






## Measurement of Street Greenness and Interface Permeability Based on Street View Image Analysis

Lei Zhang<sup>1</sup>, Xin Huang<sup>2\*</sup>, Hua Zhong<sup>3</sup>

<sup>1</sup> School of Architecture and Planning, Anhui Jianzhu University, Hefei 230601, China

<sup>2</sup> Department of Architecture, Tongji Zhejiang College, Jiaxing 314051, China

<sup>3</sup> School of the Built Environment and Architecture, London South Bank University, London SE1 0SW, UK

Corresponding Author Email: [huangxin@tjzj.edu.cn](mailto:huangxin@tjzj.edu.cn)

Copyright: ©2024 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ts.410404>

### ABSTRACT

**Received:** 26 February 2024

**Revised:** 13 May 2024

**Accepted:** 2 June 2024

**Available online:** 31 August 2024

#### Keywords:

*street view images, deep learning, street greenness, interface permeability, urban ecological environment*

With the rapid advancement of urbanization, the spatial quality and ecological environment of urban streets have garnered increasing attention. Street greenness and street interface permeability are critical indicators for assessing the ecological environment and spatial quality of streets, playing a significant role in enhancing urban livability. Street view image analysis provides an intuitive and effective method for evaluating these metrics. However, existing research methods face challenges in handling large-scale street view images and complex scenarios, particularly concerning the precision and efficiency of semantic segmentation. This study aims to employ deep learning techniques to perform semantic segmentation on street view images, thereby measuring street greenness and interface permeability. The findings will offer scientific support for urban planning and the formulation of green city policies, holding substantial theoretical and practical value.

## 1. INTRODUCTION

Against the backdrop of accelerated urbanization, the spatial quality and ecological environment of urban streets have gradually become a focus of attention in both academic research and planning practice [1-5]. As an important component of urban public spaces, streets' greenness and interface permeability directly affect residents' quality of life and the urban ecological environment. Street view images, as a crucial data source reflecting the actual conditions of streets, have been widely applied in urban spatial analysis and evaluation in recent years [6-9]. However, how to effectively use street view images to measure street greenness and interface permeability remains a topic worthy of in-depth study.

The significance of researching street greenness and interface permeability lies in its ability to provide data support for urban planning and design, as well as to offer scientific evidence for governments to formulate green city policies and enhance urban livability [10-14]. By analyzing street view images, the greening level and interface permeability of street spaces can be assessed more intuitively, which is of great importance for improving urban microclimates, reducing urban heat island effects, enhancing street vitality, and promoting sustainable urban development [15, 16].

Although some progress has been made in the measurement of street greenness and interface permeability, current research methods still have some shortcomings [17-20]. For example, traditional methods often rely on manual annotation and limited datasets, which are not only time-consuming and labor-intensive but also difficult to adapt to the analysis needs

of large-scale street spaces. Moreover, existing algorithms face challenges in the precision and efficiency of semantic segmentation and feature extraction in complex street view images, making it difficult to comprehensively reflect the real conditions of street environments [21-24].

This paper aims to innovatively achieve accurate measurement of street greenness and interface permeability through deep learning-based semantic segmentation of street view images. The main content of the paper is divided into two parts: first, deep learning technology is applied to street view images for semantic segmentation, accurately identifying street greening and interface elements; second, the identified results are analyzed to measure the greenness and interface permeability of streets, and corresponding optimization strategies are proposed. This research not only addresses the shortcomings of current research methods but also provides an efficient and scalable analysis tool for urban planning, with significant theoretical and practical value.

## 2. SEMANTIC SEGMENTATION OF STREET VIEW IMAGES BASED ON DEEP LEARNING

Street view images contain a large amount of complex visual information, such as buildings, trees, roads, and pedestrians, making it difficult to directly measure greenness and interface permeability and accurately identify and extract relevant environmental elements. Through semantic segmentation technology, different elements in the image can be classified and labeled to clearly distinguish key features such as green areas and street interfaces. However, traditional

image analysis methods face issues of low accuracy and long processing times when dealing with large-scale data. In contrast, deep learning technology, especially convolutional neural networks (CNNs), excels in pattern recognition and feature extraction from images, significantly improving the efficiency and accuracy of street view image analysis. By achieving precise semantic segmentation through deep learning, a solid foundation is laid for the subsequent measurement of greenness and interface permeability, ensuring the scientific and reliable nature of the analysis results.

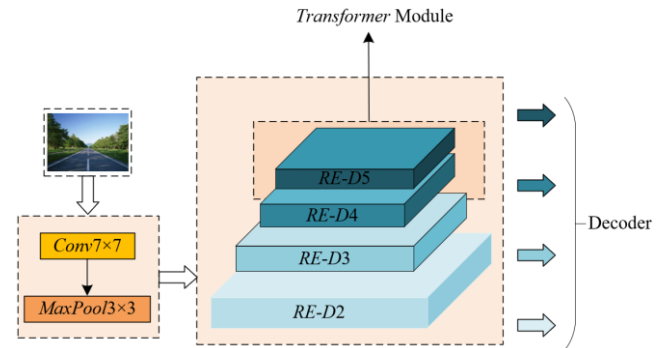
## 2.1 Network structure

This paper proposes a refined multi-scale perception and optimized contour adaptive street scene semantic segmentation network, composed of an encoder, a decoder, a Transformer module, and a segmentation module. The design of the encoder can flexibly adapt to objects of various shapes and sizes in street scenes, such as high-rise buildings, green belts, roads, and pedestrians, which is crucial for accurately segmenting key elements in street images. The feature refinement module (FRM) and feature alignment module are integrated into the feature pyramid network (FPN) to ensure that when accessing multi-level feature information, detailed information can be finely processed and multi-level features can be aligned. The FRM, by refining feature information, helps to improve the segmentation accuracy of green areas and building boundaries, while the feature alignment module solves the problem of feature inconsistency caused by scale changes by aligning feature information at different levels, which is particularly critical in complex street scenes. The Transformer module processes deep encoded feature information through the Miti-DETR decoder, using 100 query objects to perform refined decoding and classification of street scenes, ensuring precise identification of key elements in the street view. The segmentation module combines these decoded pixel-level feature information with mask embedding information to generate accurate segmentation masks, which are then matched with category information to finally generate the semantic segmentation map. This series of processing steps ensures that the network can accurately segment green areas and building interfaces in complex and diverse street scenes, thereby providing reliable data support and a foundation for the subsequent measurement of greenness and interface permeability. The network structure will be detailed below.

### (1) Encoder

The encoder adopts the DC-ResNet101 structure, which effectively enhances the network's adaptability to complex street scenes through two types of bottleneck structures: the standard convolution stacked bottleneck structure and the deformable convolution stacked bottleneck structure. The structure is shown in Figure 1. Specifically, the encoder first processes the 2D input image through the Res1 level, extracting the initial feature map Res1-F1 through operations of a  $7 \times 7$  convolutional layer, group normalization layer (GN), ReLU activation function layer, and  $3 \times 3$  max pooling layer. This step effectively captures low-level features of the street scene, such as edges and basic shapes. Subsequently, the feature map is further processed through the RE, RE3, RE4, and RE5 levels. In these levels, RE and RE3 adopt the standard convolution stacked bottleneck structure, focusing on enhancing the scale perception capability of the feature map,

while RE4 and RE5 adopt the deformable convolution stacked bottleneck structure. By introducing deformable convolutions, the network's ability to perceive complex and variable objects in street scenes is enhanced. Throughout the encoding process, five levels respectively extract five scales of feature maps: RE-D1, RE-D2, RE-D3, RE-D4, and RE-D5. These feature maps provide multi-level visual information of the street scene, ranging from low-level to high-level and from local to global.



**Figure 1.** Encoder of the street view image semantic segmentation network

### (2) Decoder

Figure 2 presents the decoder of the street view image semantic segmentation network. The decoder first inputs the multi-level feature maps processed by the encoder into the FRM, where detailed information in the feature maps is enhanced through refinement processing, resulting in the refined feature map set  $D_{u}^f = \{D_{2}^f, D_{3}^f, D_{4}^f, D_{5}^f\}$ , and these feature maps are then upsampled by a factor of 2. This step ensures that the feature maps retain high-quality detailed features after upsampling, corresponding to important elements in street scenes such as greening and building boundaries, thereby better capturing and reflecting subtle changes in the street. Furthermore, the decoder performs feature alignment on the upsampled feature map set  $D_{u}^i = \{D_{3}^i, D_{4}^i, D_{5}^i\}$  and the corresponding refined feature map set  $D_{u}^f = \{D_{2}^f, D_{3}^f, D_{4}^f\}$  in three stages. The feature alignment module can adjust and align feature information between feature maps of different scales, ensuring that no feature loss or error accumulation occurs during the fusion process, which is of great significance for complex multi-scale features in street scenes. Through this calibration and parallel feature fusion, the decoder can effectively integrate multi-level feature information, generating more refined and accurate decoded feature maps. Finally, the decoder further processes the fused features through a  $3 \times 3$  convolutional layer to generate decoded feature maps of different scales: RE-F2, RE-F3, RE-F4, and RE-F5. These decoded feature maps possess high resolution and rich semantic information, contributing to the precise segmentation of greening areas and building interfaces in street scenes.

### (3) Transformer module and segmentation module

To enable the model to maintain precise object localization while accommodating rich semantic information when processing various complex objects in street scenes, the Transformer module adopts the Miti-DETR decoder, which works in parallel with the decoder integrated with the FRM and feature alignment module. The focus is on the decoding of category information, mask embedding information, and pixel-level decoded feature information. The Miti-DETR

decoder consists of six decoder layers, capable of parallel decoding of the mask embedding information and category information of 100 objects, including the "no object" category,

from the encoded features Res5-F5 through multi-layer perceptrons (MLP) and fully connected layers. The module structure is shown in Figure 3.

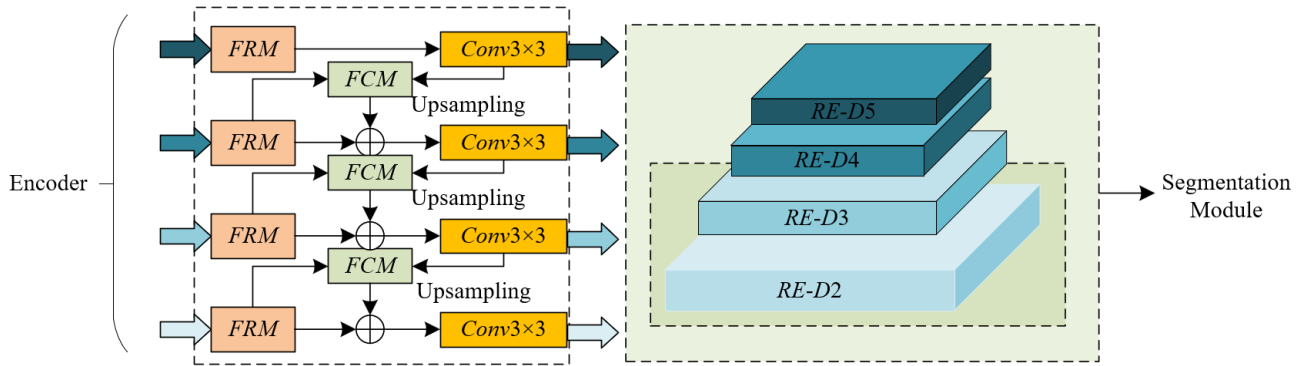


Figure 2. Decoder of the street view image semantic segmentation network

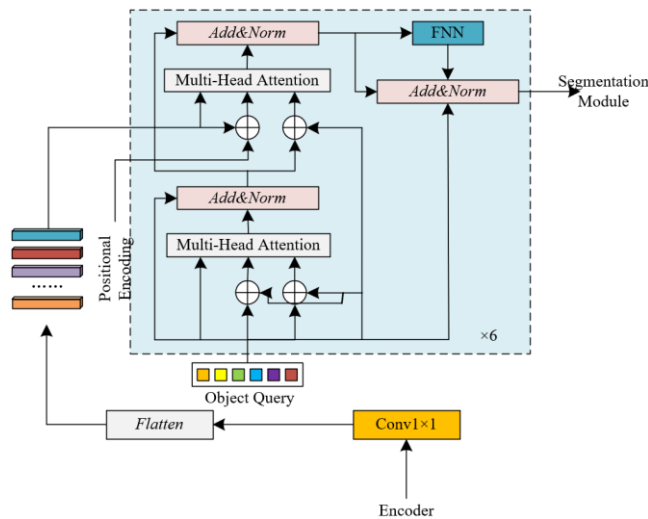


Figure 3. Transformer module of the street view image semantic segmentation network

To accurately segment various elements in the street, such as green vegetation, buildings, roads, and sidewalks, the segmentation module adopts the decoded feature Rest-D2 as pixel-level feature information, embedding the decoded mask information into the 2D pixel-level decoded feature map to generate 100 object mask maps. This process ensures that the model achieves high accuracy in segmenting objects of different scales and categories. Subsequently, the segmentation module matches the mask maps with category information to further generate a semantic segmentation map of 19 categories. Figure 4 illustrates the structure of the segmentation module in the street view image semantic segmentation network.

## 2.2 Deformable convolution stacked bottleneck structure

The measurement and analysis of street greenness and interface permeability rely on the accurate identification and segmentation of street elements. The standard convolutional layers of traditional CNNs, due to their fixed geometric structures, cannot effectively handle the diversity of objects in street scenes, especially in terms of scale variations, posture transformations, and geometric deformations. This limitation makes it difficult for CNNs to accurately capture the true shapes and boundaries of objects when dealing with complex street scenes, thereby affecting the accuracy of segmentation tasks.

To address this issue and enhance the model's adaptability and representation capabilities for various complex objects in street scenes, this paper adopts a deformable convolution stacked bottleneck structure as the basic unit of the encoder. Deformable convolutions introduce a mechanism for learning offset information within the convolutional layer, allowing the network to flexibly adjust the sampling points' positions based on the shape and location of the object. This dynamic adjustment enables the model to overcome the limitations of traditional CNNs in modeling complex deformable objects, allowing for more precise capture of various complex elements in street scenes, such as curved trees, uniquely shaped buildings, and variable road structures. Additionally, when deep CNNs extract feature information from salient objects, they often face difficulties in optimizing nonlinear components as the network depth increases. To address this, this paper introduces skip connections within the ResNet architecture, effectively alleviating the optimization challenges of deep networks by preserving the propagation of linear components.

Specifically, the deformable convolution stacked bottleneck structure consists of a  $1 \times 1$  convolutional layer, a skip connection, and a  $3 \times 3$  deformable convolutional layer. The  $1 \times 1$  convolutional layer is responsible for adjusting the feature channels, ensuring that the number of channels in the feature map can accommodate the subsequent processing steps of the network. The skip connection strategy retains linear information within the network by directly adding the input to the output, alleviating the vanishing gradient problem that may occur during the optimization of deep networks. The  $3 \times 3$  deformable convolutional layer, the core module of the deformable convolution stacked bottleneck structure, employs

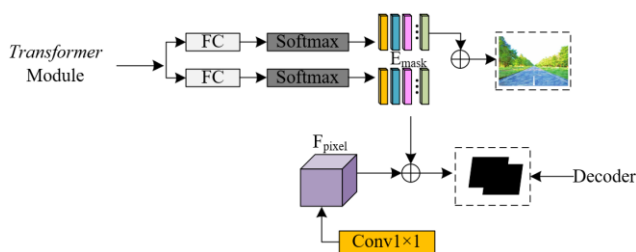


Figure 4. Segmentation module of the street view image semantic segmentation network

a 3×3 convolutional kernel and learns the offset information of each sampling point in the 2D space through a standard convolutional layer. Specifically, this offset information includes adjustments in the x-axis and y-axis directions, ensuring that the convolutional kernel can adaptively adjust the sampling points' positions according to the shape of the object. Through bilinear interpolation, these sampling points can precisely adapt to the complex geometric shapes of objects, ultimately extracting more accurate feature information through the 3×3 convolutional layer.

### 2.3 FRM

During the process of image feature extraction, the gradual downsampling operation compresses high-resolution feature maps into multi-channel low-resolution feature maps, inevitably introducing irrelevant information that can interfere with the segmentation results in both the channel domain and the spatial domain. To enhance the model's ability to capture important features in street scenes, a FRM is introduced to filter out these irrelevant information and reinforce the features relevant to the segmentation task.

The specific architectural design of this module is based on attention mechanisms in both the channel and spatial domains, with the aim of emphasizing relevant feature information critical to street scene semantic segmentation while retaining global feature information. First, the channel domain attention mechanism combines global max pooling (GMP) and global average pooling (GAP) strategies. This complementary representation learning method allows the further emphasis of important feature weights without losing overall features. Additionally, the FRM achieves local cross-channel interactive learning through a one-dimensional convolutional layer, thereby readjusting and expressing the weight proportions of feature channels. This adjustment enables the network to more accurately capture key features related to street greenness and interface permeability analysis. In the spatial domain, the FRM adopts the same global pooling strategy as in the channel domain and processes and weights the spatial domain features through a 7×7 convolutional layer. This approach effectively filters out background noise or other irrelevant spatial feature information, ensuring that the network focuses more on areas critical to street scene segmentation, such as green vegetation and building boundaries. Specifically, the channel domain attention mechanism first uses GMP and GAP in parallel to map the 2D feature map  $F$  into one-dimensional vectors  $I_u=\{i_1,\dots,i_z\}$  and  $N_u=\{n_1,\dots,n_z\}$ . Then, these two vectors are processed through one-dimensional convolutional layers to learn local cross-channel interaction information. This step enables the network to recognize and enhance the correlations between feature channels. After merging this interaction information, it is processed through a Sigmoid activation function layer to generate a feature map  $D_z$  that better expresses important features within the channel domain. This feature map effectively emphasizes channel features crucial to the street scene semantic segmentation task in the channel domain. Assuming the GMP layer is denoted by  $d_{GM}$ , the GAP layer by  $d_{GA}$ , the 7×7 one-dimensional convolutional layer by  $d_{se}$ , and the Sigmoid activation function layer by  $\sigma$ , the expression for  $D_z$  is:

$$D_z = D \times \sigma \left( d_{se} \left( d_{GM} (D) \right) + d_{se} \left( d_{GA} (D) \right) \right) \quad (1)$$

In the spatial domain attention mechanism, the FRM first applies mean and max operations on the feature map  $F_c$  processed by the channel domain attention to generate single-channel feature maps  $D_{AVG}$  and  $D_{MAX}$ . These two feature maps capture important statistical information within the spatial domain. Then, these two feature maps are concatenated along the channel dimension, generating a feature map  $D_{CO}$  that integrates important feature information within the spatial domain. Finally, the  $D_{CO}$  is processed through a 7×7 convolutional layer and a Sigmoid activation function layer, generating the final feature map that focuses on spatial domain-related features as shown in the following equation. This process further eliminates irrelevant information within the spatial domain and enhances important spatial features related to street scene segmentation. Assuming the max operation along the channel dimension is denoted by  $d_{MAX}$ , the average operation along the channel dimension by  $d_{AVG}$ , and the Sigmoid activation function layer by  $\sigma$ , the final feature map is expressed as:

$$D_i = D_z \times \sigma \left( d_{cp} \left( \text{CONCAT} \left( d_{AVG} (D_z), d_{MAX} (D_z) \right) \right) \right) \quad (2)$$

### 2.4 Feature alignment module

In the model, the Feature Alignment Module is designed to effectively reduce the spatial misalignment caused by upsampling, enhancing the ability to capture object contours, especially when processing small targets rich in detail in street scenes. In traditional CNNs, although the depth of the network can enhance the representation of salient object features, it also leads to insufficient capture of small target features. Therefore, multi-level feature fusion in the FPN has become a common solution, effectively representing multi-scale objects by integrating features from different levels through a top-down pathway. However, this upsampling operation can cause spatial misalignment when high-level feature maps are upsampled and fused with low-level feature maps, negatively impacting the precise segmentation of object contours.

To overcome this issue, the Feature Alignment Module introduces the concept of deformable convolutions, adjusting the positions of corresponding pixels in the nearby high-level upsampled feature map  $D_u^i$  based on the low-level feature map  $D_{u-1}$ , thereby resolving the spatial misalignment problem. Specifically, the Feature Alignment Module first identifies the positions that need adjustment in the upsampled feature map and dynamically adjusts these pixel positions based on information from the low-level feature map, ensuring better spatial alignment between the two. This adjustment not only ensures precise spatial fusion of high- and low-level feature maps but also preserves the integrity and accuracy of object contours during feature fusion.

Specifically,  $D_u^i$  and  $D_{u-1}$  are fused by channel concatenation. To avoid the negative effects of spatial misalignment, the feature channel number is first adjusted through a 1×1 convolutional layer, and then the 3×3 convolutional layer is used to learn the offset information and weight information of sampling points in the 2D space. The standard convolutional layer typically uses a 3×3 convolution kernel and employs  $8 \times 3j^2$  feature channels, of which  $2j^2$  feature channels are used to learn the offset information of sampling points, while  $j^2$  feature channels learn the weight

information of sampling points through the Sigmoid activation function layer. Compared with sharing the same sampling point offset and weight information across all feature channels, the Feature Alignment Module further divides the feature channels into eight groups, each independently learning its sampling point offset and weight information. This grouping strategy allows more refined handling of multi-scale object feature representation, ensuring better alignment during high- and low-level feature fusion while conforming to object shape characteristics. Through bilinear interpolation combined with the weight information and offset information of sampling points, the Feature Alignment Module can precisely adjust the positions of sampling points in the feature map, thereby extracting more accurate and representative feature information in the subsequent standard  $3 \times 3$  convolutional layer. Assuming the  $1 \times 1$  convolutional layer is denoted by  $d_z$ , the  $3 \times 3$  convolutional layer by  $d_p$ , and the channel dimension in the 2D space by  $a$ -axis is denoted by  $d_{CH}$ , the module processing process is represented by the following formula:

$$\begin{aligned} & \Delta p_u^a, \Delta p_u^b, \Delta l_u \\ &= d_{CH} \left( d_p \left( d_z \left( \text{CONCAT} \left( D_u^i, D_{u-1} \right) \right) \right) \right) \end{aligned} \quad (3)$$

$$\Delta p_u = \text{CONCAT} \left( \Delta p_u^a, \Delta p_u^b \right) \quad (4)$$

$$\hat{D}_u^j = d_f \left( d_i \left( D_u \right), \Delta p_u, \sigma \left( \Delta l_u \right) \right) \quad (5)$$

## 2.5 Transformer module

The Transformer architecture, due to its powerful self-attention mechanism, excels in capturing long-range dependencies and global features in complex scenes. However, the depth representation learning ability of a purely self-attention network is relatively weak, which can lead to a loss of detail information when extracting deep features, posing a significant challenge for street scenes requiring high-precision segmentation. Miti-DETR introduces a residual self-attention network, effectively preserving non-attention feature information, allowing the deep network to maintain strong representation learning capabilities during propagation. In street scene semantic segmentation, the introduction of Miti-DETR not only enhances the feature-capturing ability in complex street scenes but also significantly improves the model's performance in handling multi-scale objects and detail information. By preserving and utilizing non-attention feature information in deep networks, Miti-DETR addresses the potential degradation issues of pure self-attention networks in deep learning, ensuring higher accuracy and reliability in the analysis of street greenness and interface permeability.

The Transformer module employs six Miti-DETR decoder layers, whose structure is based on DETR with added skip connections to retain the propagation of non-attention feature information. To achieve precise localization and classification of multi-scale objects in street scenes, the Miti-DETR decoder receives encoded features  $D_5^F$ , object queries  $a_{py}$ , positional encoding  $a_{or}$ , and the output from the previous decoder layer as inputs to capture local area features in the image while retaining the positional information of each area. The encoded features  $D_5^F$  are passed through a  $1 \times 1$  convolutional layer to reduce the number of feature channels to 256, and then

reshaped into a one-dimensional sequence, which is combined with the positional encoding  $a_{or}$  to assign positional information to each vector in the sequence. To ensure that the model can accurately identify and classify various objects in street scenes and generate corresponding mask information, the input information is decoded into feature information  $D_{RE}$  of 100 query objects in the six Miti-DETR decoder layers. This feature information is further processed by a fully connected layer and a Softmax activation function to obtain the category relevance and irrelevance information  $D_{CL}$  for the 100 objects. Meanwhile,  $D_{RE}$  is also processed by a MLP and ReLU activation function to generate mask embedding information  $R_{MA}$ .

In the model, the design of the segmentation module directly impacts the segmentation accuracy and the ability to handle complex street scenes. It needs to effectively fuse high-level semantic information and low-level spatial information, enabling the model to handle complex street scenes, particularly those containing various objects and complex backgrounds. Specifically, the segmentation module takes the decoded feature map as input, passing it through a  $1 \times 1$  convolutional layer to reduce the number of feature channels to 256, generating pixel-level feature information  $D_{PI}$ . This step aims to reduce computational complexity while retaining image detail information, making the subsequent mask embedding operation more efficient. Subsequently, the 100 object mask embedding information  $R_{MA}$  extracted from the image queries is embedded into the pixel-level feature information  $D_{PI}$ , generating a 2D mask map  $R_{MA}$ . Specifically, each 256-dimensional vector in the  $R_{MA}$  contains detailed information about the mask, and through a dot product operation with  $D_{PI}$ , the mask information is embedded into the corresponding pixel-level features. This embedding operation ensures that the model can accurately capture and represent objects and regions in the image at the pixel level, thereby improving segmentation accuracy. Next, the generated mask information  $D_{MA}$  undergoes category relevance learning with the category information  $D_{CL}$ . Through this learning process, the model can accurately match the mask information with the object categories and generate a semantic segmentation map  $D_{SE}$ . This segmentation map not only includes the category information of each object in the image but also precisely defines the position and shape of each object in the image.

## 3. MEASUREMENT AND ANALYSIS OF STREET GREENNESS AND INTERFACE PERMEABILITY

Through accurate semantic segmentation of street view images, the model can accurately classify and label different regions in the image, with green vegetation areas being segmented and distinguished from other category regions. Based on this segmentation result, the measurement of street greenness can be further achieved through pixel calculation methods.



**Figure 5.** Street view image with a greenness rate higher than 35%



**Figure 6.** Street view image with a greenness rate lower than 5%

Through the semantic segmentation model, all pixels in the street view image that belong to green vegetation can be extracted. These pixels represent the parts of the image that display green vegetation. Next, the total number of green vegetation pixels is divided by the total number of pixels in the entire image to calculate the proportion of green vegetation coverage in the image, which is the greenness rate. This ratio reflects the proportion of green vegetation in the entire street scene in a specific shooting direction. Assuming the green area is represented by  $S$ , the street length by  $W$ , the average building height by  $H$ , and  $LSL$  as the greenness rate, the commonly used calculation formula for comparing greenness effects is:

$$LSL = \frac{S}{W * H} \times 100\% \quad (6)$$

If based on the semantic segmentation results of the street view image, the calculation formula can be simplified as follows:

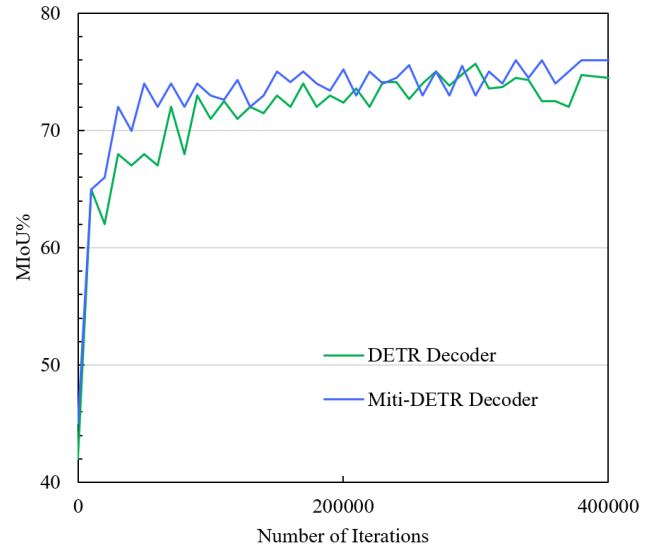
$$LSL = \frac{S}{SYM} \times 100\% \quad (7)$$

where,  $S$  is the green area,  $SYM$  is the field of view area, and  $LSL$  is the greenness rate. Figures 5 and 6 show street view images with a greenness rate higher than 35% and lower than 5%, respectively.

Street interface permeability is mainly measured by calculating the ratio of the number of pixels representing doors and windows in the ground floor buildings to the total number of pixels in the ground floor interface in the street view image. This ratio reflects the permeability of the street building interface, i.e., the interaction between the ground floor buildings and the street environment. The trained semantic segmentation model can accurately identify and label the areas in the street view image that belong to the ground floor building interface, including doors, windows, and other related structures. Through precise segmentation of these areas, the model can distinguish the pixels of door and window areas from the total pixels of the ground floor interface area. Next, the ratio of the total number of pixels of doors and windows to the total number of pixels in the ground floor interface area is calculated as the street interface permeability. Assuming the permeability of a street is represented by  $STL$ , the number of pixels corresponding to the doors and windows of the ground floor buildings at a sampling point by  $D_u$ , the number of pixels corresponding to the ground floor interface at the sampling point by  $D$ , and the total number of sampling points on the street by  $v$ , the calculation formula is:

$$STL = \frac{\sum_{u=1}^v D_u / D}{v} \times 100\% \quad (8)$$

#### 4. EXPERIMENTAL RESULTS AND ANALYSIS



**Figure 7.** Comparison of ablation experiment results of Miti-DETR

As shown in the experimental data in Figure 7, the Miti-DETR decoder consistently outperformed the DETR decoder across multiple iterations. From 0 to 200,000 iterations, the performance of the Miti-DETR decoder gradually increased, starting at 45 points and reaching 74 points at 200,000 iterations, significantly surpassing the DETR decoder's 73 points. Additionally, at 400,000 iterations, the Miti-DETR decoder's score fluctuated between 74 and 76 points, repeatedly exceeding the DETR decoder's highest score of 74.8 points, and eventually stabilized at 76 points. Overall, the Miti-DETR decoder scored higher than the DETR decoder at every stage, particularly in the later iterations, where it demonstrated more stable and superior performance. Analyzing the experimental results, it can be concluded that the Miti-DETR decoder exhibits stronger feature extraction and decoding capabilities in deep learning models, especially in large-scale iterative training, where its ability to recognize key elements in street view images is more accurate. This is due to the introduction of multi-scale perception and optimized contour strategies in Miti-DETR, enabling the model to better capture and decode complex street view features. Moreover, the Miti-DETR decoder refines the classification and recognition of street scenes using 100 query objects, effectively improving the accuracy and stability of the model.

**Table 1.** Comparison results of different networks on the street view image validation set

Network	Backbone Network	MIoU/%
<i>DeepLabV3+</i>	<i>ResNet101</i>	75.69
<i>Auto-DeepLab</i>	<i>DeepLab V3+</i>	76.21
<i>RetinaNet</i>	<i>ResNet101</i>	76.36
<i>Faster R-CNN with ResNet34</i>	<i>ResNet34</i>	76.22
<i>Panoptic FPN</i>	<i>ResNet101</i>	76.54
<i>DETR with ViT-L/16</i>	<i>ViT-L/16</i>	79.12
Proposed Network	<i>DC-ResNet101</i>	79.28

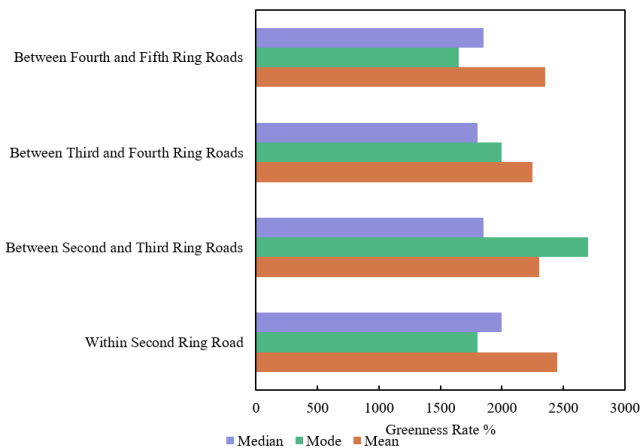
Based on the results from the validation set in Table 1, the performance differences of various networks in the street view image semantic segmentation task can be observed. The

network proposed in this paper, using DC-ResNet101 as the backbone network, achieved a Mean Intersection over Union (MIoU) of 79.28%, the highest among all the compared models. DETR with ViT-L/16 followed closely, reaching a MIoU of 79.12%. In contrast, traditional CNN models, such as DeepLabV3+(75.69%), Auto-DeepLab (76.21%), RetinaNet (76.36%), Faster R-CNN with ResNet34 (76.22%), and Panoptic FPN (76.54%), all had MIoU scores below 79%. This indicates that the proposed network demonstrates a stronger performance advantage in semantic segmentation tasks. From the data analysis, it is clear that the proposed network, utilizing the DC-ResNet101 backbone, outperforms other models in the street view image semantic segmentation task. Its high MIoU indicates that the network has higher precision in identifying

street greenery and interface elements, particularly in complex street view images. The superior performance of the proposed network may be attributed to its innovative integration of multi-scale perception and optimized contour adaptive methods, which better capture detailed features in street view images. Additionally, the proposed network provides a more accurate analytical foundation for measuring street greenness and interface permeability, highlighting its potential value and advantages in practical applications. In comparison, while other models also performed well in specific tasks, they did not reach the overall performance level of the proposed network, further validating the effectiveness and innovation of the proposed approach.

**Table 2.** Performance comparison of the various networks in street view image semantic segmentation

Network	DeepLabV3+	Auto-DeepLab	RetinaNet	Faster R-CNN with ResNet34	Panoptic FPN	DETR with ViT-L/16	Proposed Network
Road	96.45	96.32	98.25	97.25	98.25	97.26	97.35
Sidewalk	79.36	78.25	87.15	82.65	82.36	80.14	85.36
Building	89.36	91.25	95.36	91.48	96.35	93.22	92.36
Wall	48.23	48.25	56.33	51.48	57.23	57.44	57.69
Fence	42.13	45.26	53.25	53.66	52.36	54.26	64.26
Pole	49.36	54.26	57.23	64.25	59.26	61.68	69.26
Traffic Light	26.33	65.24	58.14	70.21	66.59	67.69	74.26
Traffic Sign	61.24	71.26	70.33	78.69	73.21	72.15	80.36
Vegetation	89.36	91.25	90.23	92.15	91.26	91.25	91.23
Terrain	52.36	58.36	59.32	61.24	60.22	62.58	64.15
Sky	92.26	95.36	95.21	94.21	95.36	96.87	94.26
Person	69.26	79.25	85.26	91.26	90.26	85.36	94.25
Rider	48.23	58.36	56.14	63.21	57.93	54.32	67.59
Car	93.26	92.15	97.26	97.36	94.26	96.36	95.31
Truck	56.32	76.12	79.36	76.24	82.36	81.28	82.36
Bus	73.26	86.33	93.21	88.25	93.26	92.85	90.32
Train	62.31	70.15	66.48	67.26	67.36	68.23	79.36
Motorcycle	47.26	56.31	63.26	64.15	64.89	66.32	66.33
Bicycle	72.32	73.12	82.15	77.36	80.36	83.69	79.54
MIoU/%	64.23	73.21	75.15	76.54	76.26	77.41	79.26



**Figure 8.** Comparison of the mean, mode, and median of greenness rates for each ring road in Hefei's old town in summer 2023

From the data in Table 2, it can be seen that the proposed network performs the best in overall performance on street view image semantic segmentation, achieving a MIoU of 79.26%. Specifically, in several key object categories, the proposed network achieved higher segmentation accuracy. For example, in the segmentation of objects such as traffic lights, traffic signs, poles, fences, pedestrians, and motorcycles, the proposed network's accuracy was significantly higher than that

of other networks, particularly in the segmentation of complex objects like traffic lights (74.26%), traffic signs (80.36%), poles (69.26%), and pedestrians (94.25%). This indicates that the proposed network has higher accuracy and robustness in identifying important elements in street scenes. The data analysis suggests that the proposed network demonstrates excellent segmentation capabilities when processing complex street view images, especially for small and difficult-to-recognize objects. Compared to other networks, the proposed network performs outstandingly in multiple object categories, benefiting from its innovative multi-scale perception and optimized contour methods, which allow the network to more accurately capture key features in street view images. This high-precision semantic segmentation capability not only improves the accuracy of street greenness and interface permeability measurements but also provides more reliable technical support for practical applications such as urban planning and intelligent transportation systems.

Hefei's old town (within the moat) is a mature urban living area. After years of changes and accumulation, the street greenness rate and street interface permeability are relatively stable, making it suitable for studying the corresponding patterns between the two. This paper crawled Baidu street view images and measured the street greenness rate and street interface permeability for 20 streets. As shown in Figure 8, the greenness rates of different ring roads in Hefei's old town showed some variation in the summer of 2023. The average

greenness rate within the second ring road was the highest at 2450, while the average rate between the third and fourth ring roads was the lowest at 2250. In terms of mode, the greenness rate between the second and third ring roads was the highest at 2700, while the mode between the fourth and fifth ring roads was the lowest at 1650. The median showed a more consistent trend, with the median within the second ring road at 2000, the median between the second and third ring roads and between the fourth and fifth ring roads both at 1850, while the median between the third and fourth ring roads was the lowest at 1800. These data indicate significant differences in the distribution of greenness rates across different ring roads, with the closer the ring road is to the city center, the higher the average and median greenness rates. The analysis of the greenness rate data shows that the greenness rate within the second ring road is significantly higher than in other ring roads, indicating a higher degree of greening in the city center area, likely due to the bias in urban planning or the historically better foundation

of urban greening. In contrast, the greenness rate between the third and fourth ring roads was relatively low in terms of average, mode, and median, reflecting a lack of greening planning in this area during urban expansion or the influence of other construction priorities. Additionally, the mode between the second and third ring roads is 2700, much higher than in other areas, suggesting that this area contains a few streets with very high greenness density, while the mode between the fourth and fifth ring roads is 1650, indicating that the greening distribution in this area is relatively sparse. Overall, these data reveal an uneven distribution of greening coverage across different ring roads in Hefei's old town, suggesting that in future urban greening planning, priority should be given to optimizing the greening between the third and fourth ring roads and between the fourth and fifth ring roads to improve the overall greening quality of the entire old town.

**Table 3.** Street conditions in Hefei's old town ring roads based on greenness rate classification

Greenness Rate Classification	Location	Street Length (km)	Street Length (km)
Very Poor Greening (<5%)	Within Second Ring Road	1.12	0.27
	Between Second and Third Ring Roads	3.25	0.56
	Between Third and Fourth Ring Roads	5.82	0.66
	Between Fourth and Fifth Ring Roads	8.86	0.61
Poor Greening (5%-15%)	Within Second Ring Road	46.23	12.36
	Between Second and Third Ring Roads	68.26	12.25
	Between Third and Fourth Ring Roads	135.26	15.36
	Between Fourth and Fifth Ring Roads	194.25	13.25
Average Greening (15%-25%)	Within Second Ring Road	178.32	44.58
	Between Second and Third Ring Roads	265.12	47.26
	Between Third and Fourth Ring Roads	425.32	48.23
	Between Fourth and Fifth Ring Roads	678.15	46.26
Good Greening (25%-35%)	Within Second Ring Road	128.36	32.69
	Between Second and Third Ring Roads	167.25	29.58
	Between Third and Fourth Ring Roads	226.36	25.36
	Between Fourth and Fifth Ring Roads	425.23	28.14
Very Good Greening (>35%)	Within Second Ring Road	46.25	11.53
	Between Second and Third Ring Roads	60.32	10.26
	Between Third and Fourth Ring Roads	78.15	9.25
	Between Fourth and Fifth Ring Roads	145.36	10.32



**Figure 9.** Comparison diagram of results from the proposed method and manual measurements

From the data in Table 3, it is evident that there are significant differences in the street greening levels across the different ring roads in Hefei's old town. Streets with poor greening (5%-15%) and average greening (15%-25%) occupy the majority of each ring road, especially between the third and fourth ring roads and the fourth and fifth ring roads, where the street lengths with average greening are 425.32 km and 678.15 km, respectively, occupying the majority of the streets in these areas. In contrast, although the streets within the second ring road are relatively short in overall length, they account for a larger proportion of streets with good greening (25%-35%) and very good greening (>35%), with lengths of 32.69 km and

11.53 km, respectively, reflecting a higher greening coverage rate in the city center. Although the street length with very good greening between the fourth and fifth ring roads is 145.36 km, the highest among all the ring roads, the overall street greening level remains relatively low.

From the data analysis, it can be seen that the greening layout of Hefei's old town shows a clear trend of gradually decreasing from the center to the periphery, particularly within the second ring road and between the second and third ring roads, where the higher greening levels benefit from historical urban planning and maintenance. However, as the city expands outward, the streets between the third and fourth ring



roads and the fourth and fifth ring roads, despite some improvement in greening levels, still have a large number of streets with a greening rate below 25%, indicating a deficiency in greening development in these areas. Particularly, streets with poor greening are mainly concentrated in the outer ring areas, suggesting that these areas need focused greening improvements. Additionally, although the streets between the fourth and fifth ring roads have a greater absolute length of high greening rate streets, this proportion is still insufficient relative to the total street length in that area.

**Table 4.** Street conditions in Hefei's old town ring roads based on interface permeability classification

Interface Permeability Classification	Year	Street Length (km)	Proportion (%)
Very Low Street Openness (<5%)	2021	6.82	3.45
	2023	1.46	0.72
Low Street Openness (5%-15%)	2021	86.36	42.36
	2023	36.25	18.45
Average Street Openness (15%-25%)	2021	69.24	35.26
	2023	118.26	59.32
High Street Openness (25%-35%)	2021	27.36	13.85
	2023	37.26	19.26
Very High Street Openness (>35%)	2021	8.23	4.21
	2023	4.02	2.06

Figure 9 presents a comparison between the results obtained from the proposed method and those from manual measurements. Table 4 shows the street conditions in Hefei's old town ring roads based on interface permeability classification. From the data in Table 4, it can be seen that there were significant changes in the distribution of interface permeability of streets in Hefei's old town between 2021 and 2023. In 2023, the street length and proportion of streets with very low and low street openness decreased significantly. The street length with very low street openness decreased from 6.82 km (3.45%) in 2021 to 1.46 km (0.72%) in 2023, while the length of streets with low street openness decreased significantly from 86.36 km (42.36%) to 36.25 km (18.45%). Meanwhile, the proportion of streets with average street openness increased from 35.26% in 2021 to 59.32% in 2023, indicating an overall improvement in street openness. In addition, the street length with high street openness also increased from 27.36 km (13.85%) to 37.26 km (19.26%). From the data analysis, it can be concluded that the street openness in Hefei's old town improved significantly between 2021 and 2023, especially with a notable decrease in the length and proportion of streets with low openness, indicating that urban planning and renovation efforts have achieved significant results in enhancing street interface permeability. Particularly, the significant reduction in the proportion of streets with low openness, along with a substantial increase in streets with average and high openness, can be attributed to the positive impact of the renovation of old urban areas and the optimization of public spaces. However, despite the overall improvement in street openness, the proportion of streets with very high openness remained low in 2023 (2.06%), suggesting that further enhancement of high street openness levels remains an important direction for future urban planning. Overall, these data indicate that there has been a significant improvement in the street interface permeability in Hefei's old town, but there is still room for further improvement, especially in achieving a layout with higher street openness.

## 5. CONCLUSION

This paper innovatively achieved accurate measurement of street greenness and interface permeability through a deep learning-based street view image semantic segmentation method. The research content is divided into two main parts: first, deep learning technology was used for semantic segmentation of street view images, successfully identifying street greening and interface elements; second, based on the identification results, this paper conducted systematic analysis to measure street greenness and interface permeability, further proposing targeted optimization strategies. Through a series of experiments, such as Miti-DETR ablation experiments and validation set comparisons of different networks, the proposed method demonstrated excellent performance in the semantic segmentation and precision measurement of street view images, validating its effectiveness and reliability in measuring street greenness and openness. The research results show that the proposed method can accurately reflect the actual greening and openness conditions of streets when measuring the greenness and interface permeability of Hefei's old town ring roads in the summer of 2023. The experimental data indicate significant regional differences and temporal changes in street greenness and interface permeability, providing scientific support for urban planning and street view optimization. Particularly in comparison with manual measurement results, the proposed method significantly improved accuracy and efficiency, proving the potential of deep learning-based semantic segmentation methods in urban greening and street view assessment.

This research provides innovative methods and technical means for urban planning and greening optimization, especially in the rapid measurement and analysis of street greenness and interface permeability, with important application value. Through the use of deep learning models, this paper significantly improved the efficiency and accuracy of measurements based on traditional manual methods, providing more refined reference data for urban greening management and street design. However, there are some limitations to the research. First, the acquisition and processing of street view images depend on the quality and coverage of the data, and differences in data across cities or regions may affect the accuracy and generalizability of the measurement results. Second, the performance of deep learning models relies on the quality and quantity of training data, and the current model's performance in extreme environments or complex street views still needs improvement. Future research directions could focus on expanding the diversity and coverage of datasets, improving the robustness and generalization ability of the model, and incorporating more urban planning parameters, such as pedestrian flow density and traffic conditions, to more comprehensively evaluate and optimize urban street views. With these improvements, further enhancement of the practical application of street view image semantic segmentation technology in urban planning is anticipated.

## ACKNOWLEDGEMENTS

This work is supported by Scientific Research Start-up Project for Introducing Talents (PhD) of Anhui Jianzhu University (Grant No.: 2023QDZ21); and the Tongji Zhejiang College Research Launch Project (Grant No.: KY0221516).

## REFERENCES

- [1] Istrate, A.L., Bosák, V., Nováček, A., Slach, O. (2020). How attractive for walking are the main streets of a Shrinking city? *Sustainability*, 12(15): 6060. <https://doi.org/10.3390/su12156060>
- [2] Lundberg, K., Popovski, H., Young, A. (2024). Justice in the streets? Towards a criminology of more-than-spatial justice. *Criminology & Criminal Justice*, 2024: 17488958241257834. <https://doi.org/10.1177/17488958241257834>
- [3] Zhang, P., Zhao, Q., Gao, J., Li, W., Lu, J. (2019). Urban street cleanliness assessment using mobile edge computing and deep learning. *IEEE Access*, 7: 63550-63563. <https://doi.org/10.1109/ACCESS.2019.2914270>
- [4] Bénit-Gbaffou, C. (2016). Do street traders have the 'right to the city'? The politics of street trader organisations in inner city Johannesburg, post-Operation Clean Sweep. *Third World Quarterly*, 37(6): 1102-1129. <https://doi.org/10.1080/01436597.2016.1141660>
- [5] Ping, P., Kumala, E., Gao, J., Xu, G. (2020). Smart street litter detection and classification based on faster R-CNN and edge computing. *International Journal of Software Engineering and Knowledge Engineering*, 30(4): 537-553. <https://doi.org/10.1142/S0218194020400045>
- [6] Zhang, L. M., Chao, W.W., Liu, Z.Y., Cong, Y., Wang, Z.Q. (2022). Crack propagation characteristics during progressive failure of circular tunnels and the early warning thereof based on multi-sensor data fusion. *Geomechanics and Geophysics for Geo-Energy and Geo-Resources*, 8: 172. <https://doi.org/10.1007/s40948-022-00482-3>
- [7] Shi, Y., Campbell, D., Yu, X., Li, H. (2022). Geometry-guided street-view panorama synthesis from satellite imagery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 10009-10022. <https://doi.org/10.1109/TPAMI.2022.3140750>
- [8] Liu, X., Hu, Q., Ai, M., Zhao, P., Yu, D. (2017). An effective spherical panoramic LoD model for a mobile street view service. *Transactions in GIS*, 21(5): 897-915. <https://doi.org/10.1111/tgis.12247>
- [9] Han, Y., Zhong, T., Yeh, A.G., Zhong, X., Chen, M., Lü, G. (2023). Mapping seasonal changes of street greenery using multi-temporal street-view images. *Sustainable Cities and Society*, 92: 104498. <https://doi.org/10.1016/j.scs.2023.104498>
- [10] Hu, A., Yabuki, N., Fukuda, T., Kaga, H., Takeda, S., Matsuo, K. (2023). Harnessing multiple data sources and emerging technologies for comprehensive urban green space evaluation. *Cities*, 143: 104562. <https://doi.org/10.1016/j.cities.2023.104562>
- [11] Zhang, L. M., Cong, Y., Meng, F. Z., Wang, Z. Q., Zhang, P., Gao, S. (2021). Energy evolution analysis and failure criteria for rock under different stress paths. *Acta Geotechnica*, 16(2): 569-580. <https://doi.org/10.1007/s11440-020-01028-1>
- [12] Abdel Sater, R., Mus, M., Wyart, V., Chevallier, C. (2023). A zero-cost attention-based approach to promote cleaner streets: A Signal Detection Theory approach in Parisian streets. *PLOS ONE*, 18(4): e0284272. <https://doi.org/10.1371/journal.pone.0284272>
- [13] Gou, A., Zhang, C., Wang, J. (2022). Study on the identification and dynamics of green vision rate in Jing'an district, Shanghai based on Deeplab V3+ model. *Earth Science Informatics*, 15(1): 163-181. <https://doi.org/10.1007/s12145-021-00691-6>
- [14] Sandhyavitri, A., Wira, J., Martin, A. (2018). Green technology as a strategy in managing the black spots in Siak Highway, Indonesia. *Materials Science and Engineering*, 345(1): 012037. <https://doi.org/10.1088/1757-899X/345/1/012037>
- [15] Cheng, L., Chu, S., Zong, W., Li, S., Wu, J., Li, M. (2017). Use of Tencent street view imagery for visual perception of streets. *ISPRS International Journal of Geo-Information*, 6(9): 265. <https://doi.org/10.3390/ijgi6090265>
- [16] Helbich, M., Danish, M., Labib, S.M., Ricker, B. (2024). To use or not to use proprietary street view images in (health and place) research? That is the question. *Health & Place*, 87: 103244. <https://doi.org/10.1016/j.healthplace.2024.103244>
- [17] Yao, Y., Zhang, J., Qian, C., Wang, Y., Ren, S., Yuan, Z., Guan, Q. (2021). Delineating urban job-housing patterns at a parcel scale with street view imagery. *International Journal of Geographical Information Science*, 35(10): 1927-1950. <https://doi.org/10.1080/13658816.2021.1895170>
- [18] Toikka, A., Willberg, E., Mäkinen, V., Toivonen, T., Oksanen, J. (2020). The green view dataset for the capital of Finland, Helsinki. *Data in Brief*, 30: 105601. <https://doi.org/10.1016/j.dib.2020.105601>
- [19] Park, J., Jeon, I.B., Yoon, S.E., Woo, W. (2021). Instant panoramic texture mapping with semantic object matching for large-scale urban scene reproduction. *IEEE Transactions on Visualization and Computer Graphics*, 27(5): 2746-2756. <https://doi.org/10.1109/TVCG.2021.3067768>
- [20] Wanniarachchi, W.A.M., Wu, W. (2021). Permeability evolution of rock-concrete interfaces in underground lined storage systems. *International Journal of Rock Mechanics and Mining Sciences*, 143: 104792. <https://doi.org/10.1016/j.ijrmm.2021.104792>
- [21] Zhang, L.M., Wang, X.S., Cong, Y., Wang, Z.Q., Liu, J. (2023). Transfer mechanism and criteria for static-dynamic failure of granite under true triaxial unloading test. *Geomechanics and Geophysics for Geo-Energy and Geo-Resources*, 9: 104. <https://doi.org/10.1007/s40948-023-00645-w>
- [22] Ahmadi, E., Cortez, R., Fujioka, H. (2017). Boundary integral formulation for flows containing an interface between two porous media. *Journal of Fluid Mechanics*, 816: 71-93. <https://doi.org/10.1017/jfm.2017.42>
- [23] Liu, Z., Liu, B., Zhou, J., Shen, W., Shao, J. (2022). Triaxial direct shear property and permeability change of interface between high-performance concrete and claystone with water injection and chemical leaching. *Construction and Building Materials*, 356: 129307. <https://doi.org/10.1016/j.conbuildmat.2022.129307>
- [24] Gärdebjer, S., Gebäck, T., Andersson, T., Fratini, E., Baglioni, P., Bordes, R., Larsson, A. (2016). The impact of interfaces in laminated packaging on transport of carboxylic acids. *Journal of Membrane Science*, 518: 305-312. <https://doi.org/10.1016/j.memsci.2016.06.045>