Multimedia Tools and Applications (2019) 78:22077–22098
https://doi.org/10.1007/s11042-019-7430-x

Semantic segmentation using reinforced fully convolutional densenet with multiscale kernel

Sourour Brahimi[1] · Najib Ben Aoun[1,2] · Alexandre Benoit[3] · Patrick Lambert[3] · Chokri Ben Amar[1,4]

Received: 21 February 2018 / Revised: 28 December 2018 / Accepted: 25 February 2019 /
Published online: 9 April 2019

Abstract
In recent years, semantic segmentation has become one of the most active tasks of the computer vision field. Its goal is to group image pixels into semantically meaningful regions. Deep learning methods, in particular those who use convolutional neural network (CNN), have shown a big success for the semantic segmentation task. In this paper, we will introduce a semantic segmentation system using a reinforced fully convolutional densenet with multiscale kernel prediction method. Our main contribution is to build an encoder-decoder based architecture where we increase the width of dense block in the encoder part by conducting recurrent connections inside the dense block. The resulting network structure is called wider dense block where each dense block takes not only the output of the previous layer but also the initial input of the dense block. These recurrent structure emulates the human brain system and helps to strengthen the extraction of the target features. As a result, our network becomes deeper and wider with no additional parameters used because of weights sharing. Moreover, a multiscale convolutional layer has been conducted after the last dense block of the decoder part to perform model averaging over different spatial scales and to provide a more flexible method. This proposed method has been evaluated on two semantic segmentation benchmarks: CamVid and Cityscapes. Our method outperforms many recent works from the state of the art.

Keywords Semantic Segmentation · Fully Convolutional DenseNet · Wider Dense Block · MultiScale kernel prediction

1 Introduction

Semantic segmentation is a fundamental part in the computer vision field. It aims at partitioning the image into semantically meaningful parts and classifying each part into one class. Semantic segmentation has been used in many applications such as video action and

✉ Najib Ben Aoun
najib.benaoun@ieee.org

Extended author information available on the last page of the article.

event recognition [9, 12, 13, 17, 40, 52], image search engines [8, 24, 29, 39, 41, 53], image and video coding [10, 11, 25], medical imaging [7, 33], augmented reality [2], autonomous robot navigation [37] and autonomous driving [19].

Many semantic segmentation systems have been developed by applying different methods such as: Thresholding [5, 6], edge detection [20], region growing [44], graph partitioning [26, 43, 60], sparse coding [62] and Convolutional Neural Network (CNN) [4, 21, 32, 38, 50, 59, 61]. Thresholding-based methods [5, 6] segment the image by thresholding the image pixel values and grouping similar pixels together. But the problem of this method is to fix the appropriate threshold which gives a good segmentation. While, edge detection methods [20] detect region edges inside the image. Then, image regions are recognized based on their detected edges. However, one limitation of edge detection based methods is to detect blurred edges and overlapped objects. Besides, region growing methods [44] are based on the assumption of having similar values for neighboring pixels in one region. The problem about these methods is how to identify the degree of similarity between two adjacent pixels in addition to the fact that an object is generally composed of different connected regions. Furthermore, graph partitioning methods [26, 43, 60] consist on grouping the graph nodes into two or more partitions based on certain criteria. One problem of graph methods is the grouping criteria as well as the choice of a number of segments which influences the quality of segmentation. Moreover, sparse coding [62] is introduced as a high level image region representation. Based on this representation, image regions are identified. Nonetheless, sparse coding methods did not always succeed in extracting discriminative region features.

Over the past few years, CNN [4, 21, 32, 38, 50, 59, 61] has made a great progress in semantic segmentation due to its high capacity for data learning. As a result, many CNN variants have been developed such as Fully Convolutional Network (FCN) [38], deep fully convolutional neural network architecture for semantic pixel-wise segmentation (SegNet) [4], Wide Residual Network (ResNet) [58] and Fully Convolutional DenseNet (FC-DenseNet) [32]. Recently, due to its powerful CNN architecture, FC-DenseNet has given very promising results in comparison with the state of the art methods on many semantic segmentation benchmarks.

In this paper, we propose an encoder-decoder method called Reinforced Multiscale fully convolutional DenseNet (RM-DenseNet) where we increase the width of the network by integrating some Wider Dense Blocks (WDBs). These WDBs consist in making the dense block wider by conducting a set of dense blocks recurrently connected together. In fact, each DB takes not only the output of the previous layer but also the initial input of the WDB. Indeed, these recurrent connections extend the contextual field of view and increase the depth of our network without augmenting the number of parameters. Besides, they allow our network to go back to an earlier time to pick up information that may have been otherwise forgotten. In addition, it emulates the human brain system and helps to strengthen the extraction of the target features. Moreover, inspired by [3], we have conducted a MultiScale Convolutional (MSConv) layer after the last DB of the decoder part. This MSConv conducts three parallel convolutional layers with different kernels (1×1 , 3×3 and 5×5) which aggregates the prediction in different sizes of spatial context. In addition, the MSConv layer has made more flexible and powerful method. Our RM-DenseNet has been evaluated on two semantic segmentation benchmarks: CamVid [18] and Cityscapes [22] datasets and has given better results than the state of the art methods.

The remainder of our paper is organized as follows. We review the recent semantic segmentation works in Section 2. In Section 3, we detail our proposed approach. Then, the experimental results are presented for the two semantic segmentation datasets in Section 4. Finally, we conclude this paper and give some future directions in Section 5.

2 Related works

Due to the importance of the semantic segmentation field, many attempts have been developed. In this section, we have focused our study on the deep learning methods since they have recently demonstrated a substantial success in many applications ranging from image processing to semantic segmentation. Consequently, we will present different CNN variants [4, 21, 32, 38, 50, 59, 61] for semantic segmentation task since they have shown their good performance and given the best segmentation results in recent works (See Table 1). In order to make this section clear, we have classified the CNN methods into two categories. The first category concerns the CNN methods that have been developed for the classification task and extended to the semantic segmentation. The second category groups the encoder-decoder based CNN methods. These methods are composed of two main parts: encoder and decoder. The encoder part is similar to the architecture of the conventional CNN methods without neither the fully connected layers nor the classification layer. While the decoder part is added in order to map the low resolution feature maps of the encoder to complete input resolution feature maps for pixel-wise classification.

2.1 Classification oriented CNN methods adapted to the semantic segmentation task

As for the first category of CNN methods, image segmentation is conducted using adapted version of the classification oriented CNN methods [14–16]. In [38], Long et al. have introduced a Fully Convolutional Networks (FCN) method. This method removes the fully connected layers which give classification scores and replaces it with convolutional layers with very large receptive fields to capture the global context of the scene and output spatial heat maps. The FCN has been built upon three CNN methods: AlexNet [35] to get the FCN-AlexNet version, VGG-16 [47] to get the FCN-VGG16 version and GoogLeNet [48] to get the FCN-GoogleNet version. The FCN-AlexNet architecture is illustrated in Fig. 1. Another method is ReSeg [50] which transforms the ReNet [51] classification method to use it for the semantic segmentation. The architecture of ReSeg is composed of four Recurrent

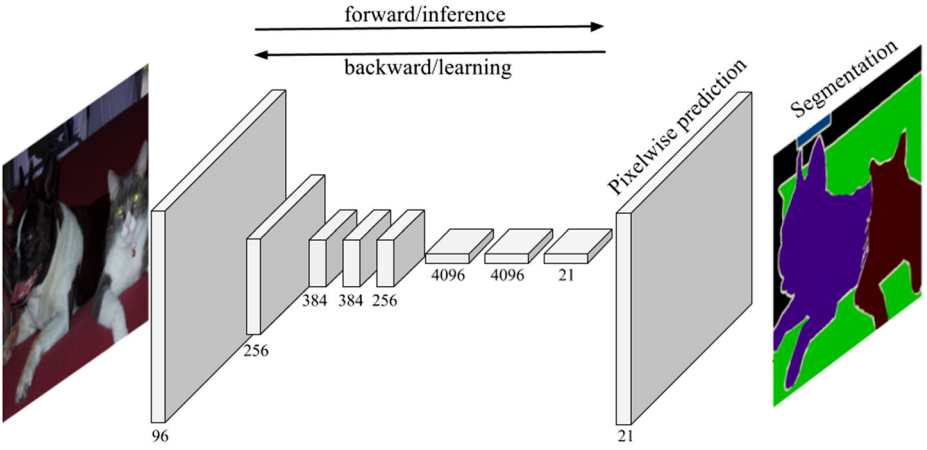


Fig. 1 Example of Fully Convolutional Networks (FCN) architecture

Neural Network (RNN) which retrieve the contextual information by sweeping the image horizontally and vertically in both directions. Then, the last feature map is re-sized by one or more max-pooling layers. Finally, a soft-max layer is used to presume the probability distribution over the classes for each pixel.

Besides, Pyramid Scene Parsing Network (PSPNet) is introduced in [61] as an extension of the classification ResNet [31] method. PSPNet used the pretrained ResNet to get the feature map from the last convolutional layer. Then, to get a different sub-region representation, PSPNet used a pyramid parsing module. Moreover, up-sampling and concatenated layers are coming after the pyramid parsing to form the final representation of feature map. Finally, the final pixel prediction is obtained by a convolutional layer. PSPNet method produces additional contextual information.

In addition, in [54], an end-to-end trainable deep bidirectional LSTM (Long-Short Term Memory) named Bi-LSTM is proposed. This method produces a deep CNN and two separate LSTM networks. It exploits the two separate LSTM in order to learn hierarchical visual language embeddings. It is a deep network which exploits future and history context information. Furthermore, Cheng et al. [57] have improved the Bi-LSTM method by adding multi task learning. It takes the advantages of Bi-LSTM model to learn hierarchical visual language and it employs multi task learning to increase its generality. Then, in [55], a regularized deep neural network (RE-DNN) is proposed. This method studies the highly non-linear semantic correlation between text and image. It includes visual, textual and joint models for visual semantic representation learning, textual semantic representation learning and cross-modal mapping respectively. It is composed of five layers which are divided into three parts: one-third for image modality, one-third for text modality and the last one-third of the network for multimodal joint modeling. Furthermore, Cheng W. et al. [56] proposed a CNN based framework in order to exploit the multimodal video representation in action recognition. This framework contains four modules: spatial CNN, temporal CNN, acoustical CNN and a fusion layer. The fusion layer contains both early fusion and late fusion with Neural Network and Support Vector Machine. It is added on the top of CNNs to learn a joint

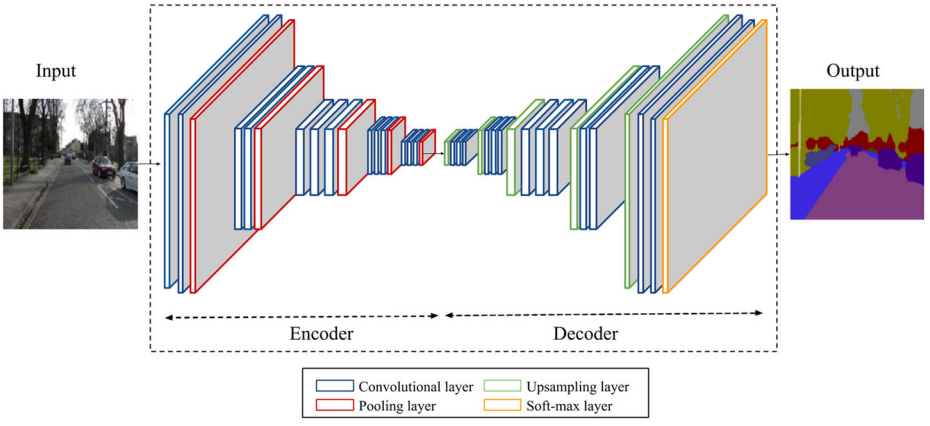


Fig. 2 Example of SegNet architecture

video representation. This method takes into consideration the audio information with the implementation of sophisticated fusion methods.

2.2 Encoder-decoder architecture based methods

Despite their success, the extensions of the conventional CNN methods have problems on learning to decode low-resolution images to pixel-wise predictions for segmentation. That is why encoder-decoder architectures have been proposed in many recent CNN methods. In particular, SegNet [4] as an example of encoder-decoder models (see Fig. 2) which is composed of two symmetric parts where the decoder is an exact mirror of encoder. The encoder part is composed of 13 convolutional layers inspired from VGG-16 [47] method. Besides, a corresponding decoder part with 13 layers also maps the low-resolution feature maps of the encoder. The final decoder output is a soft-max classifier that produces class probabilities for each pixel independently.

Following the same encoder-decoder architecture, Simon J. et al. proposed a FC-DenseNet [32] method which transforms the existing classification model DenseNet [27] into fully convolutional one. FC-DenseNet is composed of 11 dense blocks (DBs) where 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 (see Fig. 3a). 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 (see Fig. 3b). The layer between the encoder and the decoder is referred to as a 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 with stride 2 (see Fig. 3c). The transposed convolution consists on upsampling

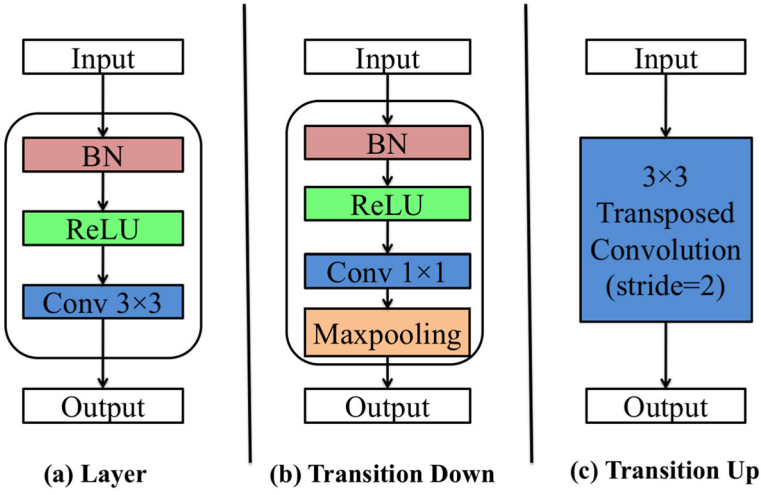


Fig. 3 Building blocks of RM-DenseNet: **a** Layer used in the model, **b** Transition Down (TD) layer and **c** Transition Up layer (TU)

Table 1 Different semantic segmentation methods

Segmentation methods	Architecture	Contributions
FCN [38]	VGG-16 [47]	Fully convolutional
SegNet [4]	VGG-16 [47]	Encoder-Decoder
Dilation [59]	VGG-16 [47]	Dilated convolutions
ReSeg [50]	VGG-16 [47] + ReNet [51]	RNN for semantic segmentation task
PSPNet [61]	ResNet [31]	Pyramid pooling module
DeepLab [21]	VGG-16 [47] / ResNet [31]	Conditional Random Fields
FC-DenseNet [32]	DenseNet [27]	Dense connections for semantic segmentation task

the previous feature maps. These feature maps are then concatenated to the ones coming from the skip connection to form the input of a new DB. Finally, a 1×1 convolutional layer followed by Softmax classification method is used to give the per class distribution at each pixel. This method has achieved good results in semantic segmentation on CamVid [18] and Gatech [45] datasets.

As it can be seen from Table 1, the architecture as well as the main contribution of each CNN based method are presented. The majority of the presented methods are based on the VGG-16 [47] architecture. However, among the reported methods, the FC-DenseNet [32] has given the best results for many image segmentation benchmarks. That is what encourages us to build our proposed CNN method upon the FC-DenseNet.

3 Proposed approach

Our proposed method presents a new CNN architecture built upon the successful FC-DenseNet [32] method while using WDBs and an MSConv layer. The WDB improves the classical dense blocks by increasing the width of DB in the encoder part by building a recurrent structure. In fact, the recurrent connections inside the WDB are added to emulate the human visual system and to integrate the context information with fixed number of parameters. In addition, our method conducts an MSConv layer which is inspired from [3] after the last DB of the decoder part to aggregate the prediction in different sizes of spatial context. This network architecture leads to a new semantic segmentation method called Reinforced MultiScale fully convolutional DenseNet (RM-DenseNet) (see Fig. 4).

3.1 Reinforced multiscale fully convolutional densenet architecture

Our RM-DenseNet follows the FC-DenseNet pipeline with six WDBs. This architecture is built from 269 convolutional layers: one convolutional layer in the input, 162 layers in the encoder part, 61 layers in the bottleneck, 43 layers in the decoder part as well as one MSConv layer and one convolutional layer at the end. First, the input image is passed through a standard convolutional layer with 3×3 receptive field. Then, 5 WBDs have been applied where each one of them contains one convolutional layer required for the summation operation and 4 DBs connected between each other with recurrent connections (See Table 2). In addition, as shown in Fig. 3, each WDB is followed by a TD. Each TD is composed of BN, ReLU, a 1×1 convolutional layer, dropout with $p = 0.2$ and a max pooling

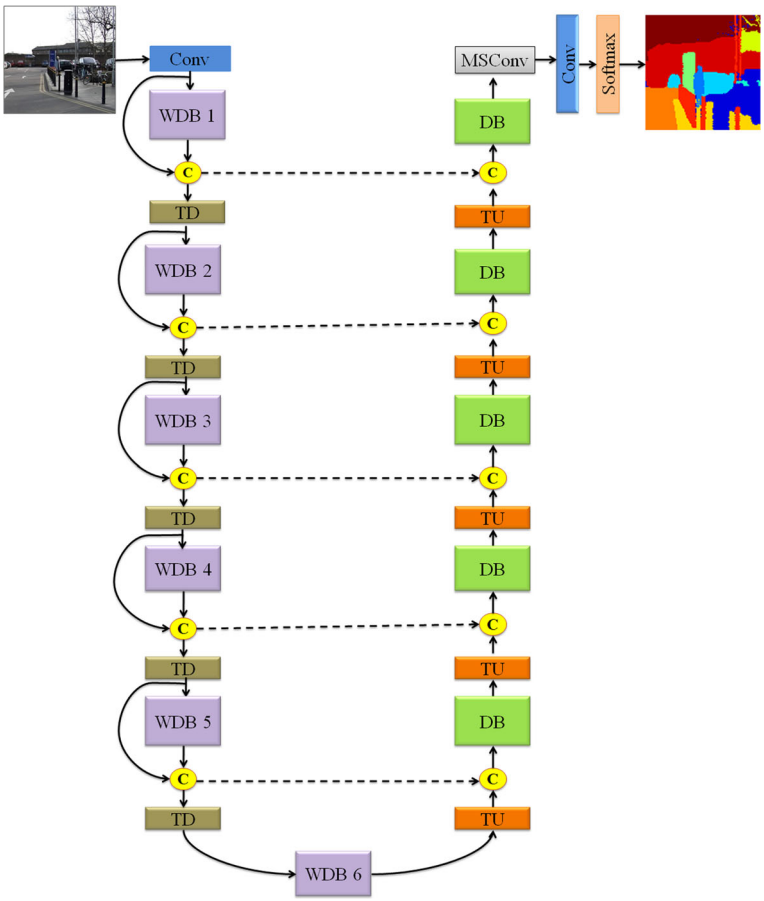


Fig. 4 Diagram of our Reinforced MultiScale fully convolutional DenseNet

of size 2×2. Besides, a bottleneck between the encoder and the decoder with one WDB is conducted. Afterwards, 5 DBs are used in the decoder part which each of them is preceded by a TU. Each TU is composed of a 3×3 transposed convolution with stride 2 to compensate the pooling operation (See Fig. 3). In order to perform model averaging over several scales, MSConv layer is conducted after the last DB. Finally, a convolutional layer with

Table 2 Wider dense block parameters

Block name	Number of DB	Number of layer in each DB
WDB 1	4	4
WDB 2		5
WDB 3		7
WDB 4		10
WDB 5		12
WDB 6		15

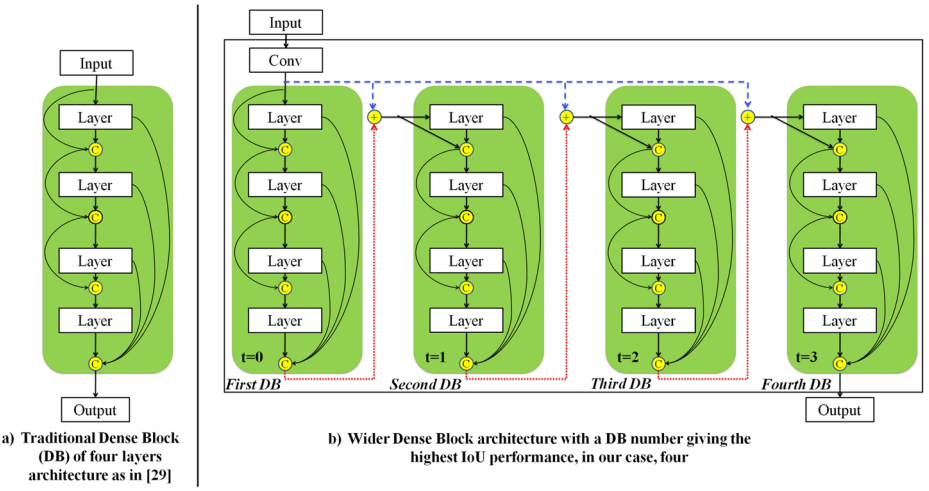


Fig. 5 Wider Dense Block architecture

1×1 receptive field and a Soft-max layer are used to provide the per class distribution at each pixel.

3.2 Wider dense block

As shown in Fig. 5, each WDB is composed of one convolutional layer (needed for the summation operation) and four DBs designed following recurrent structure (see Fig. 5b). It is unfolded for t time steps with t= 0 represents only a standard feed forward connection which is denoted in Fig. 5 by the dashed lines. However, for the three other time steps the input of the DB is the summation of the initial input of the WDB and the output of the previous DB which is represented in Fig. 5 by dotted lines. In order to conduct this summation, the initial WDB input and the output of the previous DB should have the same dimension. Similarly to the traditional DB [32], the output of a DB with n number of layers is $n \times k$ feature maps (where k is the number of applied filters). In order to make the initial WBD input of a same size, a convolutional layer is applied on it which outputs also $n \times k$ feature (see Fig. 5b).

The optimal WDB width has been experimentally determined as four (see Tables 3 and 6). Indeed, increasing the width by using recurrent structure allows our network to gain several advantages. Firstly, the integration of the recurrent connections allows our network to become more suitable to the context information [23, 36] since not only the initial states is stored in the internal memory but also the previous states. Secondly, the network will

Table 3 mIoU of FC-DenseNet with different time steps of WDB on CamVid dataset

Model	Time steps	mIoU (%)
FC-DenseNet [32]	–	66.90
FC-DenseNet+WDB	2	68.93
	4	**69.13**
	6	69.01

The bold numbers indicate the highest result and the best time steps

be able to take into account previous processing that one could interpret as ”previous time step” to pick up some information that may have been otherwise forgotten. Actually, this approach explicitly reuses the earlier processing that processed the data at a lower field of view and is combined with a wider one. This is implicitly a multiscale processing that relies on similar processing structures using the same parameters. In addition, it does not need additional computational parameters thus helping to avoid the difficulty in training such as the over-fitting. Therefore, our network becomes deeper with no additional parameters because of the weights sharing. Finally, the recurrent structure strengthens the extraction of the target features and ameliorates the segmentation performance.

3.3 MultiScale convolutional layer

Our network has been also boosted by using MSConv layer which is inspired from [3] (see Fig. 6). This MSConv Layer performs 3 parallel convolutions using different kernels with 1×1, 3×3 and 5×5 receptive fields contrarily to FC-DenseNet [32] method that uses only one kernel with 1×1 size. This will lead to three different feature maps which will be concatenated together into one feature map. These three parallel convolutional layers allow our model to aggregate the predictions at different scales while giving only one prediction output. This ensures more flexibility of our network to presume more information and to improve the segmentation accuracy.

4 Experimental results

This section provides an experimental study of our RM-DenseNet. The proposed method was initialized using HeUniform [30] and trained with RMSprop [49], with an initial learning rate of 0.001. It was evaluated on two semantic segmentation datasets: CamVid [18] and Cityscapes [22]. To evaluate the methods on these two datasets, the Mean Intersection over Union (mIoU) metric is used. The IoU determines the similarity between the predicted region and the ground-truth 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:

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

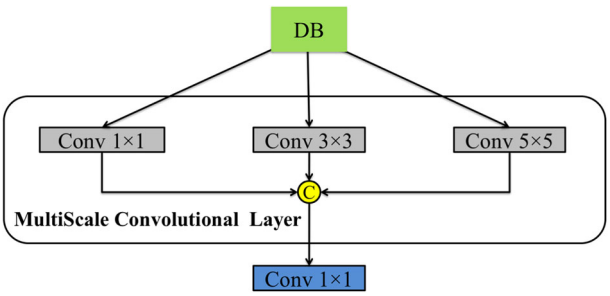


Fig. 6 MultiScale convolutional layer

Fig. 7 Samples from CamVid dataset

where $\wedge$ is a logical *and* operation, while $\vee$ is a logical *or* operation. We compute IoU by summing over all the pixels i of the dataset. Besides, our RM-DenseNet method was implemented using the publicly available TensorFlow Python API [1].

4.1 CamVid dataset

One of the most commonly used semantic segmentation dataset is Cambridge-driving Labeled Video Database (CamVid) [18] which presents 32 semantic classes. However, in order to compare our system to recent methods [4, 32, 34, 38, 42, 50, 59] only 11 classes have been used for our experiments: sky, building, pole, road, sidewalk, tree, sign, fence, car, pedestrian and cyclist. This dataset contains 701 semantic segmentation frames: 367 frames used to train the network, 233 for testing and 101 for validation. The evaluation of the methods on this dataset is done with the validation set. The size of each frame is 360 × 480. Figure 7 visualizes samples from CamVid dataset. Our RM-DenseNet method was trained with image crops of 224 × 224. Our model can run at about 180ms per image on a GPU.

For our RM-DenseNet architecture, the appropriate width is determined experimentally in order to reach the best mIoU. So, we have run our method with three different widths: two, four and six. As it can be seen from Table 3, the mIoU of our method saturates when the WDB reaches 4 time steps with mIoU equal to 69.13%. However, six time steps have given low result than four time steps which is costly in time without any improvement in mIoU. That is why, we did not go further for this dataset. Besides, in comparison with the FC-DenseNet [32] method, making the architecture wider by adding WDB with four time steps has increased the segmentation accuracy by a factor of 2.2%.

Table 4 The contribution of WDB and MSConv layer on CamVid evaluation set

Model	WDB	MSConv	mIoU (%)
FC-DenseNet [32]	×	×	66.90
FC-DenseNet+WDB	√	×	69.13
FC-DenseNet+MSConv	×	√	68.11
RM-DenseNet	√	√	**69.59**

The bold number indicates the highest result

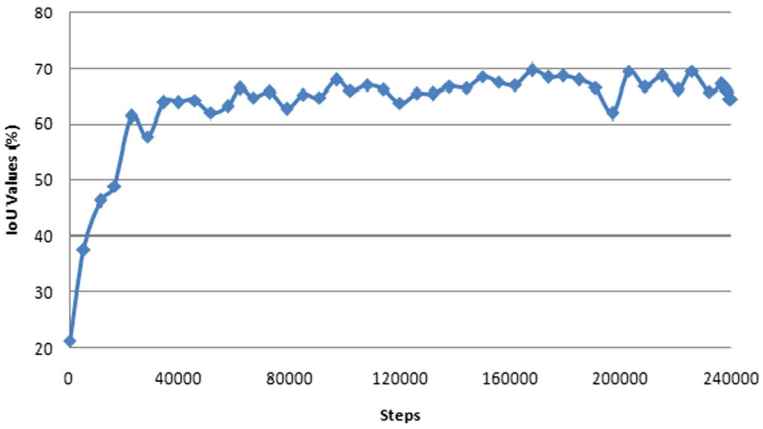


Fig. 8 Testing curve for CamVid evaluation set per step number

To improve the performance of our method, we have added an MSConv layer after the last DB which conducts three parallel convolutions with different kernel sizes. The impact of this layer can be seen in Table 4. In fact, it gives a mIoU score gain of 1.21% in comparison with the FC-DenseNet [32] standard and reaches 68.11%. As a result, the WDB as well as the MSConv layer have contributed significantly to our RM-DenseNet method with outperforms the FC-DenseNet method [32] by a factor of 2.69%.

Figure 8 illustrates the different mIoU of our method per network step number which are varying from 1 to 240000. As a result, the maximum mIoU score has been reached with 225985 steps with 256 batch size.

Table 5 gives the mIoU scores of our method in comparison with other efficient methods in the literature. In terms of mIoU, ENet [42] has given a lower result than other methods. In addition, FCN-8 [38] which is an extension of the classification oriented VGG-16 [47] method 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 [50] which takes the advantages of RNN, gives low results less than 59%. For the SegNet [4] which uses VGG-16 classification based model with encoder-decoder technique and which is substantially deeper than those are mentioned previously, it has provided 60.1% mIoU. Whereas, despite the improvement obtained by using Baysien filters within the Baysien SegNet [34] method, the result is still limited comparable to our RM-DenseNet because of the speed degradation problem. However, Dilation [59], which has incorporated long spatio-temporal regularization to the output of FCN-8 to boost their performance, has given promising result with 65.3% mIoU scores. Among the state of the art methods, FC-DenseNet [32] has given the best mIoU score (66.9%). It is based essentially on DenseNet [27] classification method. That is why our RM-DenseNet method followed the same architecture while using WDB and integrating MSConv layer. Our method improves the mIoU score by a substantial margin of 2.69% compared to the FC-DenseNet method and it gives 69.6%. In terms of per class mIoU, our RM-DenseNet outperforms all the other methods on all the classes except for the class ”sign” where Dilation [59] method performs better due to its dilated convolution operator which expands the receptive field without losing resolution or coverage. This proves more the effectiveness of our RM-DenseNet.

Table 5 Results on CamVid evaluation set

Model	parameters (M)	Building	Tree	Sky	Car	Sign	Road	Pedestrian	Fence	Pole	Sidewalk	Cyclist	mIoU
ENet [42]	–	n/a											55.6
FCN-8 [38]	134.5	77.8	71.0	88.7	76.1	32.7	91.2	41.7	24.4	19.9	72.7	31.0	57.0
ReSeg[50]	–	n/a											58.8
SegNet [4]	29.5	68.7	52.0	87.0	58.5	13.4	86.2	25.3	17.9	16.0	60.5	24.8	60.1
Bayesian SegNet [34]	29.5	n/a											63.1
Dilation [59]	140.8	82.6	76.2	89.0	84.0	**46.9**	92.2	56.3	35.8	23.4	75.3	55.5	65.3
FC-DenseNet [32]	9.4	83.0	77.3	93.0	77.3	43.9	94.5	59.6	37.1	37.8	82.2	50.5	66.9
RM-denseNet	9.5	**84.9**	**79.7**	**94.8**	**84.4**	45.9	**95.3**	**62.2**	**37.7**	**37.8**	**83.7**	**58.4**	**69.6**

The bold numbers indicate the highest results

Fig. 9 Samples from Cityscapes dataset

4.2 Cityscapes dataset

Cityscapes dataset [22] consists of 5000 images split into three sets: 2975 images for training, 500 for validation and 1525 for testing. The evaluation of the methods on this dataset is done with the validation set. It has a high images resolution of 2048 × 1024. The images belong to 19 classes. Figure 9 visualizes samples from Cityscapes dataset. Our model can run at about 3s per image on a GPU due to the big size of the images.

In order to determine the optimal time steps number for the WDB, we have run our method with two, four and six time steps. The best segmentation accuracy of our method is gotten when WDB is conducted with four time steps which gives an mIoU equal to 79.50% (see Table 6). Besides, including the WDBs to the standard FC-DenseNet method has significantly increased the mIoU by 1.18% which the confirms more the contribution of the WDB.

Table 7 illustrates the impacts of both WDB and MSConv layer. Similarly to the integration of the WDB to the FC-DenseNet architecture, adding the MSConv layer has improved the performance by an mIoU equal to 0.28%. These experimental results confirm the importance of using wider network as well as the MSConv layer.

In order to obtain the best network steps number for this dataset, our method has been conducted with different steps number ranging from 1 to 500000 (Fig. 10). The optimal mIoU has been reached when the number of steps was 431425 with 256 batch size.

Table 8 reports a comparative study between our method and the state of the art methods. Similarly, on the CamVid dataset, ENet [42] and FCN-8 [38] have given weak results. Moreover, Dilation [59] method has given a 67.1 % mIoU score. Furthermore, different ResNet [31] based models such as DeepLab [21], wide-ResNet [58] and PSPNet [61] have given 70.4%, 78.4% and 80.2% respectively. Indeed, our RM-DenseNet method outperforms all state of the art methods and gives 80.3% due to the use of WDB and MSConv layer. This result confirms one more time the strengths of our method. In fact, in terms of per class

Table 6 mIoU of FC-DenseNet with different time steps of WDB on CityScapes dataset

Model	Time steps	mIoU (%)
FC-DenseNet [32]	–	78.32
FC-DenseNet+WDB	2	78.97
	4	**79.50**
	6	79.19

The bold numbers indicate the highest result and the best time steps

Table 7 The contribution of WDB and MSConv layer on CityScapes evaluation set

Model	WDB	MSConv	mIoU (%)
FC-DenseNet [32]	×	×	78.92
FC-DenseNet+WDB	√	×	79.50
FC-DenseNet+MSConv	×	√	79.20
RM-DenseNet	√	√	**80.32**

The bold number indicates the highest result

mIoU, our RM-DenseNet outperforms all the other methods on all the classes except for the class "trafficsign" where PSPNet [61] method has given better result due to its capability to embed difficult scenery context features. As a result, our RM-DenseNet has proven its effectiveness and good performances for the CityScapes dataset.

5 Discussion

Our RM-DenseNet method has given very promising results on the two datasets. These results confirm the robustness of our architecture which includes the WDBs and the MSConv layer. In fact, thanks to the recurrent connectivity inside the WDB, our architecture has gained several advantages. Firstly, this recurrent structure allows our network to go back to an earlier time to pick up information that may have been otherwise forgotten. Secondly, our network becomes deeper with no additional parameters because of weights sharing. In addition, it allows our model to propagate in width with the same number of parameters which avoids the problem of system complexity and reduces the time consumption. Moreover, it allows the best accumulation and extraction of information. The results obtained in section 4 prove the success of adding recurrent connections (See Tables 4 and 7).

Moreover, the RM-DenseNet architecture has been enriched with the MSConv layer which has led to a significant improvement in the segmentation accuracy. Actually, the MSConv layer aggregates information from three 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. The WDB and the MSConv layer have provided a big gain for our RM-DenseNet method and helped in surmounting several constraints for other CNN methods (See Tables 4 and 7).

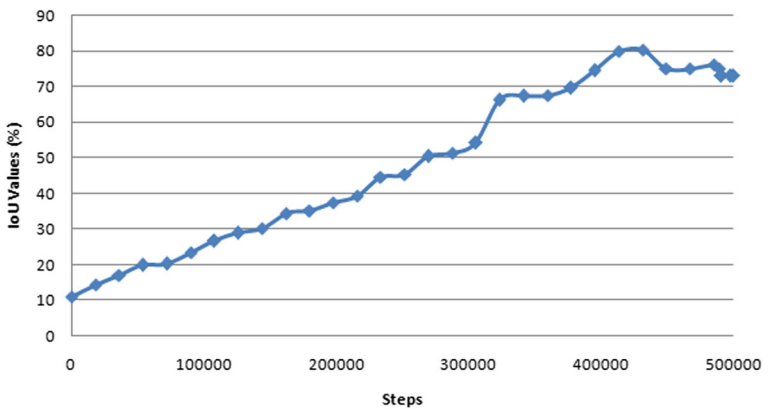


Fig. 10 Testing curve for Cityscapes per step number

Table 8 Results on CityScapes dataset

Model	road	sidewalk	building	wall	fence	pole	trafficlight	trafficsign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle	mIoU
ENet[42]									n/a											58.3
FCN-8[28]	97.4	78.4	89.2	34.9	44.2	47.4	60.1	65.0	91.4	69.3	93.9	77.1	51.4	92.6	35.3	48.6	46.5	51.6	66.8	65.3
Dilation[59]	97.6	79.2	89.9	37.3	47.6	53.2	58.6	65.2	91.8	69.4	93.7	78.9	55.0	93.3	45.5	53.4	47.7	52.2	66.0	67.1
DeepLab[21]	97.9	81.3	90.3	48.8	47.4	49.6	57.9	67.3	91.9	69.4	94.2	79.8	59.8	93.7	56.5	67.5	57.5	57.7	68.8	70.4
Wide-ResNet[58]	n/a																			78.4
PSPNet [61]	98.6	86.6	93.2	58.1	63.0	64.5	75.2	**79.2**	93.4	72.1	95.1	86.3	71.4	96.0	73.5	90.4	80.3	69.9	76.9	80.2
RM-denseNet	**98.7**	**86.8**	**93.5**	**58.3**	**63.3**	**64.5**	**75.2**	78.2	**93.7**	**72.4**	**95.2**	**86.3**	**71.4**	**96.2**	**73.7**	**90.8**	**80.7**	**70.1**	**77.1**	**80.3**

The bold numbers indicate the highest results

By comparing our method with recent state-of-the-art ones, our method emulates the visual system of the human brain by integrating recurrent connections inside the wider dense blocks which are abundant in the visual system. Besides, it follows a very deep CNN architecture which has proven its success in recent methods such as Segnet [4], Wide-ResNet [58], PSPNet [61] and FC-DenseNet [32]. Therefore, our RM-DenseNet has significantly improved the segmentation accuracy compared to all reported methods for both CamVid and Cityscapes datasets (see Tables 5 and 8).

Furthermore, the computational cost of our architecture has been identified by computed the time cost for each image. For CamVid dataset, each image requires about 180 ms since the resolution of images in this dataset is small. However, it takes 3 seconds for each CityScapes images since they are of large-scale. Besides, as it can be seen from Table 5, our RM-DenseNet with deeper and wider architecture has less number of parameters than all other methods and very comparable to FC-DenseNet [32].This will encourage more the use of our model.

The proposed measures evaluate inference time on the original architecture. When it comes to the deployment of such an architecture in a production system, specific hardware optimization is applied to speedup inference time. For example, the TensorRT library from NVIDIA applies network pruning and operators fusion. Such approach is very appropriate to networks based on Dense Blocks since the numerous connections can be pruned very efficiently thus strongly reducing the number of operations and thus processing speed. However, we do not evaluate on such optimized networks since it is application and hardware dependent so that we rely on the original architecture measures.

6 Conclusion and future work

In this paper, we have presented a Reinforced MultiScale fully connected DenseNet method which uses wider dense blocks and multiscale convolutional layers. The wider dense blocks improve the classical dense blocks by increasing the width of DB in the encoder part by building a recurrent structure. By this recurrent structure, the depth of our network will be increased without increasing of the number of parameters. Moreover, a multiScale 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 good results. For our future work, we plan to improve our RM-DenseNet method by optimizing its architecture.

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.

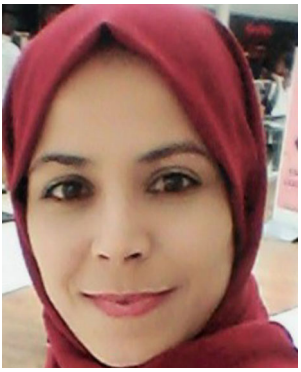
References

1. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Ghemawat S Tensorflow: Large-scale machine learning on heterogeneous distributed systems. Publicly available at: https://tensorflow.org
2. Alhaija H-A, Mustikovela S-K, Mescheder L, Geiger A, Rother C (2017) Augmented reality meets deep learning for car instance segmentation in urban scenes. In: British machine vision conference, vol 3
3. Audebert N, Le Saux B, Lefevre S (2016) Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. In: Asian conference on computer vision, pp 180–196

4. Badrinarayanan V, Kendall A, Cipolla R (2015) Segnet: a deep convolutional encoder-decoder architecture for image segmentation. arXiv:1511.00561,2015
5. Batenburg K-J, Sijbers J (2009) Adaptive thresholding of tomograms by projection distance minimization. Pattern Recogn 42(10):2297–2305
6. Batenburg K-J, Sijbers J (2009) Optimal threshold selection for tomogram segmentation by projection distance minimization. IEEE Trans Med Imaging 28(5):676–686
7. Ben Ahmed O, Benois-Pineau J, Allard M, Ben Amar C, Catheline G (2014) Classification of Alzheimer's disease subjects from MRI using hippocampal visual features. Multimedia Tools and Applications 74(4):1249–1266
8. Ben Aoun N, Elarbi M, Ben Amar C (2010) Multiresolution motion estimation and compensation for video coding. In: ICSP, pp 1121–1124
9. Ben Aoun N, Elghazel H, Ben Amar C (2011) Graph modeling based video event detection. In: IIT, pp 114–117
10. Ben Aoun N, Elghazel H, Hacid M-S, Ben Amar C (2011) Graph aggregation based image modeling and indexing for video annotation. In: CAIP, pp 324–331
11. Ben Aoun N, Elarbi M, Ben Amar C (2012) 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.), 23–40
12. Ben Aoun N, Mejdoub M, Ben Amar C (2014) Graph-based approach for human action recognition using spatio-temporal features. J Vis Commun Image Represent 25(2):329–338
13. Ben Aoun N, Mejdoub M, Ben Amar C (2014) Graph-based video event recognition. In: ICASSP, pp 1566–1570
14. Brahimi S, Ben Aoun N, Ben Amar C (2018) Boosted convolutional neural network for object recognition at large scale. NeuroComputing 330:337–354
15. Brahimi S, Ben Aoun N, Ben amar c (2016) Improved very deep recurrent convolutional neural network for object recognition. In: SMC, pp 2497–2502
16. Brahimi S, Ben Aoun N, Ben amar c (2016) Very deep recurrent convolutional neural network for object recognition. In: ICMV
17. Brahimi S, Ben Aoun N, Ben Amar C, Benoit A, Lambert P (2018) Multiscale fully convolutional densenet for semantic segmentation. In: International conference on computer graphics, visualization and computer vision
18. Brostow G-J, Fauqueur J, Cipolla R (2009) Semantic object classes in video: a high-definition ground truth database. Pattern Recogn Lett 30(2):88–97
19. Chen B-K, Gong C, Yang J (2017) Importance-aware semantic segmentation for autonomous driving system. In: Proceedings of the international joint conference on artificial intelligence, pp 1504–1510
20. Chen L-C, Barron J-T, Papandreou G, Murphy K, Yuille A-L (2016) Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform. In: IEEE conference on computer vision and pattern recognition, pp 4545–4554
21. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille A-L (2014) Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. arXiv:1606.00915
22. Cordts M, Omran M, Ramos S, Scharwachter T, Enzweiler T, Benenson R, Franke U, Roth S, Schiele B (2015) The cityscapes dataset. In: CVPR workshop on the future of datasets in vision
23. Dinarelli M, Tellier I (2016) Improving recurrent neural networks for sequence labelling. arXiv:1606.02555
24. Boughrara H, Chtourou M, Ben Amar C (2012) MLP neural network based face recognition system using constructive training algorithm. In: International conference on multimedia computing and systems (ICMCS), pp 233–238
25. El'Arbi M, Ben Amar C, Nicolas H (2006) Video watermarking based on neural networks. In: IEEE international conference on multimedia and expo (ICME), pp 1577–1580
26. Fabijanska A, Goclawski J (2014) New accelerated graph-based method of image segmentation applying minimum spanning tree. IET Image Process 8(4):239–251
27. Gao H, Zhuang L, Kilian Q-W (2016) Densely connected convolutional networks. arXiv:1608.06993v3
28. Garcia-Garcia A, Orts-Escolano S, Oprea S, Villena-Martinez V, Garcia-Rodriguez J (2017) A review on deep learning techniques applied to semantic segmentation. arXiv:1704.06857
29. Guedri B, Zaied M, Ben Amar C (2011) Indexing and images retrieval by content. In: International conference on high performance computing and simulation (HPCS), pp 369–375
30. He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision, pp 1026–1034
31. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778

32. Jégou S, Drozdzal M, Vazquez D, Romero A, Bengio Y (2017, July) The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation. In: Computer vision and pattern recognition workshops (CVPRW), pp 1175–1183
33. Kayalibay B, Jensen G, Smagt P (2017) CNN-based segmentation of medical imaging data. arXiv:1701.03056v2
34. Kendall A, Badrinarayanan V, Cipolla R (2015) Bayesian segnet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. arXiv:1511.02680
35. Krizhevsky A, Sutskever I, Hinton G (2012) Imagenet classification with deep convolutional neural networks. In: NIPS, pp 1097–1105
36. Lai S, Xu L, Liu K, Zhao J (2015, January) Recurrent convolutional neural networks for text classification. In: AAAI, vol 333, pp 2267-2273
37. Lin J, Wang W-J, Huang S-K, Chen H-C (2017) Learning based semantic segmentation for robot navigation in outdoor environment. In: Fuzzy systems association and 9th international conference on soft computing and intelligent systems (IFSA-SCIS), pp 1–5
38. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3431–3440
39. Mejdoub M, Fonteles L, Ben Amar C, Antonini M (2008) Fast indexing method for image retrieval using tree-structured lattices. In: International workshop on content-based multimedia indexing (CBMI), pp 365–372
40. Mejdoub M, Ben Aoun N, Ben Amar C (2015) Bag of frequent subgraphs approach for image classification. Intell Data Anal 19(1):75–88
41. Othmani M, Bellil W, Ben Amar C, Alimi AM (2010) A new structure and training procedure for multi-mother wavelet networks. Int J Wavelets Multiresolution Inf Process 8(1):149–175
42. Paszke A, Chaurasia A, Kim S, Culurciello E (2016) Enet: A deep neural network architecture for real-time semantic segmentation. arXiv:1606.02147
43. Pourian N, Karthikeyan S, Manjunath B-S (2015) Weakly supervised graph based semantic segmentation by learning communities of image-parts. In: Proceedings of the IEEE international conference on computer vision, pp 1359–1367
44. Qin A-K, Clausi D-A (2010) Multivariate image segmentation using semantic region growing with adaptive edge penalty. IEEE Trans Image Process 19(8):2157–2170
45. Raza S-H, Grundmann M, Essa I (2013) Geometric context from video. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 3081–3088
46. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg A-C, Fei-Fei L (2015) Imagenet large scale visual recognition challenge. Int J Comput Vis 115(3):211–252
47. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
48. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: CVPR, pp 1–9
49. Tieleman T, Hinton G (2012) rmsprop adaptive learning. In: COURSERA: neural networks for machine learning
50. Visin F, Ciccone M, Romero A, Kastner K, Cho K, Bengio Y, Matteucci M, Courville A (2016) Reseg: a recurrent neural network-based model for semantic segmentation. In: The IEEE conference on computer vision and pattern recognition (CVPR) workshops, pp 426–433
51. Visin F, Kastner K, Cho K, Matteucci M, Courville A-C, Bengio Y (2015) Renet: a recurrent neural network based alternative to convolutional networks. arXiv:1505.00393v3
52. Wali A, Ben Aoun N, Karray H, Ben Amar C, Alimi AM (2010) A new system for event detection from video surveillance sequences. In: ACIVS, pp 110–120
53. Wan J, Wang D, Hoi S-C-H, Wu P, Zhu J, Zhang Y, Li J (2014) Deep learning for content-based image retrieval: a comprehensive study. In: ACM international conference on multimedia, pp 157–166
54. Wang C, Yang H, Bartz C, Meinel C (2016) Image captioning with deep bidirectional LSTMs. In: Proceedings of the 2016 ACM on multimedia conference, pp 988–997
55. Wang C, Yang H, Meinel C (2015) Deep semantic mapping for cross-modal retrieval. In: Tools with artificial intelligence (ICTAI), pp 234–241
56. Wang C, Yang H, Meinel C (2016) Exploring multimodal video representation for action recognition. In: Neural networks (IJCNN), pp 1924–1931

57. Wang C, Yang H, Meinel C (2018) Image captioning with deep bidirectional LSTMs and multi-task learning. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 14(2s):40:1–40:20
58. Wu Z, Shen C, Hengel A-V-D (2016) Wider or deeper: revisiting the resnet model for visual recognition. arXiv:1611.10080
59. Yu F, Koltun V (2015) Multi-scale context aggregation by dilated convolutions. arXiv:1511.07122
60. Zhang K, Zhang W, Zeng S, Xue X (2014) Semantic segmentation using multiple graphs with Block-Diagonal constraints. In: AAAI, pp 2867–2873
61. Zhao H, Shi J, Qi X, Wang X, Jia J (2017) Pyramid scene parsing network. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 2881–2890
62. Zou W, Kpalma K, Ronsin J (2012) Semantic segmentation via sparse coding over hierarchical regions. In: Image processing (ICIP), pp 2577–2580



Sourour Brahimi received the B.S degree in Computer management and the M.S. master degree in business intelligence from the Higher Institute of Management Gabes (ISGG), in 2011 and 2014 respectively. She is now preparing her PhD in the Research Groupsin Intelligent Machine (REGIMLab), ENIS, Sfax (Tunisia). She is a member of IEEE and IEEE signal processing Society. Since 2015, she joined the Higher Institute of Management of Gabes (ISGG) as a lecturer. Her research interests include Computer Vision, Image classification and pattern recognition.

Najib Ben Aoun is an Assistant Professor in the College of Computer Science and Information Technology of Al-Baha University, KSA since 2015. He was teaching as Assistant Professor National School of Electronics and Telecommunications of the University of Sfax, Tunisia during 2014-2015. He received the B.S degree in computer sciences from the Higher Institute of Applied Sciences and Technologies of Sousse (ISSATS) in 2006. He obtained his M.S degree and the Ph.D degree in Computer Systems Engineering from the National Engineering School of Sfax (ENIS), Tunisia, in 2008 and 2014, respectively. He is now pursuing his post-doctoral researches within the REsearch Groups in Intelligent Machines (REGIM-Lab), Tunisia. His main research interests are focused on issues related to Computer Vision, Data Science and Biometrics. He published a list of papers in international conferences and journals. Dr. Najib served as a TC member and reviewer for several high reputed conferences (ICASSP'2012-2017, VCIP'2016, EUSIPCO'2015-2016, ICME'2015, MMSP'2013-2015, SoCPaR'2014, MELECON'2012, ISCI'2012, ICCAIE'2011) and journals (IEEE Transaction on MultiMedia, the Journal of Pattern Recognition, the Journal of Visual Communication and Image Representation, the Journal of Machine Vision and Applications and the Journal of Frontiers of Information Technology & Electronic Engineering). Dr. Najib is currently an ACM, IAPR, MIRLabs, IAES, IEEE and IEEE Signal Processing Society member. He was the treasurer of the SPS Tunisia Chapter in 2015-2016(Vice Chair in 2013-2014). He obtained an IEEE appreciation award for his contribution as Vice-chair of the SPS Tunisia Chapter in 2013-2014.

Alexandre Benoit received PhD degree in electronics and computer science from the University of Grenoble, INP in 2007. Starting 2008, he is an associate professor at Université Savoie Mont Blanc at LISTIC lab. His main research interest is related to Image and Video understanding. He actively participates to multimedia indexation challenges such as TRECVid within the IRIM French laboratory consortium. He develops features extraction methods for low levels image description and data fusion processes for high semantic level image and video indexing and segmentation. He contributes to the open source OpenCV library by providing a specific spatio-temporal filtering module, bioinspired.

Patrick Lambert received the PhD degree in signal processing in 1983 from the National Polytechnic Institute of Grenoble, France. He is currently a Full Professor at the School of Engineering of University Savoie Mont Blanc, France and a member of the Informatics, Systems, Information and Knowledge Processing Laboratory (LISTIC), Annecy, France. His research interests are in the field of image and video analysis, and actually dedicated to non-linear color filtering and automatic image understanding.

Chokri Ben Amar received the B.S. degree in Electrical Engineering from the National Engineering School of Sfax (ENIS) in 1989, the M.S. and PhD degrees in Computer Engineering from the National Institute of Applied Sciences in Lyon, France, in 1990 and 1994, respectively. He spent one year at the University of "Haute Savoie" (France) as a teaching assistant and researcher before joining the higher School of Sciences and Techniques of Tunis (ESSTT) as Assistant Professor in 1995. In 1999, he joined the Sfax University (USS) as Assistant Professor, and since 2011 as a full professor in the Department of Computer Sciences and Applied Mathematics of the National Engineering School of Sfax. Since September 2018, he is a full professor at the college of Computers and Information technology of Taief University in Saudi Arabia.

His research interests include Computer Vision and Image and video analysis. These research activities are centered on intelligent algorithms and their applications to data Classification and approximation, Pattern Recognition, Watermarking and image and video indexing and securing.

He is a senior member of IEEE since 2008. He founded the IEEE Signal Processing Society (SPS) Tunisia Chapter on January 2009, and he is actually the chair of this Chapter. During this period, the chapter organized five IEEE Distinguished Lectures and other technical and professional activities. He is the current advisor of the IEEE SPS Student Chapter in ENIS since 2010.

Affiliations

Sourour Brahimi[1] · Najib Ben Aoun[1,2] · Alexandre Benoit[3] · Patrick Lambert[3] · Chokri Ben Amar[1,4]

Sourour Brahimi
sourour.brahimi.TN@ieee.org

Alexandre Benoit
alexandre.benoit@univ-smb.fr

Patrick Lambert
patrick.lambert@univ-smb.fr

Chokri Ben Amar
chokri.benamar@ieee.org

1 REGIM-Lab.: REsearch Groups in Intelligent Machines, University of Sfax, National Engineering School of Sfax (ENIS), BP 1173, Sfax, 3038, Tunisia

2 Department of Computer Science, College of Computer Science and Information Technology, Al-Baha University, Al Baha, Saudi Arabia

3 LISTIC-Lab: Univ. Savoie Mont Blanc, LISTIC, Polytech Annecy Chambéry, 5 ch. de Bellevue, Annecy-le-Vieux, 74940, Annecy, France

4 Department of Computer Engineering, College of Computers and Information Technology, Taif University, Taif, Saudi Arabia