International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-10, August 2019

Automatic Road Segmentation from High Resolution Satellite Images Using Encoder-Decoder Network

Naveen Pothineni, Praveen Kumar Kollu, Suvarna Vani Koneru

Abstract*: Road network segmentation from high resolution satellite imagery have profound applications in remote sensing. They facilitate for transportation, GPS navigation and digital cartography. Most recent advances in automatic road segmentation leverage the power of networks such as fully convolutional networks and encoder-decoder networks. The main disadvantage with these networks is that they contain deep architectures with large number of hidden layers to account for the lost spatial and localization features. This will add a significant computational overhead. It is also difficult to segment roads from other road-like features. In this paper, we propose a road segmentation architecture with an encoder and two path decoder modules. One path of the decode module approximates the coarse spatial features using upsampling network. The other path uses Atrous spatial pyramid pooling module to extract multi scale context information. Both the decoder paths are combined to fine tune the segmented road network. The experiments on the Massachusetts roads dataset show that our proposed model can produce precise segmentation results than other state-of-the-art models without being computationally expensive.

***Index Terms*: Convolutional networks, Encoder-Decoder, Road Network, Segmentation.**

I. INTRODUCTION

Road Extraction from satellite images is a complex and essential task in geographic information system (GIS). It has profound applications in transportation networks, disaster management, cartography and GPS navigation. Most often the road network is added as an overlay layer for navigation in GIS applications such as Google Maps. They are also used to find the change detection where the road deterioration can be found out by taking the road structure from two different time points. Recently they are applied in finding the road structures in rural areas or inaccessible geographical areas.

Most of the Road extraction approaches are either Manual or Automatic. Manual segmentation of roads requires a lot of labor and often very time consuming. It is almost not feasible and cost efficient for manual road extraction. Although automatic and semi-automatic approaches are better than the manual approaches they also have several shortcomings due to the complexity of the road structures. They suffer from occlusion of objects. Also the variations in viewpoints, sensors and resolution of capturing are also the reason for performance degradation of the methods. With the advancements in the area of deep learning especially in image classification they have become the best viable option for image segmentations tasks by achieving close to real and state of the art performance. Recently there have been many proposed approaches based on deep learning techniques. They have better performance and accuracy than the previous methods. But the problem with these approaches is that it is very difficult to train deeper neural networks especially with a large number of hidden layers as after some layers the signal will vanish almost completely. Also the architecture can accurately classify at only some regions of the images due varying sizes of operations.In this research, a novel architecture for accurate road segmentation by using a single encoder and two path decoder network is being proposed. The contributions of the proposed architecture are in three-fold. First, by using Depthwise separable convolutions throughout the network will increase the computation efficiency. Second, by leveraging Atrous spatial pyramid pooling(ASPP) additional context information with multiscale feature maps are obtained. These can help the network attain the lost context information. Third, the upsampling decoder module and ASPP module combination ensures that the network has the necessary spatial and semantic localizations to produce smooth roads network maps.

II. RELATED WORK

Initial application of deep neural networks for road network extraction was proposed by [1]. Their work employs restricted Boltzmann machines for automatic segmentation of road areas. Fully convolutional networks [2,3] interprets classification networks as fully convolutional networks that produces segmentation maps. The classification is performed at pixel level between two classes i.e. road and non-road. Most of the literature on road segmentation has been focused employing U-Net and its variants [4]. U-Net was initially proposed for medical image segmentation [5] and since been used in remote sensing applications. The architecture contains skip connections between the contraction and expansion path for semantic and spatial propagation. [6] combined residual units with U-Net architecture called residual U-Net (ResUnet) for semantic road segmentation. It helps facilitate information propagation and can handle vanishing gradient problem.

Manuscript published on 30 August 2019.
***Correspondence Author(s)**
Naveen Pothineni, CSE Department, Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada, India.
Praveen Kumar Kollu, CSE Department, Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada, India.
Suvarna Vani Koneru, CSE Department, Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada, India.

III. METHODOLOGY

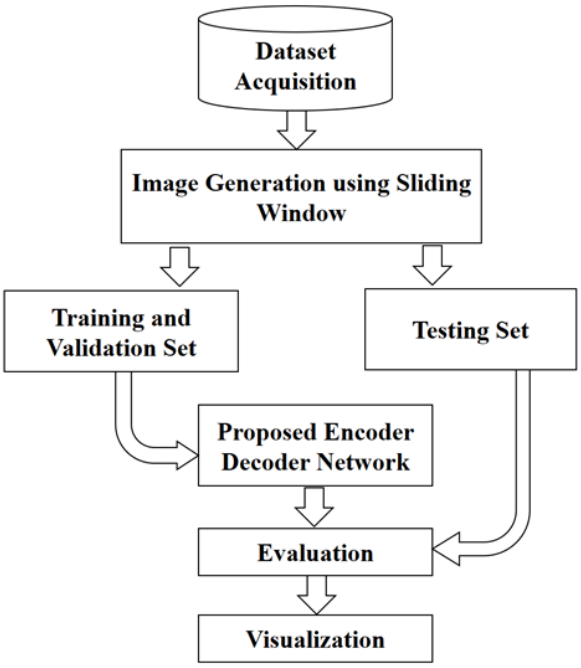


Fig 1. Proposed Methodology Diagram

A. Dataset

The Massachusetts road dataset is widely used benchmark dataset for road segmentation models. It contains images captured from Massachusetts region. The dataset contains a total of 1171 satellite images for training with a resolution of 1500 x 1500. Some of the images in the dataset contains blank regions which are not suitable for training. For initial preprocessing we have manually removed the images and their corresponding mask labels that contains more than 50% of blank area. As the number of images is very limited for generalizing the model, sliding window approach was used to generate more training samples.

B. Image Generation

First the images and the masks were upsampled in to 1536 x 1536 across all channels. Then a sliding window with a size of 256 x 256 is used to segment the images with no overlapping. This approach generated 27108 training samples and their corresponding mask labels. Random sampling is also used to generate around 5000 samples. The validation and test set were also generated using the same approach which generated 564 and 1764 samples respectively. The sliding window approach has two advantages: 1) We can generate the segmented masks in their original dataset resolution and 2) There is no need to use data augmentation while training which is computationally expensive. All the mask labels for training are normalised in to 0,1 range.

C. Atrous Convolution

Atrous convolutions (also called Dilated Convolutions) are type of convolution operations which have an additional parameter called dilation rate. Dilation rate is the stride between the values in a single kernel. They are used to control the wider field-of-view while taking the same number of parameters. They can minimize multiple convolutions in the network which are computation heavy. The atrous convolution for an input vector x and output vector y for each location i is given as:

$$y[i] = \sum_{k} x[i + r.k]w[k] \quad (1)$$

Where w is the convolution filter and r is the dilation rate. The value r = 1 signifies a standard convolution operation. Fig.2 shows the atrous convolution operation.

Atrous spatial pyramid pooling (ASPP) is the atrous convolution version of spatial pyramid pooling, where atrous convolutions with different dilation rates are fused together. ASPP can adjust the field of view at multiple scales which helps capture the context feature maps.

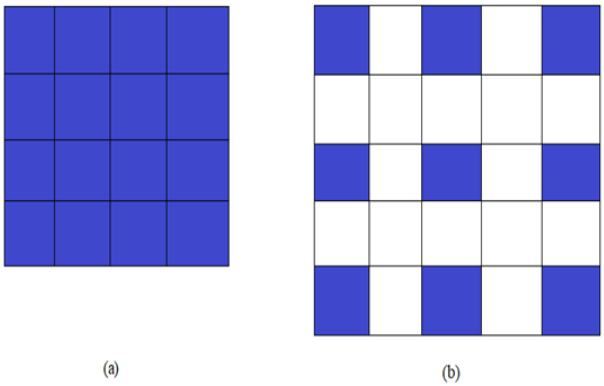


Fig 2. Atrous convolution. (a) Standard convolution with r =1 and (b) Atrous convolution with r = 6

D. Encoder Decoder Network

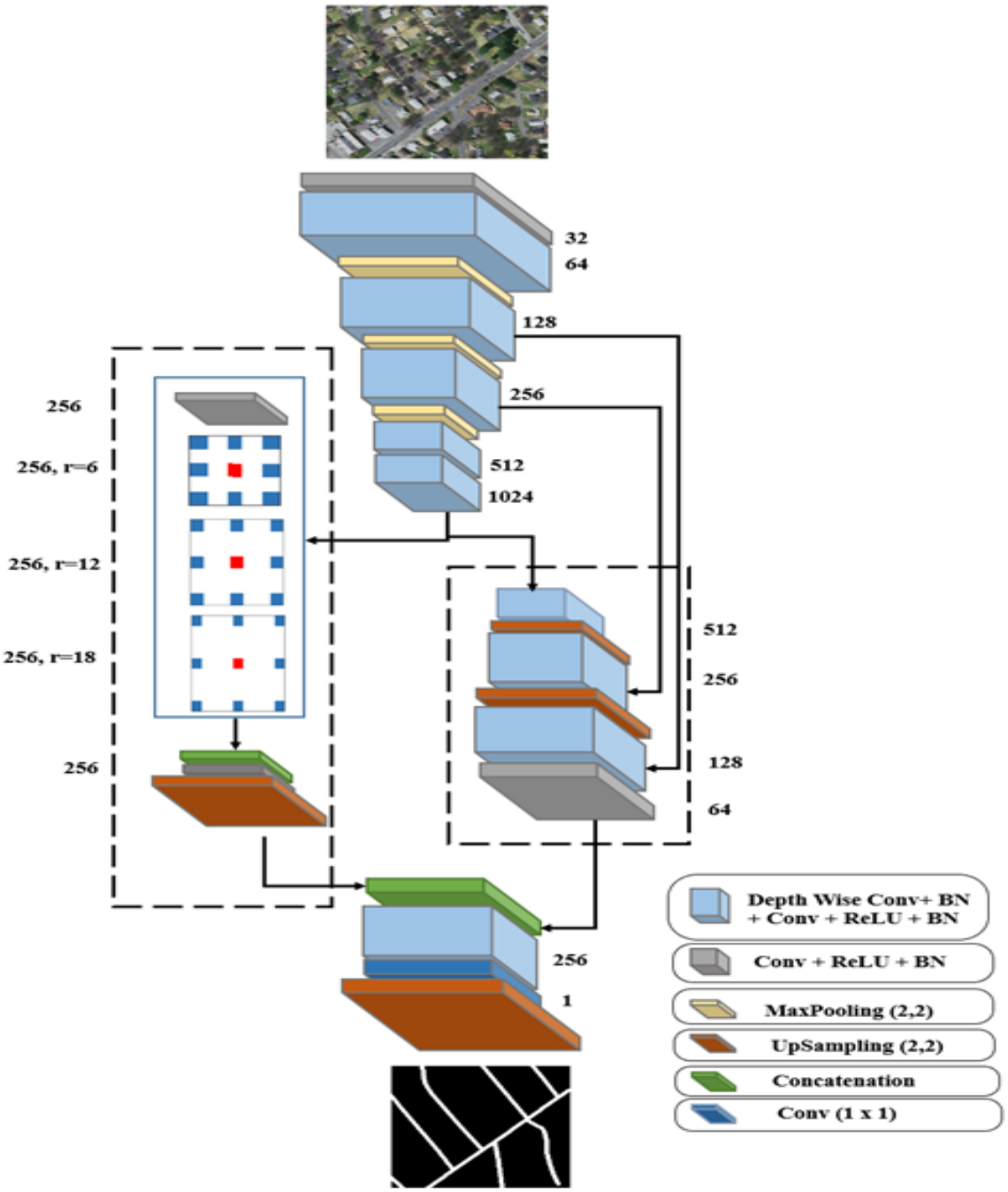
The architecture of the proposed model is shown in Fig 3. The network contains a minimal architecture with a single encoder and two path decoders. After the initial convolution the encoder part contains five Depth Wise Convolution block with increasing number of filters. Each block contains Depthwise convolution, Batch Normalization, Normal Convolution, Batch Normalization and ReLU activation layers respectively. Depth Wise convolution contains a constant filter size of (3 x 3) throughout the network while the convolution layer contains (1 x 1) filter size. After each block MaxPooling layer is applied to reduce the dimensionality. The encoder part extracts the semantic features from the images. These feature maps are given to the two decoder paths.ASPP decoder path contains a normal convolution and three atrous convolutions. It is noted that all the convolution layers take same feature maps from the encoder part. Dilation rates of 6, 12, 18 are used in the atrous convolutions. All these convolutions are concatenated to fuse the features. A normal convolution is performed along with the upsampling layer to make features compatible with the other decoder part. ASPP decoder path adds the context information and localization features.

Retrieval Number: J91320881019/19©BEIESP
DOI: 10.35940/ijitee.J9132.0881019
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

The Upsampling decoder path follows similar structure to the encoder part. It contains three Depth Wise convolution blocks with Upsampling layer between them. Apart from the initial block other two blocks contains skip connections directly from the encoder. This resembles a U-Net like architecture. Finally, both the decoder parts are concatenated. Both parts contain the semantic and spatial features required for accurate segmentation. Through the network a upsampling of factor 2 is used which is not an aggressive size. After the final two convolutions final layer in the network contains a convolution layer with a single filter of size (1 x 1) along with a sigmoid layer which produces the final segmentation map.

Fig 3. Proposed Encoder-Decoder Network Architecture



IV. IMPLEMENTATION DETAILS

TensorFlow framework was used to implement the proposed and other three baseline networks. Model optimization plays a crucial role in achieving best results. Adam is used as the optimizer to train the network. It has considerably faster training and convergence rates. The initial learning rate was fixed as 0.001. If there is no significant decrease in loss

value, the learning rate was reduced by a factor of (0.1). 37108 images were used to train the network and tested on 1764 images. Training was performed on NVIDIA GTX 1080-Ti GPU with a batch size of 10.

Retrieval Number: J91320881019/19©BEIESP
DOI: 10.35940/ijitee.J9132.0881019
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

A. Evaluation Metrics

To validate the results of the segmentation, mean Dice coefficient, mean intersection over union (meanIOU) and accuracy measures are calculated. The equations for these measures are given as:

$$meanIOU = Mean\left(\frac{TP}{TP+FP+FN}\right) \quad (2)$$

$$DiceCoefficient = \frac{2*TP}{(TP+FP)+(TP+FN)} \quad (3)$$

$$MeanAccuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

where TP, TN, FP and FN are True Positive, True Negative, False Positive and False Negative respectively.

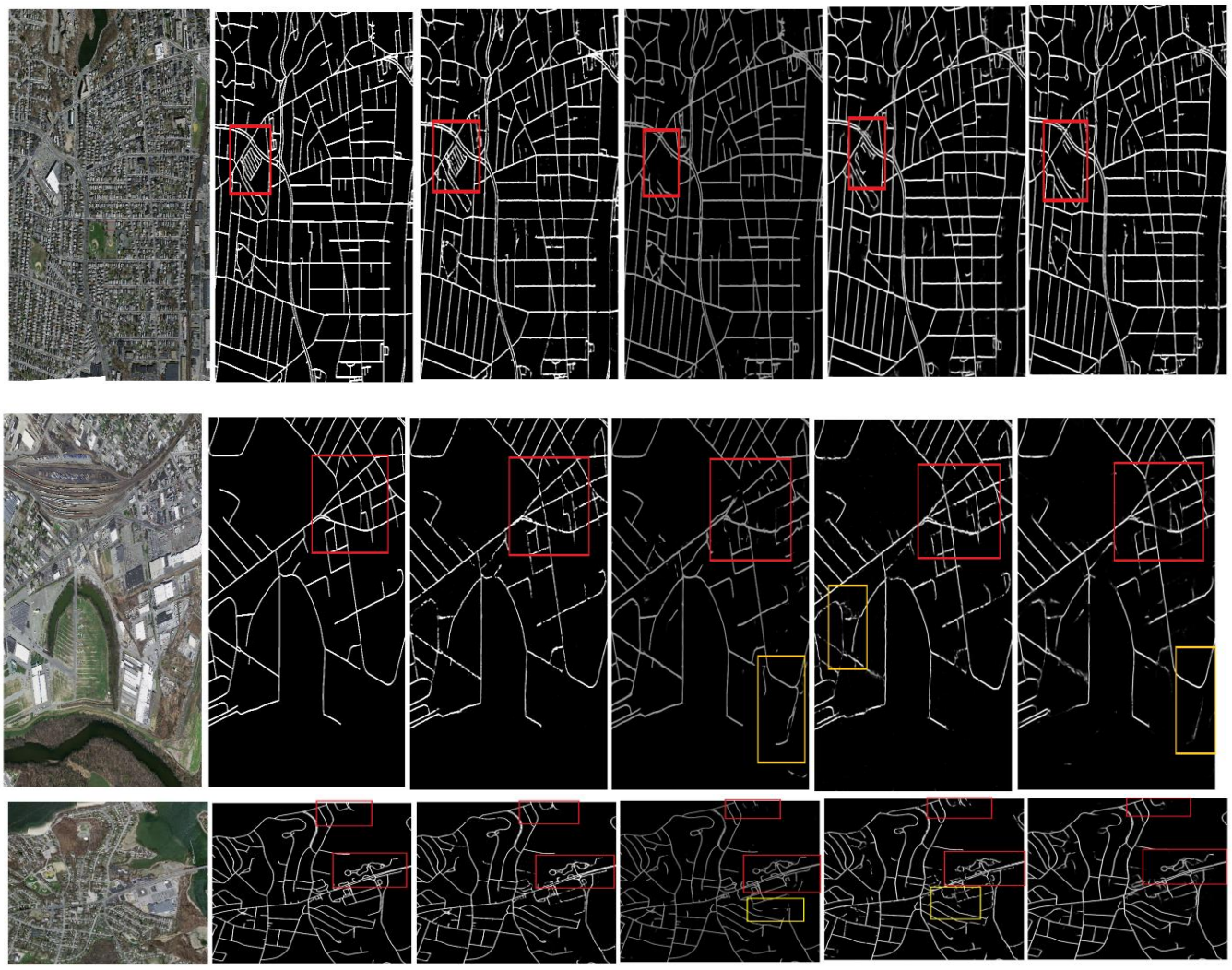


Fig 4. Predictions results on Test set. From left to right

a) Satellite Image b) Ground Truth c) Our proposed method d) U-Net e) ResUnet f) SegNet. Superior performance of our proposed model in shown in red box and yellow box shows the wrongly predicted non-road pixels as road pixels.

V. RESULTS

The Evaluation results summary is given in Table 1. It can be seen from the results that our proposed network achieves superior results in meanIOU and Mean Dice Coefficient. Although ResUnet achieves better result in mean Accuracy, it is marginal to the proposed network result. As given in Table 2 our proposed approach achieves these results with least number of training parameters, which makes our model more efficient and computationally inexpensive. The U-Net architecture with 30.8 Million trainable parameters had least performance on the evaluation metrics. Predictions performed on all the models and the corresponding ground truth labels are given in Figure 4. Our model was able to predict the complex structures and interconnections in road network. U-Net gave predictions with least noise but was not able to predict complex

structures. ResUnet and SegNet gave consistent results but also gave segments that were not in the ground truth labels. From these results it can be seen that our method has given better road segments with cleaner edges. The ASPP module in our method helped the network in considering the context information as to better distinguish between roads and other similar objects.

Retrieval Number: J91320881019/19©BEIESP
DOI: 10.35940/ijitee.J9132.0881019
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

Table 1. Performance on Evaluation Metrics of Testing Set

Method	meanIOU	Mean DC	Mean Accuracy
U-Net	0.7242	0.7912	0.7527
ResUnet	0.7761	0.8172	0.7947
SegNet	0.7644	0.7950	0.7811
Ours	0.7836	0.8218	0.7985

Table 2. Approximate training parameters (in Millions)

Method	Parameters
U-Net	30.8
ResUnet	8.2
SegNet	29.4
Ours	1.6

VI. CONCLUSION

In this paper, an encoder and two path decoder network architecture is proposed. The two path decoders can effectively capture the semantic and spatial information to produce accurate road segments. The ASPP decoder module can adjust the filters field of view and retains greater context information. This context information is combined with the upsampling decoder module for spatial localization. The experiments show that our proposed network produces superior results than the state of the art models. Using depthwise separable convolutions throughout the network significantly increases efficiency and minimizes computational costs. The proposed model contains fewer parameters than other deeper models. Our work can be utilized to create faster and accurate road maps for various applications such as GPS navigation and digital cartography.

REFERENCES

1. Mnih, Volodymyr & E. Hinton, Geoffrey. 2010. ``Learning to Detect Roads in High-Resolution Aerial Images." European Conference on Computer Vision. Lecture Notes in Computer Science. Berlin, Heidelberg.
2. Alshehhi. Rasha , Prashanth Reddy Marpu, Wei Lee Woon, Mauro Dalla Murai. 2017. ``Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks." ISPRS Journal of Photogrammetry and Remote Sensing. 130 : 139--149.
3. Buslaev, Alexander V., Selim S. Seferbekov, Vladimir I. Iglovikov and Alexey A. Shvets. 2018. ``Fully Convolutional Network for Automatic Road Extraction from Satellite Imagery.' 'IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Salt Lake City, USA. June 18–22.
4. Constantin, A, J. Ding and Y. Lee. 2018 ``Accurate Road Detection from Satellite Images Using Modified U-net." IEEE Asia Pacific Conference on Circuits and Systems (APCCAS). Chengdu: 423-426.
5. Ronneberger, O., P. Fischer, and T. Brox. 2015. ``U-Net: Convolutional Networks for Biomedical Image Segmentation." International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany. October 5–9.
6. Z Zhang, Q. Liu and Y. Wang. 2018. ``Road Extraction by Deep Residual U-Net." IEEE Geoscience and Remote Sensing Letters. 15 (5): 749-753.