CHAPTER 2

LITERATURE REVIEW

2.1 Overview of Literature Review

By and large, urban greeneries in Malaysia are managed only very indirectly through the management of open space coverage. However, this is hardly an effective measure to optimize the effects of tree coverage in the urban environment. Thus, various studies on urban forests were published to investigate various effects of urban greeneries, e.g. in ecology, biodiversity, accessibility, etc. In this project, the Green View Index (GVI) measured by using street view images along with machine learning algorithms is proposed as another metric to augment the urban planning process in Malaysia.

While previous studies have successfully established the methods to measure and compute GVI in cities, there is still room for improvements in this area, particularly in the context of Malaysian urban areas.

In this literature review, we will study the research questions in several different sections. Firstly, we will establish the background context by reviewing the existing studies of urban greeneries coverage in Malaysia. Later, we dive into the literature discussing the importance and the methods to measure GVI. Lastly, algorithms and models developed in previous studies to detect and compute GVI from the street views are explored to find out the most efficient way to adopt for this project. The extracted insights will provide a foundation for our study, which aims to investigate the GVI measurement in Johor Bahru city center.

2.2 Research on Urban Greeneries in Malaysia

In Malaysia, the management of urban greeneries are based crudely on the measurement of open and green space (land specified as public area, loosely correlated with green area) coverage according to the Department of Town and Country Planning, Peninsular Malaysia

(JPBD). Several standards are established, e.g., 2 hectares of open space are reserved for each 1000 urban populations, green areas to be established as buffer zones to limit urban development. While these measurements are essential for sustainable urban planning, they are not sufficient to optimize the urban greeneries.

While there is a lack of urban greeneries management practice, there are various research being done in Malaysia attempting to fill in the gap. In a systematic review done on urban forestry research with PRISMA framework (Rajoo et. al., 2021), there is a consistent growth in the research concerning urban forests in Malaysia from 2007 onwards. Nevertheless, out of the 43 records reviewed, only 4 of them were focused on the spatial analysis of urban green space (Kanniah, 2017; Masum et. Al., 2017; Kasim et. Al., 2019, Nor et. Al., 2019). This shows an untapped opportunity where the spatial analysis on urban green coverage in Malaysia can be investigated.

On the topic of urban green spatial analysis in Malaysia, the research done are typically conducted based on aerial images on macro or micro planning scale. For instance, on a high-level master planning scale, Kasim et. al. (2019) published a study that documented the changes in urban green spaces between 2002, 2012 and 2017 with the use of high-resolution aerial imagery; Kanniah (2017) made use of time-series Landsat satellite imagery to monitor green cover changes in Kuala Lumpur from 2001 to 2016. On the lower-level planning, Ludin and Rusli (2009) monitored the quality and distribution of open spaces in Johor Bahru Tengah Municipal Council with remote sensing data.

However, all of the studies mentioned above are investigated with aerial top-down view for instrumental planning. However, top-down view studies are abstract and not directly informative for the environmental design of the everyday human experience, making it not relatable to the end-user experience. To the best of our knowledge, there is no objective study done from the perspective of urban environment users, which is a crucial measurement central to the planning of experience of the users in the urban environments. This leaves a research opportunity for such a project to happen.

2.3 Green View Index (GVI) with Street Views

Conventionally, aerial remote data has been used for the task of trees mapping or any other similar land surveying tasks. Even with the advancement of new techniques such as remote sensing methods such as LiDAR, aerial top-down view remains to be the main perspective in which urban greeneries coverage is measured. However, we have all known that the top-down view is hardly how we humans experience our environment as we perceive it in a perspective view, making the studies being hard to relate for the end-user experience of urban dwellers. This is especially relevant in urban environments where man-made facilities are juxtaposed with the mixture of trees, forming urban treescapes with massively different appearances for the same green area.

Many studies have attempted to study the visual impact of urban trees. However, a lot of them were qualitative and subjective in nature (J. Yang et. al., 2009). One of the most popular methods used were ranking, in which the participants of the survey were shown pictures or videos of urban forests and requested to score the pictures. It does not take much to understand the limitations of the studies: qualitative, subjective and unscalable. J. Yang et. al. (2009) established a new metric called Green View Index (GVI) which measures the amount of greenery that people can see on the ground at different locations in a city, laying the foundation for all the future quantified studies of visual effect of urban treescapes.

The index calculation was simple but ingenious. On strategically sampled points on a target site, eye-level photographs were taken at each point and the GVI index of each point was then interpreted by calculating the ratio between areas containing foliage over the areas of whole photos. In short, it can be defined as the ratio of greenery within the people's field of view with a range between 0 and 1, of which 0 represents no greeneries at all and 1 means that the image is full of greeneries. Due to the streets being the main public space where urban dwellers experience, Green View Index (GVI) is typically measured by using images taken from the street. The formula to calculate GVI is shown in Equation (2.1) below. M refers to the total number of horizontal directions taken on a single point, N refers to the number of vertical view angles for each sample site; $Area_g$ refers to the area of a GSV image covered by greeneries, while $Area_g$ refers to the total area of a single GSV image.

$$GVI = \frac{\sum_{j=1}^{N}\sum_{i=1}^{M} Area_g}{\sum_{j=1}^{N}\sum_{i=1}^{M} Area_t} \tag{2.1}$$

Equation (2.1) shows the formula to calculate GVI.

After the establishment of GVI, it has since become the foundation for future researchers to conduct quantitative studies on urban treescapes for landscape and urban planning studies. With the development of services like Google Street View, GVI has become even more viable as a technique to map urban green view and become an increasingly popular metric for urban green space research. Table 2.1 displays several examples of the use of GVI in various studies and their applications.

Table 2.1 Examples of other GVI studies and their applications.

Application	Study	Discussion
Investigate green view distribution	View-based greenery: A three-dimensional assessment of city buildings' green visibility using Floor Green View Index (Yu et. al. 2016)	GVI is modified to Floor Green View Index (FGVI) to quantify the area of visible urban vegetation from a certain floor of building.
	How green are the streets? An analysis for central areas of Chinese cities using Tencent Street View (Long and Liu, 2017)	Analyze street greeneries in 245 central area in Chinese Cities and detect patterns of street green view distribution.
	Treepedia 2.0: Applying Deep Learning for Large-scale Quantification of Urban Tree Cover (Cai et. al., 2018)	Develop deep learning method to compute GVI with significantly higher accuracy.
City walkability	Analyzing the effects of Green View Index of neighborhood streets on walking time using Google Street View and deep learning (Ki and Lee, 2021)	Use deep learning to compute GVI and find out its relationship with walking time.
Socioeconomic investigation	Who lives in greener neighborhoods? The distribution of street greenery and its association with residents' socioeconomic conditions in Hartford, Connecticut, USA (Li et. al., 2015)	Aggregate GVI at the block group level to compare differences in GVI based on the socioeconomic status of an area.

2.4 GSV configurations for Street Views Collection

One of the greatest limitations of J. Yang et. al.'s study was the excessive resources required for data collection on large scale as it involved extensive manual labor to take street view photos. With the advancement of mapping services, it has become a lot easier for us to retrieve our needed street view images via Google Street View API without needing to manually take the photos of tree coverage on the street. Other than the ability to collect a larger amount of data easily, the use of Google Street View (GSV) also allows us to be more precise on our camera settings to get unbiased data for urban street treescape mapping. However, the use of GSV requires its users to have a clear understanding on the image taking configurations for the image collection.

One of the good references for this is the street-level urban greenery assessment conducted by Li et. al., 2017. They sampled 258 points of street view data collection randomly across its study site, with at least 100m intervals on average between each point. To ensure each sampling point includes all the green areas that a pedestrian can possibly see, they took 3 panoramic views upwards, straight and downwards. In the Google Street View API parameters, the "heading" was set to 0, 60, 120, 180, 240, and 300 respectively to capture a full 360-degree panorama; the "pitch" was set to −45, 0, and 45 to capture different views and 18 images were taken at each sampling point.

In fact, other than the method mentioned above, there have been many street view image configurations used in different research for different purposes. Dong et. al. (2018) reviewed multiple street view configurations used by previous research. In their review, they used Tencent Static Image (TSV) service to simulate all the GSV configurations coming up in other research.

Table 2.2 Different configurations of GSV

Configuration Name	Description	Discussion
GVI4	GVI computed with 4 horizontal TSV images with heading angle of 90◦. (Long and Liu, 2017)	Greater numbers of headings allow for a lower field of vision (zoomed in), but also takes up more computational resources to compute.
GVI	GVI computed with 6 horizontal TSV images with heading angle of 60◦. (Zhang and Dong, 2018)	
GVI8	GVI computed with 8 horizontal TSV images with heading angle is 45◦ (Dong et. al., 2018)	
GVI18	Similar approach as method used by Li et. al. GVI computed with 6 horizontal TSV images with heading angle of 60◦ and 3 pitches for each heading. 18 images taken in total. (Li et. al., 2015)	Very comprehensive, but computationally expensive.
PGVI	GVI computed with cylindrical panorama stitched together from 6 horizontal TSV pictures. The distortion for PGVI might affect accuracy of GVI. (Cheng et. al., 2017)	Cylindrical panorama cause distortion in views.

While there are many configurations used, it is important that one experiments with different configurations and chooses a configuration that can minimizes distortion and overlap to reduce error in GVI estimation.

2.5 Prediction Models for GVI

Yang et. al. (2009) made GVI calculations by manually selecting areas containing foliage using Adobe Photoshop selection, which is time and resource consuming and requires lots of automation.

2.5.1 Pixel Segmentation Model

Li et. al. (2015) assessed the tree coverage in each image by extracting the green pixels with the Pixel Segmentation method, As pixels consisting of vegetation have typically higher reflectance values in the green band than red and blue, pixels consisting of greeneries can be extracted by selecting only the pixels with higher values in green band than red and blue. Nonetheless, it risks inaccuracies by including non-tree green objects and omitting shaded parts of trees that appears to be not green enough.

Figure 2.1 example of green pixels that are not actual greeneries.

2.5.2 Deep Learning Models to Predict GVI

With the advancement of deep learning and computer vision models, semantic segmentation can be deployed to detect and mask the greeneries in street view images effectively. One of the most common deep learning algorithms for the task is convolutional neural network (CNN). Convolutional neural networks (CNNs) are a type of specialized neural network for processing data with a grid-like topology (LeCun et al., 1997) that have been widely used for image classification and other computer vision tasks by learning hierarchical representations of the data through the use of convolutional layers (Katole et al., 2015).

The basic building block of CNN is a convolutional layer. It extracts features from the input data by applying a set of learnable filters to the input data. These filters are designed to learn different features at different scales, such as edges, corners, textures, and patterns to produce a set of feature maps. By stacking up the convolutional layers in a neural network, the architecture allowed the use of hierarchical representations to effectively capture the spatial relationships between pixels in an image and learn complex patterns and features by itself as shown in Figure 2.2.

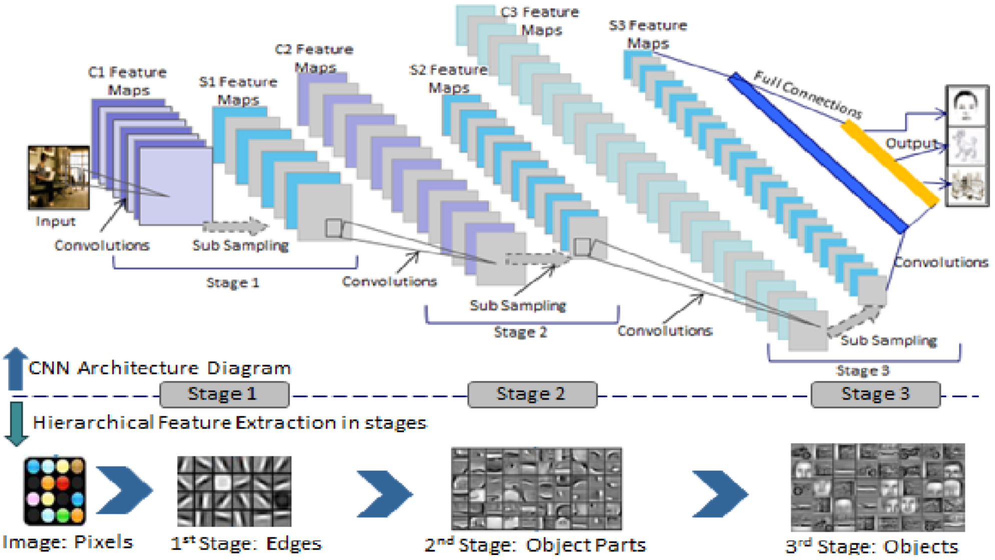


Figure 2.2 Diagram illustrating working principles of CNN.

Source: https://www.semanticscholar.org/paper/Hierarchical-Deep-Learning-Architecture-For-10K-Katole-Yellapragada/f78e280123b1c0c68f84da3cc6c66615f6e7cebd/figure/0

On top of CNN, many deep learning models are developed for image segmentation, e.g. ResNet (He et. al., 2017), FCN (Long et. al., 2015), SegNet (Badrinarayanan et. al., 2017), etc.

While CNN is the most commonly used architecture for image segmentation, there are many different models based on different architecture as well. For instance, recurrent neural networks (RNNs) based models, attention-based models, encoder-decoder based models, and others (Minaee et. al., 2022). As there are too many deep learning models developed for image segmentation, it is beyond the scope of this study to conduct an exhaustive review on every single model. So, this review will include only models that have been used to compute GVI.

Table 2.3 Summary of the deep learning models used to compute GVI.

Model	Method	Training and Calibration	Prediction Output
DCNN semantic segmentation (Cai et. al., 2018)	Pyramid Scene Parsing Network (PSPNet) (Zhao et. al., 2016) with 65,818,363 parameters	Pre-trained on full Cityscapes dataset, then trained on Cityscapes dataset, and finally on 320 GSV images collected by Cai et. al.	Pixel-segmented GSV image
DCNN end-to-end (Cai et. al., 2018)	Deep Residual Network (ResNet) (He et. al., 2017) with 28,138,601 parameters	Pre-trained on ImageNet dataset, then trained on Cityscapes dataset, and finally on 320 GSV images collected by Cai et. al.	Single GVI value between 0 and 1
HRNet-OCR (Zhang and Hu, 2022)	HRNet-OCR model (Yuan et. al., 2019) with 10,500,000 parameters	Trained on Cityscape dataset without calibration	Pixel-segmented GSV image
FCN-8 (Yu et. al., 2021)	Details of models not provided	Trained on ADE20K dataset, fine tuned with GSV images collected by author.	Pixel-segmented GSV image

DeepLabV3+ (Xia et. al., 2021)	Based on DeepLabV3+ model proposed by Chen et. al., 2018)	Trained on Cityscapes dataset, fine tuned with GSV images collected by author.	Pixel-segmented GSV image

As the details of FCN-8 model are not included in the study, the accuracy measure is therefore not included in the accuracy comparison of models in Table 2.4.

Table 2.4 Accuracy comparison between both models

Model	Mean IOU (%)	Mean Absolute Error (%) compared with true GVI	Pearson Correlation Coefficient with true GVI	5-95% of GVI Estimation Error
Pixel Segmentation (Li et. al.)	44.7	10.1	0.708	-26.6, 18.7
DCNN semantic segmentation (Cai et. al., 2018)	61.3	7.83	0.830	-20.0, 12.37
DCNN end-to-end (Cai et. al., 2018)	NA	4.67	0.939	-10.9, 7.97
HRNet-OCR (Zhang and Hu, 2022)	80.6	NA	NA	NA
DeepLabV3+ (Xia et. al., 2021)	78.37	NA	NA	NA

According to the comparison of accuracies, we can see that HRNet-OCR (Zhang and Hu, 2022) has the highest mean IOU (%) and DCNN end-to-end model (Cai et. al., 2018) has the lowest mean absolute error compared with the true GVI. Nonetheless, due to the unavailability to retrieve HRNet-OCR model proposed, DeepLabV3+ model, the next best model in terms of mean IOU and DCNN are developed and studied in this project.

2.5.2.1 DCNN end-to-end model

DCNN end-to-end model proposed by (Cai et. al., 2018) is based on a a 50 layered deep residual network (ResNet) architecture (He et. al., 2015) by adding 3 more layers of dense connections at the end with the final layer consisting of a single sigmoid unit. Instead of the two-step process of pixel-wise segmentation of greeneries and GVI computation, this model is designed to directly estimate GVI as its output. Therefore, a sigmoid function is used for the final layer as the logistic regression function returns value between 0 and 1, which is the same range as the GVI.

To understand the deeper mechanism of this model, it is important to get know about the ResNet architecture. ResNet is a modified version of the plain convolutional neural network to deal with the degradation problem that is common amongst very deep neural network with many layers, i.e., the accuracy of network gets saturated with increasing depth and degrades rapidly beyond the saturation point. This impedes the improvements of model performance by adding more layers to the model. He et. al. suggests that the issue is caused by the solvers having difficulties in approximating identity mappings by multiple non-linear layers. This is because “if the added layers can be constructed as identity mappings, a deeper model should have training error no greater than its shallower counterpart”, which was proven not to be the case in their experiment.

ResNet architecture deals with the challenge by introducing shortcut connections that directly connecting the input data to the output of the stacked layers that acts as identity mapping that plain deep neural network has problem with. Formally, denoting the desired underlying mapping as $H(x)$, we let the stacked nonlinear layers fit the residual mapping of $F(x) = H(x)-x$ instead of the desired mapping of $H(x)$. The original mapping is recast into $F(x)+x$. While both functions asymptotically approximate the desired function, the residual mapping is easier to optimize as it is easier to just push $F(x)$ to 0 in the case of identity mapping. The building block of residual learning is displayed in Figure 2.3, followed by a diagram displaying how ResNet works in Figure 2.4.

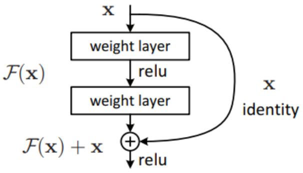

x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 2.3 Building block of residual network.

source: [1512.03385] Deep Residual Learning for Image Recognition (arxiv.org)

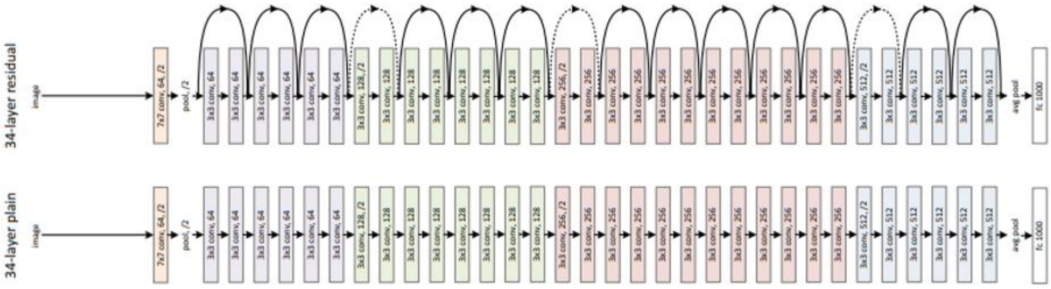


Figure 2.4 Residual network architecture diagram vs plain network.

source: [1512.03385] Deep Residual Learning for Image Recognition (arxiv.org)

Moving back to the discussion on DCNN end-to-end model, it has 28,138,601 parameters trained and fine-tuned with several datasets. The process of model training was first initialized with weights for ResNet that have been pretrained on the ImageNet dataset, then pre-trained with the transformed Cityscapes dataset and associated true GVI labels, and finally trained on small labelled GSV dataset collected by Cai et. al.

2.5.2.2 HRNet-OCR Model

High-Resolution Network-Object Contextual Representation (HRNet-OCR) is a stacked image segmentation model that combines HRNet and OCRNet used by Zhang et. al. in 2022 for GVI computation. The motivation for choosing such a method for semantic segmentation is that the HRNet can be used to find out meaningful semantic features and the OCRNet explicitly transforms the pixel classification problem into an object region classification problem (Yuan et al., 2020).

HRNet serves as the backbone of the model as the computation of GVI is a position-sensitive vision problem that can be improved significantly with high-resolution representations. Nevertheless, existing deep convolutional neural networks (DCNNs) frameworks are based on low-resolution representation subnetwork that is formed by connecting high-to-low resolution convolutions in series, and recover the high-resolution representation from the encoded low-resolution representation. Compared with DCNNs, HRNet improves the performance in position-sensitive image segmentation task by maintaining high-resolution representations through the whole process, allowing for a semantically richer and spatially more precise representation (Wang et. al., 2020). The improvements are enabled via two key characteristics: (i) Connect the high-to-low resolution convolution streams in parallel maintain the high resolution instead of recovering high resolution from low resolution, and accordingly the learned representation is potentially spatially more precise*; (ii) Repeated fusions of representations from multi-solution streams generate reliable high-resolution representations with strong position sensitivity*. In short, HRNet achieves both complete semantic information and accurate location information by parallelizing multiple branches of the resolution, coupled with the constant interaction of information between different branches (Sun et al., 2019). *The architecture of HRNet is illustrated in Figure 2.5.*

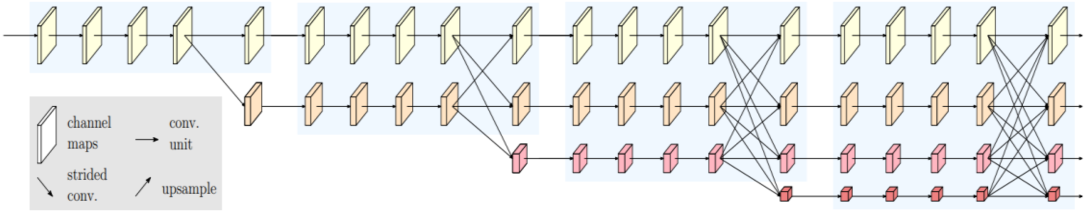


Figure 2.5 HRNet basic architecture diagram.

source: 1908.07919v2.pdf (arxiv.org)

Meanwhile, object contextual representation (OCR) is attached to the HRNet to enhance the performance of the image segmentation model. The main idea of OCR is consistent with the original definition of the semantic segmentation problem, i.e. the class of each pixel is the class of the object to which the pixel belongs (Yuan et. al., 2021). In other words, OCR takes the context of each pixel into account when assigning a class label to it. In contrast with the previous relational context schemes that consider the contextual pixels separately and only exploit the relations between pixels and contextual pixels or predict the relations only from pixels without considering the regions, the proposed approach structures the contextual pixels into object regions and exploits the relations between pixels and object regions (Yuan et. al., 2021). By incorporating object information into the network, the resultant model can better capture the context of each pixel in the image and improve the accuracy of semantic segmentation.

On top of the HRNet backbone, the contextual pixels are divided into a set of soft object regions with each corresponding to a class, i.e., a coarse soft segmentation learned under the supervision of the ground-truth segmentation. The representation of each object region is then estimated by aggregating the representations of the pixels in the corresponding object region. Lastly, the representation of each pixel is augmented with the object-contextual representation (OCR), which is the weighted aggregation of all the object region representations with the weights calculated according to the relations between pixels and object regions.

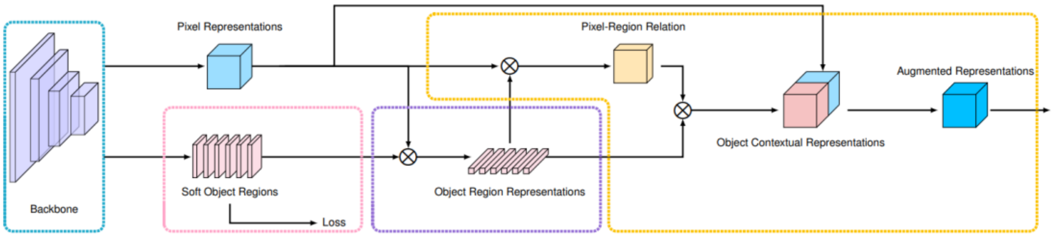
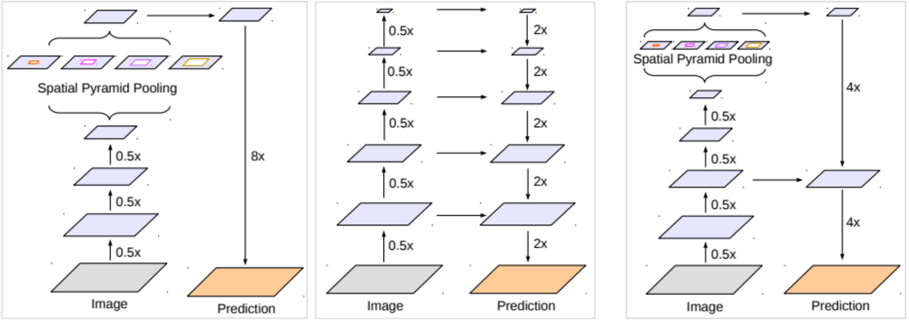

Pixel Representations
Pixel-Region Relation
Augmented Representations
Object Contextual Representations
Soft Object Regions
Object Region Representations
Backbone
Loss

Figure 2.6 OCR basic architecture diagram.
source: https://arxiv.org/pdf/1909.11065.pdf

Zhang et. al. downloaded 5000 images with the correlated fine label in Europe open source labelled dataset Cityscapes to develop their segmentation model used for GVI calculation. The entire dataset is divided into train, validation, and test sets with the ratio of 75:10:15 respectively.

2.5.2.3 DeepLabV3+ Model

DeepLabV3+ model is a deep learning model that integrates spatial pyramid pooling module in an encode-decoder structure used for image segmentation tasks, where the former extracts various levels of contextual information with pooling at different resolutions and the latter to find out sharp object boundaries (Chen et. al., 2018). This can be illustrated clearly in Figure 2.7.



(a) Spatial Pyramid Pooling (b) Encoder-Decoder (c) Encoder-Decoder with Atrous Conv

Figure 2.7 Components and basic architecture to develop DeepLabV3+ model.
Source: 1802.02611v3.pdf (arxiv.org)

Spatial Pyramid Pooling is pooling layers applied after the convolution layers to extract features at various resolutions to capture rich contextual features at different levels as shown in Figure 2.7 (a). Meanwhile, encode-decoder is made up of an encoder module that reduces the feature maps to extract semantic information and a decoder module that gradually up-samples the features as shown in Figure 2.7 (b). The combination of encoder and decoder forms a versatile structure that enables models to capture semantic information and make predictions accurately. On top of

On top of the integration of spatial pyramid pooling and the encode-decoder structure, atrous convolution, i.e. convolution layers with different dilation rates are used instead of conventional convolution layer to capture contextual information of multiple scales. The combination of components mentioned are illustrated in Figure 2.7 (c) to form DeepLabV3+ architecture.

2.6 Performance Measure for GVI Prediction

Cai et. al. (2018) proposed two metrics to evaluate the performance of tree cover estimation, i.e. mean Intersection over Union (IOU) to measure the accuracy of the location of labeled greeneries pixels, and mean absolute error (MAE) for the accuracy of overall GVI.

In computer vision, the mean IOU is commonly used for object detection and segmentation. It is defined as the ratio between the area where predicted objects or pixels overlap with the target object over the area of the union of both (Padilla et. al.), see Figure 2.8. In the context of GVI computation, the predicted and target values are both the pixels consisting of greeneries in the street view images.

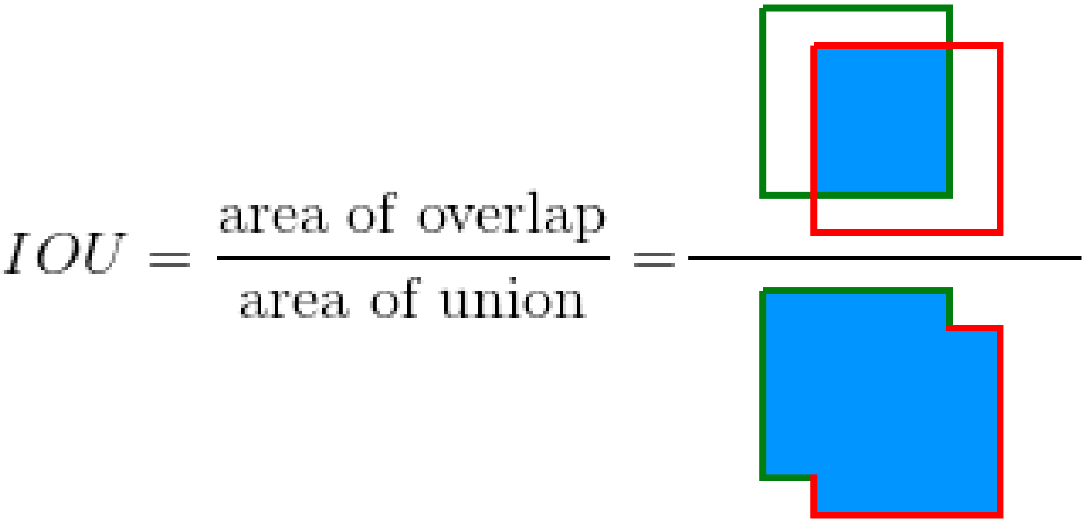


Figure 2.8 Illustration of the intersection over union (IOU)
Source:
https://www.researchgate.net/publication/343194514_A_Survey_on_Performance_Metrics_for_Object-Detection_Algorithms

The IOU of greeneries pixels can be calculated with formula shown in Equation (2.2). n is the number of images in test data, TP_i refers to true positive predicted greeneries label in image i, FP_i refers to the false positive predicted greeneries label in image i, FN_i refers to the false negative predicted greeneries label in image i.

$$IoU = \frac{1}{n} * \sum_{i=1}^{n} \frac{TP_i}{TP_i + FP_i + FP_i}$$

(2.2)

Equation (2.1) shows the formula to calculate IOU.

As for the accuracy of GVI value, it can be measured with the usual evaluation metrics used in regression. For instance, root mean square error (RMSE) proposed and mean absolute error (MAE) by (Dong et. al., 2018) and (Cai et. al., 2018) respectively. A model with the lowest MAE or RMSE has the highest accuracy. Their formulas are as shown below as Equation (2.3), Equation (2.4). $\widehat{y_\iota}$ refers to the predicted label of a pixel, y_i refers to the true label of a pixel. M refers to the number of pixels in a GSV image, n refers to the number of images in test set.

$$RMSE_j = \sqrt{\sum_{i=1}^{M} \frac{(\widehat{y_\iota} - y_i)^2}{M}}, \qquad \overline{RMSE} = \frac{1}{n}\sum_{j=1}^{n} RMSE_j$$

(2.3)

Equation (2.3) shows the equation to calculate RMSE.

$$MAE_j = \frac{1}{M}\sum_{i=1}^{M} |y_i - \widehat{y_\iota}|, \qquad \overline{MAE} = \frac{1}{n}\sum_{j=1}^{n} |MAE_j|$$

(2.4)

Equation (2.4) shows the equation to calculate MAE.

Other than measuring the accuracy of prediction, Pearson correlation coefficient (r) is used to evaluate whether the predicted GVI can accurately model the underlying patterns in the true GVI. The calculation of Pearson correlation coefficient is demonstrated in Equation (2.5), where r is the correlation coefficient, x_i is the predicted GVI value for image i, $\overline{x}$ is the mean predicted GVI, y_i is the true GVI for image i, $\overline{y}$ is the mean true GVI.

$$r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}}$$

(2.5)

Equation (2.5) shows the formula to calculate Pearson correlation coefficient (r).

Finally, there is another metric used to evaluate the variance and distribution of the difference between predicted and true GVI, that is 5-95% Estimation Error. For this metric, closer the central value is to 0 and smaller the range of the value is deemed to be better performing.

2.7 Interpretation of GVI

GVI is defined to be the green view that one can see at a single point, typically on a street level. By combining the GVI extracted from many sampling points, one can connect the dots and come up with a comprehensive plot of the visibility of urban greeneries in the studied site. For instance, the Treepedia project by the MIT Senseable City Lab is mapping GVI across many major cities in the world to explore their green distributions as shown in Figure 2.9.

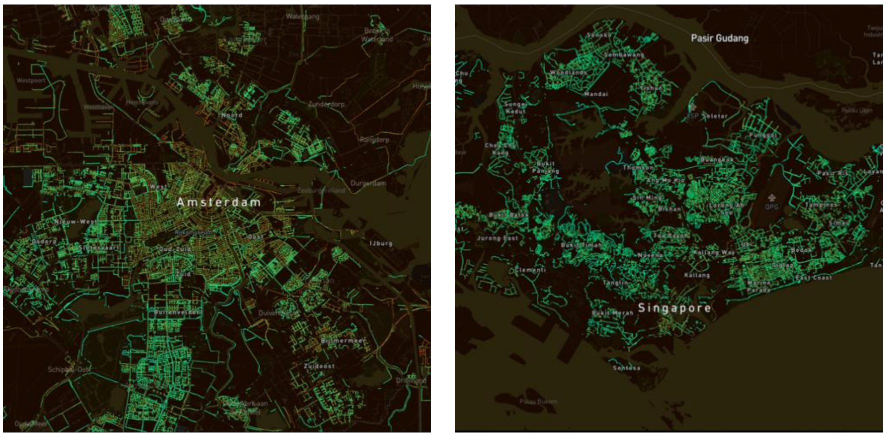


Figure 2.9 Map plots of GVI for Amsterdam and Singapore as an illustration. Source:Treepedia :: MIT Senseable City Lab

Other than directly mapping GVI on streets, the data can also be aggregated at an area-level by mean or median to ease the extraction of meaningful information by stakeholders such as planners and local governments. Nevertheless, there is a big potential to bias if the GVI is simply aggregated by the mean of points per area due to the difference in density of data collection. Kumakoshi et. al. (2020) modified the GVI calculation to propose a Standardized Green View Index (sGVI) as a weighted aggregation of GVI scores in a study area. The formula is shown in Equation (2.6). i represents the point of GVI calculation, while l_i represents the total length of links (streets) that the point i is associated with, and l is the total length of all links in the zone.

$$sGVI = \sum_{i=1}^{n} GVI_i * \frac{l_i}{l} \tag{2.6}$$

Equation (2.6) shows the formula to calculate sGVI.

On top of the mapping of GVI or sGVI values on map, multiple secondary data such as the width of road (Dong et. al., 2018), ethnic distribution (Li et. al., 2015) can also be used to investigate the effects or reasons behind differences in GVI.

1.8 Issues

Overall, there are multiple issues that can be identified from the literature review. To the best of our knowledge, there are no GVI studies in Malaysia, resulting in a missing potential to measure urban greeneries coverage from the users' perspectives. As the computation of GVI was labor intensive (manual selection of greeneries), it is important that we choose an effective method to automate the extraction of greeneries from the street view images. In the literature review, we have come across multiple methods such as Pixel Segmentation (Li et. al. 2015) and deep learning models (Cai et. al., 2018) based on convolutional neural networks (CNN) such as PSNet and ResNet that were used by previous researchers for the task. It is important that we explore these methods and choose the most accurate and interpretable model to compute our GVI. Finally, it is also essential that we can visualise the computed GVI so that they can be interpreted effectively in meaningful ways to assist spatial planning of the cities.