ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

Deep Semantic Segmentation Models in Computer Vision

Paolo Andreini[1] and Giovanna Maria Dimitri[1]

DIISM - Universitá degli Studi di Siena
Via Roma 56, Siena, Italy

Abstract. Recently, deep learning models have had a huge impact on computer vision applications, in particular in semantic segmentation, in which many challenges are open.
As an example, the lack of large annotated datasets implies the need for new semi-supervised and unsupervised techniques. This problem is particularly relevant in the medical field due to privacy issues and high costs of image tagging by medical experts.
The aim of this tutorial overview paper is to provide a short overview of the recent results and advances regarding deep learning applications in computer vision particularly for what concerns semantic segmentation.

1 Introduction

Deep Learning (DL) has represented a real revolution in the computer science related research areas in recent years [1]. Several important applications have benefited from the use of DL, as for example the fields of bioinformatics [2, 3, 4, 5], forensic analysis [6, 7] and natural language processing [8, 9].
Computer visions tasks, which were so far considered almost impossible to be solved, were finally tackled using deep learning methodologies in a successful way. For instance successful applications of DL in computer vision can be found in the tasks of object detection [10, 11], image reconstruction [12] and image generation [13]. The success of DL, for such complicated computer vision tasks, relies in fact on the capability of performing the necessary machine learning approaches without the need of specifying the relevant input features to be used by the Deep learning architectures.
Among the various research tasks, which has been successfully addressed through the use of machine learning techniques, one important research challenge is represented by image segmentation. With such task we intend the need of labelling each pixel (or voxels in the case of 3D) of an image according to the relevant object/element of the image that is shown in that area.
More formally we can use the definition used in [14, 15] where image segmentation is defined as:

> *"An image segmentation is the partition of an image into a set of non overlapping regions whose union is the entire image. The purpose of segmentation is to decompose the image into parts that are meaningful with respect to a particular application" [14]*

Semantic segmentation is therefore an inherently multi-disciplinary tasks, which finds applications for example in the fields of biomedical image processing, image forensic or anomaly detection [15]. These tasks brings a set of inherently challenging research issues to be addressed. For instance there is the need of having the availability of big amount of labelled datasets, with which is possible to train the relevant segmentation network [16, 17].

Moreover there is the need on relying on computational capabilities, able to efficiently address the processing of large quantities of images and segmentation tasks, and a plethora of new efficient learning methods have been proposed in this direction [18].

This overview and tutorial paper is structured as follows. In Section 2 we revise briefly the main concepts concerning Deep Learning background and foundations. In Section 3 we report examples of application of Semantic Segmentation, in several fields. In Section 4 we summarize the main contributions of the special Session Deep Semantic Segmentation Models in Computer Vision. Furthermore in Section 5 we report conclusions and future perspectives of this field of research with future possibilities for deep learning applications in the semantic field.

2 Deep Learning

Deep learning is a branch of machine learning characterized by the use of models composed of multiple processing levels. At the heart of deep learning is the ability to learn multiple levels of data abstraction. These methodologies have greatly improved the state of the art in a wide variety of computing domains. While previous machine learning models were limited in their ability to process raw data, requiring complex, carefully designed feature extraction procedures, often based on domain–specific skills, deep learning models are representative learning methods. In fact, they allow you to learn multiple levels of representation directly from the raw data. Deep learning has demonstrated high abilities to discover complex structures in high–dimensional data that allow its application in many different fields. One of the most successful applications of deep learning is semantic segmentation.

3 Semantic Segmentation

Semantic segmentation can be defined as the task of predicting the semantic category of each pixel of an input image. The goal of semantic segmentation is to characterize the image by dividing it into multiple meaningful areas and can be a significant step in visual understanding. In fact, it plays a significant role in many different applications, from medical image analysis to robotics. In recent years, deep learning-based models have become the most common solution for semantic segmentation, often outperforming previous methods on popular benchmarks. In particular, Convolutional Neural Networks (CNNs) [19] are among the most successful and widely used architectures in deep learning for computer vision tasks and are widely used in semantic segmentation. However, by requiring pixel-level classification, semantic segmentation is generally a more challenging task than other popular computer vision tasks such as image classification. In fact, to obtain the best performance, a semantic segmentation model must determine not only the semantic category of the different elements in the scene but also their exact position. This makes it difficult to use typical CNNs in semantic segmentation, due to the use of down-sampling layers in the network. Indeed, common CNN architectures designed for image classification typically use multiple levels of down-sampling to rapidly broaden the receptive range, increase the level of abstraction, and reduce the computational requirements of the network [20],[21],[22]. However, the down-samplings progressively reduce the spatial resolution of the internal representation of the network (i.e. feature maps). To perform semantic segmentation, CNN models are then typically modified to produce a pixel-level prediction. In particular, they often assume an encoder-decoder structure, in which the encoder produces a spatially reduced representation of the input while the decoder retrieves the resolution to obtain an image-level prediction.

Fully convolutional networks [23] represent a milestone in this sense. The network employs a common CNN originally designed for image classification and adapt it for the semantic segmentation task. In particular the fully connected layers of the network are replaced with 1 by 1 convolutions allowing to use inputs of arbitrary size. In this way the CNN backbone can be used as an encoder where, however, the spatial resolution is considerably reduced.

To produce predictions at the image level, a decoder with up-sampling layers is employed. The decoder employs transposed convolutions to increase the resolution and skip connections to better incorporate fine details. By employing this structure he network parameters can be learned end to end.

After the FCN, many different segmentation networks have been proposed in which most of the conceptual design ideas are retained. As and example in [24] the segmentation model has a symmetrical structure. Similar to [23] the encoder is constituted by a typical CNN for image classification (i.e. a VGG [21]). Instead, the decoder uses unpooling to track the activations back to the image space and transposed convolutions to obtain dense feature maps. A fundamental contributions in the definition of segmentation models is given by the U-Net [25]. The network structure shows a typical U-shaped architecture which gives the network name. The encoder has a typical CNN structure and, for each downsampling stage in the encoder, the decoder has a corresponding up-sampling layer obtained through transposed convolutions. One of the main characteristics of the network is the particular type of skip connections that allow a flow of information from the encoder to the decoder. In particular, for each decoder level, feature maps of the encoder at the same resolution, are concatenated.

By using concatenation the number of feature maps in the encoder and decoder can be different giving a huge flexibility to the network design. The network initially applied on biomedical segmentation applications has proved very good performance in many different applications. One of the main characteristics of the previously discussed architectures is the loss of resolution in the encoder part of the network which leads to different decoding strategies to retrieve details make predictions at the image level.

To reduce the problem at its root, the deeplab [26], [27], [28] family of network architectures employs an additional strategy. In particular, some layers in the encoder are replaced with dilated convolutions. The idea of dilated convolutions, is to expand the field of view of convolutional filters by exploiting a dilation rate. The dilation rate expands the context that the filter takes into account but allows to keep the same number of parameters. The use of dilated convolutions allows to limit the reduction of the encoder spatial resolution avoiding the need for aggressive up-sampling in the decoder. Furthermore, the semantic segmentation task is made difficult by the need to analyze objects with different dimensions. Indeed, segmentation networks must collect and merge information at different scales. The most straightforward way to obtain multi-scale fusion is is to feed multi-resolution images separately across multiple networks and aggregate the output feature maps [29]. Instead a different approach consist in extracting features at different scales through a pyramid module (which can employ pooling, dylated convolutions or a combination of both) and fusing the so obtained representations. PSPNet [30], DeepLab [27], [28] and SMANet [31], for example, employ this approach.

A different solution is proposed in HR-Net [32], which maintains high resolution representations without the use of an encoder-decoder structure. The network connects high and low resolution convolution streams in parallel and repeatedly exchanges information between the different resolutions.

Finally, more recently, vision transformers are employed in place (or combined) with CNNs to perform semantic segmentation. A significant example is [33] in which vision transformers are used in place of convolutional networks as a backbone for the segmentation network. Tokens from various stages of the vision transformer are converted into image-like representations at different resolutions. The token are then progressively combined and transformed into an image level prediction using a convolutional decoder. In their work, [33] suggest that the dense vision transformer may allow for more accurate and globally consistent predictions than CNN-based networks.

Another aspect to consider when evaluating a segmentation network is the computational burden. This can be assessed by considering two factors: the speed of completion of the algorithm and the amount of computational memory required. Computational efficiency is a very important aspect of any algorithm that needs to be implemented on a real system. Often this goal is obtained by defining more efficient convolutional operations, changing the network architecture [34],[35].

A survey on efficient models capable of deployment on low-memory embedded systems while meeting the constraint of real-time inference is provided in [36].

The most popular approach to training semantic segmentation models is the use of full supervision where the ground truth is given by an image-level resolution segmentation map While the fully supervised methods enable state-of-the-art performance, annotating each training image pixel-by-pixel is very expensive and time-consuming. For this reason, to reduce the cost of annotation, other approaches have been proposed: the unsupervised approach uses unlabeled images, the semi-supervised approach uses a combination of labeled and unlabeled images, the weakly supervised approach uses spatially less informative annotations than the pixel-level. Of these approaches, weak supervision generally offers the best results. Weak supervision can consist of several types of annotations: bounding-boxes [37],[**?**], points [38] or image-level labels [37],[39]. Finally, many approaches to reducing the need for large annotated data sets rely on the same form of image augmentation. For example, some recent approaches employ generative adversarial networks to produce synthetic images and corresponding annotations, that can be used to enlarge existing datasets, from very small annotated training datasets [40], [41], [42].

4 Contributions from ESANN 2022

Contributions of the special session of ESANN 2022 on "Deep Semantic Segmentation Models in Computer Vision" covered several different applications of deep learning models for semantic segmentation.
In [43] the authors present a novel application of deep learning approaches to oocytes segmentation. The work is related to the biomedical field of Medical Assisted Procreation (MAP). The use of semantic segmentation deep learning models in this context can efficiently help the human operator responsible for assessing the healthiness of a oocyte to be fertilized and returned to the uterus. Other two interesting contributions showing applications of deep learning for the task of semantics segmentation in biomedical related fields, in particular related to the dermatology fields, are represented by [44, 45].
In [44] the authors presented a short survey and overview of the most used methodologies in this fields, together with the relevant datasets. Most recent models were compared.
In [45] instead, the author presented deep semantic model application to the case of skin lesions detection. The paper presents a novel application of convolutional neural networks based architecture to the case of skin lesions detection. The paper presents a weakly supervised approach in order to extract segmentation label maps of 43000 images from the reference ISIC database, which is used to train the deep learning architectures proposed.
A further interesting biomedical application presented in our special session is [46]. In [46] the authors presented a comparison of two well known deep learning segmentation architectures (the DeepLab and the MobileNet) in order to segment mice kidney images to detect glomeruli. The analysis of such biological structure is in fact fundamental to decide the transplantability of the organ in humans.
However there is a lack of availability of publicly available datasets. Therefore the presence of good performances in a non-human datasets, could help in tackling efficiently such task and could be the first step in order to efficiently compute transfer learning approaches to humans. Lastly in [47] address instead an interesting non–biomedical related research tasks. In particular detection and

localization of GAN manipulated images is performed in [47] using EfficientNet–B4 architectures.
The detection part is tested on several generated multi–spectral datasets from numerous world regions and several GAN architectures. Instead the localization test is performed on an inpainted images dataset, with promising results shown by the experiments performed.

5 Conclusions

Semantic segmentation with deep learning models represent a fundamental approach for the success of this research tasks.
Several are the challenging aspects to be addressed by deep learning semantic models: computational efficiency, presence of a sufficiently large training set, as well as collection of well labelled datasets on which is possible to training the experimental setup of the architecture proposed.
In this brief overview–tutorial paper we have revised the foundations of deep learning methods, of semantic segmentation applications as well as applications contributed in our special session. Future challenges of the field are numerous, as also the contributions to our special sessions can show.

References

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[2] Giorgia Giacomini, Caterina Graziani, Veronica Lachi, Pietro Bongini, Niccolò Pancino, Monica Bianchini, Davide Chiarugi, Angelo Valleriani, and Paolo Andreini. A neural networks approach for the analysis of reproducible ribo–seq profiles. *Algorithms*, 15(8):274, 2022.

[3] Seonwoo Min, Byunghan Lee, and Sungroh Yoon. Deep learning in bioinformatics. *Briefings in bioinformatics*, 18(5):851–869, 2017.

[4] Monica Bianchini and et al. Deep neural networks for structured data. In *Computational Intelligence for Pattern Recognition*, pages 29–51. Springer, 2018.

[5] V. Cicaloni and et al. Interactive alkaptonuria database: investigating clinical data to improve patient care in a rare disease. *The FASEB Journal*, 33(11):12696–12703, 2019.

[6] Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, and Cuong M Nguyen. Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding*, page 103525, 2022.

[7] Christian Galea and Reuben A Farrugia. Forensic face photo-sketch recognition using a deep learning-based architecture. *IEEE Signal Processing Letters*, 24(11):1586–1590, 2017.

[8] Daniel W Otter, Julian R Medina, and Jugal K Kalita. A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems*, 32(2):604–624, 2020.

[9] Li Deng and Yang Liu. *Deep learning in natural language processing*. Springer, 2018.

[10] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *arXiv preprint arXiv:1905.05055*, 2019.

[11] Rachel Huang, Jonathan Pedoeem, and Cuixian Chen. Yolo-lite: a real-time object detection algorithm optimized for non-gpu computers. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 2503–2510. IEEE, 2018.

[12] Samir Teniou and Mahmoud Meribout. A multimodal image reconstruction method using ultrasonic waves and electrical resistance tomography. *IEEE Transactions on Image Processing*, 24(11):3512–3521, 2015.

[13] Giovanna Maria Dimitri, Simeon Spasov, Andrea Duggento, Luca Passamonti, Nicola Toschi, and Pietro Lió. Unsupervised stratification in neuroimaging through deep latent embeddings. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1568–1571. IEEE, 2020.

[14] Robert M Haralick and Linda G Shapiro. *Computer and robot vision*, volume 1. Addison-wesley Reading, 1992.

[15] Saeid Asgari Taghanaki, Kumar Abhishek, Joseph Paul Cohen, Julien Cohen-Adad, and Ghassan Hamarneh. Deep semantic segmentation of natural and medical images: a review. *Artificial Intelligence Review*, 54(1):137–178, 2021.

[16] Hongshan Yu, Zhengeng Yang, Lei Tan, Yaonan Wang, Wei Sun, Mingui Sun, and Yandong Tang. Methods and datasets on semantic segmentation: A review. *Neurocomputing*, 304:82–103, 2018.

[17] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016.

[18] Marin Oršić and Siniša Šegvić. Efficient semantic segmentation with pyramidal fusion. *Pattern Recognition*, 110:107611, 2021.

[19] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[23] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[24] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[26] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[27] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[28] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[29] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 648–656, 2015.

[30] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[31] Simone Bonechi, Monica Bianchini, Franco Scarselli, and Paolo Andreini. Weak supervision for generating pixel–level annotations in scene text segmentation. *Pattern Recognition Letters*, 138:1–7, 2020.

[32] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5386–5395, 2020.

[33] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021.

[34] Vijay Badrinarayanan, Ankur Handa, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *arXiv preprint arXiv:1505.07293*, 2015.

[35] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[36] Christopher J Holder and Muhammad Shafique. On efficient real-time semantic segmentation: A survey. *arXiv preprint arXiv:2206.08605*, 2022.

[37] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1742–1750, 2015.

[38] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. Whatâs the point: Semantic segmentation with point supervision. In *European conference on computer vision*, pages 549–565. Springer, 2016.

[39] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1796–1804, 2015.

[40] Paolo Andreini, Giorgio Ciano, Simone Bonechi, Caterina Graziani, Veronica Lachi, Alessandro Mecocci, Andrea Sodi, Franco Scarselli, and Monica Bianchini. A two-stage gan for high-resolution retinal image generation and segmentation. *Electronics*, 11(1):60, 2021.

[41] Giorgio Ciano, Paolo Andreini, Tommaso Mazzierli, Monica Bianchini, and Franco Scarselli. A multi-stage gan for multi-organ chest x-ray image generation and segmentation. *Mathematics*, 9(22):2896, 2021.

[42] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10145–10155, June 2021.

[43] B.T. Corradini and et al. A deep learning approach for oocytes segmentation and analysis. In *ESANN 2022 proceedings European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. ESANN Ciaco, 2022.

[44] D. Cuza and et al. Deep semantic segmentation in skin detection. In *ESANN 2022 proceedings European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. ESANN Ciaco, 2022.

[45] Simone Bonechi. A weakly supervised approach to skin lesion segmentation. In *ESANN 2022 proceedings European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. ESANN Ciaco, 2022.

[46] Duccio Meconcelli, Simone Bonechi, and Giovanna Maria Dimitri. Deep learning approaches for mice glomeruli segmentation. In *ESANN 2022 proceedings European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. ESANN Ciaco, 2022.

[47] Lydia Abady, Giovanna Maria Dimitri, and Mauro Barni. Detection and localization of gan manipulated multi-spectral satellite images. In *ESANN 2022 proceedings European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. ESANN Ciaco, 2022.