International Journal of Scientific Research and Engineering Development-– Volume 7 Issue 1, 2024

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Urban Scene Analysis using PSPNet and Efficient Channel Attention Mechanism for Autonomous Driving

Abdu Naim[1], Md Sohag Mia[2], Md Jabir Al Hujaifa[3]

(School of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing, China)[1,2]
(School of Information Engineering, Nanjing University of Information Science and Technology, Nanjing, China)[3]
Email: naimabdu, shuvo2018{@nuist.edu.cn}[1,2]
Email : mdjabiralhujaifa@gmail.com}[3]

Abstract:

In the realm of computer vision, semantic segmentation is a prominent topic. Scene parsing, a fundamental aspect of computer vision, involves segmenting images into semantic categories like sky, road, person, and more, providing a comprehensive understanding of the image. The challenge lies in assigning categories to each pixel, especially in diverse scenarios. This work introduces an enhanced Pyramid Scene Parsing Network (PSPNet) utilizing a proposed pyramid pooling module and an advanced scene parsing network. Leveraging global context information through region-based context aggregation, our model employs an Efficient Channel Attention (ECA) mechanism for improved understanding of urban scenarios. The improved PSPNet excels in pixel-level prediction, demonstrating superior performance on scene parsing challenges across diverse datasets. Validation on the Cityscapes Dataset yields impressive results, with the model achieving 72% mIoU in the training set and 67% mIoU in the validation set, showcasing its efficacy in urban scene analysis.

Keywords — PSPNet, Deep Convolutional Neural Networks, Semantic Segmentation, Scene Parsing, Computer Vision.

I. INTRODUCTION

Before the advent of deep learning, classical machine learning techniques like SVM, Random Forest, and K-means Clustering were used to solve the problem of image segmentation. But as with most image-related problem statements deep learning has worked comprehensively better than the existing techniques and has become a norm now when dealing with semantic segmentation. Being able to move efficiently and safely in driverless vehicles has been a hot research topic in recent years, and many companies and research centres are trying to come up with the first completely practical driverless car model. The purpose is to conduct real-time video segmentation tasks for scene interpretation since they have a direct impact on vehicle steering and braking for safer movements. The entire control mechanism of AVs is shown in figure 1. The initial strategy for visual scene understanding is semantic segmentation. This is a very promising field with a lot of possible benefits such as an increase in safety, fewer costs, comfortable travel, increased mobility, and reduced environmental footprint[1]. Semantic segmentation is the process of assigning each pixel of

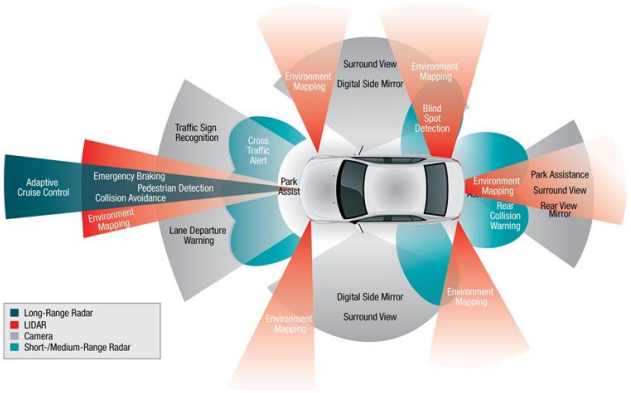

Fig 1. Autonomous Driving Control System Flowchart.

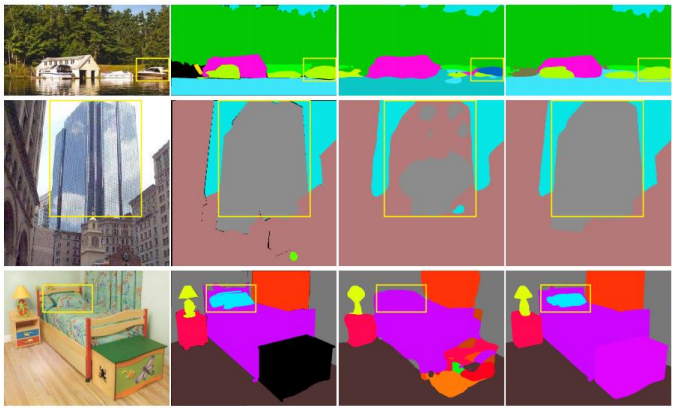

Fig 2. Scene parsing issues observed on the ADE20K dataset. The first row shows the issue of a mismatched relationship, i.e., cars go across the water far less frequently than boats. The second row displays categories of misunderstanding, such as "building" being easily mistaken with "skyscraper." The third row depicts classes that aren't visible. In this case, the pillow is quite similar in color and texture to the bedsheet. FCN is prone to misclassifying these inconspicuous things.

the received image to one of the predefined classes. These classes represent the segment labels of the image, e.g., roads, cars, signs, traffic lights, or pedestrians [2]. Therefore, semantic segmentation is sometimes referred to as “pixel-wise classification”. The main advantage of semantic segmentation is situation understanding. Scene understanding has various benefits in robotics applications [3] and the most prominent benefit is in autonomous driving [1], [4], [5]. To ensure the acquisition and recognition of the surrounding environment by the Autonomous Vehicle (AVs), the perception module of the self-driving system needs to obtain large amount of environmental information through various sensors (like Cameras, Lidar, Radars, etc.), including the status of the self-driving, traffic flow information, road conditions, pedestrians etc. Segmentation has also been used in medical applications and augmented reality [6]. The first prominent work in deep [2] semantic segmentation was fully convolutional networks (FCNs) [7], which proposed an end to-end method to learn pixel-wise classification. That method paved the road to subsequent advances in segmentation accuracy. Multi-scale approaches [8], context-aware models, and temporal models [9], introduced different directions for improving accuracy. All of the above approaches focused on the accuracy and robustness of segmentation. Although deep convolutional neural network (CNN)-based algorithms improve dynamic object perception, they encounter obstacles when dealing with a wide range of scenarios and a large vocabulary. This thesis reviewed numerous challenges for complex-scene parsing by looking at the prediction results of the FCN baseline supplied in ADE20K [10].

(i) Mismatched Relationships: Understanding complex scenarios relies on universal and crucial contextual interactions. Visual patterns, as seen in Figure 2's first row, can lead to misclassification when contextual data is lacking. For example, the FCN predicts a boat as a "car" in the yellow box, despite the rarity of cars crossing rivers.

(ii) Category Confusion: The ADE20K dataset presents challenging class label pairings, such as field and earth, mountain and hill, and various structures like wall, home, building, and skyscraper. Figure 2's second row illustrates FCN labeling an object as both a 'skyscraper' and a 'building,' highlighting the need to leverage category relationships for more accurate classification.

(iii) Inconspicuous Classes: Traditional FCNs overlook size differences in scene objects, leading to inconsistent predictions across scales. In Figure 2's third row, the pillow's similarity to the sheet highlights this issue. Ignoring the global scene category may omit the pillow. To improve performance for small or large objects, focus on sub-regions with inconspicuous-category data. The network's narrow receptive field, limiting attention to specific sub-regions while disregarding the overall scene category, contributes to the problem. The lack of contextual linkage, narrow receptive field, and limited global knowledge are key factors. To achieve accurate scene perception, it's crucial to correctly predict the image context, particularly when identifying a boathouse by a river. Existing FCN-based models face a challenge in utilizing global scene category hints. The traditional spatial pyramid pooling method, used for complex scene comprehension, lacked proper techniques. The proposed spatial pyramid pooling network (PSPNet) overcomes this by incorporating global properties, enhancing pixel-level functionality. The combination of local and global clues improves prediction accuracy, supported by a supervised loss optimization technique. This work outlines three key contributions, emphasizing the importance of establishing objectives before delving into the paper.

1. In an FCN-based pixel prediction framework, proposed a pyramid scene parsing network to include complex scenery context data.

2. Build an efficient deep ResNet optimization approach based on highly supervised loss, it is a kind of deeply supervised training strategy for training a very deep network.
3. Implemented Efficient Channel Attention (ECA) module with ResNet to teach our model what and where should focus on.

II. RELATED WORKS

An Autonomous driving has grown in popularity in recent years, and semantic segmentation has played a key part in detecting barriers and identifying road conditions. Traditional methods for pixel classification in images involve creating robust handcrafted features and using classifiers like Random Forest or boosting-based models. Post-processing techniques, such as conditional random fields (CRF), have been developed to enhance initial segmentation results and improve accuracy by reducing per-pixel prediction noise from classifiers. However, deep learning, particularly with deep convolutional neural networks (DCNN), has significantly advanced segmentation accuracy, outperforming traditional methods and achieving state-of-the-art performance across various visual tasks over the years. Semantic image segmentation deals with the assignment of class labels for every pixel of an image based on the class it belongs to [11]. It has multiple applications in the fields of medical imaging and autonomous vehicles. Segmentation has been widely used to classify biomedical images to segment neuron structures. Ronneberger et al. [12] introduced an encoder-decoder (U-Net) type of architecture for biomedical image segmentation to improve localization accuracy, and detect brain tumours [11], for the purpose of colon crypt segmentation, etc. In recent years, autonomous driving has gained much popularity and semantic segmentation has played an important role in perceiving obstacles and recognizing road conditions [1]. Some traditional methods focus on designing powerful handcrafted features and using random forest or boosting-based classifiers for predicting the class of image pixels. With the use of DCNN, it has been possible to achieve state-of-the-art performance for various visual tasks. Such as, by performing supervised training of a large network on the ImageNet dataset. Different deep architectures modified for training in different domains have been introduced since the advancement of deep learning-based segmentation methods. A network with a sliding window setup to predict pixel labels was suggested by [13] which was slow in processing and less accurate. Various other implementations involved the use of features from different layers of the architecture as discussed in [14]. Relevant work done by[15], to add fully-connected random fields to CNNs led to a significant upgrade in the segmentation performance. Various other approaches involving the use of a pyramid architecture to concatenate various feature maps also proved well. The DeepLab v1 and DeepLab v2 paved the way for DeepLab v3+ [16] which incorporates advanced elements from the previous two implementations. Besides, some recent developments have been shown in scene parsing and semantic segmentation areas. Pixel-level prediction tasks like scene parsing and semantic segmentation make significant progress thanks to strong deep neural networks [9], which were inspired by replacing the fully-connected layer in classification with the convolution layer [17]. [18] As discussed above, there are a variety of potential methods to be implemented for the task of semantic segmentation.

III. Proposed method

In this part we discussed about our proposed improved PSPNet model.

A. PSPNet Overall Architecture

With the pyramid pooling module, the proposed pyramid scene parsing network (PSPNet) [19] is illustrated in Figure 2. Given an input image in Figure 2(a), I use a pre-trained ResNet152 [20] model with the dilated network strategy [21] to extract the feature map. The final feature map size is 1/8 of the input image, as shown in Figure 2 (b). On top of the map, the pyramid pooling module is used to show in (c) to gather context information. Using a 4-level pyramid, the pooling kernels cover the whole, half of, and small portions of the image. They are fused as the global prior. Then concatenate the prior with the original feature map in the final part of (c). It is followed by a convolution layer to generate the final prediction map in (d). To explain the structure, PSPNet provides an effective global contextual prior for pixel-level scene parsing. The pyramid pooling module can collect levels of information, more representative than global pooling. In terms of computational cost, the proposed PSPNet does not much increase compared to the original dilated FCN network. In end-to-end learning, the global pyramid pooling module and the local FCN feature can be optimized simultaneously. First, given an input image,

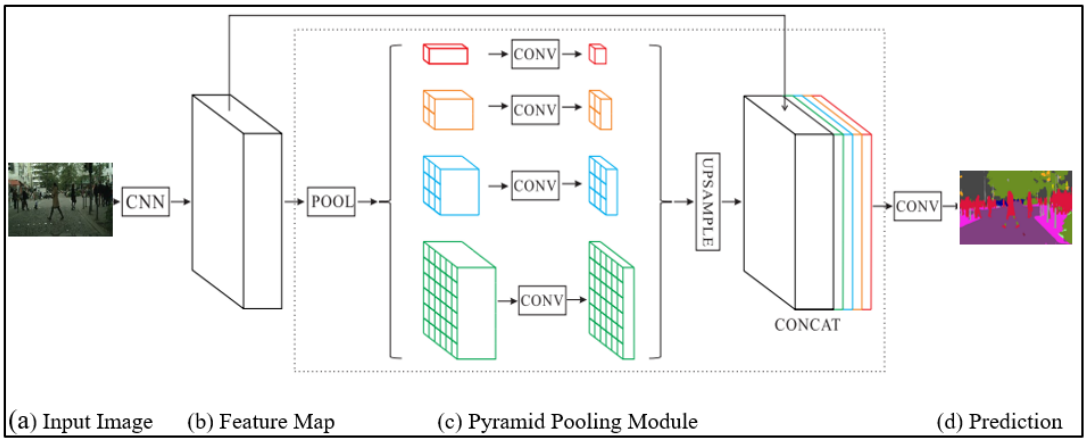

Fig 3. Architecture of the Pyramid Scene Parsing Network (PSPNet).

and then use CNN to get the feature map of the last convolutional layer, then a pyramid parsing module is applied to harvest different sub-region representations, followed by up-sampling and concatenation layers to form the final feature representation, which carries both local and global context information. Finally, the representation is fed into a convolution layer to get the final per pixel prediction.

B. Pyramid Pooling Module

A hierarchical global prior, containing information with different scales and varying among distinct sub-regions, is proposed to further decrease context information loss between different sub-regions. As shown in Figure 3, it's termed a pyramid pooling module for the global scene before the deep neural network's final-layer-feature-map is built. Under four distinct pyramid scales, the pyramid pooling module merges features. Global pooling to yield a single bin output is the coarsest level, depicted in red. The next pyramid level divides the feature map into sub regions and creates pooled representations for various places. The feature map with various sizes is produced by the pyramid pooling module at various levels. If the level size of the pyramid is N, a 1×1 convolution layer is employed after each pyramid level to decrease the dimension of context representation to $1/N$ of the original one to retain the weight of the global feature. Then, using bilinear interpolation, directly upsample the low-dimension feature maps to achieve the same size feature as the original feature map. Finally, as the final pyramid pooling global feature, multiple tiers of features are concatenated. The number of pyramid levels and the size of each level may also be changed. They have something to do with the size of the feature map given into the pyramid pooling layer. In a few strides, the structure abstracts distinct sub-regions by using varying-size pooling kernels. As a result, the multi-stage kernels should keep a suitable representation gap. The proposed pyramid pooling module has four levels, each having bin sizes of $1 \times 1, 2 \times 2, 3 \times 3$, and 6×6.

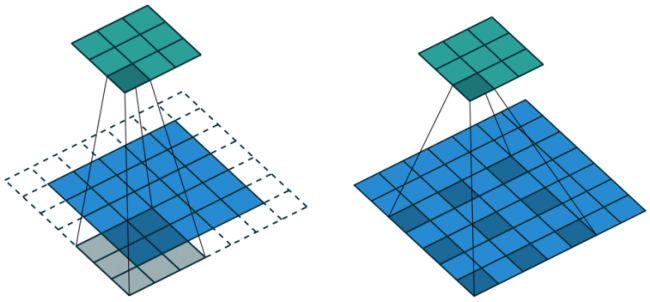

Fig 4. In left, Standard Convolution (l=1), in right side, Dilated Convolution (l=2).

C. Dilated Residual Networks (DRNs)

A unique dilated residual network is proposed as the backbone network in PSPNet in this paper. Convolutional networks for image classification gradually lower picture resolution until the image is represented by small feature maps with no discernible spatial organization. A loss of spatial sharpness like this might diminish image classification performance and make model transfer to downstream applications that need exact scene information more challenging. Dilation,

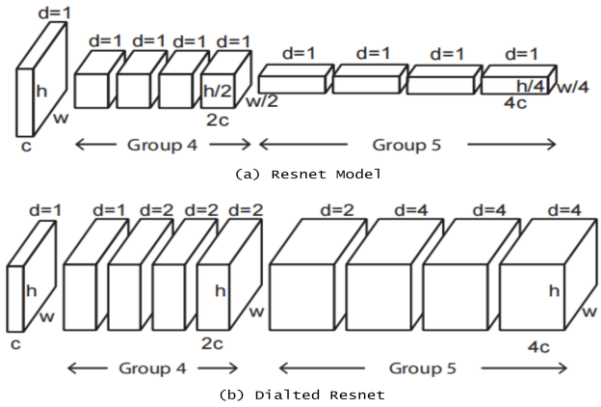

Fig 5. Converting a ResNet into a Dilated Residual network (DRN).

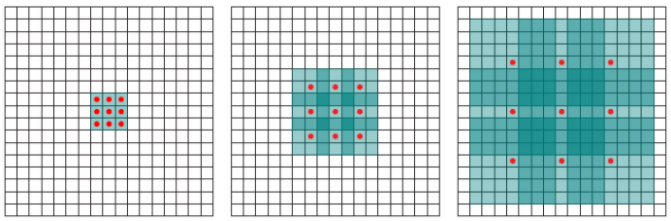

Fig 6. Diagram of Receptive Field Mechanism.

which raises the resolution of output feature maps without diminishing the receptive field of individual neurons, can help solve these challenges. It shows that dilated residual networks (DRN) [21] outperform their non-dilated counterparts in image classification without increasing the model's depth or complexity. then gridding artifacts introduced by dilation, develop an approach to removing these artifacts ('degridding'), and show that these further increases the performance of DRNs. In addition, it also shows that the accuracy advantage of DRN's is further magnified in downstream applications such as object localization and semantic segmentation. Here, equation 1 is Standard Convolution and equation 2 is Dilated Convolution.

$$(F \times k)(p) = \sum_{s+t=p} F(s)k(t) \quad (1)$$

$$(F \times lk)(p) = \sum_{s+lt=p} F(s)k(t) \quad (2)$$

Where, $F(s)$ = Input, $k(t)$ = Applied Filter, $*l = l-$dilated convolution, $(F \times lk)(p)$= Output.
The left one is the standard convolution. The right one is the dilated convolution. We can see that at the

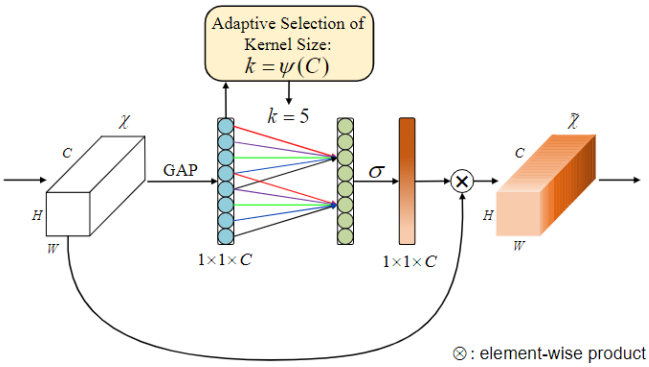

Fig 7. Diagram of our efficient channel attention (ECA) module. Given the aggregated features obtained by global average pooling (GAP), ECA generates channel weights by performing a fast 1D convolution of size k, where k is adaptively determined via a mapping of channel dimension C.

summation, it is $s + lt = p$ that we will skip some points during convolution. field is larger compared with the standard one.

D. Efficient Channel Attention (ECA)

To reduce computing expenses, a deep learning and computer vision method known as the Efficient Channel Attention (ECA) [22]module enhances feature maps inside particular channels. The ECA module has shown promise in reducing overfitting and enhancing the discriminative power of neural networks. It is flexible and works with many different network topologies, including deep residual networks and CNNs. The Efficient Channel Attention (ECA) module provides an appealing collection of features that distinguishes it as a notable addition to deep neural networks. ECA is notable for its lightweight and low computational overhead, making it ideal for real-time applications. Its dynamic channel-wise recalibration function enables networks to modify channel importance adaptively based on the input feature map, improving the model's ability to collect subtle information. Its versatility is a crucial feature since ECA can easily integrate into different CNN designs, allowing for easy experimentation to see how it affects overall performance. Furthermore, ECA's localized attention mechanism distinguishes itself by effectively capturing local channel-wise dependencies, with a focus on specific regions along the channel dimension. Importantly, it emphasizes efficiency without sacrificing the network's representational capability, all while

Fig 8: Real image frame (left) vs Ground Truth (middle) vs Predicted Segmentation (right).

retaining a less parameterized attention mechanism. The ECA module intends to improve the efficiency of the attention mechanism in CNNs, allowing models to be more performant and computationally efficient. Because of its versatility and lightweight, it is a preferred candidate for improving various CNN designs in computer vision. ECA module can be expressed by mathematically: The first step is to apply global average pooling on the feature map $Y \in R^{W\times H\times c}$ to gain the vector $Y_{avg} \in R^{1\times 1\times c}$ in order to aggregate the channel information.

$$Y_{avg} = GAP(y) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} Y_{i,j} \quad (3)$$

Where $GAP(.)$ apprises global average pooling. Y apprises the input feature map. H and W apprises the length and width of the feature map.

$$W = \sigma\big(C1D_k(Y)\big) \quad (4)$$

Where σ apprises the sigmoid activation function. $C1D$ apprises the one-dimensional convolution. k apprises convolution kernel size.

$$C = \emptyset(k) = 2^{(y\times k-b)} \quad (5)$$

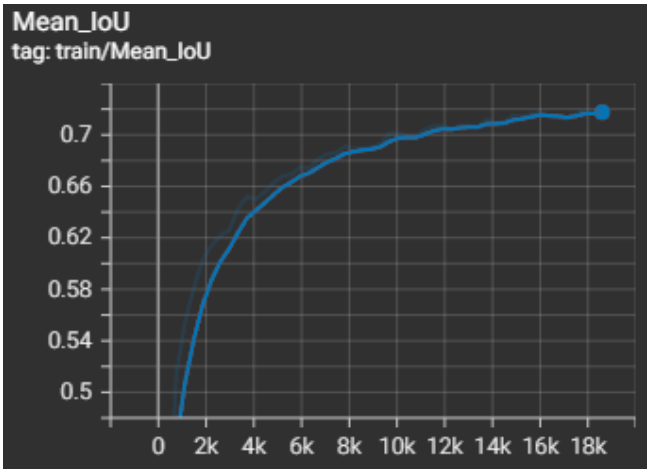

Fig 9: Training mIoU vs number of Training images.

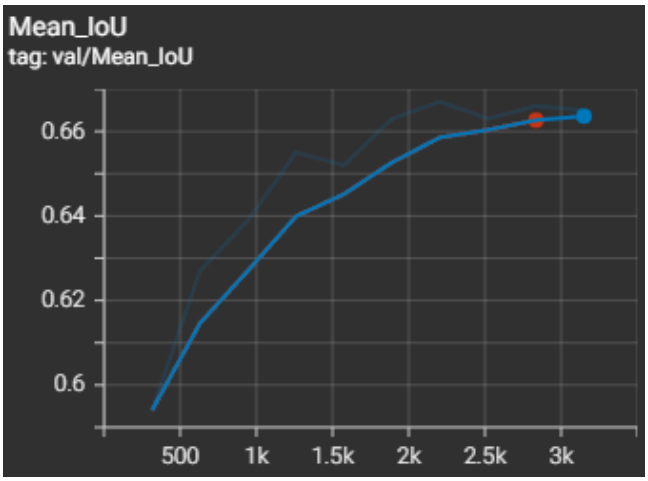

Fig 10: Validation mIoU vs number of Validation images.

Where C apprises the channel size of feature map. y & b apprises parameters to 2 and 1. k apprises convolution kernel.

$$k = \varphi(c) = \left|\frac{log_2(c)}{y} + \frac{b}{y}\right|_{odd}$$

Where $|t|_{odd}$ apprises nearest odd number of t.

IV. Experiments

In this section, we will demonstrate the experimental results and relevant topics.

A. Dataset, Evaluation Metrics and Experimental Setup

Cityscapes [23] is a semantic urban scene understanding dataset. It contains 5,000 high-quality

Table 1: Overall Training and Validation Result.

	Training Set	Validation Set
Accuracy	0.971	0.952
mIoU	0.72	0.67
Loss value	0.121	0.145

Table 3: Comparison of Results with Existing Methods.

Methods	FCN	Dilation	CRF-RNN	PSPNet
mIoU	58.3	60.3	57.6	65.8
Methods	DeepLab	UNet	UPerNet	**Ours**
mIoU	66.1	65.4	64.2	**67.0**

Table 2: Initialization Parameters Settings for Training.

Shape	Batch size	Momen -tum	LR	Epochs	Weight decay
400×380	8	0.9	0.01	50	$1e^{-4}$

pixel-level finely annotated images collected from 50 cities in different seasons. The images are divided into sets with numbers 2,975, 500, and 1,525 for training, validation, and testing. It defines 19 categories containing both stuff and objects. Also, 20,000 coarsely annotated images are provided for two settings in comparison, i.e., training with only fine data or with both the fine and coarse data. The dataset includes images taken in different seasons (spring, summer, fall), as well as under different weather conditions. Furthermore, all images have a fixed resolution of 2048×1024 pixels. The detection model we trained and tested on a workstation whose parameters were as follows: Intel(R) Xeon(R) Gold 6226R CPU@2.90GHz 2.89 GHz (2 processors) CPU, 128 GB DDR4 random-access memory (RAM), NVIDIA GeForce RTX 3090 with 24 GB VRAM GPU, and Ubuntu 20.04 OS. The initialization parameters of the network are shown in. To build a comparable basis, we focus on the same metrics as described in the work of [7]. IoU is defined as the ratio of intersection of ground truth and predicted segmentation outputs over their union. If we are calculating for multiple classes, the IoU of each class is calculated and their mean is taken.

B. Results and Discussion

The proposed PSPNet model is successful in scene parsing and semantic segmentation for traffic images as shown in figure 8. We evaluate it in Cityscapes datasets. I got the following result after implementing the said model. The pixel accuracy increases too since the more the images, the better can the model learn and generalize the objects in the scene. The training pixel accuracy was found to be about 97% whereas the validation pixel accuracy was more than 95%. It first starts training the network with a large learning rate and then slowly reducing/decaying it until local minima are obtained. The learning rate is reducing over time (represented with a green line), since the learning rate is large initially, we still have relatively fast learning toward as tending toward minima learning rate gets smaller and smaller, end up oscillating in a tighter region around minima rather than wandering far away from it. It is empirically observed to help both optimization and generalization.
The Loss versus several images in the training set as well as in the validation set respectively. Both graphs depict that as the number of images increases, the error loss decreases too since the more the images, the better can the model learn, and memorize, and thus fewer errors and loss is produced. The training loss was decreased to about 12% whereas the validation loss was about 14%. We also made sure no overfitting is occurring. The PSPNet model was finely tuned and debug before being implemented. Then it was trained and validated for 50 epochs which gave the following results including metric mean IoU (mean Intersection over Union over all class labels) Accuracy of the training images, validation images as well as each class label, at the end of the 50th epoch. Table 4 shows the IoU accuracy of each class label. The higher the accuracy, the more likely the objects are to be segmented accurately in the images and the videos. We tried adding important class labels as of now and more labels can be added in the future to give better scene understanding. Figure 9 and figure 10 show the Mean IoU versus the number of images in the training set as well as the validation set respectively. Both graphs depict that as the number of images increases, the mean IoU (average IoU of all the class labels) increases too since more the images, the better can the model learn and generalize the objects in the scene and thus can segment more accurately in the test images and videos. The training mean IoU was found to be more than 72% whereas the validation pixel accuracy was more than 67%. Now the following segmented results are found on test image frames. Figure 8 shows the result of the proposed PSPNet model on various image frames. The semantic segmentation yields good results for most objects in traffic scenes, with some misclassifications. Improvements can be achieved through model fine-

tuning, increased training data, and additional class labels. Accurate segmentation is observed for visible objects like cars, pedestrians, roads, buildings, and signs, but challenges arise with unlabelled objects and adverse weather conditions. Despite challenges, the overall outcome is satisfactory for traffic scene segmentation.

III. CONCLUSIONS

In conclusion, our enhanced PSPNet excels in pixel-level categorization for diverse urban scenes. The model, featuring a novel pyramid pooling module and advanced scene parsing network with ECA mechanism, demonstrates superior performance on Cityscapes Dataset, achieving 72% mIoU in the training set and 67% mIoU in the validation set. This research contributes significantly to semantic segmentation models, particularly in urban scene analysis.

ACKNOWLEDGMENT

We express our gratitude to the School of AI for generously providing GPU resources, enabling us to conduct the experiments and analysis essential to this research. This support played a crucial role in the successful implementation and validation of our enhanced PSPNet model.

REFERENCES

[1] Q. Sellat, S. Bisoy, R. Priyadarshini, A. Vidyarthi, S. Kautish, and R. K. Barik, "Intelligent Semantic Segmentation for Self-Driving Vehicles Using Deep Learning," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/6390260.

[2] W. Zhou, S. Lv, Q. Jiang, and L. Yu, "Deep Road Scene Understanding," *IEEE Signal Process. Lett.*, vol. 26, no. 4, pp. 587–591, Apr. 2019, doi: 10.1109/LSP.2019.2896793.

[3] V. Vineet *et al.*, "Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, May 2015, pp. 75–82. doi: 10.1109/ICRA.2015.7138983.

[4] C. Y. Lin, Y. C. Chiu, H. F. Ng, T. K. Shih, and K. H. Lin, "Global-and-local context network for semantic segmentation of street view images," *Sensors (Switzerland)*, vol. 20, no. 10, 2020, doi: 10.3390/s20102907.

[5] Z. Yang *et al.*, "Small Object Augmentation of Urban Scenes for Real-Time Semantic Segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 5175–5190, 2020, doi: 10.1109/TIP.2020.2976856.

[6] O. Miksik *et al.*, "The Semantic Paintbrush: Interactive 3D Mapping and Recognition in Large Outdoor Spaces," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 3317–3326. doi: 10.1145/2702123.2702222.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *CoRR*, vol. abs/1411.4, 2014, [Online]. Available: http://arxiv.org/abs/1411.4038

[8] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *4th Int. Conf. Learn. Represent. ICLR 2016 - Conf. Track Proc.*, 2016.

[9] A. Gurita and I. G. Mocanu, "Image segmentation using encoder-decoder with deformable convolutions," *Sensors*, vol. 21, no. 5, pp. 1–27, 2021, doi: 10.3390/s21051570.

[10] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Semantic Understanding of Scenes through the {ADE20K} Dataset," *CoRR*, vol. abs/1608.05442, 2016, [Online]. Available: http://arxiv.org/abs/1608.05442

[11] W. Weng and X. Zhu, "INet: Convolutional Networks for Biomedical Image Segmentation," *IEEE Access*, vol. 9, pp. 16591–16603, 2021, doi: 10.1109/ACCESS.2021.3053408.

[12] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11211 LNCS, pp. 833–851, 2018, doi: 10.1007/978-3-030-01234-2_49.

[13] T. Zhou, W. Wang, E. Konukoglu, and L. Van Goo, "Rethinking Semantic Segmentation: A Prototype View," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2022-June, pp. 2572–2583, 2022, doi: 10.1109/CVPR52688.2022.00261.

[14] P. Lu, S. Xu, and H. Peng, "Graph-Embedded Lane Detection," *IEEE Trans. Image Process.*, vol. 30, pp. 2977–2988, 2021, doi: 10.1109/TIP.2021.3057287.

[15] Z. Tian, C. Shen, H. Chen, and T. He, "{FCOS:} Fully Convolutional One-Stage Object Detection," *CoRR*, vol. abs/1904.0, 2019, [Online]. Available: http://arxiv.org/abs/1904.01355

[16] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2018, doi: 10.1109/TPAMI.2017.2699184.

[17] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, 2017, doi: 10.1109/TPAMI.2016.2644615.

[18] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 4510–4520, 2018, doi: 10.1109/CVPR.2018.00474.

[19] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 6230–6239, 2017, doi: 10.1109/CVPR.2017.660.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *CoRR*, vol. abs/1512.0, 2015, [Online]. Available: http://arxiv.org/abs/1512.03385

[21] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 636–644, 2017, doi: 10.1109/CVPR.2017.75.

[22] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," *CoRR*, vol. abs/1910.0, 2019, [Online]. Available: http://arxiv.org/abs/1910.03151

[23] M. Cordts *et al.*, "The Cityscapes Dataset for Semantic Urban Scene Understanding," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 3213–3223, 2016, doi: 10.1109/CVPR.2016.350.

