Fully Convolutional Neural Network with Relation Aware Context Information for Image Parsing

Basim Azam
Centre for Intelligent Systems, School of Engineering and Technology, Central Queensland University,
Brisbane, Australia
b.azam@cqu.edu.au

Ranju Mandal
Centre for Intelligent Systems, School of Engineering and Technology, Central Queensland University,
Brisbane, Australia
r.mandal@cqu.edu.au

Brijesh Verma
Centre for Intelligent Systems, School of Engineering and Technology, Central Queensland University,
Brisbane, Australia
b.verma@cqu.edu.au

***Abstract*— Image parsing is among the core tasks in the field of computer vision. The automatic pixel-wise segmentation offers great potential in terms of application adaptability. Traditional convolutional networks have produced better segmentation maps however the research is continued for integration of context information with neural network approaches. In this paper, we propose an image parsing framework that explores the traditional convolutions in fully convolutional networks and learns rich semantic contextual information using the adjacent and spatial modules to generate probability maps. The implicit fusion of the probability maps generated enhances the accuracy of segmentation labels. The proposed framework improves the segmentation accuracy on the CamVid dataset achieving global accuracy of 89.8 %. A comprehensive comparison with state-of-the-art approaches demonstrates that the proposed network exhibits the capability to adapt to the dataset specific information and has the potential to outperform cutting-edge segmentation models.**

Keywords— Image parsing, context information, semantic segmentation, deep learning, neural network.

I. INTRODUCTION

Image parsing is the task of assigning each pixel of an image to its specific object category. Pixel wise segmentation is among the core tasks in the field of computer vision. Image parsing helps identify the shape and boundary of an object with explicit and rich label information as compared to the object detection problem [1, 2]. The segmentation algorithms have successfully been developed over the years. The modern segmentation architectures are completely based on the fully convolutional networks (FCN) [3, 4]. FCN explores the traditional arrangement of convolutional neural network (CNN) to produce pixel wise segmentations, it uses the idea deconvolutions and instead of the fully connected layer at the end of CNN the FCN restores the size of original image by up-sampling the convolutional layers.

The FCNs have demonstrated a compact and effective representation of characteristic for pixel-wise labels. The feature maps are generated using the convolutions and these convolutional feature maps are computed and explored with additional layers such as pooling, activations and fully connected layer to produce pixel labels [5]. The contextual information is of prime importance for scene parsing task in complex scenarios. The images that includes object of varying sizes and similar attributes are complex in terms of producing pixel-wise labels. One of the commonly occurring complexity in the street view images is the misclassification of vehicle objects as a single class [6]. Also, the sky and sea observe a lot of misclassified pixels due to the similarity in color information and wide-ranging appearances. However, such problems can be treated with specific encoding of context information in segmentation architectures. Additionally, the pixel wise production of labels is affected by occurrence of alike objects at varying scales.

A systematic way to fuse the local and global contextual information in addition to convolutional neural network is being explored efficiently in the paper. The study presents a unique image parsing framework that learns contextual relation information of objects for scene parsing. The contextual attributes help producing the pixel-wise labels and achieving higher accuracy values by taking into consideration the neighborhood information. The original contributions of this paper are as follows:

- A novel image parsing framework exploring relation aware global and local information along with fully convolutional network to produce pixel wise labels.
- Spatial context module and adjacent context module to improve and attain high accuracy of pixel labelling using contextual information.
- A comprehensive comparison and analysis with existing approaches, presents enhanced performance of the proposed architecture on the CamVid benchmark dataset.

The rest of the paper is organized as follows. Section 2 briefly describes relevant work in the literature. Section 3 presents the architecture of the proposed network. The experimental results and analysis are presented in section 4 while section 5 concludes the paper.

II. RELATED WORKS

The field of computer vision has observed an upsurge in the performance of algorithms, directed by the recent accomplishments of deep neural networks. The pixel-wise segmentation frameworks have also evolved over the years. The literature review section provides an insight to the existing image parsing frameworks that utilize FCN, ultimately highlighting the potential gap and bridging the information with proposed architecture.

The main research that helped defining the path for the investigations carried out in recent years revolves around deep learning-based techniques such as FCN. The researchers have approached the segmentation problem with extensive usage of FCN based techniques. These techniques can be categorized as: 1) Encoder-decoder architectures and 2) Structurally hierarchical architectures.

A. Encoder-Decoder Architectures

FCNs have facilitated the research being carried out to accurately parse an image, several segmentation architectures rely entirely upon FCN. Such models make use of the convolutional layers to compute semantic attributes eventually producing the feature maps. Most of the encoder-decoder architectures comprise of the convolutional neural networks as an encoder while the decoder is mostly built of similar recurring layers in reverse order to capture the context. The decoder path of the architecture restores the resolution to match the input and produces prediction maps. ERFNet and ESPNet [7, 8] make use of a delicate feature extraction unit as encoder to produce pixel wise label information.

In a similar fashion, the extensively used U-Net [9] acquires the context information using a contracting path as encoder module while the localization is done using similar symmetric expansion path as decoder component of architecture. U-Net architecture originally was developed for research in the area of medical however it has been used extensively for other segmentation tasks as well [10-12]. Similarly, the SegNet [13] comprises of the encoder-decoder formation of layers. Traditional convolutional layers compute the feature information as encoder path while the decoder formation explores the information computed using the max pooling layers. Even though these architectures have performed well for segmentation tasks, they are still inefficient to compute the relation information between the object classes occurring within an image. Consequently, the area for the development and investigation of image parsing architectures to explore the relational information to precisely parse an image is highlighted.

B. Structurally Hierarchical Architectures.

The conventional encoder-decoder architectures have helped produce the object labels of individual pixels. The loss of information, during the down-sampling of the feature maps produced using convolutional layers, disturbs the overall performance of such frameworks. To this end, the architectures with multiple pathways improve upon the loss of information. The initial pathway acquires the semantic information from the pixels while the other pathway obtains spatial information. BiSe-Net [14] has segregated dual pathway architecture, and the DRANet [5] explores combination of position attention and channel attention module to compute the semantic information. Despite the channel wise consideration of semantic and spatial attributes, both architectures condense the channels thus reducing the quality of segmentation labels.

Although the modern segmentation architectures rely heavily upon FCN, however the architectures build upon the base network with additional consideration of multi-scale input in the form of pyramids. The pyramid approach improves the computation of fine details and contextual information. Such architectures produce information at multiple levels considering the input at several scales. Eventually, the features computed are merged in a supervised fashion to improve the segmentation accuracy. The research work [1] and [15] employs the pyramids for segmentation purposes, however they are limited by various challenging aspects and the huge memory requirement is one of the additional limitation.

The modern segmentation architectures take care of certain limitations by utilizing dilated convolution with extended field of view. Atrous or dilated [16] convolutions compute the feature maps in wide ranging field of view but this does not reduce the spatial dimension. The dilation refers to usage of convolving filters with greater size to compute spatial feature maps with enhanced contextual information. The dilated convolutions help improve the segmentation accuracy.

Lately, the Generative Adversarial Networks (GANs) [17] have been used for pixel-wise segmentation. A traditional GAN consist of two architectures, a generator and discriminator competing to improve until the convergence. The generator module adapts the data to put discrimination module on trial, while the discriminative module keeps on improving to distinguish the data coming from generator and noise. The pioneering work in GANs investigate Multilayer Perceptron (MLPs), which are too naïve for segmentation purposes. The developments in the GAN architectures permit the stable training for Deep Convolutional Generative Adversarial Networks (DCGANs) [18]. The DCGANs make use of the small strides and convolutions in discriminator and generator modules respectively. The semantic segmentation architecture using the generative-discriminative approach (SegGAN) produces labelled segmentation maps from a colored input image. The adversarial network in this case discriminates between the segmentation output and input true labels for each pixel. GANs are known for the instability throughout the training process specifically high-resolution imagery is highly undesired for usage with GANs. The Wasserstein’s gradient plenty helps cope with this limitation.

The models developed for pixel wise segmentation produce the labels by considering the attributes of the features and do not consider the whole context of image or the information occurring in the neighborhood. However, the label of an individual pixel has high correlation with the labels present around it, thus the image parsing problem can be reformulated to adapt the contextual relation awareness for further research.

III. Proposed Approach

In this section, we present proposed image parsing framework in detail. The overall structure of the network is presented in figure 1. The architecture is based upon the tradition FCN network to extract features from the input image, these feature maps are explored in the softmax layer just before the traditionally final convolution layer to produce probabilities. In addition, the spatial and adjacent occurrences of similar pixels are learned in respective context module. Lastly, the fusion of the probabilities computed in a supervised manner helps produce the improved pixel labels.

A. The Main Architecture

The proposed image parsing architecture explores the input images with help of three different modules, the FCN architecture, spatial context module and the adjacent context module. The FCN module deals with the traditional exploration of input images, extraction of feature maps using the receptive fields finally producing coarse output maps. The coarse feature maps formed are connected back to the pixels by reformulating an equivalent modified network to produce pixelwise prediction. However, there is a slight variation than the traditional FCN that is instead of producing dense pixel labels as output we explore the softmax probabilities before the final convolutional layer. The spatial context module predicts the probability of class label based on the long-rage

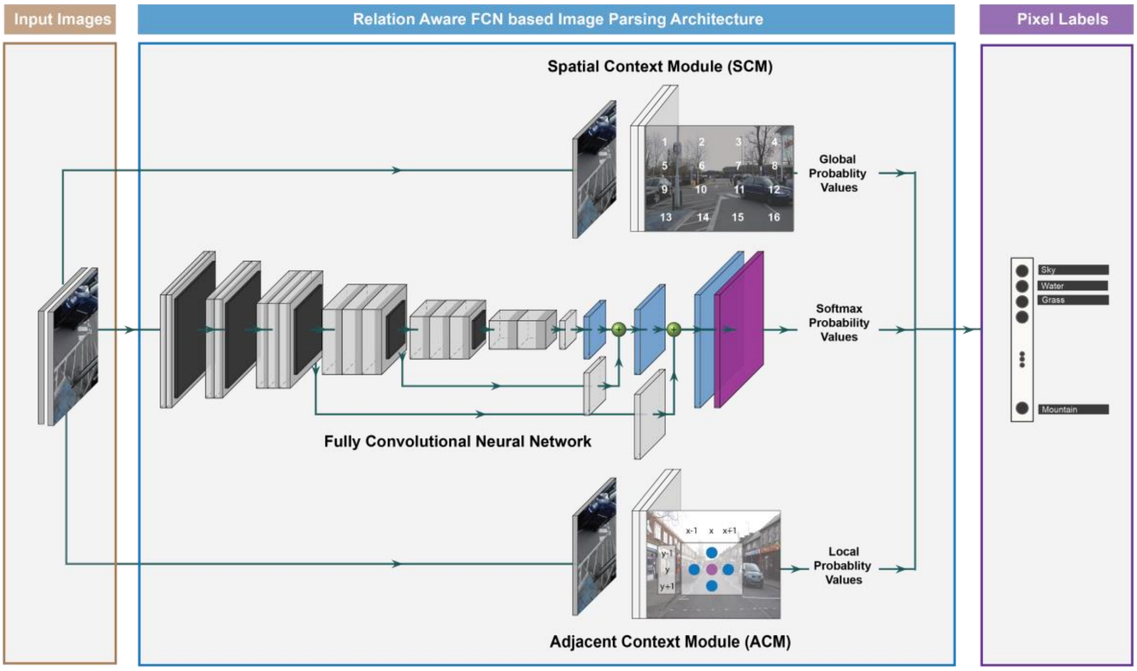
Input Images
Relation Aware FCN based Image Parsing Architecture
Pixel Labels
Spatial Context Module (SCM)
1 2 3 4
5 6 7 8
9 10 11 12
13 14 15 16
Global Probablity Values
Softmax Probability Values
Fully Convolutional Neural Network
x-1 x x+1
y-1 y y+1
Local Probablity Values
Adjacent Context Module (ACM)
Sky
Water
Grass
Mountain

Fig. 1. The overview of the relation-aware FCN based image parsing framework with integrated Spatial Context Module and Adjacent Context Modules.

context properties of the superpixels, while the adjacent context module anticipates the probability of class labels based on the neighborhood superpixel information. These probability values estimated are finally fused together to compute the pixel labels.

B. *Spatial Context Module (SCM)*

The SCM explores the context information in the form of spatial blocks from the superpixels computed using the input images. The spatial module assimilates the likelihood of class label for the pixel values by learning prior occurrence of object pixels. The term spatial context represents the contextual occurrences of pixels within spatial block of information. To calculate the global probability value, the votes are casted by each superpixel present within the spatial block to compute the class probability of an individual superpixel.

Each individual pixel with a specific class label that occurs in a different block helps generate a matrix to indicate the probability of pixels with class labels. The normalized matrix is computed for entire dataset in the form of block pairs. The pixel co-occurrence frequency among the object in various blocks is represent in the matrix. The values computed represent the confidence of class label by considering the spatial contextual information from the rest of superpixels in every block.

C. *Adjacent Context Module (ACM)*

The adjacent contextual information refers to occurrence of adjacent superpixel information. The ACM module reflects the prior local context information learned in terms the adjacent superpixels. The term adjacent reflects the connected neighborhood superpixels, and the local probability value is computed by the votes casted by each neighboring superpixel.

Every adjacent superpixel with a specific class label contribute towards computation of a matrix from the training data. The local matrix is formulated using the pairs of neighboring superpixels present at any part of the image. The local probability represents the confidence value of class labels for each superpixel given the contextual information provided by the neighboring superpixels.

IV. Experiments and Results

The proposed architecture is experimentally evaluated on the CamVid dataset using the evaluation metrics that include the dice coefficient score, the jaccard similarity score and global accuracy.

A. *Platform Specifications*

To implement the FCN, ACM and SCM modules MATLAB has been used. The images, labels and train-test set ratios are generated within the environment. The High-Performance Computing (HPC) facility available at Central Queensland University is used for the computations. The parameters have been maintained consistently throughout the experimentations. The training parameters are presented in table I.

TABLE I. THE TRAINING PARAMETER USED TO TRAIN/TEST THE IMAGE PARSING FRAMEWORK.

Parameter	Value
Optimizer	SGDM
Momentum	0.9
Learning Rate	10^{-3}
Max Epochs	30
Mini Batch Size	16
L2 Regularization	0.005
Learning Rate Scheduling	Piecewise

B. *CamVid Dataset*

The Cambridge-driving Labelled video database includes the video sequences of road driving scenarios. The captured videos are comprehensive and complex as they are captured at different times of the day including bright illuminous and darker (around dusk) times. The sequences comprise of

TABLE I. TABLE 1. RESULTS ON THE CAMVID DATASET IN TERMS OF CLASS-WISE ACCURACY, MEAN JACCARD SCORE AND GLOBAL ACCURACY IN COMPARISON WITH THE STATE-OF-THE ART SEGMENTATION TECHNIQUES.

Approach / Class	Building	Tree	Sky	Car	Sign	Road	Pedestrian	Fence	Pole	Sidewalk	Cyclist	Mean IOU	Global Accuracy
ReSeg [19]	86.6	84.7	93.0	87.3	48.6	98.0	63.3	20.9	35.6	87.3	43.5	58.8	88.9
SegNet[13]	68.7	52.0	87.0	58.5	13.4	86.2	25.3	17.9	16.0	60.5	24.8	46.4	62.5
B. Segnet [20]	-											63.1	86.9
FCN-8 [21]	77.8	71.0	88.7	76.1	32.7	91.2	41.7	24.4	19.9	72.7	31.0	57.0	88.0
Dilation-8 [22]	82.6	76.2	89.0	84.0	46.9	92.2	56.3	35.8	23.4	75.3	55.5	65.3	79.0
DeepLab-L [16]	81.5	74.6	89.0	82.2	42.3	92.2	48.4	27.2	14.3	75.4	50.1	61.6	-
FCN-Comb[23]	79.7	77.2	85.7	86.1	45.3	94.9	45.9	69.0	25.2	86.2	57.9	-	88.8
Proposed Approach	**93.2**	**86.4**	**94.5**	**86.7**	31.9	96.3	26.9	**71.5**	27.1	83.7	**62.2**	**64.1**	**89.8**

information rich scenarios including buildings, roads, vehicles, and humans. The ten-minute footage is split into 701 images, these images are annotated into 32 different categories. As the researchers have merged similar classes into 11 object categories, we follow the same convention and make use of majority of labelled pixels in the CamVid database and split the database into 80% for training while the rest 20% for evaluations.

C. Evaluation Metrics

The evaluation metrics used to represent the results include jaccard coefficient, dice coefficient and global accuracy. The dice coefficient score evaluates the closeness of the predicted boundary pixel label with the ground truth label. The dice coefficient between predict label L_{pr} and ground truth label L_{gt} can be represented as

$$Dice \ = \ 2 * |\ Lpr \cap Lgt||\ Lpr| + |Lgt| \qquad (1)$$

The Jaccard Coefficient also known as Intersection-Over-Union (IoU) is the ratio of overlap area between the predicted labels and the ground-truth labels to area represented by union of both labels. The Jaccard similarity score of predicted labels L_{pr} and ground truth labels L_{gt} can be described as:

$$Jaccard \ = \ |\ Lpr \cap Lgt||Lpr \cup Lgt| \qquad (2)$$

D. Performance Comparison

Table II presents the obtained results for the proposed image parsing approach with state-of-the-art pixel wise segmentation methods. The architecture has been trained using the conventional convolutional layer and the traditional up sampling layers. The results present improvement in the segmentation accuracy. The proposed architecture improves upon the state-of-the-art models.

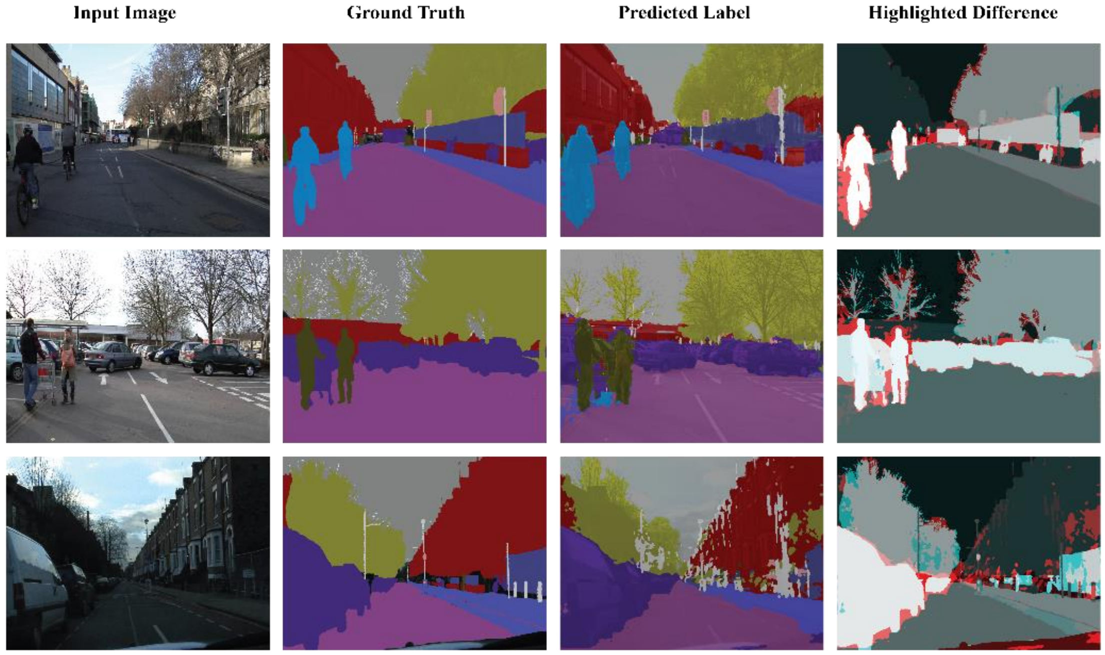

Fig. 2. Examples of pixel-wise segmentation results on the CamVid dataset. The differences between the ground truth labels and predicted labels are highlighted using cyan and red areas.

TABLE II. TABLE 2. THE HIGH-LEVEL PERFORMANCE OVERVIEW OF THE PROPOSED ARCHITECTURE WITHOUT RELATION INFORMATION FOR EACH CLASS.

	Building	Tree	Sky	Car	Sign	Road	Pedestrian	Fence	Pole	Sidewalk	Cyclist
Jaccard Score	0.76	0.76	0.90	0.77	0.39	0.91	0.44	0.46	0.21	0.70	0.60
Dice Score	0.58	0.70	0.89	0.69	0.48	0.76	0.60	0.49	0.51	0.71	0.55
Accuracy	0.78	0.87	0.93	0.90	0.79	0.93	0.87	0.81	0.71	0.88	0.88

TABLE III. TABLE 3. RESULTS OF THE ABLATION STUDY ON CAMVID DATASET.

	Conv. Architecture	ACM	SCM	Proposed Architecture
Building	0.78	☑	☑	0.93
Tree	0.87	☑	☑	0.86
Sky	0.93	☑	☑	0.94
Car	0.90	☑	☑	0.87
Sign	0.79	☑	☑	0.32
Road	0.93	☑	☑	0.96
Pedestrian	0.87	☑	☑	0.27
Fence	0.81	☑	☑	0.72
Pole	0.71	☑	☑	0.40
Side Walk	0.88	☑	☑	0.83
Cyclist	0.88	☑	☑	0.62
Dice Score	-			0.65
Jaccard Score	-			0.63
Weighted Jaccard Score	-			0.80
Global Accuracy	-			0.87
Dice Score	-	☑	☑	0.67
Jaccard Score	-	☑	☑	0.64
Weighted Jaccard Score	-	☑	☑	0.81
Global Accuracy	-	☑	☑	0.89

The CamVid dataset consist of the video frames and to some extent includes the temporal information as well. The incorporation of adjacent and spatial context modules helps to build the final pixel labels based on the prior information learned. And the temporal info from the data aids the ACM and SCM modules. The final pixel labels produced are slightly improved than the labels obtained using the softmax probabilities from traditional FCN architecture.

Fig. 2 represents the qualitative pixel-wise segmentation results obtained on the CamVid dataset. The overlayed ground-truth pixel labels help identify the improvement in the pixel labels. The red and cyan areas in the fourth column represent regions where the pixel-wise segmentation deviates from the expected ground truth.

The pixel-wise segmentation results exhibit better overlap scores for classes such as road, sky, and building. These classes have higher number of training pixels as well. Whereas the objects that relatively smaller like cars and pedestrians are less accurate. The Jaccard score metric provides validation of the visual result. The object categories road, building and sky have higher jaccard similarity scores, on the other hand pedestrian and car have comparatively low scores.

E. *Ablation Study*

The ablation studies are performed to present the enhanced segmentation outputs and effectiveness of the proposed context modules among the overall architecture. The ablation study is carried out on CamVid dataset. Both ACM and SCM modules operate in conjunction with traditional architecture, the segmentation results are also evaluated after the incorporation of modules. The results quantify the contributions towards final pixel labels. Table III represents the high-level performance overview of the proposed architecture. The per class metric helps to identify the overall good performance for pixel-wise labelling however the classes with lower pixel representations in the dataset (cyclist, pedestrian, and car) have not been segmented well. The availability of more data may help produce better results for such classes.

Table IV presents the results of the ablation study on the CamVid dataset. The ACM and SCM modules increase the global accuracy by 0.17 and 1.94% and the average jaccard score 0.09 and 1.4%. The use of the contextual modules increases the dice score by 2.9%. The proposed architecture has an overall advantage of 1.27% in weighted jaccard scores. The segmentation outputs contain a few errors which identify the fact that the ACM and SCM modules work better on the object categories with higher number of pixels in the training set. The comparative experimental analysis identifies the advantage of the proposed architecture. It has an enhanced capability of adapting to the specific dataset, which is because of the adjacent and spatial superpixel occurrences.

V. CONCLUSION

In this paper, we have proposed a novel image parsing architecture for pixel-wise segmentation tasks, which utilizes an implicit adjacent context module and spatial context module to improve the pixel-wise segmentation of base convolutional neural network. We have presented extensive experiments on the CamVid benchmark dataset, these experiments demonstrate that the integration of context modules improve the cutting-edge segmentation techniques and promises that further developments can achieve state-of-the-art performances on other benchmarks. The proposed model can adapt to the nature of dataset and learn the occurrences of object categories using the ACM and SCM modules. We aim to integrate backpropagation within the modules and use the proposed architecture on various benchmark datasets.

VI. Acknowledgments

This research was supported under Australian Research Council's Discovery Projects funding scheme (project number DP200102252).

References

[1] Farabet, C., Couprie, C., Najman, L., Lecun, Y.: Learning hierarchical features for scene labeling. IEEE Transactions on Pattern Analysis and Machine Intelligence 35, pp. 1915-1929, 2013.

[2] Socher, R., Lin, C.C.-Y., Ng, A.Y., Manning, C.D.: Parsing natural scenes and natural language with recursive neural networks. Proceedings of the 28th International Conference on International Conference on Machine Learning, pp. 129–136, 2011.

[3] Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A., Agrawal, A.: Context encoding for semantic segmentation. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition pp. 7151-7160, 2018.

[4] Shuai, B., Zuo, Z., Wang, B., Wang, G.: Scene segmentation with DAG-recurrent neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1480-1493, 2018.

[5] Fu, J., Liu, J., Jiang, J., Li, Y., Bao, Y., Lu, H.: Scene segmentation with dual relation-aware attention network. IEEE Transactions on Neural Networks and Learning Systems, pp. 1-14, 2020.

[6] Xiong, Y., Liao, R., Zhao, H., Hu, R., Bai, M., Yumer, E., Urtasun, R.: UPSNet: A unified panoptic segmentation network. Proceedings of 2019 Conference on Computer Vision and Pattern Recognition, pp. 8818-8826, 2019.

[7] Romera, E., Álvarez, J.M., Bergasa, L.M., Arroyo, R.: ERFNet: Efficient residual factorized convNet for real-time semantic segmentation. IEEE Transactions on Intelligent Transportation Systems, pp. 263-272, 2018.

[8] Du, J., Song, J., Cheng, K., Zhang, Z., Zhou, H., Qin, H.: Efficient spatial pyramid of dilated convolution and bottleneck network for zero-shot super resolution. IEEE Access, pp. 117961-117971, 2020.

[9] Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. Proceedings of the 18th International Conference on Medical Image Computing and COmputer-Assisted Intervention, pp. 234-241, 2015.

[10] Xie, H., Lin, C., Zheng, H., Lin, P.: An UNet-Based head shoulder segmentation network. In: IEEE International Conference on Consumer Electronics-Taiwan, pp. 1-2, 2018.

[11] Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: UNet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Transactions on Medical Imaging, pp. 1856-1867, 2020.

[12] Baheti, B., Innani, S., Gajre, S., Talbar, S.: Eff-UNet: A novel architecture for semantic segmentation in unstructured environment. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1473-1481 , 2020.

[13] Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 39, pp. 2481-2495, 2017.

[14] Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In: European Conference on Computer Vision, pp. 334-349, 2018.

[15] Lin, G., Shen, C., Hengel, A.v.d., Reid, I.: Exploring context with deep structured models for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 40, pp. 1352-1366, 2018.

[16] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence 40, pp. 834-848, 2018.

[17] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems pp. 2672–2680, 2014.

[18] Lu, Y., Li, J.: Generative adversarial network for improving deep learning based malware classification. In: Winter Simulation Conference, pp. 584-593, 2019.

[19] Visin, F., Ciccone, M., Romero, A., Kastner, K., Cho, K., Bengio, Y., Matteucci, M., Courville, A.: ReSeg: A recurrent neural network-based model for semantic segmentation. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops pp. 426-433 , 2015.

[20] Kendall, A., Badrinarayanan, V., Cipolla, R.: Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. British Machine Vision Conference. pp. 1-12 , 2015.

[21] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 39, pp. 640-651, 2014.

[22] Yu, Fisher, Koltun, Vladlen: Multi-scale context aggregation by dilated convolutions. In: 4th Internation Conference on Learning Representations, pp 1-13, 2016.

[23] Wu, Y., Yang, T., Zhao, J., Guan, L., Li, J.: Fully combined convolutional network with soft cost function for traffic scene parsing. In: International Conference on Intelligent Computing, pp. 725-731, 2017.