International Journal of Advances in Applied Sciences (IJAAS)
Vol. 15, No. 3, September 2026, pp. 975~989
ISSN: 2252-8814, DOI: 10.11591/ijaas.v15.i3.pp975-989

Recent advances in computer vision segmentation for urban tree detection using street view imagery: a literature review

Muhammad Fareez Rashdan[1], Nurbaity Sabri[2,3], Shafaf Ibrahim[1,3], Nor Masri Sahri[2]
[1]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Malaysia
[2]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Cawangan Melaka Kampus Jasin, Melaka, Malaysia
[3]Center of Vision, Algorithm, and Management Analytics, Universiti Teknologi MARA, Shah Alam, Malaysia

Article Info

Article history:

Received Sep 25, 2025
Revised Jun 5, 2026
Accepted Jun 17, 2026

Keywords:

Computer vision
Deep learning
Street view imagery
Tree segmentation
Urban forestry

ABSTRACT

Segmentation of trees in urban settings is important for tracking vegetation and managing green infrastructure in urban areas. However, there are numerous factors that make it difficult to obtain precise results, such as differences in tree species, season, and environmental variables. Street-view imagery can provide a valuable database to be used for the detection and identification of trees. Unfortunately, the process of labelling images manually is time-consuming and inconsistent. In recent years, deep learning models have provided better automation and precision. However, difficulties in obtaining results persist due to factors such as occlusion, background, and changes in lighting conditions. Therefore, this paper presents an overview of the current trends in the use of deep learning for the detection and segmentation of trees. Specifically, the following aspects will be discussed in this work: preprocessing of images, semantic and instance segmentation, models based on deep learning, and transformers. This literature review also analyzes current datasets, evaluation measures, and applications, offering knowledge about existing challenges and future directions for further studies. The article is designed to be a source of reference for scientists and practitioners working on problems related to urban vegetation classification and computer vision.

Corresponding Author:

Shafaf Ibrahim
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA
Shah Alam, Selangor, Malaysia
Email: shafaf2429@uitm.edu.my

1. INTRODUCTION

Trees in urban areas play a vital role in sustainable urban development, especially in terms of air purification, lowering the heat in urban areas, capturing carbon, and ecological resiliency [1]. With the increasing trend of data-driven urban planning in recent years, there is a need for accurate detection and segmentation of trees in urban settings to help manage green spaces, monitor environmental changes, and implement climate change adaptation strategies. In the smart city concept, the use of vegetation data has become increasingly important in decision-making and urban planning. Specifically, it should be noted that street-level imagery can serve as an alternative source of vegetation data in contrast to aerial and satellite imagery [2].

Accurate segmentation is essential because it enables pixel-level delineation of individual tree structures, facilitating accurate canopy cover estimation, evaluation of the state of trees, species-specific study, and spatial distribution analysis [3]. As compared to simple detection, segmentation provides more specific data that is necessary for urban forestry analysis, decision-making, and its incorporation into

geographic information systems. The increased level of detail makes it possible to use it for advanced purposes like automatic inventory creation, growth tracking, and quantifying ecosystem services, thereby improving the overall effectiveness of urban green infrastructure management in smart city ecosystems.

However, tree segmentation in street view images is particularly challenging. Trees of the same species can be very different from one another depending on the season or species; trees are occluded by buildings and vehicles, and there are variations in the lighting conditions [4]. In contrast to the remote sensing images, the street-level images contain a lot of distractions in the background and distortions due to the perspective, which makes it a lot more difficult to classify the image at the pixel level. Massive datasets from services like Google Street View (GSV) are great to support large-scale analysis of urban vegetation, but the background annotations are manual, and to do them on a large scale is just not practical [5]. The above-mentioned limitations mean that there is a need for more advanced algorithms to perform pixel segmentation in complex and varied urban environments. There have been notable improvements in the accuracy of tree detection due to the recent developments in computer vision, particularly the deep learning-based semantic and instance segmentation models [6]. A significant segmentation performance enhancement, as well as increased generalization, has been observed as the field of research has moved from traditional techniques to convolutional neural networks (CNNs), and more recently to transformer-based architectures and models [7]. The introduction of the segment anything model (SAM) is a significant development in the field, as it has proven exceptional zero-shot generalization and segmentation customizability across a range of different visual tasks. With flexibility to interact with SAM using points, bounding boxes, or masks, the dependence on massive, manually annotated datasets is drastically reduced [8]. SAM is scalable and adaptable, rendering it ideal for segmentation of urban trees within complex street images, especially in smart city data processes, where scalable and adaptable solutions are needed.

In spite of these kinds of developments, few studies have been done on the utilization of lightweight versions of SAM, such as efficient vision transformer (EfficientViT)-SAM. EfficientViT-SAM also possesses architectural advancements that enhance computational effectiveness and speed up inference, which are more appropriate in real-time and big-city applications. Nevertheless, empirical analysis and its application in city street-level tree segmentation have yet to be implemented [9]. This is one of the possible research directions to explore how to balance EfficientViT-SAM to be accurate and efficient in real-world urban settings, particularly in smart city systems with high performance and low computational cost. Modern pipelines typically combine preprocessing, model training, inference, and quantitative evaluation steps and are often optimized to handle large-scale data and real-time inference scenarios [10]. Recently, the trend has been to target more computational efficiency that facilitates its integration with smart city platforms, geographic information systems, and AI-based urban analytics dashboards.

Although vegetation segmentation in remote sensing has been extensively studied [11], [12], little has been done to analyze the techniques used to study street-level imagery. Issues with street-view environments include perspective-based object occlusion of detailed tree boundaries that have not yet been fully addressed in aerial research. Basic models such as SAM and variants have emerged at a rapid pace since 2022-2026. However, reviews of the application of these models to urban tree segmentation with street-level imagery are still lacking. This is a niche area that is not well covered in most existing studies related to segmentation or remote sensing. Therefore, an in-depth analysis is needed to examine trends, performance using the most popular datasets, and the trade-off between accuracy and efficiency in urban applications. A literature review will help in understanding the effectiveness of such segmentation methods on trees in street-level imagery. It will also consider the trade-off between yield and low computational power. The objective is to offer an idea of what works best in urban-scale applications.

This paper presents a literature review of methodologies for automated tree segmentation in street-level imagery data. It incorporates new methods, looks at evaluation metrics and benchmarking practices, and discusses new trends such as transformer-based segmentation, foundational models such as SAM, and the potential of efficient architectures such as EfficientViT-SAM in the context of smart city applications. It also shows the engineering concerns that must be considered to scale such systems to real-world urban systems, and that tree segmentation is a critical aspect of applied computer vision for building smart and sustainable cities. The key highlights of this research are: i) it offers an overview of the advanced methods of segmentation to detect trees and extract them in street-level images and ii) it identifies the current limitations and recommends possible future research opportunities to enhance the techniques of urban tree segmentation. This is how the rest of the paper is structured. Section 2 will describe the process of reviewing existing research and how to classify various approaches to tree segmentation. This will aid in learning the methods employed. Section 3 will compare methods that consider the new research directions and will explain the benefits and drawbacks of each of the methods. It will also consider how these methods work in real-life situations. There will be a conclusion of the paper in section 4, and future directions of research are outlined.

2. METHOD

2.1. Systematic review framework

This literature review presents a straightforward systematic review of the literature on urban tree segmentation in street-level imagery. The process of literature review includes four phases: i) literature search, ii) screening and eligibility evaluation, iii) extraction of data, and iv) comparative analysis. By following this structured approach, the review ensures a transparent, reproducible, and comprehensive synthesis of existing methods and their performance metrics.

2.1.1. Literature search strategy

To find as many relevant studies as possible, an extensive literature review was carried out in large academic databases, including Scopus, Web of Science (WoS), IEEE Xplore, ScienceDirect, and SpringerLink. In doing so, this study focused only on papers published between 2015 and 2025, with an increased focus on papers published between 2023 and 2025. The intention behind this step was to take into account all the recent trends concerning computer vision segmentation. The literature search was conducted with a set of relevant keywords, including urban tree segmentation, street view imagery, semantic segmentation, instance segmentation, deep learning, transformer-based segmentation, urban vegetation mapping, and green view index (GVI). These keywords, combined with the aid of the Boolean operators (AND, OR) served to reduce the search results and ensure that the studies found were very closely related to the scope of the research.

2.1.2. Inclusion and exclusion criteria

Inclusion-exclusion criteria were used to select the appropriate papers to be included in the literature review to ensure their relevance and quality. The following were the inclusion criteria: articles had to be published in a journal or conference, and had to be about the topic of tree or vegetation segmentation in street-level imagery, including that of GSV and Mapillary, and use a deep learning-based segmentation model or some other sophisticated segmentation method. In the exclusion criteria, they did not emphasize street-level images, relying only on aerial or satellite images; published the article on a site, not any experiment, purely theoretical research, and no measures to evaluate the performance of the model. Articles that were pre-printed were not normally provided unless of great significance and interest to the subject.

2.2. Comparative analysis framework

A systematic review of the selected studies was conducted to create a systematic comparison, and a framework was created based on four main dimensions: preprocessing strategy, type of segmentation, model structure, dataset, training strategy, and evaluation measures and performance. It offers a standardized way of comparing and contrasting different approaches, and it is possible to compare different approaches more objectively despite the differences in data sources, model designs, and evaluation methods.

Segmentation is a fundamental task in computer vision, the main goal of which is to divide an image into identifiable and meaningful parts. The pixels are assigned labels depending on common characteristics such as color, texture, or closeness [13]. There are several types of segmentation based on the degree of detail and the aim of isolating certain objects [14]. Semantic segmentation classifies all the pixels in an image into a single category but does not differentiate between different objects [15]. An example is that in a street view image, all of the pixels composing a tree are labeled as tree, but nothing is done to distinguish between trees. This form of segmentation can help comprehend the composition of the scene, such as the extent to which trees occupy an image [16]. Instance segmentation is extended and labels pixels by category and differentiates between instances of the category [17]. This approach is suitable when the objective is to count or track individual objects in a scene [18]. Panoptic segmentation is a hybrid of semantic and instance segmentation. It assigns to every pixel a class label and instance ID, distinguishing between all objects and also labeling the background [19]. In a street image, it would therefore label each tree and pedestrian with a distinct label and label the road, sky, and buildings as a background class. In this manner, you get the whole picture of the entire scene.

2.3. Taxonomy of segmentation approaches

The reviewed studies sorted segmentation methods into four main categories, which are traditional image processing methods, CNN-based semantic segmentation, instance- and region-based segmentation, and transformer-based and foundation models. This classification represents the history of the evolution of segmentation methods with the transition to modern deep learning models. Over the last several years, the number of segmentation methods to be used in tree segmentation tasks has been investigated and published in research journals, including both classic and recent methods, such as deep learning-based approaches. These methods are broadly divided into categories, as shown in Figure 1.

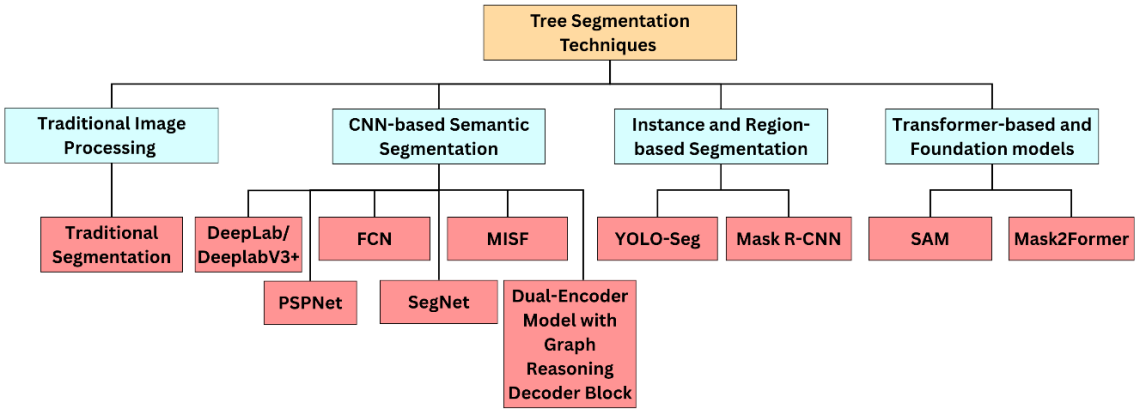

Figure 1. Image segmentation techniques

2.3.1. Traditional segmentation

Classical image processing methodologies, which focus on dividing an image into semantically coherent regions by using low-level visual features like color, intensity, texture, and edges without deep learning models, are referred to as traditional segmentation. Some of the common algorithms in this paradigm include thresholding, clustering, edge detection, region growing, and watershed algorithms [20]. These methods rely on manually written rules or set filters to specify object edges or homogeneous regions that can be very parameter-sensitive to be optimized to fit a specific dataset or imaging environment. Although the traditional methods of segmentation may be reasonable in small or simplified cases, they are not as strong in dealing with complex scenes that have illumination variation, scale, as well as partial obscuration, particularly when contrasted with the contemporary deep learning-based segmentation techniques [21].

2.3.2. Convolutional neural networks-based semantic segmentation

Deep labelling version 3 (DeepLabV3) and DeepLabV3+ are deep learning models that are used in semantic segmentation tasks [22]. DeepLabV3 applies atrous (dilated) convolutions to obtain context information of different scales without loss of spatial resolution. It also features an atrous spatial pyramid pooling (ASPP) module in order to retrieve features in various receptive fields [23]. DeepLabV3+ extends this by adding a decoder module that focuses on the segmentation results, especially at the points of object edges, resulting in more accurate results. The encoder can be built on backbones like residual network (ResNet) or extreme inception (Xception), which can enable the network to learn rich semantic features [24]. This, along with other factors, makes DeepLabV3+ particularly effective in situations of intricate segmentation of scenes with great fine details.

Pyramid scene parsing network (PSPNet) is a semantic segmentation network in deep learning, which is useful in acquiring both local and global context information [25]. The core of the model is the pyramid pooling module (PPM) that gathers the contextual features of various sizes: 1 × 1, 2 × 2, 3 × 3, and 6 × 6 regions, and concatenates the features with the original feature map [26]. This enables the network to make sense of the overall arrangement of the scene and remember small details. PSPNet typically uses a powerful backbone like ResNet to extract features as well as to work on complex segmentation with hierarchical world contexts and pixel-level forecasts [27].

Fully convolutional networks (FCNs) represent a novel approach to semantic segmentation that proposes a semantic architecture of end-to-end pixel-based classification [28]. Unlike conventional CNNs, FCNs lack fully connected layers and use only convolutional layers, with the ability to accept inputs of arbitrary dimensions and generate outputs that are spatially dense [29]. The backbones (visual geometry group (VGG) or ResNet) do feature extraction and are followed by learned upsampling (deconvolution) layers that successively reintroduce spatial resolution. FCNs make use of skip connections to combine fine-grained spatial information obtained in earlier network layers with high-level semantic information obtained in later layers [30] to achieve better localization. This mixture of multi-level information was a change of direction in the history of fully trainable, dense prediction architectures.

Segmentation network (SegNet) is a convolutional encoder-decoder network that was trained to do semantic segmentation [31]. Its encoder is usually founded on a hierarchical classification network, such as VGG, that obtains hierarchical feature representations with the help of convolution and pooling

algorithms [32]. The unique characteristic of SegNet is a decoder design that repurposes the max-pooling indices of the encoder in non-linear upsampling. This approach helps to preserve spatial accuracy in the process of reconstructing the segmentation map, which leads to better boundary delineation and fewer trainable parameters than architectures that use solely learned upsampling. The last layer generates pixel-level predictions of classes that enable SegNet to be useful in tasks that involve accurate segmentation with comparatively low-energy computations [33].

The multi-vision image segmentation and coordinate fusion (MISF) architecture primarily comprises two main stages, which are image segmentation and coordinate fusion for tree size attribute calculation. It begins by capturing images of a standing tree from multiple perspectives using a camera setup that often involves three binocular subsystems. Deep learning-based semantic segmentation methods such as SEMD are then used to accurately delineate the main body of the tree from the background and automatically identify key edge feature points in these segmented images [34]. Finally, the information from these segmented multi-view images is subjected to coordinate fusion, allowing for the reconstruction of the tree's 3D spatial data and the automatic calculation of its height, diameter at breast height (DBH), and crown width.

The dual-encoder model with a graph reasoning decoder block is an image segmentation architecture that utilizes two specialized encoders, which are a basic encoder with VGG blocks and ASPP for multi-scale semantic and contextual features, and an edge encoder with directional convolutions for detailed structural information of elongated objects. The innovative graph reasoning decoder block applied in the final decoding stage addresses discontinuous or blurry edges by first fusing features with attention, then projecting them onto a graph performing relational reasoning with graph convolution networks (GCNs) to infer connections between object parts, and finally reprojecting the refined graph features back into a coherent segmentation map [35].

2.3.3. Instance and region-based segmentation

YOLO-Seg is an extension of you only look once object detector to real-time segmentation [36]. The architecture incorporates the functionality of segmentation into the YOLO system by including a decoder module that generates pixel-level segmentation masks in addition to the standard bounding box outputs [37]. YOLO-Seg ensures high inference rates by relying on the high-speed and efficient feature extraction backbone of YOLO, and can detect objects jointly and fine-granularly [38].

The powerful deep learning architecture, mask region-based convolutional neural network (Mask R-CNN), is a faster region-based convolutional neural network (Faster R-CNN) variation with an additional branch for pixel-wise instance segmentation. It uses a backbone network, which is usually ResNet, to extract features hierarchically and a region proposal network (RPN) to generate region candidates in the form of bounding boxes [39]. The model in both of the proposed regions performs three functions at once: classification of object categories, fine-tuning of bounding box coordinates, and a high-resolution binary mask with the assistance of a small fully convolutional subnet [40]. This auxiliary mask prediction head enables precise object boundaries, and thus Mask R-CNN is most suitable in areas that require accurate detection in addition to fine geometric details about instances.

2.3.4. Transformer-based and foundation models

One of the main components in transformer-based segmentation models is self-attention that facilitates the extraction of global contextual dependencies in images [41]. Such models as vision transformer (ViT) divide images into patches and encode them as sequences of tokens to overcome the local receptive field limitations of CNNs, that enable them to comprehend images as a whole in terms of both semantics and space [42]. The rich and context-aware features are extracted by a transformer encoder and sent to a special decoder to generate pixel-level segmentation masks. A good example of this approach is SAM, a three-part decoupled architecture intended to be a high-performance and real-time segmentation model. Its core is a powerful image encoder that is based on a ViT pre-trained with masked autoencoders (MAE) techniques that encode high-resolution images into a small embedding that can be stored in memory to be reused in the future. This embedding is then paired with a prompt encoder, which projects sparse inputs, like points, bounding boxes, or text, into a common feature space. Lastly, there is a lightweight mask decoder that uses a two-way cross-attention to combine the image and prompt data. To cope with the ambiguity of a single point, the decoder is specially created to provide multiple valid masks at the same time to effectively represent the hierarchy between the sub-parts and the entire object [43].

Masked-attention mask-former (Mask2Former) is a state-of-the-art segmentation framework that combines instance, semantic segmentation, and panoptic segmentation into a single framework [44]. It is based on the transformer-based MaskFormer architecture by having a set-based global loss and masked attention to directly predict a fixed set of segmentation masks and their corresponding class labels. It is based on a powerful backbone, such as ResNet or Swin Transformer, to extract features, and its transformer

decoder part trains on learnable queries to produce mask embeddings [45]. This single architecture makes it easier to train and infer, and to perform well on a variety of segmentation tasks.

Although there is a vast selection of deep learning-based segmentation models that have successfully been used to address tree segmentation tasks, the majority of current research is centered on traditional CNN models, instance segmentation models like Mask R-CNN and YOLO-Seg, or more recently, transformer-based models such as SAM and Mask2Former. Although such methods have proven to be very effective in terms of segmentation accuracy (SA) and generalization, they can be computationally complex and have high processing demands. Variants that are efficient, e.g., EfficientViT-SAM, have proven to be a promising direction as they combine the powerful generalization properties of foundation models with lightweight and computationally efficient designs. Nevertheless, recent research indicates that little has been done regarding the use of EfficientViT-SAM in the distinct task of tree segmentation, especially with street-level images. It lacks the specific research on its effectiveness, trade-offs between robustness and efficiency compared with existing methods. It underscores an important research gap and encourages further investigation of EfficientViT-SAM to perform the urban tree segmentation task, particularly in the context of a situation that must be highly accurate and operate in real-time and within the framework of a smart city.

2.4. Dataset and training analysis

The field of teaching a machine learning or deep learning model to correctly classify every pixel in an image is called data training, where patterns in a set of annotated training data are learned. This is usually a dataset consisting of image and ground truth mask pairs, with each pixel classified by its category, such as road, tree, and building [46]. Popular semantic segmentation training datasets consist of ADE20K, Cityscapes, and Mapillary Vistas, commonly used in activities like computing the GVI in cities [47]. In the meantime, datasets such as Jekyll dataset are specifically created to segment individual trees, thus making it appropriate to perform fine-grained analysis of vegetation [48].

2.4.1. Cityscapes

The Cityscapes dataset is a leading standard in semantic urban scene perception, providing a wide range of 5,000 high-resolution images of streets in 50 European cities. Developed initially by German scientists, it aids in studying urban environments in a precise manner by recording them consistently. Since the image, along with precise labels on 30 types of objects, is provided, the analysis of green spaces becomes extremely accurate. Rather than mere buildings or walking paths, there are plants and land surfaces in the spring, summer, and fall months. This seasonal distribution makes it more useful in forming models to divide visual information into significant sections. Due to clear labeling work, investigators may subdivide large groups, such as vegetation, into smaller groups, such as bushes or forest stands. This detail provides avenues into a greater understanding of the distribution of plants in urban regions [49].

2.4.2. ADE20K

The ADE20K dataset is a rich source of semantic knowledge, with more than 27,000 pictures that cover a variety of indoor and outdoor settings, such as cityscapes. It is highly effective in measuring GVI and also analyzing the quality of the landscapes by offering various landscape subclasses, including trees, grass, and rivers, with pixel-level annotations of 150 different semantic categories, which makes it highly effective [50]. The dataset is also the foundation of the MIT scene parsing benchmark (SceneParse150) that offers an established platform to train and test complex scene parsing algorithms. Because of the presence of object- and part-level segmentation, models that have been trained using ADE20K can be post-processed to isolate certain green objects or other elements of the environment, allowing detailed and advanced landscape architecture analysis in a variety of settings.

2.4.3. Mapillary Vistas

The Mapillary Vistas dataset is a massive and varied dataset of 25,000 high-resolution street-level shots that is intended to enhance knowledge about urban environments, self-driving, and smart city design. The dataset is unique in its worldwide view and represents imagery of six continents taken in diverse seasonal, weather, and lighting conditions. It has a high-density and fine-grained annotation of 66 object classes with specific polygons, and instance-specific labelling of 37 classes [51]. The detail it provides renders it a crucial work towards the creation of strong computer vision models that can maneuver and analyze various complicated real-world scenarios.

2.4.4. Jekyll dataset

The Jekyll dataset is a large collection of urban tree data, which is a collection of over 3,000 high resolution RGB images of trees taken using mobile devices in ten cities in China. The dataset includes

images taken in each of the four seasons, spring, summer, autumn, and winter, in order to capture seasonal variation, and is systematically divided into separate subsets to support a variety of applications in computer vision. These subdivisions are a stem genera subset that composed 7,675 images annotated with 28 species, a tree segmentation subset that consists of 3,949 images annotated with pixel-level binary masks, and a stem segmentation subset where the delineation of individual stems is done using LabelMe software [52]. The dataset can be used to develop domain-specific models that can identify trees with high precision and map urban forests with high fidelity due to its genus-level taxonomic labeling and polygon-based annotation.

A comparative study of commonly used datasets in tree segmentation is provided in Table 1. Unlike general urban scene datasets, i.e., Cityscapes, ADE20K, and Mapillary Vistas, the Jekyll dataset is better suited to single tree segmentation due to its specialized creation and high-resolution annotations. Though the aforementioned datasets provide pixel-level semantic labels that include general categories like vegetation, terrain, and landscape features, the categories are aggregations of plant life, thus failing to distinguish between individual trees and other vegetative features. They therefore suit more applications like GVI computation, which focuses on measuring total visible green cover, as opposed to determining individual tree occurrences.

In comparison, the Jekyll dataset is directly adapted to urban tree analysis and offers granular binary masks and polygon-based annotations that can precisely describe discrete parts of trees, such as both canopy and trunk structures. Its species-level data and multi-seasonal imagery also increase its potential to perform fine-grained analysis, e.g., tree-level segmentation, structural evaluation, and species identification. Consequently, the Jekyll dataset is a better starting point to formulate and test models that can be used to segment individual trees with high precision. The other datasets, on the other hand, are more suited to general vegetation mapping and GVI-based activities.

Table 1. Comparison of tree segmentation datasets

Dataset	No. of images	Scene type	Annotation type	No. of classes	Tree-specific detail	Suitable application	Key strengths
Cityscapes	5,000	Urban street (Europe)	Pixel-level (fine and coarse)	30	Low (vegetation grouped)	GVI, urban segmentation	High-quality annotations, consistent conditions, benchmark dataset
ADE20K	27,000	Indoor and outdoor scenes	Pixel-level + part-level	150	Moderate (includes tree class)	GVI, scene parsing, landscape analysis	Large diversity, rich semantic categories, detailed annotations
Mapillary Vistas	25,000	Global street-level	Polygon-based + instance labels	66 (37 instance-level)	Moderate	GVI, urban analytics, autonomous driving	Global diversity, varied conditions, fine-grained annotations
Jekyll dataset	3, 949	Urban tree-focused	Pixel-wise masks + polygon (LabelMe)	28 Species	High (individual trees, trunks)	Individual tree segmentation, species analysis	Tree-specific design, multi-seasonal data

2.5. Evaluation metrics analysis

The evaluation stage refers to the process of measuring the accuracy of a segmentation model by comparing the segmentation masks it predicts to the ground truth labels [53]. This comparison enables the measurement of how accurately the model classifies each pixel into its respective class, thereby providing an objective assessment of segmentation quality. This is done with quantitative measures of the quality of pixel-wise classification.

2.5.1. Intersection over union

The intersection over union (IoU) is a performance measure that is widely utilized in image segmentation and object detection. It measures how well data is predicted by the ratio of the area overlap between the predicted segmentation mask and ground truth mask [54], expressed as (1), where TP is the number of pixels correctly predicted to be part of the object, FP is the number of pixels incorrectly predicted to be part of the object but background, and FN is the number of pixels that are correctly part of the object but are predicted as background.

$$IoU = \frac{TP}{(TP+FP+FN)} \quad (1)$$

2.5.2. Mean absolute error

Mean absolute error (MAE) is an analysis measure that is used to calculate the mean of absolute differences between the estimated values and the real ground truth ones [55]. In image segmentation, MAE is often used in regression-based segmentation tasks or when evaluating continuous-valued outputs, such as saliency maps, depth maps, or vegetation indices like the GVI [56], as calculated in (2), where y_i is the ground truth value for a pixel, $\hat{y}_i$ refers to the predicted value for a pixel, and n is the total number of pixels.

$$MAE = \frac{1}{N}\sum_{n=1}^{n} |y_i - \hat{y}_i| \quad (2)$$

2.5.3. Segmentation accuracy, over-segmentation rate, and under-segmentation rate

SA quantifies how well the algorithm correctly identifies and delineates objects compared to a perfect ground-truth segmentation as calculated in (3). Meanwhile, the over-segmentation rate (OR) indicates when a single, true object is incorrectly split into multiple smaller segments, and the under-segmentation rate (UR) refers to instances where several distinct objects are wrongly merged into a single segment [57], as calculated in (4) and (5), respectively.

$$\text{SA} = 1 - \frac{|R_S - T_S|}{R_S} \times 100\% \quad (3)$$

$$OR = \frac{O_S}{R_S+O_S} \quad (4)$$

$$UR = \frac{U_S}{R_S+O_S} \quad (5)$$

Where R_s is the ground truth, T_s is the area segmented, $|R_s - T_s|$ is the number of pixels that were incorrectly segmented, O_s is the number of pixels that were wrongly included in the segmentation result but should not have been, and U_s is the number of pixels that were incorrectly excluded from the segmentation result but should have been included.

2.5.4. Green view index

GVI is an environmental metric that measures the visible proportion of greener vegetation, especially trees, shrubs, and grasses, in street-level imagery [58]. Also used extensively in urban studies, environmental analysis, and city planning, GVI assesses the perceptible vegetation in urban settings from a pedestrian perspective, as established by (6), where the green pixels are the image segmentation of parts of the image that are plants, such as trees or grass. Whereas, the total pixels refer to the number of pixels in the street-level image, excluding sky and distant background, given the approach taken.

$$GVI = \frac{Number\ of\ Green\ Pixels}{Total\ Pixels} \times 100\% \quad (6)$$

Based on the evaluation metrics presented, the IoU proves to be more effective in individual tree segmentation, where it measures the extent of spatial agreement between the predicted and reference masks by comparing pixels between the estimated and reference masks. Whereas continuity or proportional vegetation data is better captured by MAE and GVI, the sensitivity needed in alignment of boundaries to accommodate complex urban canopy shapes and occlusions is offered by the IoU. Although measures, including SA, OR, and UR, provide clues on the errors in particular categories, including fragmentation or merging, they do not provide the balance of a holistic IoU that simultaneously takes into consideration missed areas and false positives to guarantee proper object-level localization.

3. RESULTS AND DISCUSSION

Recent studies have demonstrated substantial progress in tree segmentation techniques that leverage street-view imagery. The majority of this research has concentrated on estimating the GVI as a means of quantifying urban greenery from a pedestrian-level perspective. In contrast, relatively fewer studies have specifically addressed the more challenging task of segmenting individual trees within such imagery.

Table 2 provides a concise summary of the key methodological developments and representative works in both research directions.

Table 2. Overview of segmentation techniques for urban tree detection from street-level imagery

Author	Model/method	Dataset	Performance	Limitation
Aikoh *et al.* [59]	DeepLab (vs. manual Photoshop)	Cityscapes and ImageNet	High correlation with manual GVI (r = 0.97, 0.95); fast processing (~51s vs 30 min); and strong agreement with GSV (r = 0.966)	Sensitive to viewpoint differences, seasonal variation, and image resolution
Xia *et al.* [60]	DeepLabV3+	Cityscapes	Higher mIoU (78.37%), lower RMSE (2.75%), and faster (0.29 s) than PSPNet	Affected by seasonal variation, sampling distance, and lack of time-series data
Ki and Lee [61]	FCN8s (TensorFlow)	Cityscapes	Accuracy = 0.8456; effective for large-scale GVI; captures 3D greenery; and links to walking behavior	Segmentation errors, temporal inconsistency in GSV, and missing environmental variables
Sánchez and Labib [62]	Mask2Former (Transformer-based)	COCO, Cityscapes, ADE20K, and Mapillary Vistas	mIoU: 57.4% (small), 59.0% (medium/large) and scalable global analysis	Image availability, quality variability, seasonal/climate differences, and NDVI mismatch
Yang *et al.* [63]	YOLO-SegNet (YOLOv8 + SegFormer)	Dataset Jekyll	mIoU = 92.0%; pixel accuracy = 95.9%; and strong across species	Average metrics may not reflect species-specific performance
Lumnitz *et al.* [64]	Mask R-CNN + depth + triangulation	COCO and COCO Stuff	>70% detection and location error = 4–6 m	Errors in occlusion, overlapping trees, small/shadowed objects; and limited training diversity
Mo *et al.* [34]	MISF (SEMD + multi-vision + SURF)	Private dataset	Errors = 1.89% (height), 2.42% (DBH), and 3.15% (crown width)	Sensitive to occlusion and camera calibration errors
Zhou *et al.* [35]	Dual-encoder + graph reasoning decoder	Dataset Jekyll	+4.3% mIoU and +5.7% F-score improvement	Reduced performance for distant tree branches
Yang *et al.* [65]	Adaptive mean shifting + image abstraction	None	>92% segmentation accuracy	Performance drops for distant trees and complex environments
Tilki *et al.* [66]	Grounded SAM (Grounding DINO + SAM)	SA-1B	IoU ≈ 0.50 for trees; multi-task detection + segmentation	Difficulty in complex urban vegetation segmentation

3.1. Studies on green view index estimation

A number of studies have used deep learning-based semantic segmentation to predict the GVI using street-level images. Aikoh *et al.* [59] compared the calculation of GVI manually in Adobe Photoshop with an automated method based on DeepLab and found extremely high correlations, which gives reason to believe that deep learning can give an approximation of manual results. Other models like PSPNet, SegNet, and FCN-8s have been reported to have similarly high correlations. Despite the fact that some cases of minor misclassification were observed, like that of reflections, shadows, or man-made objects, the effect was minimal. The automated method proved to be far more efficient and was in good agreement with GSV data. However, other factors could affect accuracy, like divergent opinions, season, and image resolution; therefore, more validation is needed in comparison to human perceptual standards.

Xia *et al.* [60] showed how DeepLabV3+ is effective at semantic segregation of street-level imagery to evaluate urban greenery and calculate the GVI. Testing on the Cityscapes dataset showed that DeepLabV3+ outperforms PSPNet. Based on these results, the current research proposes a pedestrian visible green view index (PVGVI) framework, which uses panoramic street view and high-precision semantic segmentation to better represent urban greenery as seen from a pedestrian viewpoint, thus addressing the overestimation that comes with traditional GVI frameworks. The framework allows scalable and automated analysis through a connection to a street view API that allows cross-city comparison and allows higher-order applications. Although these benefits are present, several shortcomings still exist, such as less reliability of segmentation in different seasonal conditions, reliance on the density of sampling intervals, and limited access to longitudinal street view images.

The greenery at the street level has proven to have an influence on the pattern of human behavior. The FCN-8s model is used in the framework of the TensorFlow system by Ki and Lee [61], which uses semantic segmentation of GSV to extract vegetative components and calculate the GVI. The Cityscapes dataset was used to train the model. The findings indicate that GVI is a more polished and perceptionally-driven city vegetation index than traditional ones. The study also reveals that the association between green spaces and walking is mediated by the green space assessment method, which, according to GVI, is more strongly associated with utilitarian walking as opposed to closeness to parks. It should be mentioned that socioeconomic disparities are reflected in the fact that the lower-income groups are more susceptible to GVI. These results emphasize the necessity to implement street-level green infrastructure, particularly in disadvantaged areas, and to take into account methodological constraints such as potential misclassification of pictures by segmentation, time delays of GSV data, and the lack of crucial environmental variables.

Transformer-based models have also been enhanced as well, and semantic segmentation is more precise regarding the analysis of urban green space. Sánchez and Labib [62] applied Mask2Former to project Mapillary street-level imagery to map GVI at the global level, which showed excellent results in heterogeneous cities. Riding on this advancement, the current study uses open-access Mapillary data to correspond to FAIR principles that ensure findability, accessibility, interoperability, and reusability, and foster reproducible research, suggesting a geographically flexible and scalable framework of GVI computation. Unlike traditional top-down remote sensing models, this model models vertically oriented, human-perceived vegetation and combines GVI with the normalized difference vegetation index (NDVI) to bridge the data gaps. The transparency of the methodology is given priority, which is especially important in the case of underrepresented urban settings. However, constraints exist, such as uneven spatial coverage of imagery, image quality variability as a result of crowd-sourced collection, and temporal variability through seasonal and climatic variations. Moreover, although NDVI integration aids in data imputation, it can create inaccuracies since it has an aerial vantage point and does not match ground-level measurements, which highlights the necessity of further methodological development and empirical verification.

3.2. Studies on individual tree segmentation

Although there are many studies that have focused on approximating vegetation cover, there is a dearth of research studies that have focused their attention on the recognition of individual trees using street-level images. Yang *et al.* [63] introduced a hybrid architecture, YOLO-SegNet, based on the use of YOLOv8, which detects objects, and segmentation transformer (SegFormer), which segments semantically, with a promising mIoU and mean pixel accuracy, which is a significant example of strong performance in the isolation of individual trees in an urban environment. Though such aggregate measures are indicative of good overall segmentation potential, they can mask differences in model effectiveness in tree species. This research paper builds on this analysis by providing a species-based experiment on validation and test set results on segmentation, showing that YOLO-SegNet can achieve higher IoU scores on a variety of taxa of street trees than other models.

Lumnitz *et al.* [64] apply the Mask R-CNN to detect and segment single trees in GSV and Mapillary images. The investigation of 296 failure cases reveals that most misclassifications occur in the presence of dense trees, overlapping trees, or those that are in shadow or are in leaf-off phases, mainly because of the lack of coverage in the training data. These results highlight the difficulties that exist in separating amorphous, visually unstructured objects. Nevertheless, this can be overcome through combining various street-level viewpoints by providing complementary views on obscured buildings. Also, false positives, especially hedges, and false negatives in underrepresented categories specifically affect predictions with small segmentation masks, showing the need to use heterogeneous training sets and multi-view images to improve the detection performance.

The MISF framework proposed by Mo *et al.* [34] is focused on the estimation of structural parameters of standing trees based on the principle of multi-view image segmentation and integration of spatial coordinates. MISF uses the SEMD deep learning model of multi-scale feature fusion to increase edge preservation and better feature point localization. The three-view camera measurements of tree height, DBH, and crown width give low relative errors. The important developments of MISF are improved accuracy of measurements, multi-view image processing with a better speeded-up robust features (SURF) matching algorithm, and they have proven effectiveness in real-life contexts. To achieve a more continuous branching structure in segmented outputs, Zhou *et al.* [35] introduced a dual-encoder model with a graph reasoning decoder block. The result of this approach is an increase in mIoU and F-score over previous models, but the performance decreases in distal branches at further distances from the imaging source.

At the same time, classical methods of segmentation are still under study. Yang *et al.* [65] introduced a technique combining adaptive mean shifting with image abstraction to separate trees in the

complex background with a high SA. This shows that the traditional image processing schemes may be used even when the conditions are under control. However, the performance is poorer when target trees are far-off, or placed in very complex environments. On the same note, Tilki *et al*. [66] adopted the grounded SAM, which combines grounding DINO and the SAM to allow joint object detection and segmentation in street-level images. The approach yields variable scores in the IoU of object categories, indicating the continuing challenges in the accurate definition of vegetative objects in crowded urban environments.

3.3. Discussion

Based on the presented studies, it is possible to distinguish a clear trend concerning the evolution of methods for assessing urban greenery, namely, the transition from manual to automatic procedures using a deep learning approach. When comparing the accuracy of manual techniques, represented by the example of Adobe Photoshop, with machine learning models such as DeepLab, the obtained correlations are very high. It allows stating that not only does semantic segmentation prove its efficiency compared to human perception, but it also significantly increases the speed of the process. In this way, the possibility of making GVI assessments at a large scale becomes possible, which would be impossible under purely manual conditions. The effectiveness of different models, including PSPNet, SegNet, and FCN-8s, highlights the importance of model design in terms of providing a good balance.

The trend that becomes very clear in the literature is the increasing focus on human-centric indices and the utilization of big data. GSV and Mapillary provide efficient ways to estimate GVI on a city or worldwide level using automated pipeline approaches. At the same time, more advanced indices, such as PVGVI, signify the transition to assessment of vegetation from a pedestrian perspective instead of focusing solely on coverage-based indices. Yet, regardless of all advances in the sphere, the issue of data variations, including changes in viewpoint, seasonality, or low-quality images, still influences the quality of segmentation outcomes.

The development of transformers and foundation models such as Mask2Former and Grounded SAM has introduced an extra layer into the potential of segmentation in cities. While there is evidence of better performance across different datasets with regard to these models, the performance of these models in the case of vegetation classes highlights the difficulty of segmenting occluded and overlapping objects in the scene. Moreover, studies on tree segmentation using individuals indicate a transition from metrics that focus on the area towards object-based assessments. Hybrid methods, as represented by YOLO-SegNet, which is a combination of YOLOv8 and SegFormer, have shown impressive results, highlighting the success of merging detection and segmentation algorithms. Furthermore, instance segmentation approaches, including Mask R-CNN, are capable of detecting tree objects with high precision when combined with depth estimation. However, some difficulties still exist when processing densely packed or hidden areas, where the model may merge multiple trees into one object or miss small tree objects.

Recent approaches that combine segmentation techniques with geometry analysis methods like MISF have shown high promise for determining various tree parameters, including height, diameter, and crown width. Yet, there are certain issues related to their effectiveness that need further investigation. First, their ability to detect tree parameters is highly dependent on factors such as varying conditions at the street level, such as the quality of images, seasons, and lighting, as well as a lack of appropriate datasets. Second, the dependence on images acquired from only one sensor can be an issue for future research. Therefore, future research must focus on creating more robust models resistant to variations in conditions, utilizing multimodal information, namely combining street-level imagery with aerial/satellite imagery, and developing instance-level segmentation algorithms.

4. CONCLUSION

This literature review is a synthesis of the recent developments in the field of automated tree segmentation in street-level imagery to detect salient trends in the methodological procedures, datasets used, and architectural innovation. From the reviews, general-purpose benchmarks like Cityscapes and ADE20K are often utilized to perform semantic segmentation tasks, but domain-specific datasets like Jekyll are more often used in research focused on delineating trees on a case-by-case basis. According to the modeling perspective, deep learning architectures are the most prevalent, with DeepLabV3+ being the most popular, and then PSPNet and FCN. Traditional, non-learning-based segmentation approaches have found little use in the modern context of research because they are less effective in addressing the challenges of urban visual scenes like variability in illumination, scale differences, and partial occlusions. One such methodological improvement highlighted in this study is the combined use of EfficientViT-SAM, which alleviates the computational overhead that is common with foundation models. This methodology has zero-shot generalization properties akin to Efficient-SAM, but with significantly reduced hardware requirements,

allowing high- throughput processing to be adopted in real-world urban monitoring systems. The main contribution of this work is the development of a scalable computation model that balances high-fidelity segmentation with the functional limits of smart city infrastructure. The results reveal that transformer-based models can be optimized systematically to attain a desirable trade-off between the accuracy of segmentation and the cost-effectiveness of computation, which provides a feasible resolution to large-scale monitoring of urban forests. This ability enables a proper delineation of individual tree canopies in even highly populated urban environments to aid in evidence-based decision-making in urban planning and the strategic improvement of metropolitan green infrastructure.

ACKNOWLEDGMENTS

This research was supported by Universiti Teknologi MARA.

FUNDING INFORMATION

The authors state that there is no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Muhammad Fareez Rashdan	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓			
Nurbaity Sabri	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Shafaf Ibrahim	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Nor Masri Sahri	✓	✓	✓	✓		✓	✓	✓		✓	✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

This study does not involve human subjects or personal information requiring informed consent.

ETHICAL APPROVAL

This study does not involve human subjects or animal experiments.

DATA AVAILABILITY

The authors confirm that the data supporting the findings of this study are available within the article [and/or its supplementary materials].

REFERENCES

[1] T. Hu and C. Miao, "Urban forests as an approach to mitigate urban heat island effects: mechanisms, methods, and planning strategies," *International Journal of Environmental Science and Technology*, vol. 23, no. 3, 2026, doi: 10.1007/s13762-025-06982-5.

[2] D. Liu, Y. Jiang, R. Wang, and Y. Lu, "Establishing a citywide street tree inventory with street view images and computer vision techniques," *Computers, Environment and Urban Systems*, vol. 100, Mar. 2023, doi: 10.1016/j.compenvurbsys.2022.101924.

[3] C. Wu, S. Tang, and W. Wang, "Boundary-constrained supervoxel clustering for tree segmentation in broadleaf forests," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 179–186, Nov. 2025, doi: 10.5194/isprs-annals-X-1-W2-2025-179-2025.
[4] A. Hu, N. Yabuki, and T. Fukuda, "Multi-temporal analysis of urban vegetation using deep learning and 3D reconstruction," *Landscape Ecology*, vol. 40, no. 7, Jul. 2025, doi: 10.1007/s10980-025-02090-4.
[5] X. Huo, H. Wang, S. Chen, J. Ye, and W. Li, "Urban street tree recognition based on deep learning and multi-source data," in *Proc. 2024 Int. Symp. Space Optical Instruments and Applications (ISSOIA 2024)*, 2026, pp. 153–159, doi: 10.1007/978-981-95-1410-6_14.
[6] K. Park, D. Ki, and S. Lee, "Toward automated and comprehensive walkability audits with street view images: leveraging virtual reality for enhanced semantic segmentation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 223, pp. 78–90, May 2025, doi: 10.1016/j.isprsjprs.2025.02.015.
[7] X. Liu *et al.*, "Toward the unification of generative and discriminative visual foundation model: a survey," *The Visual Computer*, vol. 41, no. 5, pp. 3371–3412, Mar. 2025, doi: 10.1007/s00371-024-03608-8.
[8] Y. Wang, Y. Zhao, and L. Petzold, "An empirical study on the robustness of the segment anything model (SAM)," *Pattern Recognition*, vol. 155, Nov. 2024, doi: 10.1016/j.patcog.2024.110685.
[9] M. Ali *et al.*, "A review of the segment anything model (SAM) for medical image analysis: accomplishments and perspectives," *Computerized Medical Imaging and Graphics*, vol. 119, 2025, doi: 10.1016/j.compmedimag.2024.102473.
[10] S. Wang, R. Zhu, Y. Pu, M. S. Wong, Y. Xu, and Z. Qin, "Modelling sunlight and shading distribution on 3D trees and buildings: deep learning augmented geospatial data construction from street view images," *Building and Environment*, vol. 275, May 2025, doi: 10.1016/j.buildenv.2025.112816.
[11] J. Zheng, S. Yuan, W. Li, H. Fu, L. Yu, and J. Huang, "A review of individual tree crown detection and delineation from optical remote sensing images: current progress and future," *IEEE Geoscience and Remote Sensing Magazine*, vol. 13, no. 1, pp. 209–236, Mar. 2025, doi: 10.1109/MGRS.2024.3479871.
[12] L. Huang, B. Jiang, S. Lv, Y. Liu, and Y. Fu, "Deep-learning-based semantic segmentation of remote sensing images: a survey," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 8370–8396, 2024, doi: 10.1109/JSTARS.2023.3335891.
[13] D. S. Mohan and K. P. Chandar, "A comprehensive review on various image segmentation methods with applications," in *2025 3rd International Conference on Sustainable Computing and Data Communication Systems* Aug. 2025, pp. 1618–1623, doi: 10.1109/ICSCDS65426.2025.11167803.
[14] Y. Chuang, S. Zhang, and X. Zhao, "Deep learning-based panoptic segmentation: recent advances and perspectives," *IET Image Processing*, vol. 17, no. 10, pp. 2807–2828, Aug. 2023, doi: 10.1049/ipr2.12853.
[15] J. Nie, Z. Wang, X. Liang, C. Yang, C. Zheng, and Z. Wei, "Semantic category balance-aware involved anti-interference network for remote sensing semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023, doi: 10.1109/TGRS.2023.3325327.
[16] H. Ni, Q. Liu, H. Guan, H. Tang, and J. Chanussot, "Category-level assignment for cross-domain semantic segmentation in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023, doi: 10.1109/TGRS.2023.3271776.
[17] Y. Wang, W. Zhou, Q. Lv, and G. Yao, "MetricMask: single category instance segmentation by metric learning," *Neurocomputing*, vol. 500, pp. 896–908, Aug. 2022, doi: 10.1016/j.neucom.2022.05.117.
[18] H.-T. Nguyen, Y. Cao, C.-W. Ngo, and W.-K. Chan, "FoodMask: real-time food instance counting, segmentation and recognition," *Pattern Recognition*, vol. 146, Feb. 2024, doi: 10.1016/j.patcog.2023.110017.
[19] O. L. F. de Carvalho, O. A. de C. Júnior, A. O. de Albuquerque, N. C. Santana, and D. L. Borges, "Rethinking panoptic segmentation in remote sensing: a hybrid approach using semantic segmentation and non-learning methods," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2022.3172207.
[20] S. T. Bhairnallykar and V. Narawade, "Exploration of image segmentation: a comprehensive review of traditional segmentation techniques," in *2023 7th International Conference on Electronics, Materials Engineering & Nano-Technology*, Dec. 2023, pp. 1–6, doi: 10.1109/IEMENTech60402.2023.10423439.
[21] J.-H. Son, H. Song, C. Song, M. Ha, D. Kang, and Y.-S. Ha, "Condition-based synthetic dataset for amodal segmentation of occluded cucumbers in agricultural images," *Computers and Electronics in Agriculture*, vol. 238, Nov. 2025, doi: 10.1016/j.compag.2025.110800.
[22] H. Chen, Y. Qin, X. Liu, H. Wang, and J. Zhao, "An improved DeepLabV3 lightweight network for remote-sensing image semantic segmentation," *Complex and Intelligent Systems*, vol. 10, no. 2, pp. 2839–2849, Apr. 2024, doi: 10.1007/s40747-023-01304-z.
[23] G. Sunandini, R. Sivanpillai, V. Sowmya, and V. V Sajith Variyar, "Significance of atrous spatial pyramid pooling (ASPP) in DeepLabV3 for water body segmentation," in *2023 10th International Conference on Signal Processing and Integrated Networks*, Mar. 2023, pp. 744–749, doi: 10.1109/SPIN57001.2023.10116882.
[24] A. A. Sundaresan and A. A. Solomon, "Post-disaster flooded region segmentation using DeepLabV3 and unmanned aerial system imagery," *Natural Hazards Research*, vol. 5, no. 2, pp. 363–371, Jun. 2025, doi: 10.1016/j.nhres.2024.12.003.
[25] H. Kim, Y. J. Kim, J. Hong, H. Yang, S. Euna, and K. Lee, "Local context aggregation for semantic segmentation: a novel PSPNet approach," in *2024 International Technical Conference on Circuits/Systems, Computers, and Communications*, Jul. 2024, pp. 1–5, doi: 10.1109/ITC-CSCC62988.2024.10628341.
[26] Z. Wanzhen, W. Chan, and Z. Xiaoming, "Remote sensing image semantic segmentation algorithm based on improved PSPNet," in *2025 4th International Conference on Image Processing and Media Computing*, Jun. 2025, pp. 15–21, doi: 10.1109/ICIPMC66319.2025.11170680.
[27] N. Sharma, S. Gupta, A. Rajab, M. A. Elmagzoub, K. Rajab, and A. Shaikh, "Semantic segmentation of gastrointestinal tract in MRI scans using PSPNet model with ResNet34 feature encoding network," *IEEE Access*, vol. 11, pp. 132532–132543, 2023, doi: 10.1109/ACCESS.2023.3336862.
[28] N. Ali, A. Z. Ijaz, R. H. Ali, Z. Ul Abideen, and A. Bais, "Scene parsing using fully convolutional network for semantic segmentation," in *2023 IEEE Canadian Conference on Electrical and Computer Engineering*, Sep. 2023, pp. 180–185, doi: 10.1109/CCECE58730.2023.10288934.
[29] Nancy and D. Kumar, "Semantic segmentation of road scenes using fully convolutional networks," in *2025 IEEE 7th International Conference on Computing, Communication and Automation*, Nov. 2025, pp. 1–5, doi: 10.1109/ICCCA66364.2025.11325218.
[30] C. Shen *et al.*, "A cascaded fully convolutional network framework for dilated pancreatic duct segmentation," *International Journal of Computer Assisted Radiology and Surgery*, vol. 17, no. 2, pp. 343–354, Feb. 2022, doi: 10.1007/s11548-021-02530-x.
[31] F. He, W. Wang, L. Ren, Y. Zhao, Z. Liu, and Y. Zhu, "CA-SegNet: a channel-attention encoder–decoder network for histopathological image segmentation," *Biomedical Signal Processing and Control*, vol. 96, Oct. 2024, doi: 10.1016/j.bspc.2024.106590.

[32] H. Zhu *et al.*, "Deep convolutional encoder–decoder networks based on ensemble learning for semantic segmentation of high-resolution aerial imagery," *CCF Transactions on High Performance Computing*, vol. 6, no. 4, pp. 408–424, Aug. 2024, doi: 10.1007/s42514-024-00184-0.
[33] F. M. A. Mazen, Y. O. Shaker, and R. A. A. Seoud, "Attention-based SegNet: toward refined semantic segmentation of PV modules defects," *IEEE Access*, vol. 12, pp. 100792–100804, 2024, doi: 10.1109/ACCESS.2024.3431098.
[34] L. Mo, L. Shi, G. Wang, X. Yi, P. Wu, and X. Wu, "MISF: a method for measurement of standing tree size via multi-vision image segmentation and coordinate fusion," *Forests*, vol. 14, no. 5, May 2023, doi: 10.3390/f14051054.
[35] Y. Zhou, H. Wang, Y. Liu, and Y. Wu, "Enhanced tree branch segmentation in urban environments using a dual-encoder model with graph reasoning decoder block," in *2024 IEEE International Conference on Systems, Man, and Cybernetics*, Oct. 2024, pp. 4054–4060, doi: 10.1109/SMC54092.2024.10831953.
[36] Q. Wang, D. Chen, W. Feng, L. Sun, and G. Yu, "Enhancing instance segmentation in agriculture: an optimized YOLOV8 solution," *Sensors*, vol. 25, no. 17, Sep. 2025, doi: 10.3390/s25175506.
[37] H. Abudukelimu *et al.*, "DVF-YOLO-Seg: a two-stage breast mass segmentation model with enhanced feature extraction and small lesion detection," *Digital Health*, vol. 11, May 2025, doi: 10.1177/20552076251374192.
[38] Y. Jiang, M. Jiang, T. Abbas, H. Ge, X. Wen, and H. Chen, "Detection of wheat impurities using terahertz imaging technology and deep learning," *LWT*, vol. 229, Aug. 2025, doi: 10.1016/j.lwt.2025.118201.
[39] X. Xu *et al.*, "Crack detection and comparison study based on Faster R-CNN and Mask R-CNN," *Sensors*, vol. 22, no. 3, Feb. 2022, doi: 10.3390/s22031215.
[40] G. Kaur, M. Garg, and S. Gupta, "Integrated model for segmentation of glomeruli in kidney images," *Cognitive Robotics*, vol. 5, pp. 1–13, 2025, doi: 10.1016/j.cogr.2024.11.007.
[41] J. Wei and J. Tan, "TCNet: a CNN-transformer hybrid framework based on multi-scale self-attention for accurate farmland segmentation," *Smart Agricultural Technology*, vol. 14, Aug. 2026, doi: 10.1016/j.atech.2026.102041.
[42] X. Li *et al.*, "Transformer-based visual segmentation: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10138–10163, Dec. 2024, doi: 10.1109/TPAMI.2024.3434373.
[43] A. Kirillov, E. Mintun, N. Ravi, H. Mao, "Segment anything," in *2023 IEEE/CVF International Conference on Computer Vision*, Oct. 2023, pp. 3992–4003, doi: 10.1109/ICCV51070.2023.00371.
[44] C. A. Patel, D. S. Pragna, K. M. N. Babu, M. A. Kumar, K. Swaraja, and N. A. Vignesh, "Panoptic segmentation using Mask2Former with Swin Transformers," in *2023 4th International Conference on Intelligent Technologies*, Jun. 2024, pp. 1–6, doi: 10.1109/CONIT61985.2024.10626934.
[45] J.-C. Sheng, Y.-S. Liao, and C.-R. Huang, "Apply masked-attention mask transformer to instance segmentation in pathology images," in *2023 Sixth International Symposium on Computer, Consumer and Control*, Jun. 2023, pp. 342–345, doi: 10.1109/IS3C57901.2023.00098.
[46] A. R. Singh, S. D. R, and A. V. D, "Early detection of diabetic foot ulcer with multi-modal images using federated learning," in *2025 International Conference on Multi-Agent Systems for Collaborative Intelligence*, Jan. 2025, pp. 964–969, doi: 10.1109/ICMSCI62561.2025.10894410.
[47] J. Xie, M. Li, J. Wu, X. Zhang, and J. Zhang, "Semantic segmentation of building façade materials and colors for urban conservation," *npj Heritage Science*, vol. 13, no. 1, Jul. 2025, doi: 10.1038/s40494-025-01888-4.
[48] T. A.-Ramirez, A. Alfaro, J. M. Saavedra, M. Recabarren, M. P.-Donoso, and J. Delpiano, "Exploring the potential of reconstructed multispectral images for urban tree segmentation in street view images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 12112–12122, 2024, doi: 10.1109/JSTARS.2024.3419127.
[49] L. Kumar and D. K. Singh, "Comparative analysis of Vid2Vid and fast Vid2Vid models for video-to-video synthesis on the Cityscapes dataset," in *2023 International Conference on Computer, Electronics & Electrical Engineering & their Applications*, Jun. 2023, pp. 1–5, doi: 10.1109/IC2E357697.2023.10262586.
[50] D. H. Lee, H. Y. Park, and J. Lee, "A review on recent deep learning-based semantic segmentation for urban greenness measurement," *Sensors*, vol. 24, no. 7, Mar. 2024, doi: 10.3390/s24072245.
[51] R. Ankareddy and R. Delhibabu, "Dense segmentation techniques using deep learning for urban scene parsing: a review," *IEEE Access*, vol. 13, pp. 34496–34517, 2025, doi: 10.1109/ACCESS.2025.3543944.
[52] T. A.-Ramirez *et al.*, "Challenges for computer vision as a tool for screening urban trees through street-view images," *Urban Forestry & Urban Greening*, vol. 95, May 2024, doi: 10.1016/j.ufug.2024.128316.
[53] A. Hassan, R. Mumtaz, Z. Mahmood, M. Fayyaz, and M. K. Naeem, "Wheat leaf localization and segmentation for yellow rust disease detection in complex natural backgrounds," *Alexandria Engineering Journal*, vol. 107, pp. 786–798, Nov. 2024, doi: 10.1016/j.aej.2024.09.018.
[54] A. Vephasayanant, A. Jitpattanakul, P. Muneesawang, K. Wongpatikaseree, and N. Hnoohom, "YOLO-based image segmentation for the diagnostic of spondylolisthesis from lumbar spine X-ray images," *IEEE Access*, vol. 12, pp. 182242–182258, 2024, doi: 10.1109/ACCESS.2024.3507354.
[55] N. Sirjani *et al.*, "Automatic cardiac evaluations using a deep video object segmentation network," *Insights into Imaging*, vol. 13, no. 1, Dec. 2022, doi: 10.1186/s13244-022-01212-9.
[56] Y. Ma, P. Chen, M. Gong, Y. Cai, and I. Y. Jian, "The first seasonal green view index mapping dataset across Chinese cities powered by deep learning," *Scientific Data*, vol. 12, no. 1, Aug. 2025, doi: 10.1038/s41597-025-05706-1.
[57] Y. Li, Q. Qu, Y. Yue, Y. Guo, and X. Yi, "Evaluation of children's oral diagnosis and treatment using imaging examination using AI-based internet of things," *Technology and Health Care*, vol. 32, no. 3, pp. 1323–1340, May 2024, doi: 10.3233/THC-230099.
[58] M. E. S. Sabedotti, F. Duarte, P. Koutrakis, P. Santi, C. Ratti, and M. M. Nyhan, "Air pollution and greenspace exposure disparities revealed by hyperlocal exposure metrics across European cities," *Communications Sustainability*, vol. 1, no. 1, Mar. 2026, doi: 10.1038/s44458-026-00046-6.
[59] T. Aikoh, R. Homma, and Y. Abe, "Comparing conventional manual measurement of the green view index with modern automatic methods using Google Street View and semantic segmentation," *Urban Forestry & Urban Greening*, vol. 80, Feb. 2023, doi: 10.1016/j.ufug.2023.127845.
[60] Y. Xia, N. Yabuki, and T. Fukuda, "Development of a system for assessing the quality of urban street-level greenery using street view images and deep learning," *Urban Forestry and Urban Greening*, vol. 59, Apr. 2021, doi: 10.1016/j.ufug.2021.126995.
[61] D. Ki and S. Lee, "Analyzing the effects of green view index of neighborhood streets on walking time using Google Street View and deep learning," *Landscape and Urban Planning*, vol. 205, Jan. 2021, doi: 10.1016/j.landurbplan.2020.103920.
[62] I. A. V. Sánchez and S. M. Labib, "Accessing eye-level greenness visibility from open-source street view images: a methodological development and implementation in multi-city and multi-country contexts," *Sustainable Cities and Society*, vol. 103, Apr. 2024, doi: 10.1016/j.scs.2024.105262.

[63] T. Yang, S. Zhou, A. Xu, J. Ye, and J. Yin, "YOLO-SegNet: a method for individual street tree segmentation based on the improved YOLOV8 and the SegFormer network," *Agriculture*, vol. 14, no. 9, Sep. 2024, doi: 10.3390/agriculture14091620.
[64] S. Lumnitz, T. Devisscher, J. R. Mayaud, V. Radic, N. C. Coops, and V. C. Griess, "Mapping trees along urban street networks with deep learning and street-level imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 175, pp. 144–157, May 2021, doi: 10.1016/j.isprsjprs.2021.01.016.
[65] T. Yang, S. Zhou, A. Xu, and J. Yin, "A method for tree image segmentation combined adaptive mean shifting with image abstraction," *Journal of Information Processing Systems*, vol. 16, no. 6, pp. 1424–1436, 2020, doi: 10.3745/JIPS.02.0151.
[66] S. Tilki, A. Kaplan, and A. T. Zengin, "Zero-shot object detection and segmentation: a focus on street view imagery," in *2024 IEEE 3rd International Conference on Computing and Machine Intelligence*, Apr. 2024, pp. 1–5, doi: 10.1109/ICMI60790.2024.10585911.

BIOGRAPHIES OF AUTHORS

Muhammad Fareez Rashdan received a bachelor's degree in Computer Science from Universiti Teknologi MARA, Melaka, Malaysia, in 2025. He is currently pursuing a master's degree in Computer Science at Universiti Teknologi MARA, Shah Alam, Malaysia. His research interests include image segmentation. He can be contacted at email: muhammadfareezrashdan@gmail.com.

Nurbaity Sabri is a senior lecturer at the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM), UiTM Melaka. She obtained her bachelor's degree in Computer Science from UiTM Shah Alam, followed by a master's and Ph.D. in Computer Science from Universiti Teknologi Malaysia (UTM). Her research interests include image processing, machine learning, and pattern recognition. She can be contacted at email: nurbaity_sabri@uitm.edu.my.

Shafaf Ibrahim is an associate professor at the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM), Malaysia. She holds a Ph.D. in Computer Science and has over ten years of experience in academic research and postgraduate supervision. Her research interests include image processing, deep learning, and medical imaging. She can be contacted at email: shafaf2429@uitm.edu.my.

Nor Masri Sahri is a senior lecturer at the Faculty of Computer and Mathematical Sciences, UiTM Melaka. He earned his Ph.D. in Cyber Security and is an expert in software-defined networking (SDN) and cybersecurity, particularly in the internet of things (IoT). He has over 10 years of experience in both academic research and industry practices, including work in threat detection and authentication protocols. He can be contacted at email: normasri@uitm.edu.my.