Convolutional neural network based image segmentation: a review

Hina Ajmal, Saad Rehman, Umar Farooq, Qurrat U. Ain, Farhan Riaz, et al.

Hina Ajmal, Saad Rehman, Umar Farooq, Qurrat U. Ain, Farhan Riaz, Ali Hassan, "Convolutional neural network based image segmentation: a review," Proc. SPIE 10649, Pattern Recognition and Tracking XXIX, 106490N (30 April 2018); doi: 10.1117/12.2304711

Event: SPIE Defense + Security, 2018, Orlando, Florida, United States

Convolutional Neural Network Based Image Segmentation: A Review

Hina Ajmal[a], Saad Rehman[*a], Umar Farooq[a], Qurrat U. Ain[b], Farhan Riaz[a], Ali Hassan[a]
[a] Department of Computer Engineering, National University of Sciences and Technology (NUST), Islamabad, Pakistan; [b] COMSATS Institute of Information Technology, Islamabad, Pakistan

ABSTRACT

Object recognition and semantic segmentation have been the two most common problems of traditional scene understanding in the computer vision domain. Major breakthroughs were reported in the last few years because of the increased utilization of deep learning, which offer a convincing alternative by learning the problem specific features on their own. In this paper, a summary of the frequently used framework – convolutional neural networks (CNN) is discussed. Accordingly a categorization scheme has been proposed to analyze the deep networks developed for image segmentation. Under this scheme, thirteen methods from the literature have been reviewed which are classified on the basis on how they perform segmentation operation i.e. semantic segmentation, instance segmentation and hybrid approaches. These method were reviewed from different aspects like their category, the novelty in the architecture of the method, and their special features in contrast with the traditional approaches. Latest review and analysis of these segmentation approaches, which provided outstanding results for image segmentation compared to the ordinary system, reveals that deep learning is increasingly becoming an important part of image segmentation and improvement in deep learning algorithms, which could resolve computer vision problems.

Keywords: Deep Learning, Convolutional Neural Network, Semantic Segmentation, Instance based Segmentation, Hybrid Segmentation

1. INTRODUCTION

Computer vision interprets the visual world at several levels: pixels, object parts, objects and beyond. The conventional approaches to deal with problems of computer vision, such as object identification and image segmentation were based on human engineered features like Histograms of Oriented Gradients (HOG) [1, 2], Scale Invariant Feature Transform (SIFT) and, Bag-of-features and Speeded Up Robust Features (SURF) image representations as well as deformable part models [3-10].Basically, there are two main components from the computer vision perspective: the design of features and developing a learning algorithm. Designing the meaningful features is the main challenge in pixel level classification problems. Some of the related works include Random Forest based classifiers, Texton-forest and Texton-Boost [11-13].Deep learning has outperformed these methods on many applications such as image sorting and object detection. Paul and Singh [14] gave a brief review on advances in deep learning. The foundations of deep learning deal with many phases of data processing in hierarchical structures. These structures can be employed for learning the features and pattern cataloguing. Deep learning networks are based on deep architectures as well as intellectual learning algorithms [15]. A formal definition is: "deep architectures are the compositions of many layers of adaptive non-linear components"[16]. Initial deep models included: the Restricted Boltzmann Machine, Deep Belief Networks (DBN), and stacked auto-encoders. These offer advantages of node efficiency, reduction of the amount of redundant work and to ability to approximate more complex functions with the same accuracy using fewer total nodes [17-20].

Although there are several architectures for deep networks, CNN and DBN are the milestones that are very popular among the deep learning community. Recently, the CNN based model won its first image segmentation challenge in 2012 and then up until 2015 all the best performing algorithms were developed based on deep convolutional neural networks [21]. Impressed by the increased performance with using CNN for image sorting and object recognition, scientists adopted CNN for image segmentation. Recently, studies have been done for image segmentation using CNN and its variants. Figure 1 shows the evolution of Deep learning models and its achievements.

Semantic segmentation is common but it classifies the image at the pixel level and each pixel can belong to a set of predefined classes and have no notion for instance level segmentations, however, a lot of problems can be solved if instance segmentation is used. For example, it can enable a robot to segment a particular object with the aim of grasping

Pattern Recognition and Tracking XXIX, edited by Mohammad S. Alam, Proc. of SPIE Vol. 10649, 106490N · © 2018 SPIE · CCC code: 0277-786X/18/$18 · doi: 10.1117/12.2304711

it. Recently, researchers tried to use the best of both worlds and introduced instance aware semantic segmentation that outperformed other methods in scene understanding problems.

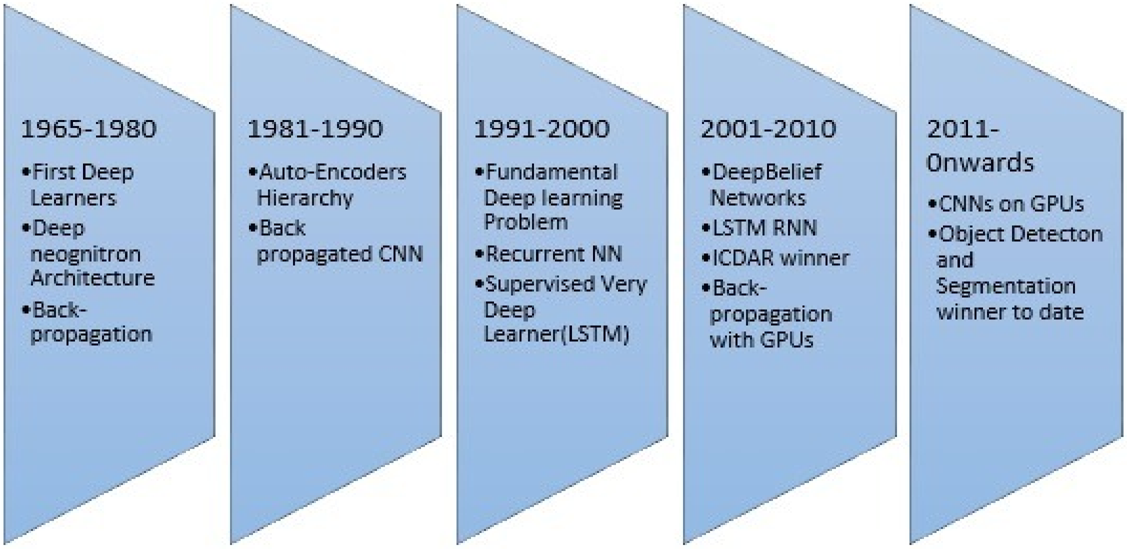

Figure 1. Evolution of Deep Learning and its achievements over a few decades

Deep learning is a progressing field and a lot of work has been done in the past few decades. Only a few surveys can be seen on deep learning based image grouping and segmentation [22-24]. In this paper, the aim is to organize the plethora of solutions for segmentation using CNN and its variants. So, first a brief introduction to how computer vision and deep learning are being used together to solve segmentation problems is provided. Accordingly, a summary of the deep learning building block CNN that is in focus nowadays and is outperforming the solutions presented earlier. A classification scheme has been proposed that organizes the recent deep learning literature for image segmentation into three categories depending on how it performs the segmentation operation: Semantic; Instance Based; and Hybrid, which combines both semantic segmentation and instance masks. Afterwards, a summary and up-to-date review of the evolution of image segmentation techniques using deep learning concepts is presented and then concluded by discussing examples of their performance on some commonly used datasets.

2. EVOLUTION OF DEEP LEARNING BASED SEGMENTATION TECHNIQUES

2.1 Deep Convolutional Neural Network Architecture

A back-propagated CNN was first proposed by LeCunn [25]. It is a hierarchical neural network consisting of three 'hidden' layers that perform the different functions of convolution and pooling, and is fully connected. The network starts with an input stage followed by hidden stages and ending with an output stage. The precision of the network is improved by using more hidden layers [26]. The output layer provides a transformed classified result for the input given. A basic CNN design is presented in the figure 2.

The convolution and pooling layers replicate the simple and complicated units in the visual cortex. The convolution layer is used for extracting localized features from the input using filters. Then the pooling operation sub-samples the results. Eventually the output of the pooling is passed through the convolution layer in order to refilter. The process continues and the results are passed into the network through multiple pooling and convolution layers. The total number of layers depends upon which network architecture is selected. The features within the network are learned by the data.

The features from the final pooling stage are then processed using a flattening level which is trailed by a fully connected convolution stage. Eventually, results from fully connected layer are passed through a Multinomial Logistic Regression (MLR) output layer [27, 28]. The precise order and type of the layers depends on the network architecture.

Using conventional neural networks for images is not very practical because you have to deal with millions of multiplications of very large size matrices. One way to avoid computational complexity is to use 2D convolutions.

Training a set of filters is more convenient to handle than to learning a matrix of very large dimensions. The convolutions are similar to neural networks with restrictions of local connectivity and weight sharing [29].

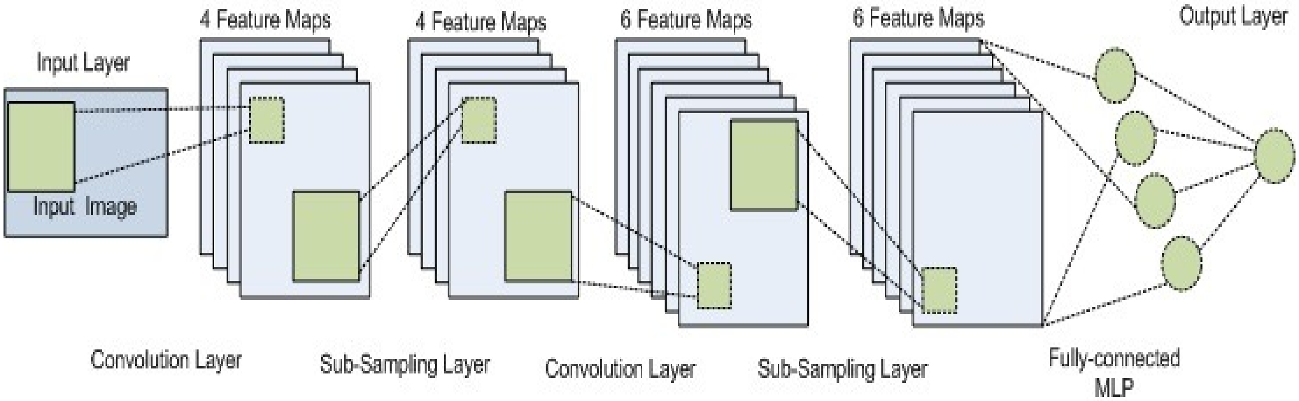

Figure 2. A Basic Convolutional Neural Network Architecture

According to the Universal Approximation Theorem [30], a neural network consisting of one hidden stage is enough to construct any function. Though, it has been determined that networks with fewer hidden layers need a larger number of neurons than networks with a larger number of hidden layers [31].

Recently it has been clearly revealed that deeper networks can perform much better than shallow networks [32]. In 2006, Greedy layer-wise pre-training was introduced and scholars were capable to use deeper networks [18]. Before greedy layer wise approach, no algorithm existed that could train the deep networks efficiently. Many recent networks have many multiple layers. For example, GoogleNet and VGGnet have 22 and 19 layers, respectively [33, 34]. The convolutions in the architectures usually use a non-linearity operation after each stage. The typical feed-forward networks use sigmoid and tangent hyperbolic non-linearity but modern CNN use nonlinearity i.e.

$$\text{ReLU(x)} = \max(0, x) \tag{1}$$

The deep networks using this non-linearity can be trained faster than the networks using conventional non-linearity [35]. Later a new non-linearity function known as leaky-ReLU was proposed [36]:

$$Leaky\_\operatorname{Re}LU(x) = \max(0, x) + \alpha \min(0, x) \tag{2}$$

Leaky ReLU uses α as a fixed parameter but it was suggested that if $\propto$ is learnt, it leads to a better model [37].

This algorithm in general learns by minimizing a cost function. Backpropagation algorithm is used for training the neural networks that depends on the chain rule to accelerate the gradient computation for Gradient Descent (GD) [38]. The GD is not practical for large datasets. Stochastic Gradient Descent (SGD) performs generalization in the case of large datasets [39, 40] but convergence is slow. To respond to this, SGD with batches of small numbers of data points has been used.

2.2 Deep Networks based Image Segmentation

The biggest advantage of the CNNs for computer vision related applications is that the system is trained end-to-end as a whole, from the pixel level to the final classes, and thus lessens the need to create a feature extractor [41]. The basic disadvantage is its requirement for labelled training examples.

The problems related to structured estimate where every pixel in the image is required to be classified were first tackled using hierarchical features from a network but even this method needs post processing for removing the over-segmented parts [42]. After that many researchers indulged in using CNN and its variants to get better results for image segmentations. Figure 3 is a brief summary of the recent developments in segmentation techniques based on deep learning.

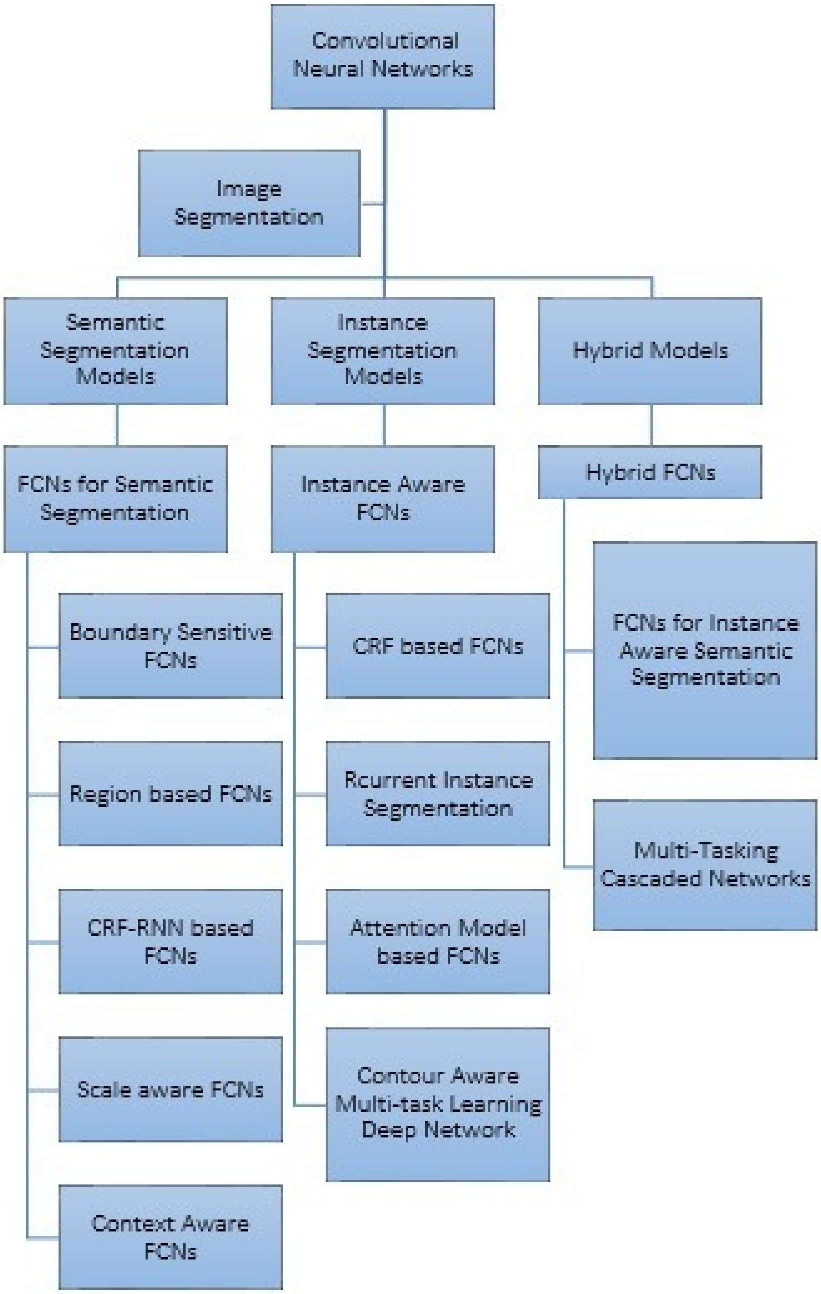
Convolutional Neural Networks
Image Segmentation
Semantic Segmentation Models
Instance Segmentation Models
Hybrid Models
FCNs for Semantic Segmentation
Instance Aware FCNs
Hybrid FCNs
Boundary Sensitive FCNs
Region based FCNs
CRF-RNN based FCNs
Scale aware FCNs
Context Aware FCNs
CRF based FCNs
Rcurrent Instance Segmentation
Attention Model based FCNs
Contour Aware Multi-task Learning Deep Network
FCNs for Instance Aware Semantic Segmentation
Multi-Tasking Cascaded Networks

Figure 3. Proposed Hierarchy of Deep Learning based Image Segmentation Techniques

3. SEMANTIC SEGMENTATION

Semantic segmentation deals with the task of assigning every pixel the class label. Although pixel level labelling provides explicit descriptions of the images, it is far more computationally complicated. Recently, advances in semantic segmentations have been made using deep learning concepts.

3.1 Fully Convolutional Networks

A fully convolutional network has been proposed that is trained end-to-end and pixel to pixel, and has outperformed conventional semantic segmentation [43]. The main vision is to develop fully a connected network, shown in Figure 4 that can operate on an input of any size and can produce results with effective learning. The classification networks (GoogleNet, VGGnet, AlexNet) [44] are transformed into fully convolutional networks and their learned models are used for the segmentation task. This model is trained for pixel-wise predictions using supervised pre-training.

The learning and prediction are done using the entire image at a time by utilizing dense feed-forward computation. Up-sampling layers incorporated in the network allow the network to learn and predict with subsampled pooling. Semantic segmentation has some problems regarding the coherence between semantics and position. Global information answers the 'what' questions and local information is about 'where'. This model introduces a skip architecture for segmentation that merges the data form the deep layers with the appearance information form the shallow layers and consequently it increases the spatial accuracy of the output.

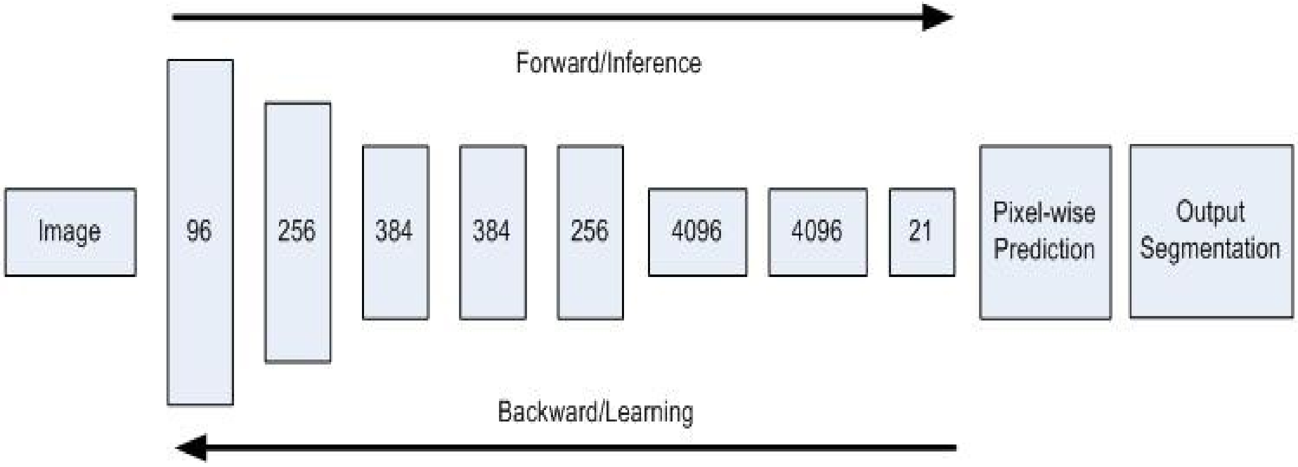

Figure 4. FCN for semantic segmentation

3.2 Boundary Neural Fields in FCNs

The FCNs for semantic segmentation have receptive fields of very large size and also contain many pooling layers. Both receptive fields and pooling layers are responsible for low spatial resolution and blurring effects in the deep layers. This adversely affects the boundaries around the objects. At first, this problem was handled in the post-processing steps by means of a dense Conditional Random Field (CRF) but this introduced new constraints that are troublesome to integrate in the existing architecture [45].

A global energy model - Boundary Neural Field (BNF) is presented in [46] that combines the FCN results with boundary detections. This architecture uses a single FCN for finding the boundaries and then utilizes global optimization to predict segmentations. The fundamental novelty of this algorithm is the use of boundary neural fields. Like the conventional global methods, the BNF fields are defined using the unary and pairwise potentials. Minimizing the global energy fields gives the segmentation. The most commonly used global energy functions are Conditional Random Fields (CRF) [47], Markov Random Fields (MRFs) [48] and Graph Cuts [49]. But it is very difficult to find local optima in all these methods, so a new energy function is used in the BNF model that overcomes this shortcoming. The energy function proposed has two terms. The first one gives the unary energy and the other one gives the pairwise energy. The results of this model showed that the boundary based global minimization provided better inference than other global inference methods like Iterated Conditional Mode (ICM) or Belief Propagation [48].

3.3 Region-based FCNs

Fully convolutional networks are based on using squares of fixed size for making predictions but in the real world objects can be of arbitrary form and of varying sizes. A modified region based FCN was developed that is capable of end-to-end training and deals with free-form objects of different sizes [50]. In this model a region to pixel phase is added before the final layer that allows the region to pixel mapping. This layer does not consider regions having less candidates for all classes as they do not affect the output labelling. Another addition is the use of a differentiable ROI pooling layer. This pooling layer is used for making the model operable on free form objects.The algorithm specifically is very precise at object boundaries.

3.4 CRF-RNN based FCNs

In [51], a model combining the powers of convolution networks and Conditional Random Fields (CRFs) based modelling was presented. CRFs are used for probabilistic graphical modelling. The CRF with Gaussian pairwise potentials is formulated as a Recurrent Neural Network. This CRF-RNN network is plugged into a convolutional network. This integration allows refining of the outputs obtained from a traditional CNN in forward processing and allows training of the network end to end using the back propagation algorithm. This algorithm offers the capability to avoid post-processing steps for item delineation.

3.5 Scale-Aware FCNs

Multi-scale features are an essential factor for semantic segmentation. Some models using such features are [52, 53]. The network structures that can utilize the multi-scale features are skip-net and share-net. Attempts were made to incorporate multi-scale features in FCN using an attention model in [54]. The FCN is modified to a share net. Soft attention is used for generalization of the pooling layers for multiple scales. This model weights the multiscale features respect to the concerned item in the image. The output is a pixel-wise weight map. The weighted sum of all the score maps produced by FCN is used for classification. One of the advantages of the model is that it makes it possible to analyze the significance of features at different locations and measures.

3.6 Context-Aware FCNs

Contextual information plays a vital role in scene understanding tasks. Spatial context can be determined by two means: first the link between the object and the background; and second the relation between the object and its neighboring items. A CRFs based CNN has been proposed to utilize the patch-patch and patch-background context for semantic segmentation [55]. In this network, the patch-wise contextual relationships are modelled using CRFs and the background-patch context is exploited by providing the network with multi-scale input. The pairwise potential functions in CNN are learned to model semantic compatibility between different patches. One specialty of this model is that it uses the piecewise training of the CRFs to prevent any redundancy in the inference and thus provides efficient learning.

4. INSTANCE BASED SEGMENTATION

Instance segmentation is the delineation of different objects of interest in an image. Segmenting the image at the instance level is imperative in a wide range of uses like autonomous driving and visual question answering. Instance segmentation is more puzzling than semantic segmentation as it demands the clustering of the pixels to be evaluated.

At first, category independent region proposals are classified using CNN [56] and afterwards the same architecture is modified to get the category-specific predictions [57]. Since generating the accurate region proposal is a challenging task in itself, a proposal free network has been described for the instance level segmentation problem [58]. This end-to-end trained network outputs the bounding box for each instance and the confidence score for different classes of each pixel.

4.1 Instance FCNs

FCNs have proved very helpful for semantic segmentation but they have no notion of object instances. The FCNs are now able to predict instance level segments. Instance aware FCNs have been developed [59], an example being shown in figure 5. This is an end-to-end trained model that can segment the candidate object. Similar to the traditional FCN for semantic segmentation, each pixel is still a classifier but in contrast with an FCN it gives a score map for every category. Thus the method computes an instance sensitive score map. Unlike other methods used for instance segmentation this model does not include a complex layer for mask resolution but rather it uses a local coherence for making predictions about instance categories.

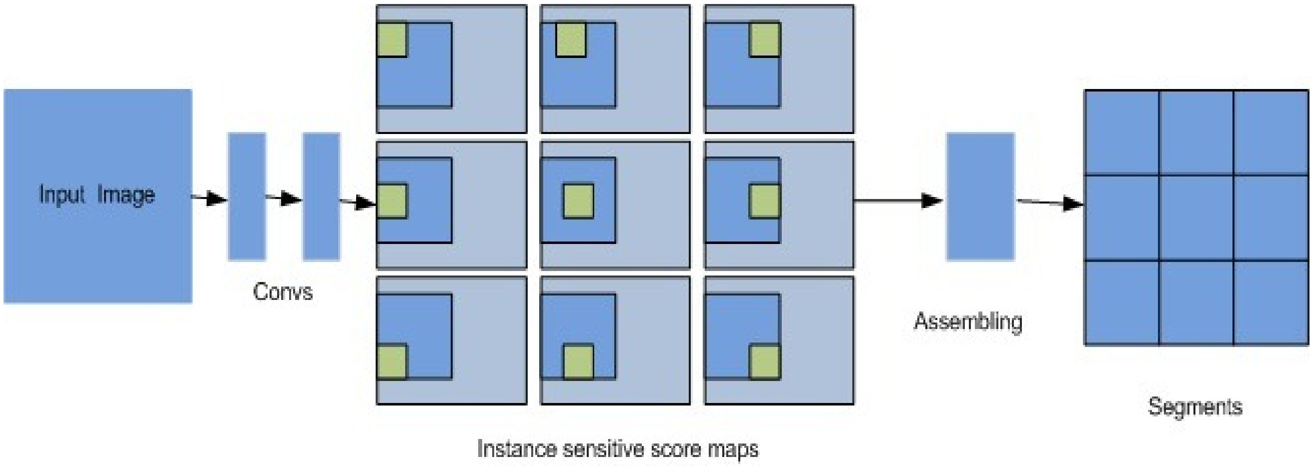

Figure 5. Illustration of Instance-sensitive FCN reproduced from [59]

4.2 CRF based FCNs

It has been stated that "Segmentation lies at the intersection of Object Detection" [60]. These authors proposed a semantic segmentation model modified with high order CRFs for instance level segmentation. One of the core characteristics of this model is that it works in a bottom up manner. This network used end-to-end trainable high order potentials. This a different approach to instance segmentation where semantic segmentation is performed at the category level before different instances of the objects are considered. The high order CRFs used for instance segmentation are differentiable and are obtained as a result of object detection.

4.3 Recurrent Instance Segmentation

A new end-to-end trained segmentation method has been proposed to segment instances sequentially one at a time [61][45]. The recurrent neural network is a basic building block of this model that finds an object in a sequence and segments it. To keep a record of which pixels have been evaluated, this method uses a spatial memory, named a Long Short Term Memory (LSTM) [62, 63]. The use of LSTM allows the network to deal with occlusion problems. A principal loss function is designed that correctly features the properties of the instances of the objects.

4.4 Attention Model based FCNs

An end-to-end RNN model modified with an attention mechanism that can replicate human-like counting has been presented [64]. This network considers visual attention and addresses counting and instance segmentation jointly. The method also resolves the dimensionality issue as it uses a temporal chain with the output being only one instance at a time. It allows us to make predictions about occluded objects by performing non-maximal suppression using the already segmented object. This network has been trained to produce Region of Interests (ROIs) sequentially as well as detailed object segmentations within each ROI. This attention based model performs better than methods that operate on the entire image.

4.5 Contour Aware FCNs

A contour aware deep network has been proposed based on the united multi-task learning framework [65]. Contextual features at multiple levels are explored with end-to-end trained FCN to handle large appearance variation. The proposed model allows to input an image and output the probability map of the input image of the same resolution with only a single forward pass. It also includes a supervision method to avoid the difficulty of disappearing gradients in the learning of the network. This method depicts obvious contours for separated clustered object instances. Deep learning based techniques in most of the cases examined outperformed the standard instance segmentation methods.

5. HYBRID APPROACHES FOR SEGMENTATION

In this section we review some approaches in which standard deep networks are modified for instance aware semantic segmentation in order to make use of the best of both the worlds i.e. semantic segmentation and instance level segmentation. For this purpose either we can incorporate the idea of instances in models for semantic segmentation or vice versa. In [66], a semantic segmentation CNN architecture was modified for instance segmentation. This so-called multi-task Network Cascade was proposed for instance aware semantic segmentation. As the name suggests, these networks are cascaded and are intended to share the features from CNN. A new algorithm for training this cascaded network had also to be developed. It is a single stage framework but can be enhanced for multiple stages. This model is able to differentiate the instances and categorize the objects.

Recently, instance aware FCNs were modified to inherit all the merits of FCNs for semantic segmentation and instance mask generation [67][49], an example being shown in Figure 6.1. The network utilizes a box instead of sliding windows. The network simultaneously performs instance mask prediction and classification. This architecture is entirely shared between these two tasks and among all ROIs. ROI based computation is time-saving as no warping and resizing operations are involved.

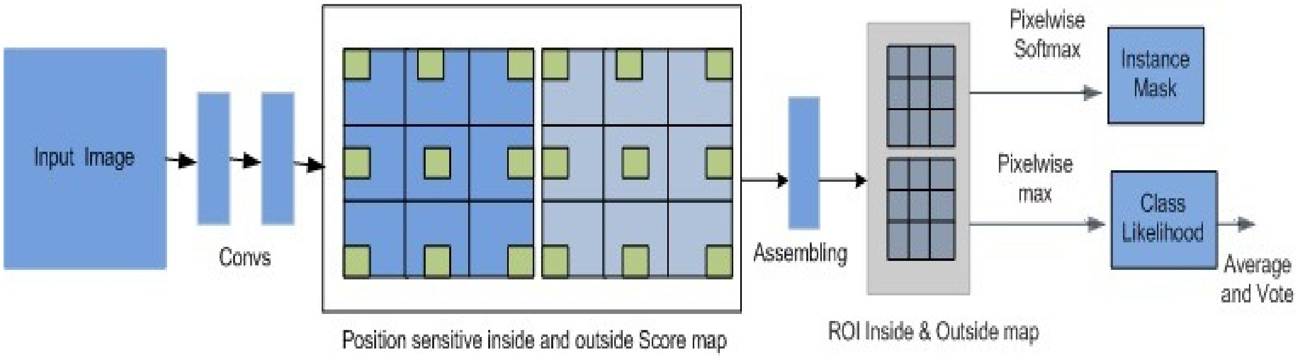

Figure 6. Instance FCN modified for semantic, modified from [67]

6. SUMMARY OF DEEP LEARNING BASED IMAGE SEGMENTATION TECHNIQUES

In table 1, all the abbreviations used in the next tables are mentioned. In Table 2, a brief impression of all the approaches is given. The Category column indicates the level of operation for segmentation. Model gives a short name for the method. In the next columns the novelty of the architecture and special features are summarized.

Table 1. Acronyms used in paper

Acronym	**Stands for**
FCN	Fully Convolutional Networks
EE	End-to-end Training
BNF	Boundary Neural Field
CRF-RNN	Conditional Random Fields-Recurrent Neural Network
CN	Convolutional Network
AM	Attention Model/Mechanism
RIS	Recurrent Instance Segmentation
AM-FCN	Attention model based FCN
ROI	Region of Interest
FCISS	Fully Convolutional Instance-aware semantic segmentation
ISS	Instance Semantic Segmentation

Table 2. Deep Learning Based Image Segmentation Techniques Summary

S No.	Category	Model	Novelty of the Model Architecture	Characteristics of the Approach
1	Semantic	FCNs	Fully connected, supervised EE, pixel wise prediction	Arbitrary size inputs and outputs
2	Semantic	BNF Model	Global energy model incorporated with FCN	Handles the blurring and low spatial resolution of deep layers
3	Semantic	Region based FCN	EE trained network labelling pixels based on fixed patches	Spatial support for appearance measurements
4	Semantic	CRF-RNN	A CN combined with an RNN for CRF prediction	No post processing required
5	Semantic	Scale aware FCN	A FCN combined with an AM capable of weighting multiscale features	Analysis of features at multiple scales and locations
6	Semantic	Context aware FCN	Piece-wise learning of CRF integrated CN	Uses patch-patch and patch-background context
7	Instance	Instance aware FCN	FCN trained to give instance-level segment score maps	Use of local coherence for instance prediction
8	Instance	CRF-FCN	FCN with high order differentiable CRF with EE	Localize object instance
9	Instance	RIS	A RNN combined with FCN to sequentially segment the instances	Segments one instance at a time, Occlusion handling
10	Instance	AM-FCN	A FCN with AM to detect ROI in sequence	Imitates human like counting
11	Instance	Contour-aware FCN	A EE-FCN with AM for multi-level contextual features	Handles appearance variation and vanishing gradients problem
12	Hybrid	FCISS	A FCN for semantic segmentation integrated with instance proposal	Output includes segmentation as well as classification
13	Hybrid	ISS	A multi-task network, cascaded structure, EE	Object detection is a by-product

7. PERFORMANCE MEASURES AND DATASETS

For the assessment of the performance of object segmentation and classification algorithms, appropriate datasets and performance measures are necessary. Datasets are responsible for individual assessment of efficiency and when used with accuracy measures, they simplify objective assessment. They also offer a reference ground for algorithm development. In the following section, we reproduced the performance for the most commonly used datasets that have been used for evaluating algorithms we reviewed i.e. PASCAL VOC [68], NYUD [69], SIFT Flow [70][52], MS COCO [71], MICCAI segmentation Challenge Dataset [72].The most commonly used evaluation criteria that are employed for semantic segmentation are mean of the region intersection over union (meanIU) and mean Average Precision (meanAP). The meanIU is used for evaluating semantic segmentation and meanAP is used for instance level segmentation.
Table 3 gives best accuracies of the methods for semantic segmentation. It indicates on which datasets the algorithm was evaluated and the performance for the corresponding dataset is given in terms of mean IU. Similarly, Table 4 provides the performance of different models for several datasets for Instance level segmentation using the mean AP and Table 5 does the same for hybrid methods.

Table 3. Performance of Models for Semantic Segmentation using mean IU for different datasets

Model	PASCAL VOC	PASCAL CONTEXT	NYUD	SIFT flow	MS COCO
FCN	62.7	N	34.0	39.50	N
BNF	77.60	N	N	N	N
Region based FCN	N	32.50	N	49.90	N
CRF-CNN FCN	74.70	39.28	N	N	N
Scale-aware FCN	71.42	N	N	N	35.78
Context-aware FCN	78	43.30	40.60	44.90	N

Table 4. Performance of Models for Instance Segmentation using mean AP

MODEL	PASCAL VOC	CITYSCAPES	MSCOCO
Instance FCN	52.60	N	39.20
CRF based FCN	58.30	N	N
RIS	50.10	N	N
AM based FCN	N	46.8	N
Contour-aware FCN	N	N	N

*contour aware FCN are designed for histology images and tested on the MICCAI challenge dataset
N=Not used for testing

Table 5. Performance of Hybrid methods

MODEL	PASCAL VOC (mAP @ 0.5)
FCISS (Fully Convolutional Instance-Aware Semantic Segmentation)	65.70
ISS (Instance-Aware Semantic Segmentation)	63.50

8. CONCLUSION

Segmentation is an essential part of scene understanding. Image segmentation is done mostly by extracting the features and designing a model that can be trained and then used to predict the segmentation. Designing the meaningful features is a challenging task. In recent years, great breakthroughs have been made to design models that can learn the best suited features on their own. Deep learning became popular and a lot of research has been performed into the method. There are only a few survey papers in the literature recently published that focus on the deep learning trends. To the best of our knowledge, this is the first review that categorized the deep learning techniques applied for segmentation in terms of whether it segmentation has been done pixel-wise or instance-wise. This paper covers the most recent state-of-the-art segmentation based on techniques deep learning. We first provided some details about foundations of deep learning and how it performed well in object detection and recognition. Then we gave necessary background knowledge about convolutional networks. We covered thirteen methods from literature that are divided into 3 types: semantic, instance based, hybrid, according to proposed classification scheme. These method were reviewed from different aspects like their category, the novelty in the architecture of the method, and their special features in comparison with the traditional approaches. We then present briefly about different datasets and performance measures used for evaluating these techniques and provided a summary of the accuracies. To conclude the results of the deep learning based models were shown to outperform the traditional segmentation methods and deep learning has proved extremely powerful and in future we expect it to be very helpful in providing ingenious solutions to segmentation problems.

REFERENCES

[1] N. Dalal, and B. Triggs, "Histograms of oriented gradients for human detection." 1, 886-893.

[2] W. T. Freeman, and M. Roth, "Orientation histograms for hand gesture recognition." 12, 296-301.

[3] H. Bay, T. Tuytelaars, and L. Van Gool, “Surf: Speeded up robust features,” Computer vision–ECCV 2006, 404-417 (2006).

[4] G. Csurka, C. Dance, L. Fan *et al.*, "Visual categorization with bags of keypoints." 1, 1-2.

[5] P. F. Felzenszwalb, R. B. Girshick, D. McAllester *et al.*, “Object detection with discriminatively trained part-based models,” IEEE transactions on pattern analysis and machine intelligence, 32(9), 1627-1645 (2010).

[6] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories." 2, 2169-2178.

[7] D. G. Lowe, "Object recognition from local scale-invariant features." 2, 1150-1157.

[8] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” International journal of computer vision, 60(2), 91-110 (2004).

[9] F. Perronnin, J. Sánchez, and T. Mensink, “Improving the fisher kernel for large-scale image classification,” Computer Vision–ECCV 2010, 143-156 (2010).

[10] J. Sivic, and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos." 1470.

[11] L. Deng, “A tutorial survey of architectures, algorithms, and applications for deep learning,” APSIPA Transactions on Signal and Information Processing, 3, (2014).

[12] J. Shotton, T. Sharp, A. Kipman *et al.*, “Real-time human pose recognition in parts from single depth images,” Communications of the ACM, 56(1), 116-124 (2013).

[13] J. Shotton, J. Winn, C. Rother *et al.*, “Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context,” International Journal of Computer Vision, 81(1), 2-23 (2009).

[14] S. Paul, and L. Singh, "A review on advances in deep learning." 1-6.

[15] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation." 1-8.

[16] Y. Bengio, and Y. LeCun, “Scaling learning algorithms towards AI,” Large-scale kernel machines, 34(5), 1-41 (2007).

[17] G. E. Hinton, “Training products of experts by minimizing contrastive divergence,” Neural computation, 14(8), 1771-1800 (2002).

[18] G. E. Hinton, S. Osindero, and Y.-W. Teh, “A fast learning algorithm for deep belief nets,” Neural computation, 18(7), 1527-1554 (2006).

[19] P. Vincent, H. Larochelle, I. Lajoie *et al.*, “Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion,” Journal of Machine Learning Research, 11(Dec), 3371-3408 (2010).

[20] N. Le Roux, and Y. Bengio, “Representational power of restricted Boltzmann machines and deep belief networks,” Neural computation, 20(6), 1631-1649 (2008).

[21] O. Russakovsky, J. Deng, H. Su *et al.*, “Imagenet large scale visual recognition challenge,” International Journal of Computer Vision, 115(3), 211-252 (2015).

[22] P. Druzhkov, and V. Kustikova, “A survey of deep learning methods and software tools for image classification and object detection,” Pattern Recognition and Image Analysis, 26(1), 9 (2016).

[23] B. Zhao, J. Feng, X. Wu *et al.*, “A survey on deep learning-based fine-grained object classification and semantic segmentation,” International Journal of Automation and Computing, 1-17 (2017).

[24] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea *et al.*, “A Review on Deep Learning Techniques Applied to Semantic Segmentation,” arXiv preprint arXiv:1704.06857, (2017).

[25] Y. LeCun, B. Boser, J. S. Denker *et al.*, “Backpropagation applied to handwritten zip code recognition,” Neural computation*,* 1(4), 541-551 (1989).

[26] D. C. Ciresan, U. Meier, J. Masci *et al.*, "Flexible, high performance convolutional neural networks for image classification." 22, 1237.

[27] C. Kwak, and A. Clayton-Matthews, “Multinomial logistic regression,” Nursing research*,* 51(6), 404-410 (2002).

[28] A. A. L. L. Sandoval-Mejıa, and Y. E. Wang, [Multinomial Logistic Regression], (2016).

[29] C. M. Bishop, [Pattern recognition and machine learning] springer, (2006).

[30] K. Hornik, “Approximation capabilities of multilayer feedforward networks,” Neural networks*,* 4(2), 251-257 (1991).

[31] Y. Bengio, “Learning deep architectures for AI,” Foundations and trends® in Machine Learning*,* 2(1), 1-127 (2009).

[32] S. Srinivas, R. K. Sarvadevabhatla, K. R. Mopuri *et al.*, “A taxonomy of deep convolutional neural nets for computer vision,” arXiv preprint arXiv:1601.06615, (2016).

[33] K. Simonyan, and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, (2014).

[34] C. Szegedy, W. Liu, Y. Jia *et al.*, "Going deeper with convolutions." 1-9.

[35] V. Nair, and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines." 807-814.

[36] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models." 30.

[37] K. He, X. Zhang, S. Ren *et al.*, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification." 1026-1034.

[38] L. Mason, J. Baxter, P. L. Bartlett *et al.*, "Boosting algorithms as gradient descent." 512-518.

[39] L. Bottou, [Large-scale machine learning with stochastic gradient descent] Springer, (2010).

[40] T. Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms." 116.

[41] P. Sermanet, D. Eigen, X. Zhang *et al.*, “Overfeat: Integrated recognition, localization and detection using convolutional networks,” arXiv preprint arXiv:1312.6229, (2013).

[42] C. Farabet, C. Couprie, L. Najman *et al.*, “Learning hierarchical features for scene labeling,” IEEE transactions on pattern analysis and machine intelligence*,* 35(8), 1915-1929 (2013).

[43] E. Shelhamer, J. Long, and T. Darrell, “Fully convolutional networks for semantic segmentation,” IEEE transactions on pattern analysis and machine intelligence*,* 39(4), 640-651 (2017).

[44] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks." 1097-1105.

[45] L.-C. Chen, G. Papandreou, I. Kokkinos *et al.*, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” arXiv preprint arXiv:1606.00915, (2016).

[46] G. Bertasius, J. Shi, and L. Torresani, "Semantic segmentation with boundary neural fields." 3602-3610.

[47] P. Krähenbühl, and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials." 109-117.

[48] M. F. Tappen, and W. T. Freeman, "Comparison of graph cuts with belief propagation for stereo, using identical MRF parameters." 900.

[49] Y. Boykov, O. Veksler, and R. Zabih, “Fast approximate energy minimization via graph cuts,” IEEE Transactions on pattern analysis and machine intelligence*,* 23(11), 1222-1239 (2001).

[50] H. Caesar, J. Uijlings, and V. Ferrari, "Region-based semantic segmentation with end-to-end training." 381-397.

[51] S. Zheng, S. Jayasumana, B. Romera-Paredes *et al.*, "Conditional random fields as recurrent neural networks." 1529-1537.

[52] B. Hariharan, P. Arbeláez, R. Girshick *et al.*, "Hypercolumns for object segmentation and fine-grained localization." 447-456.

[53] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich, "Feedforward semantic segmentation with zoom-out features." 3376-3385.

[54] L.-C. Chen, Y. Yang, J. Wang *et al.*, "Attention to scale: Scale-aware semantic image segmentation." 3640-3649.

[55] G. Lin, C. Shen, A. van den Hengel *et al.*, "Efficient piecewise training of deep structured models for semantic segmentation." 3194-3203.

[56] R. Girshick, J. Donahue, T. Darrell *et al.*, "Rich feature hierarchies for accurate object detection and semantic segmentation." 580-587.

[57] B. Hariharan, P. Arbeláez, R. Girshick *et al.*, "Simultaneous detection and segmentation." 297-312.

[58] X. Liang, Y. Wei, X. Shen *et al.*, “Proposal-free network for instance-level object segmentation,” arXiv preprint arXiv:1509.02636, (2015).

[59] J. Dai, K. He, Y. Li *et al.*, "Instance-sensitive fully convolutional networks." 534-549.

[60] A. Arnab, and P. H. Torr, “Bottom-up instance segmentation using deep higher-order crfs,” arXiv preprint arXiv:1609.02583, (2016).

[61] B. Romera-Paredes, and P. H. S. Torr, "Recurrent instance segmentation." 312-329.

[62] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, “Learning precise timing with LSTM recurrent networks,” Journal of machine learning research*,* 3(Aug), 115-143 (2002).

[63] M. Sundermeyer, R. Schlüter, and H. Ney, "LSTM neural networks for language modeling."

[64] M. Ren, and R. S. Zemel, “End-to-end instance segmentation and counting with recurrent attention,” arXiv preprint arXiv:1605.09410, (2016).

[65] H. Chen, X. Qi, L. Yu *et al.*, “DCAN: Deep contour-aware networks for object instance segmentation from histology images,” Medical image analysis*,* 36, 135-146 (2017).

[66] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades." 3150-3158.

[67] Y. Li, H. Qi, J. Dai *et al.*, “Fully convolutional instance-aware semantic segmentation,” arXiv preprint arXiv:1611.07709, (2016).

[68] R. Mottaghi, X. Chen, X. Liu *et al.*, "The role of context for object detection and semantic segmentation in the wild." 891-898.

[69] N. Silberman, D. Hoiem, P. Kohli *et al.*, “Indoor segmentation and support inference from rgbd images,” Computer Vision–ECCV 2012, 746-760 (2012).

[70] C. Liu, J. Yuen, A. Torralba *et al.*, “Sift flow: Dense correspondence across different scenes,” Computer vision–ECCV 2008, 28-42 (2008).

[71] T.-Y. Lin, M. Maire, S. Belongie *et al.*, "Microsoft coco: Common objects in context." 740-755.

[72] K. Sirinukunwattana, J. P. Pluim, H. Chen *et al.*, “Gland segmentation in colon histology images: The glas challenge contest,” Medical image analysis*,* 35, 489-502 (2017).