International Journal of Scientific Research and Engineering Development-– Volume 7 Issue 2, 2024

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Enhanced Semantic Segmentation for Autonomous Driving Using UNet and Deformable Networks

Abdullah Al Bary Voban[1], Md Sohag Mia[2], Sayed Masuk Ahmed[3]

(School of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing, China)[1,2]
(School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, China)[3]
Email: voban, shuvo2018{@nuist.edu.cn}[1,2]
Email : sayedmasuk@hotmail.com}[3]

Abstract:

Semantic segmentation stands as a significant subject in computer vision. Scene parsing, an essential facet, entails dividing images into semantic categories such as sky, road, person, and others, thereby offering a holistic comprehension of the image. The difficulty lies in categorizing each pixel, particularly in varied scenarios. This work introduces an enhanced encoder-decoder type model UNet. The U-Net consists of an encoder, which captures context by reducing input size, and a symmetric decoder for precise localization, our model employs an Efficient Channel Attention (ECA) mechanism for improved understanding of urban scenarios and also employs Deformable Convolutional Network (DCN) to understand various shaped objects. The enhanced UNet excels in pixel-level prediction, displaying superior performance in various semantic segmentation challenges. Notably, on the Cityscapes Dataset, the model attains a remarkable 74.3% mIoU in the training set and 69.8% mIoU in the validation set, underscoring its effectiveness in semantic segmentation.

Keywords — UNet, Autonomous Driving, Semantic Segmentation, Attention Network, Computer Vision.

I. INTRODUCTION

Prior to the rise of deep learning, classical machine learning methods such as SVM, Random Forest, and K-means Clustering were employed for image segmentation challenges. However, for most image-related tasks, deep learning has proven significantly more effective than traditional techniques and has become the standard approach for semantic segmentation. The pursuit of efficient and secure navigation in autonomous vehicles has been a focal point in recent research, with various companies and research centres striving to develop the first practical driverless car model. Real-time video segmentation is crucial for interpreting scenes, directly influencing the steering and braking of vehicles for safer movements. Figure 1 illustrates the entire control mechanism of autonomous vehicles. The primary approach to achieve visual scene understanding is through semantic segmentation, a highly promising field with potential

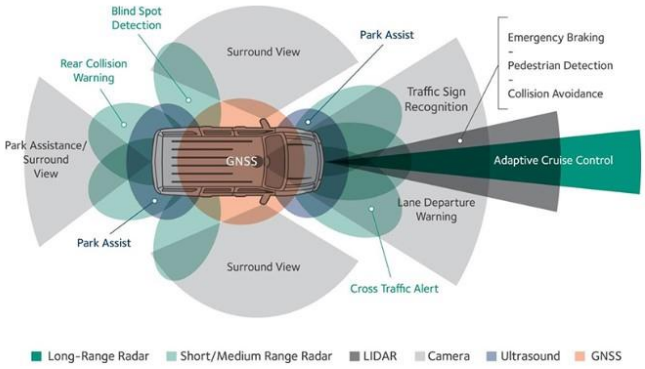

Fig 1. Autonomous Driving Control System.

benefits including enhanced safety, reduced costs, comfortable travel, increased mobility, and a decreased environmental footprint[1]. Semantic segmentation is the process of assigning each pixel of the received image

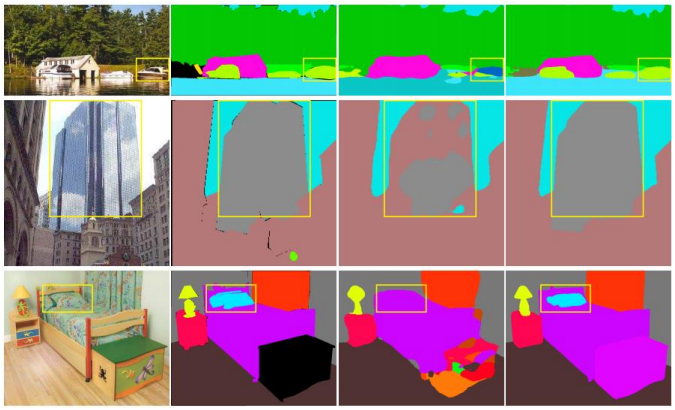

Fig 2. The ADE20K dataset highlights various scene parsing challenges. In the first row, a mismatched relationship is observed, where cars crossing water are less common compared to boats. The second row illustrates categories prone to confusion, like mistaking "building" for "skyscraper." In the third row, there are classes that remain unseen, such as a pillow blending with a bedsheet due to similar color and texture, a situation where FCN may misclassify such elements.

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]. Hence, semantic segmentation is often described as "pixel-wise classification." Its primary advantage lies in facilitating situation understanding. Scene comprehension offers various advantages in robotics applications. [3] and the most prominent benefit is in autonomous driving [1], [4], [5]. For Autonomous Vehicles (AVs) to effectively perceive and recognize their surrounding environment, the perception module of the self-driving system must gather extensive environmental data from various sensors such as cameras, LiDAR, radars, etc. This includes information on the vehicle's status, traffic flow, road conditions, pedestrians, and more. 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], This method introduced an end-to-end approach for pixel-wise classification, which subsequently paved the way for advancements in segmentation accuracy. Multi-scale approaches [8], context-aware models, and temporal models [9], introduced different directions for improving accuracy. The aforementioned approaches prioritized segmentation accuracy and robustness. Despite the improvement in dynamic object perception achieved by deep convolutional neural network (CNN)-based algorithms, challenges arise when handling diverse scenarios and a large vocabulary. This paper examined various challenges in parsing complex scenes by analyzing the prediction outcomes of the FCN baseline provided in ADE20K [10].

(i) Mismatched Relationships: Comprehending complex scenarios depends on essential contextual interactions. Visual patterns, portrayed in the top row of Figure 2, can lead to misclassification without contextual information. For example, despite cars rarely crossing rivers, FCN mistakenly categorizes a boat as a "car" within the highlighted yellow region.

(ii) Category Confusion: The ADE20K dataset includes challenging class label pairs, like field and earth, mountain and hill, and various structures such as wall, home, building, and skyscraper. In Figure 2's second row, FCN labels an object as both 'skyscraper' and 'building,' stressing the need to consider category relationships for better classification accuracy.

(iii) Inconspicuous Classes: Conventional FCNs fail to consider size differences among objects in scenes, resulting in inconsistent predictions across scales. In Figure 2's third row, the similarity between the pillow and the sheet highlights this issue. Neglecting the global scene category may lead to missing the pillow. To enhance detection for objects of varied sizes, target sub-regions with less prominent-category data. The network's narrow receptive field, focusing on specific sub-regions while ignoring the overall scene category, exacerbates the problem. The lack of contextual linkage, narrow receptive field, and limited global knowledge are key factors. Accurate scene perception requires correctly predicting image context, especially in 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 Improved encoder-decoder UNet overcomes this by incorporating global properties, enhancing precise localization 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 encoder-decoder based pixel prediction framework, reduces the spatial size of the

input to capture the context to include complex scenery context data.
2. Build an efficient deep ResNet optimization approach based on Deformable Convolutional Network, it helps the CNN model to detecting different size of objects in various objects and complex scenes.
3. Integrated Efficient Channel Attention (ECA) module with ResNet to guide our model what and where should focus on.

II. RELATED WORKS

Autonomous driving has gained popularity, with semantic segmentation playing a crucial role in barrier detection and road condition identification. Traditional pixel classification methods in images typically rely on creating strong handcrafted features and utilizing classifiers like Random Forest or boosting-based models. To refine initial segmentation outcomes and enhance accuracy by minimizing per-pixel prediction noise from classifiers, post-processing techniques like conditional random fields (CRF) have been introduced. Deep learning, especially with deep convolutional neural networks (DCNN), has notably improved segmentation accuracy, surpassing traditional methods and excelling in various visual tasks. Semantic image segmentation involves assigning class labels to each pixel based on its corresponding class 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]. Traditional methods emphasize designing robust handcrafted features and employing classifiers like random forest or boosting-based models to predict image pixel classes. The adoption of DCNN has enabled the attainment of state-of-the-art performance across diverse visual tasks. Supervised training on the ImageNet dataset, using large networks, has been a common approach. Various deep architectures tailored for specific domains have emerged alongside the progress 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 integrates advanced elements from prior implementations. Additionally, recent advancements in scene parsing and semantic segmentation have been notable. Tasks involving pixel-level prediction, such as scene parsing and semantic segmentation, have made significant strides due to robust deep neural networks [9], which were inspired by replacing the fully-connected layer in classification with the convolution layer [17]. [18] As mentioned earlier, several potential methods can be employed for semantic segmentation tasks.

III. Proposed method

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

A. UNet Overall Architecture

With the encoder-decoder module, the proposed UNet [11] is illustrated in Figure 2. Given an input image in Figure 2(a), I use a pre-trained ResNet152 [19] model with the dilated network strategy [20] to extract the feature map. the network architecture, which consists of 3×3 convolutions performed three times, then a Rectified Linear Unit (ReLU) and a 2×2 max pooling operation for downsampling. The number of feature channels doubles with each downsampling step. Upsampling is accomplished via 2×2 convolution (up-convolution), which reduces the number of feature channels supplied as input to that layer by half. It also has to skip connections that concatenate the encoder architecture's output (downsampling), as well as two 3×3 convolutions and a ReLU layer. These skip connections transfer localization data from the design's downsampling component to the upsampling section. The U-Net architecture uses an overlap-tile approach for training and prediction. The output size is smaller than the input size due to the use of unpadded convolutions. The input image is divided into patches of a size that the model can handle. Every patch's segmentation map is predicted by the model, which is then concatenated to

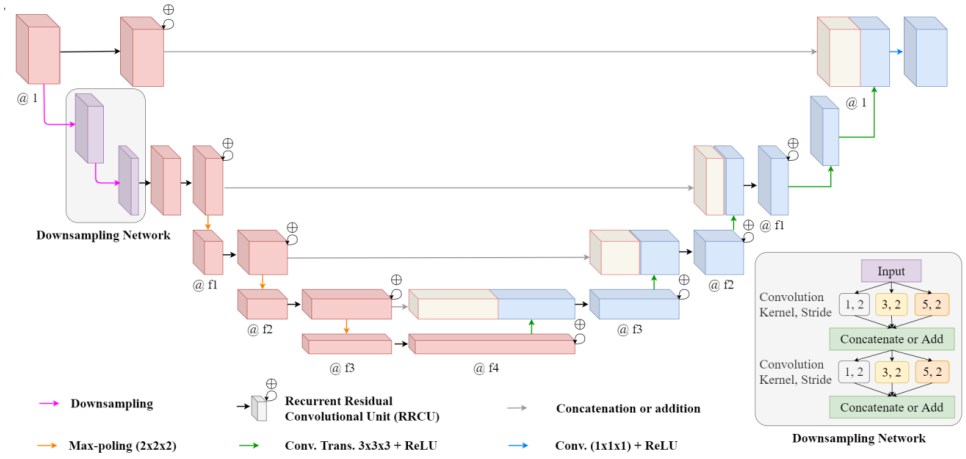

Fig 3. Architecture of the Encoder-Decoder Network (UNet).

generate the final segmentation output. U-Net also performs well with data augmentation as it can generalize well to random deformations applied to the input image and the corresponding output segmentation map the U-Net implementation also includes a weight map that is applied to the output of the network which is helpful in the separation of class boundaries touching each other. The weight map is computed as shown below:

$$w(x) = w_c(x) + w_0 . \exp\left(-\frac{(d_1(x) + d_2(x))^2}{2\sigma^2}\right)$$

Where $w_c : \Omega \to \mathbb{R}$ is the weight map to balance the class frequencies $d_1 : \Omega \to \mathbb{R}$ denotes the distance to the border of the nearest cell and $d_2 : \Omega \to \mathbb{R}$ the distance to the border of the second nearest cell.

B. Dilated Residual Networks (DRNs)

This paper introduces a distinctive dilated residual network as the foundational network in UNet. Typically, convolutional networks for image classification reduce image resolution gradually, resulting in small feature maps lacking clear spatial organization. Such loss of spatial sharpness can potentially reduce image classification accuracy and complicate model transfer to downstream applications requiring precise scene details. Dilation, which increases the resolution of output feature maps without compromising the receptive field of individual neurons, offers a solution to these challenges. It shows that dilated residual networks (DRN) [20] outperform their non-dilated counterparts in image classification without increasing the model's depth or

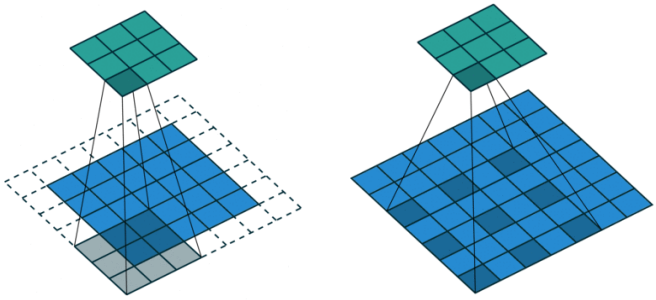

Fig 4. In left, Standard Convolution (l=1), in right side, Dilated Convolution (l=2).

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 summation, it is $s + lt = p$ that we will skip some

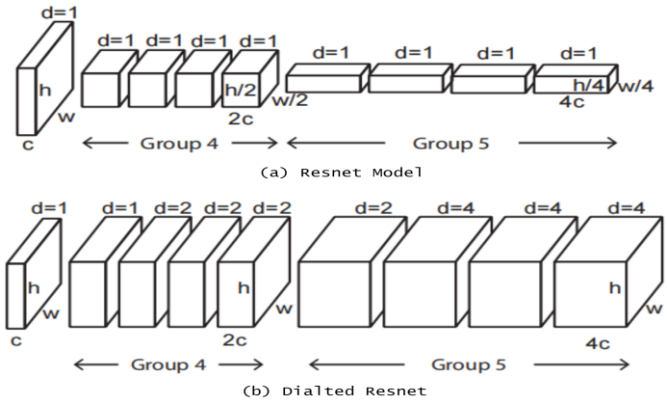

Fig 5. Converting a ResNet into a Dilated Residual network (DRN)

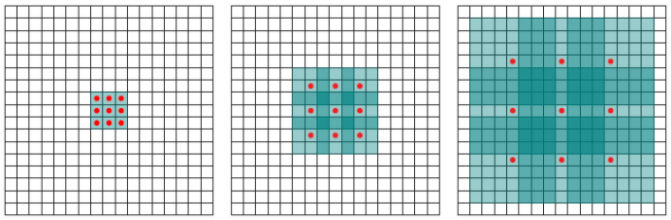

Fig 6. Diagram of Receptive Field Mechanism.

points during convolution. field is larger compared with the standard one.

C. *Efficient Channel Attention (ECA)*

To reduce computing expenses, a deep learning and computer vision method known as the Efficient Channel Attention (ECA) [21]module enhances feature maps inside particular channels. The ECA module exhibits potential in mitigating overfitting and enhancing the discriminative capacity of neural networks. It is adaptable and compatible with various network architectures, including deep residual networks and CNNs. Noteworthy for its lightweight nature and minimal computational burden, the Efficient Channel Attention (ECA) module stands out as a valuable addition to deep neural networks, particularly for real-time applications. Its dynamic channel-wise recalibration function enables networks to adjust channel importance according to input feature maps, thereby enhancing the model's ability to capture subtle details. Its versatility is another key attribute, as ECA seamlessly integrates into different CNN designs, facilitating straightforward experimentation to evaluate its impact on overall performance. Moreover, ECA's localized attention mechanism stands out for effectively capturing local channel-wise dependencies, particularly

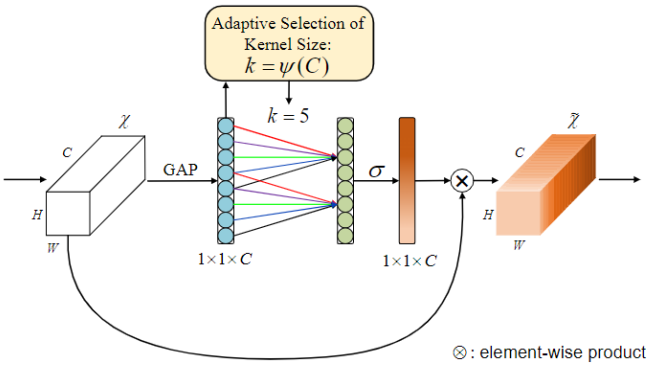

Fig 7. Diagram of our efficient channel attention (ECA) module. Given the aggregated features obtained by global average pooling (GAP), ECA generates channel weights by performing a fast 1D convolution of size k, where k is adaptively determined via a mapping of channel dimension C.

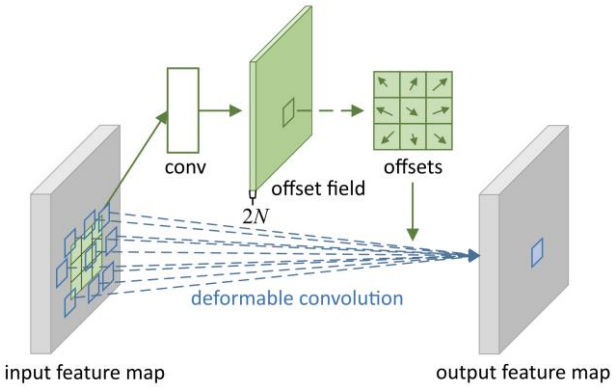

Fig 8. Architecture of Deformable Network.

focusing on specific regions along the channel dimension. Crucially, it prioritizes efficiency without compromising the network's representational capacity, all while maintaining a less parameterized attention mechanism. The primary goal of the ECA module is to enhance the efficiency of attention mechanisms in CNNs, thereby enabling models to be more effective and computationally economical. Due to its versatility and lightweight nature, it is a favoured option for enhancing diverse CNN designs in computer vision tasks. 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 $Y_{avg} \in R^{1\times1\times c}$ in order to aggregate the channel information.

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

Fig 9: Real image frame (left) vs Ground Truth (middle) vs Predicted Segmentation (right).

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

$$W = \sigma\big(C1D_k(Y)\big) \tag{4}$$

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

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

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.

D. Deformable Convolutional Network (DCN)

Deformable Convolutional Networks (DCNs) augment conventional convolutional neural networks (CNNs) by incorporating deformable convolutional layers, which dynamically adjust sampling positions

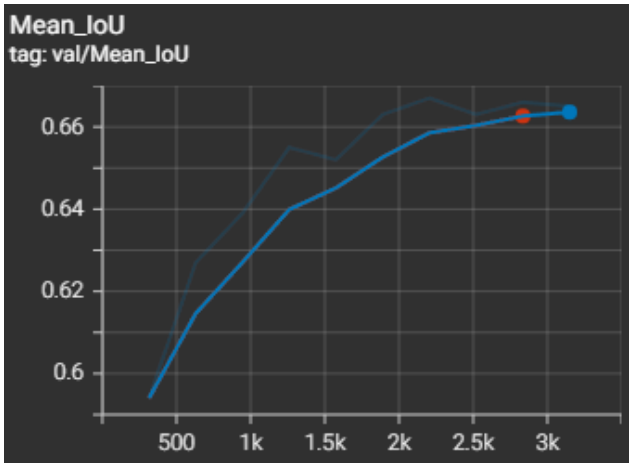

Fig 10: Training mIoU vs number of Training images.

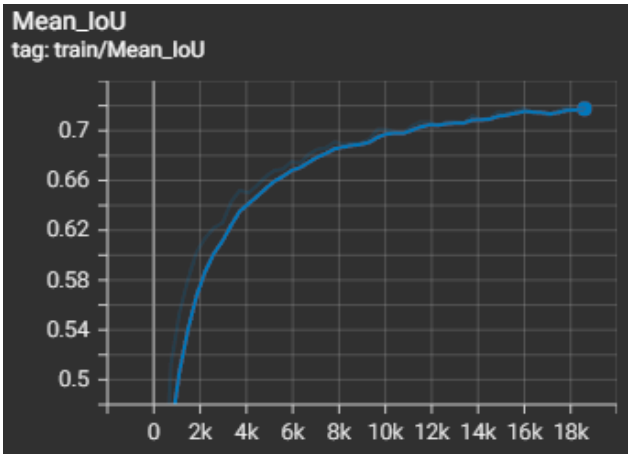

Fig 11: Validation mIoU vs number of Validation images.

based on learned offsets. In a deformable convolutional layer, the output feature map $Y_{i,j}$ at spatial location (i,j) is computed as:

$$Y_{i,j} = \sum_{m,n} X_{i+m+\Delta^x_{i+m,j+n},\, j+n+\Delta^y_{i+m,j+n}} . W_{m,n}$$

Here, X represents the input feature map, W denotes the convolutional filter weights, and $\Delta^x_{i+m,j+n}$ and $\Delta^y_{i+m,j+n}$ denote the learned offsets for each spatial location in the feature map. Both the filter weights and offsets are optimized during training using methods such as stochastic gradient descent (SGD). DCNs (Fig 8) exhibit enhanced performance in tasks like object detection and semantic segmentation, offering promising advancements in the field of computer vision.

Table 1: Overall Training and Validation Result.

	Training Set	Validation Set
Accuracy	0.977	0.964
mIoU	0.743	0.698
Loss value	0.118	0.142

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	**69.8**

Table 2: Initialization Parameters Settings for Training.

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

IV. Experiments

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

A. Dataset, Evaluation Metrics and Experimental Setup

Cityscapes [22] is a semantic urban scene understanding dataset. It contains 5,000 high-quality dataset comprises finely annotated images from 50 cities, divided into 2,975 training, 500 validation, and 1,525 testing images. It includes 19 categories, covering both stuff and objects, and offers 20,000 coarsely annotated images for comparison between training with fine data only and using both fine and coarse data. The dataset captures various seasons and weather conditions, providing a comprehensive benchmark for semantic segmentation tasks. 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 UNet demonstrates effectiveness in scene parsing and semantic segmentation for traffic images, as illustrated in Figure 9. Evaluation was conducted on Cityscapes datasets, yielding promising results. The pixel accuracy improved as the model had access to a larger volume of images, facilitating better learning and generalization of scene objects. The training pixel accuracy was found to be about 97.7% whereas the validation pixel accuracy was more than 96.4%. 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 curves for both training and validation sets illustrate a decreasing trend as the number of images increases, indicating improved model learning and memorization with larger datasets, leading to fewer errors and reduced loss. Specifically, the training loss decreased to approximately 12%, while the validation loss stabilized around 14%. To prevent overfitting, rigorous fine-tuning and debugging of the UNet model were undertaken before training. The model underwent training and validation for 50 epochs, resulting in metrics such as mean Intersection over Union (IoU) and accuracy for both training and validation images, as well as per-class IoU accuracy, detailed in Table 4. Higher IoU accuracy suggests more accurate segmentation of objects in images and videos. Additionally, we included essential class labels, with potential for further expansion to enhance scene understanding. Figures 9 and 10 depict the Mean IoU against the number of images in the training and validation sets, respectively. The graphs demonstrate that as the number of images increases, the mean Intersection over Union (IoU) also increases, reflecting the model's improved ability to learn and generalize objects in the scene for more accurate segmentation in test images and videos. The training mean IoU surpassed 74.3%, while validation pixel accuracy exceeded 69.8%. Subsequent segmentation results on test image frames, depicted in Figure 8, showcase the UNet model's performance. While the model effectively segments most objects in traffic scenes, some misclassifications occur. Enhanced fine-tuning, augmented training data,

and additional class labels offer avenues for improvement. Notably, accurate segmentation is achieved for visible objects like cars, pedestrians, roads, buildings, and signs, despite challenges posed by unlabelled objects and adverse weather conditions. Overall, the segmentation outcome is satisfactory for traffic scene analysis.

III. CONCLUSIONS

In summary, our improved UNet demonstrates exceptional pixel-level categorization capabilities across varied urban environments. Leveraging a novel Encoder-Decoder module, advanced scene parsing network with ECA mechanism, and Deformable Network, the model showcases superior performance on the Cityscapes Dataset, attaining 74.3% mIoU in the training set and 69.8% mIoU in the validation set. This study represents a significant contribution to semantic segmentation models, particularly in the domain of 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.

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/1608.0, 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] 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

[20] 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.

[21] 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

[22] 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.

