid and KITTI datasets. Complex obstructions might be present in the unstructured environment. Challenges in the unstructured environment are ambiguous road boundaries that have muddy terrain, which is also drivable. The diversity of vehicles and pedestrians is high in IDD and they can be found on the road at any arbitrary location, with no strict adherence to traffic rules. There are information boards in

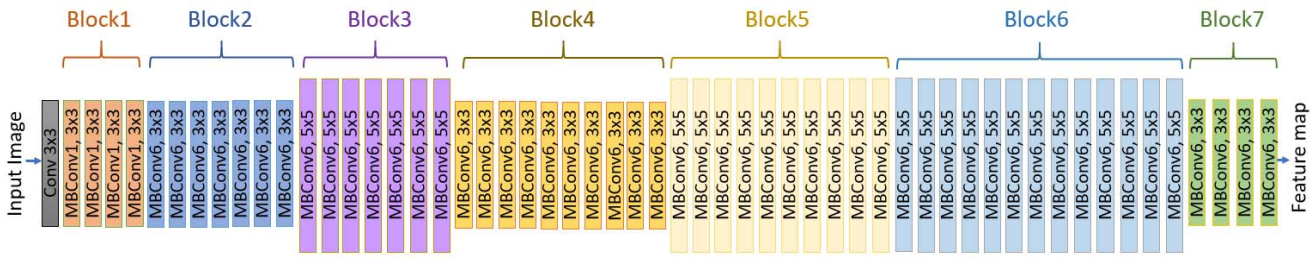

Figure 3: Architecture of EfficientNetB7 with MBConv as basic building blocks. The overall architecture can be divided into seven blocks which are shown in different colours. The basic building block of the network is MBConv (mobile inverted bottleneck convolution). Each MBConv**X** block is shown with the corresponding filter size and the **X**=1 and **X**=6 denote the standard ReLU and ReLU6 activation function respectively

the India driving dataset which have advertisements and other information related to the nearby area. Lightening condition includes variations like mid-day, dawn and dusk.

IDD-Lite is less than 50MB in size, has the resolution of 320×227 and contains 7 classes. It is composed of 1404 training images, 204 validation images and 408 testing images. It captures the complexity and diversity of Indian conditions with 7 carefully chosen classes viz: Drivable, Non drivable, Living things, Vehicles, Road Side Objects, Far objects and Sky. The segmentation results on the unknown test data were submitted to the online server for fair evaluation.

4. Technical Approach

In this section, we will briefly discuss the encoder-decoder architecture for semantic segmentation, EfficientNet which is used as feature extractor in encoder and UNet decoder.

4.1. Encoder-Decoder Architecture

A simple encoder-decoder network for semantic segmentation is shown in Figure 1. Usually, the encoder-decoder module consists of an encoder comprising of a CNN which extracts the features from original image. It progressively downsamples the image and reduces feature map resolution to capture high-level details of the image. This is done with the state-of-the-art convolutional neural networks like ResNet [11], MobileNet [12], InceptionResNetV2 [27] etc. Typically these CNN architectures progressively reduce the input resolution of the image to obtain the final feature map. It is challenging to rebuild the segmentation map of size of the original image from the smaller feature map. The decoder module consists of a set of layers that upsamples the feature map of encoder to recover spatial information.

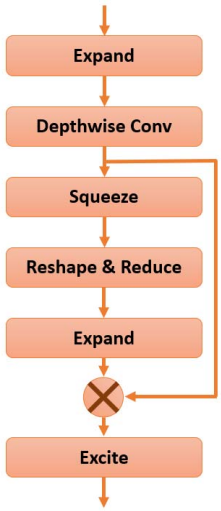

Figure 4: MBConv: Basic building block of EfficientNet

4.2. EfficientNet

Development of CNN architectures depend on the available resources, and then the scaling occurs to achieve improved performance when there is an increase in resources. For example, ResNet18 can be scaled to ResNet101 by adding more number of layers. The traditional practice for scaling the model is upsurge the CNN width or depth or the input image resolution and had been done arbitrarily. This practice involves tedious manual tuning and yet yields sub-optimal performance sometimes. Tan *et al*. [28] proposed a novel compound scaling method which uniformly scales the network depth, width and resolution for improved performance with a fixed set of scaling factors. A new baseline architecture called EfficientNetB0 was designed initially and it is scaled up to generate family of EffcientNet by compound scaling method. Powered by this approach, there are eight variants of the EfficientNets, namely EfficientNetB0 to EfficientNetB7. Scaling the

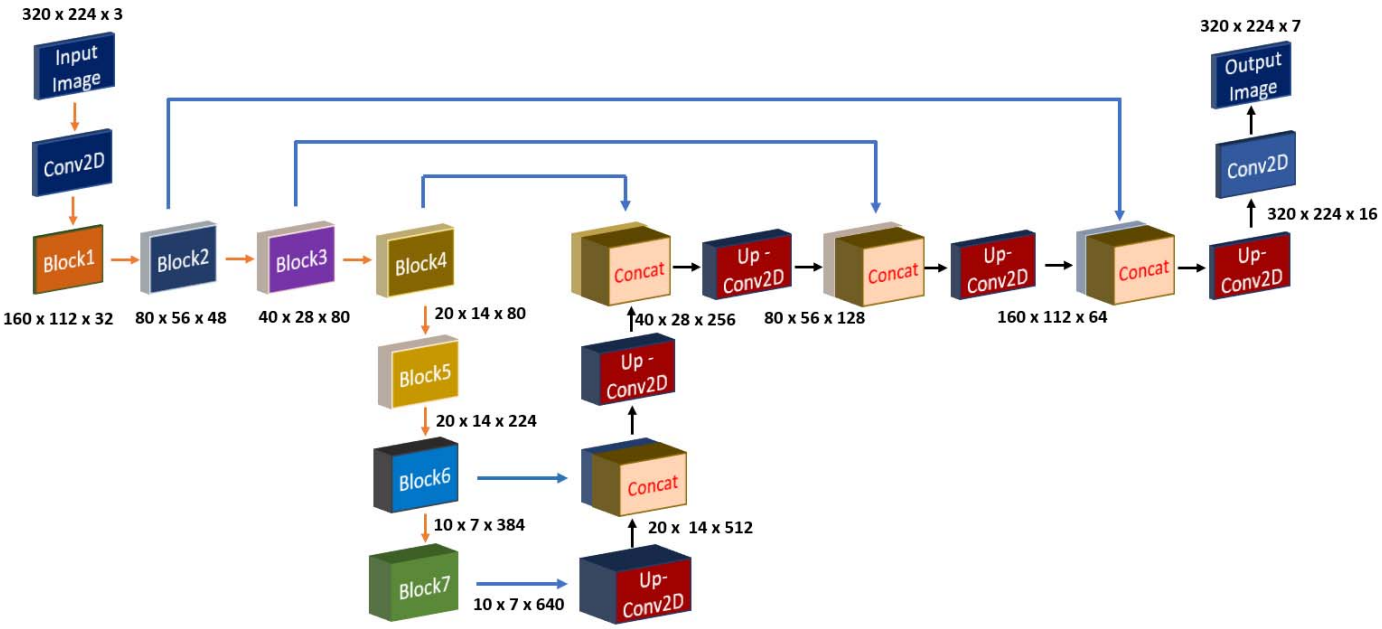

Figure 5: Architecture of proposed Eff-UNet with EfficientNetB7 framework for semantic segmentation. The details of Block1-7 are shown in Figure 3. The decoder consists of a sequence of Upconvolution and Concatenation layers to obtain the segmentation map

network systematically improves model performance balancing all compound coefficients of the architecture width, depth, and image resolution.

The basic building block of the EfficientNet architecture is mobile inverted bottleneck convolution (MBConv) [21] with squeeze and excitation optimization. Idea of MBConv is shown in Figure 4. The family of EfficientNet network has different number of these MBConv blocks. As we go from EfficientNetB0 to EfficientNetB7, depth, width, resolution and model size goes on increasing and the accuracy is also improved [28]. The best performing model EfficientNetB7 outperforms earlier state-of-art CNNs in terms of accuracy on ImageNet, and is also 8.4× smaller and 6.1× faster than the best existing CNN [28]. Network architecture of EfficientNetB7 is illustrated in Figure 3. It can be divided into seven blocks based on filter size, striding and number of channels. In our research, we used EfficientNetB5, and EfficientNetB7 as an encoder with UNet decoder and achieved best performance with EfficientNetB7.

4.3. UNet Decoder

UNet is a symmetric U shaped fully convolutional neural network originally developed for biomedical image segmentation [20]. UNet has two paths. The first path is contraction path also called as encoder which is basically a stack of convolution, activation and pooling layers to capture the context in the input image. The output of encoder which is smaller than the input, is progressively expanded in the expansion path. The expansion path or the decoder enables precise localization with transposed convolutions. The expansion pathway combines the high level features and spatial information by a sequence of upconvolutions and concatenation with corresponding feature maps from the contracting path. As the low level feature maps from encoder carry better spatial information useful in the analysis of complex scenes having multiple objects and their relative configuration, intermediate low level feature maps from Efficientnet and intermediate high level feature maps from UNet decoder are combined. The large number of feature channels in the upsampling part allows the network to propagate context information to higher resolution layers.

In conventional UNet, expansion path is nearly symmetric to the contracting path. In our work, we propose to use EfficientNet as an encoder in contracting path instead of conventional set of convolution layers. The decoder module is similar to the original UNet. Details of the proposed architecture are illustrated in Figure 5. The original input image size is 320×227 but we resized the images to 320×224 for further processing. The number of levels, resolution and number of channels of each feature map is also shown in the Figure 5. The detailed architecture of blocks in encoder can be found in Figure 3. We first bilinearly upsample the feature map of last logit of the

Table 1: Result comparison in terms of mean Intersection over Union (mIoU) on IDD Lite validation and test dataset

Network Architecture	Validation mIoU	Test mIoU
Dilated ResNet18 [33]	0.5503	-
ERFNet [19]	0.6614	-
DeepLabV3+ with ResNet18 Encoder	0.6304	0.5614
DeepLabV3+ with ResNet50 Encoder	0.6425	0.5733
UNet with ResNet34 Encoder	0.6781	0.6009
UNet with ResNet50 Encoder	0.6859	0.6076
UNet with InceptionResNetV2 Encoder	0.7247	0.6175
UNet with EfficientNetB5 Encoder	0.7072	0.6087
UNet with EfficientNetB7 Encoder	**0.7376**	**0.6276**

Table 2: Classwise IoU on IDD Lite validation dataset with EfficientNetB7 and UNet decoder

C1: Drivable	C2: Non Drivable	C3: Living Things	C4: Vehicles	C5: Roadside Objects	C6: Far Objects	C7: Sky	**mIoU**
0.9486	0.5012	0.6196	0.8131	0.5499	0.7754	0.9555	**0.7376**

encoder by factor of two and then concatenate with the feature map from encoder having same spatial resolution. It is followed by 3×3 convolution layers before again upsampling by the factor of two. The process is repeated till the segmentation map of size equal to input image is reconstructed. The proposed architecture is asymmetric unlike the original UNet. Here, the contracting path is deeper than the expansion path. Inclusion of powerful CNN like EfficientNet as encoder improves overall performance of the algorithm.

5. Results and Discussion

The proposed semantic segmentation architecture is developed in Tensorflow 2.0. The network is trained on image size 320×227 by resizing it to 320×224 , with a batch size of 4 and for 10 epochs. The pre-trained weights of EfficientNetB7 on ImageNet are used for initialization in encoder and they are further finetuned. As the training data is small, we augmented the dataset with various transformations like brightness, contrast, saturation, shear etc. to prevent the model from overfitting. Training is carried out using ADAM optimizer with a learning rate of 0.0001 and sum of Jaccard and binary cross-entropy is used as loss function. The results are submitted to the online server which are evaluated in terms of mean Intersection over Union (mIoU). Intersection over Union (IoU) also called as Jaccard index for one class is computed as follows:

$$IoU = \frac{TP}{TP + FP + FN} \tag{1}$$

where, TP, FP, FN mean number of True Positive, False Positive and False Negative pixels respectively.

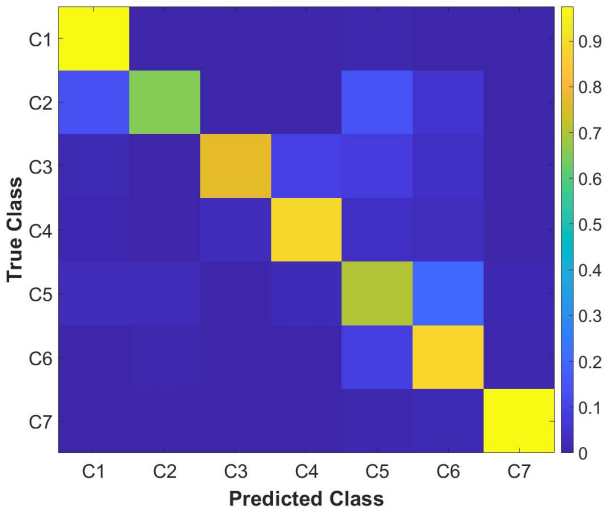

Figure 6: Confusion matrix on the IDD Lite validation dataset

We implemented various architectures for this task of semantic segmentation and the results are summarized in Table 1. Results with Dilated ResNet18 and ERFNet on validation data were released by the organizers as baseline performance [16]. Our experiment started with the DeepLabV3+ framework. Firstly, we used ResNet18 and ResNet50 as encoder for feature extraction. We carried out experiment with the legendary UNet decoder with different encoder backbones like ResNet34 [11], ResNet50 [11], InceptionResNetV2 [27], EfficientNetB5 [28] and EfficientNetB7 [28]. It can be observed from Table 1 that UNet decoder performs better than DeepLabV3+ decoder. It is because of more inclusion of low level features from

Figure 7: Result of semantic segmentation on IDD Lite validation dataset with proposed architecture. First column shows the input images depicting different scenarios from unstructured environment. Second and third column shows the ground truth and predicted segmentation map respectively where different colours signify different classes.

encoder in UNet which is useful in the analysis of complex scenes with multiple and dense objects. As the EfficientNetB7 model outperforms other CNN architectures for image classification [28], combining its effectiveness with UNet achieves benchmark performance which won the first prize in IDD Lite segmentation challenge.

We further analyze the classwise segmentation performance of the best performing model on the seven classes. As the online evaluation server returns a single value of mIoU over all the seven classes, we obtain classwise IoU on the validation dataset and is shown in Table 2. The confusion matrix for the same is also shown in Figure 6. It can be observed that IoU for few classes is low e.g. the non drivable area. This is because the notion of the drivable area is ambiguous in IDD dataset. Roadside area can have muddy terrain which can also be drivable to some extent. On the other hand, roads are also sometimes covered by dirt or mud making boundaries between drivable and non drivable area ambiguous. The road side objects class accounts for various objects like curb, wall, fence, guard rail, billboard, traffic sign, traffic light and pole. Their low IoU score is due to the low pixel count of these objects in the training data. Figure 7 shows the ground truth segmentation maps of few images along with their predicted segmentation maps. As the ground truth of test data is not available, these results are from validation dataset for visual interpretation. It should be noted that the proposed architecture achieves satisfactory performance in various scenarios like rural area, dense traffic etc.

6. Conclusion and Future Scope

The development of semantic segmentation architecture for complete scene understanding of the surrounding in intelligent transportation systems is a very challenging task especially in constrained and unstructured environment of developing countries. We have proposed a novel approach for pixel level segmentation on recent IDD Lite semantic segmentation challenge. In our research, we propose to use EfficientNet as the encoder feature extractor and decoder of UNet which incorporates both high level features and low-level spatial information together for precise segmentation. The EfficientNet maintains the compound scaling of the network, which achieves an improved performance. The proposed approach achieved first rank in the challenge which proves effectiveness of the proposed approach.

The IDD Lite dataset is a sub-sampled version of the full IDD dataset having more number of higher resolution images and classes. In future, we will evaluate the results of the proposed approach on the full IDD dataset. We hope that the similar performance of proposed approach can be translated to this larger dataset with high resolution images.

Acknowledgement

This publication is an outcome of the R & D work undertaken project under the Visvesvaraya PhD Scheme of Ministry of Electronics & Information Technology, Government of India, being implemented by Digital India Corporation with reference number: PhD-MLA/4(67/2015-16).

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:2481–2495, 2016. 3

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

[3] Gabriel J. Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recogn. Lett.*, 30(2):88–97, Jan. 2009. 3

[4] Gabriel J. Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla. Segmentation and recognition using structure from motion point clouds. In *Computer Vision – ECCV 2008*, pages 44–57, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. 2

[5] L. 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*, 40(4):834–848, April 2018. 3

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin P. Murphy, and Alan Loddon Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *ICLR*, abs/1412.7062, 2015. 3

[7] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017. 3

[8] 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. 3

[9] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, July 2017. 3

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 3

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016. 3, 4, 6

[12] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *ArXiv*, abs/1704.04861, 2017. 3, 4

[13] Lubor Ladicky, Chris Russell, Pushmeet Kohli, and Philip H. S. Torr. Associative hierarchical crfs for object class image segmentation. In *ICCV*, pages 739–746. IEEE Computer Society, 2009. 2

[14] Wei Liu, Andrew Rabinovich, and Alexander C. Berg. Parsenet: Looking wider to see better. *ArXiv*, abs/1506.04579, 2015. 3

[15] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3061–3070, June 2015. 3

[16] Ashutosh Mishra, Sudhir Kumar, Tarun Kalluri, Girish Varma, Anbumani Subramanian, Manmohan Chandrakar, and C V Jawahar. Semantic segmentation datasets for resource constrained training. *7th National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*, October 2019. 2, 3, 6

[17] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 1520–1528, Washington, DC, USA, 2015. IEEE Computer Society. 3

[18] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *ArXiv*, abs/1606.02147, 2017. 3

[19] 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. 3, 6

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 2, 3, 5

[21] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018. 5

[22] 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. 3

[23] Jamie Shotton, Matthew Johnson, and Roberto Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*. IEEE Computer Society, 2008. 2

[24] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *Computer Vision – ECCV 2006*, pages 1–15, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. 2

[25] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 3

[26] Paul Sturgess, Karteek Alahari, Lubor Ladicky, and Philip H. S. Torr. Combining appearance and structure from motion features for road scene understanding. In *BMVC*, 2009. 2

[27] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and 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. 4, 6

[28] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2019. 2, 3, 4, 5, 6, 8

[29] The Economist. "Roads are becoming more deadly in developing countries". 2017. Accessed on: Feb. 25, 2020. 1

[30] The World Health Organization. "Global Status Report on Road Safety". 2018. Accessed on: Jan. 15, 2020. 1

[31] Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and CV. Jawahar. Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019. 2, 3

[32] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *CoRR*, abs/1511.07122, 2016. 3

[33] 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. 3, 6

[34] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In *The European Conference on Computer Vision (ECCV)*, September 2018. 3

[35] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. *CoRR*, abs/1612.01105, 2016. 3