Semantic Segmentation Using Deep Learning

Shubham Singh[1], Sajal Kaushik[2], Rahul Vats[3], Arihant Jain[4], and Narina Thakur[5]

[1]*Bharati Vidyapeeth's College of Engineering, A-4, Paschim Vihar, New Delhi 110063*

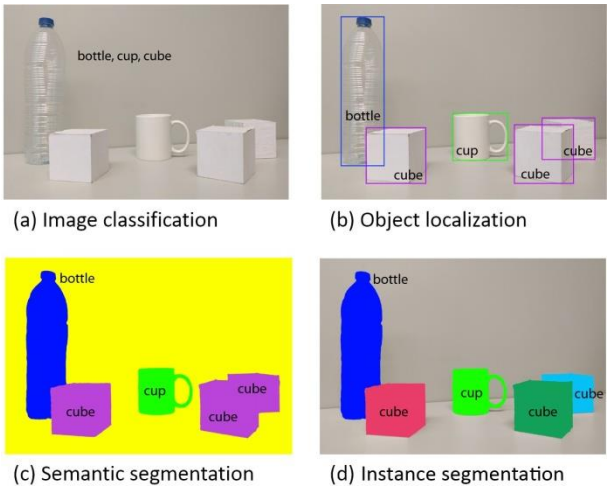
***Abstract*—** Semantic image segmentation is an essential com- ponent of modern autonomous driving systems, as an accurate understanding of the surrounding scene is crucial to navigation and action planning. Current state-of-the-art approaches in semantic image segmentation rely on pre-trained networks that were initially developed for classifying images as a whole. While these networks exhibit outstanding recognition performance, they lack localization accuracy. Therefore, additional memory intensive units have to be included in order to obtain pixel- accurate segmentation masks at the full image resolution. To alleviate this problem we Implemented various Standard Models such as GCN, DeepLabV3, PSPNet and FC-Densenet on CamVid image frames dataset, try to optimize them and then we proposed a novel FRRN based architecture that exhibits strong localization and recognition performance. We combine multi- scale context with pixel-level accuracy by using four (two as of in FRRN) processing streams within our network: One stream carries information at the full image resolution, enabling precise adherence to segment boundaries. Other streams undergoes a sequence of pooling operations to obtain robust features for recognition. The two streams are coupled at the full and half image resolution using residuals. Our approach achieves an intersection-over-union score of 0.87 on the CamVid dataset.

I. INTRODUCTION

Semantic segmentation is an important aspect of image analysis task and a key problem in Computer vision. It describes the process of associating each pixel of an image with a class label like car, bus, road, pole, etc. Semantic segmentation is widely used in autonomous driving, medi- cal image segmentation, Geo-Sensing, Facial segmentation, Precision Agriculture, Human-Machine interaction, Image search engines and many more. These problems have been solved using traditional Machine Learning and Computer Vision techniques but advancements in Deep learning tech- nology have created lot of space to improve them in terms of accuracy and efficiency.

Semantic segmentation is more informative than image classification and object localization. While image classifica- tion tells about the presence of an object in image and object localization locates the objects by making bounding boxes around them before classification, semantic segmentation classify each and every pixel of objects in image. Also, Instance segmentation is similar to semantic segmentation but it also classify different instances of a class within a image, like two cars in a image.

Semantic segmentation not only predicts classes of objects but also tells about spatial location of those classes in image. Further, different instances of same class can also be classified and also components of already segmented classes can also be classified. But this paper focus only on general



(a) Image classification (b) Object localization

(c) Semantic segmentation (d) Instance segmentation

Fig. 1.

per-pixel classification, i.e., same labels are given to different instances of same class and if they are overlapped their boundary is not distinguished as shown in Fig. 1 (c).

While image segmentation groups similar pixels of class together, in Video segmentation disjoint sets of consecutive and homogeneous frames are segmented that exhibit coher- ence in both motion and appearance. To segment dynamic scenes of a video in high quality, deep learning models paved way to achieve better performance than the traditional algo- rithms. Video segmentation is useful in activity recognition and other visual enhancements.

II. RELATED WORK

A. *BiSeNet: Bilateral Segmentation Network for Real-time Semantic Segmentation*

BiseNet performs real time semantic segmentation by taking into account the contextual and spatial features. The spatial path with a small stride preserves the spatial informa- tion and generate high-resolution features. And the Context Path with a fast downsampling strategy is used to get enough receptive field which runs parallelly to the spatial path. These fusion of two paths results in better accuracy without loss of speed termed Feature Fusion Model (FFM). Also a Attention Refinement Model refines the features of each stage by using global average pooling.[1]

B. *SegNet: A Deep Convolutional Encoder-Decoder Archi- tecture*

In this semantic pixel wise segmentation is done termed as SegNet. The architecture consist of a Encoder similar to the convolutional layers in the VGG16 network and a Decoder followed by pixel-wise classification layer. Here Encoder performs convolutions of the given input to get set of features which are normalized in batches. Further ReLu is applied in input data element wise which is pooled by max pooling followed by sub sampling of the result.[2]

C. *MobileNets for Semantic Segmentation*

This model is based on depth-wise separable convolutions. It is a type of factorized convolution which factorize a standard convolution into a depth-wise convolution and a 11 convolution called a point-wise convolution. A standard convolution simultaneously filters and combines input into a new set of outputs in a single step, whereas the depth-wise separable convolution does this in two layers, a separate layer for filtering and a separate layer for combining.[3]

D. *RefineNet: Multi-Path Refinement Networks for High- Resolution Semantic Segmentation*

It is a generic multi-path refinement network which uses long-range residual connections for high-resolution predic- tions. Here, multiple paths over which information from different resolutions and via potentially long-range connec- tions, is assimilated using a generic building block, termed RefineNet. The deep layers captures high level semantic features thus are refined using fine-grained features resulted from earlier convolutions.[4]

E. *ICNet for Real-Time Semantic Segmentation*

Image Cascade Network (IcNet) incorporates multi- resolution branches under proper label guidance. Here, cas- cade image inputs i.e., images with varying resolution are used and cascade feature fusion (CFF) unit is employed. Input image with full resolution is downsampled by factors of four and two, which acts as a cascade input to branches of high and medium resolution. CFF is used for combining cascade features from inputs of various resolution.[5]

III. DATASET AND PREPROCESSING

The dataset is taken from the Cambridge-driving Labeled Video(CamVid) Database. It is collection of videos with object class semantic labels. The CamVid dataset consists of: the original video sequences, the list of class labels and the hand labeled frames.

It provides ten minutes of 30Hz footage with correspond- ing semantically labeled images at 1Hz.

Dataset consist of 6 directories: train, train labels, test, test labels, val and val labels. Labelled directories consist of labelled data and other directories consist of actual images without labels.[6]

A. Scaling

This step is done to make sure that images have same size and aspect ratio. After this, we scale the image as per the model.

B. Normalization

It is used to makes the convergence faster during the time of training the model. It is done by subtracting the mean from each pixel and then dividing the result by the standard deviation as a result distribution will resemble Gaussian curve with mean at zero. The pixel intensity will now lie in range [0,1].

IV. APPROACHES

In this paper, we implement U-net, Dilated U-net and PSPnet.

A. U-net

U-net architecture consist of 2 paths which are basically known to be as encoder and decoder. Encoder is also known as contraction path and decoder is also known as symmetric expanding path.

Encoder captures the context in the image. Here, convolu- tion blocks followed by a maxpool downsampling to encode the input image into feature representations at multiple different levels is applied. Hence, it is stack of convolutional and max pooling layers. It is also called downsampling.

Decoder consists of upsample and concatenation followed by regular convolution operations. It enable precise localiza- tion using transposed convolutions.

Hence it is an end-to-end fully convolutional network. It contains Convolutional layers only and does not contain Dense layer because of which it can accept image of any size.

In U-net, pooling layers increase the field of view and are able to aggregate the context while discarding the where information and semantic segmentation requires the exact alignment of class maps and thus, needs the where informa- tion to be preserved. Hence, U-net is preferred here.[7]

We have hyper-tweaked the U-net model to improve the IoU metric.

B. Dilated U-net

In this model, the simple convolutions are replaced by dilated convolutions. Dilated convolutions are also called atrous convolutions.

For 1D signal x[i], the y[i] output of a dilated convolution with the dilation rate r and a filter w[s] with size S is formulated as:
Let x[i] be a 1D signal, r be the dilation rate and filter w[s] with size S. The output y[i] of the dilated convolutions is:

$$y[i] = \sum_{s=1}^{S} x[i + r.s]w[s]$$

Dilated convolutions bring translated variant in input as:

$$f(g(x)) = g(f(x))$$

where g(.) is convolution operation and f(.) is translation operation.

This will help in reducing the parameters massively be- cause receptive fields grow aggressively. Along with this, pooling helps in pixel wise classification. But pooling layer decreases the resolution of the input images as a result dilated U-net model does not works well.[8]

C. *PSPnet*

PSPnet is abbreviation for Pyramid Scene Parsing Net- work. This model propose pyramid pooling module to join the context of the image hence, called PSPnet. Dilated convolutions are used to modify Resnet and a pyramid pooling module is added to it. Pyramid pooling module captures information by applying large kernel pooling layers. This module concatenates the feature maps from ResNet with upsampled output of parallel pooling layers with kernels covering whole, half of and small portions of image.

Also an auxiliary loss is applied after the fourth stage of ResNet (i.e input to pyramid pooling module), called as intermediate supervision.

The resolution of image is also preserved in PSPnet because it uses large pooling layer.[9]

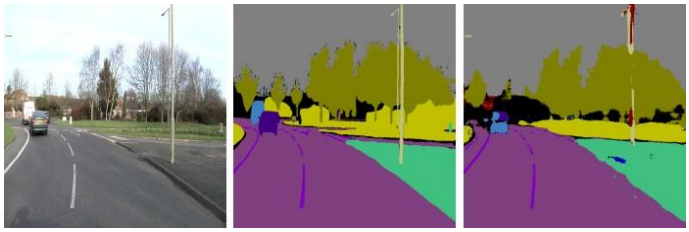


Fig. 2. Input Image - Ground Truth - Predicted Image (PSPNet)

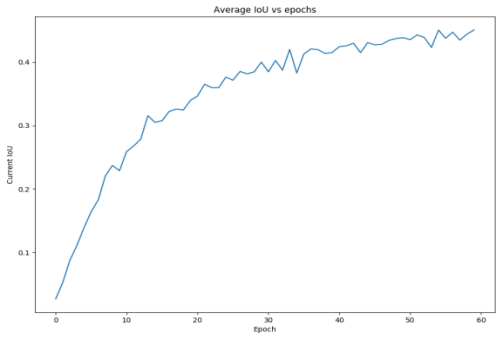


Fig. 3. IoU vs Epoch (PSPNet)

D. *Fully Convolutional DenseNets*

It is a a CNN with Densely Connected Convolutional Networks, called as DenseNets. It is based upon the fact that if each layer is directly connected to every other layer in a feed-forward fashion then the model will become more

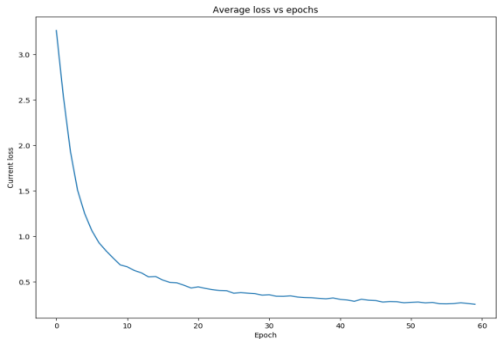


Fig. 4. Loss vs Epoch (PSPNet)

accurate and will be easier and efficient to train. They have various advantages such as they reduce the vanishing-gradient problem, strengthen feature propagation, encourage feature reuse, and reduce the number of parameters.

FCNs are built from a downsampling path, an upsam- pling path and skip associations. Skip associations help the upsampling path recover spatially detailed data from the downsampling path, by reusing features maps.

The objective of the model is to further exploit the feature reuse and avoiding the feature explosion at the upsampling path of the network.[10]

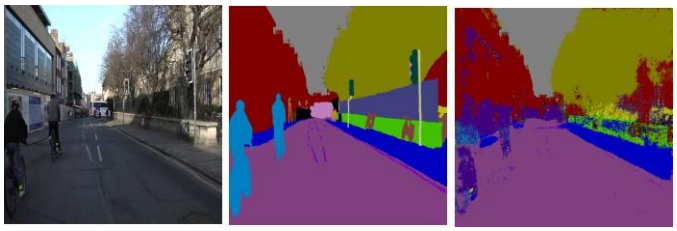


Fig. 5. Input Image - Ground Truth - Predicted Image (FC DenseNet)

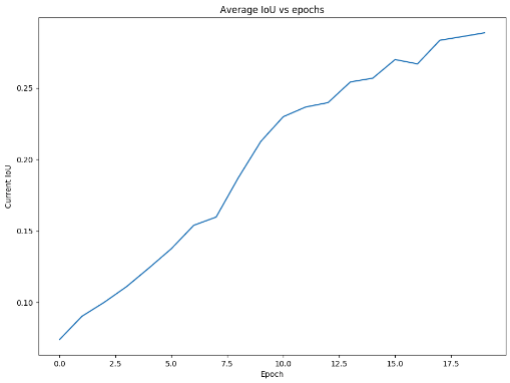


Fig. 6. IoU vs Epoch (FC DenseNet)

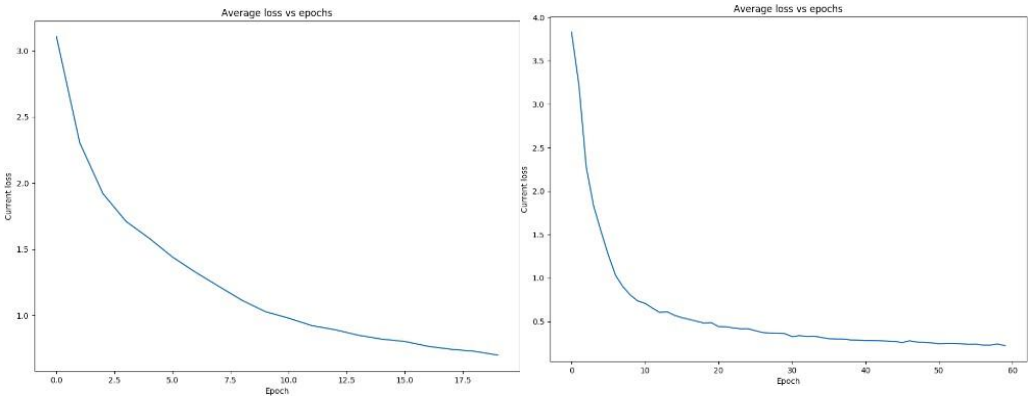


Fig. 7. Loss vs Epoch (FC DenseNet)

E. Global Convolutional Network(GCN)

In this, an encoder - decoder architecture with very large kernels convolutions is proposed. Kernel size is increased to spatial size of feature map. Such convolutions are used because fully connected layers are not able to perform semantic segmentation well. Also, large kernels have very large receptive field and model gather information from much smaller area. Large kernels have a lot of parameters and are computationally expensive. So, in order to avoid that convo- lutions are approximated. This approximated convolution is called global convolution. Encoder used is ResNet(without any dilated convolutions). Decoder consist of GCNs and deconvolutions.[11]

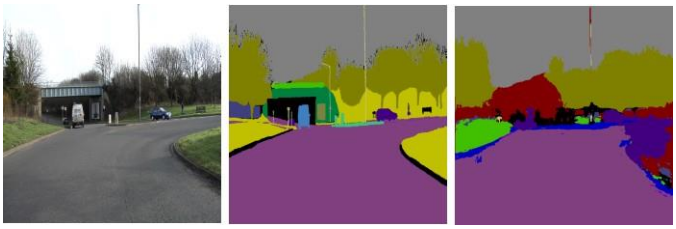


Fig. 8. Input Image - Ground Truth - Predicted Image (GCN)

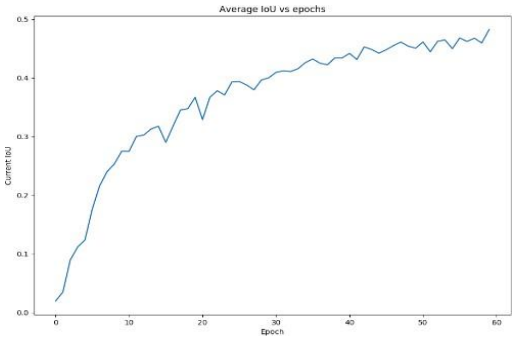


Fig. 9. IoU vs Epoch (GCN)

F. DeeplabV3

DeepLab is Google,s open sourced model of semantic seg- mentation where concept of atrous convolution is introduced which is a generalized form of the convolution operation. It uses atrous convolution with rates 6, 12 and 18. Here, rate is a parameter that controls the effective field of view of the convolution. With inspiration from success of Spatial pyramid pooling, Atrous spatial pyramid pooling was made where four parallel atrous convolutions with different atrous rates, i.e. 1 x 1 convolution and 3 x 3 atrous convolution with rates [6, 12, 18], are applied on top of the feature map, as it is effective to resample features at different scales for accurately and efficiently classifying regions of an arbitrary scale. Bilinear upsampling is used to scale the features to the correct dimensions.

In the later version i.e, DeeplabV3+ a Decoder module on top of the regular DeepLabV3 model is added. Here, instead of using bilinear upsampling,encoded features are upsampled and then concatenated with corresponding low level features from the encoder module having same spatial dimension. [12]

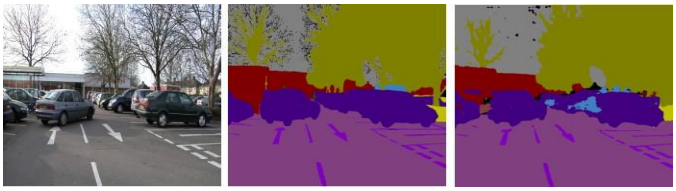


Fig. 11. Input Image - Ground Truth - Predicted Image (DeepLabV3)

G. Optimized FRRN

In this FRRN based Model, we have added more FRRU units which try to capture more of local features at pixel level for better classification of accuracies. MultiScale extracted features are passed in to each of the units that extract the features then they are passed at each of the streams.

Our design is motivated by the need to have networks that can jointly compute good high-level features for recognition

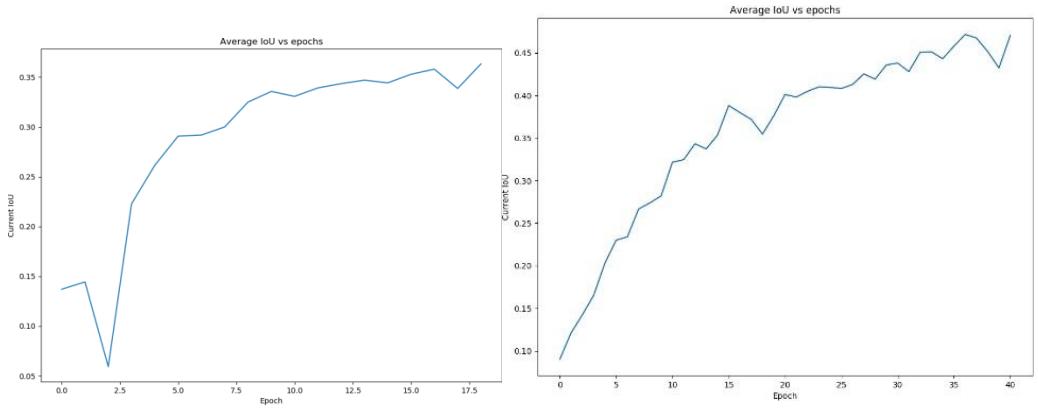


Fig. 12. IoU vs Epoch (DeepLab v3)

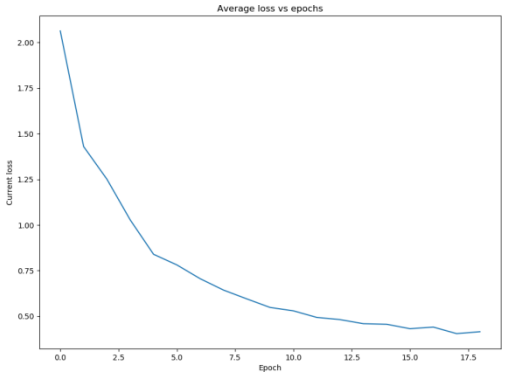


Fig. 13. Loss vs Epoch (DeepLab v3)

and good low-level features for localization. Regardless of the specific network design, obtaining good high level features requires a sequence of pooling operations.

The pooling operations reduce the size of the feature maps and increase the networks receptive field, as well as its robustness against small translations in the image.

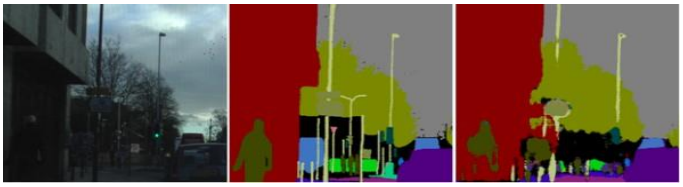


Fig. 14. Input Image - Ground Truth - Predicted Image (Optimized FRRN)

V. EXPERIMENTAL SETUP AND RESULTS

A. Evaluation Criteria And Procedure

Intersection over Union(IoU) is used as a evaluation cri- teria. It is similar to Jaccard index. It is used to compare the diversity and similarity of two images or sets. It is defined

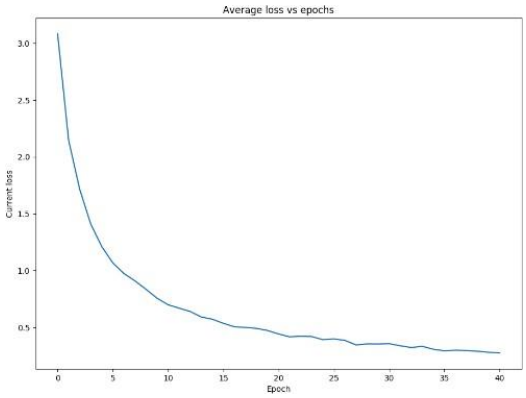


Fig. 16. Loss vs Epoch (Optimized FRRN)

as the size of intersection of two images divided by the size of union of two images.

$$IoU(X,Y) = \frac{X \cap Y}{X \cup Y}$$

where X and Y are two sets or images.

B. Experimental Results

Approach	Evaluation Criteria			
	Precision	Recall	F1 Score	IoU
PSPNet	0.74	0.74	0.74	0.81
FC-DenseNet	0.74	0.77	0.79	0.79
GCN	0.80	0.84	0.86	0.57
DeepLabV3	0.72	0.63	0.64	0.81
Our approach	0.84	0.82	0.82	0.87

VI. CONCLUSIONS

This Paper try to show the results of Various Deep learning Models on the Camvid analyzing various parameters with respect to the increasing the epochs including intersection over union score, validation Score and much more.

REFERENCES

[1] Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N. (2018). Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 325-341).
[2] Badrinarayanan, V., Kendall, A., Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmenta- tion. IEEE transactions on pattern analysis and machine intelligence, 39(12), 2481-2495.
[3] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... Adam, H. (2017). Mobilenets: Efficient convolu- tional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.
[4] Lin, G., Milan, A., Shen, C., Reid, I. (2017). Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1925-1934).
[5] Zhao, H., Qi, X., Shen, X., Shi, J., Jia, J. (2018). Icnet for real-time semantic segmentation on high-resolution images. In

Proceedings of the European Conference on Computer Vision (ECCV) (pp. 405-420).

[6] Brostow, G. J., Fauqueur, J., Cipolla, R. (2009). Semantic object classes in video: A high-definition ground truth database. Pattern Recognition Letters, 30(2), 88-97.

[7] Ronneberger, O., Fischer, P., Brox, T. (2015, October). U-net: Convo- lutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted inter- vention (pp. 234-241). Springer, Cham.

[8] Yu, F., Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122.

[9] Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J. (2017). Pyramid scene parsing network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2881-2890).

[10] Jgou, S., Drozdzal, M., Vazquez, D., Romero, A., —& Bengio, Y. (2017). The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 11-19).

[11] Peng, C., Zhang, X., Yu, G., Luo, G., —& Sun, J. (2017). Large Kernel Matters–Improve Semantic Segmentation by Global Convolutional Network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4353-4361).

[12] Chen, L. C., Papandreou, G., Schroff, F., —& Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation.arXiv preprint arXiv:1706.05587.

[13] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In CVPR, 2016.

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In European Conference on Computer Vision.

[15] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, "Indoor segmentation and support inference from RGBD images", Proc. 12th Eur. Conf. Comput. Vis. , pp. 746-760, 2012.

[16] G. Neuhold, T. Ollmann, S. Rota Bul, and P. Kontschieder. The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes. In International Conference on Computer Vision (ICCV), 2017.

[17] Brostow GJ, Fauqueur J, Cipolla R, (2009) Semantic object classes in video: A high-definition ground truth database. Pattern Recognit Lett 30:8897.

[18] J. Fu, J. Liu, Y. Wang, and H. Lu. Stacked deconvolutional network for semantic segmentation. arXiv preprint arXiv:1708.04943, 2017

[19] G. Lin, A. Milan, C. Shen, and I. Reid, Refinenet: Multi-path refine- ment networks for high-resolution semantic segmentation, in CVPR, 2017.

[20] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. 2018.

[21] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. TPAMI, 2017.

[22] Romera, E., Alvarez, J. M., Bergasa, L. M., Arroyo, R. (2018). Erfnet: Efficient residual factorized convnet for real-time semantic segmenta- tion. IEEE Transactions on Intelligent Transportation Systems, 19(1), 263-272.

[23] Lin, G., Shen, C., Van Den Hengel, A., —& Reid, I. (2018). Exploring context with deep structured models for semantic segmentation. IEEE transactions on pattern analysis and machine intelligence, 40(6), 1352- 1366.

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

[25] R. P. K. Poudel, U. Bonde, S. Liwicki, and C. Zach. ContextNet: Exploring context and detail for semantic segmentation in real-time.
In BMVC, 2018

[26] K. Rakelly, E. Shelhamer, T. Darrell, A. A. Efros, and S. Levine, Few- shot segmentation propagation with guided networks, arXiv preprint arXiv:1806.07373, 2018.

[27] S. Jain and J. Gonzalez. Fast semantic segmentation on video us- ing block motion-based feature interpolation. In ECCV International Workshop on Video Segmentation, 2018

[28] Tao Yang, Yan Wu, Junqiao Zhao, and Linting Guan. Semantic segmentation via highly fused convolutional network with multiple soft cost functions. arXiv preprint arXiv:1801.01317, 2018.

[29] Gharghabi, S., Yeh, C. C. M., Ding, Y., Ding, W., Hibbing, P., LaMunion, S., ... —& Keogh, E. (2019). Domain agnostic online semantic segmentation for multi-dimensional time series. Data mining and knowledge discovery, 33(1), 96-130.

[30] Chiu, H. P., Samarasekera, S., Kumar, R., Villamil, R., Murali, V., —& Kessler, G. D. (2019). U.S. Patent Application No. 16/101,201.

[31] Gharghabi, S., Yeh, C. C. M., Ding, Y., Ding, W., Hibbing, P., LaMunion, S., ... —& Keogh, E. (2019). Correction to: Domain agnostic online semantic segmentation for multi-dimensional time series. Data Mining and Knowledge Discovery, 1-2.

[32] Desai, A. D., Gold, G. E., Hargreaves, B. A., —& Chaudhari, A. S. (2019). Technical Considerations for Semantic Segmenta- tion in MRI using Convolutional Neural Networks. arXiv preprint arXiv:1902.01977.

[33] Rakelly, K., Shelhamer, E., Darrell, T., Efros, A., —& Levine, S. (2018). Conditional networks for few-shot semantic segmentation.

[34] Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X., —& Cottrell, G. (2018, March). Understanding convolution for semantic segmentation. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 1451-1460). IEEE.

[35] Shimoda, M., Sada, Y., —& Nakahara, H. (2019, April). Filter-Wise Pruning Approach to FPGA Implementation of Fully Convolutional Network for Semantic Segmentation. In International Symposium on Applied Reconfigurable Computing (pp. 371-386). Springer, Cham.

[36] Li, H., Xiong, P., Fan, H., —& Sun, J. (2019). DFANet: Deep Feature Aggregation for Real-Time Semantic Segmentation. arXiv preprint arXiv:1904.02216.