U.Porto Journal of Engineering, 8:6 (2022) 132-159
ISSN 2183-6493
DOI: 10.24840/2183-6493_008.006_0010

Received: 12 May, 2022
Accepted: 18 June, 2022
Published: 28 November, 2022

State of the Art Techniques to Advance Deep Networks for Semantic Segmentation - A Systematic Review

Aakanksha[1], Arushi Seth[2], Shanu Sharma[3]

[1]Department of Computer Science & Engineering, Amity School of Engineering & Technology, Amity University Uttar Pradesh, India (itaakanksha@gmail.com) ORCID 0000-0002-6953-7805; [2]Department of Computer Science & Engineering, Amity School of Engineering & Technology, Amity University Uttar Pradesh, India (setharushi@hotmail.com) ORCID 0000-0001-7635-1964; [3]Department of Computer Science & Engineering, ABES Engineering College, Ghaziabad, Uttar Pradesh, India (shanu.sharma16@gmail.com) ORCID 0000-0003-0384-7832

Abstract

In recent times, the computer vision community has seen remarkable growth in the field of scene understanding. With such a wide prevalence of images, the importance of this field is growing rapidly along with the technologies involved in it. Semantic Segmentation is an important step in scene understanding which requires the assignment of each pixel in an image to a pre-defined class and achieving 100% accuracy is a challenging task, thereby making it an active research topic among researchers. In this paper, an extensive study and review of the existing Deep Learning (DL) based techniques used for Semantic Segmentation is carried out along with a summary of the datasets and evaluation metrics used for it. The study involved the meticulous selection of relevant research papers in the field of interest by search based on several defined keywords. The study begins with a general and broader focus on Semantic Segmentation as a problem and further narrows its focus on existing Deep Learning (DL) based approaches for this task. In addition to this, a summary of the traditional methods used for Semantic Segmentation is also presented. The contents of this study are organized to provide ease of access to the relevant literature available for the problem of Semantic Segmentation, with a concentrated focus on DL-based methods. Since the problem of scene understanding is being vastly explored by the computer vision community, especially with the help of Semantic Segmentation, we believe that this study will benefit active researchers in reviewing and studying the existing state-of-the-art, as well as advanced methods for the same.

Author Keywords. Semantic. Segmentation. Image. Deep Learning. Scene Understanding.

Type: Research Article
Open Access Peer Reviewed CC BY

1. Introduction

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

With this study, our focus is on analyzing Semantic Segmentation approaches for scene understanding. Semantic Segmentation is a process of assigning a meaningful label to each pixel based on the context of the environment (Lateef and Ruichek 2019). It is a very useful step for a variety of computer vision applications, where it is important to understand the context of the operating environment, for e.g., robotics (Kim and Seok 2018), self-driving cars (Kaymak and Uçar 2019), etc. Scene understanding is a computer vision problem that contains the process of interpreting a scene captured through devices like cameras, microphones, contact sensors, etc., to get an in-depth understanding of them (Aarthi and Chitrakala 2017; Xiao et al. 2013). A scene shows a real-world situation that 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. The data of a scene can be expressed using various features like color, texture, light intensity, etc. thus, the process of creating a good understanding of a scene requires proper extraction of features from an image 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 them accordingly. It can adapt to unforeseen data and perform robustly in such situations (Li, Socher, and Fei-Fei 2009).

Scene understanding has applications in various fields. It is used in the medical field for medical image analysis which includes getting clinically meaningful information from the image, where the extracted data can be used by doctors for diagnosis (Ker et al. 2017). Scene understanding also has its application in road detection and urban scene understanding (Brust et al. 2015), in which objects in images are classified and labeled, which is then used for detecting roads and understanding urban scenes. Scene understanding also has application 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 et al. 2016). Researchers are implementing deep network models in every possible field, including image classification (Lee et al. 2018), object detection (Verschae and Ruiz-del-Solar 2015), image generation, etc. Deep learning allows us to model the high-level features of an image into compact representations for efficient manipulations, as well as analysis of the input images (Garcia-Garcia et al. 2018). As scene understanding is a complex problem involving several sub-tasks such as object detection, Semantic Segmentation, etc., deep learning models can efficiently handle these tasks.

In this paper, a systematic study on Semantic Segmentation is presented. 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 a focus on Semantic Segmentation for scene understanding is considered for analyzing the various deep learning models for Semantic Segmentation. Furthermore, different benchmark datasets along with evaluation metrics are also presented.

The study presented in this paper done for providing the following key contributions:

- An extensive study of the traditional as well as deep learning-based techniques employed for the task of Semantic Segmentation is presented.
- 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 few metrics valuable for evaluating the performance of different techniques/models are presented.

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

2. Background: From Segmentation to Semantic Segmentation

2.1. Image segmentation

For the analysis of an image, image segmentation is a fairly 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 that 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 to identify and find objects and boundaries in digital images (Ripon et al. 2017). Thus, it is concluded that image segmentation is the process of attributing a label to each pixel that holds certain similar characteristics like texture, color or intensity, etc. Segmentation acts as a reliable transformation technique that determines the success of analyzing an image; however, it is a challenging task to obtain a precise partitioning of an image. Some of the popular traditional approaches for performing segmentation, as well as their applications, are briefly discussed below.

2.1.1.Thresholding

Thresholding is a process of segmenting an image by setting a threshold value and comparing all the image pixels with the set threshold value. This method segments the object from the background by setting all the pixels having a value less than the threshold to one value (maybe white) and all the pixels having a value greater than the threshold to another value (maybe black). This method gives the best results for high-contrast images. As in the thresholding-based approach, setting a proper threshold value is the most important step, a lot of work has been done to automatically extract the optimum threshold value. Two methods for automatic threshold selection using an approximation of histogram are presented by Ramesh, Yoo, and Sethi (1995). Here one method determines the threshold by minimizing the sum of the square error, while the other method minimizes the variance of the histogram. Another method proposed by 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 the measurement of bright and dark areas, which defines each pixel in a region in terms of its membership value. To date, researchers have explored thresholding in various fields, such as the approach used for detecting cancer by segmenting the images using a combination of fuzzy entropy and thresholding on medical images (Maolood, Al-Salhi, and Lu 2018). Researchers have also tried to combine thresholding with other image processing methods;

for e.g., in Al-Azawi (2013), authors combined fuzzy-based image processing with a histogram thresholding technique for image segmentation.

2.1.2. Edge-based Segmentation

Another very common approach for segmentation is edge-based segmentation. Edges are discontinuity in the pixel values, which are identified from the differences in pixel values in two adjacent regions. This discontinuity helps in identifying the shapes of objects in the image. Edges can be identified using filters and convolutions on the image matrix; some of the common filters used for edge detection are the Sobel operator and Robert cross operator (Karthicsonia and Vanitha 2019). Edge detection methods are used in image segmentation and object recognition (Ramadevi et al. 2010) and to identify abnormalities in the images, especially medical images. In Padmapriya, Kesavamurthi, and Ferose (2012), the use of an edge-based segmentation approach is presented to determine the thickness of the urinary bladder wall. The method projected is used to collect information about bladder abnormalities and the extent of abnormalities.

2.1.3. Region-based segmentation: approach

Region-based segmentation is another approach of segmentation that extracts region-based features from the images to define different classes. This method is very useful in noisy images where edges cannot be identified (Lalaoui and Mohamadi 2013). Two famous approaches for region-based segmentation are splitting & merging and region growing. In the former approach, a uniformity criterion 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 the merging of adjacent regions based on the same uniformity criteria. The region-growing method starts by defining a seed region which can be a single pixel or a block of pixels. The neighbors 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 with another region. An extensive review of various region-based segmentation methods can be found in Lalaoui and Mohamadi (2013). Region-based segmentation is often used for identifying tumors, veins, etc., in medical images, for finding targets in aerial images and for finding people in surveillance images, etc. Gould, Gao, and Koller (2009) presented a region-based approach that combines object detection and segmentation, which performs background classification based on pixel features and object detection using a representation of regions. The model defined here gives a unified description of the scene depicted in the image.

2.2. Semantic Segmentation

Semantic Segmentation 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 and Ruichek 2019; Kim and Seok 2018). Meaningful segmentation allows the progression from low-level or crude image processing transformations, involving conversions of greyscale or color images into several other images to high-level image description creation concerning features, objects, layouts, and scenes (Liu, Deng, and Yang 2019; Gupta et al. 2015). Semantic Segmentation techniques can be 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, Deng, and Yang 2019). While many modern researchers are focusing their work on deep neural networks, some modern researchers are also combining traditional methods with new concepts (Xiao et al. 2012; Guo et al. 2016). The researchers are trying to improve the accuracy by enhancing traditional methods with different concepts like fuzzy logic (Guo et al. 2016). Some of the popular traditional methods for performing Semantic Segmentation are discussed below.

2.2.1.Features and classification based segmentation

Features play an important role in the analysis of an image and in performing meaningful segmentation on an image. Until now, a range of features has been explored by researchers, including color, texture, Histogram of Oriented Gradients (HoG), scale-invariant feature transforms, SURF, and many more (Zaitoun and Aqel 2015; Lateef and Ruichek 2019; Kim and Seok 2018). Image segmentation based on features referred to as visual descriptors can be found in Ripon et al. (2017), where the extracted features are used for generating a classification model 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 in segmenting 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 center of these clusters is then calculated and the distance of each pixel from these centers is also calculated. This process is widely used with small datasets. The use of K-means for image segmentation is presented in Shan (2018). A color-based segmentation method using K-means clustering is proposed in Muthukannan and Moses (2010), where the 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. Furthermore, various supervised classification techniques were also explored in the literature for Semantic Segmentation (Sharma et al. 2008; Savkare and Narote 2012; Sakthivel, Nallusamy, and Kavitha 2014; Wang, Wang, and Bu 2011). A text features-based medical image segmentation is presented in Sharma et al. (2008), where extracted features are used to design ANN-based classifiers to classify soft tissues. Furthermore, the use of an SVM pixel classifier for image retrieval, object detection, and medical imaging can be seen in Savkare and Narote (2012), where the SVM classifier is used to classify malaria-infected erythrocytes, which then helps in the detection of parasite life stages. Furthermore, the problem of object detection and Semantic Segmentation for indoor scene understanding have also been explored by Gupta et al. (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. A combination of different approaches was explored by researchers in Sakthivel, Nallusamy, and Kavitha (2014), where Support Vector Machines 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 the SVM classifier and pixel-level information are combined to return better results, thus improving the quality of image segmentation.

2.2.2. Markov Random Network (MRF) and Conditional Random Field (CRF) based Segmentation

The conditional random field is another method used for segmentation. This method is used where contextual information affects the prediction. It helps to work on data where the label classes are dependent on each other for e.g., the class label for a pixel depends on the label of its neighbor 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, McCallum, and Pereira 2001). In He and Kayaalp (2008), the authors present a conditional random fields framework, which explains the framework along with its comparison with Hidden Markov models and maximum entropy Markov models. In Lafferty, McCallum, and Pereira (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 and Triggs (2007), authors used CRF for scene segmentation, where CRFs partition an image into semantic-level regions and assign the class labels to these regions. Here, the model combines the local features and the features associated with a larger section for semantic image labeling.

3. Deep Learning for Semantic Segmentation

3.1. 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 et al. 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 et al. 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. Autoencoders are a class of ANNs that are used to learn data coding in an unsupervised fashion. GANs also follow an unsupervised learning method wherein two neural networks are improving each other’s performance by contesting with each other.

In this section, we are presenting 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 and its use for computer vision, the reader is referred to Guo et al. (2016).

3.2. Deep Neural 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 Le Cun et al. (1990) in the year 1990 for the recognition of handwritten digits. However, they seemed to gain popularity after the introduction of AlexNet by Krizhevsky, Sutskever, and Hinton (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 in the following subsection.

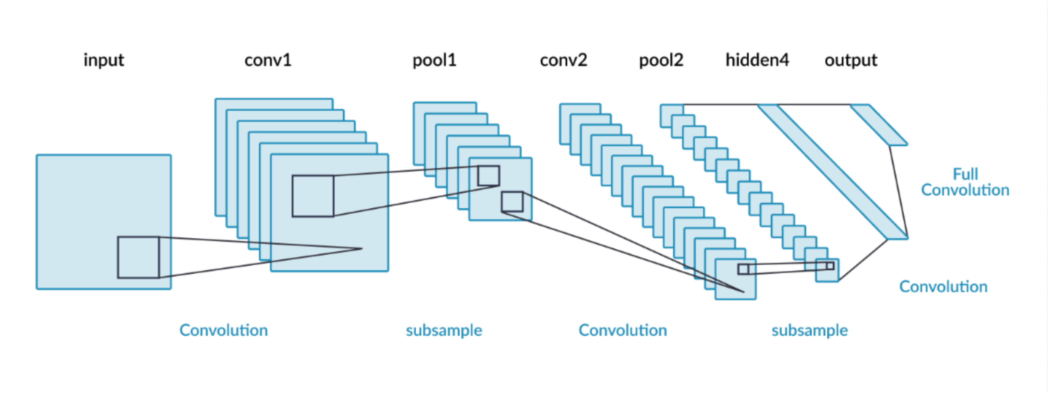


Figure 1: The basic architecture of a convolutional neural network consists of two convolution layers, two pooling layers, and two fully connected layers (Krizhevsky, Sutskever, and Hinton 2017)

3.2.1.Fully Convolutional Network (FCN)

FCN was brought into existence by Long, Shelhamer, and Darrell (2015). In this network, some exclusive convolutional layers were incorporated for performing Semantic Segmentation. The network was designed such that when an image of random size is fed to the FCN, a finally segmented image of the same size is then 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, and 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 up-sampling to produce a similar-sized image as the input. Up-sampling involves convolutions with strides less than one. It is sometimes known as deconvolution as it results in input having a smaller size than output. Using this method, the network is then trained using the concept of pixel loss. Additionally, several skip connections were introduced in this network to connect high-level feature-mapped representation to highly precise concentration at the top of the model.

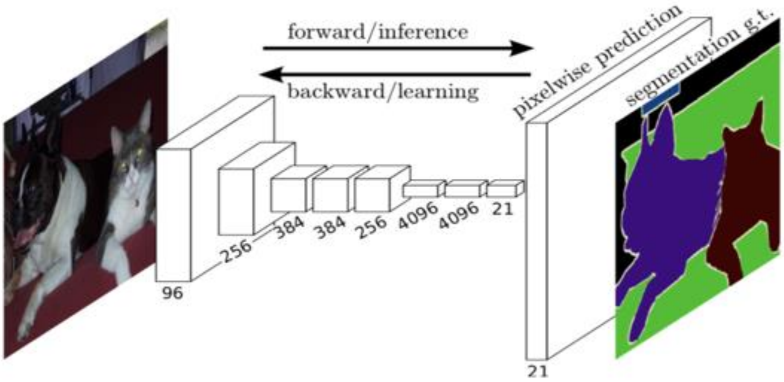


Figure 2: The architecture of the Fully Convolutional Network (Long, Shelhamer, and Darrell 2015)

3.2.2. ParseNet

ParseNet was created as an improvement to the Fully Convolutional Network model proposed by Liu, Rabinovich, and Berg (2015). It was observed that the FCN model does not consider the global context of the image as it further goes into the deeper layers by focusing on details in the produced feature mappings. ParseNet presents an exclusively convolutional network that predicts values for each pixel simultaneously and does not take regions as inputs to preserve the global context and information of the image. A module is used to take feature mappings as the input and the initial course of action makes use of a model to produce feature mappings that are condensed to just one globally accessible feature vector with a singular pooling layer. It is this vector that undergoes the process of normalization using the L2 Euclidean Norm and is further expanded or un-pooled to generate novel feature mappings of equal size to the original. The next step involves the L2-normalization of all the initial feature maps. Finally, the last step deals with the concatenation of feature mappings generated by the last 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 for the aforementioned module substitutes for the convolutional layers, as presented in Figure 3.

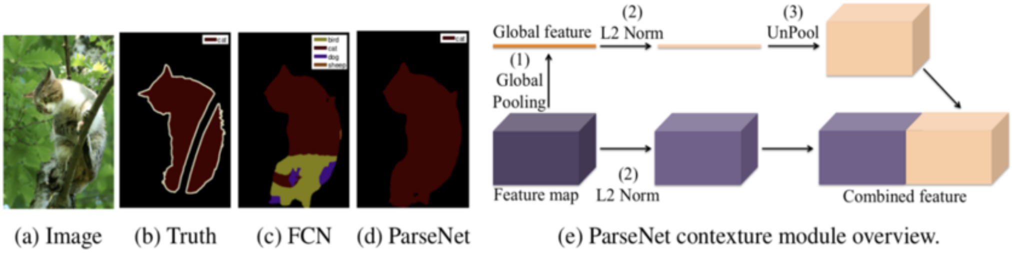


Figure 3 (a-d): Comparison between the FCN and ParseNet Output, (e) Architecture of ParseNet (Liu, Rabinovich, and Berg 2015)

3.2.3. Convolutional and deconvolutional networks

This end-to-end network comprises 2 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 to retain the information located 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 (Ronneberger, Fischer, and Brox 2015; Badrinarayanan, Kendall, and Cipolla 2017).

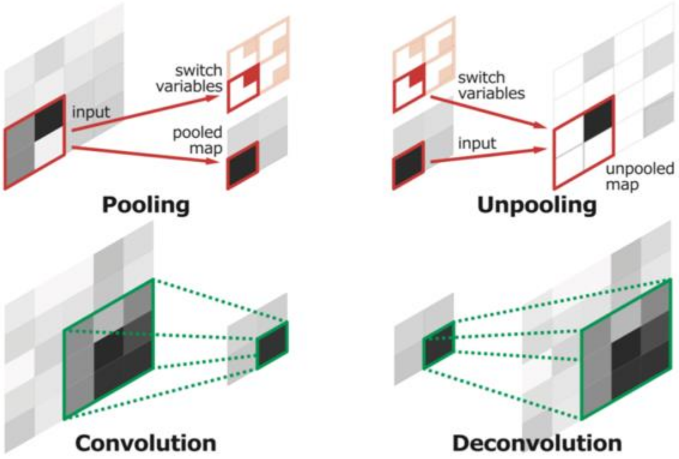


Figure 4: Visualization of convolutional and deconvolutional layers

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

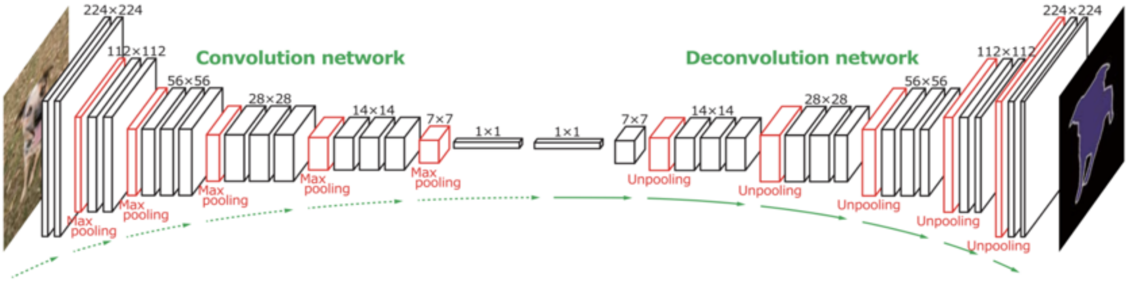


Figure 5: The architecture of the Convolutional and Deconvolutional Network (Badrinarayanan, Kendall, and Cipolla 2017)

3.2.4. U-Net

U-Net was created as an extension of the Fully Convolutional Network by Ronneberger, Fischer, and Brox (2015), mainly to cater to biological microscopy images. It consists of two parts, first is the contracting part, which works out features & the second is expanding part, which localizes the spatial patterns in the image, as presented in Figure 6. The contracting part, also known as down-sampling, 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 down-sampling part of the network are duplicated in the up-sampling 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 been greatly extended for its use in other recent architectures. It is worth noting that this model does not employ any fully connected layers and thus have a lessened number of parameters which makes it more applicable for smaller dataset of images.

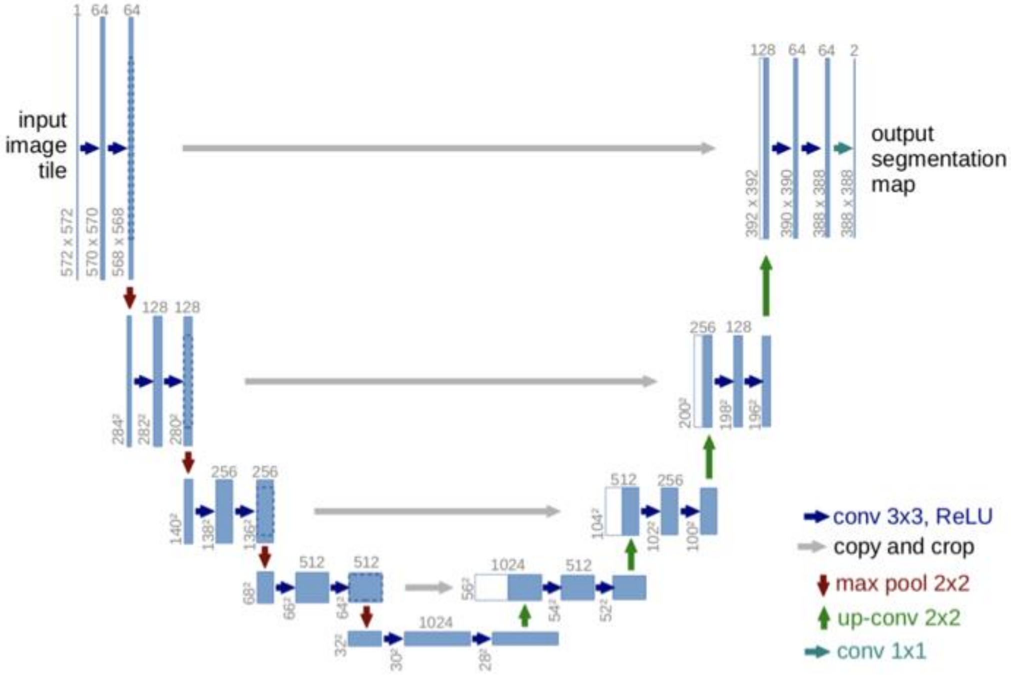
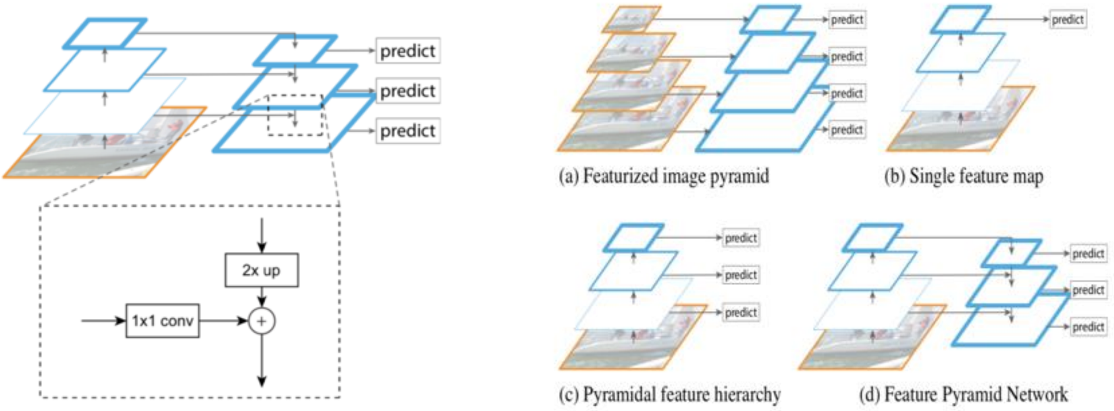


Figure 6: Architecture of U-net (Ronneberger, Fischer, and Brox 2015)

3.2.5.Feature Pyramid Network (FPN)

The FPN was created by Lin et al. (2017). 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 of this image is done using convolutional layers, which are followed by down-sampling 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 is the features utilized for the pyramid level. The top-down pipeline involves the up-sampling of final feature mappings along with their un-pooling. The un-pooling is done by modifying them with feature mappings obtained from the bottom-up pathway using 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.



(a) Featurized image pyramid

(b) Single feature map

(c) Pyramidal feature hierarchy

(d) Feature Pyramid Network

Figure 7: Detailed top-down pathway process with horizontal connections (Lin et al. 2017)

4. Benchmark Datasets and Evaluation Metrics for Semantic Segmentation

4.1. 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 that are representative 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.

4.1.1.Stanford Background Dataset (Gould, Fulton, and Koller 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, as 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 the sky, horizontal and vertical (Li, Socher, and Fei-Fei 2009). Samples of the dataset are shown in Figure 8.

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.

Table 1: Labels of Stanford background dataset (Gould, Fulton, and Koller 2009)

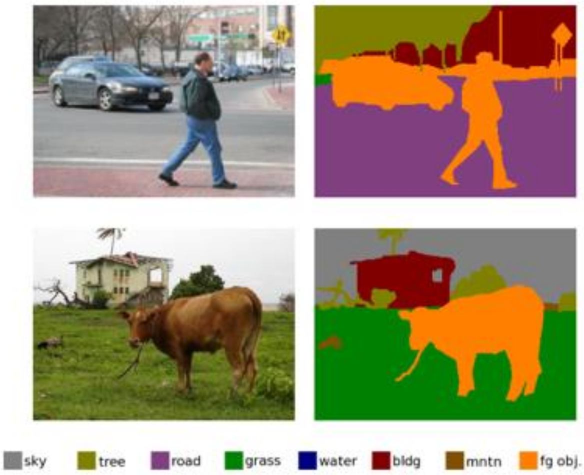
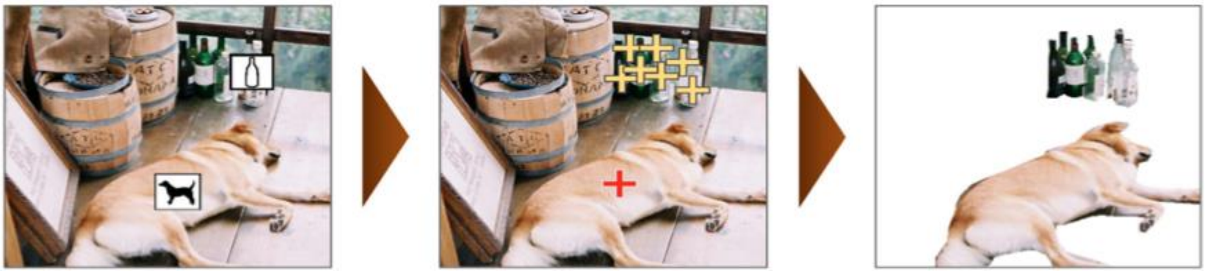


Figure 8: Example of images and semantic labels in the Stanford background dataset

4.1.2.Microsoft COCO dataset (2015 version) (Lin et al. 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.



(a) (b) (c)

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

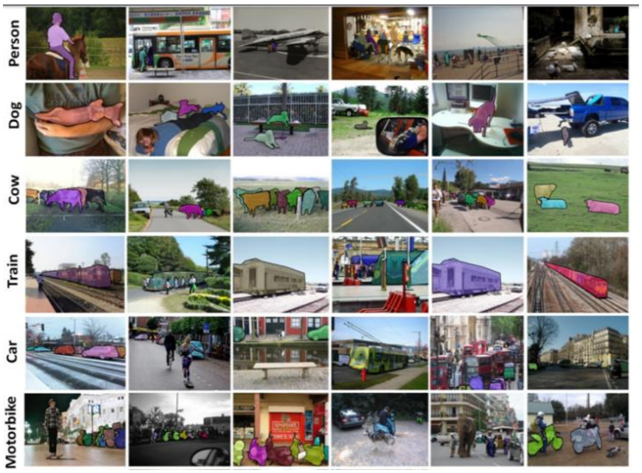


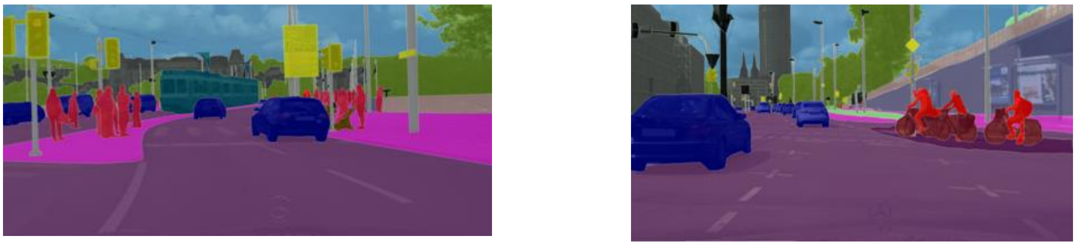
Figure 10: Example of images and classes in the COCO dataset (Lin et al. 2014)

4.1.3. Cityscapes Dataset (Cordts et al. 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. Here * represents that the annotations based on a single instance are done, whereas + denotes that the mentioned label is not included for any kind of evaluation and is thus treated as void. 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 a special focus on those consisting of an enormous number of dynamic objects and variations in layouts of the scene and 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.

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+

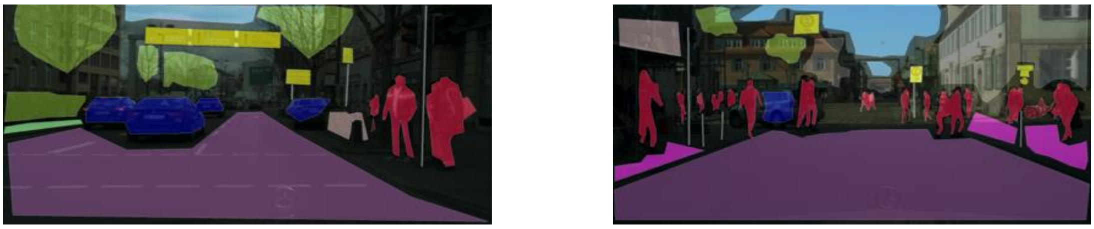
Table 2: Classes present in the Cityscapes Dataset under their respective groups



(a)

(b)

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



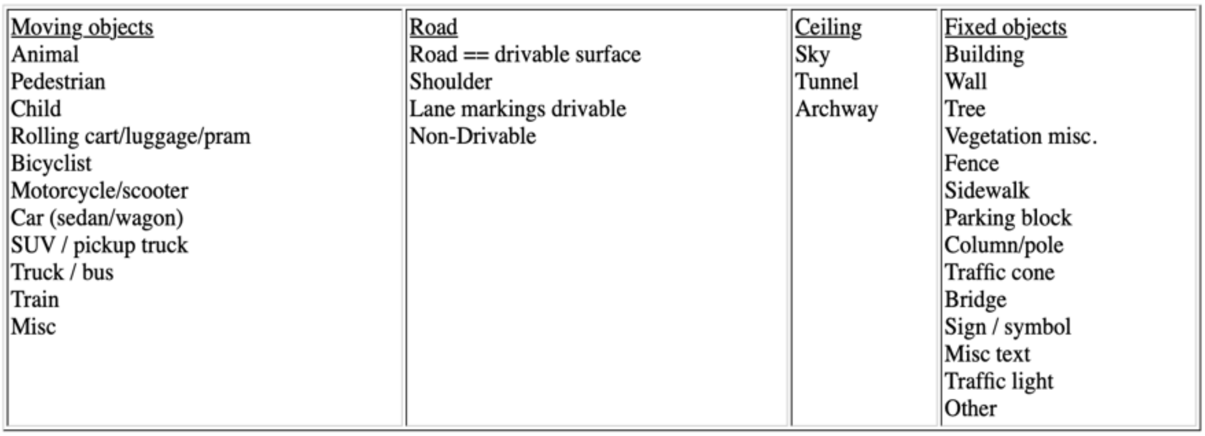
(a)

(b)

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

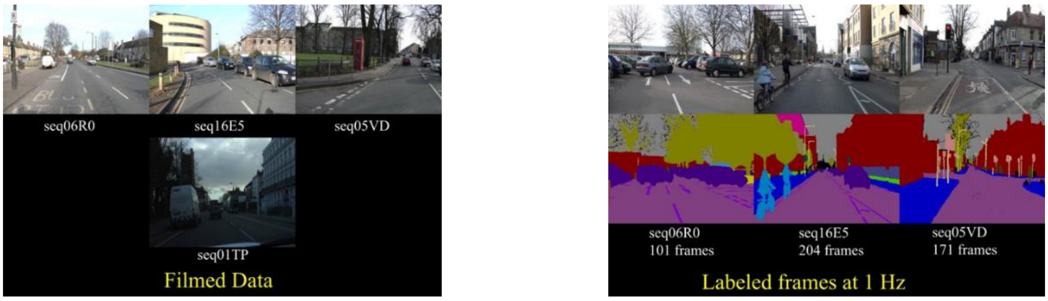
4.1.4.CamVid Dataset (Brostow, Fauqueur, and Cipolla 2009)

CamVid is short for Cambridge-driving Labeled Video Dataset, which 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 class sets. It consists of 700 training and manually-annotated images of urban scenes. Some of the semantic classes used in this dataset are mentioned in Table 3 and sample data are presented in Figure 13.



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

Table 3: Semantic classes in CamVid Dataset



(a) (b)

Figure 13: (a) Filmed data as recorded in the CamVid Dataset (b) Annotated frames in the CamVid dataset (Brostow, Fauqueur, and Cipolla 2009)

4.1.5.KITTI Semantic Segmentation Benchmark (2018 version) (Abu Alhaija et al. 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 Abu Alhaija et al. (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.

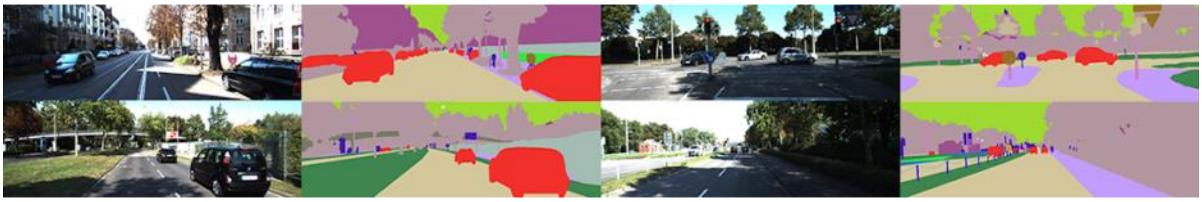


Figure 14: Filmed and segmented frames in the KITTI dataset (Abu Alhaija et al. 2018)

4.1.6. NYUDv2 (Silberman et al. 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 that has useful functions for dealing with the images and labels. Figure 15 shows some samples from the dataset.

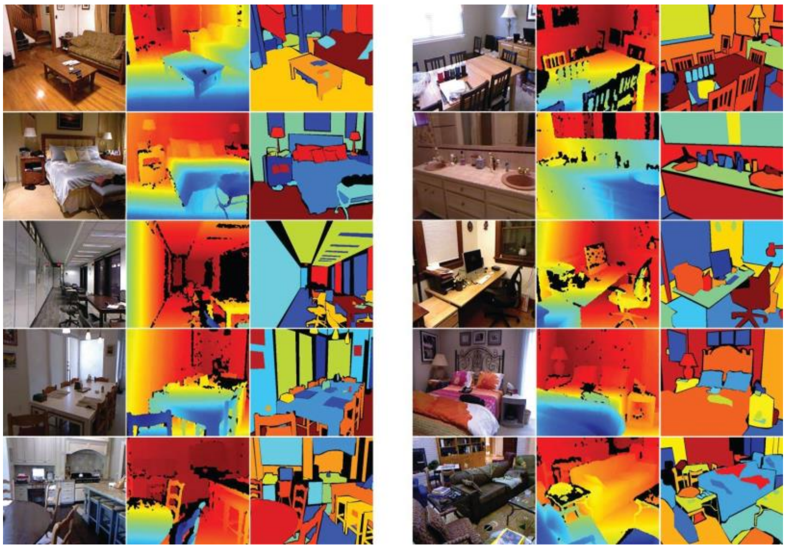


Figure 15: Samples of the RGB images, raw depth images and segmented images (Silberman et al. 2012)

4.1.7. PASCAL VOC 2012 (Everingham et al. 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 into Person, Animal, Vehicle and Indoor.

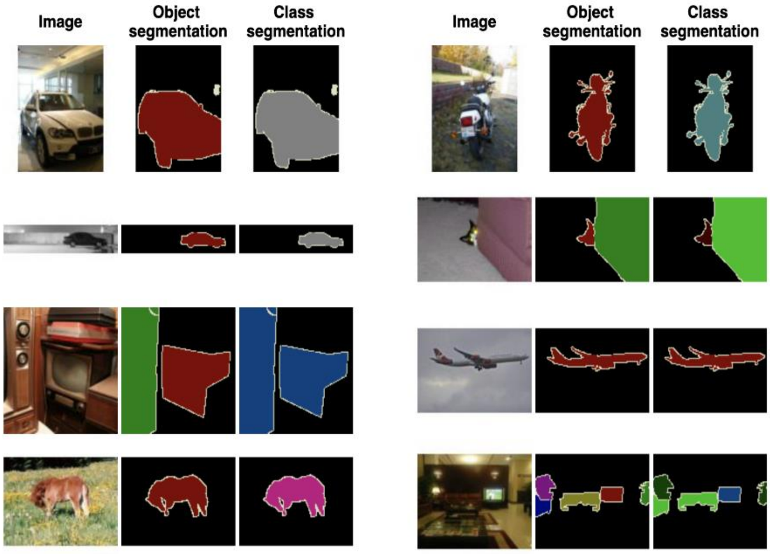


Figure 16: The training image, object segmentation and class segmentation examples from the PASCAL VOC dataset (Everingham et al. 2010)

4.1.8.SUN Database (Xiao et al. 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.

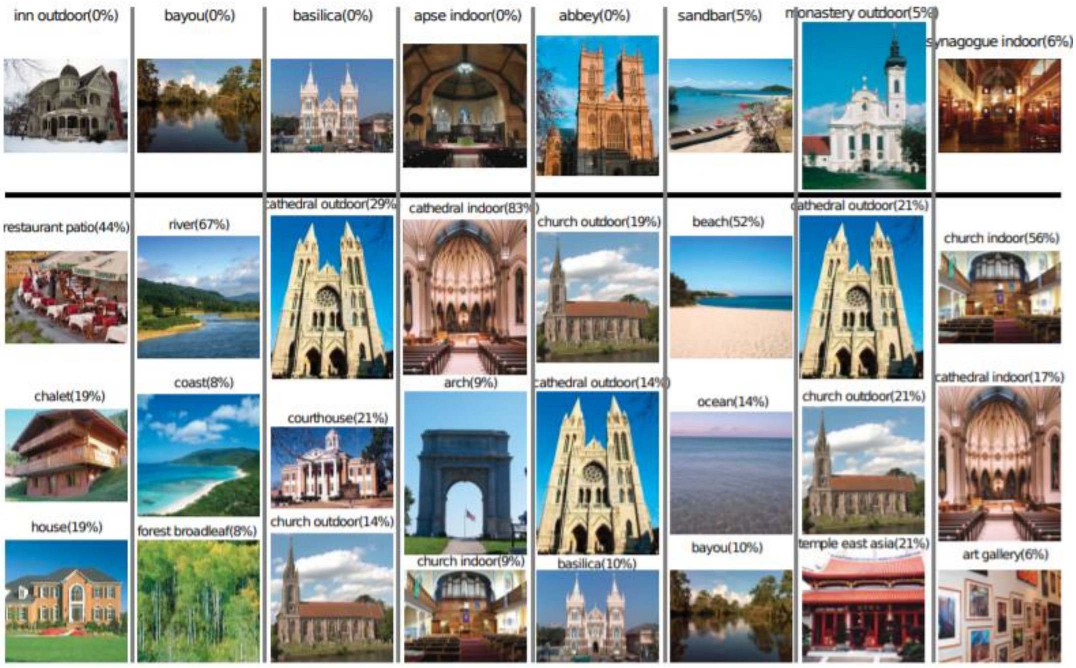


Figure 17: Images of some SUN categories, with the percentage of human recognition rate mentioned (Xiao et al. 2010)

4.1.9.SYNTHIA (Ros et al. 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.

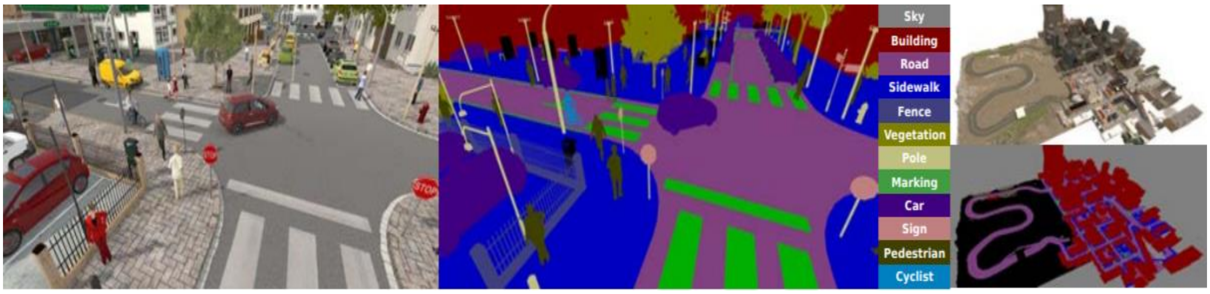


Figure 18: Sample frame from SYNTHIA dataset (Left) with its semantic labels (center) and a general view of the city (right) (Ros et al. 2016)

4.1.10. LabelMe (Russell et al. 2008)

LabelMe is a database and an online annotation tool that provides functionalities like querying images, browsing databases, 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 shows the object-part relationship using polygon annotations.

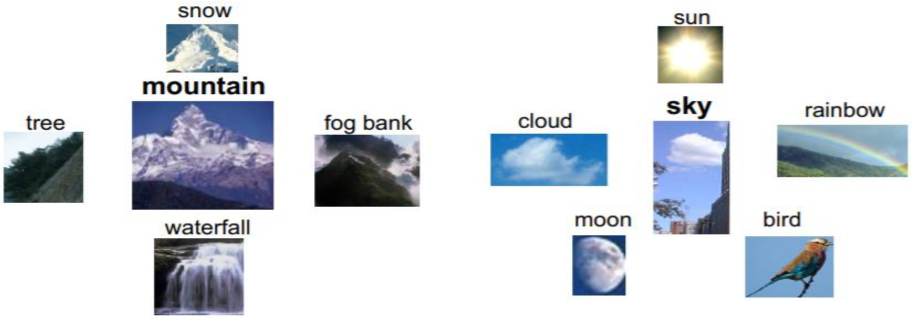


Figure 19: Sample images in the dataset, object is in the center of its parts (Russell et al. 2008)

4.2. Datasets comparative summary

Here, we present a summarized evaluation of all large-scale and benchmark datasets designed for analyzing the Semantic Segmentation and scene understanding algorithms. Important parameters and the design choices that were kept in mind with regard to the focus of the dataset are summarized in Table 4.

Dataset Name	**Purpose**	**Year**	**Classes**	**Data**	**Synthetic/ Real**	**Samples**
Stanford Background (Gould, Fulton, and Koller 2009)	Outdoor	2009	8	2D	Real	715 (572 training images and 143 test images)
COCO (Lin et al. 2014)	General	2015	91	2D	Real	328,000 images 165,482(training images) 81,208(validation images) 81,434(test images)
Cityscape (Cordts et al. 2016)	Urban	2016	30	2D	Real	2975 (training) 1525 (testing) 500 (validation)
CamVid (Brostow, Fauqueur, and Cipolla 2009)	Urban/ Driving	2008	32	2D	Real	700 (training)
KITTI (Abu Alhaija et al. 2018)	Urban/Driving	2018	34	2D	Real	200 (training) 200 (testing)
NYUDv2	Indoor	2012	40	2.5D	Real	1449 795 (training)

(Silberman et al. 2012)						654 (validation)
PASCAL VOC 2012 (Everingham et al. 2010)	General	2012	21	2D	Real	11,530 1464 (training) 1449 (validation)
SUN (Xiao et al. 2010)	Outdoor scenes	2010	899	2D	Real	130,519
SYNTHIA (Ros et al. 2016)	Urban/ Driving	2016	13	2D	Synthetic	13400
LabelMe (Russell et al. 2007)	General	2006	183	2D	Real	111490

Table 4: Benchmark and large-scale datasets for Semantic Segmentation and scene understanding

4.3. 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 that 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 et al. 2018; Liu, Deng, and Yang 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 be measured by finding the ratio of correctly classified pixels (regardless of class) to the total number of classes. For a particular class, pixel accuracy can be calculated using Equation (1).

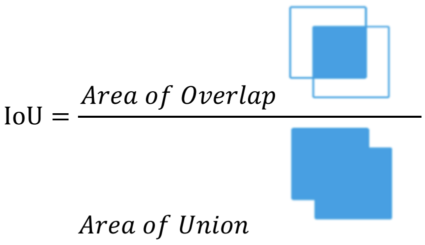
$$P_{acc} = \frac{\sum n_{ii}}{\sum t_i} \qquad \textbf{(1)}$$

Where $\boldsymbol{n_{ii}}$ are the number of pixels class i predicted belong to class i and $\boldsymbol{t_i}$ 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 et al. 2016).

$$\text{IoU} = \frac{Target \ \cap \ Prediction}{Target \ \cup \ Prediction} \qquad \textbf{(2)}$$

This can easily be understood from the visualization below:

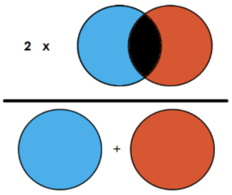


- **Weighted IoU:** Weighted IoU can also be measured by taking the average IoU of each class weighted by the number of pixels in that class. This metric is useful when images have imbalanced classes.

- **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, Deng, and Yang 2019).

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

The following illustration makes it easy to understand:



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

5. 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 are explained below.

A general review of scene understanding can be found in Aarthi and Chitrakala (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 a better understanding of the scene understanding process as well as to present a comparison between several state-of-the-art strategies, including this parameter. Furthermore, an extensive review of deep learning techniques for Semantic Segmentation is given by Guo et al. (2018), Liu, Deng, and Yang (2019) and Garcia-Garcia et al. (2018). In Garcia-Garcia et al. (2018), 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. Liu, Deng, and Yang (2019) 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 that are sought after in this domain. It also highlights some challenges faced by researchers in the usage of these datasets to encourage the reader to make informed decisions about selecting one that is most suitable to their requirements and goals.

Furthermore, 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 state-of-the-art performance at the time for the task of Semantic Segmentation. In Liu, Rabinovich, and Berg (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 that 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, Fischer, and Brox (2015), where the contracting part was responsible for feature and context mapping, whereas the expanding part was used for accurate localization. A deep convolution network-based architecture SegNet is described in Badrinarayanan, Kendall, and Cipolla (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 this paper, an analysis of SegNet is done, and the network is compared with other segmentation architectures which share the same encoder but different decoder. 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. Lin et al. (2017) 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. Furthermore, a fully convolutional neural network architecture called BlitzNet is proposed by Dvornik et al. (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 et al. 2014) and PASCAL VOC (Everingham et al. 2010) datasets. A novel method is proposed by Li et al. (2017) for the task of scene understanding by modeling 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 by Xiao et al. (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 concept parsing based on the annotations of the datasets. In Zhang et al.

(2018), the authors presented 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 et al. (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 the segmentation of objects at different scales. Thirdly, deep convolutional neural networks are combined with a fully connected Conditional Random Field to improve the localization performance of object boundaries. Valada, Mohan, and Burgard (2020) introduce a multimodal approach to the problem of Semantic Segmentation along with proposing a unimodal network called AdaptNet++ for computationally efficient performance. Furthermore, the comparative analysis of different approaches is summarized in Table 5 (see Appendix A).

6. Conclusion

The method of Semantic Segmentation for scene understanding is gaining immense popularity due to its efficiency in obtaining the correct classification for each pixel of the image, which further makes it easy for the image to be semantically understood. 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, the authors achieved the following key points:

- A thorough background of segmentation to Semantic Segmentation is presented for a 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 in 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.
- Besides a brief review of traditional approaches, a comprehensive review of recent progress on deep learning-based Semantic Segmentation is also presented.

References

Aarthi, S., and S. Chitrakala. 2017. "Scene understanding - A survey". In *International Conference on Computer, Communication, and Signal Processing: Special Focus on IoT, ICCCSP 2017*, 1-4. IEEE. https://doi.org/10.1109/ICCCSP.2017.7944094.

Abu Alhaija, H., S. K. Mustikovela, L. Mescheder, A. Geiger, and C. Rother. 2018. "Augmented reality meets computer vision: Efficient data generation for urban driving scenes".

International Journal of Computer Vision 126, no. 9: 961-72. https://doi.org/10.1007/s11263-018-1070-x.

Al-Azawi, M. A. N. 2013. "Image thresholding using histogram fuzzy approximation". *International Journal of Computer Applications* 83, no. 9: 36-40. https://doi.org/10.5120/14480-2781.

Badrinarayanan, V., A. Kendall, and R. Cipolla. 2017. "SegNet: A deep convolutional encoder-decoder architecture for image segmentation". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, no. 12: 2481-95. https://doi.org/10.1109/TPAMI.2016.2644615.

Brostow, G. J., J. Fauqueur, and R. Cipolla. 2009. "Semantic object classes in video: A high-definition ground truth database". *Pattern Recognition Letters* 30, no. 2: 88-97. https://doi.org/10.1016/j.patrec.2008.04.005.

Brust, C. A., S. Sickert, M. Simon, E. Rodner, and J. Denzler. 2015. "Convolutional patch networks with spatial prior for road detection and urban scene understanding". In *Proceedings of the 10th International Conference on Computer Vision Theory and Applications - Volume 3: VISAPP*, 510-17. https://doi.org/10.5220/0005355105100517.

Chen, L. C., G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. 2018. "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, no. 4: 834-48. https://doi.org/10.1109/TPAMI.2017.2699184.

Cordts, M., M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. 2016. "The cityscapes dataset for semantic urban scene understanding". In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3213-23. IEEE. https://doi.org/10.1109/CVPR.2016.350.

Dvornik, N., K. Shmelkov, J. Mairal, and C. Schmid. 2017. "BlitzNet: A real-time deep network for scene understanding". In *Proceedings of the IEEE International Conference on Computer Vision*, 4174-82. https://doi.org/10.1109/ICCV.2017.447.

Everingham, M., L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2010. "The pascal visual object classes (VOC) challenge". *International Journal of Computer Vision* 88, no. 2: 303-38. https://doi.org/10.1007/s11263-009-0275-4.

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

Gould S., T. Gao, and D. Koller. 2009. "Region-based segmentation and object detection". In *Proceedings of the 22nd International Conference on Neural Information Processing Systems (NIPS'09)*, 655-63. Curran Associates Inc., Red Hook, NY, USA: ACM.

Gould, S., R. Fulton, and D. Koller. 2009. "Decomposing a scene into geometric and semantically consistent regions". In *Proceedings of the IEEE International Conference on Computer Vision*, 1-8. IEEE. https://doi.org/10.1109/ICCV.2009.5459211.

Guo, Y., Y. Liu, T. Georgiou, and M. S. Lew. 2018. "A review of semantic segmentation using deep neural networks". *International Journal of Multimedia Information Retrieval* 7, no. 2: 87-93. https://doi.org/10.1007/s13735-017-0141-z.

Guo, Y., Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew. 2016. "Deep learning for visual understanding: A review". *Neurocomputing* 187: 27-48. https://doi.org/10.1016/j.neucom.2015.09.116.

Gupta, S., P. Arbeláez, R. Girshick, and J. Malik. 2015. "Indoor scene understanding with RGB-D images: Bottom-up segmentation, object detection and semantic segmentation". *International Journal of Computer Vision* 112, no. 2: 133-49. https://doi.org/10.1007/s11263-014-0777-6.

He, Y., and M. Kayaalp. 2008. "Biological entity recognition with conditional random fields". *Annual Symposium proceedings / AMIA Symposium*: 293-97.

Karthicsonia, B., and M. Vanitha. 2019. "Edge based segmentation in medical images". *International Journal of Engineering and Advanced Technology* 9, no. 1: 449-51. https://doi.org/10.35940/ijeat.A9484.109119.

Kaymak, Ç., and A. Uçar. 2019. "Semantic image segmentation for autonomous driving using fully convolutional networks". In *2019 International Conference on Artificial Intelligence and Data Processing Symposium, IDAP 2019*, 1-8. IEEE. https://doi.org/10.1109/IDAP.2019.8875923.

Ker, J., L. Wang, J. Rao, and T. Lim. 2017. "Deep learning applications in medical image analysis". *IEEE Access* 6: 9375-79. https://doi.org/10.1109/ACCESS.2017.2788044.

Kim, W., and J. Seok. 2018. "Indoor semantic segmentation for robot navigating on mobile". In *International Conference on Ubiquitous and Future Networks, ICUFN*, 22-25. IEEE. https://doi.org/10.1109/ICUFN.2018.8436956.

Krizhevsky, A., I. Sutskever, and G. E. Hinton. 2017. "ImageNet classification with deep convolutional neural networks". *Communications of the ACM* 60, no. 6: 84-90. https://doi.org/10.1145/3065386.

Lafferty, J., A. McCallum, and F. C.N. Pereira. 2001. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data". In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01)*, 282-89. https://dl.acm.org/doi/10.5555/645530.655813.

Lalaoui, L., and T. Mohamadi. 2013. "A comparative study of Image Region-Based Segmentation Algorithms". *International Journal of Advanced Computer Science and Applications* 4, no 6: 198-206. https://doi.org/10.14569/IJACSA.2013.040627.

Lateef, F., and Y. Ruichek. 2019. "Survey on semantic segmentation using deep learning techniques". *Neurocomputing* 338: 321-48. https://doi.org/10.1016/j.neucom.2019.02.003.

Le Cun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. 1990. "Handwritten digit recognition with a back-propagation network". In *Advances in Neural Information Processing Systems 2 (NIPS 1989)*, 396-404.

Lee, S. J., T. Chen, L. Yu, and C. H. Lai. 2018. "Image classification based on the boost convolutional neural network". *IEEE Access* 6: 12755-68. https://doi.org/10.1109/ACCESS.2018.2796722.

Li, L. J., R. Socher, and L. Fei-Fei. 2009. "Towards total scene understanding: Classification, annotation and segmentation in an automatic framework". In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2036-43. IEEE. https://doi.org/10.1109/CVPR.2009.5206718.

Li, Y., W. Ouyang, B. Zhou, K. Wang, and X. Wang. 2017. "Scene graph generation from objects, phrases and region captions". In *2017 IEEE International Conference on Computer Vision (ICCV)*, 1270-79. IEEE. https://doi.org/10.1109/ICCV.2017.142.

Lin, T.-Y., M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. 2014. "Microsoft COCO: Common objects in context". In *Computer Vision – ECCV 2014*, 740-55. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-10602-1_48.

Lin, T. Y., P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. 2017. "Feature pyramid networks for object detection". In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 936-44. IEEE. https://doi.org/10.1109/CVPR.2017.106.

Liu, W., A. Rabinovich, and A. C. Berg. 2015. "ParseNet: Looking wider to see better". Preprint, submitted January 15, 2015. [Computer Vision and Pattern Recognition]. https://arxiv.org/abs/1506.04579.

Liu, X., Z. Deng, and Y. Yang. 2019. "Recent progress in semantic image segmentation". *Artificial Intelligence Review* 52, no. 2: 1089-106. https://doi.org/10.1007/s10462-018-9641-3.

Long, J., E. Shelhamer, and T. Darrell. 2015. "Fully convolutional networks for semantic segmentation". In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 431-40. IEEE. https://doi.org/10.1109/CVPR.2015.7298965.

Lu, D., and Q. Weng. 2007. "A survey of image classification methods and techniques for improving classification performance". *International Journal of Remote Sensing* 28, no. 5: 823-70. https://doi.org/10.1080/01431160600746456.

Maolood, I. Y., Y. E. A. Al-Salhi, and S. Lu. 2018. "Thresholding for medical image segmentation for cancer using fuzzy entropy with level set algorithm". *Open Medicine* 13, no. 1: 374-83. https://doi.org/10.1515/med-2018-0056.

Messer, K. D., M. Costanigro, and H. M. Kaiser. 2017. "Labeling food processes: The good, the bad and the ugly". *Applied Economic Perspectives and Policy* 39, no. 3: 407-27. https://doi.org/10.1093/aepp/ppx028.

Muthukannan, K., and M. M. Moses. 2010. "Color image segmentation using k-means clustering and Optimal Fuzzy C-Means clustering". In *2010 International Conference on Communication and Computational Intelligence (INCOCCI)*, 229-34. IEEE. https://ieeexplore.ieee.org/document/5738735.

Padmapriya, B., T. Kesavamurthi, and H. W. Ferose. 2012. "Edge based image segmentation technique for detection and estimation of the bladder wall thickness". *Procedia Engineering* 30: 828-35. https://doi.org/10.1016/j.proeng.2012.01.934.

Ramadevi, Y., T. Sridevi, B. Poornima, and B. Kalyani. 2010. "Segmentation and object recognition using edge detection techniques". *International Journal of Computer Science & Information Technology* 2, no. 6: 153-61. https://doi.org/10.5121/ijcsit.2010.2614.

Ramesh, N., J. H. Yoo, and I. K. Sethi. 1995. "Thresholding based on histogram approximation". *IEE Proceedings: Vision, Image and Signal Processing* 142, no. 5: 271-79. https://doi.org/10.1049/ip-vis:19952007.

Ripon, K. S. N., S. Newaz, L. E. Ali, and J. Ma. 2017. "Bi-level multi-objective image segmentation using texture-based color features". In *2017 20th International Conference of Computer and Information Technology (ICCIT)*, 1-6. IEEE. https://doi.org/10.1109/ICCITECHN.2017.8281795.

Ronneberger, O., P. Fischer, and T. Brox. 2015. "U-net: Convolutional networks for biomedical image segmentation". In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 234-41. https://doi.org/10.1007/978-3-319-24574-4_28.

Ros, G., L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez. 2016. "The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes". In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3234-43. https://doi.org/10.1109/CVPR.2016.352.

Russell, B. C., A. Torralba, K. P. Murphy, and W. T. Freeman. 2008. "LabelMe: A database and web-based tool for image annotation". *International Journal of Computer Vision* 77, no. 1-3: 157-73. https://doi.org/10.1007/s11263-007-0090-8.

Sakthivel, K., R. Nallusamy, and C. Kavitha. 2014. "Color image segmentation using SVM pixel classification image". *World Academy of Science, Engineering and Technology, International Journal of Computer and Information Engineering* 8, no. 10: 1924-30. https://doi.org/10.5281/zenodo.1099796.

Savkare, S. S., and S. P. Narote. 2012. "Automatic system for classification of erythrocytes infected with malaria and identification of parasite's life stage". *Procedia Technology* 6: 405-10. https://doi.org/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, no. 1: Article number 81. https://doi.org/10.1186/s13640-018-0322-6.

Sharma, N., A. Ray, S. Sharma, K. Shukla, S. Pradhan, and L. Aggarwal. 2008. "Segmentation and classification of medical images using texture-primitive features: Application of BAM-type artificial neural network". *Journal of Medical Physics* 33, no. 3: 119-26. https://doi.org/10.4103/0971-6203.42763.

Silberman, N., D. Hoiem, P. Kohli, and R. Fergus. 2012. "Indoor segmentation and support inference from RGBD images". In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 746-60. https://doi.org/10.1007/978-3-642-33715-4_54.

Srinivas, S., R. K. Sarvadevabhatla, K. R. Mopuri, N. Prabhu, S. S. S. Kruthiventi, and R. V. Babu. 2016. "A taxonomy of deep convolutional neural nets for computer vision". *Frontiers in Robotics and AI* 2: Article 36. https://doi.org/10.3389/frobt.2015.00036.

Valada, A., R. Mohan, and W. Burgard. 2020. "Self-supervised model adaptation for multimodal semantic segmentation". *International Journal of Computer Vision* 128, no. 5: 1239-85. https://doi.org/10.1007/s11263-019-01188-y.

Verbeek, J., and B. Triggs. 2007. "Scene segmentation with conditional random fields learned from partially labeled images". In *Advances in Neural Information Processing Systems 20 - Proceedings of the 2007 Conference*, 1553-60. https://hal.inria.fr/inria-00321051v1.

Verschae, R., and J. Ruiz-del-Solar. 2015. "Object detection: Current and future directions". *Frontiers in Robotics and AI* 2: Article 29. https://doi.org/10.3389/frobt.2015.00029.

Wang, X. Y., T. Wang, and J. Bu. 2011. "Color image segmentation using pixel wise support vector machine classification". *Pattern Recognition* 44, no. 4: 777-87. https://doi.org/10.1016/j.patcog.2010.08.008.

Xiao, J., J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. 2010. "SUN database: Large-scale scene recognition from abbey to zoo". In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3485-92. IEEE. https://doi.org/10.1109/CVPR.2010.5539970.

Xiao, J., J. Hays, B. Russell, G. Patterson, K. Ehinger, A. Torralba, and A. Oliva. 2013. "Basic level scene understanding: categories, attributes and structures". *Frontiers in Psychology* 4. https://doi.org/10.3389/fpsyg.2013.00506.

Xiao, J., B. C. Russell, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. 2012. "Basic level scene understanding: From labels to structure and beyond". In *SIGGRAPH Asia 2012 Technical Briefs*, Article number 36. https://doi.org/10.1145/2407746.2407782.

Xiao, T., Y. Liu, B. Zhou, Y. Jiang, and J. Sun. 2018. "Unified perceptual parsing for scene understanding". In *Lecture Notes in Computer Science*, 432-48. Springer. https://doi.org/10.1007/978-3-030-01228-1_26.

Zaitoun, N. M., and M. J. Aqel. 2015. "Survey on image segmentation techniques". *Procedia Computer Science* 65: 797-806. https://doi.org/10.1016/j.procs.2015.09.027.

Zhang, Z., X. Zhang, C. Peng, X. Xue, and J. Sun. 2018. "ExFuse: Enhancing feature fusion for semantic segmentation". In *Lecture Notes in Computer Science*, 273-88. Springer. https://doi.org/10.1007/978-3-030-01249-6_17.

Appendix A

Authors	Year	Methodology	Datasets Used	Analysis
Long, Shelhamer, and Darrell (2015)	**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, Rabinovich, and Berg (2015)	**2015**	A model called ParseNet is proposed, which has modules that work on the feature mappings of the image rather than regions of an image.	**PASCAL VOC 2012, PASCAL-Context, SiftFlow**	The inclusion of the global spatial context of the image was considered in addition to the Fully Convolutional Network approach.
Ronneberger, Fischer, and Brox (2015)	**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**	**The 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, Kendall, and Cipolla (2017)	**2017**	A network SegNet is used for pixel-wise Semantic Segmentation of road and indoor scenes. A comparison of SegNet and other segmentation architectures is made using different performance measures.	**CamVid and SUN RGB-D**	**A 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 et al. (2017)	**2017**	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 et al. (2017)	**2017**	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**	**The proposed architecture jointly performs object detection and Semantic Segmentation, which increases the accuracy as both tasks benefit from each other. There is weight sharing between the tasks, which enhances the learning process.**

Li et al. (2017)	**2017**	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 et al. (2018)	**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 et al. (2018)	**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**	**A 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 et al. (2018)	**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 their surroundings better.**
Valada, Mohan, and Burgard (2020)	**2020**	An architecture is proposed for 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 leverage multiple modalities, which allow for learning richer and better representations that are robust to challenges like appearance changes etc.**

Table 5: Analysis of the deep learning-based Semantic Segmentation methods