FuzzyNet: Context Encoding and Spatial Fuzzy Refinement Network in Semantic Segmentation

Ariyo Oluwasanmi
School of Information and Software Engineering
University of Electronic Science and Technology of China
Chengdu, China
ariyo@std.uestc.edu.cn

Ebere Eziefuna
School of Biomedical Engineering
University of Electronic Science and Technology of China
Chengdu, China
magnusonyinye@yahoo.com

Favour Ekong
School of Information and Software Engineering
University of Electronic Science and Technology of China
Chengdu, China
fazzyabb@yahoo.com

Edward Baagyere
School of Information and Software Engineering
University of Electronic Science and Technology of China
Chengdu, China
eddiesyoung200@yahoo.com

Zhiguang Qin
School of Information and Software Engineering
University of Electronic Science and Technology of China
Chengdu, China
qinzgg@uestc.edu.cn

***Abstract*—This paper addresses the pertinent object localization problem in deep convolutional neural networks by introducing a spatial fuzzy post-processing function which allows the smooth transition of object edges within individual pixel's neighborhood. We accomplish the task of semantic segmentation by first computing class weights as a means of avoiding class bias or imbalance training. Our proposed FuzzyNet runs a convolutional encoder-decoder network architecture with the following novel features: (i) It incorporates a new Global Context Spatial Module (GCSM) (ii) It exploits the atrous spatial pyramid structure for enriching the semantic encoding (iii) It incorporates the transfer of lower level features connected to higher levels with contextual spatial feature maps (iv) It effectively achieves an attention component with an extensive focus on objects of interest. Thus, the fusion of spatial fuzzy function enables normalization of intensity variation at different object boundaries, avoidance of poor localization and ultimately resulting in quality semantic segmentation. The evaluation of our proposed FuzzyNet model achieves improved performance with respect to the accuracy and object boundary refinement on the PASCAL VOC 2012 and CamVid benchmark datasets.**

***Index Terms*—semantic segmentation, deep convolutional network, fully convolutional network, fuzzy function, refinement network, deep learning**

I. Introduction

Dense classification or semantic segmentation which is a huge branch of computer vision, involving pixel level classification and labeling has seen considerable usefulness ranging from scene parsing, GeoSensing and autonomous driving to robotics [1]. Though recent advances in convolutional neural network (CNN) has increased performances in tasks such as image recognition and classification, it however still suffers from poor localization due to loss of spatial feature from downsampling, and ultimately, coarse edge in pixel-to-pixel labeling [2].

Consequently, Fully Convolutional Networks (FCN) which is made up of downsampling and upsampling Encoder-Decoder model have been adopted to maintain the spatial resolution of convolutional feature maps [3]. Often, the encoder is mainly convolutional layers stacked to learn feature maps and then scaled up to the input image size by a decoder model. Regularly, the decoder adopts bilinear interpolation or deconvolution learning to scale up image resolution [4]. This encoder-decoder idea helps to complement each other's weaknesses. First, most decoder try to retain spatial encoding from low level feature in the lower convolutional stack to other higher layers by connecting residual blocks [5], providing contextual relationships from layers to layers. Hence, supplementing the decoder's generation of spatially rich high-resolution feature maps for accurate sematic segmentation and comparison [6]. Similarly, this ensures consistent spatial alignment of image edges, curves and object boundaries.

Notwithstanding, the task of semantic segmentation remains daunting, resulting in coarse edges and blurry object boundaries [7]. This is principally because of heterogeneous and complex pictorial content as displayed in Fig. 1 and 2. Understandably, contextual information and spatial relationships in

This work was supported by the NSFC-Guangdong Joint Fund (Grant No. U1401257), National Natural Science Foundation of China (Grant No. 61300090, 61133016 and 61272527), science and technology plan projects in Sichuan Province (Grant No. 2014JY0172) and the opening project of Guangdong Provincial Key Laboratory of Electronic Information Products Reliability Technology (Grant No. 2013A061401003).

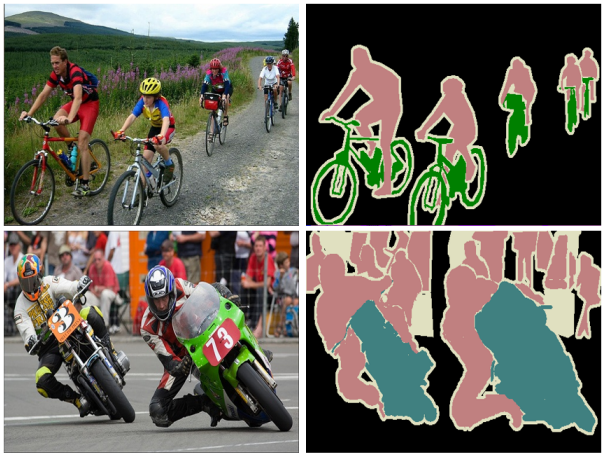

Fig. 1. Illustration of complex scenes in the PASCAL VOC 2012 dataset.

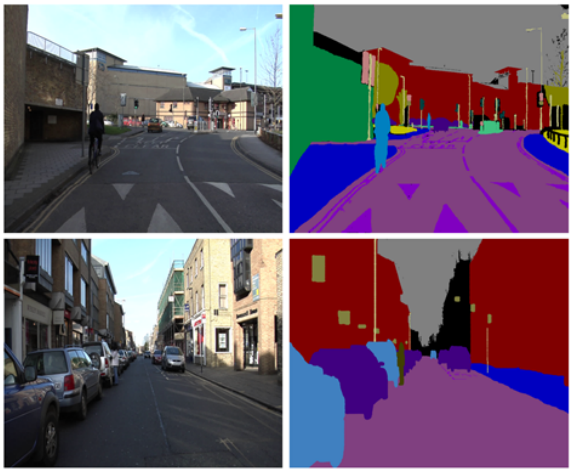

Fig. 2. Illustration of complex scenes in the CamVid dataset.

sematic segmentation remains a major concern. This underlies imperatively the weightiness of global context and the transfer of spatial coordinate along all stacks of encoder and decoder branches. Our algorithm first implements class weighting to avoid class imbalance or bias training on more dominant classes in the dataset. Additionally, a new spatial path and global context transfer network is presented in this work, termed Global Context Spatial Module (GCSM).

This efficiently eliminates spatial information loss alongside permitting contextual information flow from lower level convolutional feature to other higher levels. Conjointly, we explore a post-processing spatial fuzzy function to solve the inherent problem of object boundary localization. Our proposed model has a number of contributions as follows:

- Imbalance regularization: We highlight the adverse effect of imbalance class in segmentation. Our label classes are effectively computed using label median frequency to balance dominancy in classes and improve training against bias learning as depicted in Table 1 and Fig. 3.
- Spatial and contextual connectivity: We propose a novel semantic information transfer network from low-level high-resolution features to high-level low-resolution features. Explicitly, our designed Global Spatial Context and Global Atrous Module proves the preservation of spatial features and attention exertion.
- Smoothness: We effectively extend fully connected networks with post processing fuzzy spatial function at acceptable cost, achieving improved localization devoid of coarse invariance.

The organization of this paper is as follows: Section 2 visits summary of backbone approaches implemented in semantic segmentation, while Section 3 describes our preliminary values. Section 4 gives a detailed description of our model with its implementation in section 5 and Section 6 highlights the summary of the paper under the conclusion.

II. Related Work

The Fully Convolution Network (FCN) has been adopted as the backbone model for semantic segmentation because of the inherent invariant property of the classical CNN as well as its ability to combat its spatial consequence with a deconvolution network [8]. Furthermore, FCN have seen enhanced application in other computer vision tasks such as spot the difference, image restoration and depth estimation [9]. Nonetheless, due to the architecture's alternating convolution and pooling computation, FCN possess low resolution feature map prediction drawback which have been addressed to varying degrees by several techniques. For instance, [10] used basic bilinear interpolation to derive finer details of the coarse heat maps generated from classical CNN to high resolution feature maps good for semantic segmentation. They achieved this by direct upsampling and aggregating present layer output with preceding features in the decoder network. The authors in [11] utilize data augmentation to overcome training dataset limitation, hence, permitting larger network design which consists of contracting path for transmitting context information to the upsampled layers. SegNet [12] implemented a VGG network as the encoder, alongside a decoder layer having pooling indices transferred from corresponding encoder layer to upstream decoder network. These pooling connections then forfeits the need to learn upsample feature maps, achieving better segmentation result. The concept of Atrous convolution which introduces certain spaces between kernel values was utilized by [13] to achieve dense feature map generation. These dilated convolutions improve feature resolution by enlarging filters' field of view.

In addition, a dilated pooling with different rate was introduced to robustly capture image context, allowing the multi-scale pooling layers to be summed at a later layer, enriching its contextual information, thus improving performance [14]. A chained residual pooling consisting of aggregation of convolution and indices transfer in [15] could encode contextual information efficiently, refining the model with residual connections

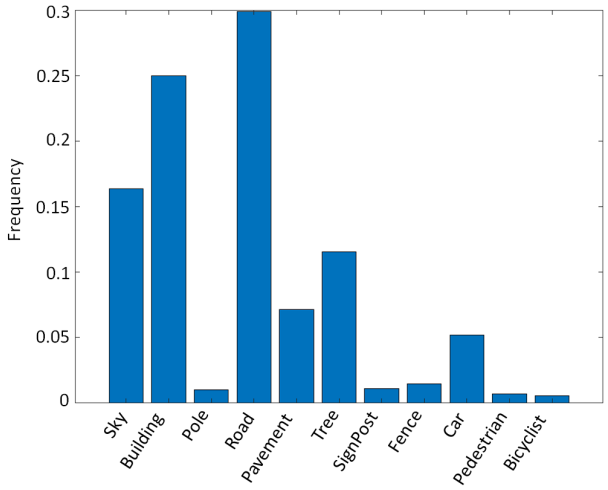

Fig. 3. CamVid Dataset class pixel frequency count

along the subsample terrain. In addressing the problem of localization, [16] utilized a globally constructed network using large convolution filter size to improve boundary refinement. Alongside modifying classical CNN with atrous convolutions, [17] added a post processing probabilistic model to refine their end-to-end model, ensuring a more refined segmentation output. Furthermore, global context encoding was exploited using L2 normalization to fuse different scales of pooling features in a bid to stabilize training and improve accuracy [18]. With the aim of getting a more compact feature resolution, authors in [19] used a densely connected multi scale atrous pooling model to achieve large scale contextual semantic information in a denser range. A model designed to take advantage of early sampling such that a more compressed output is achieved to eliminate redundancy [20]. This architecture is particularly fast for inference with high accuracy. More semantic concept and information was encoded from low level features to higher levels by augmenting the transfer of feature resolution along the upsampling layers with real time inference [21].

III. FuzzyNet

Here, we epitomize our propose FuzzyNet model which is a new Fully Convolution Network (FCN) framework of encoder-decoder network with a novel context-spatial path and post processing refinement function as indicated in Fig. 4 and 5. We discuss its effectiveness for achieving refined semantic segmentation task. The network is initiated by processing dataset class labels to avoid biased training.

A. Structural Design

1) Class weighting: In most segmentation datasets, the pixels by class labels differ for each classes. In most cases, some classes have very large pixel count while others have low counts. Such scenario if unchecked may result in class bias or imbalance training as dominant classes are at an advantage. The class distribution of the CamVid dataset by class pixel count as illustrated in Table 1 and Fig. 3 portrays the imbalance nature of the various classes.

TABLE I
Class pixel label distribution in the CamVid Dataset

Class	Pixel Count
Sky	7.6801e+07
Building	1.1737e+08
Pole	4.7987e+06
Road	1.4054e+08
Pavement	3.3614e+07
Tree	5.4259e+07
SignPost	5.2242e+06
Fence	6.9211e+06
Car	2.4437e+07
Pedestrian	3.4029e+06
Bicyclist	2.5912e+06

To extenuate the effect of class imbalance on our training, we weighted the class labels by the median of total class pixel to the total number of images in each class [2].

2) Encoder: We take advantage of CNN's invariance property and ability to extract features from images. By implementing a stacked layer of convolutions, we obtain a fully learned filter weights which are capable of recognizing different objects as well as classifying their pixels into a class for segmentation purposes [7]. For our encoder model, we adopted the ResNet architecture [22] which was able to learn an identity function of the encoder's sub-networks. That is, given an input x and output $H(x)$ of a particular sub-network, the underlying difference or residual $F(x)$ is also learned such that the pertinent curse of dimensionality and vanishing gradient problems of deep CNN are subjugated during backpropagation. However, we added the hole algorithm of [23] by using atrous or dilated rate convolutions for dense computation.

3) Decoder: Owing to subsampling operation, the encoder outputs somewhat coarse feature maps that must be semantically refined to achieve fine-grained localization accuracy in pixel space [17]. This is addressed by upsampling to higher resolution through bilinear interpolation, taking weighted average of neighboring pixels, hence relinquishing computational cost for upsampling learning. Also, skip connections from previous layers along the model are strategically embedded into the decoder to establish contextual and spatial reconstruction [6].

4) Atrous Convolution and Pooling: Atrous convolution which introduces specified spacing between kernel values conclusively allows increase in receptive field of view. This is achieved without multiple convolutions, advertently resulting in efficient and denser feature maps extraction. Allowing fewer parameters, faster computation and an output of larger resolution. Alongside, atrous pooling is implemented for capturing multiscale objects and their sizes, retaining contextual information densely. The popular and effective atrous spatial pyramid pooling is designed by concatenating different atrous rates of pooling labels to avoid biased training.

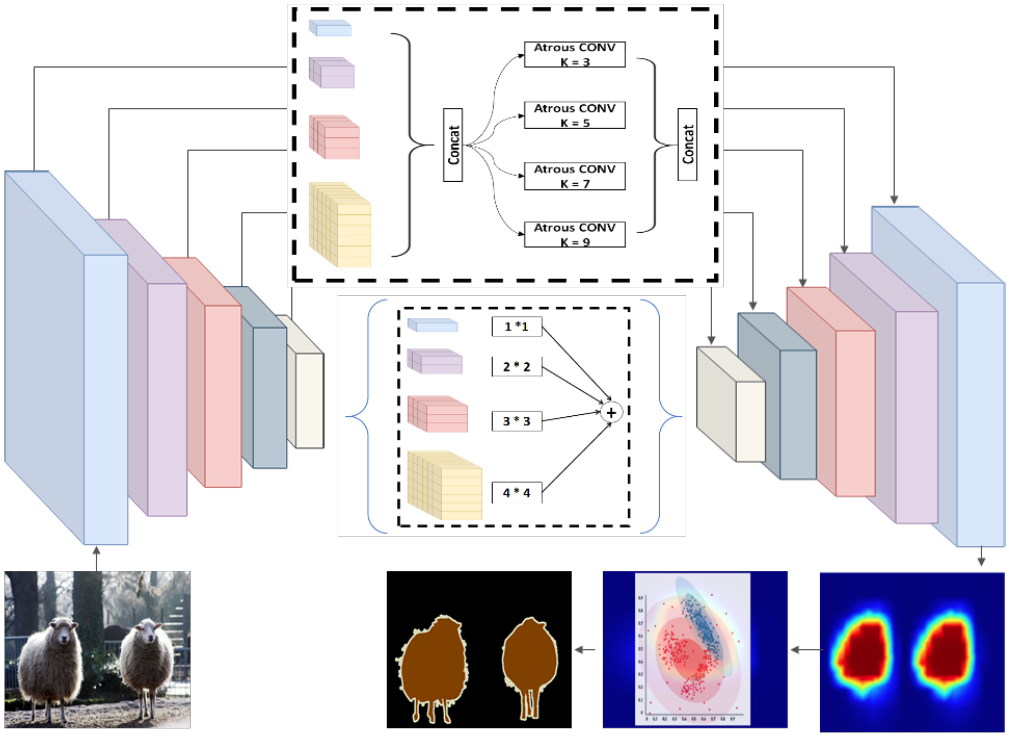

Fig. 4. Our proposed model with our tailored backbone and post processing function.

B. Spatial and Context Extraction Module

1) Global Atrous Module: Since bottlenecks obtain compressed representation of input image with reduced feature dimensionality, the means of conveying such representation to the decoder must be especially thoughtful, both contextually and spatially for productive feature reuse [3]. Bolstering that, we introduce a pyramidal atrous convoluted bottleneck with four different rates 1, 2, 3 and 4 receptively. As different atrous convolution rate captures varying low level features, it ensures the sustenance of multi-scale object sizes, incorporating their spatial context, and eventually improving the task of semantic segmentation.

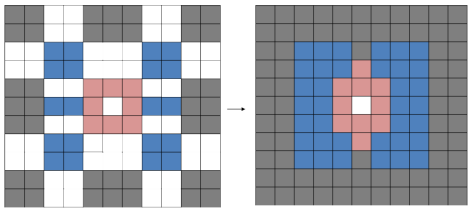

Fig. 5. Fuzzy function mapping with spatial neighborhood cluster.

2) Global Spatial Context: Efficient use of skip connections from previous layer is very decisive in semantic segmentation tasks, especially prior to downsampling. This effort helps to fine-grain convoluted features with captured context which results in improved localization [4]. Here, we establish a new context path which is able to propagate contextual information along feature layers with their spatial tendencies and boundaries. Our spatial context path utilizes global ambiance for overall scene understanding and interpretation, thus, helping to classify object with basic understanding of scenery and cues as relating to global information such as nearby objects. This is achieved by employing spatial pyramid pooling path which is specially designed to enrich convolution operations.

C. Refinement Function

1) Spatial Fuzzy Function: Unlike classification, coarse feature maps are not pleasing for segmentation. Here, a post-bilinear interpolation spatial fuzzy function which individually evaluates the class of pixels is implemented. This is done by considering a certain neighboring region of 10*10 window size [24] as described in Fig. 5. Each pixel is refined by taking the euclidean distance to surrounding pixel. The individual pixel optimization particularly eliminates intensity inhomogeneity, thereby refining object boundaries and effectively improving localization. For instance, given $X = (x1, x2, x3, \ldots\ldots, xN)$ pixels ranging to N, the objective function J is defined as:

$$J = \sum_{j=1}^{N} \sum_{i=1}^{c} \sum_{k \epsilon NB(X_j)} U_{ij} \parallel x_j - v_i \parallel^2 . \quad (1)$$

Where U_{ij} is represented as pixel X_j association to the i_{th} class, $NB(X_j)$ is represented as a square window around pixel X_j. The saddle point or local minimum v_i of the i_{th} class is then represented as :

$$U_{ij} = \frac{1}{\sum_{k=1}^{c} (x_j - v_i / x_j - v_k)^{2/(m-1)}}. \quad (2)$$

$$V_i = \frac{\sum_{J=1}^{N} U_{ij} m}{\sum_{J=1}^{N} U_{ij} m}. \quad (3)$$

The fuzziness is controlled by parameter $m = 2$ as often used in literature. A pixel is then finally classified to a particular class in which it has the highest probability. The added advantage of this is that objects become even more homogenous by removing spurious blobs and smoothing connecting pixels with neighborhood function. Though the refinement outcome have a degree of fuzzy membership to all the classes, we ultimately assign each pixel to the class with the highest probability. As such, the global information of a scene is included in determining the class of an object, this is especially useful at object's boundaries and edges.

D. Architecture

For encoding contextually rich global and spatial information, we propose FuzzyNet for semantic segmentation. First, the datasets are preprocessed by weighting the class labels for training enhancement. Afterwards, a pretrained ResNet 50 network is transferred as the encoder model, helping to achieve high classification of objects with a bilinear model as the decoder. In between, we designed a global atrous module with different atrous rate for conveying spatial contextual features to the up-sample layers. Specifically, our global spatial context module channels global information from low-level high-dimensional feature maps to high-level low-dimensional feature maps along the model. The combination of the global atrous module and global spatial context module essentially incorporates larger receptive field and richer spatial information for increased performance. Subsequently, the output of the model is further spatially ingrained with a fuzzy function to refine objects boundaries, considering a specified neighbored.

IV. EXPERIMENTS

Experimental results and analysis of our model is compared to other state-of-the-art approaches. Pascal VOC 2012 and CamVid benchmark datasets are used for evaluating and result comparison in accomplishing contextual information identification and labeling.

A. Dataset

The PASCAL (Pattern Analysis, Statistical Modeling and Computational Learning) VOC (Visual Object Classes) [25] is the mostly used semantic segmentation dataset. It consists of 1464, 1449 and 1456 training, validation and testing data respectively with a total of 20 classes and another for the image background. Also, the CamVid dataset [26] is a 960 * 720 resolution scene understanding dataset containing 701 annotated images of training, testing and validation sets of 367, 233 and 101 respectively. Furthermore, during training, the datasets are augmented to increase learning capability by flipping, translation, rotation, scaling and combinations between 10 and -10 degrees and random scaling between 0.5 and 2 during the training process.

B. Training

Cross entropy loss function is used for finding performance cost and we performed back propagation using mini-batch stochastic gradient descent as the optimizer with a batch size of 12, momentum of 0.99, weight decay of 0.0001 and a base learning rate of 1e-10. The poly learning rate of 10-1, 10-2, 10-3 and 10-4 was adopted as in [13] at 1, 30, 60, and 90 epochs respectively the training iteration is set to 1e+5 for PASCAL VOC and CamVid dataset respectively for 120 epochs.

C. Evaluation Metric

The mean Intersection over Union (mIoU) is used for qualitative performance evaluation. Given a number of class C, ground truth mask G and predicted output P, the IOU quantifies the percentage overlap of G to P for all C [27].

D. Ablation study

1) Baseline: We use the FCN model [10] as our backbone, it is a contemporary fully connected network of encoder-decoder model. The FCN model utilized here is built on ResNet50 Network instead of the VGG16 Net used in [10] without our Global Spatial Context, Global Atrous Module and Spatial Fuzzy Refinement Function, yielding a mIoU of 67.5% on the PASCAL VOC 2012 dataset and 61.3% on the CamVid dataset.

2) GCSM: Accordingly, we added the Global Spatial Context (GSC) and Global Atrous Module (GAM) to our backbone, this is termed as Global Context Spatial Module (GCSM).

Though our backbone accomplishes the task of segmentation, the addition of spatial and global module increases accuracy as they both enrich semantic context by incorporating spatial information which improves segmentation. Also, the multi scale convolution at the bottleneck improves labelling of small and large objects. Transferring contextual information from different receptive field appears to be an important tactics for capturing objects of different sizes. Qualitative and quantitative experimental evaluation of our models are highlighted. Table 2 & 3 display the quantitative results of our model on the PASCAL VOC 2012 and CamVid dataset while Fig. 6 depicts the model's segmentational accuracy.

TABLE II
EVALUATION OF ABLATION STUDIES ON CAMVID TEST SET

Method	Building	Tree	Sky	Car	Sign	Road	Pedestrian	Fence	Pole	Sidewalk	Bicyclist	mIoU(%)
ApesNet [28]	76.0	80.2	**95.7**	84.2	**52.3**	93.9	59.9	43.8	**42.6**	87.6	46.1	
ENet [20]	74.7	77.8	**95.1**	82.4	**51.0**	95.1	**67.2**	51.7	35.4	86.7	34.1	51.3
SegNet [12]	88.8	**87.3**	92.4	82.1	20.5	**97.2**	57.1	49.3	27.5	84.4	30.7	55.6
LinkNet [29]	**88.8**	85.3	92.8	77.6	41.7	96.8	57.0	**57.8**	37.8	88.4	27.2	55.8
FCN8 [10]	77.8	71.0	88.7	76.1	32.7	91.2	41.7	24.4	19.9	72.7	31.0	57.0
AttentionM [3]	88.4	84.5	93.4	84.9	48.8	95.6	61.8	54.8	38.4	**90.5**	47.7	60.1
DeepLab - LFOV [17]	81.5	74.6	89.0	82.2	42.3	92.2	48.4	27.2	14.3	75.4	50.1	61.6
Dilation8 [30]	82.6	76.2	89.0	84.0	46.9	92.2	56.3	35.8	23.4	75.3	55.5	65.3
BiseNet [31]	83.0	75.8	92.0	83.7	46.5	94.6	58.8	53.6	31.9	81.4	54.0	68.7
FuzzyNet (Ours)	83.2	77.1	91.7	**85.6**	47.3	93.1	59.4	54.2	38.7	81.9	**56.1**	**69.9**

TABLE III
EVALUATION OF ABLATION STUDIES ON PASCAL VOC 2012 TEST SET

	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Mbike	Person	Plant	Sheep	Sofa	Train	Tv	mIoU
FCN8 [10]	76.8	34.2	68.9	49.4	60.3	75.3	74.7	77.6	21.4	62.5	46.8	71.8	63.9	76.5	73.9	45.2	72.4	37.4	70.9	55.1	62.2
PSP-CRF [32]	81.4	35.7	71.8	65.8	73.1	82.1	79.0	80.5	32.3	65.6	40.7	70.3	68.1	71.2	72.9	29.3	74.4	45.0	77.4	63.4	65.4
Zoom Out [33]	85.6	37.3	83.2	62.5	66.0	85.1	80.7	84.9	27.2	73.2	57.5	78.1	79.2	81.1	77.1	53.6	74.0	49.2	71.7	63.3	69.6
DeepLab1 [17]	84.4	54.5	81.5	63.6	65.9	85.1	79.1	83.4	30.7	74.1	59.8	79.0	76.1	83.2	80.8	59.7	82.2	50.4	73.1	63.7	71.6
DeConvNet [4]	89.9	39.3	79.7	63.9	68.2	87.4	81.2	86.1	28.5	77.0	62.0	79.0	80.3	83.6	80.2	58.8	83.4	54.3	**80.7**	65.0	72.5
GCRF [34]	85.2	43.9	83.3	65.2	68.3	89.0	82.7	85.3	31.1	79.5	63.3	80.5	79.3	85.5	81.0	60.5	85.5	52.0	77.3	65.1	73.2
DPN [35]	87.7	59.4	78.4	64.9	70.3	89.3	83.5	86.1	31.7	79.9	62.6	81.9	80.0	83.5	82.3	60.5	83.2	53.4	77.9	65.0	74.1
Piecewise [36]	90.6	37.6	80.0	67.8	74.4	92.0	**85.2**	86.2	39.1	81.2	58.9	83.8	83.9	84.3	84.8	62.1	83.2	58.2	80.8	72.3	75.3
FuzzyNet (Ours)	**91.2**	**61.1**	**85.5**	**68.1**	**75.3**	**92.5**	84.6	**87.9**	**40.2**	**88.4**	**66.4**	**86.5**	**87.2**	**85.6**	**85.2**	**64.7**	**88.2**	**61.5**	80.2	**73.0**	**78.3**

First, the Atrous convolutions ensures that receptive field of views are enlarged without increment in number of parameters, effectively reducing about 2% of computational time compared to the baseline. Conjointly, it incorporates larger context which extracts long range information and reduces blurriness as depicted in Fig. 6.

- PASCAL VOC 2012 Dataset: Likewise, the introduction of atrous spatial pyramid pooling allows control of the feature resolution responses by segmenting objects at different scales. Compared to the baseline which doesn't include Atrous Spatial Pyramid Pooling (ASPP), object classes having big scales such as airplane, train and boat have increased 3% segmentation accuracy as displayed in Table 2. Furthermore, smaller objects are observed to have about 2% improvement. We argue that the atrous spatial pooling multiple scaling sizes encode image context in dimensions which is beneficial to objects with several ranges on size unlike classical pooling which is limited to a fixed size. Evaluating on the test set scores 74.8%. The visualization effects are well depicted in Fig 6.
- CamVid Dataset: Our model effectively demonstrates pixel-wise prediction accuracy compared to the baseline as very large object class such as fence, pavement, road, building and sky are well labeled due to increased receptive field. Also, the smaller object classes benefit from the smaller receptive field scales. Imperatively, our context module boosts an improvement of 6% to the baseline for encoding densely connected features between the layers. The different variation and transformation at each layer are adequately complemented by enriched spatial information from previous feature-maps which infuses dense connections, representing a global context which preserves and captures enough spatial information. The consideration of different feature levels is complimented with scaling capable of extracting necessary information. This attention inclusion achieves an improvement boost of 65.7% on the test set, outperforming several other recent approaches.

3) FUZZYNET: In addition to GCSM, we included the Spatial Fuzzy Refinement Function. From the results shown, it demonstrates clearly that our proposed model achieves the highest result as shown in Fig. 6.

- PASCAL VOC 2012 Dataset: Mostly, the pixels at the edges of objects or the object boundary returns noisy predictions due to misalignment with other objects' pixels or the background regions. Introducing the post processing spatial fuzzy function ensures 9% blob elimination compared to our baseline and 4% compared to our GCSM model. Because of pooling which occurs during convolution, the reduced resolution often misaligns labels especially at object boundaries and intersection with other objects and background. As such, localization at the edges becomes noisy and blurry. Apparently, incorporating a

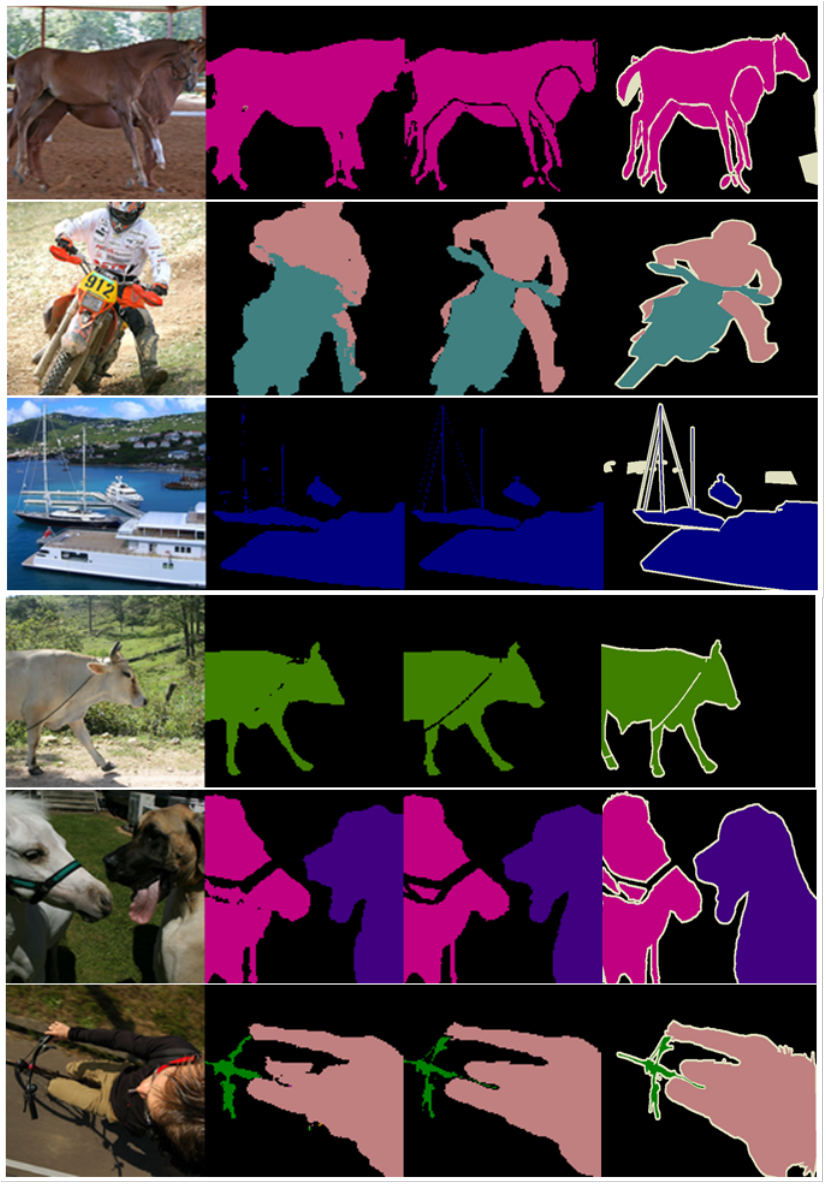

Fig. 6. Visualization results from our proposed FuzzyNet architecture on PASCAL VOC set. From left to right: input image, GCSM prediction, FuzzyNet prediction and ground truth.

fuzzy neighborhood window allows the boundary pixels to ensure segmentation into different classes by finding the Euclidean distance of the pixel to the constituent classes around. The fuzziness allows sharing of pixel membership based on proximity but ultimately classifying each pixel to a single class with which the highest probability is attributed. This ensures increment in localization accuracy as well as elimination of blurs to a test evaluation of 78.3% mIoU. The boundary and localization increment based on the fuzzy iterations are depicted in Fig. 6.

- CamVid Dataset: At the object's boundary, pixels inconsistencies affected localization severely with the baseline, this is due to consistent resolution reduction in convolution. We fine-tune the image boundaries with iterative pixel distance computation for each class. Different window sizes ranging from 3 to 15 are tried. smaller window sizes show little improvement over the baseline while window size from 10 and above shows same results. We therefore settled for window size of 10. The iterations are set to 50 after several experiments which indicates consistent results after the 50th iteration. Our proposed model shows a well refined boundary which translates into improved localization accuracy of 4% increase compared to SGMFCN and 9% compared to the baseline. The results of this is as well depicted in Table 3 and Fig. 6 with a mIOU of 69.9%.

E. Implementation

We trained our model using MatLab R2018b and Keras 2.2.4 on Intel Core (TM) i7-8700 CPU, 16GB RAM computer with a single NVIDIA GeForce GTX 1080 Ti graphics card. This took a total period of 7 hours.

V. CONCLUSION

In this paper, we achieved semantic image segmentation task by applying deep convolutional neural network backbone alongside a novel global context spatial module with a refinement function. Particularly, we exploit the concept of human visual cortex mechanism to encode attention as well as preserve neural nets spatial capability. Using a supervise means of learning, our model learns prelabeled dataset to extract low level and high-level features of images, subsequently classifying such pixels into their individual classes. To overcome the effect of convolutional pooling, we refined our model's output by introducing a spatial fuzzy function. We investigated the efficiency and effectiveness of our design by running ablation studies and evaluation using widely recognized PASCAL VOC 2012 and CamVid dataset, yielding highly competitive segmentation results.

REFERENCES

[1] X. Liu, Z. Deng, and Y. Yang, "Recent progress in semantic image segmentation," *Artificial Intelligence Review,* pp.1–18, 2018.

[2] L. Perez, and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," 2017 [Online]. Available: https://arxiv.org/abs/1712.04621,

[3] L. Fan, W. Wang, F. Zha, J. Yan, "Exploring New Backbone and Attention Module for Semantic Segmentation in Street Scenes," *IEEE Access,* PP. 1-1. 10.1109/2880877, 2018.

[4] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," *2015 IEEE International Conference on Computer Vision (ICCV),* pp. 1520–1528,2015.

[5] S. Jegou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* pp. 1175–1183, 2017.

[6] H. Zhou, A. Han, H. Yang and J. Zhang, "Edge gradient feature and long distance dependency for image semantic segmentation," In *IET Computer Vision,* 2018.

[7] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," In *ECCV,* 2018.

[8] L. Fan, H. Kong, W. C. Wang, and J. Yan, "Semantic segmentation with global encoding and dilated decoder in street scenes," *IEEE Access,* 6:50333–50343, 2018.

[9] J. Dai and X. Tang, "ResFusion: deeply fused scene parsing network for RGB-D images," In *IET Computer Vision,* 2018, 12, pp. 1171-1178.

[10] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *2015 IEEE Conference on Computer Vision and Pattern Recognition(CVPR),* pp. 3431–3440, 2015

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," In *MICCAI,* 2015.

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

[13] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *CoRR,* abs/1706.05587, 2017.

[14] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 6230–6239,2017.

[15] G. Lin, A. Milan, C. Shen, and I. D. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 5168–5177, 2017.

[16] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters improve semantic segmentation by global convolutional network," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 1743–1751, 2017

[17] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," In *International Conference of Learning Representations (ICLR),* 2015.

[18] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," *CoRR,* abs/1506.04579, 2015.

[19] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for semantic segmentation in street scenes," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition,* pp. 3684–3692, 2018

[20] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *CoRR,* abs/1606.02147, 2016.

[21] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "Icnet for real-time semantic segmentation on high-resolution images," In *ECCV,* 2018.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp.770–778, 2016

[23] 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 Transactions on Pattern Analysis and Machine Intelligence,* 40:834–848, 2018.

[24] C. Keh-Shih, T. Hong-Long, C. Sharon, W. Jay, and C. Tzong-Jer, "Fuzzy c-means clustering with spatial information for image segmentation," *Computerized Medical Imaging and Graphics,* Volume 30, Issue 1, Pages 9-15, 2006.

[25] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.,* vol. 88, no. 2, pp. 303 338, Sep. 2009.

[26] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in *Proc. Eur. Conf. Comput. Vis. (ECCV),* Marseille, France, Oct. 2008, pp. 44 57.

[27] N. Goyette, P. M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, "Change detection.net A new change detection benchmark dataset," In *IEEE CVPR Workshops,* pages 1–8. , 2012

[28] C. Wu, H. Cheng, S. Li, H. Li and Y. Chen, "ApesNet: a pixel-wise efficient segmentation network for embedded devices," In *14th ACM/IEEE Symposium on Embedded Systems For Real-time Multimedia,* pages 1–7. , 2016

[29] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," *IEEE Vis. Commun. Image Process.,* pp. 1-4. Dec. 2017.

[30] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," [Online]. Available: https://arxiv.org/abs/1511.07122. 2015

[31] C. Yu, J. Wang, G. Peng, C. Gao, G. Yu and Nong Sang, "BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation," *European Conference on Computer Vision.,* 2018

[32] L. ZhangH, Li, P. Shen, G. Zhu, J. Song, S. Shah, M. Bennamoun, and L. Zhang, "Improving semantic image segmentation with a probabilistic superpixel-based dense conditional random field," *IEEE Access* vol. 6 pp. 15297-15310 2018.

[33] M. Mostajabi P. Yadollahpour G. Shakhnarovich, "Feedforward semantic segmentation with zoom-out features," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* pp. 3376-3385 2015.

[34] R. Vemulapalli, O. Tuzel, M.-Y. Liu, and R. Chellapa, "Gaussian conditional random field network for semantic segmentation," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pages 3224–3233, 2016.

[35] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang, "Semantic image segmentation via deep parsing network," In *Proceedings of the IEEE International Conference on Computer Vision,* pages 1377–1385, 2015.

[36] G. Lin, C. Shen, A. van den Hengel, and I. Reid, "Efficient piecewise training of deep structured models for semantic segmentation," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pages 3194–3203, 2016.