Semantic Image Segmentation: A Comprehensive Overview

Shubhangi Ranjan
Information Management
ranjan4@illinois.edu

ABSTRACT

One of the major challenges in the long-standing history of computer vision has been semantic segmentation, which is the capacity to divide an unknown image into several components and objects (e.g., beach, ocean, sun, dog, swimmer). The goal of semantic segmentation is to segment an input image based on semantic information and predict the semantic category of each pixel based on a label set without necessarily knowing the precise identities of all the objects in the scene. Particularly, people are capable of doing visual segmentation without even being aware of what the nature of the items are (for example, in satellite imagery or medical X-ray scans, there may be several objects which are unknown, but they can still be segmented within the image typically for further investigation). As modern life becomes more intellectualised, even more applications, such as metaverse applications, autonomous driving, video surveillance, and so on, require inferring relevant semantic information from images for subsequent processing. A crucial step in our visual understanding process is segmentation, which can be used to both enhance or supplement current computer vision techniques as well as provide us with a powerful model for understanding the world. Prior to the development of deep learning, image segmentation problems were solved using traditional machine learning techniques like SVM, Random Forest, and K-means Clustering. But like with the majority of image-related problem statements, deep learning has significantly outperformed the competition and is now the standard method for semantic segmentation. In this paper, we will review some of the parallel techniques which are being used to solve the problem.

KEYWORDS

Semantic segmentation, image segmentation, scene understanding, convolutional neural network, metaverse.

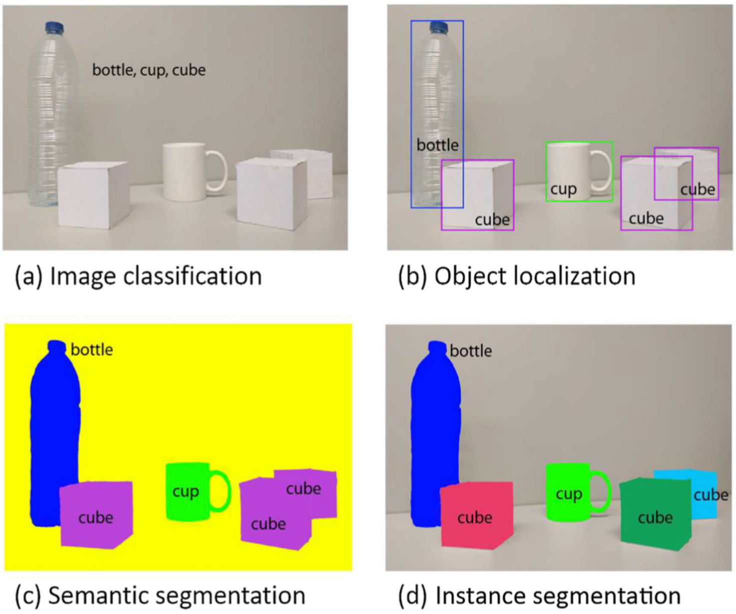
1 INTRODUCTION

Semantic segmentation is applied to still 2D images, video, and even 3D or volumetric data. It is now one of the most important problems in computer vision. In the grand scheme of things, semantic segmentation is one of the high-level tasks that pave the way for complete scene understanding. The fact that an increasing number of applications benefit from inferring knowledge from imagery emphasises the significance of scene understanding as a core computer vision problem. Autonomous driving, human-machine interaction, computational photography, image search engines, and augmented reality are just a few examples. Image segmentation can even be applied to radiation, image-guided therapies, and enhanced radiological diagnostics in the field of medical image analysis.Various traditional computer vision and machine learning techniques have been used in the past to address such a problem. Despite these methods' widespread use, the deep learning revolution has turned the tables and allowed many computer vision issues, including semantic segmentation, to be solved using deep architectures, typically Convolutional Neural Networks (CNNs), which are outperforming other methods in terms of accuracy and occasionally even efficiency.

Early approaches to image segmentation could divide images into regions using only basic colour and low-level textural information. Such techniques could be combined with machine learning methods such as Support Vector Machine or Random Forest, typically in a process involving the segmentation of an image into superpixels that are then classified, but performance was limited by the accuracy of a naive to semantic information segmentation method as well as the limitations of the machine learning algorithm used. Deep-learning techniques became popular in the 2010s, thanks in part to the availability of massively parallel GPUs and large labelled datasets, and convolutional neural networks capable of combining colour, texture, and semantic information to produce significantly more accurate results emerged. Early vision models built on deep learning were frequently centred on classification, in which a single image is given a single semantic label. The more precise problem of detection, which involves classifying and locating several items inside an image, later attracted a lot of research interest. Semantic segmentation finally emerged as a result of this pursuit for paradigms for understanding scenes at ever-finer spatial scales. Every pixel in an input image is given a class probability by a CNN, with the class with the highest likelihood assumed to be the pixel's predicted class label. Each element of the class probability vector indicates the likelihood that the given pixel belongs to one of a predetermined number of semantic classes. These concepts have since been expanded upon in works such as instance segmentation, which labels pixels not only by semantic class but also by specific instances of that class, volumetric segmentation, which labels each element of a 3D scene, which may be represented as a point cloud, mesh, or voxel grid, and video segmentation, which tracks semantic classes over a temporal sequence of images. Fig.1 shows the aforementioned evolution.

The modern deep learning architectures approach semantic segmentation as a natural step in the progression from coarse to fine inference rather than an isolated field. Even if the origin can be identified at classification, the next step towards fine-grained inference is localization or detection, which provides not only the classes but also extra details about the geographical location of those classes, such as centroids or bounding boxes. Given that, it follows naturally that semantic segmentation is the next stage in achieving fine-grained inference; its objective is to make dense predictions inferring labels for every pixel; in this way, each pixel is labelled with the class of its surrounding item or region.

Ultimately, the per-pixel labelling problem can be expressed as follows: Find a technique to give each of the components of the set of random variables X = "x1," "x2,"..., "xN" a state from the label space L = "l1, l2,..., lk." Each label designates a distinct category or item, such as an aeroplane, car, traffic sign, or background. The k potential states in this label space are typically increased to k + 1 while treating l0 as background or a void class. In most cases, X is a 2D image made up of W H = N pixels. However, that collection of random variables can be expanded to any dimensions, including hyperspectral or volumetric data.



(a) Image classification (b) Object localization

(c) Semantic segmentation (d) Instance segmentation

Fig.1. Evolution of object recognition or scene understanding from coarse-grained fine-grained inference: classification, detection & localization, semantic segmentation, and instance segmentation.[4]

2 **BACKGROUND**

Deep semantic segmentation systems frequently rely on a set of common networks, strategies, and design choices. Deep networks have contributed so significantly to the area that they are now widely accepted as standards. Due to their significance, numerous segmentation schemes are presently using them as building parts. A handful of them as well as certain methods will be reviewed.

2.1 **AlexNet**

Convolutional Neural Networks (CNNs) have long been the preferred model for object recognition because they are reliable, manageable, and even simpler to train than other models. When applied to millions of pictures, they don't exhibit overfitting at any concerning levels. Their performance is nearly equivalent to that of similar-sized traditional feedforward neural networks. The only issue is that applying them to high resolution images is challenging. The deep CNN, Alexnet solves for this with GPU-optimised and reduced training times while enhancing performance was required at the ImageNet scale[4]. It won the ILSVRC-2012 with a TOP-5 test accuracy of 84.6% whereas the nearest rival, which used conventional techniques rather than deep structures, only managed a 73.8% accuracy. The architecture depicted was relatively straightforward. It has five convolutional layers, three fully connected layers, dropout, max-pooling layers, and Rectified Linear Units (ReLUs) as non-linearities.

2.1.2 Dataset

ImageNet is a dataset with over 15 million high-resolution photos that have been assigned to 22,000 classes. Web-scraping photos and crowdsourced human labels are the solution. Researchers at the ImageNet Large-Scale Visual Recognition Challenge are challenged to attain the lowest top-1 and top-5 mistake rates in this competition using a subset of the ImageNet images (the top-5 error rate would be the percentage of images where the correct label is not one of the model's five most likely labels). Data is not an issue in this competition; there are over 1.2 million training images, 50,000 validation images, and 150 000 testing images. By removing the central 256x256 patch from each image, the authors imposed a fixed resolution of 256x256 pixels.

2.2 **VGG**

The University of Oxford developed the CNN model known as Visual Geometry Group (VGG)[5]. One of the deep CNN models and configurations they proposed was submitted to the (ILSVRC)-2013. Popularly known as VGG-16, consisting of 16 layers of weight, rose to prominence after achieving 92.7% TOP-5 test accuracy in ImageNet at the (ILSVRC)-2013. The use of a stack of convolution layers with small receptive fields in the first layers rather than a small number of layers with large receptive fields is the primary distinction between VGG-16 and its forerunners. The decision function becomes more discriminative and the model is simpler to train as a result of having fewer parameters and more non-linearities between them.

3 **METHODS**

3.1 **FCN: FULLY CONVOLUTIONAL NETWORKS**

Researchers have been exploring the potential of deep learning techniques for pixel-level labelling issues like semantic segmentation due to the unceasing success of these methods in a variety of high-level computer vision and remote sensing tasks, particularly supervised approaches like Convolutional Neural Networks (CNNs) for object detection or image classification. The main benefit of these deep learning techniques, which gives them an edge over conventional methods, is the capacity to learn appropriate feature representations for the task at hand, such as pixel labelling on a specific dataset, in an end-to-end manner as opposed to using hand-crafted features that demand domain expertise, effort, and frequently too much fine-tuning to make them work on a particular scenario. The Fully Convolutional Network (FCN) by Long et al.[1] is the precursor of the state-of-the-art deep learning algorithms for semantic segmentation that are currently the most effective. The genius of that strategy was to use pre-existing CNNs as potent visual models with the capacity to learn feature hierarchies. They replaced the fully connected layers with convolutional ones to convert the existing and well-known classification models AlexNet, VGG (16-layer net), GoogLeNet, and ResNet into fully convolutional ones that produce spatial maps rather than classification scores. To create dense per-pixel labelled outputs, those maps are upsampled using fractionally strided convolutions (also known as deconvolution). Since this work demonstrated how CNNs can be effectively trained end-to-end for this issue, learning how to produce dense predictions for semantic segmentation with inputs of any size, it is regarded as a milestone in the field. On common datasets like PASCAL VOC, this strategy significantly outperformed conventional methods in terms of segmentation accuracy while maintaining inference speed. The FCN is the cornerstone of deep learning used for semantic segmentation for all of these reasons as well as other important contributions.

3.1.1 Architecture

Typically, a shrunk input image is processed using convolution layers and fully connected layers in classification, yielding one predicted label for the input image. Instead of "Dense" layers (as in conventional CNNs), this network uses 1x1 convolutions to execute the function of fully linked layers (Dense layers). Since it is no longer constrained by the “flattened” operation, we can take arbitrary sized inputs and receive outcomes that still follow the same size. Additionally, the output will not be a single label if the image is not shrunk. Instead, the output is smaller than the size of the original image (due to the max pooling). We can determine the pixel-wise output if we upsample the signal from above (label map). Upsampling basically means that the spatial dimensions are enlarged to the original dimensions, thus allowing for pixel wise classification. Figure 1 illustrates how the loss is computed and back propagated over the network. In order to deal with the loss of information during the encoding and decoding processes, a technique called skip connections is used. It aims to combine the different semantic information for localised information. Combining fine layers and coarse layers lets the model make local predictions that respect global structure. This enables the understanding of “what” and “where” questions.

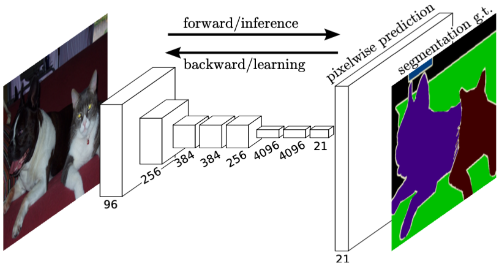


Figure 1: Fully convolutional networks can learn to make dense predictions for per-pixel tasks such as semantic segmentation effectively.[1]

3.1.2 Datasets Evaluation Metrics

PASCAL VOC: This challenge includes a ground-truth annotated image dataset as well as five competitions: classification, detection, segmentation, action classification, and person layout. The segmentation one is particularly intriguing because it aims to predict the object class of each pixel in each test image. There are 21 categories divided into vehicles, household, animals, and others: aeroplane, bicycle, boat, bus, car, motorcycle, train, bottle, chair, dining table, potted plant, sofa, TV/monitor, bird, cat, cow, dog, horse, sheep, and per- son. If the pixel does not belong to any of those classes, the background is also taken into account. The dataset is divided into two subsets: training and validation, each of which contains 1464 and 1449 images. The challenge's test set is private. Since this dataset is presumably the most popular for semantic segmentation, nearly every notable method in the literature is submitted to its performance evaluation server to validate against their private test set. Methods can be trained using either the dataset alone or with additional information. Furthermore, its leaderboard is open to the public and can be viewed online[6].

NYUDv2: 1449 indoor RGB-D photos were recorded using a Microsoft Kinect device and are included in this database. In both the training (795 photos) and testing (654) splits, it offers per-pixel dense labelling (at the category and instance levels) that was condensed into 40 indoor item classes. This dataset is particularly noteworthy because it was collected indoors, which makes it ideal for several domestic robotic jobs. However, the use of deep learning architectures is hampered by its relatively limited scale in comparison to other current datasets[6].

SIFT Flow: It contains 2688 fully annotated images drawn from the LabelMe database. The majority of the photographs are based on eight different outdoor scenes, which include streets, mountains, fields, beaches, and buildings. Images are 256 256 pixels in size and belong to one of 33 semantic classes. Unlabeled pixels are treated as such, as are pixels labelled with a different semantic class[6].

3.1.3 Evaluation Metrics

Mean IU: The Intersection over Union (IoU) metric, also known as the Jaccard index, quantifies the percentage overlap between the target mask and our prediction output. This metric is closely related to the Dice coefficient, which is frequently used as a training loss function. Simply put, the IoU metric counts the number of pixels shared by the target and prediction masks divided by the total number of pixels shared by both masks.

Pixel accuracy: Another metric for evaluating semantic segmentation is to simply report the percentage of pixels in the image that were correctly classified. Pixel accuracy is commonly reported separately for each class as well as globally across all classes. When evaluating per-class pixel accuracy, we are essentially evaluating a binary mask; a true positive represents a pixel that is correctly predicted to belong to the given class (based on the target mask), whereas a true negative represents a pixel that is correctly identified as not belonging to the given class.

3.1.4 Baseline Results

Since this work demonstrated how CNNs can be effectively trained end-to-end for this issue, learning how to produce dense predictions for semantic segmentation with inputs of any size, it is regarded as a milestone in the field. The inherent spatial invariance of the FCN model does not account for useful global context information, there is no instance-awareness by default, efficiency is still far from real-time execution at high resolutions, and it is not entirely suitable for unstructured data such as 3D point clouds or models. Despite its strength and flexibility, these features prevent the FCN model from being applied to certain problems and situations. On common datasets like PASCAL VOC, this strategy significantly outperformed conventional methods in terms of segmentation accuracy while maintaining inference efficiency. We observe that FCN-8s is the best in Pascal VOC 2011, FCN-16s is the best in NYUDv2 and the FCN-16s is the best in SIFT Flow.

3.2 **DILATED CONVOLUTIONS**

The problem of semantic segmentation demands the integration of data from diverse spatial scales. Additionally, it suggests balancing regional and global information. To attain good pixel-level precision, fine-grained or local information is essential. On the other hand, in order to be able to clarify local uncertainties, it is also crucial to incorporate information from the image's overall context. Vanilla CNNs have trouble striking this equilibrium. The global context data is discarded by pooling layers, which help the networks achieve some level of spatial invariance and reduce computational cost. Due to the fact that the receptive field of its units can only increase linearly with the number of layers, even pure CNNs without pooling layers have limitations.

3.2.1 Architecture

In general, context information that extends far beyond pixel-level appearance becomes aware of semantics and serves as a useful supplementary source for developing semantic segmentation models. In 2011, Lucchi et al. [2] stated that semantic segmentation accuracy can be boosted by appropriately introducing context information.Yu et al. [2] proposed DilatedNet in 2015 to aggregate multi-scale contexts using dilated convolution. They maintain resolution despite supporting exponentially expanding receptive fields. By receptive fields we mean, any pixel in the original image that is affecting the current pixel in the current layer.

Dilated convolutions, then, are just standard convolutions with upsampled filters. That upsampling factor is controlled by the dilation rate l. The dilation rate l controls that upsampling factor. The larger dilated rate convolution has a larger receptive field while introducing no extra computations as shown in Figure 2. This means that dilated convolutions allow efficient dense feature extraction on any arbitrary resolution. The dilated convolution with a dilated rate of 1 is a special case that degenerates to the traditional convolution. DilatedNet extracts contexts from multiple scales using five different dilated rates which are powers of 2. It is then convolved with 3 × 3 filters. This allows us to have a huge receptive field. The same is depicted mathematically as follows where k is the filter. $F_{i+1} = F_i *_{2^i} k_i \quad \text{for } i = 0, 1, \ldots, n-2.$ It is easier to see the size of the receptive field of each element in F_{i+1} is $(2^{i+2}-1) \times (2^{i+2}-1)$ The receptive field is a square of exponentially increasing size. The same is illustrated in Figure 2.

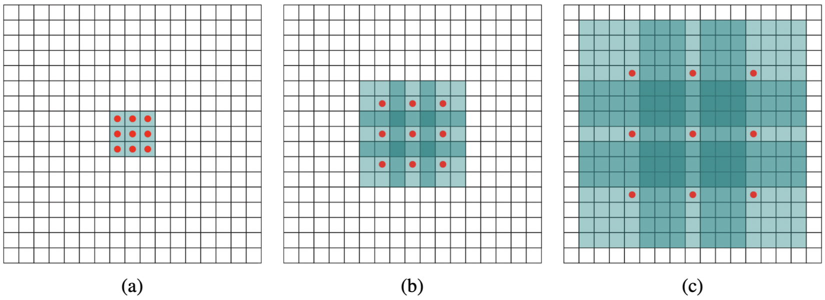

(a)
(b)
(c)

Figure 2: Systematic dilation supports exponential expansion of the receptive field without loss of resolution or coverage. (a) F_1 is produced from F_0 by a 1-dilated convolution; each element in F_1 has a receptive field of 3×3. (b) F_2 is produced from F_1 by a 2-dilated convolution; each element in F_2 has a receptive field of 7×7. (c) F_3 is produced from F_2 by a 4-dilated convolution; each element in F_3 has a receptive field of 15×15. The number of parameters associated with each layer is identical. The receptive field grows exponentially while the number of parameters grows linearly. [2]

Context Network: A context module is constructed based on the dilated convolution. The context module uses 33 convolutions with various dilation factors across 7 layers. There are 1, 1, 2, 4, 8, 16, and 1 dilations. The final one uses 1-1 convolutions to map the number of channels to match the input channel count. As a result, the number of channels on both the input and the output are equal. Additionally, it can be put into other convolutional neural networks. While the large context module contains increasing numbers of channels from 1C as input to 32C at the seventh layer, the simple context module only has 1 channel (1C) throughout the module.

3.2..2 Dataset

COCO: Due to the fact that one of its components is segmentation-focused, it has a variety of obstacles, with detection being the most pertinent for this subject. More than 82,783 images are provided for training, 40,504 images are provided for validation, and the test set for this challenge, which has more than 80 classes, includes more than 80,000 images. The test set is specifically divided into four separate subsets or splits: test-dev (20,000 images) for additional validation and debugging, test-standard (20,000 images) is the competition's default test data and is used to compare cutting-edge methods, test-challenge (20,000 images) is the split used for the challenge when submitting to the evaluation server, and test-reserve (20,000 images) is a split used to guard against potential overfitting in the test set.

3.2.3 Training & Baseline Results

The front-end module utilised is VGG-16. The context module is inserted after the final two pooling and striding layers have been totally eliminated. Additionally deleted is the padding from the intermediate feature maps. Only a 33-pixel width is added by the authors to the input feature maps. In our tests, the outcomes from zero padding and reflection padding were comparable. In addition, rather than using the usual random initialization, a weight initialization that takes into account the number of input and output channels is employed. The MIoU for DilatedNet for the PASCAL VOC 2012 test set is 67.6%.

3.3 **ADVERSARIAL NETWORKS**

The Generative Adversarial Network (GAN) has been successful in a variety of applications, including style transfer, picture inpainting, and text to image synthesis. It is inspired by game theory: two models, a generator and a critic, are competing with each other while making each other stronger at the same time. ANet, a network made up of both a segmentation network and an adversarial network introduced by Luc et al. [3], is the ground-breaking research for using GAN in semantic segmentation. The segmentation network (generator) divides the input image into a number of non-overlapping sections for forward propagation. The adversarial network (discriminator) separates the output of the generator from the ground-truth label mappings as shown in Figure 2. It is trained using a hybrid loss that is a weighted sum of two terms. The first term in the loss is the conventional multi-class cross entropy loss used in semantic segmentation. It encourages the segmentation model to predict the right class label at each pixel location independently. The second loss term is based on an auxiliary adversarial convolutional network. This loss term is large if the adversarial network can discriminate the output of the segmentation network from ground-truth label maps. In the back-propagation, the discriminator and generator, which behave in a min-max-game manner, are alternately subjected to the adversarial loss. As an input to the discriminator, the probability maps are directly provided since there is a risk of trivial separation between ground truth and ground truth generated label maps. In order to avoid this problem, the first strategy proposed aims to segment the input image with each of the probability maps; by element-wise multiplication with each of the RGB channels. The second strategy is to produce noisy segmentation maps from the ground truth by setting the distribution for each spatial location such that there is a high enough probability on the ground truth label. It aims to replicate the distribution that was produced by the segmentation network. The pixel-level categorization loss is utilised as a strong constraint since semantic segmentation has special requirements.

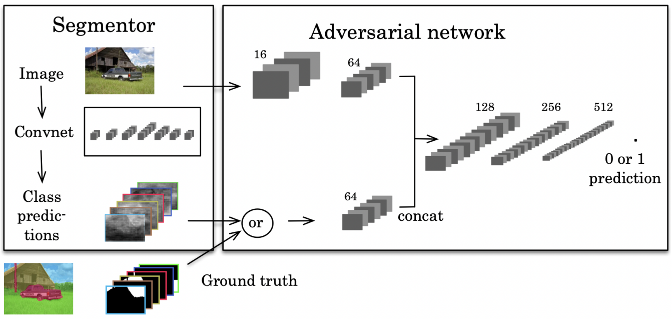


Figure 3: Overview of the proposed approach. Left: segmentation net takes RGB image as input, and produces per-pixel class predictions. Right: Adversarial net takes label map as input and produces class label (1=ground truth, or 0=synthetic). Adversarial optionally also takes RGB image as input.[3]

3.3.1 Baseline Results

Significant gains were observed through all the commonly used metrics. On PASCAL VOC, smaller but consistent improvements were observed on validation and test sets. It reported a Mean IoU of 54.3%. Adversarial training significantly reduced overfitting. Qualitatively, it smoothes and strengthens the predictions across large areas and removes smaller predictions with sharper class boundaries.

3.3.2 Dataset & Training

Stanford Dataset: LabelMe, MSRC, PASCAL VOC, and Geometric Context are some of the public datasets that the outdoor scene photos in this dataset were imported from. The collection includes 715 images (320 x 240 pixels each) containing at least one foreground object and the location of the horizon. For the purpose of testing approaches for semantic scene understanding, the dataset is pixel-wise labelled (horizon location, pixel semantic class, pixel geometric class, and image region).

The multi-scale segmentation network was patch-wise trained from scratch on this dataset. An input label map and its associated RGB image are provided to the adversarial network. The segmentation net generates the label map, which is either the ground truth corresponding to the image. The output resolution of the segmentation net is matched by downsampling the ground truth label maps, which are then given to the adversarial in a one-hot encoding format. To enable various low level representations for the two distinct signals, the picture and the label map are initially processed separately by two branches. Each input signal should ideally have around the same number of channels, to prevent one signal from overpowering the other when fed to later levels.

4 COMPARISONS

We observed three different approaches to semantic segmentation.

One major observation during the study was that FCN is considered groundbreaking in convolution network based architectures and its results were baseline for convolution network based architectures. However, the labels predicted by FCN lack a basic spatial relationship or contextual information, resulting in a coarser segmentation result. The FCN only uses the current pixel and neighbourhood pixels to predict each label, completely ignoring the relationships between the labels. In other words, the FCN only makes proper use of the prior probability between pixels and ignores the likelihood knowledge of the label. The pixel accuracy and meanIoU for FCN are 71.32% and 29.39% respectively for the validation set of PASCAL context dataset.[8]

Dilated convolutions reported a slight improvement in the standard evaluation metrics.They reported 73.55% & 32.31% on the standard evaluation metrics. They created a convolutional network module that can analyse rescaled images and collect multi-scale contextual data without sacrificing resolution. The advantage to this architecture is that at any resolution, the module can be plugged into existing designs. It can be concluded that while FCNs are a consolidated approach for semantic segmentation, they lack several features such as context modelling that help increasing accuracy.

The adversarial approach poses a lot of advantages in comparison to the convolutional network approaches. Firstly, the adversarial model is versatile enough to find mismatches in a variety of higher-order statistics between the model predictions and the ground truth without the need for explicit definition. Secondly, since the model does not contain any higher-order terms or recurrence within the model itself, it is efficient once trained.

5 CONCLUSION AND FUTURE SCOPE

A detailed overview of deep learning-based semantic segmentation techniques has been provided in this study. The architecture, experiments, dataset and their training methodology has also been discussed in detail. Our critical observations were that accuracy and efficiency are both significant for evaluating a semantic segmentation method. However, the gains in these two aspects are still contradictory to each other for all the current semantic segmentation methods.[9] Also, coping with common camera frame rates and striking a balance between accuracy and runtime will aid in real-time semantic segmentation. Pruning is a promising research line that aims to simplify a network, making it lightweight while keeping the knowledge, and thus the accuracy, of the original network architecture. Secondly, for accurate semantic segmentation, high-quality training data are considered to be a prerequisite. However, obtaining high-quality training data, i.e., sufficiently labelled images with pixel-level annotation, is an inevitably laborious and time-consuming task. This heavy dependency has become another common challenge of semantic segmentation. Moreover, the experiments were conducted on carefully curated and labelled datasets like PASCAL and COCO. Similar output from the models cannot be expected on real world datasets that include occlusions and complex scenes[6].

Methods that fully utilise 3D information are beginning to emerge, but despite the development of new concepts and approaches, they still lack one of the most crucial elements: data. Large-scale datasets are essential for 3D semantic segmentation, but they are more difficult to produce than their lower dimensional equivalents.[7] Even though there is some promising research already, more, better, and more diverse data are still needed. Finally, deep learning has proven to be very effective in addressing this issue, thus in the future years, we can anticipate a frenzy of innovation and the spawning of new research directions.

6 **REFERENCE**

[1] Long, J., Shelhamer, E., & Darrell, T. (1970, January 1). *Fully convolutional networks for semantic segmentation*. Page Redirection. Retrieved December 12, 2022, from https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Long_Fully_Convolutional_Networks_2015_CVPR_paper.html

[2]Yu, F., & Koltun, V. (2015, November 23). *Multi-scale context aggregation by dilated convolutions*. arXiv.org. Retrieved December 12, 2022, from https://arxiv.org/abs/1511.07122v1 ICLR, 2016

[3]Luc, P., Couprie, C., Chintala, S., & Verbeek, J. (2016, November 25). *Semantic segmentation using adversarial networks*. arXiv.org. Retrieved December 12, 2022, from https://arxiv.org/abs/1611.08408 NIPS, 2016

[4]Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., & Garcia-Rodriguez, J. (2017, April 22). *A review on Deep Learning techniques applied to semantic segmentation*. arXiv.org. Retrieved December 12, 2022, from https://arxiv.org/abs/1704.06857

[5]ShijieHaoabYuanZhouabYanrongGuoPersonab, A. links open overlay, ShijieHaoab, a, b, YuanZhouab, YanrongGuoPersonab, (2020, April 13). *A brief survey on semantic segmentation with Deep Learning*. Neurocomputing. Retrieved December 12, 2022, from https://www.sciencedirect.com/science/article/abs/pii/S0925231220305476

[6]Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena Martínez, V., Martínez González, P., & Garcia-Rodriguez, J. (2018, September 1). *A survey on Deep Learning techniques for image and video semantic segmentation*. RUA. Retrieved December 12, 2022, from https://rua.ua.es/dspace/handle/10045/75753

[7]*Deep semantic segmentation of natural and Medical Images: A Review*. DeepAI. (2019, October 16). Retrieved December 12, 2022, from https://deepai.org/publication/deep-semantic-segmentation-of-natural-and-medical-images-a-review

[8]Wu, H., Zhang, J., Huang, K., Liang, K., & Yu, Y. (2019, March 28). *FASTFCN: Rethinking dilated convolution in the backbone for semantic segmentation*. [1903.11816] FastFCN: Rethinking Dilated Convolution in the Backbone for Semantic Segmentation. Retrieved December 12, 2022, from http://export.arxiv.org/abs/1903.11816

[9]*An end-to-end bayesian segmentation network based on a generative ...* (n.d.). Retrieved December 12, 2022, from https://www.researchgate.net/publication/338483913_An_End-To-End_Bayesian_Segmentation_Network_Based_on_a_Generative_Adversarial_Network_for_Remote_Sensing_Images/fulltext/5e1740b74585159aa4c08691/An-End-To-End-Bayesian-Segmentation-Network-Based-on-a-Generative-Adversarial-Network-for-Remote-Sensing-Images.pdf