International Journal of Grid and Distributed Computing
Vol. 14, No. 1, (2021), pp. 1501–1512

Semantic Segmentation of Satellite Images: A Survey

Shailendra Singh[1] and Sheetal Girase[2] [1]*School of CET, MIT WPU, Pune, Maharashtra, India*
[2]*School of CET, MIT WPU, Pune, Maharashtra, India*
[1]*shailendrasng77@gmail.com*, [2]*Sheetal.Girase@mitwpu.edu.in*

Abstract

In the era of AI, Computer Vision plays an important role in various applications which are beneficial for the society. Semantic Segmentation is one such sub domain of Computer Vision which has a number of applications such as Autonomous Driving, Medical Image Diagnosis, Satellite Image Processing etc. In simple words, Semantic Segmentation is the process of assigning pixels to different classes present in a visual imagery. It is important for researchers working in this field to know about some of the widely used Semantic Segmentation models. Our main focus is on Semantic Segmentation of Satellite Images. Studying Satellite Images and using it for better understanding of our planet is possible due to advancements in Computer Vision and Deep Learning along with availability of low cost high performance GPUs. This paper provides literature survey of various Semantic Segmentation models that can be used for processing various Satellite Images. The paper also discusses about Satellite Image processing techniques, its challenges, various Satellite Images datasets and different evaluation metrics used for the purpose of evaluation of these Semantic Segmentation models.

Keywords: *Computer Vision, Deep Learning, Satellite Images, Semantic Segmentation.*

1. Introduction

Satellite imagery depicts the Earth's surface at various spectral, temporal, radiometric, and increasingly detailed spatial resolutions, as is determined by each collection system's sensing device and the orbital path of its reconnaissance platform [49]. The potential details provided by the imagery is referred to as resolution of that imagery. In remote sensing there are three types of resolution namely Spatial, Spectral and Temporal resolution.

Spatial resolution: This refers to the size of the smallest feature displayed in a satellite image. For example, a spatial resolution of 250m means that one pixel represents an area 250 by 250 meters on the ground.

Spectral Resolution: It refers to the ability of a satellite sensor to measure specific wavelengths of the electromagnetic spectrum.

Temporal resolution: It refers to the capability for satellites to provide images of the same geographical area i.e. the time between images.

Satellite images can also be categorized into High-Resolution, Very High-Resolution and Low Resolution images based on the number of pixels they have.

Satellite images are captured in various wavelengths. For instance, below are two pictures of the same location. Left side image is an image that our eyes would normally see and this image is called true color image. On the other hand, right side image is made up of three wavelengths only and is called false color image.

Since the dawn of space age this satellite data is used in various domains like forecasting, Agriculture, Forestry, landscaping, geology, education and so on. Wide adaptation of Machine Learning and Deep Learning Techniques and the availability of open, free satellite imagery data has opened up many avenues for study of Semantic Segmentation. Semantic segmentation of satellite imagery is used to extract road networks, water bodies, trees, detect buildings for various environmental purposes.

ISSN: 2005-4262IJGDC

Figure 1. Satellite images of different wavelengths

2. Semantic Segmentation Models

Biao Li et al. [1] provided survey for semantic segmentation and talked about models such as FCN, SegNet and Enet. **FCN**: First work was done by Shelhamer et al. [2]. They introduced the first deep learning method of convolutional networks for semantic segmentation. Using this method, dataset can be trained in end-to-end fashion and can achieve results as pixel-to-pixel. Their main task was to propose a CNN that can be used for semantic segmentation tasks and which can produce the output with the similar size as that of the input image. The key idea in FCN-based methods [2][14-15] is that they learn a mapping from pixels to pixels, without extracting the region proposals [6]. However, these networks are commonly used for local tasks rather than global tasks (i.e. semantic segmentation [2] or object detection [15] instead of object classification [16]). The problem in FCN approach is that by propagating through several convolutional and pooling layers, the resolution of the output feature maps is down-sampled which makes the output low in resolution which further results in fuzzy object boundaries. **SegNet**: Badrinarayanan et al. [3] proposed an encoder-decoder architecture. The base of the network is the unpooling layers. The method has got two main advantages: One was that the proposed architecture can save a considerable amount of computation resources and also can save some amount of training time. Second advantage was that the method can compute faster due to less number of parameters during training. There were two parts in this network: encoder network and decoder network. There are 13 convolutional layers in the encoder architecture which are coming from VGG16 [10] architecture (pre -trained) . This has been done to fetch coarse information from the input image. The decoder network has been used for final pixel-wise classification. The output image was of the same size as that of input image. **E-net**: Many of the previous architectures used VGG16 [10] architecture as the base architecture to fetch coarse information from the input image which can be further used for the segmentation purpose. This architecture deals with a large number of parameters due to which it is difficult to perform semantic segmentation on handheld devices such as phones, tablets etc. For this purpose, Paszke et al. [4] proposed a new structure coming from the idea of another work[5]. The new structure has less number of parameters and also it is efficient like the previous proposed architectures. It also made possible the task of semantic segmentation on mobile devices. They named their new architecture Enet (Efficient Neural Network). They claimed that it can run 18 times faster than the previously proposed models and also it requires 79 less parameters with 25 FLOPs. They compared their model with previous models on various datasets such as CamVid, Cityscapes and SUN. They also claimed that the final results showed that E-Net can maintain the balance between accuracy, time and computational resources.

Yanming Guo et al. [6] provided a survey for semantic segmentation and talked about **RCNN** [7] and its limitations. RCNN first uses selective search [8] approach to extract large number of object regions and then it computes CNN features for all of them individually. At final stage, it classifies every region using linear SVM approach. RCNN can be built on top of any CNN structures such as AlexNet [9], VGG [10], GoogLeNet [11] and ResNet [12][6]. In image segmentation, it extracts mainly two features namely full region feature and foreground feature and also we can get better performance if we concatenate both the features. However, RCNN also suffers from some drawbacks which are:

Feature not compatible with segmentation task.

Feature does not contain enough spatial information which is required to generate precise boundary.

Using it for segment-based tasks will take time and can greatly affect the final performance.

Andrew King at el. [17] talked about models such as VGG16, ResNet, Dilation8. **VGG16** architecture

ISSN: 2005-4262IJGDC

was proposed by Simonyan and Zisserman [10] for the task of image classification. The VGG16 architecture represents a significant improvement over previous networks by its use of small 3 × 3 kernel filters instead of the larger kernel filters common at that time [17]. The VGG16 architecture consists of total 16 layers out of which 13 are convolutional layers and 3 are fully connected layers. As the CNN grow deeper, the gradient updates become vanishingly small in the upper layers of the network, presenting significant difficulties during the training process [17]. This is termed as the vanishing gradient problem. This problem is addressed by He et al. [12] using their new architecture which they named **ResNet**. ResNet makes use of residual blocks which further makes use of skip connections which passes information directly from first layer to the last layer of the block. This will allow the gradient to be preserved across several CNN layers. Yu and Koltun [18] proposed a new FCNN architecture which is termed as **Dilation8.** Dilation8 is based on the FCN8s architecture [2] improving on its results. Dilation8 removes some of the max pooling layers in VGG16 in order to preserve spatial resolution [17]. Their approach only downsamples the image to 1/8 of its original size as opposed to 1/32 in the FCN8s architecture which is proposed by Shelhamer et al. [2].

Shervin Minaee at el. [20] talked about ParseNet, U-Net, V-Net, DeepLabv1, DeepLabv2, DeepLabv3, DeepLabv3+, ReSeg, ReNet and RAN. Liu et al. [21] proposed a model called **ParseNet** to overcome the problem with FCN which ignores global context information. ParseNet adds global context to FCNs by using the average feature for a layer to augment the features at each location [20]. Ronneberger et al. [22] proposed the **U-Net** for segmentation of biomedical images. The U-Net architecture comprises two parts, a contracting path to capture context and a symmetric expanding path that enables precise localization [20]. The contracting part has similar FCN-like architecture that extracts features with 3 × 3 convolutions. The expanding part uses up-convolution or deconvolution which reduces the number of feature maps while increasing their dimensions. Feature maps from the down-sampling part of the network are copied to the up-sampling part to avoid losing pattern information. Finally, a 1×1 convolution processes the feature maps to generate a segmentation map that categorizes each pixel of the input image. **V-Net** is another well-known FCN-based model which was proposed by Milletari et al. [23] for the segmentation of medical images. For model training, they introduced a new objective function based on the Dice coefficient, enabling the model to deal with situations in which there is a strong imbalance between the number of pixels in the foreground and background [20]. The proposed network was trained end-to-end on MRI volumes that depicts prostate and it learns to predict segmentation once for the whole volume. **DeepLabv1** [24] and **DeepLabv2**

[19] are some of the most popular image segmentation methods which are developed by Chen et al [19][24]. The DeepLabv2 has 3 key features which are: the use of dilated convolution to address the decreasing resolution in the network which is caused by max-pooling and striding processes, second is Atrous Spatial Pyramid Pooling (ASPP) which probes an incoming convolutional feature layer with filters at multiple sampling rates, thus capturing objects as well as image context at multiple scales to robustly segment objects at multiple scales and third is improved localization of object boundaries by combining methods from deep CNNs and probabilistic graphical models [20]. Chen et al. [25] also proposed **DeepLabv3** which combines cascaded and parallel modules of dilated convolutions [20]. The parallel convolution modules are grouped into the ASPP. A 1 × 1 convolution and batch normalisation are then added in the ASPP. All the outputs are concatenated and are processed by another 1 × 1 convolution that creates the final output with logits for each pixel. In 2018, Chen et al.

[26] released a new version **Deeplabv3+** which uses an encoder-decoder architecture, including atrous separable convolution, composed of a depthwise convolution (spatial convolution for each channel of the input) and pointwise convolution (1 × 1 convolution with the depthwise convolution as input) [20] and they used the DeepLabv3 framework as encoder architecture.

Visin et al. [27] proposed an RNN-based model for semantic segmentation called **ReSeg** which is based on another work, **ReNet** [28] and it was developed for image classification. To perform image segmentation with the ReSeg model, ReNet layers are stacked on top of pre-trained VGG-16 convolutional layers that extract generic local features. ReNet layers are then followed by up-sampling layers to recover the original image resolution in the final predictions. Contrary to other

works where convolutional classifiers are trained to learn the important semantic features of labeled objects, Huang et al. [29] proposed a semantic segmentation method using reverse attention mechanism approach. Their proposed **Reverse Attention Network (RAN)** architecture trains the model to also capture the features that are not associated with a target class. This RAN architecture is a three-branch network that performs direct and reverse attention learning processes simultaneously.

ISSN: 2005-4262IJGDC

3. Semantic Segmentation Models for Satellite Images

For finding complex patterns and classification of satellite images not only Machine learning methods but also Convolutional Neural Networks have proven to be useful [31]. These Semantic Segmentations techniques can be applied on various satellite images to understand the features present in the images and also for comparing them.

Aleksandr A. Tiurin et al. [30] proposed an algorithm for Semantic segmentation of satellite images based on the architecture of the convolutional neural network U-Net [22] which was mainly used for segmenting biomedical images. This algorithm constructs raster masks of objects belonging to a single class. This algorithm provides a continuous result without stitches for satellite images of unlimited sizes, despite the limitation of the U-Net input data size. They chose U-Net because of its various advantages such as low inference time, less training data needed, easy to understand architecture, high speed learning.

Mohammad Yousuf Saifi et al. [32] used U-Net and FCN in their work for satellite image segmentation. They have used FCN along with U-Net so as to improve result and decrease the pressure on the machine. FCN is used as the main architecture whereas U-Net is used as the pre-trained model. They used SpaceNet dataset which consists of commercial satellite images. They used their proposed method for segmenting water, buildings, roads, corps in the input satellite images. The main purpose of this work is to perform semantic segmentation on satellite images which are taken from urban areas.

Manami Barthakur et al. [33] considered using SegNet [3] in which the output of K-means clustering algorithm is used as input and the label of the particular region of interest (ROI) in the image are used as target. Though this method does not reply on feature extraction, region growing or splitting methods to configure and train the SegNet [3], it suffers from high training time. They have segmented the image in mainly three classes which are sea, grass and house. The experimental results show that this proposed method is reliable and also suitable for real world scenarios.

J. Gonzalez et al. [13] has proposed a method where they have used two models for the segmentation task. The task is to segment water bodies using these two models. The models are namely model-1 and model-2. The main architecture for both the models is U-Net. In first model, they trained all the HR (high resolution) images and predicted the output. They used this same model for the knowledge transfer process for transfering the knowledge to model-2. In model-2, they used all the VHR (very high resolution) images in which they used the corresponding labels of the images to generate the the segmented output. If the labels are not there for some images, that image is send to model-1 by downsampling the image and through model-1 its corresponding label is generated which is again fed to model-2 by upsampling the labeled image to get the segmented result. They have used mainly 2 datasets to evaluate their models namely PeruSat -1 and Sentinel-2 dataset. The U-Net they have used is a modified one which is known as TernausNet [44]. Overall they have compared two different deep learning models for segmenting water bodies using satellite images in Peru.

Ming Wu et al. [34] proposed a method named AD-LinkNet for the semantic segmentation of satellite images. Image segmentation is the process of assigning a label to each pixel in an image, same-labeled pixels with same characteristic [35]. There is a long tradition of using computer vision techniques for satellite image understanding [36-37]. Since the fully convolutional network (FCN) [2] has shown numerous improvements in semantic segmentation, many researchers [38-40] have made efforts by making use of FCN. The network model designed in this paper is based on FCN and UNet. Unet [22] uses Transposed-conv [41] as its upsampling structure on the basis of FCN, connects the features of the network Encoder part to the Decoder part and combines low-level information with high-level information [34]. LinkNet [42] with ResNet [5] as Backbone is one of the baselines of their method. LinkNet also uses the U-shape structure and it replaces the convolutional structure of every level of its encoder and decoder with res-block. They proposed AD-LinkNet to leverage context

ISSN: 2005-4262IJGDC

information which will benefit satellite image segmentation by introducing channel-wise attention [43].

Lunhao Duan et al. [45] proposed a new network named MSR-Net (MultiScale Refinement Network) which is a deep convolutional neural network. They have used two datasets for training and testing their model, which are Gaofen Image dataset(GID) and DeepGlobe dataset. They have also compared their results with the existing state-of-the-art architectures such as Unet, SegNet, DeepLabv3+. They achieved 84.42% accuracy on the DeepGlobe dataset whereas on GID dataset, they achieved 94.54% accuracy. They claimed that their proposed network gives sharper boundaries, providing sufficient details when the water bodies are linear and planar in nature. The MSR-Net focuses on the multiscale information for segmentation improving existing networks at the same time. Instead of the traditional one-off manner that concatenates features and conducts segmentation on one uniform scale, the MSR-Net adopts a new multiscale refinement scheme that makes full use of the multiscale features for more accurate water-body segmentation [45]. A novel erasing-attention module has also been used for an effective feature embedding during the refinement process. The multiscale features of the images are extracted through a backbone network and a coarse segmentation result is obtained from the top-layer feature.

4. Datasets

There are a number of Satellite images datasets available and we have tried to mention some of the most commonly used and standard datasets.

PeruSAT-1 dataset [13] is a very-high resolution dataset which has a resolution of 2.8 m per pixel in RGB and NIR bands. Each of the image is having a size of 6000 X 6000 pixels. Sentinel-2 dataset [13] is another dataset which consists of 7671 images of high resolution. The resolution of the images is 10 m per pixel in RGB and NIR bands. Each of the image is of the size 64 X 64 pixels. UC Merced Land Use [46] and EuroSat Land Use [47] are both annotated with a single label per image [46]. UC Merced Land Use contains 2100 images (each of size 256 × 256) and each labelled with one of the 21 land use classes (e.g. agricultural, airplane, residential). EuroSat Land Use contains 27000 images (each of size 64 × 64) and each labelled with one of the 10 land use classes (e.g. Annual Crop, Industrial, Residential).

DeepGlobe Land Cover [48] was released for a fully-supervised semantic segmentation challenge and is annotated with multiple labels per image and it comprises of 1146 images (sized 2448 × 2448), 803 of which are annotated with one or more of 6 classes (e.g. urban, agriculture, rangeland), as well as an unknown class. DeepGlobe road extraction dataset is a 2-tiles dataset which is taken from India, Thailand and Indonesia. The task of this dataset is to extract roads from given satellite images. This dataset consists of 6226 pairs of training data, 1243 of verified images and 1101 of test images. Each image is of size 1024*1024 and ground resolution of the image pixels is 0.5m/pixel. DeepGlobe land classification dataset is a 7-tiles dataset which includes: urban area, agricultural area, water, barren land, and unknown land. It consists of 803 pairs of training data, 171 verification images and 1101 test images. The ground resolution of the image pixels is 0.5m/pixel.

The Inner Mongolia land classification dataset is a 7-tiles dataset which is taken from the Jilin No. 1 satellite. The original resolution of the satellite is 27338*24631 with a total of 12 pictures and ground resolution of the image pixels is 0.7m/pixel [34]. Another popular dataset of satellite images is GID [50] dataset. It consists of 150 high quality images from more than 60 different cities in China with six classes of annotations: built-up, farmland, forest, meadow, waters, and unknown [45].

5. Satellite Image Processing Techniques

Satellite image processing is defined as the processing of any satellite image using computer resources for analysis and other purposes. Different techniques of Satellite image processing are shown in the figure below.

Satellite image processing techniques are as follows (in order of their requirement): **Pre-processing Techniques:** When various digital satellite images are captured using different sensors, then they also capture some of the errors related to the pixel values of the respected pixels.

ISSN: 2005-4262IJGDC

These errors need to be corrected before further processing of the satellite images. This process is commonly known as Satellite image pre-processing [57]. Some of the preprocessing techniques are:
Image filtering: Filters can be used for sharpening and smoothing the satellite images. Noise Removal: To remove the noise components captured in satellite images.
Radiometric Correction: Radiometric correction [57] is done to calibrate the pixel values and to correct the errors in those values.
Geometric correction: Geometric correction [57] is done to a satellite image to ensure that pixels in the captured satellite image are in their proper and exact position similar to real world coordinates and geometry.
Atmospheric correction: Atmospheric correction [57] is done to a satellite image in order to remove the effects of the atmosphere on the reflectance values of images captured by different sensors.
Image Enhancement: This can be defined as the modification of the pixel brightness values so that visual appearance of the image would be better [57]. We can also define this as the conversion of the image to a form which is most suited for computer understanding.

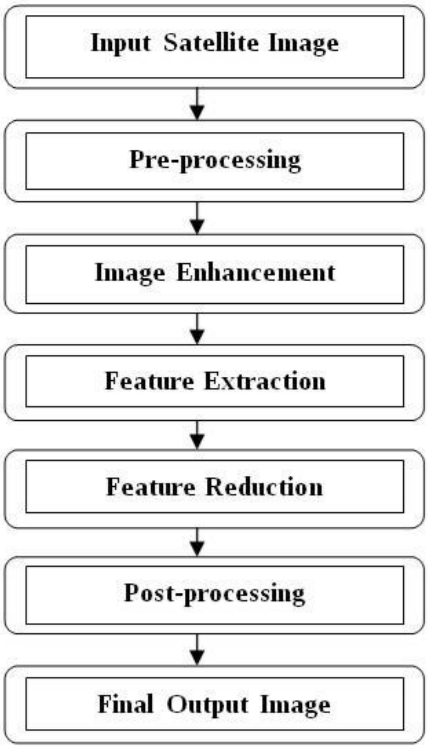

Figure 2. Satellite Image Processing

Feature Extraction Techniques: Features are those which are used to describe an object uniquely such as shape, size, location etc [54]. Feature extraction is the method of extracting high level features from satellite images so that the extracted features can be used for classification purpose. Some of the Feature extraction techniques for satellite images are:
Histogram of Oriented Gradients: Histogram of Oriented Gradients {HOG) [53] is best suited to extract the features which are used to determine the shape of objects present in satellite images. It works by computing intensities of gradients and edge direction of the objects present in satellite images.
Scale Invariant Feature Transform: In Scale Invariant Feature Transform (SIFT) [53], various key points are determined in the regions of the satellite image using the gradient information. The key points are extracted in large number from the satellite image which makes the technique effective in extracting features like small objects from the satellite image.

ISSN: 2005-4262IJGDC

Color Histogram: Color Histogram [53] is used to create a histogram for the different colors present in the satellite image. It is capable of detecting small illumination changes. Its limitation is that the texture and shape of the objects are neglected. Also, it is difficult to differentiate objects having same colors using this technique.

Local Binary Patterns: Local Binary Patterns (LBP) [54][56] is used for texture feature extraction. Texture features are used to define the spatial variation in pixel intensities values of a satellite image. We can use these texture features to identify different regions of a satellite image.

Feature Reduction Techniques: Feature reduction techniques are used for dimensionality reduction in satellite image processing. Some of the techniques for feature reduction are:

Singular Value Decomposition: Singular Value Decomposition (SVD) [57] is widely used technique for feature reduction. It provides another way to factorize a matrix into singular vectors and singular values. SVD is used both in matrix operations as well as in reduction method in satellite image processing.

Principal Component Analysis: Principal Component Analysis (PCA) [55] is a method emerging from mathematical statistics which can also be used for the purpose of dimensionality reduction.

Curvilinear Component Analysis: Curvilinear Component Analysis (CCA) [55] is feature reduction method which focuses on the preservation of the distance matrix while projecting data onto a lower dimensional space.

Isometric Feature Mapping: Isometric Feature Mapping (ISOMAP) [55] is used in the estimation of geodesic distance using the shortest path in the nearest neighbors graph. This is again a feature reduction technique which is used in satellite image processing.

Post-processing Techniques: Satellite Image Classification [57] is a post-processing technique. It is the process of labeling a group of pixels based on their pixel values. In simple words, it is a method to extract the information from the satellite image in order to describe what is present in the image. Some mostly used Satellite Image Classification techniques are as follows:

Support Vector Machine algorithm: Support Vector Machine (SVM) [52] is a technique which can be used for the classification of satellite images. It makes use of support vectors to differentiate between multiple classes. It is a supervised classification technique, hence can be used for classification tasks to produce results with better accuracy.

Artificial Neural Network: Artificial Neural Networks (ANN) [52] are widely used for the purpose of classification of satellite images. They have the advantage of learning on their own using a number of neurons which make them most used technique for image classification. Moreover, they are capable of providing results with higher accuracy. Decision Tree algorithm: Decision tree (DT) [52] is one more technique which can be used for satellite image classification. The construction of decision tree requires supervised training, therefore it is necessary to have a training dataset.

K-Nearest Neighbor algorithm: K-Nearest Neighbor (KNN) [52] is a supervised technique used for satellite image classification. It computes a group of k objects in training set which are closest to the test objects. Based on the majority of neighbors in a class, the assignment of label take place.

K-means algorithm: K -means is an [52] unsupervised clustering technique which is based on the observations nearest to a cluster centroid to form k clusters. As it is a unsupervised technique, number of classes has to be determined as we do not know how many different variety of clusters will form.

Ensemble: This approach combines various classifiers together to give better results. For example, SVM, DT and KNN can be combined to give better results for classification.

6. Satellite Image Processing Challenges

Satellite image processing is a challenging task as one has to deal with a number of difficulties during the processing stage. To start with, there are lot of errors in the captured images which makes the preprocessing of the image bit tricky like captured noise, uneven brightness values and intensity values of various pixels etc. Reading and displaying the image is again a tricky task as the satellite images come in various formats like .tiff, .jp2. Hence, displaying the true color image after all the preprocessing is a bit tedious task. Extraction of the most important features associated with satellite images for the classification and analysis purpose is again a challenge as important features can be different with respect to different applications and datasets of satellite images. Also, higher computation power is always a need

ISSN: 2005-4262IJGDC

for satellite image processing. One has to be equipped with powerful GPUs (Graphics Processing Units) to process available satellite image datasets. General purpose CPUs (Central Processing Units) are not capable of processing these satellite images.
Different shapes and sizes of the various objects present in the satellite images need to be taken care of, when you need to classify those objects. Some errors get generated when a specific image is captured in daylight. For example, sunlight can be reflected through water bodies and when the sensor captures image at that moment, water bodies may look white in color similar to clouds. This leads to another challenge which is, to distinguish different objects present in the satellite image.

7. Evaluation Metrics

Alberto Garcia-Garcia at el. [51] provided some of the commonly used metrics for evaluating the models. For the significant results from semantic segmentation system, its performance must be evaluated rigorously. For fair comparisons with various existing methods evaluation metrics like Execution Time, Memory Footprint, Accuracy are used which are described below:
Execution time: Execution time is an important metric which tells about the exact time for the inference pass of various systems . Although execution time provides a critical information, it also depends on the hardware and the conditions on which the model has been trained and tested. It is important for some researchers to know the exact execution time of a particular model to get an idea about how the model will perform for a specific application.
Memory footprint: Memory usage can also be considered as an essential aspect for evaluating the segmentation models. When we talk about high-end Graphics Processing Units (GPUs) which are used for segmentation task, they also do not provide a large amount of memory. The peak and average memory footprint of a method with complete description of execution conditions can be very helpful. **Accuracy:** Many evaluation criteria have been proposed to assess the accuracy semantic segmentation models. We will provide some of the commonly used accuracy measures. A total of k + 1 classes are there and p_{ij} is the amount of pixels of class i inferred to belong to class j. In other words, p_{ii} represents the number of true positives (TP), while p_{ij} and p_{ji} are usually interpreted as false positives (FP) and false negatives (FN) respectively.
Pixel Accuracy (PA): Computation of a ratio of the amount of correctly classified pixels to the total number of pixels.

$$PA = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}} \quad [51] \qquad (1)$$

Mean Pixel Accuracy (MPA): Correctly classified pixels is computed on the per-class basis and averaged over the total number of classes present.

$$MPA = \frac{1}{k+1}\sum_{i=0}^{k} \left(\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}}\right) \quad [51] \qquad (2)$$

Intersection over Union (IoU): For each class, IoU is the ratio of correctly classified pixels to the total number of ground truth and predicted pixels in that class [33].

$$IoU = \frac{TP}{TP+FP+FN} \quad [33] \qquad (3)$$

ISSN: 2005-4262IJGDC

Mean Intersection over Union (MIoU): The IoU is computed on a per-class basis and then finally averaged.

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k}\left(\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}}\right) \quad [51] \qquad (4)$$

Frequency Weighted Intersection over Union (FWIoU): It weighs each class significance depending on their frequency of appearance.

$$\frac{1}{\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}} \sum_{i=0}^{k}\left(\frac{\sum_{j=0}^{k} p_{ij}\; p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}}\right) \quad [51] \qquad (5)$$

8. Conclusion

A literature survey has been done where we provided some of the common semantic segmentation models used and also the segmentation models which are commonly used for satellite imagery. We also provided some of the well known satellite imagery datasets, satellite image processing techniques, related challenges and in the last section we provided various evaluation metrics used for the evaluation of these models. We have tried to cover some of the significant work done by various researchers in semantic segmentation and also in semantic segmentation of satellite imagery. This work has been done to give researchers a more clear understanding of various models used for the purpose of semantic segmentation of satellite images. Future work can be designing a model for a specific semantic segmentation task like detection of roads, vehicles, water bodies, forest, urban areas and many other applications that fall under satellite image segmentation.

References

[1] B. Li, Y. Shi, Z. Qi and Z. Chen, "A Survey on Semantic Segmentation," IEEE International Conference on Data Mining Workshops (ICDMW), Singapore, 2018, pp. 1233-1240, doi: 10.1109/ICDMW.2018.00176.
[2] J. Long, E. Shelhamer, T. Darrell, “Fully convolutional networks for semantic segmentation,” IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 39(4):640-651.
[3] V. Badrinarayanan, A. Kendall, R. Cipolla, “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Scene Segmentation,” IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, PP(99):2481-2495.
[4] A. Paszke, A. Chaurasia, S. Kim, E. Culurciello, “ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation,” arXiv:1606.02147 [cs.CV], 2016.
[5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” arXiv:1512.03385 [cs.CV], 2015.
[6] Y. Guo, Y. Liu, T. Georgiou *et al.*, “A review of semantic segmentation using deep neural networks,” *Int J Multimed Info Retr* **7,** 87–93, 2018, https://doi.org/10.1007/s13735-017-0141-z.
[7] R. Girshick, J. Donahue, T. Darrell, J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” In: CVPR, 2014.
[8] J. R. R. Uijlings, K. E. A. Van De Sande, T. Gevers, A.W. Smeulders, “Selective search for object recognition,” Int J Comput Vis 104(2):154–171, 2013.
[9] A. Krizhevsky, I. Sutskever, G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” In: NIPS, 2012.

ISSN: 2005-4262IJGDC

[10] K. Simonyan, A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” In: ICLR, 2015.
[11] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, “Going deeper with convolutions,” In: CVPR, 2015.
[12] K. He, X. Zhang, S. Ren, J. Sun, “Deep residual learning for image recognition,” In: CVPR, 2016.
[13] J. Gonzalez, K. Sankaran, V. Ayma, C. Beltran, "Application of Semantic Segmentation with Few Labels in the Detection of Water Bodies from Perusat-1 Satellite’s Images," IEEE Latin American GRSS & ISPRS Remote Sensing Conference (LAGIRS), Santiago, Chile, 2020, pp. 483-487, doi: 10.1109/LAGIRS48042.2020.9165643.
[14] Y. Liu, Y. Guo, M. S. Lew, “On the exploration of convolutional fusion networks for visual recognition,” In: MMM, 2017.
[15] J. Dai, Y. Li, K. He, J. Sun, “R-FCN: Object Detection via region-based fully convolutional networks,” In: NIPS, 2016.
[16] S. Ren, K. He, R. Girshick, J. Sun, “Faster R-CNN: towards real-time object detection with region proposal networks,” In: NIPS, 2015.
[17] A. King, S. M. Bhandarkar, B. M. Hopkinson, “A Comparison of Deep Learning Methods for Semantic Segmentation of Coral Reef Survey Images,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2018, pp. 1394-1402.
[18] F. Yu, V. Koltun, “Multi-scale context aggregation by dilated convolutions,” In ICLR, 2016.
[19] L. -C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” IEEE transactions on pattern analysis and machine intelligence, 40(4):834–848, 2018.
[20] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, D. Terzopoulos, “Image Segmentation Using Deep Learning:A Survey,” In Computer Vision and Pattern Recognition, Apr. 2020.
[21] W. Liu, A. Rabinovich, A. C. Berg, “Parsenet: Looking wider to see better,” arXiv:1506.04579, 2015.
[22] O. Ronneberger, P. Fischer, T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.
[23] F. Milletari, N. Navab, S.-A. Ahmadi, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” in Fourth International Conference on 3D Vision (3DV), IEEE, 2016, pp. 565–571.
[24] L. -C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, “Semantic image segmentation with deep convolutional nets and fully connected crfs,” arXiv:1412.7062, 2014.
[25] L. -C. Chen, G. Papandreou, F. Schroff, H. Adam, “Rethinking atrous convolution for semantic image segmentation,” arXiv:1706.05587, 2017.
[26] L. -C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 801–818.
[27] F. Visin, M. Ciccone, A. Romero, K. Kastner, K. Cho, Y. Bengio, M. Matteucci, A. Courville, “Reseg: A recurrent neural network-based model for semantic segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 41–48.
[28] F. Visin, K. Kastner, K. Cho, M. Matteucci, A. Courville, Y. Bengio, “Renet: A recurrent neural network based alternative to convolutional networks,” arXiv:1505.00393, 2015.
[29] Q. Huang, C. Xia, C. Wu, S. Li, Y. Wang, Y. Song, C.-C. J. Kuo, “Semantic segmentation with reverse attention,” arXiv:1707.06426, 2017.
[30] A. A. Tiurin, M. I. Vorobiev, O. I. Lisov, A. M. Andrianov, E. S. Yanakova, "An Effective Algorithm for Analysis and Processing of Satellite Images for Semantic Segmentation," IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), St. Petersburg and Moscow, Russia, 2020, pp. 2018-2022, doi: 10.1109/EIConRus49466.2020.9039113.

ISSN: 2005-4262IJGDC

[31] D. Lary, A. Alavi, A. Gandomi, A. Walker, “Machine learning in geosciences and remote sensing,” Geoscience Frontiers, 2016, pp.3-10.
[32] M. Y. Saifi, J. Singla, Nikita, "Deep Learning based Framework for Semantic Segmentation of Satellite Images," Fourth International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2020, pp. 369-374, doi: 10.1109/ICCMC48092.2020.ICCMC-00069.
[33] M. Barthakur, K. K. Sarma, "Semantic Segmentation using K-means Clustering and Deep Learning in Satellite Image," 2nd International Conference on Innovations in Electronics, Signal Processing and Communication (IESC), Shillong, India, 2019, pp. 192-196, doi: 10.1109/IESPC.2019.8902391.
[34] M. Wu, C. Zhang, J. Liu, L. Zhou, X. Li, "Towards Accurate High Resolution Satellite Image Semantic Segmentation," in IEEE Access, vol. 7, pp. 55609-55619, 2019, doi: 10.1109/ACCESS.2019.2913442.
[35] L. Barghout, L. Lee, ‘‘Perceptual information processing system, ’’U.S. Patent 10 618 543, Mar. 25, 2004.
[36] G. Cheng, J. Han, ‘‘A survey on object detection in optical remote sensing images,’’ ISPRS J. Photogramm. Remote Sens., vol. 117, pp. 11–28, Jul. 2016.
[37] A. Huertas, R. Nevatia, ‘‘Detecting buildings in aerial images,’’ Comput. Vis., Graph., Image Process., vol. 41, no. 2, pp. 131–152, Feb. 1988.
[38] L. -C. Chen, Y. Yang, J. Wang, W. Xu, A. L. Yuille, ‘‘Attention to scale:Scale-aware semantic image segmentation,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., May 2016, pp. 3640–3649.
[39] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, Y. Bengio, ‘‘The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, Jun. 2017, pp. 11–19.
[40] Y. Wei, J. Feng, X. Liang, M. -M. Cheng, Y. Zhao, S. Yan, ‘‘Object region mining with adversarial erasing: A simple classification to semantic segmentation approach,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 1568–1576.
[41] M. Zeiler, D. Krishnan, G. Taylor, R. Fergus, ‘‘Deconvolutional networks,’’ in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Jun. 2010, pp. 2528–2535.
[42] A. Chaurasia, E. Culurciello, ‘‘Linknet: Exploiting encoder representations for efficient semantic segmentation,’’ in Proc. IEEE Vis. Commun. Image Process. (VCIP), Dec. 2017, pp. 1–4.
[43] L. Chen et al., ‘‘SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Aug. 2017, pp. 5659–5667.
[44] V. Iglovikov, A. Shvets, ‘‘Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation,’’ 2018, arXiv:1801.05746.
[45] L. Duan, X. Hu, "Multiscale Refinement Network for Water-Body Segmentation in High-Resolution Satellite Imagery," in IEEE Geoscience and Remote Sensing Letters, vol. 17, no. 4, pp. 686-690, April 2020, doi: 10.1109/LGRS.2019.2926412.
[46] Y. Yang, S. Newsam, “Bag-of-visual-words and spatial extensions for land-use classification,” In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2010, pp 270–279.
[47] P. Helber, B. Bischke, A. Dengel, D. Borth, “Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2019.
[48] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, R. Raskar, “Deepglobe 2018: A challenge to parse the earth through satellite images,” In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2018.
[49] S. J. Walsh, D. R. Butler, G. P. Malanson, “An overview of scale, pattern, process relationships in geomorphology: a remote sensing and GIS perspective,” Geomorphology 21, 183 205, 1998.
[50] X. -Y. Tong et al., “Learning transferable deep models for land-use classification with high-resolution remote sensing images,” Jul. 2018, arXiv:1807.05713.

ISSN: 2005-4262IJGDC

[51] G. -G. Alberto, O. -E. Sergio, O. Sergiu, V. -M. Victor, M. -G. Pablo, G. -R. Jose, "A survey on deep learning techniques for image and video semantic segmentation," 2018, doi:10.1016/j.asoc.2018.05.018.
[52] P. Muthu, S. S. Ranjani, " Classification techniques used in remote sensing satellite imageries: A survey," In National Journal of Multidisciplinary Research and Development, 2019, Volume 4; Issue 6; November 2019; Page No. 24-28.
[53] J. Babbar, N. Rathee, "Satellite Image Analysis: A Review," 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, India, 2019, pp. 1-6, doi: 10.1109/ICECCT.2019.8869481.
[54] S. Karim, Y. Zhang, M. R. Asif, S. Ali, "Comparative analysis of feature extraction methods in satellite imagery," J. Appl. Remote Sens. 11(4), 042618, 2017, doi: 10.1117/1.JRS.11.042618.
[55] L. Journaux, X. Tizon, I. Foucherot, P. Gouton, "Dimensionality Reduction Techniques: An Operational Comparison On Multispectral Satellite Images Using Unsupervised Clustering," Proceedings of the 7th Nordic Signal Processing Symposium - NORSIG 2006, Rejkjavik, 2006, pp. 242-245, doi: 10.1109/NORSIG.2006.275233.
[56] T. Vigneshl, K. K. Thyagharajan, "Local binary pattern texture feature for satellite imagery classification," 2014 International Conference on Science Engineering and Management Research (ICSEMR), Chennai, 2014, pp. 1-6, doi: 10.1109/ICSEMR.2014.7043591.
[57] D. R. Sowmya., P. D. Shenoy, K. R. Venugopal., "Remote Sensing Satellite Image Processing Techniques for Image Classification: A Comprehensive Survey," International Journal of Computer Applications (0975 – 8887), Volume 161 – No 11, March 2017.

ISSN: 2005-4262IJGDC