urban science

Article

Quantifying Urban Surroundings Using Deep Learning Techniques: A New Proposal

Deepank Verma [1,*], Arnab Jana [1] and Krithi Ramamritham [2]

[1] Centre for Urban Science and Engineering, Indian Institute of Technology Bombay, Mumbai 400076, India; arnab.jana@iitb.ac.in

[2] Department of Computer Science and Engineering, Indian Institute of Technology Bombay, Mumbai 400076, India; krithi@iitb.ac.in

[*] Correspondence: deepankverma1@gmail.com; Tel.: +91-22-2576-9331

Received: 28 July 2018; Accepted: 26 August 2018; Published: 28 August 2018

Abstract: The assessments on human perception of urban spaces are essential for the management and upkeep of surroundings. A large part of the previous studies is dedicated towards the visual appreciation and judgement of various physical features present in the surroundings. Visual qualities of the environment stimulate feelings of safety, pleasure, and belongingness. Scaling such assessments to cover city boundaries necessitates the assistance of state-of-the-art computer vision techniques. We developed a mobile-based application to collect visual datasets in the form of street-level imagery with the help of volunteers. We further utilised the potential of deep learning-based image analysis techniques in gaining insights into such datasets. In addition, we explained our findings with the help of environment variables which are related to individual satisfaction and wellbeing.

Keywords: environment perception; deep learning; WebRTC; object detection; semantic segmentation; image classification; QGIS

1. Introduction

The relationship between the existence of low-level environmental features [1,2] (the presence of elements such as colours, trees, grass, built architecture, etc.) and people's perception of such an environment has not been entirely understood. The studies on how the high-level descriptors [3] translate to the low-level environmental features are fewer in numbers [1]. It has long been established that the presence of specific elements such as vegetation, in the form of trees and grass, and the presence of water bodies and mountains in the surroundings are more preferred over the human-made elements [2,4,5]. Such findings, however, have not been converted to large-scale urban studies. Environmental preference and perception have been a topic of discussion among researchers in the domains of psychology, environmental studies, architecture, land use, and urban planning. A substantial amount of research has been devoted to the understanding of people's perception of their surroundings. Since the early 1970s, empirical studies have been conducted to assess human responses in different human-made and natural environments. These studies facilitated the design and formulation of environment variables [3] and descriptors [6] to define, assess, and compare different surroundings. Preference-based studies usually involved participants who are presented with the stimulus (such as imagery or a video depicting an urban environment), and their responses based on the assessments are recorded. These responses were usually provided in the form of ratings of the environmental variables and descriptors. The descriptors provide insights into the high-level understanding of the surroundings. Some of the high-level environmental variables are complexity, familiarity, novelty, coherence, prospect, and refuge [3], which induce the perceptions of safety, liveliness, and boredom. Various such experiments [7–12] have allowed researchers to discover the association between the environment and people's preferences for it.

City-wide studies require detailed data collection in the form of street-level imagery which covers every part of an area. Due to the limitations in human and technical resources, researchers chose fewer locations in the cities to collect these datasets, which did not represent entire cities. This affected the generalisability of the conclusions and results. Recent studies working on similar themes have overcome these limitations by utilising publicly available datasets such as Google Street View [13–15] and the crowdsourced datasets of Mapillary, Flickr, Tripoto, and so forth [16,17]. These datasets have been used for the comprehensive evaluation of the urban environment. The estimation of low-level elements present in urban surroundings has become possible with the recent advancements in machine learning, such as image classifiers and object detectors based on deep learning methods. Applications [13,18] utilising such methods and datasets have been developed which present detailed environmental analysis. Even though these datasets have been a boon to the large-scale comparison of the urban environment, such open-data services and datasets have limited coverage in cities. Moreover, the data points (street-level photographs) present in the datasets have a lower temporal resolution. The geotagged street-level images are the fixed points in continuously changing urban environments. Apart from the fixed assets such as buildings, service infrastructures, trees, and water bodies, the presence of people, traffic, and street activity depends on the time of the day. The evaluation of environmental variables such as complexity, familiarity, prospect, and refuge, which induce perceptions of safety, liveliness, or boredom, cannot be accurately checked with such temporally stagnant databases. This study is an attempt to address two voids present in the literature: (a) evaluation of the presence of low-level physical features (elements) in the surroundings at a large scale; and (b) the use of high temporal resolution datasets for the comparative analysis of the urban spaces. This study describes the process of creating an application framework which utilises mobile devices as a platform to collect photographs of the outdoor scenes. It further utilises computer vision techniques to analyse the contents of the collected street view images.

The potential of smartphones in recording responses as a part of surveys has been vastly acknowledged in urban sensing studies [19]. The availability of arrays of sensors in the smartphone device makes it suitable for urban data collection tasks. We demonstrate the creation of a web-application framework which gathers photographic and device-specific information with the help of smartphone sensors. The collected photographic datasets are analysed with the help of a deep learning-based image analysis pipeline which provides the information on (a) the presence of elements such as people, vehicles, buildings, etc.; (b) the existence of different places such as a cafeteria, promenade, etc.; and (c) the percentage of natural and human-made components in the scenes. All the inputs are stored in the database tagged with the location and the device's information.

Deep Learning-Based Visual Perception

The analysis of the visual realm has dominated perception-based studies. Erstwhile studies largely depended on people's abstraction and documentation of the outdoor scenes. The constituents of the scenes such as different physical features can now be identified and extracted through technological means. The human preferences of the urban environment largely depend upon the features [3] such as the typology and architecture of buildings, the presence of natural and human-made features and nuisance-inducing elements, etc. With every glance, human eyes capture information on the presence of physical attributes such as buildings, infrastructures, the style and design of buildings, trees, grass, parks and playgrounds, dilapidation, littering, garbage, advertisement signboards, etc. The presence of these elements may influence the viewer to ignore the scene [20–22]. In large-scale urban research, the gist of the scenes can be captured with the help of computer vision techniques. Conventional techniques such as Histogram of oriented gradient (HOG), Scale invariant feature transform (SIFT), Speeded up robust feature (SURF), thresholding, and edge detections using different filters have been used for image processing tasks including object detection and segmentation. However, these algorithms have faced scalability issues due to inherent complexity in implementation [23].

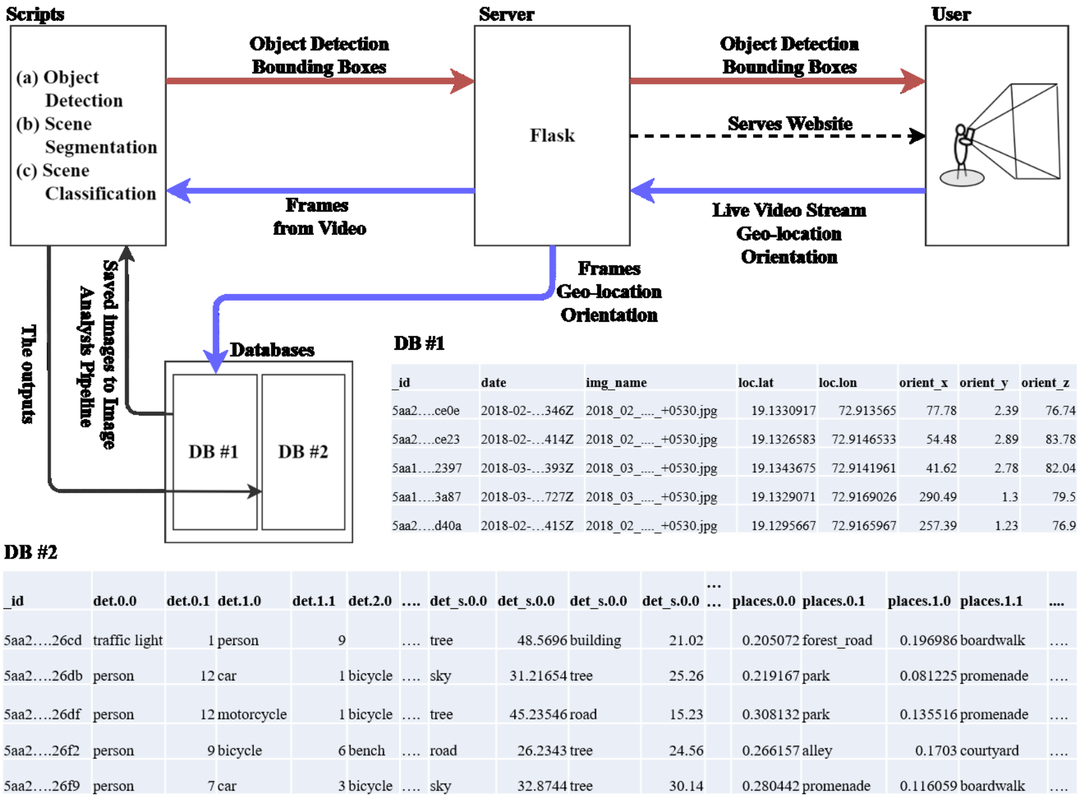
Deep learning techniques have proved their potential in processing large datasets with near-human accuracy. Convolutional neural networks (CNNs) are one of the Deep Learning (DL) techniques which deals with image analysis. CNNs are being actively explored and utilised by researchers in tasks related to object detection, semantic segmentation, and image classification. It has also been used by urban researchers to study the quality of the visual environment. Seresinhe [24] applied a CNN to train 0.2 Million images from a local image database to predict 'scenicness' and beauty in urban and natural scenes. An online game, "Scenic-or-not", was developed, in which users provided scores to the images containing scenic quality. Similarly, Dubey [25] and DeNadai [14] studied the visual attributes of the urban scenes which are determinant of the perceptions of safety and liveliness. They utilised the "Place Pulse" dataset, which consists of a perception-based ranking of street scenes to rate six different perceptual attributes such as safety, liveliness, boredom, wealthiness, depressing, and beautiful. Hyam [18] utilised the Google Vision API and Google Street View imagery to extract automatically generated contextual tags to explore green cover in the city. Similarly, Shen [13] used CNN-based image segmentation architecture to create a database of outdoor urban features, such as trees, sky, roads, etc., extracted from the Google Street View imagery. The database was used to compare the composition of street view elements present in different cities. While the majority of such studies have utilised widely available street view platforms, in contrast, this study aims to generate its own street view dataset with the help of mobile-based application and volunteers. In this way, this study intends to cover the same locations at various time intervals to comment on changing landscapes and their characteristics.

2. The Application Framework

We developed a web application involving a real-time communication service (WebRTC) to send the visual datasets over the internet in real time. Our motive towards the development of a real-time mobile-based application over the typical photo-capturing native application is twofold. First, real-time data transfer for analysis provides a scope for expanding such experiments at a larger scale to various users without any application installation and other device-specific hassles. Secondly, the mobile device can be developed to act as a visual sensor which can map changes in urban landscapes in real time and project the same in real-time dashboards. Figure 1 presents a high-level architecture of the application. We describe the workflow of the application and will cover the details in subsequent sections. The user opens up the web app through a mobile browser with an intention to collect photographs of the outdoor scenes. The application starts capturing a live video feed along with the orientation and the geolocation details from the device's sensors. The captured data from these sensors is continuously sent to the server. The server appends the incoming data stream into the database for further analysis. The server posts back the outputs, such as detected objects, gathered from the image processing pipeline.

Apart from data collection, the intention behind the creation of the app is also to provide the user with a sense of engagement and novel value addition. This application provides real-time object detection on static and moving objects to the users while storing the video frames in the database while the app is active. At present, the application is just a proof of concept, in which application scalability and production-ready features are not administered. The application is served by a local server. To overcome device-specific security warnings regarding the use of geolocation and on-device camera in mobile devices, we used the Opera browser, which allows the application to access these sensors with minor warnings. The permanent solution to this issue would be to purchase a website domain and related security certificates. The application is tested on various entry-level Lenovo (model no. K8) and Motorola (model no. G5) smartphones. The data communication latency is checked with the help of web developer tools and Python-based timer functions. The latency statistics ranged from 0.2 s to 1.8 s, based on the quality of the internet connection available to the user's mobile phone. The application forces the on-device camera to capture images with a resolution of 480 by 640 pixels, which can be modified to include higher resolutions. However, the quality of the images is a trade-off between the internet bandwidth and the performance of computer vision algorithms on such images. The architecture comprises: (a) A high-performance workstation acting as a server; (b) a WebRTC platform which maintains live video communication with the server. Individual frames from the video are used as the input for analysis. The WebRTC platform is actively developed and maintained, which is also cross-compatible in different mobile- and desktop-based browsers; (c) deep learning-based Python scripts which compute the outputs, which are then sent to the user (client) and the database; (d) a flask server framework to host the client webpage and the necessary scripts; (e) a MongoDB database to store and fetch the datasets. The server is chosen such that it can handle high web traffic, including the processing of incoming data, and provide outputs in real time. The server is configured with Intel Xeon, which has 32 GB of memory, along with Nvidia K2000 Graphics, with 2 GB of memory. The server runs Flask, which is a Python-based server framework. Flask is preferred over other such frameworks due to the ease in integrating the deep learning models which are based on Python. The server hosts the pretrained deep learning models and scripts which perform object detection, scene segmentation, and classification. The front-end interface of the application shows the video in the live feed from the device's camera, while simultaneously displaying the outputs in the form of bounding boxes with labels of the objects detected in the feed.

The server handles the incoming data at two stages: (i) real-time processing of incoming data and (ii) storage of data for detailed analysis. The real-time processing utilises a lightweight object detection module (detailed in the next subsection) and produces outputs to the client. The segmentation, classification, and the main object detection module require a large amount of computational resources and time. Therefore, these modules are implemented after data is collected and stored in the database. The two-level hierarchical structure is created due to the limitations in the present hardware setup and server configuration. MongoDB is selected as a database due to its ease in the setup process. It does not require a predefined schema for data storage and can hold different types of documents in a single collection, in which the content, size, and number of fields of the document can vary from each other. Here, DB1 (Figure 1) stores the location, device's orientation, and captured image, which are further analysed by deep learning scripts, the results of which are stored in DB2 (Figure 1). The list of details acquired by the application is as follows:


Scripts
(a) Object Detection
(b) Scene Segmentation
(c) Scene Classification
Object Detection Bounding Boxes
Server
Flask
Frames from Video
Object Detection Bounding Boxes
User
Serves Website
Live Video Stream Geo-location Orientation
Frames Geo-location Orientation
The outputs
Saved images to Image Analysis Pipeline
Databases
DB #1
DB #2

DB #1

_id	date	img_name	loc.lat	loc.lon	orient_x	orient_y	orient_z
5aa2….ce0e	2018-02-…346Z	2018_02_...._+0530.jpg	19.1330917	72.913565	77.78	2.39	76.74
5aa2….ce23	2018-02-…414Z	2018_02_...._+0530.jpg	19.1326583	72.9146533	54.48	2.89	83.78
5aa1….2397	2018-03-…393Z	2018_03_...._+0530.jpg	19.1343675	72.9141961	41.62	2.78	82.04
5aa1….3a87	2018-03-…727Z	2018_03_...._+0530.jpg	19.1329071	72.9169026	290.49	1.3	79.5
5aa2….d40a	2018-02-…415Z	2018_02_...._+0530.jpg	19.1295667	72.9165967	257.39	1.23	76.9

DB #2

_id	det.0.0	det.0.1	det.1.0	det.1.1	det.2.0	….	det_s.0.0	det_s.0.0	det_s.0.0	det_s.0.0	… …	places.0.0	places.0.1	places.1.0	places.1.1	....
5aa2….26cd	traffic light	1	person	9		….	tree	48.5696	building	21.02		0.205072	forest_road	0.196986	boardwalk	….
5aa2….26db	person	12	car	1	bicycle	….	sky	31.21654	tree	25.26		0.219167	park	0.081225	promenade	….
5aa2….26df	person	12	motorcycle	1	bicycle	….	tree	45.23546	road	15.23		0.308132	park	0.135516	promenade	….
5aa2….26f2	person	9	bicycle	6	bench	….	road	26.2343	tree	24.56		0.266157	alley	0.1703	courtyard	….
5aa2….26f9	person	7	car	3	bicycle	….	sky	32.8744	tree	30.14		0.280442	promenade	0.116059	boardwalk	….

Figure 1. Mobile application architecture and database (DB) attributes.

2.1. Device Orientation

The browser provides access to the device's onboard sensors, such as accelerometers and magnetometers, which provide the orientation of the device with respect to Earth's coordinate frame. The device returns alpha, beta, and gamma angles, which are the angles made by the device from the x, y, and z axes. The alpha value is the device's heading relative to the Earth's magnetic north. The alpha value is essential in providing the heading (direction) of a person while capturing a photograph. The values of the beta and gamma angles are required to filter out the images captured which are not focused towards the street or buildings in the proper orientation.

2.2. Geolocation

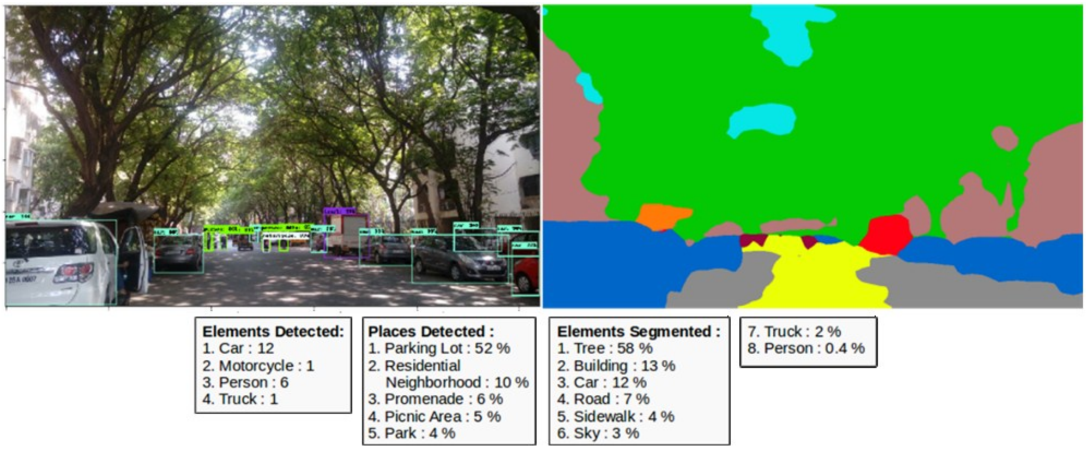
The geolocation API gives access to the installed Global Positioning System (GPS) hardware in a device. It provides locational data as a pair of latitude and longitude coordinates with accuracy in metres. The precision of the GPS is set to high accuracy to obtain continuous updates with slight changes in the user's position. The geolocation API (https://developer.mozilla.org/en-US/docs/Web/API/Geolocation) is based on HTML5, which works across all recent mobile browsers.

2.3. Images from the Video Stream

The video stream from the device is broken down into the individual frames for obtaining the information on the detected objects and places with the help of deep learning convolutional neural network-based models. The convolutional neural network is a variant of the artificial neural network, which is primarily used to perform image analysis. This model learns to detect the features from the provided set of image datasets to frame meaningful representations and associations. Upon training with large datasets, these models perform with near-human accuracy. Such neural network models are complex matrix operations designed to process datasets and provide predictions on specific tasks. To overcome the inconvenience of redesigning such matrix operations for every task, several prebuilt open libraries have been developed by the academia and industry. Some of the libraries are Tensorflow [26], Pytorch [27], Keras [28], and Caffe [29]. In this study, Tensorflow, Keras, and Pytorch are utilised to analyse image datasets.

2.3.1. Object Detection

We selected the object detection model based on Tensorflow due to two reasons. The Tensorflow object detection API provides detailed documentation on the usage of the detection pipeline in prebuilt image datasets as well as a custom user-based dataset. Furthermore, the large user community is active in suggesting options and alternatives to avoid the errors while working with the specific modules in the library. The typical object detection task requires weights trained on the deep learning architectures using a large database of images. In this study, these weights are downloaded from the Tensorflow official GitHub repository (https://github.com/tensorflow/tensorflow). The repository contains a variety of pretrained weights along with the statistics on accuracy and inference-time latency metrics. The choice of latency over the accuracy depends upon the purpose and the usage. The available models have been trained on different datasets such as COCO (Common Objects in Context) [30], Kitti [31], and Open-Images [32] using various state-of-the-art architectures such as ResNet [33] and Inception [34]. We selected COCO-based models as they included most of the objects present in the urban outdoor street view scenes, such as a person, bicycle, car, motorcycle, bus, truck, traffic light, fire hydrant, stop sign, bench, and various animals. Two models are chosen for the object detection task in this study: (a) Faster RCNN NAS [35] and (b) Mobilenets [36], in which the former is used for the more intensive object detection task and the latter for providing real-time inferences to the client. Faster RCNN NAS currently has the highest COCO mean average precision (an accuracy metric) amongst all the models available in the Tensorflow model zoo, thereby increasing its inference time, even with the high-performance graphic processing units (GPUs). The MobileNets model is the fastest detector amongst all the available object detector models in the Tensorflow library, but has lower precision for the detected objects. The outputs from MobileNets are not considered for scene evaluation and are utilised only for returning outputs to the mobile application. The inference script provided in the Tensorflow object detection repository (https://github.com/tensorflow/models/tree/master/research/object_detection) is utilised as a part of the created web-based application. Figure 2 shows results obtained from the object detection task using the Faster RCNN NAS model. The script provides a rectangular bounding box over the elements detected and the probability of the correct detections. It is common to include a probability threshold over which the particular detected element is shown and counted by the algorithm. We randomly tried various thresholds and tested them over different street images, finally agreeing on keeping it at 20%. In other words, if the probability of detection of the element is found to be above 20% by the model, the element is detected and displayed with the bounding box.



Elements Detected:
1. Car : 12
2. Motorcycle : 1
3. Person : 6
4. Truck : 1

Places Detected :
1. Parking Lot : 52 %
2. Residential Neighborhood : 10 %
3. Promenade : 6 %
4. Picnic Area : 5 %
5. Park : 4 %

Elements Segmented :
1. Tree : 58 %
2. Building : 13 %
3. Car : 12 %
4. Road : 7 %
5. Sidewalk : 4 %
6. Sky : 3 %
7. Truck : 2 %
8. Person : 0.4 %

Figure 2. Outputs from the deep learning pipeline.

2.3.2. Semantic Segmentation

Semantic segmentation refers to the partition of the image into meaningful predefined classes, such as sky, trees, vehicles, etc. It is a pixel-wise classification of the image (it classifies every pixel, which are then aggregated to represent the class). These created segments provide a high-level overview of the elements present in the image. Inferences such as the percentage of the scene covered by various elements can be obtained from the results of this model (Figure 2). Among the widely used pixel-based classification techniques are Segnet [37], FCN [38], and Deeplab [39]. We utilised a semantic segmentation technique based on PSPNet [40]. The Pyramid Scene Parsing Network (PSPNet) has shown significantly better accuracy statistics on various datasets and secured the first rank on the ImageNet scene parsing challenge in 2016. The author of PSPNet has made training weights and the architecture of the model available for use in research purposes. The PSPNet architecture is trained on Pascal VOC [41], Cityscapes [42], and ADE20K [43] datasets. These datasets have been widely used and researched in studies related to street view segmentation. This particular study utilised the model trained on ADE20K (https://github.com/Vladkryvoruchko/PSPNet-Keras-tensorflow), which consists of 150 labels related to the elements present outdoors and indoors. PSPNet trained on the ADE20K dataset is chosen due to two important factors. First, the prediction time of the model trained with the ADE20K dataset is faster than other two datasets. Secondly, the classes present in the ADE20K dataset covers classes from the COCO dataset used in object detection tasks. Other common street view elements which do not feature in the COCO dataset, such as sky, trees, lampposts, and traffic signs, are present in the ADE20K dataset. The outputs from both the detection and segmentation tasks are able to capture most of the scenic attributes from street view imagery.

2.3.3. Scene Classification

The classification of a captured image into different semantic classes related to common urban places has the potential to categorise urban landscapes at a large scale. The Places dataset [44] consists of around 10 million images depicting indoor and outdoor views. These places are classified into more than 400 classes, where each class consists of more than 5000 images. Images are trained on state-of-the-art ResNet [33], VGG [45], and GoogleNet [34] architectures and the trained weights are being made available for the research community. The scene classification model used in this study is based on Places365, which is a subset of the Places database and consists of 365 classes. It classifies the images into a list of the top five classes along with the classification probability (Figure 2). The common classes which can be detected in the scene classification are an alley, cafeteria, balcony, department stores, forest path, park, courtyard, promenade, etc. The Places database has been used by several visual perception-based studies to annotate the collected imagery into its scenic attributes. This study makes use of the working demo script provided by CSAIL (https://github.com/CSAILVision/places365), implemented in Pytorch.

3. Prototype Implementation

To test the applicability of the designed framework in real urban scenes, we selected a geographical area of 0.25 km^2 to conduct the study. The selected area is part of an academic campus. The data collection process involved usage of the created application to collect photographs from different locations by sending streaming data to the server. Approximately 2000 photographs were collected in 12 days with the help of five volunteers, who were given access to the application to conduct the survey. The streets were predefined in the selected area for conducting the survey. However, the selection of a particular location or set of locations were at the volunteer's discretion. There were no time restrictions on the collection of photographs, but several photographs were filtered out from the database which were sent to the server after a specific time (0600 to 1800 h) due to the lower visibility outdoors. Furthermore, the collected frames which had been tilted by around ±20 degrees in the y-axis were filtered out with the help of database operations. Each street view scene is processed by the Python-based scripts, which provide 15–30 attributes as outputs, such as detected objects, segmented features, and places classification, with the probability of the top five classes (Figures 1 and 2). These results are stored in the database, which creates a unique ID for every saved image and outputs along with its coordinate information. This information can be easily plotted and analysed through any Geographical Information System (GIS) based platform.

We utilised QGIS to represent our findings. QGIS provides the simple implementation of rule-based styling, labelling, and filtering, which is helpful in representing large datasets. With different attributes (results obtained from the DL pipeline) tagged to one data point, the options for its representation are endless. We calculated environment variables such as naturalness, complexity, openness, and familiarity (discussed in subsequent sections) and represented them with the help of the QGIS platform. The naturalness is calculated as the percentage of natural elements present in the scene, while complexity corresponds to the diversity of the street scenes. Similarly, openness is calculated with the sky-view factor and familiarity of the place with the image classification. Variables such as the naturalness and diversity of the scene are dependent on time, and hence show distinct outputs at different hours at the same location. Other variables such as the amount of visible sky and the places are the fixed characteristics of a particular location. The variations in data collection are recorded during different hours, as shown in Figure A1. We segregated the collected data into three segments for comparison: morning (680 images; 0600–1100 h), noon (619 images; (1100–1500 h), and evening (794 images; 1500–1800 h). The number of images collected in each slot is nearly equal. However, we represent the comparison of time-variant variables in the morning and evening slots. To ease up the representation of the data, we aggregated the results tagged with various coordinates points with the help of a hexagon grid. Each side of the hexagon in the grid has a length of 7.5 m. Figure 3 represents the number of collected data points aggregated under each hexagon. The number of data points ranges from 1–18. The aggregation of data points does result in an oversimplification of the inherent characteristics of the underlying data points. However, hex bins have been utilised in the representation of larger datasets for which minute details are not discernible with the naked eye.

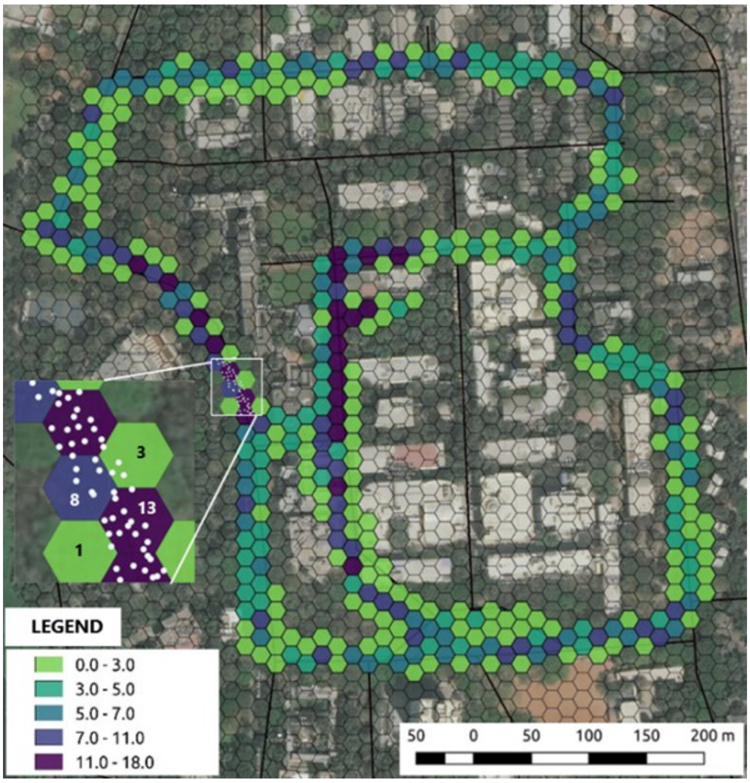


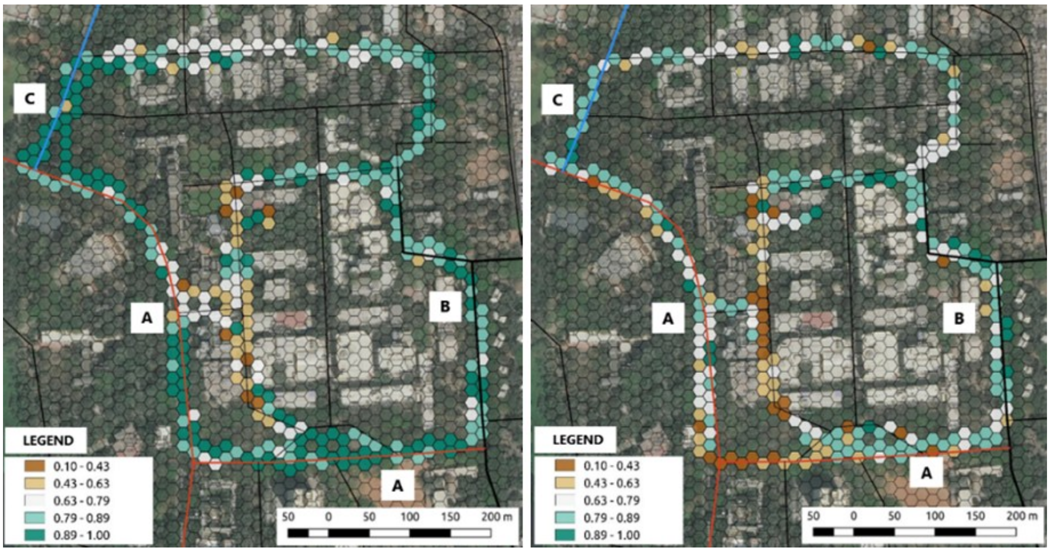
Figure 3. Map showing study area, data collection, and coverage. The zoomed-in area shows the location of observation points. Note: The legend represent number of observations.

The morning- and evening-collected datasets are not balanced for every location in the study area; therefore, hex-bin maps show slight variation at these intervals. The values calculated for various environmental variables (Figure A1) of each data point are averaged under the superimposed hex bin. Due to the uneven distribution of data points under each hexagon, the average value might generalize the output which will result in a different colour representation of the individual bin. This method of representation might produce an erroneous depiction of dataset values. However, in this particular use case, the values do not change aggressively over their neighbouring data points. Hence, in our opinion, the hex bins are still able to represent the true character of the underlying group of data points.

3.1. Percentage of Natural Versus Human-Made Elements

The presence of naturalness in the scene is associated with the restoration of attention [46,47], stress recovery [48], satisfaction, and well-being [49]. Naturalness has been studied and shown to impart long-term psychophysical health benefits. It is also studied in preference-based studies as a scene attribute related to the perception of safety, liveliness, boredom, etc. [50]. The presence of natural elements is one of the important elements of visual landscape assessment in neighbourhood satisfaction [51]. In contrast to studies which related naturalness only to the presence of vegetation involving trees and grass, this study considers the composition of the scene with various other natural elements at different intervals during the day.

The scene segmentation task, being a pixel-wise classification, provides the percentage share of each element present in the frame, such as trees, vehicles, roads, paths, etc. These percentage values are aggregated to give the share of the natural and human-made elements in a particular scene. These values show variation with respect to the time (Figure 4), which is expected due to the presence of different elements (such as people, vehicles, animals) in the frame taken at different time intervals. Among the 150 labels present in the ADE20K dataset, only 16 labels are based on natural elements (sky, tree, grass, earth, mountain, plant, water, sea, sand, river, flower, dirt track, land, waterfall, animal, and lake). The locations which show less variation with the time are the least dynamic locations regarding the movement of people and traffic (streets 'B′ and 'C′), with the opposite being true for the locations showing high variation. The street 'A′ is a major street in the campus, which shows a significant difference in morning and evening naturalness values due to traffic peak hours at evening. The legend in the map shows the percentage of pixels covered by natural classes scaled from 0 to 1.



(**a**) (**b**)

Figure 4. Percentage of naturalness during the morning (**a**) and evening (**b**). Note: The legend explains the percentage of naturalness (Very low: 0.10–0.43, low: 0.43–0.63, medium: 0.63–0.79, high: 0.79–0.89, very high: 0.89–1.0).

3.2. The Diversity of the Scene

The complexity and the visual richness of outdoor scenes have been studied in experiments related to environmental psychology and behaviour [3]. Herzog [52] defined the complexity of the scene as "how much of the scene contains many elements of different kinds". It has also been considered as a prime variable of aesthetic judgements in environmental perception studies [53]. In this study, the assessment of complexity is carried out by calculation of the diversity of urban elements present in collected street view images. Individual preferences to specific areas depend upon the presence of diversity. According to [52], exploration of the surroundings is stimulated by the diverse surroundings. Furthermore, diversity reflects how much activity is present in a scene; "if very little is going on, preference might be lower" [53]. In earlier studies, complexity was evaluated with the help of volunteers included as a part of on-site visual assessment surveys. The scene complexity has been calculated with the help of image attributes such as edges, features, and histograms to provide inputs on the diversity of the scene [54]. With the help of DL techniques, individual elements can be identified with human-level accuracy, which can emulate the findings and assessments as provided by a human observer. In this study, elements in the scene are detected via two models: (a) object detection and (b) scene segmentation. The diversity metrics are calculated for each of the models according to the different types of outputs provided by these individual tasks. Object detection detects elements in the scene as separate instances of the same detected class (for example: Car: 2; Person: 3; Motorcycle: 4), whereas segmentation provides the percentage of the area covered by each class (e.g., Car: 10%; Person: 0.08%; Motorcycle: 10%).

Hill numbers are used to calculate diversity metrics due to their ease in the interpretation of diversity values. The commonly used Shannon and Simpson's index is prone to misinterpretation of diversity due to its nonlinear nature. Hill numbers convert diversity values into effective numbers which behave linearly in relation to the changes in diversity values [55]. The comparison of Hill numbers between any two data points (photographs) would follow the linear rule, as opposed to diversity indices other than hill numbers. The general equation of a Hill number is given as:

$$^{q}{}_{D} = \left(\textstyle\sum_{i=1}^{s} p_i^{q}\right)^{1/(1-q)} \tag{1}$$

For diversity calculation of outputs from the segmentation task, the above expression can be interpreted as *D*: Diversity, *p*: Proportion of class to total detected classes, and *s*: the total number of classes detected.

q = 0 represents the richness of elements. It provides information on unique elements detected in the scene. For example, a scene having (Car: 20%; Person: 5%; Motorcycle: 35%) has the richness of 3.

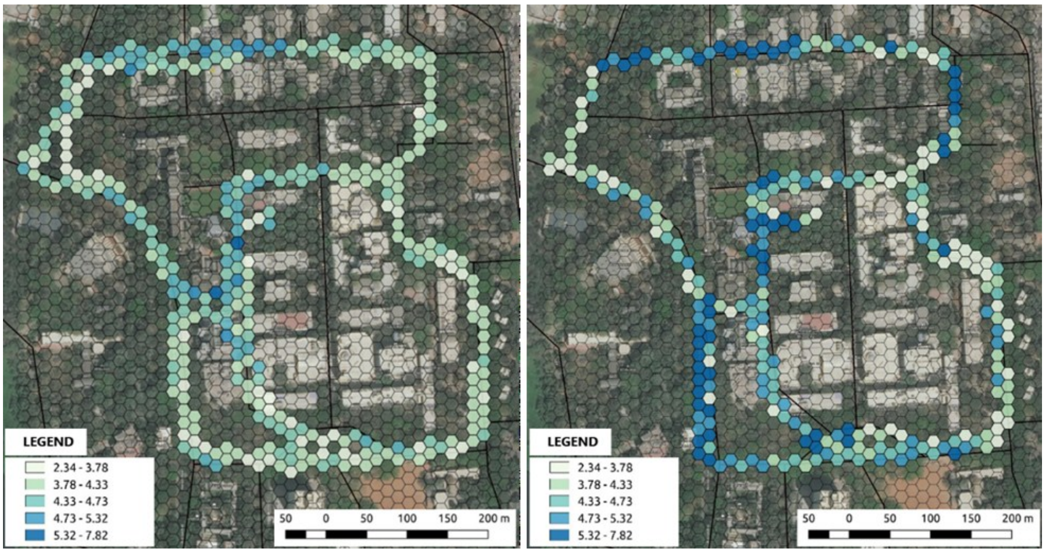
q = 1 represents the exponential of the Shannon entropy index. It weighs each class according to the frequency or weights of the present elements [55,56]. High entropy values indicate randomness, while lower values show order.

$$^{1}{}_{D} = \exp\ \left(-\textstyle\sum_{i=1}^{s} p_i \ln p_i\right) \tag{2}$$

q = 2 represents the inverse of the Simpson concentration index. It is more sensitive to dominant classes detected in the scene. The formula includes the sum of squares for which the lower values of detected elements become insignificant.

$$2_{D} = \frac{1}{\sum_{I=1}^{S} P_i^2} \tag{3}$$

Figure 5 shows the difference between entropy (exponential of the Shannon entropy index) values calculated on elements obtained from the segmentation task. The entropy values range from 2.3 to 7.8. The difference between the entropy values at morning and evening hours suggests that (a) a lower number of elements is detected in the morning hours compared to in the evening hours; (b) the amount (percentage area in a scene) of natural elements present in the scene is higher in the morning than in the evening. The natural elements such as trees and sky amount for the maximum coverage in a scene, and due to the absence of elements such as people, buses, and motorcycles during the morning hours, the variability of the scene is lower.

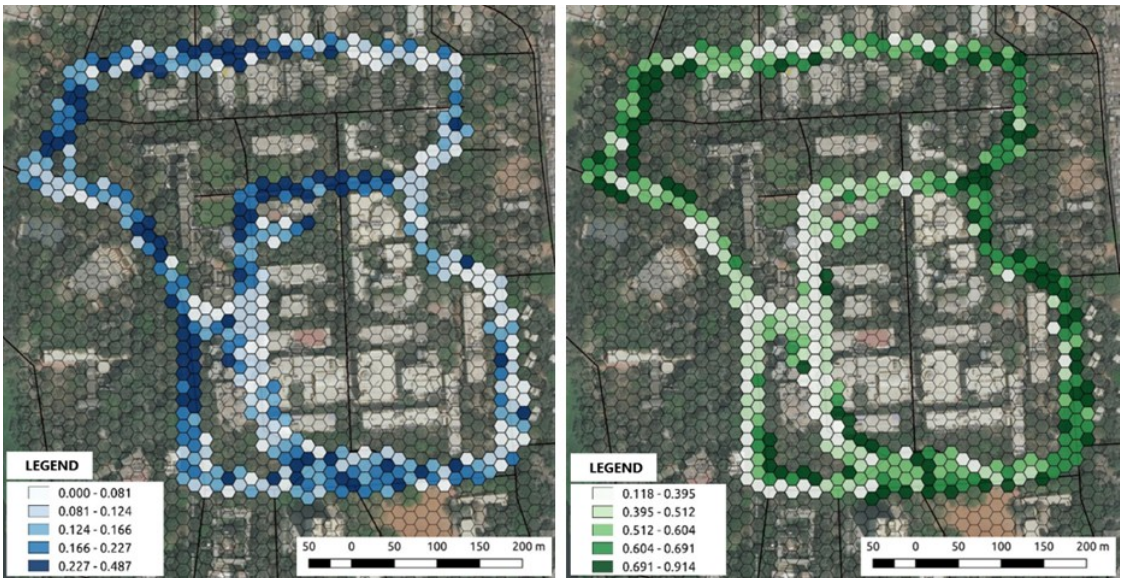


(**a**) (**b**)

Figure 5. The difference in Shannon entropy values during the morning (**a**) and evening (**b**) hours. Note: The legend explains the diversity in terms of entropy (Very low diversity: 2.34–3.78, low: 3.78–4.33, medium: 4.33–4.73, high: 4.73–5.32, very high diversity: 5.32–7.82).

3.3. Visible Sky and Green

The amount of visible sky and the green content elucidates the scene properties such as openness, depth, enclosure and expansiveness, etc. [57]. The sky view factor has been one of the essential parameters in studies related to the design and planning of urban structures [58]. The effect of visible sky and green cover has been studied explicitly in preference-based studies as it had been a prime constituent of assessment of the variables such as prospect and refuge [59]. The feelings of safety and pleasure are derived from surroundings providing a sense of enclosure and views [60]. Studies such as [13] have created tools to analyse the sky–green ratios for cities, which help urban planners to explore urban areas with interactive visual analysis. Visible sky and green percentages are extracted from the scene segmentation task. The range of values (0–0.48) in the percentage of sky view is lesser than that of the presence of a green character (0.1–0.91) of streets. The value of green cover calculated from scene segmentation should not vary with time; however, the variation in the morning, noon, and evening green percentage values are recorded (Figure A1), due to the presence of other street elements in the collected photographs at different timings. In contrast, the sky view factor only has a slight variation (Figure A1), as expected, in the values at different times. The higher percentage (Figure 6) of the visible sky and green cover at some locations can be interpreted in two ways: being due to (a) the absence of high-rise built structures and trees; and (b) the absence of human-made elements along the streets and huge green canopy cover over the street, respectively. Also, the low green cover with high sky view percentage may suggest the presence of openness, while the low green cover and low sky view percentage may provide a sense of enclosure. The assessment of greenness through street view imagery is helpful in identifying the green coverage which is within visual reach of the people.



(**a**) (**b**)

Figure 6. Maps showing sky view (**a**) and green cover (**b**) percentages. Note: The legend in (**a**) explains the percentage of sky view (Very low: 0.00–0.081, low: 0.081–0.124, medium: 0.124–0.166, high: 0.166–0.227, very high: 0.227–0.487); similarly the legend in (**b**) explains the percentage of green cover (Very low: 0.118–0.395, low: 0.395–0.512, medium: 0.512–0.604, high: 0.604–0.691, very high: 0.691–0.914).

3.4. The Places Present in the Scene

Preference to the particular scene is affected due to the presence of familiar and unfamiliar urban places in it [9,61]. The variables such as familiarity, identifiability, and novelty have been studied in urban scenes containing restaurants, coffee shops, factories, apartments, etc. The presence of particular places also acts as an element of mental, cognitive maps which people continuously utilise to travel around [62]. The presence of such places may induce the feeling of safety, liveliness, or boredom (for example, the presence of factories and alleys induce a feeling of desolation [61]). Classification of such places is done with the help of the scene classification task. The study area has a significant amount of green cover and identical built structures compared to any general urban area; therefore, the classes such as forest road, embassy, and courtyard are present in dominance. Figure 7 represents the most dominant classes present in the particular location.

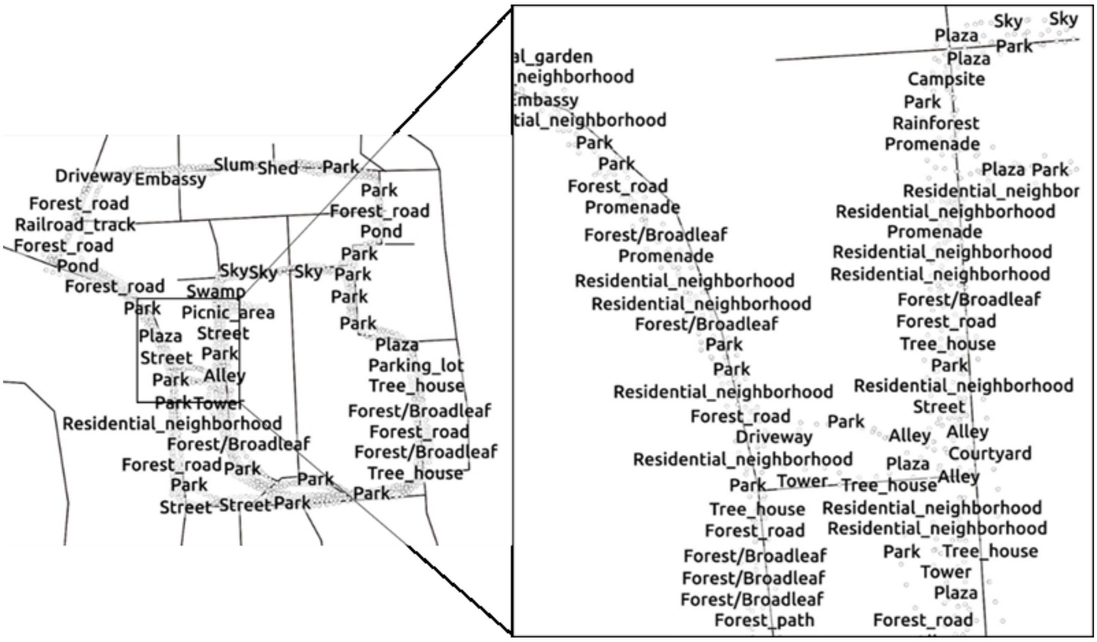


Figure 7. Snippets of maps showing places identified.

4. Data Analysis and Limitations

The differences and similarities in scenic attributes between different data points are explored through clustering algorithms such as principal component analysis (PCA) and t-distributed stochastic neighbour embedding (t-SNE). PCA is a dimensionality reduction technique which removes the redundant dimensions to keep a new set of calculated linear independent dimensions known as principal components. The values of these dimensions can be plotted in 2-D and 3-D for visualisation. Ten parameters of each data point are considered, which included three values of hill numbers, each calculated from segmentation and detection tasks; the number of total elements detected from the detection task; and the percentage of green, naturalness, and the sky view. Most of these parameters are intrinsically correlated with each other, such as the percentage of green being a subset of naturalness. Similarly, Simpson and Shannon indices of hill numbers also show high correlation. PCA was implemented in the Scikit-Learn Library (http://scikit-learn.org/) to identify key influencing parameters using 2093 data points. The parameters were normalised before processing. The scatter plot (Figure 8A) shows the plot between the first and second principal component. Three different colours indicate the timings (morning, noon, and evening) of the captured data points. The plot shows two distinct clusters significantly different in size. In order to identify characteristics of the clusters, the K-means algorithm was implemented to delineate the clusters. K-Means is an unsupervised algorithm which is used to find groups in the datasets. The smaller cluster (cluster no. 2 in Figure 8B) consists of data points for which the object detection task resulted in zero detections, due to dependent parameters such as Hill numbers producing a value of zero. The cluster consists of images collected from all the three timings. The data points in the cluster are plotted against the map for visual assessment of the association between points at different locations. Most numbers of zero-detections occurred in the morning (225), followed by the evening (162). The peculiar case of the least number (94) of zero-detections in the afternoon can be attributed to the increased street activity due to vehicles, people, and bicycles during class hours in the campus. Furthermore, the locations of zero-detections (a, b, c, and d) lie in the inner streets, which are usually deserted during the morning hours.

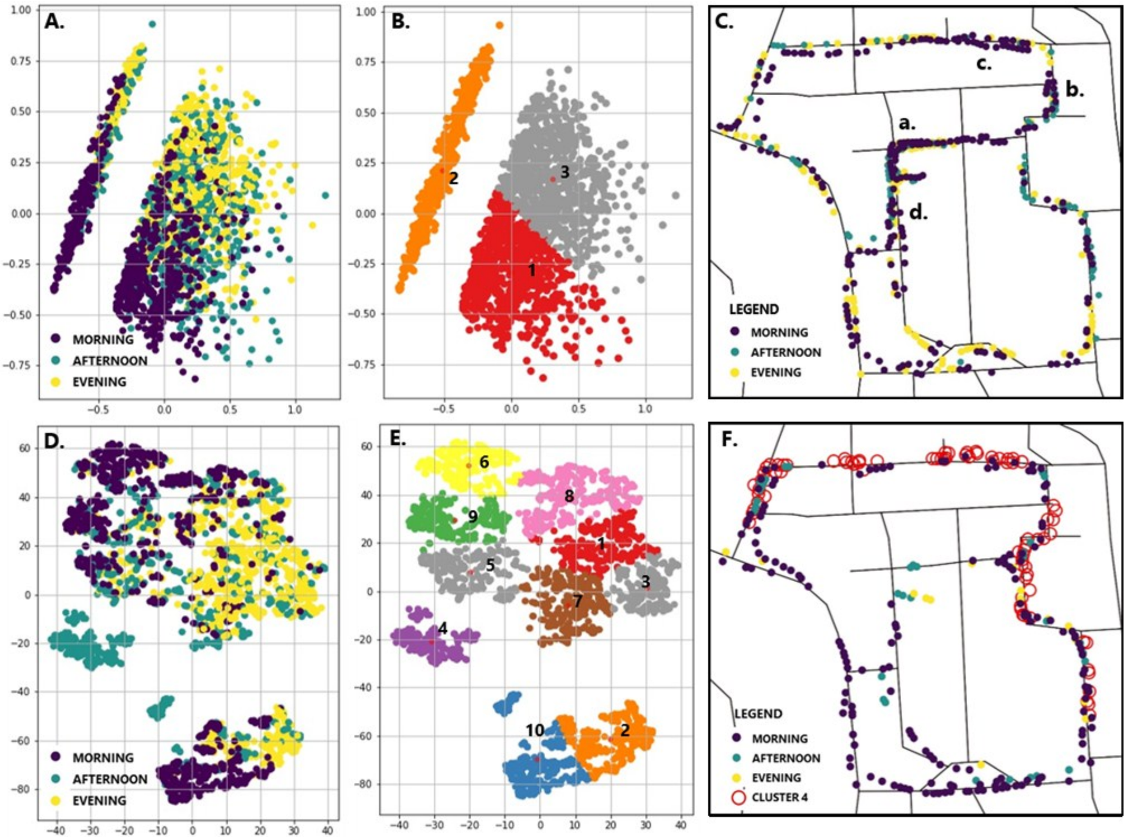


Figure 8. Identifying patterns in the dataset with Principal Component Analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) clustering.

Additionally, t-SNE clustering is implemented with the help of Scikit-Learn to gather more specific patterns and clusters in the dataset. t-SNE is a nonlinear dimensionality reduction technique which computes differences between points and tries to maintain them in low-dimensional space [63]. However, t-SNE does not preserve distances or density information, which makes it challenging to analyse and interpret the appropriateness of the clusters calculated by this algorithm. The clusters (Figure 8D) are not perfectly resolved and discernible. To inquire about a particular group of data points having similar attributes, we performed K-means to create clusters of the t-SNE results. We are aware of the fact that K-means, being a distance-based clustering algorithm, may produce errors while delineating borders of the t-SNE clusters due to loss of distance-specific information while implementing the t-SNE algorithm. Cluster nos. 1, 2, 3, and 7 do consist of a large share of points collected in the evening, while 6, 9, and 10 are the same for morning data. The clusters 10 and 2 contain the points which have the properties of PCA cluster 2. Some of the data points collected in the afternoon produce a visually distinct cluster (no. 4). On plotting the same on the adjoining map, it shows the location of these points. The distinctness of the cluster is due to the similarity between the values of environmental variables. The particular locations of the street have similar characteristics regarding the on-street parking, high canopy cover, and low sky view percentages. Similarly, cluster no. 6 was randomly chosen and plotted in the map. Unlike the cluster no. 4, which shows distinctness in the values at specific locations, the precise characteristics of the cluster are difficult to ascertain; however, all the included points fall in the range consisting of low diversity (Shannon_s in Figure A1) below 3.5 and richness (Richness_s in Figure A1) below 15, but do vary in variables such as the percentage of sky view and total detection of elements (Figure A1). The adjoining map shows that the locations of the points are scattered all over the area, which does not suggest that any particular street has similar data points.

The results in the study are based on the unequal observations taken at different locations, due to which, some of the inferences regarding the diversity and variation differ. However, PCA and t-SNE-generated clusters show that the dataset is not randomly dispersed, where attributes of points similar to each other provide meaningful associations. The data collection is conducted by a small number of participants in a smaller and managed institutional environment. A typical urban area, on the other hand, is more dynamic and diverse regarding the traffic movement and includes a different set of factors which may be interrelated as well as random. Reflecting such details in the experiment is achievable with the large user base and computational power. The lighting conditions, even during daylight, are scarce in certain pockets overshadowed by green canopy cover or tall buildings, due to which, particular objects may not be detected even if they are present. Occlusion is a widely researched phenomenon in computer vision, in which the detection of the object depends upon the visibility of the objects to the camera sensor. The detection of objects is compromised if they are partially hidden behind other objects. The models used in the study are not tested in the field for the occlusion and related technical insufficiencies. Due to this, the outputs may vary regarding the choice of the street for surveying purposes. The web-based mobile application is based on open-source modules and frameworks which are widely researched and utilised both in academia and industries. Such an application can be expanded significantly with the use of technologies and expertise not available to the authors. The individual modules of deep learning methods are a constant subject of research and do possess high chances of the development of newer and efficient models. The deep learning models used in the study may be replaced by different methods and practices over time.

5. Discussion

Urban perception studies aim to understand the relationship between the presence of scenic elements and the perceptual attributes of the environment, such as safety, liveliness, boredom, wealthiness, depressing, and beautiful. Earlier studies have established several constructs, such as the presence of naturalness stimulating satisfaction, or the liveliness present in the street increasing safety; however, none of these have been validated by large-scale perception-based research. The collection of data and the identification of low-level physical features has always been a key determinant in such studies. This study provided an example to conduct rapid data collection and visual data analysis with the state-of-the-art methods and tools. Special focus was given to create a dataset which is spatially and temporally rigorous. Mapping the transient nature of urban spaces is essential in studying the urban environment in a detailed manner.

The analysis of low-level features present in the street view scenes has an immense potential to aid researchers and city managers to focus on bottom-up urban planning approaches. The tasks of asset management and spatial planning regulations are largely based on land use land cover (LULC) maps, which provide information regarding the form and function of the city. This study proposed a methodology to generate maps of low-level physical features as visible from the individual's perspective. The visual access to different environmental attributes such as the greenness, built character, and type of urban places may help in the quantification of physical features which cannot be mapped in LULC maps. Automated map generation and its revision can be achieved with the help of such an application. Apart from data collection, the application can be extended to include real-time user-provided ratings to the places based on the different perceptual attributes. Identification of environmental variables coupled with the user's perception of the space may provide a hint on the socioeconomic conditions of the neighbourhood. It may help to identify the disparity in perceptions which may be due to criminal activities, health, high population density, and ethnic composition.

The quality of life has been studied as the relationship of a person with the environment, where the access to greenery and open spaces plays an important role in determining satisfaction and happiness. Furthermore, the analysis of visual aesthetics such as the identification of dominant colors, landscaping, built façades, and dilapidation of built structures is helpful in studying and creating aesthetically pleasing experiences. In the longer term, such a framework deployed with a large user base may ensure the rapid mapping and auditing of such changes in the environment. It will help in providing data-driven suggestions to the urban management about the places in need of redevelopment and retrofitting.

6. Conclusions and Future Work

We presented a framework aimed towards data collection and representation of the findings. This study is an attempt to quantify environmental attributes through technological means. We provided examples to assess high-level variables such as naturalness, complexity, familiarity, and openness with computer vision techniques. We focused on discussing various environmental variables and visual properties which influences people's preferences and play an important role in an individual's quality of life. Rigorous statistical and spatial analysis of the obtained results is the way forward for future research. This study, although limited to smaller premises, holds the potential to be scaled up at the city level. The web application presented in this study can be redesigned with personalisation and user experience features which can be scaled to involve more users for large-scale studies. City managers can utilize this framework to collect street-level datasets to gather detailed characteristics of the city with the help of a large number of users. The proposed framework can be integrated with the already existing mobile-based city guide applications. Recently, the deep learning frameworks have been deployed in the Android- and iOS-based mobile platforms, which do not require dedicated server support for data processing. Future studies may utilize this native platform support to showcase different features. This study is purely based on identifying visual cues in the surroundings. However, the environment perception is a collective response to different human senses. Along with the visual, auditory and olfactory realms play a significant role in preference towards specific spaces. Future studies may include the collection of these multimodal datasets for the comprehensive evaluation of surroundings. Lack of tools and techniques to understand the rapidly changing urban landscapes has been a major impediment to its planning and management. Results from such studies may discover the unavailability or inadequacy of specific environment variables essential for human health and well-being.

Author Contributions: All authors have equally contributed to the research.

Funding: The authors would like to thank the Ministry of Human Resource Development (MHRD), India and Industrial Research and Consultancy Centre (IRCC), IIT Bombay for funding this study under the grant titled Frontier Areas of Science and Technology (FAST), Centre of Excellence in Urban Science and Engineering (grant number 14MHRD005).

Acknowledgments: The authors are thankful to the surveyors who used the application to collect datasets.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix

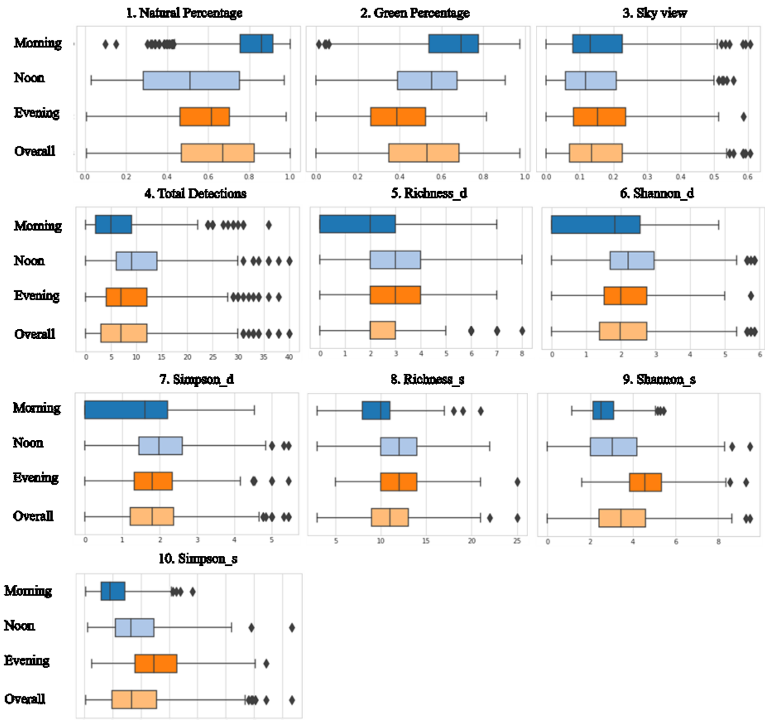


Figure A1. Boxplots showing variation in various environmental variables during the morning, noon, and evening. **Note:** 1: Natural percentage is calculated with the segmentation task by the addition of detected natural elements; 2: Green percentage is the presence of trees and grass in the scene; 3: Sky view is the percentage of visible sky in the segmentation task; 4: The total detection of elements in the scene with the help of the object detection task; 5: The Richness index of the detection task; 6: Shannon index of the detection task; 7: Simpson index of the detection task; 8: Richness index of the segmentation task; 9: Shannon index of the segmentation task; 10: Simpson index of the segmentation task.

References

1. Berman, M.G.; Hout, M.C.; Kardan, O.; Hunter, M.R.; Yourganov, G.; Henderson, J.M.; Hanayik, T.; Karimi, H.; Jonides, J. The perception of naturalness correlates with low-level visual Features of environmental scenes. *PLoS ONE* **2014**, *9*, e114572. [CrossRef] [PubMed]
2. Wang, R.; Zhao, J. Demographic groups' differences in visual preference for vegetated landscapes in urban green space. *Sustain. Cities Soc.* **2017**, *28*, 350–357. [CrossRef]
3. Linder, D.E. Public Places and Spaces. *PsycCRITIQUES* **1990**, *35*. [CrossRef]
4. Herzog, T.R.; Stark, J.L. Typicality and preference for positively and negatively valued environmental settings. *J. Environ. Psychol.* **2004**, *24*, 85–92. [CrossRef]
5. Kaplan, R.; Kaplan, S.K.R.; Kaplan, S. *The Experience of Nature: A Psychological Perspective*; Cambridge University Press: Cambridge, UK, 1989.
6. Kasmar, J. The development of a usable lexicon of environmental descriptors. *Environ. Behav.* **1970**, *2*, 153–169. [CrossRef]
7. Kaplan, S.; Kaplan, R.; Wendt, J.S. Rated preference and complexity for natural and urban visual material *. *Percept. Psychophys.* **1972**, *12*, 354–356. [CrossRef]
8. Gjerde, M. Visual Aesthetic Perception and Judgement of Urban Streetscapes. Available online: http://www.irbnet.de/daten/iconda/CIB18896.pdf (accessed on 26 August 2018).
9. Herzog, T.R.; Kaplan, S.; Kaplan, R. The Prediction of Preference for Familiar Urban Places. *Environ. Behav.* **1976**, *8*, 627–645. [CrossRef]

10. Loewen, L.J.; Steel, G.D.; Suedfeld, P. Perceived safety from crime in the urban environment. *J. Environ. Psychol.* **1993**, *13*, 323–331. [CrossRef]
11. Nasar, J.L.; Julian, D.; Buchman, S.; Humphreys, D.; Mrohaly, M. The emotional quality of scenes and observation points: A look at prospect and refuge. *Landsc. Plan.* **1983**, *10*, 355–361. [CrossRef]
12. Laumann, K.; Garling, T.; Stormark, K.M. Rating Scale Measures of Restorative Components of Environments. *J. Environ. Psychol.* **2001**, *21*, 31–44. [CrossRef]
13. Shen, Q.; Zeng, W.; Ye, Y.; Arisona, S.M.; Schubiger, S.; Burkhard, R.; Qu, H. StreetVizor: Visual Exploration of Human-Scale Urban Forms Based on Street Views. *IEEE Trans. Vis. Comput. Gr.* **2018**, *24*, 1004–1013. [CrossRef] [PubMed]
14. De Nadai, M.; Vieriu, R.L.; Zen, G.; Dragicevic, S.; Naik, N.; Caraviello, M.; Hidalgo, C.A.; Sebe, N.; Lepri, B. Are Safer Looking Neighborhoods More Lively? A Multimodal Investigation into Urban Life. In Proceedings of the 2016 ACM on Multimedia Conference, Amsterdam, The Netherlands, 15–19 October 2016; pp. 1127–1135.
15. Doersch, C.; Singh, S.; Gupta, A.; Sivic, J.; Efros, A.A. What makes Paris look like Paris? *ACM Trans. Graph.* **2012**, *31*, 1–9. [CrossRef]
16. Aiello, L.M.; Schifanella, R.; Quercia, D.; Aletta, F. Chatty maps: Constructing sound maps of urban areas from social media data. *R. Soc. Open Sci.* **2016**, *3*, 150690. [CrossRef] [PubMed]
17. Quercia, D.; Schifanella, R.; Aiello, L.M.; McLean, K. Smelly Maps: The Digital Life of Urban Smellscapes. *arXiv*, **2015**, arXiv:1505.06851.
18. Hyam, R. Automated Image Sampling and Classification Can Be Used to Explore Perceived Naturalness of Urban Spaces. *PLoS ONE* **2017**, *12*, e0169357. [CrossRef] [PubMed]
19. Ertiö, T.P. Participatory Apps for Urban Planning—Space for Improvement. *Plan. Pract. Res.* **2015**, *30*, 303–321. [CrossRef]
20. Winkel, G.; Malek, R.; Thiel, P. A Study of Human Response to Selected Roadside Environments. In Proceedings of the 1st EDRA Conference. Available online: https://trove.nla.gov.au/work/19383376?q&versionId=22775366 (accessed on 20 January 2018).
21. Nasar, J.L. New Developments in Aesthetics for Urban Design. In *Toward the Integration of Theory, Methods, Research, and Utilization*; Moore, G.T., Marans, R.W., Eds.; Springer: Boston, MA, USA, 1997; pp. 149–193.
22. Nasar, J.L. Visual Preferences in Urban Street Scenes. *J. Cross. Cult. Psychol.* **1984**, *15*, 79–93. [CrossRef]
23. Suleiman, A.; Chen, Y.H.; Emer, J.; Sze, V. Towards closing the energy gap between HOG and CNN features for embedded vision (Invited paper). In Proceedings of the 2017 IEEE International Symposium on Circuits and Systems (ISCAS), Baltimore, MD, USA, 28–31 May 2017.
24. Seresinhe, C.I.; Preis, T.; Moat, H.S. Using deep learning to quantify the beauty of outdoor places. *R. Soc. Open Sci.* **2017**, *4*, 170170. [CrossRef] [PubMed]
25. Dubey, A.; Naik, N.; Parikh, D.; Raskar, R.; Hidalgo, C.A. Deep Learning the City: Quantifying Urban Perception at a Global Scale. In *Lecture Notes in Computer Science*; Springer: Cham, Switzerland, 2016; pp. 196–212.
26. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Available online: http://download.tensorflow.org/paper/whitepaper2015.pdf (accessed on 26 August 2018).
27. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in PyTorch. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–4.
28. Keras. Available online: https://keras.io/getting-started/faq/#how-should-i-cite-keras (accessed on 26 August 2018).
29. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional Architecture for Fast Feature Embedding. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014.
30. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. In *Lecture Notes Computer Science*; Springer: Cham, Switzerland, 2014; Volume 8693, pp. 740–755.
31. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.

32. Krasin, I.; Duerig, T.; Alldrin, N.; Veit, A.; Abu-El-Haija, S.; Belongie, S.; Cai, D.; Feng, Z.; Ferrari, V.; Gomes, V.; et al. OpenImages: A Public Dataset for Large-Scale Multi-Label and Multi-Class Image Classification. Available online: https//storage.googleapis.com/openimages/web/index.html (accessed on 20 January 2018).
33. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *Multimed. Tools Appl.* **2015**, *77*, 10437–10453.
34. Rethinking the Inception Architecture for Computer Vision. Available online: https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Szegedy_Rethinking_the_Inception_CVPR_2016_paper.html (accessed on 26 August 2018).
35. Liu, C.; Zoph, B.; Shlens, J.; Hua, W.; Li, L.J.; Fei-Fei, L.; Yuille, A.; Huang, J.; Murphy, K. Progressive Neural Architecture Search. *arXiv* **2017**, arXiv:1712.00559.
36. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
37. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
38. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
39. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef] [PubMed]
40. Pyramid Scene Parsing Network. Available online: https://arxiv.org/abs/1612.01105 (accessed on 26 August 2018).
41. Everingham, M.; Eslami, S.M.A.; van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes Challenge: A Retrospective. *Int. J. Comput. Vis.* **2014**, *111*, 98–136. [CrossRef]
42. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The Cityscapes Dataset for Semantic Urban Scene Understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Basel, Switzerland, 12–15 September 2016; pp. 3213–3223.
43. Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; Torralba, A. Semantic Understanding of Scenes through the ADE20K Dataset. *arXiv* **2016**, arXiv:1608.05442.
44. Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; Torralsba, A. Places: A 10 million Image Database for Scene Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*. [CrossRef] [PubMed]
45. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
46. Kardan, O.; Demiralp, E.; Hout, M.C.; Hunter, M.R.; Karimi, H.; Hanayik, T.; Yourganov, G.; Jonides, J.; Berman, M.G. Is the preference of natural versus man-made scenes driven by bottom–up processing of the visual features of nature? *Front. Psychol.* **2015**, *6*, 1–13. [CrossRef] [PubMed]
47. Purcell, T.; Peron, E.; Berto, R. Why do preferences differ between scene types? *Environ. Behav.* **2001**, *33*, 93–106. [CrossRef]
48. Ulrich, R.S. *Behavior and the Natural Environment*; Springer: Boston, MA, USA, 1983; Volume 6.
49. Kaplan, R. The Nature of the View from Home: Psychological Benefits. *Environ. Behav.* **2001**, *33*, 507–542. [CrossRef]
50. Porzi, L.; Bulò, S.R.; Lepri, B.; Ricci, E. Predicting and Understanding Urban Perception with Convolutional Neural Networks. In Proceedings of the 23rd ACM International Conference on Multimedia, Brisbane, Australia, 26–30 October 2015; pp. 139–148.
51. Hur, M.; Nasar, J.L.; Chun, B. Neighborhood satisfaction, physical and perceived naturalness and openness. *J. Environ. Psychol.* **2010**, *30*, 52–59. [CrossRef]
52. Herzog, T.R. A cognitive Analysis of Preference for Urban Nature. *J. Environ. Psychol.* **1989**, *9*, 27–43. [CrossRef]
53. Kaplan, S. *Perception and Landscape: Conceptions and Misconceptions*; United States Department of Agriculture: Washington, DC, USA, 1979; pp. 241–248.
54. Oliva, A.; Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vis.* **2001**, *42*, 145–175. [CrossRef]

55. Jost, L. Entropy and diversity. *Oikos* **2006**, *113*, 363–375. [CrossRef]
56. Yue, Y.; Zhuang, Y.; Yeh, A.G.O.; Xie, J.-Y.; Ma, C.-L.; Li, Q.-Q. Measurements of POI-based mixed use and their relationships with neighbourhood vibrancy. *Int. J. Geogr. Inf. Sci.* **2016**, *31*, 1–18. [CrossRef]
57. Shrivastava, P.; Bhoyar, K.K.; Zadgaonkar, A.S. Bridging the semantic gap with human perception based features for scene categorization. *Int. J. Intell. Comput. Cybern.* **2017**, *10*, 387–406. [CrossRef]
58. Xu, Y.; Ren, C.; Cai, M.; Edward, N.Y.Y.; Wu, T. Classification of Local Climate Zones Using ASTER and Landsat Data for High-Density Cities. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 3397–3405. [CrossRef]
59. Appleton, J. *The Experience of Landscape*; Wiley: Hoboken, NJ, USA, 1996.
60. Dosen, A.S.; Ostwald, M.J. Evidence for prospect-refuge theory: A meta-analysis of the findings of environmental preference research. *City Territ. Archit.* **2016**, *3*, 4. [CrossRef]
61. Herzog, T.R.; Kaplan, S.; Kaplan, R. The prediction of preference for unfamiliar urban places. *Popul. Environ.* **1982**, *5*, 43–59. [CrossRef]
62. Lynch, K. *The Image of the City*; MIT Press: Cambridge, MA, USA, 1960; Volume 11.
63. Van der Maaten, L.J.P.; Hinton, G.E. Visualizing high-dimensional data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.