Multiscale Fully Convolutional DenseNet for Semantic Segmentation

Sourour BRAHIMI
REGIM-Lab: Research Groups in Intelligent Machines, University of Sfax, National School of Engineers of Sfax (ENIS), BP 1173, Sfax, 3038, Tunisia
sourour.brahimi.TN@ieee.org

Najib BEN AOUN
REGIM-Lab: Research Groups in Intelligent Machines, University of Sfax, National School of Engineers of Sfax (ENIS), BP 1173, Sfax, 3038, Tunisia
Department of Computer Science, College of Computer Science and Information Technology, AL-BAHA University, Saudi Arabia
najib.benaoun@ieee.org

Chokri BEN AMAR
REGIM-Lab: Research Groups in Intelligent Machines, University of Sfax, National School of Engineers of Sfax (ENIS), BP 1173, Sfax, 3038, Tunisia
chokri.benamar@ieee.org

Alexandre BENOIT, Patrick LAMBERT
LISTIC-Lab: Univ. Savoie Mont Blanc, LISTIC, Polytech Annecy Chambéry, 5 ch. de Bellevue, Annecy-le-Vieux, 74940, Annecy, France
{alexandre.benoit, patrick.lambert}@univ-smb.fr

ABSTRACT

In the computer vision field, semantic segmentation represents a very interesting task. Convolutional Neural Network methods have shown their great performances in comparison with other semantic segmentation methods. In this paper, we propose a multiscale fully convolutional DenseNet approach for semantic segmentation. Our approach is based on the successful fully convolutional DenseNet method. It is reinforced by integrating a multiscale kernel prediction after the last dense block which performs model averaging over different spatial scales and provides more flexibility of our network to presume more information. Experiments on two semantic segmentation benchmarks: CamVid and Cityscapes have shown the effectiveness of our approach which has outperformed many recent works.

Keywords

Semantic Segmentation, Convolutional Neural Network, Fully Convolutional DenseNet, Dense Block, MultiScale Kernel Prediction.

1 INTRODUCTION

Today, semantic segmentation represents is very active topic in the computer vision field. It aims to group image pixels into semantically meaningful regions. It has been used for many applications such as video action and event recognition [Wal10a, Ben11a, Ben14a, Ben14b, Mej15a], image search engines [Wan14a, Ben10a], augmented reality [Alh17a], image and video coding [Ben11b, Ben12a], facial expression recognition [Bou16a], image retrieval [Sim14a] and autonomous robot navigation [Lin17a]. In recent years, a big gains in semantic segmentation have been obtained through the use of deep learning. In particular, the Convolutional Neural Network (CNN) methods [Lon15a, Bad15a, Jég17a, Wu16a] have given good semantic segmentation results due to their high capacity for data learning. As a result, many CNN variants have been developed such as Fully Convolutional Network (FCN) [Lon15a], deep fully convolutional neural network architecture for semantic pixel-wise segmentation (SegNet) [Bad15a], Wide Residual Network [Wu16a] and Fully convolutional DenseNet (FC-DenseNet) [Jég17a]. Specifically, FC-DenseNet method has substantially outperformed the prior state of the art methods on many datasets of the semantic segmentation task. Today, semantic segmentation

https://doi.org/10.24132/JWSCG.2018.26.2.5

represents is very active topic in the computer vision field. It aims to group image pixels into semantically meaningful regions. It has been used for many applications such as video action and event recognition [Wal10a, Ben11a, Ben14a, Ben14b, Mej15a], image search engines [Wan14a, Ben10a], augmented reality [Alh17a], image and video coding [Ben11b, Ben12a], facial expression recognition [Bou16a], image retrieval [Sim14a] and autonomous robot navigation [Lin17a]. In recent years, a big gains in semantic segmentation have been obtained through the use of deep learning. In particular, the Convolutional Neural Network (CNN) methods [Lon15a, Bad15a, Jég17a, Wu16a] have given good semantic segmentation results due to their high capacity for data learning. As a result, many CNN variants have been developed such as Fully Convolutional Network (FCN) [Lon15a], deep fully convolutional neural network architecture for semantic pixel-wise segmentation (SegNet) [Bad15a], Wide Residual Network [Wu16a] and Fully convolutional DenseNet (FC-DenseNet) [Jég17a]. Specifically, FC-DenseNet method has substantially outperformed the prior state of the art methods on many datasets of the semantic segmentation task.

In this paper, we propose a Multiscale FC-DenseNet (MS-DenseNet) which exploits the success of FC-DenseNet [Jég17a] for the semantic segmentation. Our method is built upon the FC-DenseNet and it is reinforced by integrating a MultiScale Kernel Convolutional (MSConv) layer after the last Dense Block (DB). The idea behind the use of multiscale kernel is inspired from [Aud16a]. Indeed, this layer aggregates information from 3 parallel convolutions with different kernel sizes in order to collect different spatial contexts. Moreover, it ensures more flexibility of our network to presume more information. Our MS-DenseNet was tested on two challenging benchmarks for semantic segmentation: CamVid [Bro09a] and Cityscapes [Cor15a] datasets. It has significantly improved the segmentation accuracy compared to all reported methods for both datasets.

The rest of our paper is organized as follows. The related works are reviewed in section 2. Then, in section 3, our proposed MS-DenseNet for semantic segmentation will be described. In section 4, the experimental results are presented for the two semantic segmentation benchmarks. Finally, in section 5, conclusions and some future directions are given.

2 RELATED WORKS

Due to the importance of the semantic segmentation, different methods have been developed such as: Graph based methods [Pou15a, Zha14a], Sparse Coding based methods [Zou12a] and CNN based methods [Lon15a, Bad15a, Jég17a, Wu16a]. In this section, we will focus our study on the CNN based methods [Lon15a, Bad15a, Jég17a, Wu16a, Bra16a] since they have shown their good performance and given the best segmentation and recognition results in recent works. The powerful of each CNN variant depends on the network architecture which makes two categories. The first category concerns the CNN methods that have been developed for classification task and extended to the semantic segmentation. The second category groups the encoder-decoder based CNN methods. These CNN methods are composed of two main parts. The first encoder part is similar to the architecture of the conventional CNN methods without neither the fully connected layers nor the classification layer. While the second decoder part is added in order to map the low resolution feature maps of the encoder to complete the input resolution feature maps. This is conducted for pixel-wise classification.

For the first category, image segmentation is conducted using adapted version of the classification oriented CNN methods. Long et al. [Lon15a] have proposed a Fully Convolutional Networks (FCN) method. This method consists of replacing the fully connected layers by convolutional layers with very large receptive fields. This will allow to detect and extract the global context of the scene and output spatial heat maps. It has been built upon AlexNet [Kri12a], VGG-16 [Sim14a] and GoogLeNet [Sze15a]. Figure. 1 presents the FCN-AlexNet architecture. In addition, a ReSeg [Vis16a] method was proposed. This method has extended the ReNet [Vis15a] classification method to the semantic segmentation. It is composed of four Recurrent Neural Networks (RNNs) which retrieve the contextual information by scanning the image in both horizontal and vertical directions. Then, the last feature map is re-sized by one or more max-pooling layers. Finally, to presume the probability distribution over the classes for each pixel, a soft-max layer is used.

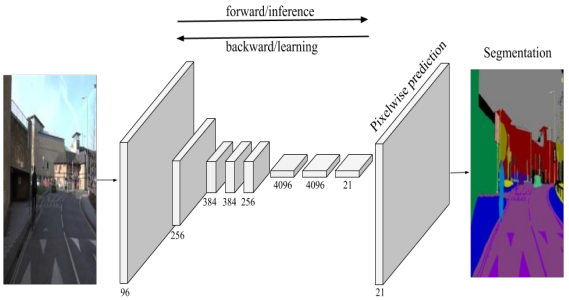

Figure 1: Fully Convolutional Networks (FCN) architecture

Despite their success, the extensions of the conventional CNN methods did not succeed to overcome the problem of learning to decode low-resolution images to pixel-wise predictions for segmentation. That is why, an encoder-decoder architecture was proposed. SegNet [Bad15a] is an example of encoder-decoder methods

(see Figure. 2). It is composed of two symmetric parts where the decoder is an exact mirror of the encoder. The encoder part is composed of 13 convolutional layers inspired from VGG-16 [Sim14a] method. Then, the encoder has a corresponding decoder part with 13 layers which maps the low resolution feature maps of the encoder. Finally, a soft-max classifier is used in order to produce class probabilities for each pixel independently.

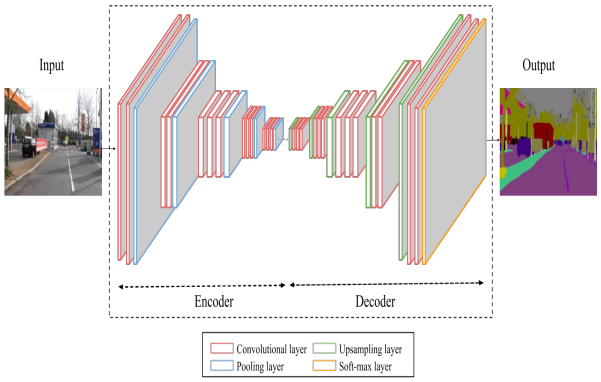

Figure 2: Example of SegNet architecture

Besides, the Efficient Neural Network (ENet) [Pas16a] has been introduced as an encoder-decoder CNN method which has a large encoder and small decoder parts. Each block in ENet architecture is composed of three convolutional layers. Batch Normalization (BN) as well as the Parametric Rectified Linear Unit (PReLU) has been placed between all convolutions. In addition, DeepLab [Che14a] applied an atrous convolutional with up-sampled filters for dense feature extraction. This method exploits deep CNN and fully connected conditional random fields in order to improve the localization performance. Their main idea is to incorporate larger context by enlarging the view field. Moreover, a Dilated convolution method is proposed in [Yu15a] with a dilated filter. This dilated filter is adapted to dense prediction without losing the resolution. It is composed of dilated convolutional layers which aggregate a multiscale contextual information.

Recently, Simon J. et al. have proposed an FC-DenseNet [Jég17a] method which transformed the existing classification model DenseNet [Gao16a] into fully convolutional one. FC-DenseNet is composed of 11 dense blocks (DBs) with five DBs in the encoder part, one DB in the BottleNeck (between the encoder and the decoder) and 5 DBs in the decoder part. In fact, each DB is composed of BN, Rectified Linear Unit (ReLU) layer and a 3×3 convolutional layer. Besides, the DB integrates direct connections from any layer to all subsequent layers. In the encoder part, each DB is followed by a Transition Down (TD) transformation which is composed of BN, ReLU, a 1×1 convolutional layer and a 2×2 max pooling operation. The layer between the encoder and the decoder is referred to as bottleneck. However, in the decoder part each DB is followed by a Transition Up (TU) transformation which is composed of a 3×3 transposed convolution and a stride equal to 2. The transposed convolution consists on upsampling the previous feature maps. Then, the feature maps outputted from the TU layer are concatenated together with the feature maps received from the skip connection. The result of this concatenation will form the input for a new dense block. Finally, a 1×1 convolutional layer followed by Softmax classification method are used to give the per class distribution at each pixel. Figure. 3 visualizes the architecture of FC-DenseNet with only 5 DB, 2 TD and 2 TU.

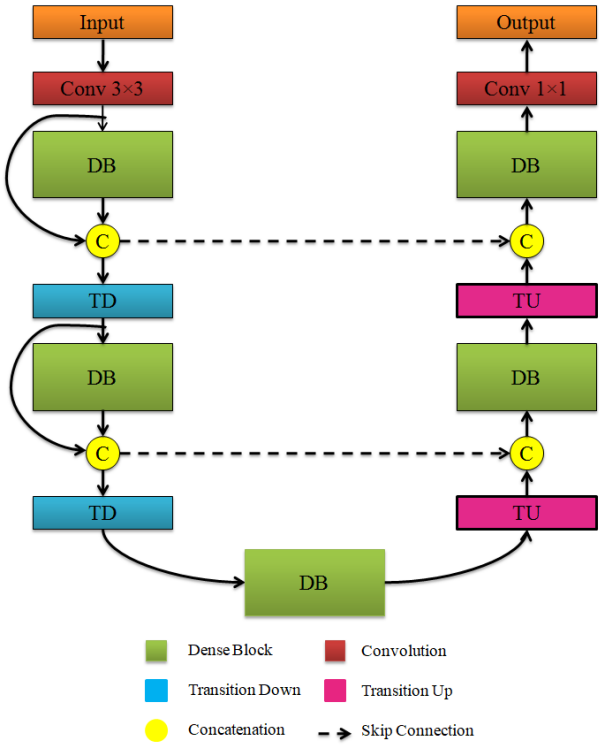

Figure 3: Fully convolutional DenseNet architecture with only five dense blocks. "c" stands for concatenation and interrupted lines are skip connections.

Among the reported methods, FC-DenseNet [Jég17a] has experimentally proven its power for many image segmentation benchmarks. That is what encourages us to build our proposed method on the FC-DenseNet.

3 PROPOSED APPROACH

The central idea of our MultiScale fully convolutional DenseNet (MS-DenseNet) is to take advantage of the FC-DenseNet [Jég17a] method while using multiscale Kernel prediction for the semantic segmentation task. Indeed, a MSConv layer is added to ensure more flex-

ibility of our network and to presume more information. It is conducted to boost the performance of our network. Table 1 details the architecture of our MS-DenseNet method.

3.1 MultiScale Fully Convolutional DenseNet Architecture

As it can be seen in Table 1, our MS-DenseNet is build from 96 convolutional layers: one convolutional layer in the input, 38 layers in the encoder part, 15 layers in the bottleneck, 38 layers in the decoder part with one MSConv layer and one convolutional layer at the end (See Table 1). First, the input image is passed through a standard convolutional layer with 3 ×3 receptive field. Then, 5 DBs are conducted in the encoder part, one DB in the BottleNeck and 5 DBs in the decoder part. As shown in (see Figure. 4), each DB is composed of BN, Rectified Linear Unit (ReLU) layer and a 3×3 convolutional layer. The DB integrates direct connections from any layer to all succeding layers. In the encoder part each DB is followed by a Transition Down (TD) transformation (see Table 1). Each TD is composed of BN, ReLU, a 1×1 convolutional layer, dropout (with p = 0.2) and a 2×2 max pooling operation (see Figure. 4). In the decoder part each DB is followed by a Transition Up (TU) transformation. Each TU is composed of a 3×3 transposed convolution (stride=2) in order to compensate the pooling operation (see Figure. 4). In order to perform model averaging over several scales, MSConv layer is conducted after the last DB. Finally, a convolutional layer with 1 ×1 receptive field and a Soft-max layer are used to determine the inclusion of each pixel to each class.

3.2 MultiScale Kernel Convolutional Layer

Following the multiscale convolutional architecture used in [Aud16a], we have applied a MSConv layer (see Figure. 5) in our method. This layer performs 3 parallel convolutions using different kernels with 1×1 , 3×3 and 5×5 receptive fields contrarily to FC-DenseNet [Jég17a] method that uses only one kernel with 1×1 size. As a result, three different feature maps will be obtained. They will be concatenated together into one feature map. By a conducting these three parallel convolutional layers, our model will aggregate the predictions at different scales while giving only one prediction output. Using MSConv layer, our network becomes more flexibile to presume more information and it will improve the segmentation accuracy.

4 EXPERIMENTAL RESULTS

In this section, we will provide the experimental details. Our proposed method was initialized using HeUniform [He15a] and trained with RMSprop [Tie12a], with an initial learning rate of 0.001. Our approach is evaluated on two datasets used as benchmarks for semantic segmentation: CamVid [Bro09a] and Cityscapes [Cor15a]. For these two datasets, the Mean Intersection over Union (mIoU) is used as a metric to measure the segmentation performance. The IoU determines the similarity between the ground-truth region and the predicted region for an object present in the image. The mean IoU (mIoU) is simply the average over all classes. The IoU is defined to a given class c, predictions (p_i) and targets (t_i), by:

$$IoU(c) = \frac{\sum_i (p_i = c \wedge t_i == c)}{\sum_i (p_i == c \vee t_i == c)} \quad (1)$$

Layer
Batch Normalization
ReLU
3×3 Convolution
Dropout p = 0.2

Transition Down (TD)
Batch Normalization
ReLU
1×1 Convolution
Dropout p = 0.2
2×2 Max Pooling

Transition Up (TU)
3×3 Transposed Convolution stride = 2

Figure 4: Different blocks of MS-DenseNet: the layer used in the model, the Transition Down (TD) and the Transition Up (TU).

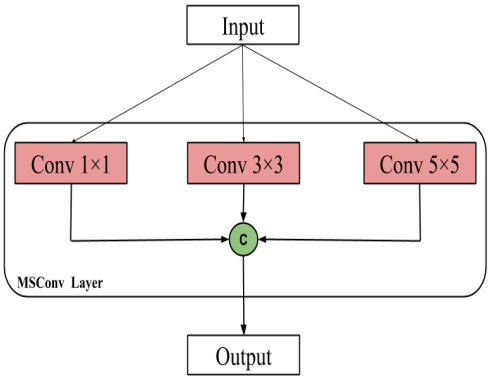

Figure 5: MultiScale Kernel Convolutional Layer

	MS-DenseNet	
	Layers	Configuration
Encoder	Convolution	3×3 Conv
	DB	4 layers
	Transition Down	
	DB	5 layers
	Transition Down	
	DB	7 layers
	Transition Down	
	DB	10 layers
	Transition Down	
	DB	12 layers
	Transition Down	
Bottleneck	DB	15 layers
Decoder	Transition Up	
	DB	12 layers
	Transition Up	
	DB	10 layers
	Transition Up	
	DB	7 layers
	Transition Up	
	DB	5 layers
	Transition Up	
	DB	4 layers
	MSConv	MultiScale Convolution
	Convolution	1×1 Conv
	Segmentation layer	Softmax

Table 1: MS-DenseNet Architecture

where $\wedge$ represents the logical "and" operation, and $\vee$ represents the logical "or" operation. The IoU is computed by summing over all the pixels i of the dataset. Besides, our MS-DenseNet method was implemented using the publicly available TensorFlow Python API [Aba16a].

4.1 CamVid dataset

Cambridge-driving Labeled Video Database (CamVid) [Bro09a] is one of the most commonly used semantic segmentation dataset with 32 semantic classes. In fact, only 11 classes have been used for our experiments: sky, building, pole, road, sidewalk, vegetation, sign, fence, car, pedestrian, cyclist and void, in order to compare our system to recent methods [Lon15a, Pas16a, Bad15a, Vis16a, Ken15a, Jég17a, Yu15a]. This dataset contains 701 semantic segmentation frames: 367 frames used to train the network, 233 for testing and 101 for validation. The size of each frame is 360×480. Figure. 6 visualizes samples from CamVid dataset. Our MS-DenseNet method was trained with image crops of 224×224. The maximum mIoU score has been reached with 225985 steps with 256 batch size.

Table 2 presents the mIoU scores of our method in comparison with the recent semantic segmentation methods in the literature. ENet [Pas16a] has given a lower result than other methods. In addition, FCN-8 [Lon15a] has also failed to give acceptable segmentation results. This can be explained by the fact that the spatial invariance does not take into account useful context execution information. Moreover, Reseg [Vis16a] which takes the advantages of RNN, gives low results less than 59%. Similarly, SegNet [Bad15a] which is an encoder-decoder based model, has given weak results because of the inefficient CNN configuration used. However, despite the improvement done by using Bayesian filters within the Bayesian SegNet [Ken15a] method, the result is still limited. This

Figure 6: Samples from CamVid dataset

network suffers from the speed degradation problem. Besides, Dilation [Yu15a], which has incorporated long spatio-temporal regularization to the output of FCN-8 to boost their performance, has given promising result with 65.30% mIoU scores. Among the state of the art methods, FC-DenseNet [Jég17a] has given the highest mIoU score (66.90%). It is based essentially on DenseNet [Gao16a] classification method. That is why our MS-DenseNet method followed the same architecture while integrating MSConv layer. Adding MSConv layer has given a very promising result. It gives an mIoU score gain of 1.21% compared to the FC-DenseNet method and reaches 68.11%. It proves more that our MS-Densenet architecture is very promising. Examples of images segmented using our MS-DenseNet method are shown in Figure. 7

Model	mIoU (%)
ENet [Pas16a]	55.60
FCN-8 [Lon15a]	57.00
ReSeg[Vis16a]	58.80
SegNet [Bad15a]	60.10
Bayesian SegNet [Ken15a]	63.10
Dilation [Yu15a]	65.30
FC-DenseNet [Jég17a]	66.90
MS-denseNet	**68.11**

Table 2: Results on CamVid evaluation set

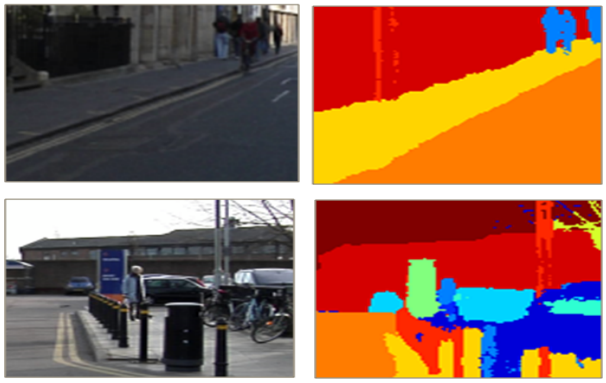

Figure 7: Qualitative results on the CamVid dataset

4.2 Cityscapes dataset

Cityscapes dataset [Cor15a] consists of 5000 images split into three sets: 2975 images for trainings, 500 for validation and 1525 for testing. It has a high image resolution 2048×1024 with 19 classes. Figure. 8 visualizes samples from Cityscapes dataset. The optimal result has been reached when the number of steps was 431425 with 256 batch size.

Table 3 presents a comparison between our method and the other reported methods performances on

Figure 8: Samples from Cityscapes dataset

CityScapes. Similarly to the CamVid dataset, ENet [Pas16a] and FCN-8 [Lon15a] have given weak results. Moreover, Dilation [Yu15a] method has given a 67.10 % mIoU score. Furthermore, different ResNet [He16a] based models such as DeepLab [Che14a], wide-ResNet [Wu16a] have given 70.40% and 78.40% respectively. Indeed, our MS-Densenet method has overcome all state of the art methods by a margin of 0.8% compared to the best reported one and gives 79.20%. This result confirms one more time the strength of our method.

Model	mIoU (%)
ENet [Pas16a]	58.30
FCN-8 [Gar17a]	65.30
Dilation [Yu15a]	67.10
DeepLab [Che14a]	70.40
Wide-ResNet [Wu16a]	78.40
MS-denseNet	**79.20**

Table 3: Results on CityScapes dataset

5 CONCLUSION AND FUTURE WORK

In this paper, a MultiScale FC-DenseNet method is proposed. It is built upon the FC-DenseNet while adding MultiScale kernel Convolutional layer. In fact, a MultiScale Kernel Convolutional layer is integrated after the last dense block in order to give a rich contextual prediction as well as to improve the results. Our method has been experimentally validated on two semantic segmentation benchmarks and has shown very promising results. Our plan for the future work is to improve our MS-DenseNet by optimizing its architecture.

6 ACKNOWLEDGEMENTS

The research leading to these results has received funding from the Ministry of Higher Education and Scientific Research of Tunisia under the grant agreement number LR11ES48. LISTIC experiments have been

made possible thanks to the MUST computing center of the University of Savoie Mont Blanc.

7 REFERENCES

[Wal10a] Wali, A., Ben Aoun, N., Karray, H., Ben Amar, C., and Alimi, A. M., A New System for Event Detection from Video Surveillance Sequences. In ACIVS, pp. 110-120, 2010.

[Wan14a] Wan, J, Wang, D, Hoi, S.C.H., Wu, P, Zhu, J, Zhang, Y and Li, J, Deep learning for content-based image retrieval: A comprehensive study. In ACM international conference on Multimedia, pp.157-166, 2014.

[Ben12a] Ben Aoun, N., Elarbi, M., and Ben Amar, C., Wavelet Transform Based Motion Estimation and Compensation for Video Coding. Advances in Wavelet Theory and Their Applications in Engineering, Physics and Technology, Dr. Dumitru Baleanu (Ed.), pp. 23-40, 2012.

[Ben14a] Ben Aoun, N., Mejdoub, M., Ben Amar, C., Graph-based approach for human action recognition using spatio-temporal features. Journal of Visual Communication and Image Representation, 25 (2): 329-338, 2014.

[Alh17a] Alhaija, H.A., Mustikovela, S.K., Mescheder, L., Geiger, A., and Rother, C., Augmented Reality Meets Deep Learning for Car Instance Segmentation in Urban Scenes. In British Machine Vision Conference, Vol. 3,2017.

[Ben11a] Ben Aoun, N., Elghazel, H., and Ben Amar, C., Graph modeling based video event detection. In IIT, pp. 114-117, 2011.

[Lin17a] Lin, J., Wang, W.J., Huang, S.K., and Chen, H.C., Learning based semantic segmentation for robot navigation in outdoor environment. In IFSA-SCIS, pp. 1-5, 2017.

[Pou15a] Pourian, N., Karthikeyan, S., and Manjunath, B.S., Weakly supervised graph based semantic segmentation by learning communities of image-parts. In Proceedings of the ICCV, pp. 1359-1367, 2015.

[Mej15a] Mejdoub, M., Ben Aoun, N. and Ben Amar, C., Bag of frequent subgraphs approach for image classification. Intelligent Data Analysis 19 (1): 75-88, 2015.

[Zha14a] Zhang, K., Zhang, W., Zeng, S., and Xue, X., Semantic Segmentation Using Multiple Graphs with Block-Diagonal Constraints. In AAAI, pp. 2867-2873, 2014.

[Ben14b] Ben Aoun, N., Mejdoub, M., and Ben Amar, C., graph-based video event recognition. In ICASSP, pp. 1566-1570, 2014.

[Zou12a] Zou, W., Kpalma, K., and Ronsin, J. Semantic segmentation via sparse coding over hierarchical regions. In ICIP, pp. 2577-2580, 2012.

[Lon15a] Long, J., Shelhamer, E., and Darrell, T., Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE CVPR, pp. 3431-3440, 2015.

[Bra16a] Brahimi, S., Ben Aoun, N., Ben Amar, C., Very deep recurrent convolutional neural network for object recognition. In ICMV, 2016.

[Ben10a] Ben Aoun, N., Elarbi, M., and Ben Amar C., Multiresolution motion estimation and compensation for video coding. In ICSP, pp. 1121-1124, 2010.

[Kri12a] Krizhevsky, A., Sutskever, I., and Hinton, G., Imagenet classification with deep convolutional neural networks. In NIPS, pp. 1097-1105, 2012.

[Sze15a] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A., Going deeper with convolutions. In CVPR, pp.1-9, 2015.

[Bad15a] Badrinarayanan, V., Kendall, A., and Cipolla, R., Segnet: A deep convolutional encoder-decoder architecture for image segmentation, 2015. arXiv preprint arXiv:1511.00561, 2015.

[Ben11b] Ben Aoun, N., Elghazel, H., Hacid, M.S., and Ben Amar, C., Graph aggregation based image modeling and indexing for video annotation. In CAIP, pp. 324-331, 2011.

[Jég17a] Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., and Bengio, Y., The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In CVPRW, pp. 117-1183, 2017, July.

[Aud16a] Audebert, N., Le, Saux, B., and Lefevre, S., Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. In Asian Conference on Computer Vision, pp. 180-196, 2016.

[Wu16a] Wu, Z., Shen, C., and Hengel, A.V.D., Wider or deeper: Revisiting the resnet model for visual recognition, 2016. arXiv preprint arXiv:1611.10080.

[Bro09a] Brostow, G.J., Fauqueur, J., and Cipolla, R., Semantic object classes in video: A high-definition ground truth database. Pattern Recognition Letters, 30(2): 88-97, 2009.

[Cor15a] Cordts, M., Omran, M., Ramos, S., Scharwachter, T., Enzweiler, T., Benenson, R., Franke, U., Roth, S., and Schiele, B., The cityscapes dataset. In CVPR, 2015.

[Sim14a] Guedri, B., Zaied, M., Ben Amar, C., Indexing and images retrieval by content. In : High Performance Computing and Simulation (HPCS), 369-375, 2011.

[Sim14a] Simonyan, K., and Zisserman, A., Very deep convolutional networks for large-scale image recog-

nition, 2014. arXiv preprint arXiv:1409.1556.

[Vis16a] Visin, F., Ciccone, M., Romero, A., Kastner, K., Cho, K., Bengio, Y., Matteucci, M., and Courville, A., Reseg: A recurrent neural network-based model for semantic segmentation. In CVPR, pp. 426-433, 2016.

[Vis15a] Visin, F., Kastner, K., Cho, K., Matteucci, M., Courville, A.C., and Bengio, Y., Renet: A recurrent neural network based alternative to convolutional networks, 2015. arXiv:1505.00393v3.

[Pas16a] Paszke, A., Chaurasia, A., Kim, S., and Culurciello, E., Enet: A deep neural network architecture for real-time semantic segmentation, 2016. arXiv preprint arXiv:1606.02147.

[Che14a] Chen, L.C., Papandreou, G, Kokkinos, I., Murphy, K., and Yuille, A.L., DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs, 2014. arXiv preprint arXiv:1606.00915.

[Yu15a] Yu, F., and Koltun, V., Multi-scale context aggregation by dilated convolutions, 2015. arXiv preprint arXiv:1511.07122.

[Gao16a] Gao, H., Zhuang, L., and Kilian, Q.W., Densely connected convolutional networks, 2016. arXiv:1608.06993v3.

[He16a] He, K., Zhang, X., Ren, S., and Sun, J., Deep residual learning for image recognition. In CVPR, pp. 770-778, 2016.

[He15a] He, K., Zhang, X., Ren, S., and Sun, J., Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision, pp. 1026-1034, 2015.

[Tie12a] Tieleman, T., and Hinton, G., rmsprop adaptive learning. In COURSERA: Neural Networks for Machine Learning, 2012.

[Aba16a] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., and Ghemawat, S., Tensorflow: Large-scale machine learning on heterogeneous distributed systems. Publicly available at: https://tensorflow.org

[Gar17a] Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., and Garcia-Rodriguez, J., A Review on Deep Learning Techniques Applied to Semantic Segmentation, 2017. arXiv preprint arXiv:1704.06857.

[Ken15a] Kendall, A., Badrinarayanan, V., and Cipolla, R., Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding, 2015. arXiv:1511.02680.

[Bou16a] Boughrara, H., Chtourou, M., Ben Amar, C. , Chen, L., Facial expression recognition based on a mlp neural network using constructive training algorithm, Multimedia Tools and Applications, 75(2), 709-731, 2016.