# SEMANTIC SEGMENTATION OF SATELLITE IMAGES USING TRANSFORMERS

MUHAMMAD HAZIQ FAIZ BIN MOHD RIPIN

SESSION 2021/2022

FACULTY OF COMPUTING AND INFORMATICS
MULTIMEDIA UNIVERSITY

JANUARY 2022

# SEMANTIC SEGMENTATION OF SATELLITE IMAGES USING TRANSFORMERS

BY

MUHAMMAD HAZIQ FAIZ BIN MOHD RIPIN

SESSION 2021/2022

THIS PROJECT REPORT IS PREPARED FOR

FACULTY OF COMPUTING AND INFORMATICS
MULTIMEDIA UNIVERSITY
IN PARTIAL FULFILLMENT

FOR

BACHELOR OF COMPUTER SCIENCE

B.C.S (HONS) DATA SCIENCE

FACULTY OF COMPUTING AND INFORMATICS
MULTIMEDIA UNIVERSITY

January 2022

## DECLARATION

I hereby declare that the work has been done by myself and no portion of the work contained in this Thesis has been submitted in support of any application for any other degree or qualification on this or any other university or institution of learning.

---

**Muhammad Haziq Faiz Bin Mohd Ripin**

Faculty of Computing and Informatics

Multimedia University

Date: 12:12:2022

## ACKNOWLEDGEMENTS

Thanks guys. I owe you many.

To my parents, my husband, and my daughter.

# ABSTRACT

This can be your **Management Summary** or **Abstract**. An abstract or management summary should be not more than one page in length. The abstract should allow the reader or moderator who is unfamiliar with the work to gain a swift and accurate impression of what the project is about, how it arose and what has been achieved.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

# PREFACE

The preface in a report is something that comes before the report. This section will typically set up the stage for whatever your report is going to discuss. It may give some background information on the subject.

Normally a preface it will be a three paragraph length answer. The first paragraph should be explaining what you are investigating and why. the second should be the scope of your investigation. the third should be the conclusion that your investigation brought you to.

If your report does not have any preface, you may remove it from your latex.

# CHAPTER 1

## INTRODUCTION, BACKGROUND STORY, MOTIVATIONS

## *1.1 Background Introduction*

### *1.1.1 Introduction to Semantic Segmentation*

The last few years have seen a massive surge in research regarding deep learning applications in computer vision with the most common one being object detection, where a network accept an image as an input and output either a single or multi class label. Typically, the position of detected objects are defined by rectangular coordinates that are represented by bounding boxes. However, in a lot of image processing task, such as in satellite images analysis the target output should include more accurate localization. The bounding box may have more than one objects inside it. To increase its localization, instead of assigning a set of labels to an image, semantic segmentation would label each pixel independently. After each pixel is labelled, a new image, called the mask will be produced with every pixels being coloured according to its label.

The emergence of the term “semantic segmentation” can be traced back to the 1970s (Yu et al., 2018). At that time, this terminology was equivalent to non-semantic image segmentation but emphasized that the segmented regions must contain a "seman-

tic meaning". Semantic segmentation algorithms learn information about the classes that each pixel belongs to before segmenting an image.they On the other hand, non-semantic segmentation algorithms try to detect consistent regions or region boundaries. Non-semantic segmentation can be solved using many unsupervised algorithms.

In the 1990s, “object segmentation and recognition” distinguished semantic objects of all classes from background and can be viewed as a two-class image segmentation problem. As the complete partition of foreground objects from the background is very challenging, a relaxed two-class image segmentation problem: the sliding window object detection, was proposed to partition objects with bounding boxes. However, two-class image segmentation cannot tell what these segmented objects are. As a result, the generic sense of object detection was gradually extended to multi-class image labeling, which is the present definition of semantic segmentation, to tell both where and what the objects in the scene.

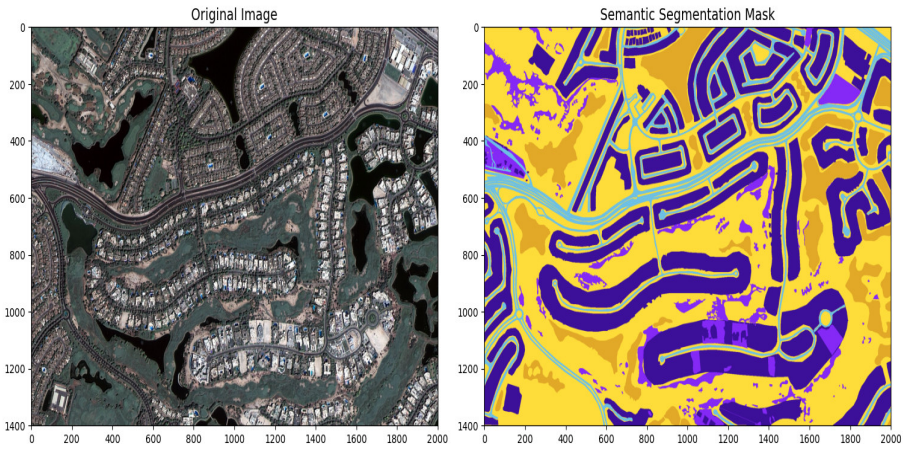


Figure 1.1: Semantic Segmentation of a Satellite Image

Figure 1.2: Object Detection of a Satellite Image

### *1.1.2 Introduction to Satellite Images*

Satellite images are images of the Earth that are collected by either drones or observation satellites. Observation satellites are satellites that are designed to observe the Earth from orbit while equipped with sensors that measure a range of electromagnetic spectrum such as UV, visible, infrared, microwave, or radio.

There are 3 types of resolution that one should consider when working with satellite images. Namely the resolutions are the spatial resolution, spectral resolution and temporal resolution.

Spatial resolution refers to the smallest feature that is displayed by an image. In most datasets it is usually represented as a single numerical value representing one side of a square pixel. For example, a spatial resolution of 10m means that a single pixel represents an area of 10 $m^2$.

Spectral resolution refers to the extent of the sensors on the satellite to detect and measure wavelengths on the electromagnetic spectrum. The finer the spectral resolution, the narrower the wavelength range for a particular channel or band. Images with high spectral resolution is important in computer vision because classes such as rock types and soil types would require an analysis at a much finer spectrum to distinguish them.

| **Spectral Bands** | **Wavelength $\mu$m** | **Description** |
|---|---|---|
| Band 1 | 400 - 450 | Least absorbed by water, and will be very useful in bathymetric studies. |
| Band 2 | 450 -510 | Provides good penetration of water. |
| Band 3 | 510 - 580 | Ideal for calculating plant vigor. |
| Band 4 | 585 - 625 | Detects the "yellowness" of particular vegetation. |
| Band 5 | 630 - 690 | Better focused on the absorption of red light. |
| Band 6 | 705 - 745 | Centered strategically at the onset of the high reflectivity portion of vegetation response |
| Band 7 | 770 - 895 | Effectively separates water bodies from vegetation, identifies types of vegetation and also discriminates between soil types |
| Band 8 | 860 - 1040 | Overlaps Band 7 but is less affected by atmospheric influence. |

Table 1.1: Different purposes of spectral bands of satellite images

Lastly, temporal resolution refers to the time period between capturing two consecutive images of the same surface area. In some literature it is also called the satellite revisit period. An image with a higher temporal resolution has a lower time period between two consecutive images. For example an image with a temporal resolution of two days means that a satellite will capture an image of the same area every two days. For some satellites, a constellation of satellites are used to increase temporal resolution. As an

example the SENTINEL-2 mission actually use 2 satellites with each having a revisit period of 10 days making the temporal resolution to be effectively 5 days. Temporal resolutions are important to detect changes that occur during a specified time period.

### *1.1.3 Applications of Semantic Segmentation of Satellite Images*

1. **Land Cover Mapping**

   Land cover mapping is the process of constructing a cover map that provide information about the Earth's surface cover pattern and land use. Such example of information provided are vegetation index and soil index. Covers maps are important for agricultural monitoring, public policy development and urban planning. Land cover mapping utilizes semantic segmentation of satellite images. Before the advances of deep learning, land cover mapping relies on traditional semantic segmentation techniques such Support Vector Machines and Random Decision Forest (Thoma, 2016). However, semantic segmentation requires a huge number of features to distinguish huge variations of land patterns. Traditional methods that only rely on low-level spectral and spatial resolution have been proven to be less optimal than its deep learning alternative due to the latter's ability to extract multilevel and multi-scale features(Yuan et al., 2020).

2. **Water Bodies Detection**

   Semantic segmentation of satellite images has long been used in detecting bod-

ies of water such as lakes and natural springs in areas where water is a scarce resource. One of the most popular traditional technique is normalized difference water index (NDWI). This technique heavily relies on IR band and measure the reflectance characteristics of water. This technique is very susceptive to noise and quite complex to develop and deploy. However, deep learning has been proven to be more reliable than NDWI, a CNN based network reached an accuracy of 99.86% (Talal et al., n.d.).

3. **Soil Erosion Detection**

   Understanding soil attributes is an important step for the construction industry. As most construction projects require excavation ((e.g.,piping, laying foundation, tunneling), soil attributes could affect the entire excavation process concerning scheduling, resource planning, procurement, claim resolution, and safety considerations. Soil classification is a process of categorizing soil based on similar attributes. The traditional involves on-site sampling and analysing the samples in a laboratory which is very time consuming and expensive. Thanks to deep learning ability to utilizes high resolution images, a CNN based network was developed to classify soil based on semantic segmentation of satellite images (Pandey, Kumar, & Chakraborty, 2021).

4. **Flood Detection and Assessment System**

   Due to climate change, flooding has quickly becoming one of the most destructive and frequent type of natural disaster, and this trend is expected to continue.

Detecting flood prior of its occurrence has been vital to save lives and minimizes financial loss. On top of that, a lot of agencies around the world require a system to assess the total destruction from flooding. The assessment method is usually done manually using aerial images. Semantic segmentation has been widely used as a tool to aid in the process of designing and deploying accurate flood detection and post-flood assessment system (Leach, Popien, Goodman, & Tellman, n.d.).

### *1.1.4 Problem Statement*

1. All of the datasets studied suffer from class imbalanced meaning that the number of classes in the samples are not equally distributed. This is a very common problem in semantic segmentation task.

2. In the majority of the papers reviewed, semantic segmentation models are not trained using High Spatial Resolution (HSR) satellite images that are balanced.

### *1.1.5 Project Objectives*

1. Identify suitable datasets of satellite images that will be used for training and validation. Multiple datasets will be evaluated and a new dataset would be constructed if necessary.

2. Design and train a transformer model to perform semantic segmentation using chosen dataset.

3. Identify appropriate evaluation metrics and use that metrics to evaluate the performance of our model so that we can provide benchmark for future experiments.

### *1.1.6 Project Scope*

The scope of the dataset used in this project will be limited to satellite images. The chosen dataset must have a spatial resolution that is small enough to avoid any losses of information. On top of that, the dataset must be bigger than 240x240 (px). The dataset must have been taken by a satellite. There is no restriction on the type of bands that the dataset can have.

The scope of the network proposed in this project must be one based on vision transformer. The vision transformer network must be able to perform semantic segmentation task of satellite images. The performance of the proposed vision transformer network shall be compared to the previous works trained on the same dataset.

### *1.1.7 Chapter Organization*

# CHAPTER 2

## LITERATURE REVIEW

This chapter would cover the literature review part of this project. The first section would elaborate on the datasets evaluated, including the data exploratory analysis of the datasets. The second and third sections would include a brief introduction to traditional methods semantic segmentation and their limitations respectively. The fourth sections would serve as a literature review of semantic segmentation using Convolutional Neural Networks. The fifth and last section would include the literature review of semantic segmentation using transformers and its advantages.

### *2.1 Satellite Images Datasets*

Table 2.1 shows the list of dataset that were evaluated and pre-processed for this project. Each of the dataset is made for semantic segmentation task.

### *2.2 Semantic Segmentation Before Deep Learning*

This section would elaborate on traditional methods of semantic segmentation, methods that do not apply any neural networks but make heavy use of domain knowledge and feature extraction methods.

| Dataset | Source | # Samples | # Classes | Size (px) | Res (m) | Band |
|---|---|---|---|---|---|---|
| Benin Cashew Plantation | Airbus Pléiades | 70 | 6 | 1,122x1,186 | 10 | MSI |
| Cloud Cover Detection | Sentinel-2 | 22,728 | 2 | 512x512 | 10 | MSI |
| Kenya Crop Trade | Sentinel-2 | 4,688 | 7 | 3,035x2,016 | 10 | MSI |
| Deep Globe Land Cover | DigitalGlobe +Vivid | 803 | 7 | 2,448x2,448 | 0.5 | RGB |
| DFC2022 | Aerial | 3,981 | 15 | 2,000x2,000 | 0.5 | RGB |
| ETCI 2021 Flood Prediction | Sentinel-1 | 66,810 | 2 | 256x256 | 5–20 | SAR |
| GID-15 | Gaofen-2 | 150 | 15 | 6,800x7,200 | 3 | RGB |
| LandCover.ai | Aerial | 10,674 | 5 | 512x512 | 0.25–0.5 | RGB |
| LoveDA | Google Earth | 5,987 | 7 | 1,024x1,024 | 0.3 | RGB |
| Potsdam | Aerial | 38 | 6 | 4,000x4,000 | 0.02 | RGB |
| Vaihingen | Aerial | 33 | 6 | 1,281–3,816 | 0.09 | RGB |
| SEN12MS | Sentinel-1/2, MODIS | 180,662 | 33 | 256x256 | 10 | SAR, MSI |

Table 2.1: Datasets for Semantic Segmentation Task

### *2.2.1 Feature Extraction*

Before we apply any of the classification method that will be discussed in the proceeding sections (SVM, Random Decision Forest, MRF, CRF) we must extract the features from an image. The accuracy of traditional semantic segmentation methods heavily depends on the selected features. The features may be the numerical value of each pixel or the feature map containing the gradient of each pixel. There are three feature extraction methods discussed in this section, namely they are Histogram of Oriented Gradients, Scale-Invariant Feature Transform and Bag of Visual Words .

1. ***Histogram of Oriented Gradients (HOG).*** HOG features interpret any given image as a discrete function $I : \mathbb{N}^2 \rightarrow \{0...255\}$ that maps a pair value $(x, y)$ which is the coordinate of the pixel to an RGB value. Then, the partial derivative in of $x$ and $y$ are calculated for every pixel. The input image is now transformed into a feature map with the gradient of each pixel. Lastly, the constructed feature map is divided into smaller patches and the direction and magnitude of the histogram is calculated for each patch. (Thoma, 2016).

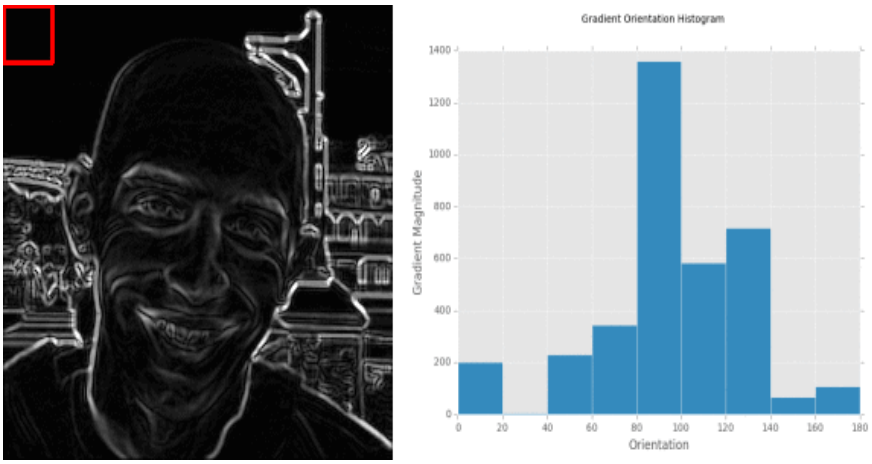


Figure 2.1: Computing a histogram of oriented gradients for the first patch of an input image.

2. ***Scale-Invariant Feature Transform (SIFT).*** SIFT is a feature extraction algorithm that was introduced in 2004. Unlike HOG, SIFT is not affected by the orientation or scale of the input image. (Thoma, 2016). An image will be divided into smaller patches and the difference-of-Gaussian (DoG)is calculated. DoG is obtained as the difference of Gaussian blurring of an image with two different $\sigma$. Next, local extrema is searched to be be assigned as potential key

points. Lastly, after the key points are discovered, an 8-bin orientation histogram is created for each patch to math the key points. The final output will be a feature map containing accepted key points. A more thorough explanation is available in the original paper (Lowe, 2004).

Figure 2.2: Each circle represents the location and orientation of SIFT keypoints

3. ***Bag of Visual Words (BOV).*** BOV construct sparse histograms that contain the frequency of features in an image. Those features are usually extracted using SIFT (Thoma, 2016). BOV is often used alongside other feature extractors such as SIFT by assigning each SIFT descriptor to the closest entry in a visual dictionary.

### *2.2.2 Random Decision Forest for Semantic Segmentation*

A decision tree is a tree where each leaf represents a class and each non-leaf nodes uses the feature inputs to decide which branch to descend to (Thoma, 2016). A random decision tree is as decision tree that is injected with some randomness during the training phase to reduce over-fitting and increase accuracy. Random Decision Forest is an unsupervised ensemble learning method that are made up of multiple independently constructed random decision trees. An in-depth explanation to semantic segmentation using Random Decision Forest is given by (Schroff, Criminisi, & Zisserman, 2008).

### *2.2.3 Support Vector Machines (SVM) for Semantic Segmentation*

In SVM the training data is represented as $(x_i, y_i)$ where $x_i$ is the feature vector, $y_i \in \{-1, 1\}$ is the class label and $i \in \{1...m\}$ where m is the number of inputs.

Assuming that the data is linearly separable, SVM is a task of solving the optimal margin classifier:

$$min_{w,b} \quad \frac{1}{2}||w||^2$$

$$s.t. \quad y^i(w^T x^i + b) \geqslant 1, i \in 1...m$$

$w$ is the linear combination of the training data $x$:

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i$$

Where $\alpha$ is the Lagrange multiplier. Not every dataset is linearly separable thus this problem can be solved by transforming the feature vectors $x$ into a higher dimension using a non-linear mapping $/psi$. Thus instead of learning using $x$, we may learn using a higher-dimensional features $\psi(x)$, this method is called the *kernel trick*. Specifically, given a feature mapping $\psi$, we define the corresponding kernel to be:

$$K(x,z) = \psi(x)^T \psi(z)$$

The SVM described above can only distinguish between binary classes.The one-vs-all strategy and the one-vs-one strategy are methods used to expand it to be a multi-class classifier.. In the one-vs-all strategy n classifiers have to be trained which can distinguish one of the n classes against all other classes. In the one-vs-one strategy $\frac{n^2-n}{2}$ classifiers are trained; one classifier for each pair of classes.

### 2.2.4 *Markov Random Field (MRF)*

MRF maps an image onto an undirected graph where each node is a pair of random variable $(x, y)$ assigned to each pixel and the edges connect adjacent pixels (Yu et al., 2018). $x$ represents the class label of a pixel and $y$ represents the RGB value of a pixel. Which means $x$ has a range of $0...n$ and y has a range of $0...255$ with $n$ being the number of classes. Every edge is assigned conditional dependencies of its connecting nodes as weight. The probability of $x, y$ can be expressed as:

$$P(x,y) = \frac{1}{Z} e^{-E(x,y)}$$

where $Z = \sum_{x,y} e^{-E(x,y)}$ and it is called the partition function whereas E is called the energy function. A commonly used energy function is $E(x,y) = \sum_{c \in C} \psi_c(x,y)$, where $\psi$ is called the clique potential (Thoma, 2016). A thorough presentation of MRF can be found in (Blake, Kohli, & Rother, 2011).

### 2.2.5 *Conditional Random Field (CRF) for Semantic Segmentation*

CRF is an extension of MRF. Instead of learning the distribution $P(x,y)$, it chooses to learn $P(x|y)$ (Thoma, 2016). There are two advantages that CRF has over MRF. The first one being it does not to estimate the distribution of x. The second advantage is the consequence of the first one, as the distribution of x is not being estimated, less

computation is required hence making CRF faster than MRF (Yu et al., 2018). CRF has the partition function $Z(x)$:

$$Z(x) = \sum_{x} P(x, y)$$

and the joint probability distribution is given as follows:

$$P(y|x) = \frac{1}{Z(x)} \prod_{c \in C} \psi_c(y_c|x)$$

CRF is often used in conjunction with neural networks as a post-processing method for semantic segmentation task. It is is used to smoothen the output mask (Teichmann & Cipolla, 2018).

## *2.3 Limitations of Traditional Methods*

1. The traditional method is simply less accurate. As an example the best traditional method back in 2015 which utilized SIFT features extraction method and Fisher Vectors, had a performance of about 25.7% error rate on semantic segmentation task on ILSVRC-2010 dataset. While AlexNet proposed by (Krizhevsky, Sutskever, & Hinton, 2012) had an error rate of 17.0% (Thoma,

2016).

2. Feature extraction method such as SIFT and Random Decision Forest require researchers to come up with a good hand-crafted feature to achieve high accuracy while good features are very hard to produce. Compared this with the automatically learned features provided by deep learning, traditional methods would require a lot more time and effort.

## *2.4 Semantic Segmentation of Satellite Images Using Convolutional Neural Networks*

Semantic segmentation task has been long dominated by Convolutional Neural network (CNN). The Fully Convolutional Network (FCN) (Long, Shelhamer, & Darrell, 2015) is the first network proven to be an effective method to extract features automatically and serves as an effective end-to-end CNN structure for semantic segmentation task. The outcome of FCN, although encouraging, appears to be coarse due to the over-simplified design of the decoder.

To tackle this problem, better CNNs were proposed such as U-Net (Ronneberger, Fischer, & Brox, 2015) which introduced the encoder-decoder framework. Since its introduction, the encoder-decoder framework has become the standard structure of satellite images segmentation network (Wang, Fang, Zhang, Li, & Duan, 2021). U-Net introduced two symmetric paths: a contracting path, which is also known as the encoder to extract local features, and an expanding path which is also called the decoder, for

extracting position. The encoder gradually apply convolutions and max pooling to reduce the resolution of the feature map, while the decoder extracts contextual information by progressively restoring the spatial resolution. At every level of the decoder skip connections are used to by concatenate the output of the decoder with the feature maps from the encoder. Figure 2.3 show the U-Net architecture. Benefiting from its translation equivariance and locality, U-Net enhances the semantic segmentation performance significantly and the encode-decoder framework has been a major influence towards the entire field.

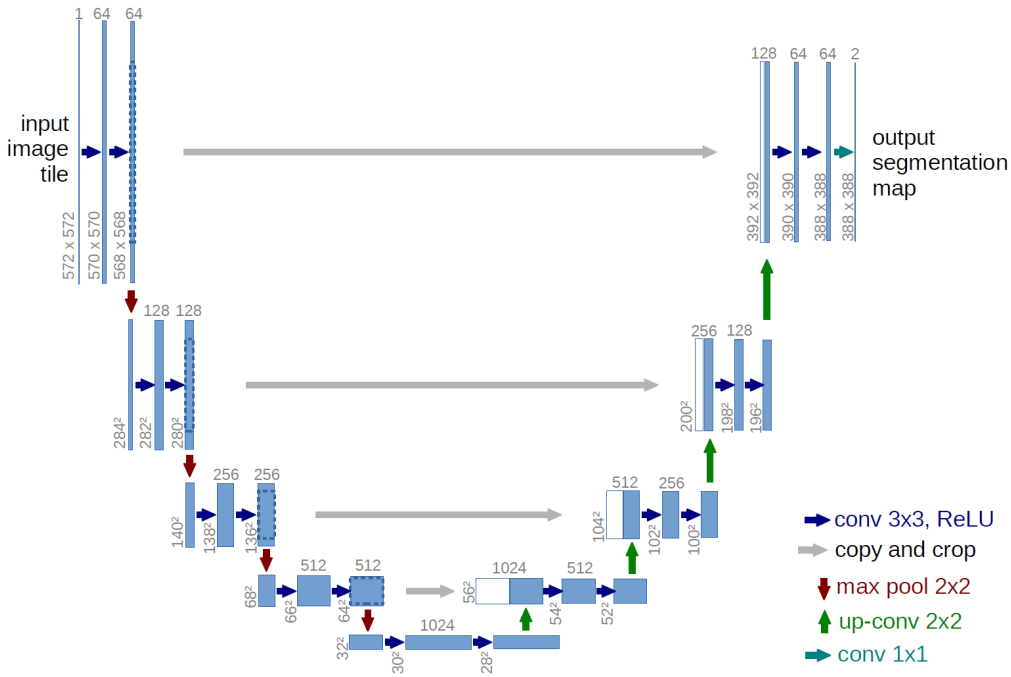


Figure 2.3: U-Net Architecture

Even though the results are promising, the long-range dependency of U-Net is limited by the locality property of the convolutional mechanism, which is critical for semantic segmentation. There are two types of methods to address the issue, either modify-

ing the convolution operation or utilizing the attention mechanism. Such examples of the first method is to enlarge the receptive fields using large kernel sizes (W. Liu, Zhang, Lin, & Liu, 2022), or utilising feature pyramids (Zhao, Shi, Qi, Wang, & Jia, 2016). On the other hand the second method focuses on integrating attention mechanisms with the encoder-decoder architecture to capture long-range dependencies of the feature maps, an example would be Attentive Bilateral Contextual network (Li et al., 2021) and Attention U-Net (Oktay et al., 2018). Although the second methods showed better performance, both methods fail to liberate the network from the dependence of the encoder-decoder structure. In the context of semantic segmentation, per-pixel classification is often ambiguous if only local information is modelled, while the semantic content of each pixel becomes more accurate with the help of global contextual information (Z. Liu, Lin, et al., 2021).

Attention U-Net (Oktay et al., 2018) is

## *2.5 Semantic Segmentation of Satellite Images Using Vision Transformers*

Transformers were introduced by (Vaswani et al., 2017) and since its inception it has been the de facto model for Natural Language Processing (NLP). Transformers that are used for image processing are called Vision Transformers to differentiate it from its NLP counterpart. Vision Transformers essentially translates 2D image-based tasks into 1D sequence-based tasks.

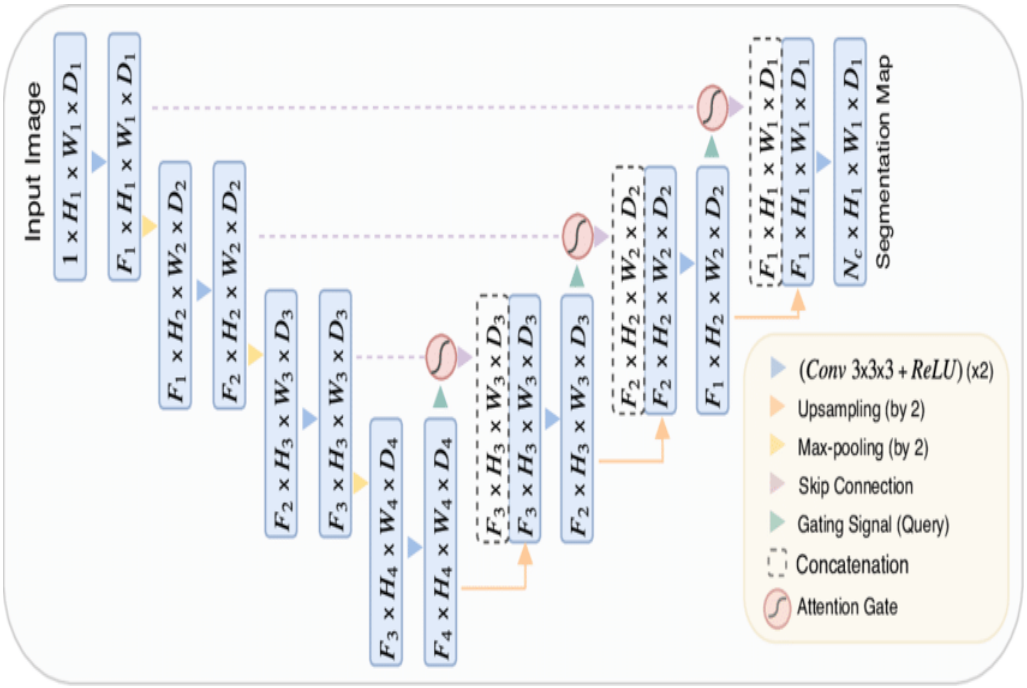


Figure 2.4: Attention U-Net Architecture

## *2.6 Advantages of Vision Transformer for Semantic Segmentation of Satellite Images*

1. ***General Modelling Capability***

   There are two aspects that gives a vision transformer general modelling capabilities. The first one being performing a task using a transformer can be interpreted as working on a fully connected graph. Any concept, can be represented by the nodes in a graph, and the relationship between concepts are represented by the graph edges.

   Every task in computer vision deals with processing two basic granular ele-

ments: pixels and objects. Thus, there are three type of relationship that can be found: pixel-to-pixel, object-to-object and pixel-to-object. The transformer's attention mechanism allows researchers to include all 3 types of relationships in one network. For examples, networks such as DETR (Carion et al., 2020), LearnRegionFeat (Gu, Hu, Wang, Wei, & Dai, 2018) and RelationNet++ (Chi, Wei, & Hu, 2020) model the relationship between object and pixel to achieve SOTA performance in semantic segmentation task.

2. ***Attention Mechanism Complements Convolution***

   Unlike convolution which is a local operation, the attention mechanism is a global one which means it can model the relationship between all the pixels in an image. This two layers complement each other very well and works such as DETR (Carion et al., 2020) and Swin Transformer V2 (Z. Liu, Hu, et al., 2021) are evidence of this claim.

3. ***Transformers are Scalable***

   Transformers has shown excellent scalability in Natural Language Processing. However, when transformers were initially used for computer vision, a lot of researchers doubted it ability to scale because all of the networks are dense as it has to process every pixel as input. Fortunately, there are recent works that shows we can improve the scalability of transformer by increasing its efficiency and reducing its computational load. Vision MoE from Google managed to match the performance of SOTA networks, while slashing the compute time into half

by using a sparse network. It managed to train a 15 billions parameter model with 90.35% accuracy on ImageNet dataset (Riquelme et al., 2021).

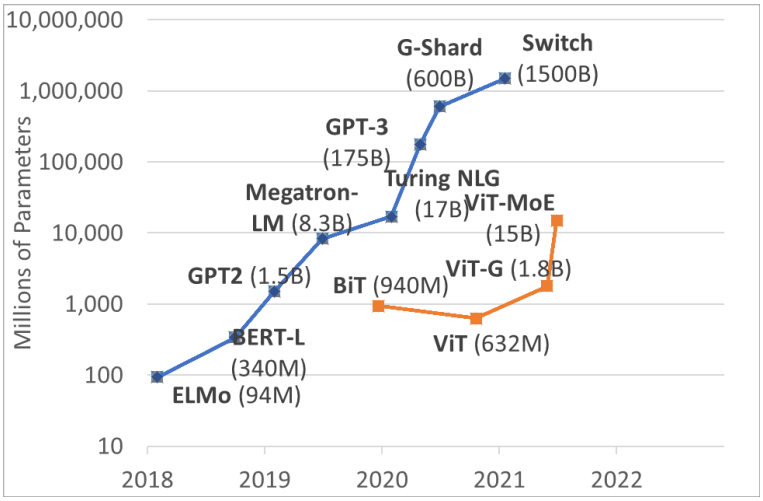


Figure 2.5: The size records of vision transformer in recent years

# CHAPTER 3

## THEORETICAL FRAMEWORK

In this chapter we would start with a brief history of artificial intelligence research that are directly or indirectly related to the innovation of transformers. Then, the second section would covers the foundations of deep learning which includes feed-forward neural networks, activation functions, loss functions and evaluation metric. The third and final section would cover the basics of transformers.

### *3.1 A Brief History of Deep Learning*

refer uday kamath

In the early 1940s, S. McCulloch and W. Pitts, using a simple electri- cal circuit called a “threshold logic unit”, simulated intelligent behavior by emulating how the brain works [179]. The simple model had the first neuron with inputs and outputs that would generate an output 0 when the “weighted sum” was below a threshold and 1 otherwise, which later became the basis of all the neural architectures. The weights were not learned but adjusted. In his book The Organization of Behaviour (1949), Donald Hebb laid the foundation of complex neural processing by proposing how neural pathways can have multiple neurons firing and strengthening over time [108]. Frank Rosenblatt,

in his seminal work, extended the McCulloch–Pitts neuron, referring to it as the “Mark I Perceptron”; given the inputs, it generated outputs using linear thresh- olding logic [212].

The weights in the perceptron were “learned” by repeatedly passing the inputs and reducing the difference between the predicted output and the desired output, thus giving birth to the basic neural learning algorithm. Marvin Minsky and Seymour Papert later published the book Perceptrons which revealed the limitations of perceptrons in learning the simple exclusive-or function (XOR) and thus prompting the so-called The First AI Winter [186].

John Hopfield introduced “Hopfield Networks”, one of the first recur- rent neural networks (RNNs) that serve as a content-addressable memory system [117].

In 1986, David Rumelhart, Geoff Hinton, and Ronald Williams pub- lished the seminal work “Learning representations by back-propagating errors” [217]. Their work confirms how a multi-layered neural network using many “hidden” layers can overcome the weakness of perceptrons in learning complex patterns with relatively simple training procedures. The building blocks for this work had been laid down by various research over the years by S. Linnainmaa, P. Werbos, K. Fukushima, D. Parker, and Y. LeCun [164, 267, 91, 196, 149].

LeCun et al., through their research and implementation, led to the first widespread application of neural networks to recognize the hand- written digits used by the U.S. Postal Service [150]. This work is a critical milestone in deep learning history, proving the utility of convolution op- erations and weight sharing in learning the features in computer vision.

Backpropagation, the key optimization technique, encountered a number of issues such as vanishing gradients, exploding gradients, and the inability to learn long-term information, to name a few [115]. Hochreiter and Schmidhuber, in their work,"Long short-term memory (LSTM)" architecture, demonstrated how issues with long-term dependencies could overcome shortcomings of backpropagation over time [116].

Hinton et al. published a breakthrough paper in 2006 titled "A fast learning algorithm for deep belief nets"; it was one of the reasons for the resurgence of deep learning [113]. The research highlighted the effective- ness of layer-by-layer training using unsupervised methods followed by supervised "fine-tuning" to achieve state-of-the-art results in character recognition. Bengio et al., in their seminal work following this, offered deep insights into why deep learning networks with multiple layers can hierarchically learn features as compared to shallow neural networks [27]. In their research, Bengio and LeCun emphasized the advantages of deep learning through architectures such as convolutional neural networks (CNNs), restricted Boltzmann machines (RBMs), and deep belief net- works (DBNs), and through techniques such as

unsupervised pre-training with fine-tuning, thus inspiring the next wave of deep learning [28]. Fei- Fei Li, head of the artificial intelligence lab at Stanford University, along with other researchers, launched ImageNet, which resulted in the most extensive collection of images and, for the first time, highlighted the usefulness of data in learning essential tasks such as object recognition, classification, and clustering [70]. Improvements in computer hardware, primarily through GPUs, increasing the throughput by almost 10× ev- ery five years, and the existence of a large amount of data to learn from resulted in a paradigm shift in the field. Instead of hand-engineered fea- tures that were the primary focus for many sophisticated applications, by learning from a large volume of training data, where the necessary features emerge, the deep learning network became the foundation for many state-of-the-art techniques.

Mikolov et al. and Graves proposed language models using RNNs and long short-term memory, which later became the building blocks for many natural language processing (NLP) architectures [184, 97]. The research paper by Collobert and Weston was instrumental in demon- strating many concepts such as pre-trained word embeddings, CNNs for text, and sharing of the embedding matrix for multi-task learning [60]. Mikolov et al. further improved the efficiency of training the word embed- dings proposed by Bengio et al. by eliminating the hidden layer and for- mulating an approximate objective for learning giving rise to "word2vec", an efficient large-scale implementation of word embeddings [185, 183]. Sutskever's research, which proposed a Hessian-free optimizer to train RNNs efficiently on long-term dependencies, was a breakthrough

in re- viving the usage of RNNs, especially in NLP [237]. Sutskever et al. in- troduced sequence-to-sequence learning as a generic neural framework comprised of an encoder neural network processing inputs as a sequence and a decoder neural network predicting the outputs based on the in- put sequence states and the current output states [238]. As a result, the sequence-to-sequence framework became the core architecture for a wide range of NLP tasks such as constituency parsing, named entity recogni- tion (NER), machine translation, question-answering, and summariza- tion, to name a few. Furthermore, even Google started replacing its monolithic phrase-based machine translation models with sequence-to- sequence neural machine translation models [272]. To overcome the bot- tleneck issues with the sequence-to-sequence framework, seminal work by Bahdanau et al. proposed the attention mechanism, which plays a crucial role in transformers and their variants [17].

## *3.2 Theoretical Introduction to Deep Learning*

### *3.2.1 Convolutional Neural Networks and Its Application in Semantic Segmentation of Satellite Images*

U-net, semua CNN jadah masuk sini.

#### *3.2.1 (a) Activation Functions*

Sigmoid, RELU, softmax

## *3.3 Introduction to Transformers*

#### *3.3.1 Encoder-Decoder Architecture*

#### *3.3.2 Sequence-To-Sequence*

#### *3.3.3 Attention Mechanism*

#### *3.3.4 Embedding and Positional Encoding*

#### *3.3.5 sdsds*

### *3.4 Transformers Architecture*

#### *3.4.1 Vision Transformer (ViT)*

Vision Transformer from (Dosovitskiy et al., 2020) is the first group of researchers that experimented with Vision Transformer by applying a standard Transformer directly to images, with the fewest possible modifications. Their model is known as Vision Transformer (ViT) as it is the first Vision Transformer. They split an image into patches and provide the sequence of linear embeddings of these patches as an input to a Transformer. The image patches were treated the same way as word tokens do in an NLP application. This methods fails to capture the translation equivariance and locality provide by the CNNs hence it is unsuitable for semantic segmentation task.

Another issue with Vision Transformer is the attention mechanism itself is $O(n^2)$ because it is a dot product thus requiring it to consume significant computational time

and memory to capture the global context, which in turn, reducing its efficiency, scaling potential and its potential for real-world applications. Figure 3.1 shows the ViT architecture.

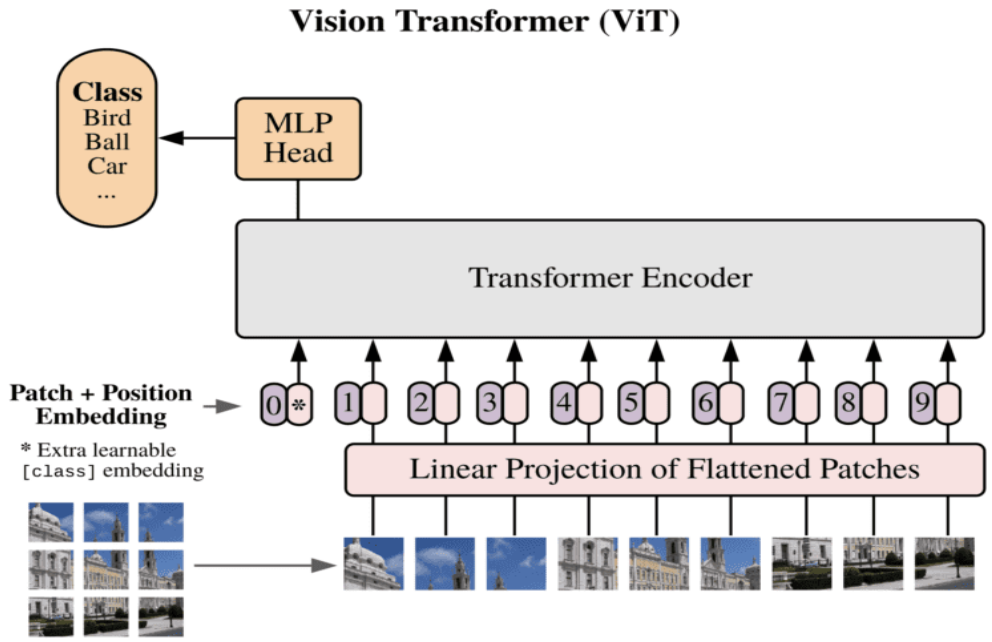


Figure 3.1: ViT Architecture

### *3.4.2 Swin Transformer*

The first version of Swin Transformer (Z. Liu, Lin, et al., 2021) presents a hierarchical feature representation scheme that demonstrates impressive performances with linear computational complexity, which makes it suitable for semantic segmentation. The detailed mechanism of Transformers and Vision Transformers are discussed in chapter 3.

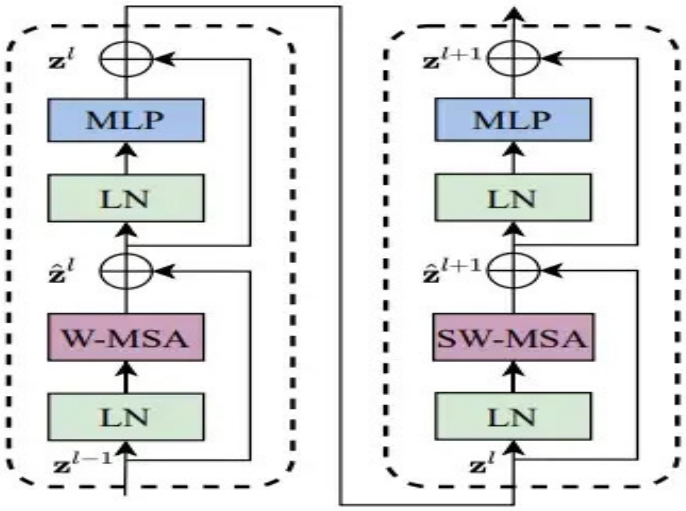


Figure 3.2: Swin Transformer V1 Architecture

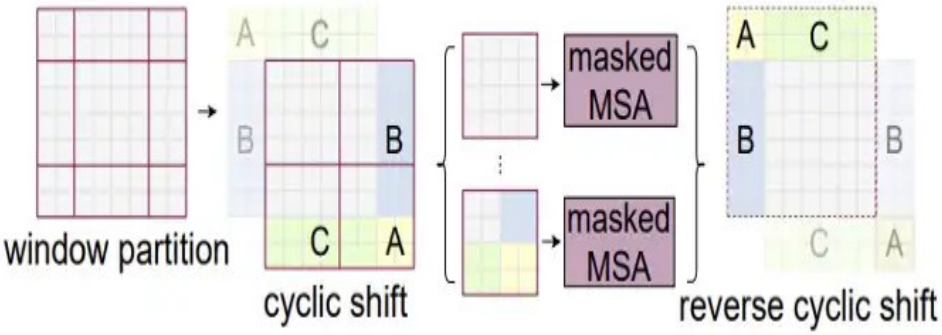


Figure 3.3: Cyclic Shifted Windows in Swin Transformer

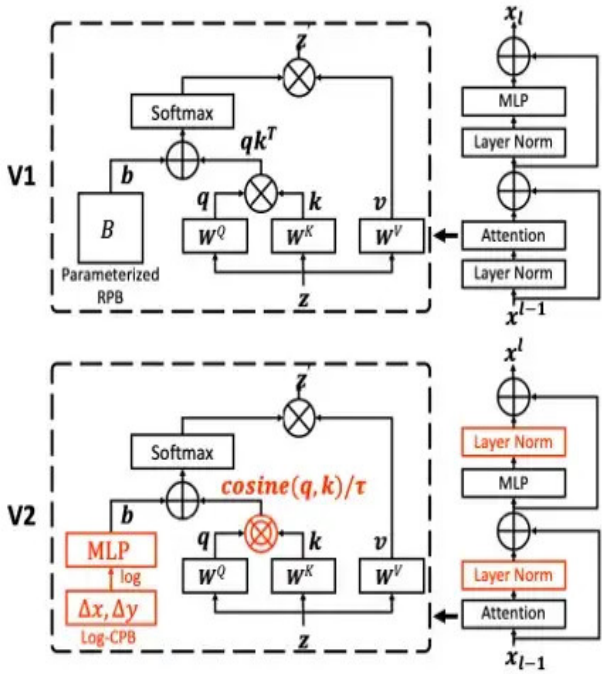


Figure 3.4: Difference Between Swin V1 and Swin V2

#### *3.4.3 SegFormer*

#### *3.4.4 MaskFormer*

#### *3.4.5 Beit*

#### *3.4.6 DeepLabV3*

### *3.5 Evaluation Metrics*

### *3.6 Limitations*

1. Imbalance class distribution.

2. Inadequate computing.

3. Quadratic computation complexity of the attention mechanism.

# CHAPTER 4

## DUMMY CHAPTER

# CHAPTER 5

## CONCLUSION

# APPENDIX A

## MANUALS, TECHNICAL SPECIFICATIONS, DOCUMENTATIONS, EXAMPLE SCENARIOS

You may want to include appendix in your report. Appendix such as manuals, technical specification, or documentations. You should **NOT** include all your source codes as appendix. Generally source code should be included in CD/DVD and **NOT** in your report.

# APPENDIX B

## APPENDIX 2: WHAT IS APPENDIX

Appendix is included in your report as it is information that is not essential to explain your findings, but that supports your analysis (especially repetitive or lengthy information), validates your conclusions or pursues a related point should be placed in an appendix (plural appendices). Sometimes excerpts from this supporting information (i.e. part of the data set) will be placed in the body of the report but the complete set of information ( i.e. all of the data set) will be included in the appendix. Examples of information that could be included in an appendix include figures/tables/charts/graphs of results, statistics, questionnaires, transcripts of interviews, pictures, lengthy derivations of equations, maps, drawings, letters, specification or data sheets, computer program information.

There is no limit to what can be placed in the appendix providing it is relevant and reference is made to it in the report. The appendix is not a catch net for all the semi-interesting or related information you have gathered through your research for your report: the information included in the appendix must bear directly relate to the research problem or the report's purpose. It must be a useful tool for the reader

# REFERENCES


Blake, A., Kohli, P., & Rother, C. (2011). *Markov random fields for vision and image processing*. Cambridge, Massachusetts: MIT Press.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. *CoRR*, *abs/2005.12872*. Retrieved from https://arxiv.org/abs/2005.12872

Chi, C., Wei, F., & Hu, H. (2020). Relationnet++: Bridging visual representations for object detection via transformer decoder. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 13564–13574). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2020/file/9d684c589d67031a627ad33d59db65e5-Paper.pdf

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, *abs/2010.11929*. Retrieved from https://arxiv.org/abs/2010.11929

Gu, J., Hu, H., Wang, L., Wei, Y., & Dai, J. (2018). Learning region features for object detection. *CoRR*, *abs/1803.07066*. Retrieved from http://arxiv.org/abs/1803.07066

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, & K. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 25). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Leach, N. R., Popien, P., Goodman, M. C., & Tellman, B. (n.d.). Leveraging convolutional neural networks for semantic segmentation of global floods with planetscope imagery. *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium, Geoscience and Remote Sensing Symposium, IGARSS 2022 - 2022 IEEE International*, 314 - 317. Retrieved from https://go.openathens.net/redirector/mmu.edu.my?url={UrlEncode(https://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=edseee&AN=edseee.9884272&site=eds-live)},Year={2022},

Li, R., Zheng, S., Zhang, C., Duan, C., Wang, L., & Atkinson, P. M. (2021). Abcnet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, *181*, 84-98. Retrieved from https://www.sciencedirect.com/science/article/pii/S0924271621002379 doi: https://doi.org/10.1016/j.isprsjprs.2021.09.005

Liu, W., Zhang, C., Lin, G., & Liu, F. (2022, sep). Crcnet: Few-shot segmentation with cross-reference and region–global conditional networks. *Int. J. Comput. Vision*, *130*(12), 3140–3157. Retrieved from https://doi.org/10.1007/s11263-022-01677-7 doi: 10.1007/s11263-022-01677-7

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., ... Guo, B. (2021). Swin transformer V2: scaling up capacity and resolution. *CoRR*, *abs/2111.09883*. Retrieved from https://arxiv.org/abs/2111.09883

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *CoRR*, *abs/2103.14030*. Retrieved from https://arxiv.org/abs/2103.14030

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *2015 ieee conference on computer vision and pattern recognition (cvpr)* (p. 3431-3440). doi:

10.1109/CVPR.2015.7298965

Lowe, D. G. (2004, November). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, *60*(2), 91–110. Retrieved from http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94 doi: 10.1023/B:VISI.0000029664.99615.94

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M. C. H., Heinrich, M. P., Misawa, K., ... Rueckert, D. (2018). Attention u-net: Learning where to look for the pancreas. *CoRR*, *abs/1804.03999*. Retrieved from http://arxiv.org/abs/1804.03999

Pandey, A., Kumar, D., & Chakraborty, D. B. (2021). Soil type classification from high resolution satellite images with deep cnn. In *2021 ieee international geoscience and remote sensing symposium igarss* (p. 4087-4090). doi: 10.1109/IGARSS47720.2021.9554290

Riquelme, C., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Pinto, A. S., ... Houlsby, N. (2021). Scaling vision with sparse mixture of experts. *CoRR*, *abs/2106.05974*. Retrieved from https://arxiv.org/abs/2106.05974

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *CoRR*, *abs/1505.04597*. Retrieved from http://arxiv.org/abs/1505.04597

Schroff, F., Criminisi, A., & Zisserman, A. (2008). Object class segmentation using random forests. In *Bmvc*.

Talal, M., Panthakkan, A., Mukhtar, H., Mansoor, W., Almansoori, S., & Ahmad, H. A. (n.d.). Detection of water-bodies using semantic segmentation. *2018 International Conference on Signal Processing and Information Security (ICSPIS), Signal Processing and Information Security (ICSPIS), 2018 International Conference on*, 1 - 4. Retrieved from https://go.openathens.net/redirector/mmu.edu.my?url={UrlEncode(https://

search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=edseee&AN=edseee.8642743&site=eds-live)},Year={2018},

Teichmann, M. T. T., & Cipolla, R. (2018). Convolutional crfs for semantic segmentation. *CoRR*, *abs/1805.04777*. Retrieved from http://arxiv.org/abs/1805.04777

Thoma, M. (2016). A survey of semantic segmentation. *CoRR*, *abs/1602.06541*. Retrieved from http://arxiv.org/abs/1602.06541

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *CoRR*, *abs/1706.03762*. Retrieved from http://arxiv.org/abs/1706.03762

Wang, L., Fang, S., Zhang, C., Li, R., & Duan, C. (2021). Efficient hybrid transformer: Learning global-local context for urban sence segmentation. *CoRR*, *abs/2109.08937*. Retrieved from https://arxiv.org/abs/2109.08937

Yu, H., Yang, Z., Tan, L., Wang, Y., Sun, W., Sun, M., & Tang, Y. (2018). Methods and datasets on semantic segmentation: A review. *Neurocomputing*, *304*, 82-103. Retrieved from https://www.sciencedirect.com/science/article/pii/S0925231218304077 doi: https://doi.org/10.1016/j.neucom.2018.03.037

Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., ... pei Zhang, L. (2020). Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sensing of Environment*, *241*, 111716.

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2016). Pyramid scene parsing network. *CoRR*, *abs/1612.01105*. Retrieved from http://arxiv.org/abs/1612.01105

**NOTES**

## PUBLICATION LIST