Deep Learning Models for Image Segmentation

Chapter · July 2021

DOI: 10.1109/INDIACom51348.2021.00027

1 author:

Sanskruti Patel
Charotar University of Science and Technology

Proceedings of the 15th INDIACom; INDIACom-2021; IEEE Conference ID: 51348
2021 8th International Conference on "Computing for Sustainable Global Development", 17th - 19th March 2021
Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM), New Delhi (INDIA)

Deep Learning Models for Image Segmentation

Sanskruti Patel
Faculty of Computer Sceicne and Applications
Charotar University of Science and Technology (CHARUSAT)
Changa, India
sanskrutipatel.mca@charusat.ac.in

***Abstract*—Artificial Intelligence and deep learning models have evolved rapidly in the last decade and successfully applied to face recognition, autonomous driving, satellite imaging, robotics, and many more. Computer vision tasks often require adequate segmentation of an image that helps to understand the patterns and information. The adequate segmentation makes the analysis of each part of an image easier. Traditional segmentation techniques are often applied for image segmentation, but they are less efficient than deep learning techniques. Using deep learning approaches, it is possible to obtain hierarchical feature representations directly from the images, and hence, it eliminates the requirement of handcrafted features. This paper covers the fundamentals of image segmentation and deep learning, deep learning models for image segmentation, some successful implementations of deep learning models for image segmentation, and available open and benchmark datasets for image segmentation tasks.**

Keywords—Deep Learning, Image Segmentation, Semantic Segmentation, Instance Segmentation

I. INTRODUCTION

Image processing refers to a set of methods used to carry out specific operations on the image to extract some meaningful information from it. Digital image processing manipulates the images using a computer system. Computer vision is the field that focuses on gaining high-level understanding from digital images. Image segmentation is the subfield of digital image processing, and computer vision provides set of methods for splitting a digital image into various fragments [1]. This process helps to analyze the image in a more meaningful form. The objective of image segmentation is to localize boundaries and objects presented in an image. For that, a label is assigned to every pixel in an image where pixels with the same label share a definite set of characteristics.

Numerous methods are available in the literature that is successfully employed for image segmentation. Some of them are edge-detection, graph-cut, k-means clustering, histogram-based methods, thresholding, and watersheds. However, in recent years, Deep Learning (DL) methods show remarkable performance for image segmentation tasks with higher accuracy [2]. Deep learning is the field of machine learning algorithms that contains models that are inspired by human brains. It is one of the revolutionary advancements in machine learning and artificial intelligence [3]. Deep learning is successfully applied to many fields, including agriculture, healthcare, video surveillance, etc. [4][5]. Image segmentation is useful in many applications such as real time object detection, medical image analysis, video surveillance, augmented reality, etc.

This paper provides a comprehensive review of different deep learning methods available for image segmentation. Section II describes the background details of image segmentation tasks. Section III describes the various deep learning models available for image segmentation tasks. Section IV covers the details regarding open datasets publically available for image segmentation. Section V explains the various applications, and section VI provides the details about different performance evaluation metrics used to evaluate the deep learning model for image segmentation. Section VII covers the opportunities and challenges in the aforesaid area.

II. BACKGROUND DETAILS

In computer vision, image segmentation plays a pivotal role. It divides the image into multiple segments that represent objects. For that, image pixels are sorted to make larger segments. These distinct segments are containing each pixel with similar attributes. The segmentation process can be of two types; semantic segmentation and instance segmentation. In semantic segmentation, all the pixels of an image are classified into meaningful classes of objects. It does not differentiate distinct instances of the same object. Semantic segmentation defines the process of coupling each pixel of an image with a class label [6].

Instance, segmentation identifies the boundaries between different objects instead of only classifying them. For that, each instance of an object contained in an image is masked independently [7]. Instance segmentation derives from the object detection process where a bounding box is drawn to detect each object instance of an image with a label for the classification task. Instance segmentation performs object detection by adding a segmentation mask for each existed example in an image.

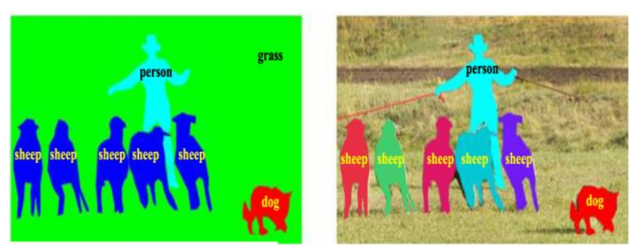

Fig. 1. Semantic Segmentation and Instance Segmentation [8]

Figure 1 distinguished the semantic segmentation and instance segmentation. There are five sheep presented in an image. Semantic segmentation classifies all the sheep as one instance as represented in figure 1 (a), whereas instance segmentation identifies each ship as represented in figure 1 (b).

III. Deep Learning Models for Image Segmentation

In artificial intelligence, machine learning comprises algorithms that make software applications more accurate while predicting outcomes. In the last decade, deep learning, a subfield of machine learning, has emerged as a useful computing paradigm that helps to solve many complex tasks of computer vision [9] [10] [11] [12]. It provides computational models that has multiple processing layers. These layers learn and represent the complex data. In deep learning, a deep neural network is used that has more number of hidden layers, and that makes a neural network "deep". These multiple layers are used for feature extraction, and each layer uses the output value of the previous layer as an Input value [10]. Advancements in deep learning methods with extensive processing and memory capacity make it very popular to solve image segmentation based applications. The established deep learning models used for image segmentation are explained as follows.

A. Fully Convolutional Network

A Fully Convolutional Network (FCN) is a convolutional neural network that transforms input image pixel to pixel categories. Segmentation is different from classification tasks as for each pixel, it requires predicting a class [13]. While classification focuses on what is in the input, segmentation adds where it is in the input image. The design of the FCN is illustrated in figure 2.

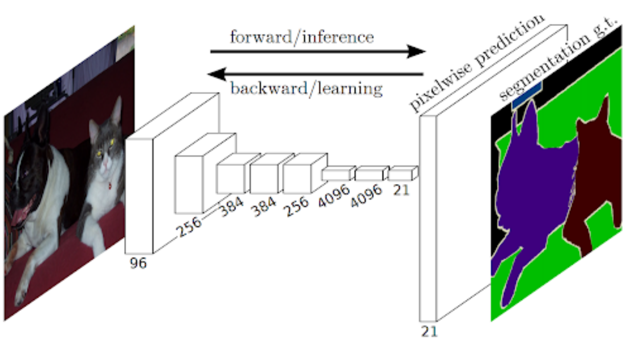

Fig. 2. Fully Convolutional Neural Network (FCN) [13]

In traditional image classification using convolutional neural networks, an input image is passed through set of convolutional layers and fully connected layers where the image size is reduced. In CNN, to obtain the deep features, it must go deeper; despite the spatial location information is also lost. To get the output size similar to input size, it is required to perform upsampling. In FCN, at last, the output is fused using element-wise addition that gives the enhanced result. For that, the intermediate layer feature map's height and width is converted back to the size of the input image using a transposed convolution layer [14].

B. Mask R-CNN

R-CNN (region-based convolutional neural network) family algorithms are widely used object detection models that fall under two-stage object detection models. It first identifies a subset of regions in an image that probably contains an object [13]. After that, it classifies the object pertains to each region. The Mask R-CNN is built on the top of the Faster R-CNN network. Therefore, along with the bounding box coordinates and class label to each object, it also provides the segmentation mask for each region containing an object [15]. Thus, analogues with the present branch for bounding box recognition, Mask R-CNN extend Faster R-CNN by adding an extension for predicting an object mask.

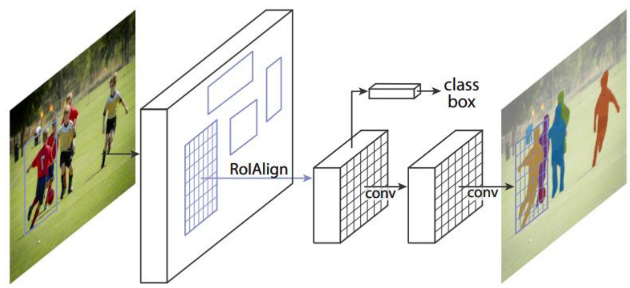

Fig. 3. Mask R-CNN for Instance Segmentation [15]

C. Encoder-Decoder Based Models

The encoder is basically an artificial neural network that takes input and produces a feature map as an output. The decoder is also a neural network with the same network structure as the encoder but in the opposite orientation. SegNet [16] developed for scene understanding applications. It consists of an encoder and decoder network with a pixel-wise classification layer. The decoder converts the low-resolution feature maps generated by the Encoder network into a full input resolution feature map. This feature map is used for pixel-wise classification. The architecture of the Encoder network of SegNet is similar to the VGG16. To make the architecture simpler, the fully connected layers of VGG16 are removed in SegNet. It also helps to reduce the number of parameters significantly. The architecture of SegNet is shown in figure 3.

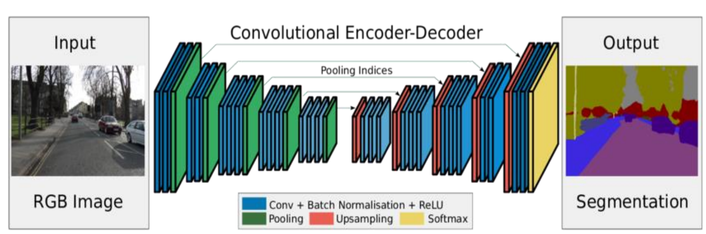

Fig. 4. An illustration of the SegNet architecture [16]

Another promising architecture based on the Encoder-Decoder framework is called U-Net [17], represented in figure 5. It was developed for biomedical image segmentation. U-Net architecture has three parts: 1) the contracting path or down-sampling path, 2) Bottleneck, and 3) The expanding path or up-sampling path. A contracting path with 3×3 convolutions captures the context from an image. The growing path increases the proportions of the feature maps while reducing the number of feature maps. Finally, the feature maps are processed using a 1×1 convolution to produce a segmentation map that classifies every pixel of the input image [17].

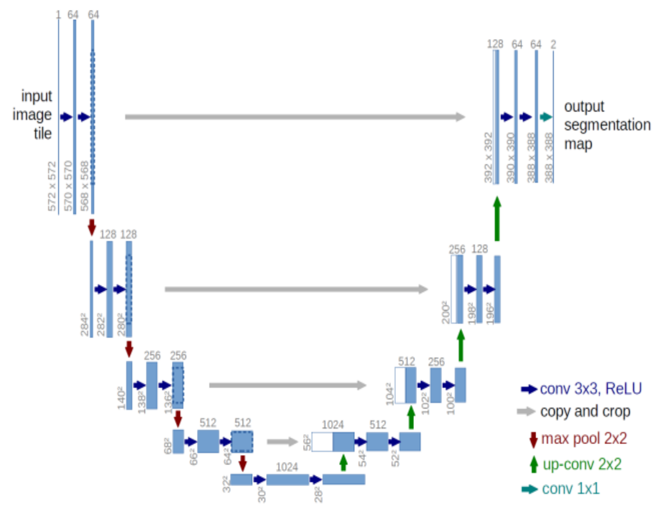

Fig. 5. U-Net: Convolutional Networks for Biomedical Image Segmentation [17]

D. DeepLab

Google designs DeepLab for performing semantic segmentation. DeepLab performs image segmentation by reducing the number of samples and data that the network processes [18]. The current version of DeepLab is DeepLabV3+, illustrated in figure 6.

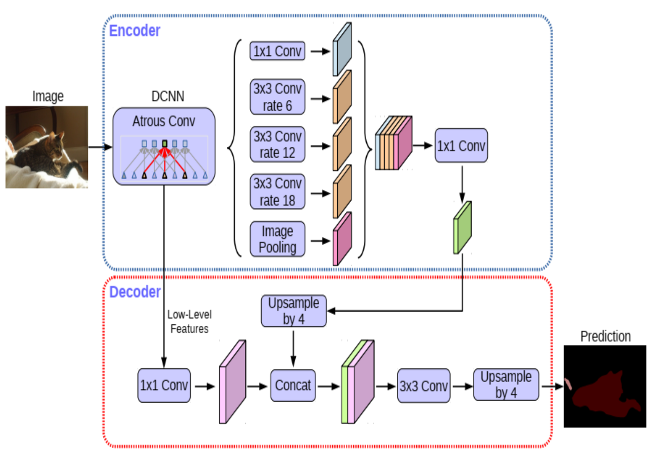

Fig. 6. DeepLabV3+ model [19]

It upsamples the output of the last convolutional layer with pixel-wise loss calculation. Instead of regular convolution operation, DeepLab uses atrus, called dilated, convolutions for upsampling. The latest DeepLabV3+ is extended with a useful decoder module that is used to improve the segmentation results, especially along object boundaries.

IV. Datasets for Image Segmentation

There are various open-source datasets available for performing image segmentation tasks. The datasets contain annotated images with more than one category.

Common Objects in Context (COCO) [20] dataset is available for segmentation, object detection, classification, and captioning. It contains around 330000 images. Among them, 200000 images are in annotated form. It includes images of 2,50,000 people with key points and provides five captions per image. The samples of annotated images in the COCO dataset is represented in figure 7.

Fig. 7. Samples of annotated images in the COCO dataset [20]

PASCAL VOC (Visual Object Classes) [19] dataset has 20 classes and more than 11,000 images. It mainly uses for classification, object detection, and segmentation task. It contains common images of daily life with annotations. It covers common objects like a bottle, aeroplane, bus, bicycle, person, boat, etc., that have a different scale, poses, occlusion, and lighting. The extended version of the PASCAL VOC dataset is PASCAL-Context [21] [22] dataset that adds additional annotations to make it more appropriate for segmentation tasks. It contains 400 labels and contains pixel-wise labels for each training image.

The different images of city scenes are contained by the Cityscapes [23][24] dataset, which is widely used for semantic segmentation as well as instance segmentation. It contains 30 classes and consists of Polygonal annotations. Images are captured from 50 cities spanning several months. Example of classes covered by the dataset is building, pole, vegetation, sky, road, parking, etc. Figure 8 shows the samples of images available in the Cityscapes dataset.

Fig. 8. Samples of annotated images in the Cityscapes dataset [23]

The YouTube-8M [25] dataset is a large-scale labeled dataset that provides 6.1 million video Ids that are categorized in 3862 classes. The video clips of the dataset are sampled uniformly from popular content on YouTube. It provides very high-quality machine-generated annotations that have a varied vocabulary of 3,800+ visual entities.

The Open Images dataset [26] is one of the largest datasets available with object location annotations. The latest version, V6, provides 15,851,536 boxes on 600 categories. Along with this, on 350 categories, it also provides 2,785,498 instance segmentations. It is a collection of 478,000 crowdsourced images that are categorized into 6,000+ categories. It is one of the widely used datasets for segmentation tasks. The sample of the Open Images dataset with object location annotations is represented in figure 9.

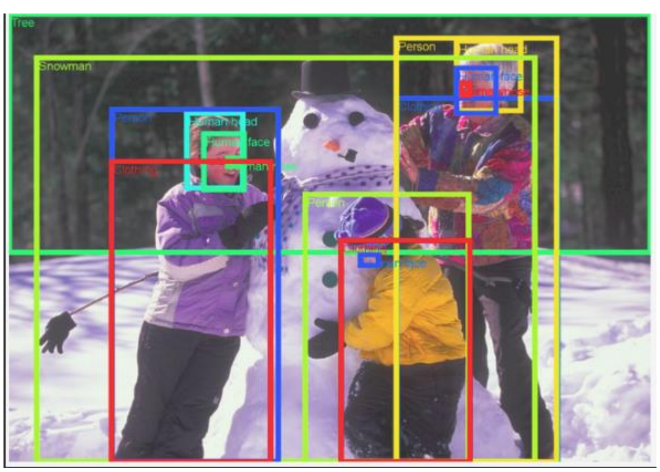

Fig. 9. The sample of Open Images dataset with object location annotations [26]

The Berkeley Segmentation [27] dataset is specifically made available for image segmentation and boundary detection. The dataset contains 12,000 hand-labeled segmentations of 1,000 Corel dataset images. These images are based on 30 human subjects.

V. IMAGE SEGMENTATION USING DEEP NETWORKS

Image segmentation is successfully applied in several areas, including face detection, medical imaging, autonomous driving, satellite image analysis, machine vision, etc. Several researchers have applied image detection models in various application areas as described in the following table 1.

TABLE I. IMAGE SEGMENTATION USING DEEP NETWORKS

Purpose	**Image Segmentation Method**	**Dataset Used**	**Performance**
Segmenting Satellite Images [28]	FCN with U-Net	SpaceNet	Precision - 0.911 Recall - 0.68 IOU - 0.644 F1 - 0.778
Brain Tumor Segmentation [29]	3D FCN	BraTS 2017 Challenge	0.710, 0.860, and 0.783 Dice Score respective to enhancing tumor, whole tumor, and tumor core
Biomedical Image Segmentation [30]	U-Net	ISBI Cell Tracking Challenge	IoU - 0.7756
Detection of Dark Spot [31]	SegNet	SAR Images of Oil Spill	Average FWIoU - 0.9724 Average MIoU - 0.9174
Segmentation of Air Tissue Boundary [32]	SegNet	rtMRI videos	Average Pixel Classification Accuracy - 99.3%
Blood Cell Images Segmentation [33]	SegNet	ALL-IDB1 database	Global Accuracy - 0.89447
Automated Marine Oil Spill Detection [34]	Mask R-CNN	SAR Images	Overall Accuracy Value - 96.6%
Precision livestock farming [35]	Mask R-CNN	Cattle Image Dataset	Mean Pixel Accuracy (MPA) - 0.92 (Cattle Segmentation Performance) Average Distance Error (ADE) - 33.56 pixel (Contour Extraction)

Purpose	Image Segmentation Method	Dataset Used	Performance
Airport Runway Semantic Segmentation [36]	DeepLabv3	Large and Medium Airport Runway	Accuracy - 96.64% Recall - 94.32%
Follicular Ultrasound Image Segmentation [37]	Improved Deeplabv3	The Follicle Ultrasound Image Dataset	Accuracy - 92.7

VI. Performance Evaluation Metrics

A variety of evaluation metrics can measure the effectiveness of any segmentation model. After training an image segmentation model, it outputs a prediction. To identify the model performance, it is required to be evaluated. The following are some of the frequently used evaluation metrics for image segmentation tasks.

A. Intersection over Union (IoU)

IoU calculates the performance using the intersection and union between the Prediction and Ground Truth values. IoU falls between 0 - 1 (0 - 100%), where 0 signifies no overlap and 1 signifies perfectly overlapping segmentation. It is defined as following figure 10.

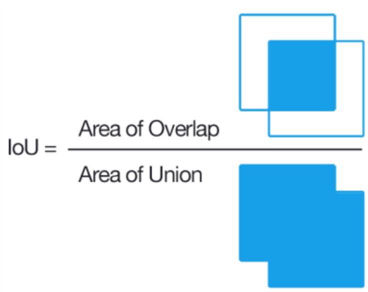

Fig. 10. Intersection over Union (IoU)

B. Pixel Accuracy

Pixel accuracy takes the ratio of correctly classified pixels with respect to total pixels. The following is the equation for defining pixel accuracy.

$$\text{Accuracy} = (TP+TN) / (TP+TN+FP+FN) \quad (1)$$

C. Precision

Precision is the ratio between True Positives and all the Positives. The following is the equation for defining precision.

$$\text{Precision} = TP/(TP + FP) \quad (2)$$

D. Recall

Recall defines as a measure that correctly identifies True Positives. The following is the equation for defining recall.

$$\text{Recall} = TP / (TP + FN) \quad (3)$$

E. F1 Score

The weighted average of Precision and Recall is known as F1 score. It is useful when there an uneven class distribution existed in the dataset. The following is the equation for defining the F1-Score.

$$F1 = 2 * ((\text{Precision} * \text{Recall})/(\text{Precision} + \text{Recall})) \quad (4)$$

F. Dice Coefficient

It is 2 * the Area of Overlap divided by the total number of pixels in both images. It is defined as following figure 11.

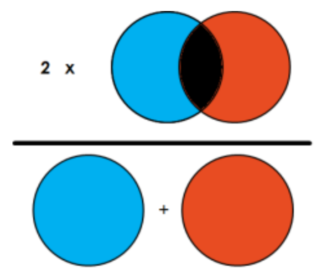

Fig. 12. Dice Coefficient

VII. Opportunities and Challenges

Deep learning methods have a great impact on computer vision applications, especially image segmentation tasks. Various researchers observe certain opportunities and challenges narrated as follows.

A. Limited Annotated Datasets

Object detection and image segmentation tasks required a huge set of annotated data. There are different open image and video datasets are available for conducting experiments. However, most of them are not in annotated form. Data annotation is a time-consuming and cumbersome task.

Also, it is required to take help from subject experts. Incomplete and inaccurate annotation significantly affects the segmentation accuracy and model performance.

B. Uneven Class Distribution

The uneven class distribution in dataset is one of the other challenges for image segmentation. Most of the training datasets do not have an exactly equal number of samples in each class; however, a small difference often does not matter. Nevertheless, severe imbalance is challenging for the model. Imbalanced distribution certainly creates a challenge for predictive modeling and has an adverse effect on results.

C. Interpretable Deep Learning Models

Despite of promising results provided by deep learning models, they work like a black-box. It is complex to understand how deep learning models learned features [38]. All deep learning models are based on deep neural network architecture that contains complex neuron structure. Also, they are considered as non-transparent models, and their predictions are not traceable by humans. A good understanding of the working of these models may lead to better development of image segmentation models.

VIII. CONCLUSION

Artificial Intelligence and deep learning models are successfully applied on many computer vision tasks. The paper presents a survey conducted for applying deep learning models for image segmentation tasks. It provides the architecture of FCN, Mask R-CNN, U-Net, SegNet and DeepLab that are successfully applied for various image segmentation tasks and provides higher accuracy and good segmentation results. The paper also provides the details of different open datasets available for experimenting with image segmentation tasks. the various performance metrics used to evaluate the model is also discussed. Moreover, it presents the research work carried out for applying deep learning models for image segmentation tasks.

REFERENCES

[1] R. Szeliski, *Computer Vision: Algorithms and Applications*. London, England: Springer, 2010.

[2] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Comput. Intell. Neurosci.,* vol. 2018, pp. 1–13, 2018.

[3] S. Patel, "A comprehensive analysis of Convolutional Neural Network models," *Int. j. adv. sci. technol*., vol. 29, no. 04, pp. 771–777, 2020.

[4] N. Ganatra and A. Patel, "A comprehensive study of deep learning architectures, applications and tools," *Int. j. comput. sci. eng*., vol. 6, no. 12, pp. 701–705, 2018.

[5] D. Chhachhiya, A. Sharma, and M. Gupta, "Designing optimal architecture of recurrent neural network (LSTM) with particle swarm optimization technique specifically for educational dataset," *Int. J. Inf. Technol.,* vol. 11, no. 1, pp. 159–163, 2018.

[6] Y. Wei et al., "Learning to segment with image-level annotations," *Pattern Recognit*., vol. 59, pp. 234–244, 2016.

[7] Z. Wu, C. Shen, and A. van den Hengel, "Bridging category-level and instance-level semantic image segmentation," *arXiv [cs.CV*], 2016.

[8] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Computer Vision – ECCV 2014*, Cham: Springer International Publishing, 2014, pp. 740–755.

[9] A. Bhandare, M.V. Bhide, P. Gokhale, R. Chandavarkar, "Applications of Convolutional Neural Network", *International Journal of Computer Science and Information Technologies*, vol. 7, no. 5, pp. 2206-2215, 2016.

[10] T. Kalaiselvi, S. T. Padmapriya, P. Sriramakrishnan, and K. Somasundaram, "Deriving tumor detection models using convolutional neural networks from MRI of human brain scans," *Int. J. Inf. Technol*., vol. 12, no. 2, pp. 403–408, 2020

[11] S. Patel, "Diabetic Retinopathy Detection and Classification using Pre-trained Convolutional Neural Networks", *International Journal on Emerging Technologies*, vol. 11, no. 3, pp. 1082–1087, 2020

[12] S. Patel and A. Patel, "Deep leaning architectures and its applications A survey," *Int. j. comput. sci. eng*., vol. 6, no. 6, pp. 1177–1183, 2018.

[13] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in 2015 IEEE *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[14] "13.11. Fully convolutional networks (FCN) — dive into deep learning 0.15.1 documentation," *D2l.ai*. [Online]. Available: https://d2l.ai/chapter_computer-vision/fcn.html. [Accessed: 22-Nov-2020].

[15] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," IEEE *Trans. Pattern Anal. Mach. Intell*., vol. 42, no. 2, pp. 386–397, 2020.

[16] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," IEEE *Trans. Pattern Anal. Mach. Intell*., vol. 39, no. 12, pp. 2481–2495, 2017.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *arXiv [cs.CV*], 2015.

[18] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *arXiv [cs.CV],* 2016.

[19] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Computer Vision – ECCV 2018*, Cham: Springer International Publishing, 2018, pp. 833–851.

[20] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," *arXiv [cs.CV]*, 2014.

[21] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[22] R. Mottaghi et al., "The role of context for object detection and semantic segmentation in the wild," in 2014 IEEE *Conference on Computer Vision and Pattern Recognition*, 2014.

[23] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in 2016 IEEE *Conference on Computer Vision and Pattern Recognition (CVPR),* 2016.

[24] M. Cordts et al., "The Cityscapes Dataset," in CVPR Workshop on The Future of Datasets in Vision, 2015.

[25] S. Abu-El-Haija et al., "YouTube-8M: A large-scale video classification benchmark," *arXiv [cs.CV]*, 2016.

[26] A. Kuznetsova et al., "The open images dataset V4: Unified image classification, object detection, and visual relationship detection at scale," *Int. J. Comput. Vis*., vol. 128, no. 7, pp. 1956–1981, 2020.

[27] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings Eighth IEEE International Conference on Computer Vision.* ICCV 2001, 2002.

[28] M. Y. Saifi, J. Singla, and Nikita, "Deep learning based framework for semantic segmentation of satellite images," in 2020 *Fourth International Conference on Computing Methodologies and Communication (ICCMC),* 2020.

[29] A. Jesson and T. Arbel, "Brain tumor segmentation using a 3D FCN with multi-scale loss," in Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Cham: Springer International Publishing, 2018, pp. 392–402.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Lecture Notes in Computer Science,* Cham: Springer International Publishing, 2015, pp. 234–241.

[31] H. Guo, G. Wei, and J. An, "Dark spot detection in SAR images of oil spill using Segnet," *Appl. Sci. (Basel),* vol. 8, no. 12, p. 2670, 2018.

[32] C. A. Valliappan, A. Kumar, R. Mannem, G. R. Karthik, and P. K. Ghosh, "An improved air tissue boundary segmentation technique for real time magnetic resonance imaging video using segnet," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[33] T. Tran, O.-H. Kwon, K.-R. Kwon, S.-H. Lee, and K.-W. Kang, "Blood cell images segmentation using deep learning semantic segmentation," in 2018 IEEE *International Conference on Electronics and Communication Engineering (ICECE),* 2018.

[34] S. Temitope Yekeen, A. Balogun, and K. B. Wan Yusof, "A novel deep learning instance segmentation model for automated marine oil spill detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 167, pp. 190–200, 2020.

[35] Y. Qiao, M. Truman, and S. Sukkarieh, "Cattle segmentation and contour extraction based on Mask R-CNN for precision livestock farming," *Comput. Electron. Agric.,* vol. 165, no. 104958, p. 104958, 2019.

[36] Z. C. Men, J. Jiang, X. Guo, L. J. Chen, and D. S. Liu, "Airport runway semantic segmentation based on dcnn in high spatial resolution remote sensing images," *ISPRS - Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. XLII-3/W10, pp. 361–366, 2020.

[37] T. Zeng and J. Liu, "Follicular ultrasound image segmentation based on improved Deeplabv3," in 2019 *3rd International Conference on Computer Engineering, Information Science and Internet Technology*, 2019.

[38] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *arXiv [cs.CV]*, 2020.