Journal of Information Technology Research
Volume 15 • Issue 1

Semantic Segmentation:
A Systematic Analysis From State-of-the-Art Techniques to Advance Deep Networks

Aakanksha, Amity University, Noida, India
Arushi Seth, Amity University, Noida, India
Shanu Sharma, ABES Engineering College, India*
https://orcid.org/0000-0003-0384-7832

ABSTRACT

Semantic segmentation was traditionally performed using primitive methods; however, in recent times, a significant growth in the advancement of deep learning techniques for the same is observed. In this paper, an extensive study and review of the existing deep learning (DL)-based techniques used for the purpose of semantic segmentation is carried out along with a summary of the datasets and evaluation metrics used for the same. The paper begins with a general and broader focus on semantic segmentation as a problem and further narrows its focus on existing DL-based approaches for this task. In addition to this, a summary of the traditional methods used for semantic segmentation is also presented towards the beginning. Since the problem of scene understanding is being vastly explored in the computer vision community, especially with the help of semantic segmentation, the authors believe that this paper will benefit active researchers in reviewing and studying the existing state-of-the-art as well as advanced methods for the same.

KEYWORDS

Annotation, CNN, Deep Learning, Deep Networks, Image, Scene, Segmentation, Semantic

INTRODUCTION

From the past few decades, tremendous growth can be seen in the computer vision community. Researchers have provided optimal solutions in different vision based tasks like image classification, object detection, object labelling, saliency estimation, image compression and many more (Lu, 2007; Verschae, 2015; Messer, 2017). Almost all the vision-based applications include a basic step of segmenting an image into meaningful regions, which is basically the process of linking each pixel in an image with a class label. Although many optimal solutions have been provided till date for segmenting an image (Guo, 2018; Zaitoun, 2015), due to unpredictable real world situations and dependency of a majority of vision applications on this step, segmentation of image is still an open research problem for the computer vision community.

In this study, our focus is on performing the semantic segmentation for scene understanding. Semantic segmentation is a process of assigning a meaningful label to each pixel based on the context of the environment (Lateef, 2019). It is a very useful step for a variety of computer vision applications

DOI: 10.4018/JITR.299388 *Corresponding Author

where it is important to understand the context of the operating environment in which the systems are operating, for example in robotics (Kim, 2018), self-driving cars (Kaymak, 2019) etc. Further, scene understanding is a computer vision application which includes analysis and perception of an image of the scene to create an overview of the event depicted in the scene (Xiao, 2013). A scene shows a real-world situation which is extracted from the environment. It includes multiple objects which are interacting with each other, thereby having some meaning. A scene can represent a variety of real-world events ranging from personal events to public events. Scene understanding is the process of interpreting scenes, which are captured through devices like cameras, microphones, contact sensors etc. to get an in depth understanding of it (Aarthi, 2017). The data of a scene can be expressed using various features like color, texture, and light intensity, thus, the process of creating a good understanding of a scene includes proper extraction of features from an image of a scene that characterizes it efficiently. It is based on the idea of vision and cognition, in which the functionality of detection, localization, recognition and understanding is performed first, followed by cognition, which is used to add functionalities like learning, adaption, finding alternatives, interpretation and analysis. The models that perform scene understanding include the capability to analyze events and modify accordingly. It can adapt to unforeseen data and perform robustly in such situations (Li, 2009). Scene understanding is included in machines to make them capable of interpreting events in a similar manner as humans do.

Scene understanding has applications in various fields. In the medical field, it is used for medical image analysis which includes getting clinically meaningful information from the image. (Ker, 2018) presents a review on the use of machine learning algorithms for the analysis of medical images, particularly CNN. Here, they show that the extracted data from images can be used by doctors for diagnosis. Along with in depth study of application of classification, localization, detection, segmentation and registration on medical images, challenges and future applications are also discussed. Scene understanding also has its application in road detection and urban scene understanding (Brust, 2015), in which objects present in images are classified and labelled, which is then used for detecting roads and understanding the urban scenes in the datasets. Scene understanding is also used in robotics to improve navigation in robots. Since the process of scene understanding is based on a general formulation, it finds its use in a plethora of applications.

In recent years, deep networks are very popular among computer vision researchers (Srinivas, 2016). Researchers are implementing deep networks models for every possible field like image classification (Lee, 2018), object detection (Verschae, 2015), image generation etc. (Kolberg, 2018). Deep learning allows us to model the high-level features of an image into compact representations, which allows for efficient manipulations as well as analysis of the input images (Garcia-Garcia, 2017). As scene understanding is a complex problem involving several sub-tasks such as object detection, semantic segmentation etc., deep learning models can efficiently handle the tasks.

In this paper, a systematic study on semantic segmentation is presented. The paper is presented in a way to provide in-depth knowledge about the semantic segmentation to the readers. The work proposed here is targeted to the researchers interested in the field of scene understanding. Here, various traditional approaches, state of the art models, and recently developed deep learning based models are discussed. Recent work done in the past five years with focus on semantic segmentation for scene understanding is considered for analyzing the various deep learning models for semantic segmentation. Further, different benchmark datasets along with evaluation metrics are also presented.

The work done in this paper is presented for providing following key contributions:

- We lay out an extensive study of the traditional as well as deep learning-based techniques employed for the task of semantic segmentation.
- An in-depth and systematic review of the related work for semantic segmentation using deep learning, with a special focus on their contributions is presented.
- Analysis of several datasets pertinent to and useful for semantic segmentation is discussed.
- Specifications of a few metrics valuable for evaluating the performance of different techniques/ models is presented.

Motivated by the need of an extensive review in the field of semantic segmentation for scene understanding, further sections of this paper are organized as follows: Section II gives a brief overview about the background concepts like segmentation, semantic segmentation and various traditional and advanced approaches for performing it. Section III is focused on giving the brief overview on deep learning and various deep networks extensively used for semantic segmentation. Further, various benchmark datasets along with their comparative summary and different evaluation metrics to test the developed models are described in Section IV. In Section V, review of some of the recent work done in deep learning based segmentation is presented. Further, major key findings are discussed in section VI and the paper is concluded in Section VII.

BACKGROUND: FROM SEGMENTATION TO SEMANTIC SEGMENTATION

Image Segmentation

For analysis of an image, image segmentation is fairly a popular step in the domain of digital image processing and computer vision. The aim of carrying out the process of segmentation is to represent the image in a simpler manner which is more abstract and meaningful thereby, making it easier to examine. It refers to the process of splitting a digital image into several distinct sections, i.e., collections of pixels, which further collectively form the objects in the image and hence, share similarities. Segmentation is usually performed in order to identify and find objects and boundaries in digital images (Ripon, 2017). Thus, it is concluded that image segmentation is the process of attributing a label to each pixel in an image in such a manner that those which are assigned the same labels hold certain characteristics like texture, color or intensity etc. in common can be easily drawn. Segmentation acts as a reliable transformation technique that determines the success of analysing an image. However, it is a challenging task to obtain a precise partitioning of an image.

(Zaitoun, 2015) presents an intensive comparative study of image segmentation techniques, broadly divided into layer-based segmentation and block-based segmentation. The techniques of block-based segmentation, further divided into region based and edge based, are explained in detail. Some of the popular traditional approaches for performing segmentation as well as their applications are briefly discussed below:

- **Thresholding:** It is a process of segmenting an image by setting a threshold value and comparing all the image pixels with the threshold value. This method segments the object from the background, by setting all the pixels having value less than the threshold to one value (may be white) and all the pixels having value greater than threshold to another value (may be black). This method gives best results when there is high contrast. As in thresholding based approach, setting a proper threshold value is the most important step, a lot of work has been done for automatically extracting the optimum threshold value. Two scheme automatic threshold selection based on approximation of histogram is presented in (Ramesh, 1995), in which one method determines the threshold by minimizing the sum of the square error while the other method minimizes the variance of the histogram. The algorithm proposed in the paper (Al-azawi, 2013), overcomes the drawbacks of taking the threshold value as the global minimum of the histogram. This is done by using membership functions for measurement of bright and dark area, which defines each pixel in a region in terms of its membership value. Till now, researchers have explored thresholding in various fields. In (Maolood, 2018), approach was used for detecting cancer by segmenting the images using combination of fuzzy entropy and thresholding on medical images. The approach can be used to detect cancer in ultrasound results, MRI and dermoscopy. The accuracy of the method is high but computation cost is not considered by the authors. Researchers have also tried to combine thresholding with other image processing methods. For example in (Al-azawi, 2013), the writers combined fuzzy based image processing with histogram thresholding technique for image segmentation.

- **Edge based segmentation:** Another very common approach for segmentation is edge based segmentation. Edges are basically discontinuity in the pixel values, which are identified from the difference in pixel values in two adjacent regions. This discontinuity helps in identifying the shapes of objects in the image. By using filters and convolutions on the image matrix edges can be identified. Some of the common filters used for edge detection are Sobel operator and robert cross operator, as discussed in (Karthicsonia, 2019) by working on medical images. (Karthicsonia, 2019) discusses various edge based segmentation methods, that is Robert cross method, Prewitt method, Sobel method, Laplacian of gaussian method and Canny method. These methods are defined along with their use in brief in the paper, though the paper does not cover the mathematical concept of these methods. Attempts have also been done to understand the interaction of image segmentation, using some edge detectors, and object recognition (Ramadevi, 2010). In (Ramadevi, 2010), the masks used by different edge detection operators are also mentioned. Along with edge detector algorithms, different algorithms, including EM algorithm, OSTU algorithm and Genetic algorithm, are also explored to explain the interaction between image segmentation and object recognition. The operators and algorithms are applied on an image to understand the segmentation effect. Edge detection method is also used in identifying abnormalities in the images, especially medical images. In (Padmapriya, 2012), a new method is proposed to identify the thickness of the bladder wall by applying automatic edge based image segmentation. The method projected is used to collect information about bladder abnormalities and the extent of abnormalities. The proposed method is explained by the authors in detail, which is supported with promising results.
- **Region based segmentation:** Region based segmentation is an approach that extracts region based features from the image and uses those features to define different classes. This method is very useful in noisy images where edges cannot be identified (Lahouaoui, 2013). Two famous approaches for region based segmentation are splitting & merging and region growing. In former approach a uniformity criteria is selected which decides if two regions need to split or merge. Initially splitting is done by dividing an image into sub parts until the splitting does not make any difference followed by merging of adjacent regions based on the same uniformity criteria. Region growing method starts by defining a seed region which can be a single pixel or a block of pixels. The neighbours of the seed region are then checked with the uniformity criteria for merging. When the criterion is not met then the region is extracted and another seed is selected to merge another region. An extensive review on various region based segmentation methods can be found in (Lahouaoui, 2013). A comparative study of different region-based segmentation techniques is presented in the paper. Eight methods are evaluated on four criteria, by applying them on synthetic MR image and real data. Region based segmentation is often used for identifying tumor, veins etc. in medical images, for finding targets in aerial images and finding people in surveillance images etc. In (Gould, 2009) authors presented a region based approach that combines object detection and segmentation which performs background classification based on pixel features and object detection using representation of regions. The model defined in the paper gives a unified description of the scene depicted in the image. This is done as the model explains every pixel in the image.

Semantic Segmentation

It is the process of assigning a meaningful label to every pixel in the image. It is different from the normal segmentation process, as in semantic segmentation a single label is assigned to multiple objects of the same class. To justify their significance for image analysis and evaluation, the regions should be markedly related to the present objects in the image or the features of interest (Lateef, 2019; Kim, 2018). Meaningful segmentation allows the progression from low-level or crude image processing transformations involving conversions of grayscale or color image into several other images to high-level image description creation with respect to features, objects, layouts and scenes (Liu, 2019;

Gupta, 2015). Segmentation techniques can be further classified as contextual or non-contextual. Contextual techniques make great use of spatial relationships that exist between the features of an image. Whereas, non-contextual techniques do not consider any such relationships and rather categorize features based on attributes such as grey-level or color. For example, clustering those pixels together which have related grey levels and are spatially close.

Traditionally, features and classification methods were used by researchers to perform semantic segmentation. Extraction of various features was popularly done for segmentation. Various supervised and unsupervised classifiers, like support vector machine and K-mean clustering respectively, were also used to perform segmentation (Liu, 2019). While many modern researchers are focusing their work towards deep neural networks, some modern researchers are also combining the traditional methods with new concepts (Jianxiong, 2012; Guo, 2016). The researchers are trying to improve the accuracy by enhancing traditional methods with different concepts like fuzzy logic (Guo, 2016). Some of the popular traditional methods for performing semantic segmentation are discussed below:

- **Features and Classification based segmentation:** Features play an important role for analysis of an image and for performing meaningful segmentation on an image. Till now a range of features have been explored by researchers which includes color, texture, Histogram of oriented gradients (HoG), scale-invariant feature transforms, SURF and many more (Lateef, 2019; Kim, 2018). Image segmentation based on features referred to as visual descriptors can be seen in (Ripon, 2017). The extracted features are used for generating classification models to provide meaningful segmentation. Based on the adopted classification techniques, the segmentation approach can be classified into supervised and unsupervised segmentation approaches.

One of the popular techniques for performing unsupervised classification is K-means clustering. Clustering helps segment the objects in an image by dividing the pixels into various clusters. It starts by randomly choosing the number of clusters that is the value of k, then the pixels are randomly allocated to these clusters. The centers of these clusters are then calculated and distance of each pixel form these centers is also calculated. This process is widely used with small datasets (Shan, 2018). Use of K-means for image segmentation is presented in (Shan, 2018). The author proposed a method that uses K-means for image segmentation and gray-gradient maximum entropy for feature extraction. Results and comparison with other algorithms given in paper shows that K-means is an efficient approach for performing segmentation on an image. Similarly, a color-based segmentation method using K-means clustering is proposed in (Muthukannan, 2010), where pixels are first divided into clusters using color and spatial features and then a specific number of clusters are merged to make a region. This approach can be used for image retrieval which would generate reliable images for locating tumors, fingerprint recognition, locating objects from satellite images etc. Further, various supervised classification techniques were also explored in literature for semantic segmentation (Sharma, 2018; Savkarea, 2012; Sakthivel, 2014; Wang, 2011). Textural features based medical image segmentation is done in (Sharma,2008). The authors present an approach for auto-segmentation and tissue characterization, which includes extracting features which are used in an ANN based classifier to classify/analyse the soft tissues. Use of SVM pixel classifier for image retrieval, object detection and medical imaging is given in (Savkarea, 2012), where SVM classifier is used to classify malaria infected erythrocytes, which further helps in detection of parasite life stages. Further, the problem of object detection and semantic segmentation for indoor scene understanding have been also explored in (Gupta, 2015), where features based on shape, size, geocentric pose and appearance are extracted for segmentation. These features are then classified using random decision tree forest and support vector machine based classifiers. Combination of different approaches was explored by researchers, in (Sakthivel, 2014), Support Vector Machine and fuzzy C-means are combined to perform color image segmentation. The features extracted are given as input to the SVM classifier which is trained using

fuzzy C-means. Here the advantages of SVM classifier and pixel level information are combined to return better results, thus improving the quality of image segmentation.

- **Markov Random Network (MRF) and Conditional Random fields (CRF) based segmentation:** Conditional Random fields is another method used for segmentation. This method is used where the contextual information affects the prediction. This method helps to work on data where the label classes are dependent on each other. For example the class label for a pixel depends on the label of its neighbour pixels also. In this method the classifier predicts value y for pixel x by considering its features and labels of all the pixels x is dependent on (Lafferty, 2001). In (He, 2008) author presents conditional random fields framework. It explains the framework along with its comparison with Hidden Markov models and maximum entropy Markov models. In (Lafferty, 2001), conditional random fields are used along with other frameworks for biological entity recognition (BER). The paper presents an approach for extracting features and then modeling and predicting the BER. In (Verbeek, 2007) authors used CRF for scene segmentation, where CRFs partition an image into semantic-level regions and assign the class labels to these regions. In the paper, the model combines the local features and the features associated over a larger section for semantic image labelling.

DEEP LEARNING FOR SEMANTIC SEGMENTATION

Deep Learning

Deep learning is a subset of a broader family of machine learning algorithms based on artificial neural networks, also commonly referred to as ANNs, and representation learning as a multilayered representation of the input data is constructed through the network (Guo, 2016; Goodfellow, 2016). Deep learning methods can be broadly divided into two categories - Supervised and Unsupervised. Supervised methods work around a loss function, which is defined based on the problem at hand, by updating the model parameters based on the values of the loss function. Unsupervised methods usually define a loss function based on the reconstruction ability of the model. The goal is to minimize the value of the loss function (Srinivas, 2016). Commonly used types of deep neural networks are as follows - Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Autoencoders (AE) and Generative Adversarial Networks (GANs) among many others. Whilst CNNs are generally used for computer vision problems, RNNs have found great use in the field of natural language processing (NLP) in which Long Short Term Memory (LSTM) networks and Gated Recurrent Unit (GRU) networks have had significant success (Goodfellow, 2016). Autoencoders are a class of ANNs which are used to learn data coding in an unsupervised fashion. GANs also follow an unsupervised learning method wherein there are two neural networks improving each other's performance by contesting with each other.

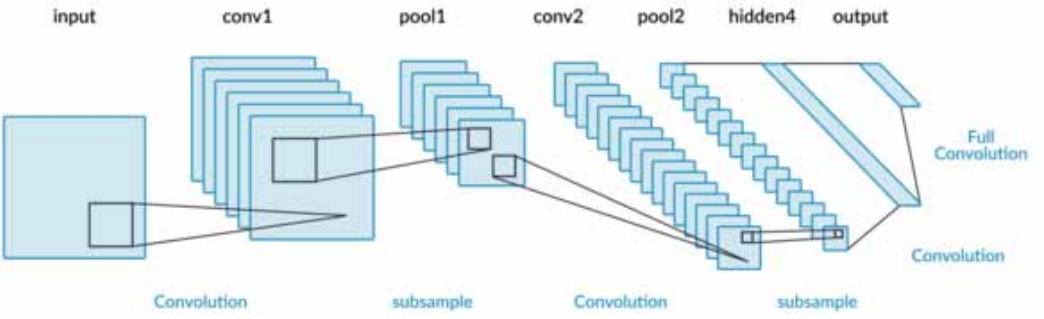
In this section, we will present a summary of the methods pertinent to and useful for semantic segmentation. These methods are heavily based on convolutional neural networks (CNNs), which are explained in more detail in the next subsection. For a detailed understanding of deep learning, the reader is referred to (Goodfellow, 2016) and for a general overview of deep learning used for computer vision refer to (Guo, 2016).

Deep Networks for Semantic Segmentation

Convolutional Neural Networks (CNNs): Convolutional neural networks form a category of deep neural networks usually applied to visual image tasks. The architecture of a CNN typically consists of multiple convolutional layers, pooling layers and activation functions, preferably non-linear. As the name suggests, these networks employ a mathematical operation called convolution, which is a specialized linear matrix operation. These networks differ from multilayered perceptrons in that they

use convolution instead of the general matrix multiplication in at least one of their layers. An example of this architecture is shown in Figure 1. These networks were introduced by LeCun et al. (Yann, 1990) in 1990 for recognition of handwritten digits. However, they seemed to gain popularity after the introduction of AlexNet by Krizhevsky et al. (Krizhevsky, 2017), after their efficient performance and win in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012. In the present day, several variations of convolutional networks are being employed for semantic segmentation, some of which are discussed in detail the following subsection.

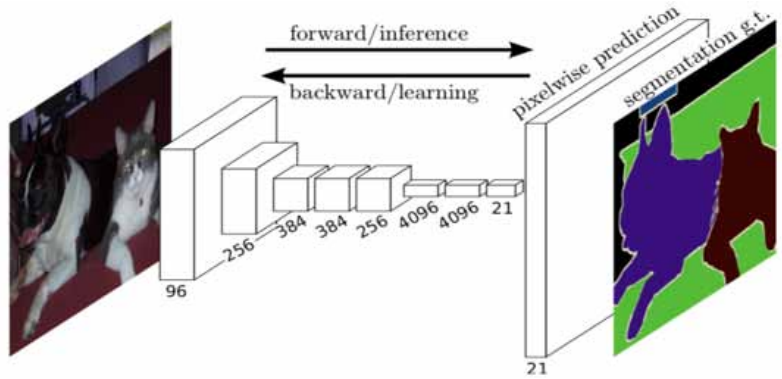
Figure 1. Basic architecture of a convolutional neural network, consisting of two convolution layers, two pooling layers and two fully connected layers (Krizhevsky, 2017)



Fully Convolutional Network (FCN)

FCNs were developed by Jonathan Long et al. in 2015 (Long, 2015). In this network, some convolutional layers were exclusively incorporated for performing semantic segmentation. The network was designed such that when an image of random size was fed to the FCN, a semantically segmented image of the same size was generated as a result as shown in Figure 2. The initial steps in developing this model consisted of modifying popular architectures such as LeNet, AlexNet, VGG16 to have the scope for an arbitrarily sized input whilst substituting the entire set of fully connected layers with convolutional layering. As the network builds multiple feature mappings from relatively small sizes and compacted representations, it is important to perform upsampling in order to produce a similar sized image as the input. Upsampling involves convolutions with strides less than one. It is sometimes known as deconvolution as it results in input having smaller size than output. Using this method, the

Figure 2. Architecture of the Fully Convolutional Network

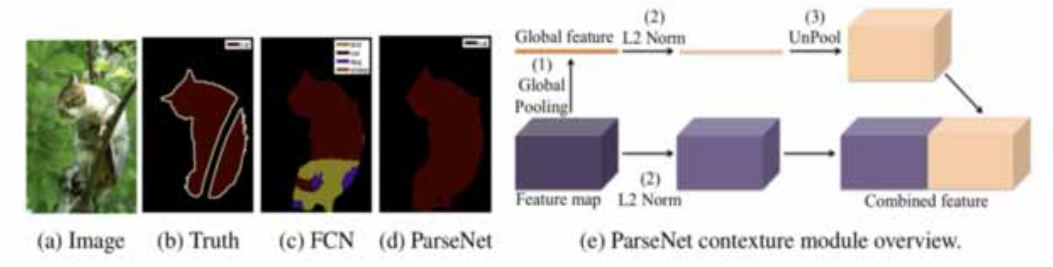


network is then trained using the concept of a pixel loss. Additionally, several skip connections were introduced in this network to connect high-level feature mapped representation to highly precise ones at the top of the model. However, FCN models do not consider the global context and information of the input images. Application of FCN can be seen in (Kaymak, 2019), where semantic segmentation is done to support autonomous driving vehicles. The experimental study done on SYNTHIA dataset using FCN architecture is a contribution towards researchers working on autonomous driving.

ParseNet

ParseNet was created as an improvement to the Fully Convolutional Network model proposed above by Liu et al. (Liu, 2015). Since the FCN model does not consider the global context of the image as it goes further into the deeper layers by focusing only on details in the produced feature mappings, ParseNet attempted to address this issue. ParseNet presents an exclusively convolutional network which predicts values for each pixel simultaneously and does not take regions as inputs in order to preserve the global context and information of the image. A module is used to take feature mappings as the input. The initial course of action makes use of a model to produce feature mappings that are further condensed to just one globally accessible feature vector with a single pooling layer. It is this vector that undergoes the process of normalization using the L2 Euclidean Norm and is further expanded or unpooled to generate novel feature mappings of equal size as the original. The next step involves the L2-normalization of all the initial feature maps. The last step deals with concatenation of feature mappings generated by the previous two steps. Normalization proves to be useful in scaling the concatenated feature map values and hence, results in a better performance. In short, the ParseNet is essentially an FCN except that the aforementioned module substitute the convolutional layers as presented in Figure 3.

Figure 3. (a-d) Comparison between the Fully Convolutional Network and ParseNet Results (e) Architecture of ParseNet



Convolutional and Deconvolutional Networks

This end to end network comprises two connected portions shown in Figure 4. The first portion is a convolutional net with the architecture of VGG16 and the second part is a deconvolutional network. The convolutional network takes an instance proposal, for instance, a bounding box produced by an object detector model as input, which is then processed and modified by a convolutional net to produce a feature vector. This vector is then input to the deconvolutional network, which then produces a pixel-wise probabilities map for every class. The deconvolutional net uses unpooling as shown in Figure 5 and utilizes the maximum activations in order to retain the information location in the maps. The 2nd net uses deconvolution as well; developing associations between a single input and several feature maps. The process of deconvolution results in an expansion of the feature maps whilst still keeping the information compact.

Upon analysis of the deconvolution feature maps, it was observed that the lower level feature maps are specific to the shape whilst the higher level maps are useful in categorizing the input proposal.

Figure 4. Visualization of convolutional and deconvolutional layers

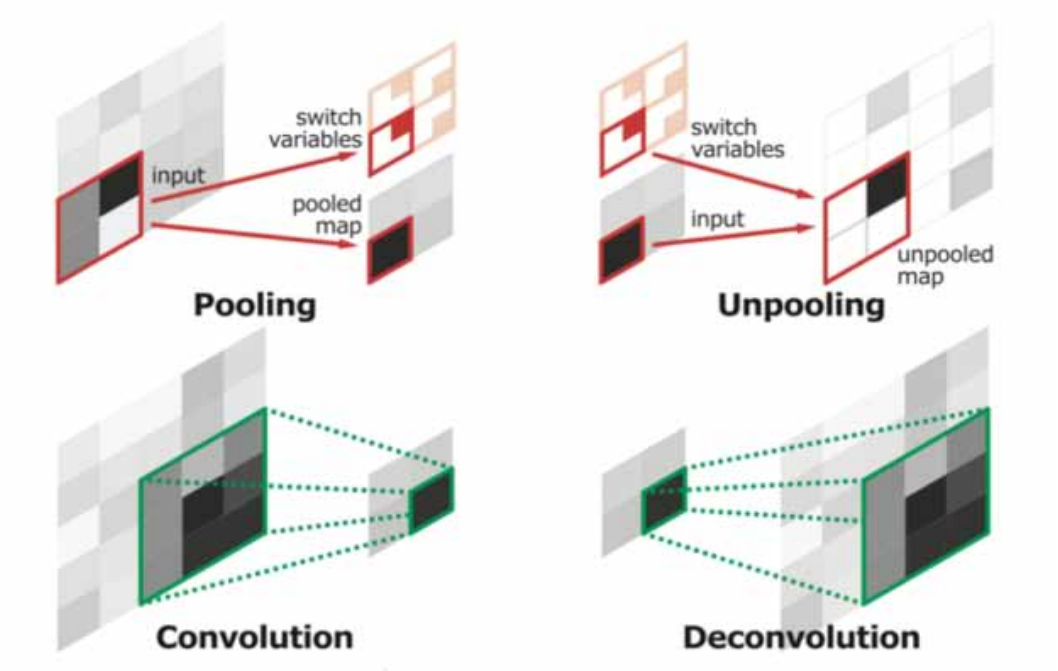
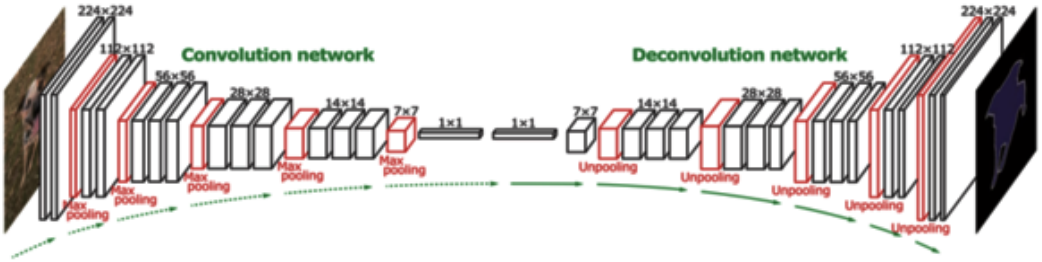


Figure 5. Architecture of the Convolutional and Deconvolutional Network

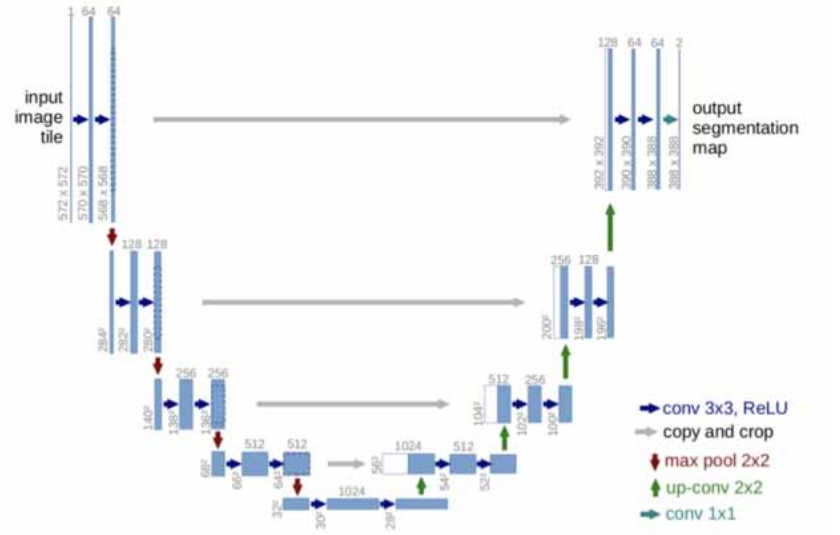


Ultimately, when all the image proposals are successfully processed by the model, the generated feature mappings are subsequently concatenated for getting the image, which is then segmented.

U-Net

U-Net was created as an extension of the Fully Convolutional Network by Ronneberger et al. in 2015 (Ronneberger, 2015), mainly to cater biological microscopy images. It consists of 2 parts - first is the contracting part, which works out features and the second is expanding part, which localizes the spatial patterns in the image as presented in Figure 6. The contracting part, also known as downsampling, possesses an FCN-like architecture, which derives features with 3x3 convolutions. The expanding part, also known as upsampling, uses deconvolution to decrease the number of feature maps whilst simultaneously projecting an increase in their width and height. Clipped feature mappings from the downsampling part of the network are duplicated in the upsampling segment in order to prevent the loss of pattern information. Lastly, a 1x1 convolution processes the generated feature mappings to produce a segmentation mapping, thus classifying every pixel to a relevant label. The U-Net has further

Figure 6. Architecture of U-net



been greatly extended for its use in other recent architectures. It is worthy to note that this model does not employ any fully connected layers and as a consequence, the no. of parameters of the network were greatly lessened. Thus, it is relatively easy for it to be trained with limited labelled samples.

Feature Pyramid Network (FPN)

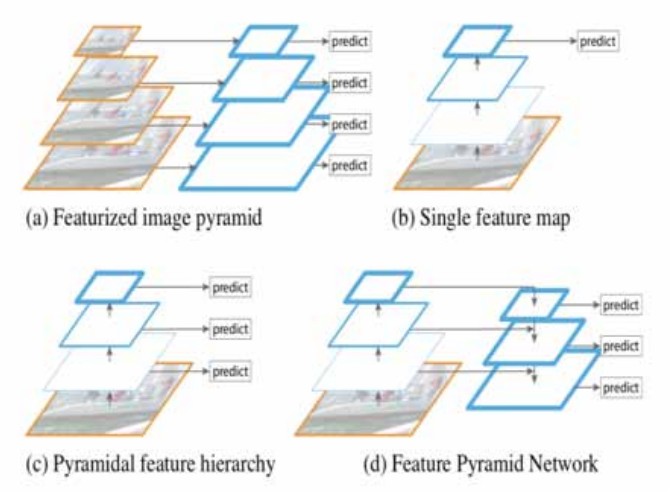
The FPN was created by T.Y. Lin et al. in 2016 (Lin, 2016). It is extensively utilized in object detection tasks and in frameworks utilizing image segmentation. The architecture is based on a bottom-up path, a top-down pipeline and horizontal connections to conjoin features of both lower and higher resolutions. An image of random size acts as the input for the bottom-up pathway. The processing on this image is done using convolutional layers which is followed by downsampling using pooling layers. Here, feature maps of the same size are grouped together to form what is known as a stage. The output generated in the last layer of every stage are the features utilized for the pyramid level. The top-down pipeline involves the upsampling of final feature mappings with unpooling by modifying them with feature mappings from the same stage obtained from the bottom-up pathway by making use of the lateral connections. These connections are responsible for integrating the feature mappings obtained from the bottom-up pathway with those from the top-down pipeline. The joined feature mappings further undergo processing by a 3x3 convolution to generate the resulting o/p of a stage. Finally, each stage in the top-down pipeline comes up with a prediction for object detection as shown in Figure 7. For the purpose of image segmentation, 2 Multi-Layer Perceptrons (MLP) are used to produce 2 masks of varying sizes over the objects.

BENCHMARK DATASETS AND EVALUATION METRICS FOR SEMANTIC SEGMENTATION

Datasets

Data is one of the most important parts of any machine learning system, especially one based on deep learning. For that reason, datasets play a crucial role in the performance of any segmentation model based on deep learning techniques. Thus, it is essential to use datasets which are representative

Figure 7. Detailed top-down pathway process with horizontal connections (Lin, 2016)



enough of the domain of the task at hand. In this section, we describe some common large-scale datasets which are popular and useful for the problem of semantic segmentation.

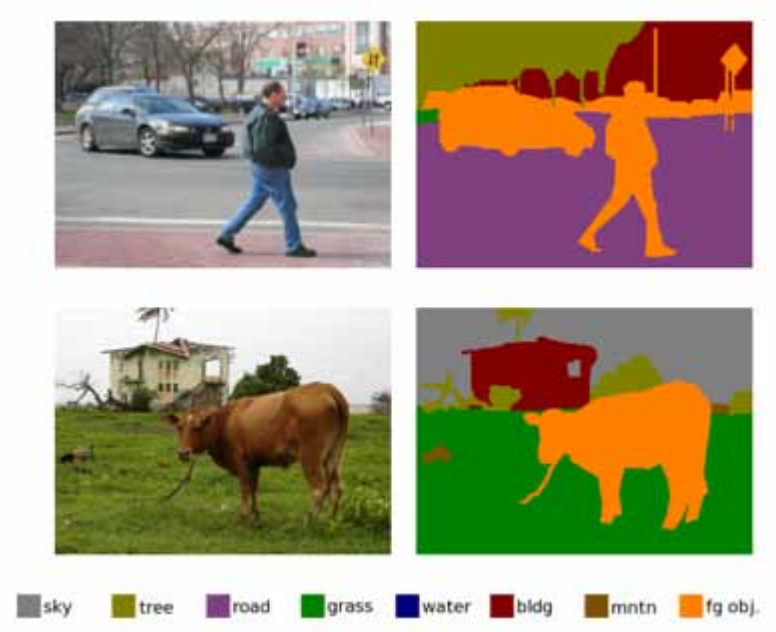
Stanford Background Dataset (Gould, 2009)

The Stanford background dataset contains images of outdoor scenes. This dataset was developed by choosing images from some public datasets, which are LabelMe, MSRC, PASCAL VOC and Geometric Context. Stanford background dataset includes 715 images for training. The size of the images is approximately 320 X 240 pixels. The images are selected in such a way that they have at least one foreground object. The labels in the dataset are horizons, regions, surfaces and layers, explained in Table 1. Some of the semantic classes mentioned in the dataset are sky, tree, road, mountain and building. Some of the geometric classes mentioned in the dataset are sky, horizontal and vertical (Li, 2009). Samples of the dataset are shown in Figure 8.

Table 1. Labels of Stanford background dataset (Gould, 2009)

Label	Description
horizons.txt	image dimensions and location of horizon
labels/*.regions.txt	integer matrix indicating each pixel's semantic class (sky, tree, road, grass, water, building, mountain or foreground object). A negative number indicates unknown.
labels/*.surfaces.txt	integer matrix indicating each pixel's geometric class (sky, horizontal or vertical).
labels/*.layers.txt	integer matrix indicating distinct image regions.

Figure 8. Example of images and semantic labels in the Stanford background dataset



Microsoft COCO Dataset (2015 Version) (Lin, 2014)

COCO stands for common objects in context. It contains images of everyday scenes captured in their natural context. The images in the dataset provide context information that is they attach context to the object in the images. There are 91 object categories in the dataset which include person, bicycle, truck, boat and traffic light. The dataset has 165,482 training images, 81,208 images for validation, and 81,434 test images. There are pixel level annotations in COCO, which can be used for scene understanding, as shown in Figure 9. This dataset is very commonly used for image recognition and segmentation. Some of the samples are presented in Figure 10.

Figure 9. (a) category labeling categories present in the image (b) marking the instances of the labeled categories (c) segmenting each object instance

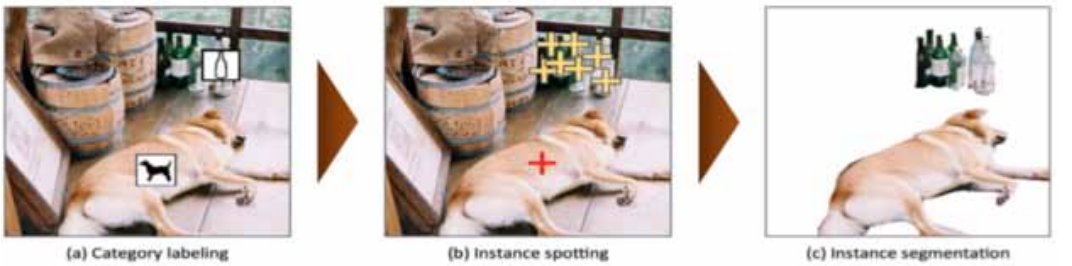
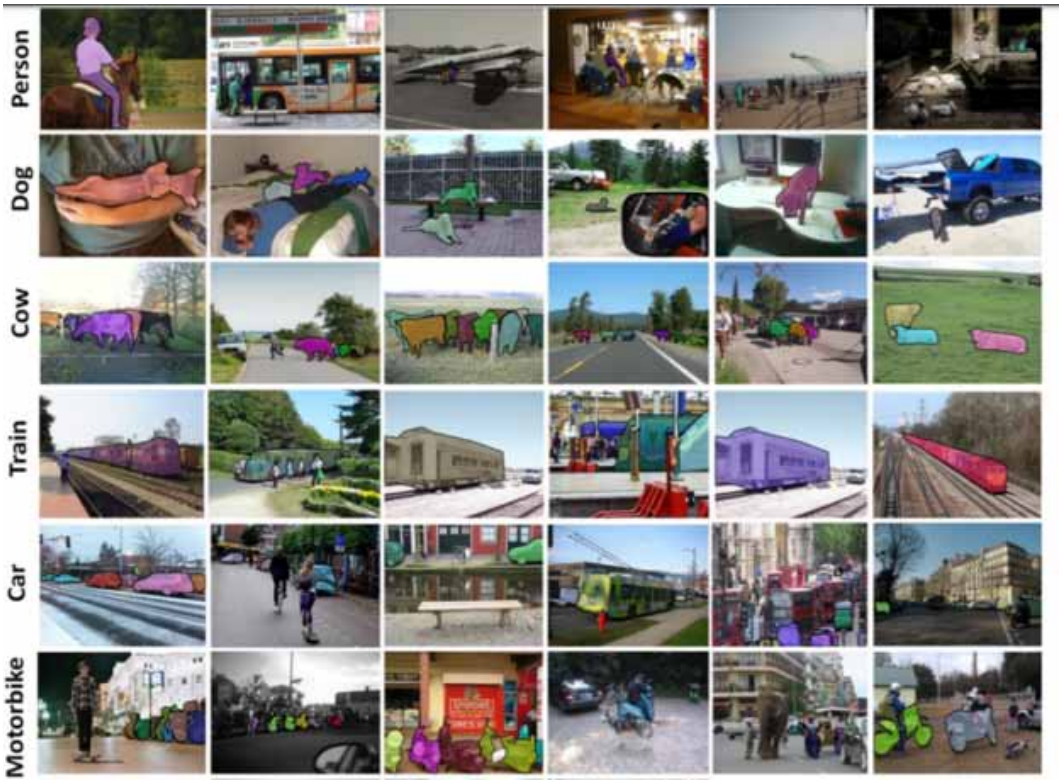


Figure 10. Example of images and classes in the COCO dataset (Lin, 2014)



Cityscapes Dataset (Cordts, 2016)

The Cityscapes Dataset is mainly centered on the semantic understanding of street scenes from urban areas which includes three different types of annotations, namely semantic, instance-wise & dense pixel annotations. It has thirty classes for which the class definitions are presented in Table 2. The diversity in the data is introduced by its collection in 50 different cities over a long tenure of several months under good/medium weather conditions. The frames were manually chosen with special focus on those consisting of an enormous no. of dynamic objects and variations in layouts of the scene and

Table 2. Classes present in the Cityscapes Dataset under their respective groups

Group	Classes
flat	road, sidewalk, parking+
human	person*, rider*
vehicle	car*, truck*, bus*, on rails*, motorcycle*, bicycle*, caravan*+, trailer*+
construction	building, wall, fence, guard rail+, bridge+, tunnel+
object	pole, pole group+, traffic sign, traffic light
nature	vegetation, terrain
sky	sky
void	ground+, dynamic+, static+

the background. This dataset provides 5000 annotated images with granular annotations and 20000 images having coarse annotations as shown in Figure 11 and Figure 12. The metadata for the images includes trailing and preceding frames in video, since each annotated image is the 20th image from a 30 frame video snippet. It also specifies the GPS coordinates and outside temperatures collected from the vehicle sensor.

Figure 11. Fine annotations (a) Frame in Zurich (b) Frame in Cologne

Figure 12. Coarse annotations (a) Frame in Dortmund (b) Frame in Erlangen

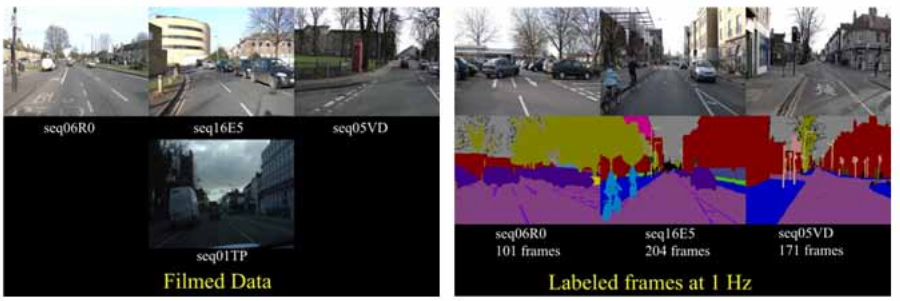
CamVid Dataset (Brostow, 2009)

CamVid is short for Cambridge-driving Labeled Video Dataset (CamVid). This dataset was created from the viewpoint of a vehicle being driven which ensures the heterogeneity in the captured data along with an increased number of samples and object classes. This dataset contains ground truth labeling for every pixel in reference to one of the 32 semantic classes set. It consists of 700 training and manually-annotated images of urban scene. Some of the semantic classes used in this dataset are mentioned in Table 3 and sample data is presented in Figure 13.

Table 3. Semantic classes in CamVid Dataset

Moving objects	Road	Ceiling	Fixed objects
Animal	Road == drivable surface	Sky	Building
Pedestrian	Shoulder	Tunnel	Wall
Child	Lane markings drivable	Archway	Tree
Rolling cart/luggage/pram	Non-Drivable		Vegetation misc.
Bicyclist			Fence
Motorcycle/scooter			Sidewalk
Car (sedan/wagon)			Parking block
SUV / pickup truck			Column/pole
Truck / bus			Traffic cone
Train			Bridge
Misc			Sign / symbol
			Misc text
			Traffic light
			Other

Figure 13. (a) Filmed data as recorded in the Camvid Dataset (b) Annotated frames in the CamVid dataset (Brostow, 2009)



KITTI Semantic Segmentation Benchmark (2018 Version) (Alhaija, 2018)

KITTI is one of the most popular datasets for its utility in mobile robotics and autonomous driving tasks. It consists of 200 labelled images available for training as well as 200 images available for testing purposes. The data format and metrics are similar to those used in (Alhaija, 2018). The annotated images consist of objects identified as one among the 34 classes defined. Figure 14 shows a sample of images available in this dataset.

Figure 14. Filmed and annotated frames in the KITTI dataset (Alhaija, 2018)

NYUDv2 (Silberman, 2012)

The NYU-Depth v2 dataset, also commonly known as the NYUDv2, consists of video sequences from various indoor scenes as captured by both RGB as well as depth cameras. It has 1449 labelled pairs of well-aligned RGB and depth images, wherein each object is labelled with a class, with respect to the 40 available classes, and an instance number. Apart from this, the dataset also has 407,024 new unlabelled frames, which were not available previously. As a whole, the dataset is comprised of labelled as well as raw images and a toolbox which has useful functions for dealing with the images and labels. Figure 15 shows some samples from the dataset.

PASCAL VOC 2012 (Everingham, 2010)

The PASCAL Visual Object Classes Challenge, more commonly known as the PASCAL VOC Challenge, is a benchmark in the visual category tasks and provides a standard dataset consisting of a ground-truth labelled set of images for five different competitions, namely Classification, Detection, Segmentation, Action Classification and Person Layout as shown in sample Figure 16. This dataset consists of 1464 images for training with 1449 images available for validation. There are 20 object classes in this dataset broadly categorized in Person, Animal, Vehicle and Indoor.

Figure 15. Samples of the RGB images, raw depth images and the segmented images (Silberman, 2012)

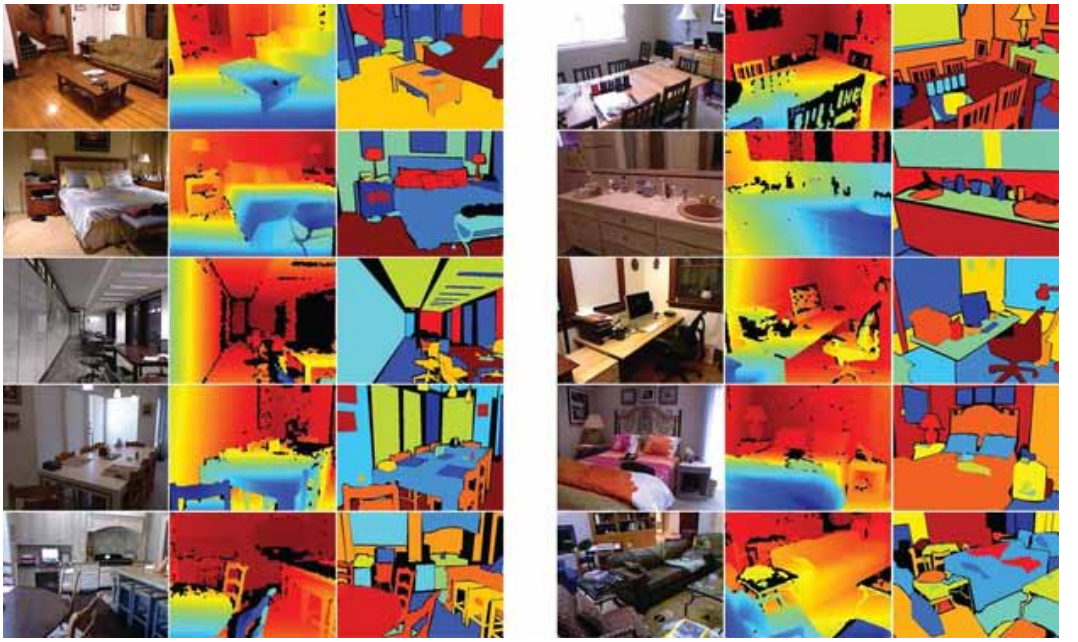
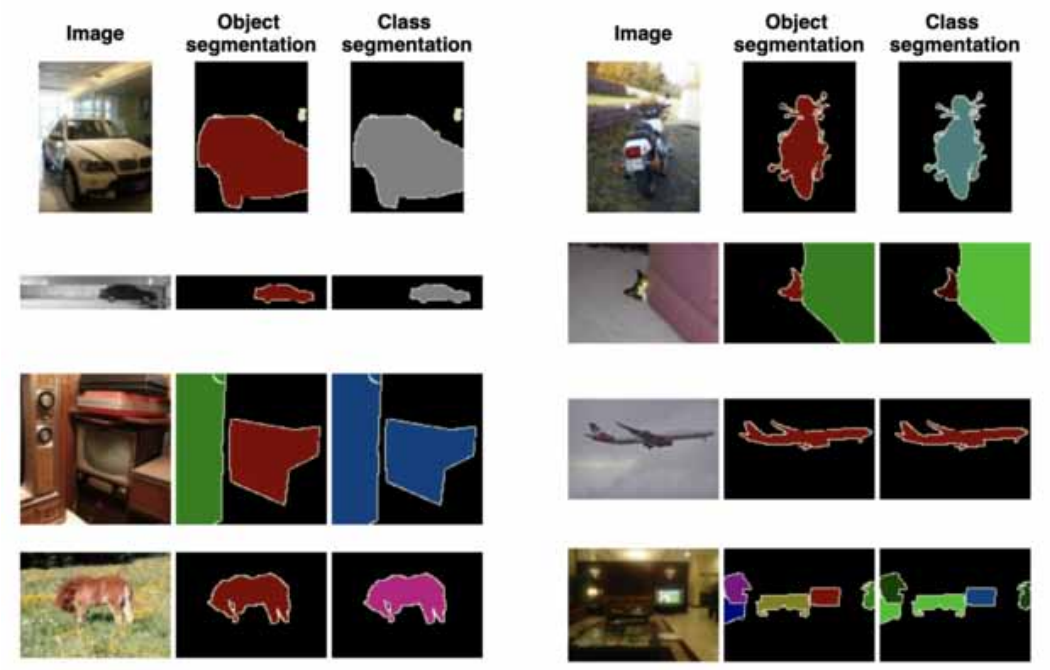


Figure 16. The training image, object segmentation and class segmentation examples from the PASCAL VOC dataset (Everingham, 2010)



SUN Database (Xiao, 2010)

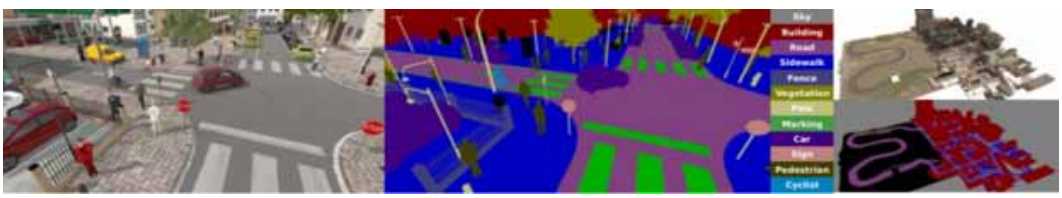
The Scene UNderstanding (SUN) database is a collection of annotated images of a variety of environmental scenes, places and objects within. The dataset, which is particularly for the advancing field of scene understanding, includes the richness of different environmental scenes belonging to different scene categories. SUN database contains 899 categories and 130,519 images. To build the dataset, all the entries in the WordNet English dictionary that directed to either the names of scenes, places or environments, were used to collect images. For each scene category images were selected using online image search engines. The objects in each image in all the categories were annotated manually. Some of the images present in the SUN dataset are shown in Figure 17.

Figure 17. Images of some SUN categories, with the percentage of human recognition rate mentioned (Xiao, 2010)

SYNTHIA (Ros, 2016)

The SYNTHetic collection of Imagery and Annotations is a dataset for scene understanding particularly for driving scenes in which a virtual world is used to generate realistic synthetic images from different viewpoints. The dataset contains 13,400 frames from the virtual city and pixel-level annotations for 13 classes. Figure 18 shows a sample frame from the dataset, showing the image in the left corner, the semantic labels of the image in the center and a general view of the city in the right corner.

Figure 18. Sample frame from SYNTHIA dataset (Left) with its semantic labels (center) and a general view of the city (right) (Ros, 2016)



MSRC v1 (Shotton, 2000)

The dataset is provided by Microsoft Research in Cambridge, which is commonly generated for scene segmentation and object recognition. Dataset contains 240 images and 9 object classes with coarse pixel wise labels. The dataset is having half images for training and half for testing that is 120 images for training and 120 images for testing.

LabelMe (Russell, 2007)

LabelMe is a database and an online annotation tool which provides functionalities like querying images, browsing database etc., while sharing images and annotations. The dataset has 30369 images divided into 183 categories and 111490 annotations in the database. Some samples from the dataset are shown in Figure 19 which show the object part relationship using polygon annotations.

Figure 19. Sample images in the dataset, object is in the center of its parts

Datasets Comparative Summary

Here, we present a summarized evaluation of all large scale and bench mark datasets designed for analyzing the semantic segmentation and scene understanding algorithms. Important parameters and the design choices that were kept in mind with regards to the focus of the dataset are summarized in Table 4.

Evaluation Metrics for Semantic Segmentation

Evaluation metrics help in analyzing the model performance. A quintessential aspect of evaluation metrics is their capability to distinguish between results obtained from various models. Following are the basic evaluation metrics which are required for evaluating semantic segmentation and scene parsing algorithms:

- **Pixel Accuracy:** Pixel accuracy can be interpolated as the percent of pixels correctly classified in the image (Garcia-Garcia, 2017; Liu, 2019). The pixel accuracy is usually computed for each class separately as well as on a global scale, i.e., across all classes. Global accuracy can

Table 4. Benchmark and large scale datasets for Semantic segmentation and Scene understanding

Dataset Name	Purpose	Year	Classes	Data	Synthetic/ Real	Samples
Stanford Background (Gould, 2009)	Outdoor	2009	8	2D	Real	715 (572 training images and 143 test images)
COCO (Lin, 2014)	General	2015	91	2D	Real	328,000 images 165,482(training images) 81,208(validation images) 81,434(test images)
Cityscape (Cordts, 2016)	Urban	2016	30	2D	Real	2975 (training) 1525 (testing) 500 (validation)
CamVid (Brostow, 2009)	Urban/ Driving	2008	32	2D	Real	700 (training)
KITTI (Alhaija, 2018)	Urban/ Driving	2018	34	2D	Real	200 (training) 200 (testing)
NYUDv2 (Silberman, 2012)	Indoor	2012	40	2.5D	Real	1449 795 (training) 654 (validation)
PASCAL VOC 2012 (Everingham, 2010)	General	2012	21	2D	Real	11,530 1464 (training) 1449 (validation)
SUN (Xiao, 2010)	Outdoor scenes	2010	899	2D	Real	130,519
SYNTHIA (Ros, 2016)	Urban/ Driving	2016	13	2D	Synthetic	13400
MSRC (Shotton, 2000)	General	2005	9	2D	Real	240
LabelMe (Russell, 2007)	General	2006	183	2D	Real	111490

be measured by finding the ration of correctly classified pixels (regardless of class) to the total number of classes. For particular class pixel accuracy can be calculated using eq. 1:

$$P_{acc} = \frac{\sum n_{ii}}{\sum t_i} \tag{1}$$

where n_{ii} are the number of pixels class *i* predicted belong to class *i* and t_i are the total number of pixels of class *i*.

- **Intersection over Union (IoU):** IoU can be defined as "the area of overlap between the predicted segmentation and the ground truth divided by the area of union between the predicted segmentation and the ground truth" (Guo, 2016):

$$IoU = \frac{Target \cap Prediction}{Target \cup Prediction} \tag{2}$$

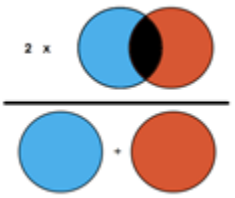
This can easily be understood from the visualization below:

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union}$$

- **Weighted IoU:** Weighted IoU can also be measured by taking average IoU of each class weighted by the number of pixels in that class. This metric is useful when image have imbalanced classes.
- **Dice Coefficient:** Dice Coefficient, also known as F1 score, is also a commonly used metric for evaluating semantic segmentation models. Simply put, it is twice the area of overlap divided by the total number of pixels in both images (Liu, 2019):

$$F1_{score} = \frac{2{*}Area\ of\ Overlap}{Total\ number\ of\ pixels\ combined} \tag{3}$$

The following illustration makes it easy to understand:



  - **Boundary F1 (BF) Score:** It is a contour matching score which indicates quality of predicted boundary in each class. This metric correlates better with human qualitative assessment.

RECENT PROGRESS IN DEEP LEARNING BASED SEMANTIC SEGMENTATION

Due to the popularity of deep learning techniques and their performance in different fields, recently researchers are trying to incorporate different deep learning techniques along with a combination of traditional methods in the field of semantic segmentation and scene understanding. Some of the recent work and reviews done in related fields in the past five years is explained below.

A general review on scene understanding can be found in (Aarthi, 2017), where authors have discussed the problem and concept of scene understanding extensively whilst presenting several strategies and techniques that are relevant to this field. It begins by presenting a description of the process of gaining meaningful insights from different scenes or visuals by highlighting some strategies with their classifications. The authors then presented some key challenges and factors that might affect the accuracy of a scene understanding model or system. Context based and semantic based analysis of 2D images is covered in great detail in order to aid better understanding of the scene understanding process as well as to present a comparison between several state-of-the-art strategies including this

parameter. An extensive review on deep learning for high dimensional data is presented in (Kolberg, 2018), where authors have described various deep networks and their usage in handling for various high dimensional data based applications. Further, review on deep learning techniques extensively for semantic segmentation is given in (Guo, 2017; Liu, 2019, Garcia-Garcia, 2017). In (Garcia-Garcia, 2017) different segmentation methods are broadly divided into the categories of traditional methods and recent deep neural network methods. The datasets used for segmentation are also briefly discussed. In (Liu, 2019), authors presented commonly observed and required terms used in this field of research as well as some important background concepts. In addition to this, a thorough evaluation is done for a multitude of datasets which are sought after in this domain. It also highlights some challenges faced by researchers in the usage of these datasets to encourage the reader in making informed decisions about selecting one that is most suitable to their requirements and goals.

Further, for performing the meaningful segmentation on an image, a Fully Convolutional Network (FCN) architecture is proposed which was composed solely of convolutional layers. The network showed the state-of-the-art performance at the time for the task for semantic segmentation. In (Liu, 2015) the authors extend the FCN model to incorporate the global context of the image whilst semantically segmenting it. In this work, the convolutional layers of the FCN get replaced by modules which take the feature mappings as input, which are initially also produced as part of the network. A contracting-expanding network was introduced by the authors in (Ronneberger, 2015) where the contracting part was responsible for feature and context mapping whereas the expanding part was used for accurate localisation. A deep convolution network based architecture SegNet is described in (Badrinarayanan, 2017). This network has an encoder and decoder network followed by a pixel-wise classifier. The decoder network of SegNet is designed such that the network is efficient in memory storage and computational time during inference. The decoder uses the max-pooling indices of the feature maps, which eliminates the need for learning to upsample. The network also uses less number of trainable parameters and can be trained end-to-end. The authors designed the network motivated specifically by road and indoor scene understanding. The datasets used in the paper are CamVid dataset for road scene segmentation and SUN RGB-D for indoor scene segmentation. In the paper, analysis of SegNet is done and the network is compared with other segmentation architecture which share the same encoder but different decoders. For comparison purposes a smaller version of SegNet is used called SegNet-Basic. To compare the performance of the networks (decoder variants) the performance measures used are global accuracy, class average accuracy, mean intersection over union and boundary F1-measure (BF). The authors also provide a CAFFE implementation of SegNet and a web demo. In (Lin, 2016), authors describe a framework to perform object detection by constructing feature pyramids with marginal extra cost. The architecture is called Feature Pyramid Network (FPN) which develops high-level semantic feature maps within deep convolutional networks. Further, a fully convolutional neural network architecture called BlitzNet is proposed in (Dvornik, 2017) to perform the task of semantic segmentation and object detection simultaneously in one forward pass. The architecture utilizes the network ResNet-50 to extract high-level features, i.e., to perform feature encoding. Then, the Single Shot Detection (SSD) approach is employed to search for bounding boxes by reducing the resolution of the generated feature maps. For the task of semantic segmentation, upsampling is performed on the feature maps using deconvolutional layers in order to generate accurate segmentation maps. The final prediction is performed by separate single convolutional layers - each for detection and segmentation - in a single forward pass. The experiments were conducted on the COCO (Lin, 2014), PASCAL VOC (Everingham, 2010) datasets. A novel method is proposed in (Li, 2017) for the task of scene understanding by modelling it as a joint problem of object detection, scene graph generation and region captioning. This is implemented using their neural network architecture called "Multi-level Scene Description Network (MSDN)" which utilizes the convolutional layers of VGG-16, primarily being used for the region proposal and recognition network. The object detection pipeline of the model follows the Faster-RCNN approach. The model proposes regions for objects, phrases and region captions, following which specialized features are extracted to construct

dynamic graphs. The experiments were conducted on the Visual Genome dataset. UPerNet, which is a framework for Unified Perceptual Parsing is presented in (Xiao, 2018) which can recognize several visual concepts simultaneously. UPerNet includes Feature Pyramid Network (FPN) and Pyramid Pooling Module (PPM) which enable the network to unify the different visual attributes. The trained network is also used to discover visual knowledge in natural scenes. A training strategy is developed to teach the model from heterogeneous datasets i.e. Broadly and Densely Labeled Dataset (Broden), which combines several datasets to incorporate different visual concepts. The authors use different evaluation metrics for different visual concepts parsing based on the annotations of the datasets. In (Zhang, 2018), the authors present a framework, ExFuse, which tackles the problem of ineffective feature fusion, by bridging the gap between high-level low-resolution and low-level high-resolution features. The framework introduces semantic information into low-level features and high resolution details into high-level features. In (Chen, 2018), the task of semantic segmentation with deep learning is discussed, by making three contributions. Firstly atrous convolution with upsampled filters is applied for dense feature extraction. Secondly the authors propose atrous spatial pyramid pooling (ASPP) for segmentation of objects at different scales. Thirdly, deep convolutional neural networks are combined with fully connected Conditional Random Field to improve localization performance of object boundaries. The authors of (Valada, 2019) introduce a multimodal approach to the problem of semantic segmentation along with proposing a unimodal network called AdaptNet++ for computationally efficient performance. Further, the comparative analysis of different approaches is summarized in Table 5.

DISCUSSION

Semantic image segmentation is an important step in scene understanding task, which makes it a key research area in computer vision society. Due to the unpredictable real-world scenario and complexity of some imaging domains like medical imaging, the proper segmentation of images is always a research issue among researchers. Due to the importance of the current research domain, through this paper, the authors presented a high-level view of the traditional methods followed by an extensive review of deep learning-based methods for the task of semantic segmentation. By preparing this paper authors achieved following key points:

- A thorough background of segmentation to semantic segmentation is presented for better understanding of the field.
- Traditional, state of the art techniques along with some advanced adopted approaches before the use of deep learning techniques are described.
- Various deep networks which were used by the researchers for semantic segmentation are summarized.
- As datasets play an important role for evaluating the performance of any proposed model, in this paper various benchmark and large scale datasets that are publicly available for testing the semantic segmentation algorithms are identified. Whilst most datasets are a collection of 2D images, some being made up of frames from video segments, there do exist a few which comprise 2.5D images, implying that the depth of the image can also be made use of for the task of semantic segmentation.
- In addition to this, some metrics have been identified to aid in the proper evaluation of the developed models.
- Beside brief review of traditional approaches, a comprehensive recent progress on deep learning based semantic segmentation is also presented.

Table 5. A summarized review of semantic segmentation related work

Ref. No.	Methodology	Datasets Used	Analysis
(Long, 2015)	An implementation of a model built exclusively of convolutional layers was introduced in this paper. A skip architecture has also been defined for transfer of information and hence, more accurate segmentation.	PASCAL VOC 2012, NYUDv2, SIFT Flow	This model showed efficient performance in making dense predictions for semantic segmentation. This approach derived from Convolutional Neural Networks proved to be the foundation of various other networks and models to follow.
(Liu, 2015)	A model called ParseNet is proposed which has modules which work on the feature mappings of the image rather than regions of an image.	PASCAL VOC 2012, PASCAL-Context, SiftFlow	Inclusion of the global spatial context of the image was considered in addition to the Fully Convolutional Network approach.
(Ronneberger, 2015)	A model called U-Net is proposed which consists of a contacting part to work out features and context and an expanding part used for accurate and precise localization.	EM Segmentation Challenge by ISBI 2012	Authors developed an efficient architecture made of convolutional layers that makes strong use of data augmentation techniques to train and evaluate the model on a relatively small dataset.
(Badrinarayanan, 2017)	A network SegNet is used for pixel-wise semantic segmentation of road and indoor scenes. Comparison of SegNet and other segmentation architectures is done using different performance measures.	CamVid and SUN RGB-D	New approach towards segmentation was understood which is more efficient in terms of memory and computational time. The way in which a decoder can be designed to improve the performance of the network was learned.
(Lin, 2016)	A model called Feature Pyramid Network (FPN) is implemented. It is a framework for building feature pyramids inside Convolutional Neural Networks, used for object detection.	COCO	A practical solution for research and applications of the feature pyramid using Convolutional Neural Network is provided. The study suggests that despite the strong representational power of deep CNN, multiscale problems should be addressed using pyramid representations.
(Dvornik, 2017)	An implementation focused on simultaneous semantic segmentation and object detection using the ResNet-50 architecture with the SSD approach for object detection and upsampling method for semantic segmentation.	COCO, PASCAL VOC 2007 and 2012	Proposed architecture jointly performs object detection and semantic segmentation which increases the accuracy as both the tasks benefit from each other. There is weight sharing between the tasks which enhances the learning process.
(Li, 2017)	An implementation focused on finding the solution as an intersection of object detection, scene graph generation and region captioning for the task of scene understanding.	Visual Genome Dataset	Understood a new perspective and approach to the problem of scene understanding as a joint problem of object detection, scene graph generation and region captioning.
(Chen, 2018)	A network DeepLab is proposed that performs semantic segmentation using atrous convolution, which is further extended to atrous spatial pyramid pooling. Deep convolutional neural networks and fully-connected conditional random fields are also combined to improve semantic segmentation and object boundaries.	PASCAL VOC-2012, PASCAL-Context, PASCAL-Person-Part, and Cityscapes.	DeepLab is a state-of-the-art method for semantic segmentation. Atrous convolution can be used to enlarge the field-of-view of filters at any DCNN layer. Combining the responses at the final DCNN layer with a fully connected CRF improves the localization performance both qualitatively and quantitatively.
(Zhang, 2018)	A framework ExFuse is presented that enhances the feature fusion process for semantic segmentation. The framework bridges the gap between low-level and high-level features to improve the quality of segmentation.	PASCAL VOC 2012 segmentation benchmark	Simple fusion of low-level and high-level features is less effective because of the gap in semantic levels. Introducing semantic details into low-level features along with introducing high resolution details into high-level features results in better fusion.
(Xiao, 2018)	A model UPerNet, which is a framework for Unified Perceptual Parsing, is used. The model is used to recognize several visual concepts simultaneously. The trained network is also used to discover visual knowledge in natural scenes.	Broden+	The model presented is able to recognize a wide range of visual concepts from images, which helps to discover rich visual knowledge from real world scenes and can help future vision systems to understand its surroundings better.
(Valada, 2019)	An architecture is proposed for a multimodal encoder streams that get fused into one intermediate representation before getting passed on to the decoder.	Cityscapes, Synthia, SUN RGB-D, ScanNet, Freiburg Forest Benchmark	The mentioned method and model leverages multiple modalities which allow for learning richer and better representations that are robust to challenges like appearance changes etc.

CONCLUSION

The method of semantic segmentation for scene understanding is gaining immense popularity due to its efficiency in obtaining correct classification for each pixel of the image, which further makes it easy for the image to be semantically understood. Semantic segmentation and scene understanding field has a plethora of applications in different fields. Due to the demand of current field various review papers are available in literature (Guo, 2018; Lateef, 2019; Garcia-Garcia, 2018; Liu, 2019). While most of the existing surveys are based on the analysis of deep learning architectures, here we are presenting a comprehensive study on semantic segmentation with respect to scene understanding field, which is still an open problem due to various complex real time scenarios. In this paper, we consider some traditional methods for semantic segmentation before reviewing the modern approaches using deep learning, with the aim of implementing a functioning system for scene understanding. In addition to this, several datasets were identified which are relevant to and useful for the problem at hand. The paper was concluded by the review of the literature available in order to gain an in-depth understanding of the process of scene understanding.

FUNDING AGENCY

The publisher has waived the Open Access Processing fee for this article.

REFERENCES

Aarthi, S., & Chitrakala, S. (2017). Scene understanding — A survey. *2017 International Conference on Computer, Communication and Signal Processing (ICCCSP)*. doi:10.1109/ICCCSP.2017.7944094

al-Azawi, A.n., M. (2013). Image Thresholding using Histogram Fuzzy Approximation. *International Journal of Computers and Applications*, *83*(9), 36–40. doi:10.5120/14480-2781

Alhaija, H. A., Mustikovela, S. K., Mescheder, L., Geiger, A., & Rother, C. (2018). Augmented Reality Meets Computer Vision: Efficient Data Generation for Urban Driving Scenes. *International Journal of Computer Vision*, *126*(9), 961–972. doi:10.1007/s11263-018-1070-x

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(12), 2481–2495. doi:10.1109/tpami.2016.2644615

Brostow, G. J., Fauqueur, J., & Cipolla, R. (2009). Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, *30*(2), 88–97. doi:10.1016/j.patrec.2008.04.005

Brust, C., Sickert, S., Simon, M., Rodner, E., & Denzler, J. (2015). Convolutional Patch Networks with Spatial Prior for Road Detection and Urban Scene Understanding. *Proceedings of the 10th International Conference on Computer Vision Theory and Applications*. doi:10.5220/0005355105100517

Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). 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. doi:10.1109/tpami.2017.2699184

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., & Schiele, B. et al. (2016). The Cityscapes Dataset for Semantic Urban Scene Understanding. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. doi: doi:10.1109/cvpr.2016.350

Dvornik, N., Shmelkov, K., Mairal, J., & Schmid, C. (2017). BlitzNet: A Real-Time Deep Network for Scene Understanding. *2017 IEEE International Conference on Computer Vision (ICCV)*. doi: doi:10.1109/iccv.2017.447

Everingham, M., Gool, L. V., Williams, C. K., Winn, J., & Zisserman, A. (2009). The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, *88*(2), 303–338. doi:10.1007/s11263-009-0275-4

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Martinez-Gonzalez, P., & Garcia-Rodriguez, J. (2018). A survey on deep learning techniques for image and video semantic segmentation. *Applied Soft Computing*, *70*, 41–65. doi:10.1016/j.asoc.2018.05.018

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Retrieved January 25, 2019, from https://www.deeplearningbook.org/

Gould, S., Fulton, R., & Koller, D. (2009). Decomposing a scene into geometric and semantically consistent regions. *2009 IEEE 12th International Conference on Computer Vision*. doi: doi:10.1109/iccv.2009.5459211

Gould, S., Gao, T., & Koller, D. (2009). Region-Based Segmentation and Object Detection. *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, 655-663.

Guo, Y., Liu, Y., Georgiou, T., & Lew, M. S. (2017). A review of semantic segmentation using deep neural networks. *International Journal of Multimedia Information Retrieval*, *7*(2), 87–93. doi:10.1007/s13735-017-0141-z

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., & Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing*, *187*, 27–48. doi:10.1016/j.neucom.2015.09.116

Gupta, S., Arbeláez, P., Girshick, R., & Malik, J. (2014). Indoor Scene Understanding with RGB-D Images: Bottom-up Segmentation, Object Detection and Semantic Segmentation. *International Journal of Computer Vision*, *112*(2), 133–149. doi:10.1007/s11263-014-0777-6

He, Y., & Kayaalp, M. (2008). Biological entity recognition with conditional random fields. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 293–297.

Karthicsonia, B., & Vanitha, M. (2019). Edge Based Segmentation in Medical Images. *International Journal of Engineering and Advanced Technology Regular Issue*, *9*(1), 449–451. doi:10.35940/ijeat.a9484.109119

Kaymak, C., & Ucar, A. (2019). Semantic Image Segmentation for Autonomous Driving Using Fully Convolutional Networks. *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*. doi:10.1109/IDAP.2019.8875923

Ker, J., Wang, L., Rao, J., & Lim, T. (2018). Deep Learning Applications in Medical Image Analysis. *IEEE Access: Practical Innovations, Open Solutions*, *6*, 9375–9389. doi:10.1109/ACCESS.2017.2788044

Kim, W., & Seok, J. (2018). Indoor Semantic Segmentation for Robot Navigating on Mobile. *2018 Tenth International Conference on Ubiquitous and Future Networks (ICUFN)*. doi:10.1109/ICUFN.2018.8436956

Kolberg, L., & Allikivi, M. L. (2018). *DCGAN for classroom images*. https://neuro.cs.ut.ee/wp-content/uploads/2018/02/DCGAN.pdf

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. doi:10.1145/3065386

Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01)*. Morgan Kaufmann Publishers Inc.

Lalaoui, L., & Mohamadi, T. (2013). A comparative study of Image Region-Based Segmentation Algorithms. *International Journal of Advanced Computer Science and Applications*, *4*(6). Advance online publication. doi:10.14569/ijacsa.2013.040627

Lateef, F., & Ruichek, Y. (2019). Survey on semantic segmentation using deep learning techniques. *Neurocomputing*, *338*, 321–348. doi:10.1016/j.neucom.2019.02.003

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., & Jackel, L. (1990). *Handwritten Digit Recognition with a Back-Propagation Network* (Vol. 2). Morgan-Kaufmann.

Lee, S., Chen, T., Yu, L., & Lai, C. (2018). Image Classification Based on the Boost Convolutional Neural Network. *IEEE Access: Practical Innovations, Open Solutions*, *6*, 12755–12768. doi:10.1109/ACCESS.2018.2796722

Li, L., Socher, R., & Fei-Fei, L. (2009). Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. *2009 IEEE Conference on Computer Vision and Pattern Recognition*. doi:10.1109/CVPR.2009.5206718

Li, Y., Ouyang, W., Zhou, B., Wang, K., & Wang, X. (2017). Scene Graph Generation from Objects, Phrases and Region Captions. *2017 IEEE International Conference on Computer Vision (ICCV)*. doi: doi:10.1109/iccv.2017.142

Lin, T., Dollar, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature Pyramid Networks for Object Detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. doi: doi:10.1109/cvpr.2017.106

Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., & Zitnick, C. L. et al. (2014). Microsoft COCO: Common Objects in Context. *Computer Vision – ECCV 2014*. *Lecture Notes in Computer Science*, 740–755. doi:10.1007/978-3-319-10602-1_48

Liu, W., Rabinovich, A., & Berg, A. C. (2015). ParseNet: Looking Wider to See Better. *Computer Vision and Pattern Recognition*. https://arxiv.org/abs/1506.04579

Liu, X., Deng, Z., & Yang, Y. (2018). Recent progress in semantic image segmentation. *Artificial Intelligence Review*, *52*(2), 1089–1106. doi:10.1007/s10462-018-9641-3

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. doi: doi:10.1109/cvpr.2015.7298965

Lu, D., & Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing*, *28*(5), 823–870. doi:10.1080/01431160600746456

Maolood, I. Y., Al-Salhi, Y. E., & Lu, S. (2018). Thresholding for medical image segmentation for cancer using fuzzy entropy with level set algorithm. *Open Medicine: a Peer-Reviewed, Independent, Open-Access Journal*, *13*(1), 374–383. doi:10.1515/med-2018-0056

Messer, K. D., Costanigro, M., & Kaiser, H. M. (2017). Labeling Food Processes: The Good, the Bad and the Ugly. *Applied Economic Perspectives and Policy*, *39*(3), 407–427. doi:10.1093/aepp/ppx028

Muthukannan, K., & Merlin, M. M. (2010). Color image segmentation using k-means clustering and Optimal Fuzzy C-Means clustering. *International Conference on Communication and Computational Intelligence (INCOCCI)*, 229-234.

Padmapriya, B., Kesavamurthi, T., & Ferose, H. W. (2012). Edge Based Image Segmentation Technique for Detection and Estimation of the Bladder Wall Thickness. *Procedia Engineering*, *30*, 828–835. doi:10.1016/j.proeng.2012.01.934

Ramadevi, Y., Sridevi, T., Poornima, B., & Kalyani, B. (2010). Segmentation And Object Recognition Using Edge Detection Techniques. *International Journal of Computer Science and Information Technology*, *2*(6), 153–161. doi:10.5121/ijcsit.2010.2614

Ramesh, N. (1995). Thresholding based on histogram approximation. *IEE Proceedings. Vision Image and Signal Processing*, *142*(5), 271. doi:10.1049/ip-vis:19952007

Ripon, K. S., Newaz, S., Ali, L. E., & Ma, J. (2017). Bi-level multi-objective image segmentation using texture-based color features. *2017 20th International Conference of Computer and Information Technology (ICCIT)*. doi: doi:10.1109/iccitechn.2017.8281795

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. Lecture Notes in Computer Science Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, 234-241. doi: doi:10.1007/978-3-319-24574-4_28

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., & Lopez, A. M. (2016). The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. doi: doi:10.1109/cvpr.2016.352

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2007). LabelMe: A Database and Web-Based Tool for Image Annotation. *International Journal of Computer Vision*, *77*(1-3), 157–173. doi:10.1007/s11263-007-0090-8

Sakthivel, K., Nallusamy, R., & Kavitha, C. (2014). Color Image Segmentation Using SVM Pixel Classification Image. *World Academy of Science, Engineering and Technology, Open Science Index 94*. *International Journal of Computer and Information Engineering*, *8*(10), 1924–1930.

Savkare, S., & Narote, S. (2012). Automatic System for Classification of Erythrocytes Infected with Malaria and Identification of Parasite's Life Stage. *Procedia Technology*, *6*, 405–410. doi:10.1016/j.protcy.2012.10.048

Shan, P. (2018). Image segmentation method based on K-mean algorithm. *EURASIP Journal on Image and Video Processing, 2018*(1). doi:10.1186/s13640-018-0322-6

Sharma, N., Ray, A., Sharma, S., Shukla, K., Pradhan, S., & Aggarwal, L. (2008). Segmentation and classification of medical images using texture-primitive features: Application of BAM-type artificial neural network. *Journal of Medical Physics / Association of Medical Physicists of India*, *33*(3), 119. doi:10.4103/0971-6203.42763

Shotton, J., & Winn, J. (2000, January 1). *Image Understanding*. Microsoft Research. Retrieved December 01, 2019, from https://www.microsoft.com/en-us/research/project/image-understanding/

Silberman, N., Hoiem, D., Kohli, P., & Fergus, R. (2012). Indoor Segmentation and Support Inference from RGBD Images. *Computer Vision – ECCV 2012*. *Lecture Notes in Computer Science*, 746–760. doi:10.1007/978-3-642-33715-4_54

Srinivas, S., Sarvadevabhatla, R. K., Mopuri, K. R., Prabhu, N., Kruthiventi, S. S., & Babu, R. V. (2016). A Taxonomy of Deep Convolutional Neural Nets for Computer Vision. *Frontiers in Robotics and AI*, *2*. Advance online publication. doi:10.3389/frobt.2015.00036

Valada, A., Mohan, R., & Burgard, W. (2019). Self-Supervised Model Adaptation for Multimodal Semantic Segmentation. *International Journal of Computer Vision*, *128*(5), 1239–1285. doi:10.1007/s11263-019-01188-y

Verbeek, J., & Triggs, B. (2007). Scene segmentation with Conditional Random Fields learned from partially labeled images. In *Proceedings of the 20th International Conference on Neural Information Processing Systems (NIPS'07)*. Curran Associates Inc.

Verschae, R., & Ruiz-Del-Solar, J. (2015). Object Detection: Current and Future Directions. *Frontiers in Robotics and AI*, *2*. Advance online publication. doi:10.3389/frobt.2015.00029

Wang, X., Wang, T., & Bu, J. (2011). Color image segmentation using pixel wise support vector machine classification. *Pattern Recognition*, *44*(4), 777–787. doi:10.1016/j.patcog.2010.08.008

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). SUN database: Large-scale scene recognition from abbey to zoo. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. doi: doi:10.1109/cvpr.2010.5539970

Xiao, J., Hays, J., Russell, B. C., Patterson, G., Ehinger, K. A., Torralba, A., & Oliva, A. (2013). Basic level scene understanding: Categories, attributes and structures. *Frontiers in Psychology*, *4*. Advance online publication. doi:10.3389/fpsyg.2013.00506 PMID:24009590

Xiao, J., Russell, B. C., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2012). Basic level scene understanding. *SIGGRAPH Asia 2012 Technical Briefs on - SA '12*. doi:10.1145/2407746.2407782

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., & Sun, J. (2018). Unified Perceptual Parsing for Scene Understanding. *Computer Vision – ECCV 2018. Lecture Notes in Computer Science*, 432–448. doi:10.1007/978-3-030-01228-1_26

Zaitoun, N. M., & Aqel, M. J. (2015). Survey on Image Segmentation Techniques. *Procedia Computer Science*, *65*, 797–806. doi:10.1016/j.procs.2015.09.027

Zhang, Z., Zhang, X., Peng, C., Xue, X., & Sun, J. (2018). ExFuse: Enhancing Feature Fusion for Semantic Segmentation. *Computer Vision – ECCV 2018. Lecture Notes in Computer Science*, 273–288. doi:10.1007/978-3-030-01249-6_17

Aakanksha is an Undergraduate Research Scholar at Amity University, Noida. Her areas of interest include image processing, pattern recognition and deep learning.

Arushi Seth is a research scholar at Amity University, Noida. Her interests include Image processing, pattern recognition and deep learning. She has a deep interest in exploring and applying deep learning techniques to various applications.

Shanu Sharma is currently working as an Assistant Professor in Department of Computer Science & Engineering at ABES Engineering College, Ghaziabad, India. She has a teaching experience of ten years. Her are of interest includes Image processing, pattern recognition, artificial intelligence and Cognitive computing. She is an active member of professional societies like IEEE, ACM, Soft Computing Research Society and lifetime member of International Association of Engineers (IAENG). She has presented and published various research papers at National and International conferences and journals and currently associated with various international conferences and journals as reviewer.