Title	Automatic extraction and quantitative analysis of building facade information at large scale using street-level images and deep learning
Author(s)	Zhang, Jiaxin
Citation	大阪大学, 2022, 博士論文
Version Type	VoR
URL	https://doi.org/10.18910/89626
rights	
Note	

Doctoral Dissertation

Automatic extraction and quantitative analysis of building facade information at large scale using street-level images and deep learning

JIAXIN ZHANG

June 2022

Graduate School of Engineering,

Osaka University

Environmental Design and Information Technology Subarea

Sustainable Environmental Design Area

Division of Sustainable Energy and Environmental Engineering

博士論文

大規模スケールに適用可能なストリートビュー画像と深層学習を用いた建物ファサード情報の自動抽出と定量分析

学籍番号 28H20807

名 前 ZHANG JIAXIN

2022 年 6 月

大阪大学 大学院工学研究科
環境エネルギー工学専攻
共生環境デザイン学講座
環境設計情報学領域

Abstract

The digital management of existing building data plays a key role in efficiently allocating resources and developing urban renewal strategies. Urban development issues such as urban building energy modeling, urban building retrofitting, urban heat island, and urban vitality are inextricably linked to the fine management of digital urban building data. With the transition of urban spatial development patterns to stock optimization, a large number of urban buildings constructed during the high-speed incremental period have generated a considerable amount of building data that needs to be managed urgently. Building facade data is an important part of urban building data. Building facades need to meet building performance requirements and pass on history and culture. Large-scale collection and management of building facade data, including building geometry data, facade colors, building functions, facade semantics, facade materials, etc., is crucial in the maintenance of the life cycle of the stock buildings. However, constructing a city-scale database of building facades is a difficult task. In particular, the automation of the building measurement process has long been a challenge that has plagued both academia and industry. Field measurements by professional surveyors are still the dominant method in the industry. This approach works well for neighborhood-scale projects but is hard to adapt to city-scale.

This study attempts to develop a three-step framework to automate the measurement of building facade data at a large scale to construct an urban facade database. The collected building facade data includes semantic segmentation, dominant colors, building functions, and window-wall semantic information. Street-level images and state-of-the-art deep learning methods are used to extract facade information. Firstly, this study develops an unwanted object elimination system that can obtain the complete building facade to decrease the loss of information caused by obstruction. Secondly, a facade instance segmentation method using the synthetic dataset from a city digital twin (CDT) is proposed, which has two benefits: It solves the segmentation

i

problem of connected buildings. Another one is the automatically generated synthetic dataset dramatically reduces the cost of data annotation. Third, an integrated multi-task facade data extraction method is proposed. As a result, building facade data, including facade dominant color, building functional classification, and window-wall semantics, will be automatically counted. Based on the above research framework, the study proposes several publicly available facade datasets for facade instance segmentation, building function classification, and window-wall semantic segmentation.

The proposed frameworks contain various building facade data types that are not recorded in existing urban geo-databases (e.g. OpenStreetMap). The experimental results are verified in several cities and show that, first, the approach can overcome the interference of street obstructions to the facade data collection. Then, the proposed CDT synthetic dataset can be effectively used for facade instance segmentation of real images, revealing the potential of the proposed synthetic dataset to replace real ones. Finally, the integrated multi-task facade parsing approach has satisfactory accuracy in facade dominant color measurement, building function classification, and semantic segmentation of walls and windows. Overall, the digital management of building assets can facilitate the efficient allocation of public resources and urban development decision-making. In the future, the collected digital facade information at the city scale will be stored in a database that allows the public, private, and research sectors to formulate urban development strategies.

Keywords: Building facades; Deep learning; Street-level images; Image inpainting; Synthetic data; Facade parsing; Quantitative analysis

Preface

This dissertation is the original work by Jiaxin Zhang under the supervision of Assoc. Prof. Tomohiro Fukuda. Three journal articles and two international conference proceedings related to this dissertation have been submitted or published. They are listed below.

Journal articles:

1. Zhang, J., Fukuda, T., & Yabuki, N. (2021). Development of a City-Scale Approach for Facade Color Measurement with Building Functional Classification Using Deep Learning and Street View Images. *ISPRS International Journal of Geo-Information*, 10(8), 551. DOI: 10.3390/ijgi10080551

2. Zhang, J., Fukuda, T., & Yabuki, N. (2021). Automatic Object Removal With Obstructed Facades Completion Using Semantic Segmentation and Generative Adversarial Inpainting. *IEEE Access*, 9, 117486-117495. DOI: 10.1109/ACCESS.2021.3106124

3. Zhang, J., Fukuda, T., & Yabuki, N. Automatic generation of synthetic datasets from a city digital twin for use in the instance segmentation of building facades. *Journal of computational design and engineering*, Under review.

International conference proceedings:

1. Zhang, J., Fukuda, T., & Yabuki, N. (2020). A Large-Scale Measurement and Quantitative Analysis Method of Facade Color in the Urban Street Using Deep Learning. *The International Conference on Computational Design and Robotic Fabrication* (pp. 93-102). Springer, Singapore. DOI: 10.1007/978-981-33- 4400-6_9

2. Zhang, J., Fukuda, T., & Yabuki, N. (2021). Image-based Cityscape Visualization with Automatic Object Removal and Facade Inpainting Using Semantic Segmentation and Generative Adversarial Networks. *SimAUD 2021 (Symposium on Simulation for Architecture and Urban Design)*

Acknowledgment

Table of contents

x

List of figures

xiii

List of tables

Chapter 1. Introduction

1.1   Research background and problem statements

With the increase in urbanization, the developable amount of urban construction land is gradually decreasing (Cao et al., 2020; M.-C. Chen et al., 2006). Therefore, the mode of urban development is changing from high-speed expansion to optimization. High-quality development is becoming the core goal of urban governance (Meijer & Bolívar, 2016). However, with the gradual aging of old urban areas, the deficiencies in public facilities, resource integration capacity, and planning and management will restrict urban development. The large number of urban buildings that have been built can hardly meet people's new demand for a high quality of life. In particular, the increase in the volume of building data brought about by various issues such as urban building renewal (Zheng et al., 2014), urban thermal environment (Ferrando et al., 2020), and urban vitality (Mouratidis & Poortinga, 2020) has posed new challenges to building data management. Therefore, there is an urgent need to digitally refine the management of existing urban data to help rationalize the allocation of resources and make urban development strategies.

Building facade data is an important part of urban architectural data because the facades should not only meet the needs of visual quality and architectural performance but also embody a city's history, express its culture and preserve its urban fabric (Degaev & Barkhi, 2019). Therefore, it is especially important to collect and manage building facade data, including building facade color, date of construction, facade material, and facade element size, on a city-wide scale and in a detailed manner. However, conducting building renewal on a large scale is a major challenge. This is especially true when it comes to automating the building measurement process. Field measurement by professional surveyors is currently the primary method in the industry.

Manual measurement is adequate for small-scale projects. Starting work on city-scale projects can, however, be difficult (Zhong et al., 2021).

The collection of building facade data is critical for the energy efficiency retrofit of buildings. By fully understanding the impact of upgrades to the individual constructions of different buildings on their thermal performance improvement, the benefits of retrofitting can be clearly assessed and understood, creating a win-win scenario for all stakeholders involved. Information from the building facade, including geometric data, building function classification, and geographic information, allows the construction of urban building energy models (UBEMs) (Ferrando & Causone, 2020) and the development of data-driven building energy retrofit strategies (Hu, 2020). When deploying retrofit programs for individual buildings, researchers collect individual building data such as building geometry (Kheiri, 2018), thermal characteristics (e.g., building materials, glazing ratios, window-door types, thermal bridging issues) (Boodi et al., 2022), and failure information (broken windows and facade defects) (Marchand et al., 2018) in order to effectively model building energy and quantify the benefits of retrofits.

Building facade data collection is also essential for urban development and renewal (M. Dai et al., 2021). The facade can convey the city's historical information. It has the role of reflecting the city's characteristics and showing the city's culture, customs, and urban landscape. Digital reconstruction and database construction of building facades with conservation value have become essential (T. Deng et al., 2021). Recently, researchers have proposed considering buildings as material databases and have focused on new buildings with building information modeling (BIM) to preserve high-value components for future use. However, these studies have done little to address the fundamental barriers to reusing materials and components in existing buildings, which lack digital records (Sultana & Storch, 2021). Urban researchers encounter difficulties when they tackle developing urban or architectural landscape

renewal programs without quantitative analysis of old building data as support (J. Wang et al., 2021).

As cities grow and spread in the population, the application of geographic information system (GIS) in urban planning provides a better understanding of various issues of a city, such as ecologies, transportation, housing, crime, aging, and other issues (X. Liu et al., 2017). By processing geospatial data from satellite imagery, aerial photography, and street view images, users can gain a detailed understanding of the land and infrastructure. The GIS is important because it can bring together the vast amount of city information necessary to balance competing priorities and solve complex building problems, such as optimizing the layout of new buildings or the digital management of building information (Zhu et al., 2018). Existing urban geo-databases with building data, such as OpenStreetMap (OSM) (OpenStreetMap, 2021), contain simple building data such as floor area, height, and the number of floors, and lack the integration of detailed indicators about building facades such as semantic information of facade components and facade dominant color. In recent years, laser scanning technology for generating BIM models has seen increased usage in identifying component attributes in existing buildings (Istenič et al., 2020; B. Wang et al., 2021). Applications for laser scanning technology in urban data collection continue to emerge (Y. Wang et al., 2019). However, the obstacles to laser scanning include the high cost of complex equipment, time-consuming and laborious handling of complex data, inconvenient field operation, and large file type storage. It is, therefore, applicable only to a small range of projects and is difficult to be used universally (Szcześniak et al., 2022).

The advantages of integrating computer vision techniques and a large range of freely available street images to extract building facade data are, in contrast, more obvious (Campbell et al., 2019). They rely on a higher-order knowledge model of facade topology and lower-level elements (roofs, windows, balconies, doors, walls, etc.) that make up the building (W. Li et al., 2020). Researchers parse building facade images

by partitioning them into semantics requirements corresponding to the elevation structures composed of lower-level elements (Kong & Fan, 2021; H. Liu et al., 2020; Ma et al., 2020). The cost of this approach is low, and the accuracy requirement is appropriate for building facade data acquisition. Therefore, supervised learning-based building facade parsing is considered to become one of the most powerful techniques for building information modeling at a city scale in the future (Gadde et al., 2016; Riemenschneider et al., 2012).

However, supervised learning-based segmentation methods require data annotation to be a laborious manual task (X. Xie et al., 2020). And the quality and quantity of the dataset largely determine the execution of the segmentation model. On the one hand, researchers want to use a large amount of accurately annotated data, and on the other hand, they often struggle with the expensive costs associated with all this data (Schumacher et al., 2019, p. 0). In addition, obstructions in the street scene are detrimental factors affecting the method of obtaining facade data based on street view images. Obstructions in front of the facade severely reduce the integrity of the data. Whether it is street 2D data acquisition or 3D model reconstruction, removing unwanted objects has been a challenge that has plagued researchers.

In summary, the large-scale automatic construction of the urban facade database can be beneficial to constructing UBEMs and developing building renewal plans. The critical issues of this study are the large-scale acquisition of facade images, the automatic extraction of facade element information using deep learning and street-level images, and the integration and utilization of facade data.

1.2   Research objective

This study attempts to develop a method that can automatically measure building facade data on a large scale to comprise an urban facade information database. The building facade data includes monolithic instance segmentation of the facade, dominant color, building function, and window-wall semantics. In this study, street-level images

will be used as the data source, and state-of-the-art deep learning will be used to extract facade information. By reviewing studies related to the large-scale collection and quantitative analysis of building facade data, the practical application value of the automated approach to building retrofitting and urban landscape renewal is explored. Based on these, the objectives are as follows.

(1) To develop a method for automatically obtaining complete building facades in street-level images, often obscured by obstacles;

(2) To develop a system for automatically generating synthetic datasets for training deep learning-based facade instance segmentation models, which would alleviate the cost of manually labeling datasets;

(3) To develop a comprehensive system of obtaining multiple data types of building facades, including facade instance segmentation, facade dominant color, building function, and window-wall semantics, in order to provide cost-effective tools for developing urban facade databases.

1.3   Research significance

The mining, analyzing, and storage of building facade data play a crucial role in the digital management of urban buildings, which can be used as data support for built environment renewal and development. This study takes the large-scale collection and quantitative analysis of building facade data as an entry point to assist city managers and researchers in practical applications and theoretical research. In practical applications, large-scale automatic extraction of building facade data can save researchers' costs and labor compared with field measurements. In terms of theoretical research, the built facade database can help urban researchers analyze urban issues, including the prediction of urban building energy consumption and the development of urban landscape renewal strategies.

1.4   Research scopes

The existing urban geo-database (such as OSM) can capture street networks, building height, and building plan outlines. This study focuses on the elements of building facades that are not available in the existing open database, such as the instance segmentation of building facades, facade dominant colors, and facade window-wall semantics, and these types of data will contribute to UBEMs and urban renewal.

The data sources in this study are street-level images, including Baidu Street View (Baidu Street View service, 2022) and Google Street View images (Google Street View service, 2022). This study is used for facade parsing methods based on deep learning for image classification, semantic segmentation, and instance segmentation. The color measurement used by Euclidean distance-based building standard color calculation.

1.5   Research framework

To implement a large-scale automated extraction system for building facades, the research framework of the system (Figure 1.1) is divided into five steps: (1) Inputting the street coordinates into the system and getting the shapefile of the road centerline from OSM. (2) Calculating the requested deflection angle for Street View Service API. The street view images for shooting the building facade vertically are obtained. (3) Automatically eliminating unwanted objects in front of the building, such as trees, cars, and people. Individual building facades are obtained using instance segmentation. (4) Performing orthogonal transformation and resizing the images based on camera-to-building distance, camera-to-edge center distance, camera zoom, and pitch angle. Necessary information for each individual building is calculated, including facade color, building function classification, and window-wall segmentation. (5) Measuring the elements of the building facade at a large scale and generating an urban facade output CSV file. An urban facade database was created to provide data-driven decision support for urban designers and stakeholders.

STEP 1

Road networks from OpenStreetMap

Sampling points of road networks

- Get the shapefile of the road centerline from OpenStreetMap.
- The sampling points are taken once at equal intervals on the road network.

STEP 2

Two adjacent road coordinates

Mock of street view

Building #1   Building #2

- Calculate the requested deflection angle for Street View Service API.
- Grab the street view images for shooting the building façade vertically.

STEP 3

Unwanted objects removal

Building #2
Building #1

Image color calibration

Facades instance segmentation

Building #2
Building #1

- Remove unwanted objects obstructing the front of the building, such as trees, cars, people, etc.
- Image color calibration.
- Individual building facades are obtained using instance segmentation.

STEP 4

Individual facades analysis

Building #1

Building #2

- Perform orthogonal transformation and resize the images based on camera-to-building distance, camera-to-edge center distance, camera zoom and pitch angle.
- Grab the street view images by shooting the building façade vertically.
- Number each building and calculate information for each individual building, including façade color, building function classification, window segmentation, and door segmentation.

STEP 5

Building façade information
Individual Building coordinates
Façade color information
Building function classification
Wall and windows semantic segmentation

Urban Façade Database

Data-driven decision making for urban regeneration

- Parsing building facades with instance segmentation.
- Mapping the city-scale façade color.
- Generating a large-scale façade output csv file.
- Provide data-driven decision support to the urban designers and stakeholders.

Input
csv file

Automatic building façade data collection
Street View Service API

Image processing
Unwanted objects removal, color calibration, facades segmentation

Data analysis
Façade color measurement, building function classification, windows and wall segmentation

Output

Figure 1.1. Research framework.

1.6   Overview of the dissertation

Figure 1.2 shows the overview of this dissertation, organized as follows.

Figure 1.2. The overview of this dissertation

The dissertation will be divided into six chapters.

Chapter 1 Introductions:

This chapter introduces the research background, problem statement, research objective, research significance, research scopes, research framework, and the overview of the dissertation.

Chapter 2 Literature review:

This chapter reviews the issues and recent research pertinent to this study. It is divided into five sections, the first of which introduces the brief problem of orientation on the achievements and shortcomings of the facade data extraction using street-level images and deep learning. The second section presents the existing methods for automatic object removal with obstructed facades completion. The third section describes the strengths and limitations of using synthetic datasets for training the instance segmentation of building facades. The fourth section reviews the applications of image-based building facade data extraction in city information modeling and building retrofitting. The fifth section summarizes the gaps in established research and outlines the overall objectives of this study.

Chapter 3 Automatic object removal with obstructed facades completion:

This chapter addresses object removal and facade inpainting. An image-based cityscape removal approach is proposed by detecting multiple classes, including pedestrians, cyclists, vegetation, and cars, as well as using generative adversarial networks (GANs) to fill in the detected areas by background textures from streetscape images.

Chapter 4 Synthetic datasets from a city digital twin for use in the instance segmentation of building facades:

This chapter develops a novel framework that can automatically produce synthetic datasets from a city digital twin (CDT). An auto-generation system for synthetic street views was built by rendering a city's digital assets into a game engine, while the system auto-generated the annotations for building facades. The hybrid dataset, along with various subsets containing different proportions of synthetic and real data, were used to train deep learning models for facade instance segmentation. Two types of synthetic data (CDT-based and virtually-based) were compared, and the results showed that the CDT synthetic data were more effective in boosting deep learning training with real-world images compared with the virtual synthetic data (no real-world counterparts). By

swapping a certain portion of the real data with the proposed CDT synthetic images, the performance could almost match what is achievable when using the real-world training set.

Chapter 5 The large-scale approach for extracting data on multiple elements of building facades:

This chapter explores an approach utilizing state-of-the-art deep learning techniques and street-level imagery to measure multiple facade elements at a large scale, including dominant color measurement, building function classification, and window-wall semantic segmentation. A street of length 500m in Osaka, Japan, is used to construct a database as an example. The results demonstrate the transferability and effectiveness of the scheme.

Chapter 6 Conclusions:

This chapter offers the study's conclusions, contributions, and limitations and presents plans for future work.

Chapter 2. Literature review: Large-scale building facades data extraction using street-level images

2.1 Background

With the increasing need for 3D building models in urban planning, urban modeling platforms, autonomous driving and game simulation, facade parsing, especially the parsing of detailed level building models developed by individual buildings in City Geography Markup Language (CityGML), has become very important in urban reconstruction. Building facade parsing aims to semantically classify the fine-grained categories of each pixel in a building facade image, where the semantic categories may include fine-grained facade components, the dominant color, and the function. Facade parsing has been receiving continuous attention from the academic community in recent years. It is much easier to search for buildings based on grammar as opposed to Red-Green-Blue (RGB) images when building images are parsed into grammar. Parsing building images into grammar can significantly reduce the storage space required for building images. Apart from that, dividing the building parts by semantics can greatly enrich the building and city information model. Regarding the 3D reconstruction of buildings, the grammar generated by facade parsing can also reduce manual labor costs. Existing facade parsing techniques typically rely on grammar rules or computer vision techniques. These methods generally improve facade parsing results by pre-processing techniques such as image correction or by embedding a priori knowledge for the algorithm.

Although previous facade parsing methods and datasets have achieved significant results (Gadde et al., 2016; Kong & Fan, 2021; H. Liu et al., 2020), the current

extraction of building facade information is based on street-level images still faces three main challenges in practical applications. (1) The existing datasets are most orthographic projections of facades (Gadde et al., 2016), and previous studies rarely considered the occlusion in front of buildings, which does not perform well for parsing building facades of complex scenes. (2) When buildings are connected or visually overlap, it is difficult to extract single buildings using semantic segmentation (M. Dai et al., 2021), which is why instance segmentation must be used. (3) The traditional facade parsing algorithm relies too much on the regular building facade structure based on a priori knowledge (Martinović et al., 2012), which is not ideal for building facades with asymmetric structures. Moreover, the application scenarios of building facade parsing are still in the early stages of exploration.

The next section introduces the literature review along the following lines. Section 2.2 presents a review of previous methods for the elimination of unwanted objects in front of building facades. Section 2.3 describes the instance segmentation methods of building facades and the feasibility of creating synthetic datasets in a cost-effective manner. Section 2.4 investigates the current practical applications of building facade parsing in 3D city modeling and building retrofitting. Section 2.5 summarizes the research gaps and research objectives.

2.2 Automatic objects removal with obstructed facades completion

In urban environments, extensive foreground occlusion exists on building facades. Analyzing the building facade without addressing the occlusion issue will result in missing a significant amount of facade information, leading to biases and a decline in the credibility of the built database.

2.2.1   *Object segmentation and removal*

Unwanted regions are detected, and the regions of interest (ROI) are eliminated and filled with surrounding textures (Y. Sun et al., 2018). The recent emergence of deep learning-based object segmentation shows the power of ROI segmentation. Objects can be detected using the convolutional neural network (CNN) semantic segmentation model and segmented according to their contours. However, when CNN-based object segmentation is applied to determine the ROI, the ROI can be a mask covering the target object or an outline of the target object (Z. Cai & Vasconcelos, 2019; Kido et al., 2021).

The ROI needs to be filled after semantic segmentation, for which there are two main approaches: observation and inpainting. Observation requires a pre-taken image of the background scene (Mori et al., 2017), which can be used as a reference to replace the foreground obstacle directly. For example, using the observation method to remove objects from the front of a building requires receiving complete information about the obscured facade, then replacing unwanted objects with parts of a known building facade. Another technique is inpainting (Criminisi et al., 2004), which uses the texture and patch information of the source image to fill the detected area. This technique does not require prior knowledge of the information behind the occlusion and uses the knowledge of the database or the texture around unwanted objects to fill the ROI. However, obtaining background images in projects where obstacles cannot be moved is challenging. Therefore, inpainting without pre-processing is more suitable for removing objects from street scenes than observation.

2.2.2   *Generative adversarial inpainting*

Existing image inpainting techniques generally fall into three categories (Elharrouss et al., 2020). (1) Inpainting by replication: These techniques attempt to explicitly borrow content or texture from the surrounding environment in order to fill in the gaps. A context copy method is an example of unsupervised learning in which

surrounding image information is used to predict the loss of contents (Nathan Mundhenk et al., 2018). However, image replication typically fails when dealing with intricate scenes. (2) Inpainting by modeling: These methods use extensive external databases to generate data-driven replacements for missing pixels. They attempt to learn to model the distribution of the training images and assume that regions surrounded by backgrounds with similar characteristics may contain similar content (Pathak et al., 2016). These methods can effectively find sample images with sufficient visual similarity to the query, but they easily fail when there are no similar examples in the database. (3) Combining the two: the third class of approaches attempts to combine the previous two in order to overcome the limitations of replication methods or modeling methods, such as generative adversarial network (GAN) methods (Yi et al., 2020; Yu et al., 2019). Not only do these methods learn to build image distributions in a data-driven manner, but they are also designed to explicitly borrow patches or features from background regions (Yi et al., 2020). However, when the training dataset and the content of the processed images do not match, the generated image quality is not satisfactory. Image inpainting works better when the dataset is customized rather than when a generic dataset is used for a specific task.

2.3 Synthetic datasets from a city digital twin for use in the instance segmentation of building facades

The instance segmentation of building facades is one of the focuses of this study. This section reviews previous studies on developing instance segmentation for building facades, using synthetic data for deep learning, and utilizing city digital twins to create synthetic datasets, and summarizes the research gap and goals.

2.3.1 Instance segmentation of building facades

Effectively performing large-scale collection and integration of building facade data in cities has been a long-standing challenge for industry and academia (Martinez

& Choi, 2017; Y. Wang et al., 2018). Parsing building facades into procedural grammars and extracting facade information using semantic segmentation plays a significant role in development tasks involving 3D buildings (M. Dai et al., 2021; Femiani et al., 2018; Rahmani & Mayer, 2018). Deep learning-based semantic facade parsing methods have yielded promising results when applied to open-source facade datasets (H. Liu et al., 2020; Ma et al., 2020). Additionally, studies have enhanced the performance of deep convolutional neural network (DCNN) models using 3D models to automatically synthesize the semantically annotated datasets of building facades. However, studies that use semantic segmentation treat all buildings as one category and do not differentiate between distinct buildings. The semantic reserve is limited in its ability to perform individual segmentation of regions (Carvalho et al., 2020), especially when several buildings are visually superimposed or in contact. Instance segmentation, as a new paradigm and the evolution of semantic segmentation, therefore, allows for a unique understanding of each item in the same class, which is necessary for precisely extracting information from building facades.

Currently, many powerful instance segmentation algorithms are emerging, such as Fast R-CNN (Girshick, 2015), Mask R-CNN (He et al., 2017), YOLACT (Bolya et al., 2019), and BlendMask (H. Chen et al., 2020). Mask R-CNN is a typical technique based on the network architecture of detection followed by segmentation that is relatively easy to train with better generalization and higher segmentation accuracy. For instance, Toda et al. (2020) used synthetic datasets to train Mask R-CNN to characterize the seed morphology of various cultivars. Carvalho et al. (2020) applied Mask R-CNN with real open-source datasets to perform instance segmentation of rural facilities for agricultural management. These studies used real or synthetic data training sets to perform object instance segmentation, but they were seldom applied to building facades. Moreover, previous studies have rarely evaluated the changes in model performance after CDT synthetic data has been added to a real dataset.

2.3.2 *Synthetic data for facade segmentation*

Many studies have shown that the performance of DCNN-based instance segmentation is affected by the network architecture and the amount of data available for training, with the latter having a greater impact on improving the accuracy of the segmentation results (X.-W. Chen & Lin, 2014). However, the acquisition and annotation of the original datasets are time-consuming and laborious, often representing a large percentage of the project budget (Sorokin & Forsyth, 2008). Consequently, several attempts have been made to reduce the reliance on data annotation, such as by using active learning (concentrating only on annotated data with high information) (Settles, 2009), semi-supervised learning (using only a little annotated data) (van Engelen & Hoos, 2020), unsupervised learning (no annotations required) (Locatello et al., 2019), and reinforcement learning (no annotations required) (Botvinick et al., 2019). However, these methods are still working to achieve performance comparable to that of supervised learning with large annotated datasets.

Recently, the use of synthetic data in the training of supervised learning models has increased considerably. Since creating synthetic datasets by computer is significantly more efficient than collecting real datasets on a manual basis, once the initial setting is established, the data is remarkably cost-effective. Visual segmentation tasks are also starting to benefit from this trend. For example, Ros et al. (2016) built a large-scale synthetic collection SYNTHIA by rendering 3D city models with semantic annotations of counterparts. They combined SYNTHIA with natural urban scene datasets for training DCNNs and showed that extending SYNTHIA in the training phase significantly improved the performance of the semantic segmentation task. Saleh et al. (2018) introduced VEIS, a virtual environment system that auto-annotates synthetic images with instance-level segmentation urban elements, such as roads, pedestrians, riders, cars, etc. However, these 3D city models are not digital copies of natural cities. There is still a gap between the distribution of streetscape features in the human-created

fictional cities and natural ones, thus leading to the low realism of the generated synthetic data.

2.3.3   *Using the game engine to create synthetic data for deep learning*

The original intention of game engines is to improve game development efficiency (Nitsche & Maureen, 2004), while some notable game engines include Unreal, Unity, CryEngine, etc. Game engines enable developers to power the physics, lighting, and interactions in their virtual worlds. They can be used to generate photorealistic synthetic datasets through physically based rendering (Z. Li & Snavely, 2018). Synthetic datasets generated by game engines for deep learning training have received increasing attention from scholars in recent years. For example, Öztürk & Erçelebi (2021) used Unity to create a large number of synthetic images of birds and UAVs for implementing a classification deep learning task. However, this approach is limited to producing synthetic data for single small-sized targets, and its effectiveness for large-scale objects in urban environments is unclear. For segmentation tasks, Poucin et al. (2021) proposed a simple method that combines the use of virtual synthetic images and real-world images to facilitate instance segmentation in urban environments but lacks the creation of synthetic data for individual building facades. For integrating multiple deep learning tasks, NVIDIA Omniverse Replicator (2021) allows users to generate physically simulated synthetic data. It provides RGB images and several ground-truth outputs, such as depth and normal information, object or category segmentation, motion segmentation, forward and backward, which can accelerate the development of autonomous vehicles and robots. In general, cost-effective synthetic data outputs with universal applicability and high fidelity are the current endeavors of game engine-based approaches.

2.3.4　City digital twins for creating synthetic datasets

The implications of CDT in academic research and industrial applications triggered extensive discussions in the fields of cities, architecture, engineering, and construction (Aheleroff et al., 2021; L. Liu et al., 2022; G. Wang et al., 2022). Many cities are beginning to experiment with creating and leveraging digital duplicates of real cities at the intersection of reality and virtuality, and a plethora of urban digital assets have been produced. However, compared to DT research in manufacturing (Niu & Qin, 2021), CDT research is still in its early stages, and there is little discussion related to the utilization of digital assets from CDT to create synthetic datasets for training CNN.

The digital assets in CDT have a high level of detail (LOD) since they replicate the physical world, simulating the materials and textures of real-world objects as closely as possible. The higher the LOD of a digital asset with well-formed surfaces, the more likely it is to be rendered as a photorealistic image. Theoretically, using synthetic data with image texture distribution close to the real one for deep learning model training can obtain satisfactory accuracy of instance segmentation. However, many studies have proven that using a synthetic dataset alone as the training set cannot accomplish competitive accuracy with the real dataset, even if it is rendered from a digital asset with high LOD (Gao et al., 2020; Saleh et al., 2018). The domain adaptation has been developed to address this problem by transferring an algorithm trained in source domains to target domains (M. Wang & Deng, 2018). For example, the Balanced Gradient Contribution (BGC) training method was introduced to improve model accuracy using synthetic data (Ros, Stent, et al., 2016). The method statistics the imagery features from two domains (synthetic and real) throughout the training process, and the results are accurate for both domains. Therefore, a real dataset is necessary for the facade instance segmentation training in order to complement the real domain.

2.4 Image-based building facade data extraction for 3D city model and urban building retrofitting

2.4.1 *Extracting building facade data for semantic enrichment of building and city information models*

The 3 Dimensional City Model (3DCM) is the result of the digitization of the city, which is composed of GIS data and BIM data at a large scale and belongs to the basic data of the new smart city development. Generally speaking, the technical route of the GIS-based 3DCM construction method used is divided into five steps. (1) building bottom contour data acquisition. The building base contour data is the boundary vector data formed by the orthographic projection of the building to the ground. (2) After getting the building bottom contour data, quality check, and post-process the data, the main process covers topological closure of polylines, conversion of closed line elements to surface elements and alignment of vector data position with image base image, and unification of coordinate conversion. (3) Based on the processed building bottom contour data, the city-level building white model is automatically generated in batch by means of parametric tool modeling and exported to OBJ, FBX, OSG, and other common formats of 3D data. (4) With data-supported intelligent applications as the entry point, the 3D model is supplemented and improved with relevant fields and attributes required for business by means of automatic links, paving the way for visualization applications. (5) City-level 3D models are integrated with Digital Orthophoto Maps (DOM) and Digital Elevation Models (DEM) within the framework of the 3DCM platform to rapidly build city-level 3D GIS scenes.

The 3DCM is based on two-dimensional geographic information and can be used to analyze the city's natural and man-made features (Chun & Kim, 2010). Users can feel a realistic and intuitive sense of the synthetic city environment through interactive operations. An important task in 3D city modeling is building facade parsing and geometric analysis to create urban geometry datasets (Kong & Fan, 2021). Automated

facade geometry extraction can be done from building images or from 3D laser scan data points. Laser scan-based methods require specialized data formats and equipment, so they cannot be used globally to create 3DCM models for additional urban renewal or building energy consumption simulations (Istenič et al., 2020). In contrast, image-based techniques and computer vision are freely accessible on a larger scale. Deep learning-based semantic segmentation methods obtain building facade geometry data, mapping materials, and GIS coordinates from street view images, and these facade data can be used to build 3DCM at city scale.

The 3D city platform reproduces the real-world (physical space) city in the virtual world (cyberspace). Several 3D city platform projects have already been developed in some cities. For example, Rennes city in France has created a 3D virtual twin of itself intended for planning future urban development (Doyle, 2019). A 3D city model platform with multiple data sources was created by the Virtual Singapore project (*Virtual Singapore*, 2022), which can be accessed by the public, private, people, and research sectors to formulate urban development strategies to address the urban challenges related to city information modeling. In Japan, the PLATEAU project (*PLATEAU*, 2022) was established to optimize the management potential of cities. The project has created massive digital assets for many cities and is an essential part of the digital infrastructure development in Japan's Society 5.0 (Fukuyama, 2018).

In general, 3D city models are an important part of digital infrastructure development. By integrating various urban activity data into the 3D city model, it achieves a high degree of integration of physical space and cyberspace and further heightens urban planning, simulation, and analysis of urban activities. The advantages of acquiring building data based on street-level images for building 3DCM methods are that (1) they do not rely on expensive equipment and specific data formats, (2) the acquired building data are fine-grained, and (3) they can be used on a large scale worldwide.

2.4.2   *Extracting building facade data for urban building retrofitting*

Existing residential building retrofits for energy efficiency are crucial to reducing global greenhouse gas emissions. In 2019, residential buildings were responsible for 15% of total greenhouse gas emissions and consumed 29% of total energy in all sectors contributing to greenhouse gas emissions in the United Kingdom (Final UK Greenhouse Gas Emissions National Statistics, 2019.; UK Housing, 2019). In this context, energy efficiency retrofits in housing as an infrastructure priority can have a significant positive effect on reducing carbon emissions. They collect data and analyze data prior to deploying energy efficiency retrofit programs for individual buildings to assess the building energy profile. Large-scale collection of facade data, including building thermal indictors (such as building materials, window-door semantics), building geometry, usage of buildings, and facade deficiency information, allow for the construction of UBEM tools (Ferrando et al., 2020). A thorough evaluation and comprehension of the advantages of retrofits can create a win-win situation for stakeholders. However, providing building data for energy-efficient building retrofits on a large scale is a major challenge, especially in terms of automation. Building facade measurements based on professional surveyors in the field is time-consuming and labor-intensive, which makes it difficult to roll out energy-efficient building retrofits at a city scale.

Urban environmental data can be collected on a large scale using in-vehicle sensors. For instance, the Google Street View service collects images with geo-data from the urban environment, which is utilized in a variety of applications, such as land use identification assistance (X. Li et al., 2015) and automatic identification of building functions (J. Zhang et al., 2021b). Automated ground building facade data collection based on street-level images for UBEM is a bottom-up approach. The field has gained momentum in recent years due to automated procedures and wider accessibility of spatial and geometric data streams. M. Dai et al. (2021) designed a street-level image

segmentation model for building facade images as a basis for an overall data analysis framework. The model is based on deep learning semantic segmentation techniques and uses an integrated learning strategy. Szcześniak et al. (2022) propose a method to automatically extract the facade hole layout of each building adjacent to the Google Street View route. The automatically generated window-to-wall ratio (WWR) of 1057 buildings in Manhattan is compared with the manually determined WWR to verify the accuracy of the method.

Existing research has revealed the potential for collecting data on urban facades using street-level imagery. By incorporating multispectral capture, building characterization will contribute directly to the automation of current building energy analysis (Martinez & Choi, 2017) and city information modeling platform (Biljecki et al., 2016) for stakeholders, including local government authorities, research institutions, and residents. Related research is at a preliminary stage, and it is worthwhile for researchers to continue exploring this further.

2.5   Summary of research gaps and goals

As mentioned above, the application of building facade parsing in city modeling and building renewal has many challenges in terms of methodology and dataset. The following is a summary of the research gaps and goals.

1) Existing methods and datasets cannot overcome the challenge of facade parsing with severe occlusions, perspective distortions, and reflections. Current deep learning methods are affected by the training dataset. Most facade datasets are small in size and low in diversity, given that producing them is time-consuming and labor-intensive.

2) It is difficult to handle facades with complex environments as background. When multiple buildings are connected in a scene, current semantic segmentation-based methods have difficulty in obtaining data for individual

buildings. Moreover, building facades usually do not have any background, only the sky in most of these open-source datasets.

3) Traditional facade parsing algorithms usually focus on the regularity and symmetry of building facades. However, these methods encounter difficulties when dealing with asymmetric, complex-shaped buildings. In addition, for different application scenarios, traditional algorithms need to continuously combine existing features to achieve optimal results, failing to achieve end-to-end learning results. The existing methods lack stability and generalizability.

Existing methods are not universal and cannot be easily applied to practical projects because of the research gaps mentioned above. The following are the objectives of this study and an overview of how these gaps were bridged.

1) There are two possible solutions to the facade obstruction problem. The first is to add a priori knowledge, such as the geometric characteristics of the window, whereby the algorithm automatically corrects the window to a rectangle when the foreground is determined to be obstructed. The second is to use the picture inpainting technique to eliminate the foreground occlusion and automatically fill in the building facade mapping. The former method can solve the facade segmentation problem but cannot solve the analysis problem that requires complete building facade data. The latter method can achieve complete facade segmentation and solve the problem of requiring complete facade information, such as color calculation of the obscured facade and facade mapping extraction. However, the quantitative evaluation of the performance of facade restoration results is a difficult problem.

2) The facade extraction methods use semantic segmentation and instance segmentation. The instance segmentation can extract building facade information one by one when multiple buildings are connected in a single image. In addition, city digital twin models are emerging, which can efficiently

generate high-fidelity synthetic data for replacing real datasets. This can greatly reduce the labor and time required for manual labeling of data. The synthetic datasets generated by the game engine have been successfully used for various computer vision tasks.

3) CNN-based segmentation algorithms have yielded promising results. This research will try to use the CNN-based building facade parsing method to extract and distinguish features of objects efficiently by learning a large amount of data to obtain higher accuracy results than can be obtained by traditional methods.

Chapter 3. Automatic object removal with obstructed facades completion

3.1 Overview of obstructed facades completion

Automatic object removal is an extensively researched and fundamental task in computer vision. Unwanted objects (e.g., pedestrians, cyclists, vegetation, and cars in front of building facades) are numerous and often obscure the scene, hindering the acquisition of building facade data. When analyzing a building facade, the obscured information will lead to computational bias and incorrect results. Many studies have been conducted to automatically remove objects from urban environments (Valada et al., 2018), ranging from filtering out areas with unwanted objects to assuming a static scene and classifying object areas as outliers (Y. Sun et al., 2017). Recently, promising results have been achieved with learning-based methods for background texture inpainting (Bescos et al., 2019; Yu et al., 2019). These methods first use semantic segmentation to detect regions containing unwanted objects at the pixel level and then use image inpainting techniques to synthesize the backgrounds of these regions (Schwarz et al., 2018). Mask based manual selection of occlusion and then complementation using surrounding textures can be labor intensive. Automatic or semi-automatic based methods for detection and elimination of unwanted objects would improve this problem and save costs. The goal of this Chapter is to automatically detect unwanted objects to be removed from the urban scene and to recover the static occluded backgrounds with a reasonable image.

With the development of DCNN and semantic segmentation datasets for urban driving scenes, significant progress has been made in the automatic segmentation of street elements; by creating target object masks, various objects can be detected with

high accuracy from street-level images, especially those that are often obscured in front of buildings (Cordts et al., 2016; Y. Zhang et al., 2019). In addition, image inpainting has many applications in urban scene complementation. For example, the impact on the urban environment before and after demolition can be assessed by eliminating entire buildings (Kido et al., 2020). Synthesis of facade textures during building renovation and digital heritage restoration (D. Dai et al., 2013). These studies fill in missing images by matching and replicating background patches to achieve object removal results (N. Zhang et al., 2019). However, traditional methods are based on copies of the surrounding textures of the target objects, and they are still prone to failure in complex and irreducible scenes (Yi et al., 2020). Recent encouraging advances in data-driven image drawing methods, which are more effective than classical methods in handling object removal for complex scenes and large occlusion rate images, have attracted the interest of researchers. However, learning-based methods require a large amount of data for training, and building diverse and high-quality databases of building facades is a challenge. In addition, it is not desired that after object removal and filling, evaluating the quality of the synthetic image will be a challenge because real textures are difficult to obtain as a reference. Several generated image quality assessment metrics, such as information fidelity criterion (IFC) (Sheikh et al., 2005), mean squared error (MSE), peak signal-to-noise ratio (PSNR), structural similarity (SSIM) index, and feature similarity indexing method (FSIM) (Sara et al., 2019), have been developed to measure the similarity between generated images and ground truth through full-reference.

This Chapter expects to remove unwanted objects from street view images with obscured facade completions. A custom occluded facade completion dataset is created. Several state-of-the-art DCNNs for image classification were selected to extract invalid data from the street view images, and then a dataset of building facades was created for learning-based image inpainting. Next, semantic segmentation is used to automatically detect regions containing unwanted objects. A GAN-based image inpainting method is proposed to provide a cost-effective tool for matching physical space with digital

objects in large-scale images by filling the missing region content of building facades with contextual concerns. Finally, qualitative and quantitative validations for evaluating the quality of the generated images are proposed.

3.2   Method and materials

Figure 3.1 depicts the workflow for automatic object removal and facade inpainting in three steps. Firstly, the building facade dataset for GAN-based image inpainting was constructed. These images can be retrieved from street view services and purified with a classifier. Secondly, a semantic segmentation algorithm based on the Cityscapes dataset can detect the street-level obstacles. Thirdly, a free-form image inpainting tool was presented to fill the blank with contextual attention.
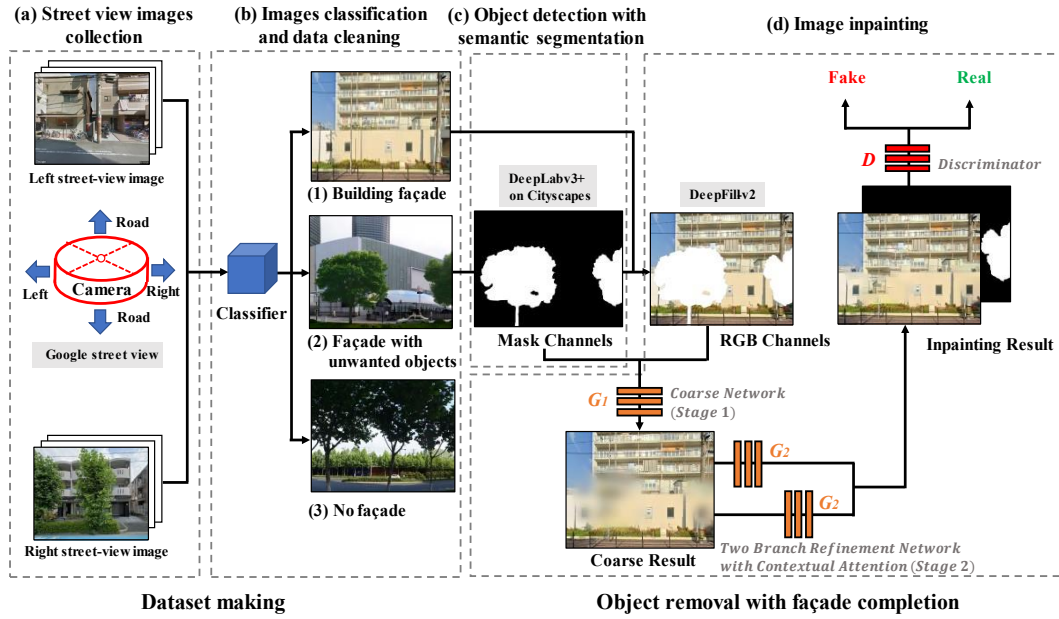
Figure 3.1. The overall workflow.

3.2.1   Dataset making

The road networks for multiple cities using open source geographic information data were extracted (Anguelov et al., 2010) to build the GAN-based image inpainting dataset, as shown in Figure 3.2a. The geographic coordinates of road points were

generated by equally sampling direction for each sampling point from the Google Map API (viewing angle is 90 degrees, the horizontal angle is 0 degrees, the compass heading of the camera is θ, and the picture size is 680×512 pixels). As shown in Figures 3.2b and 3.2c, to ensure that the angle of the crawled picture is perpendicular to the street, θ is calculated as follows:

$$\theta = arctan\left(y_A - y_B, \; x_A - x_B\right) \tag{3.1}$$

where point A (x_A, y_A) and Point B (x_B, y_B) are two adjacent points on the road centerline, and the angle θ is the deflection angle that grabs the orthographic projection of the building facade in the online street view service. The existing building facade in the urban environment can be obtained (Figure 3.2d), and these images are used in the training set for the GAN-based model. The street view image recording structure is depicted in Figure 3.3.
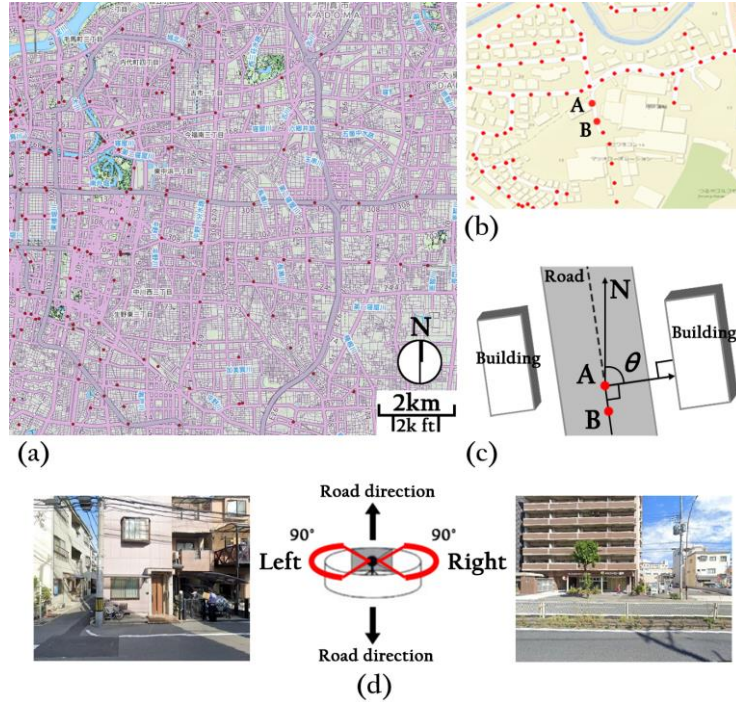
Figure 3.2. The collection method of the perpendicular street facade.

The image generative inpainting model involves learning textures from a large number of unobscured facade images. Because the collected images of street facades

contain a great deal of noise and unwanted images, an image classification algorithm is required to clean them up. 2,700 images of street facades are selected manually from street view services, with 900 images per class. Data augmentation is used to increase the diversity and size of the training sample, which prevents overfitting and improves model performance (Shorten & Khoshgoftaar, 2019). The dataset for the facade inpainting GAN is named 'Street view dataset for building facade inpainting (SVBFI).'

```
┌─────────────────┐     ┌──────────┐     ┌──────────┐     ┌────────┐
│ Street Network  ├──┬──┤ Street 1 ├──┬──┤ Point 1  ├──┬──┤  Left  │
└─────────────────┘  │  └──────────┘  │  └──────────┘  │  └────────┘
                     │  ┌──────────┐  │  ┌──────────┐  │  ┌────────┐
                     ├──┤ Street 2 │  ├──┤ Point 2  │  └──┤ Right  │
                     │  └──────────┘  │  └──────────┘     └────────┘
                     │  ┌──────────┐  │  ┌──────────┐
                     ├──┤ Street 3 │  └──┤ Point 3  │
                     │  └──────────┘     └──────────┘
                     │     … …              … …
```
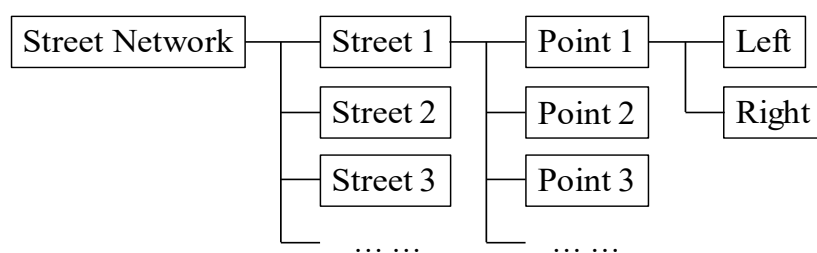
Figure 3.3. The structure of data recording.

3.2.2   *Semantic segmentation*

Semantic segmentation combines image classification and image detection to perform categorization and annotation in terms of pixel-by-pixel in an image (Lateef & Ruichek, 2019). Semantic segmentation tasks are composed of two components: the dataset and the segmentation algorithm. Unwanted objects were determined in the object segmentation dataset using Cityscapes (Cordts et al., 2016). DeepLabv3+ (L.-C. Chen et al., 2018) was used for semantic segmentation.

As shown in Figure 3.1c, DeepLabv3+ is used on the Cityscapes test set for object segmentation, and its mIoU can reach 82.1%. In this Chapter, several classes of obstacles in the streetscapes, that is, pedestrians, cyclists, vegetation, and cars, are taken as specific objects to be eliminated. Through detecting by DeepLabv3+ on the Cityscapes, they are mask images in the input image of the inpainting GAN model, as illustrated in Figure 3.4.
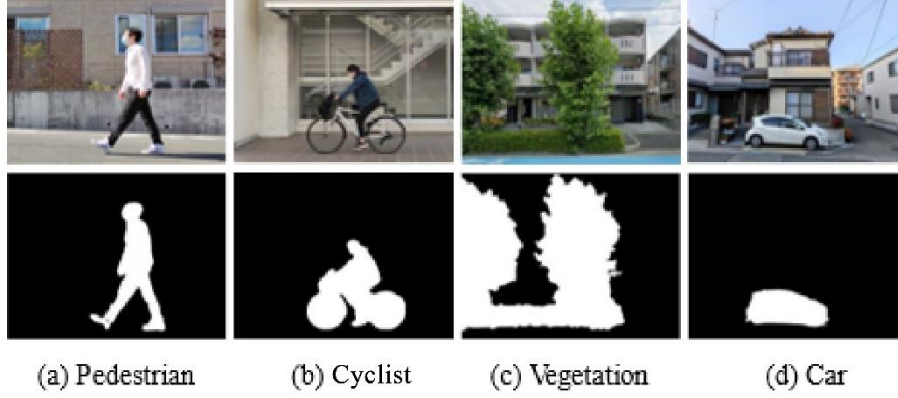
Figure 3.4. Identified and eliminated labels, including pedestrian, cyclist, vegetation, and car.

3.2.3   *Image inpainting*

The facade inpainting method uses the open-source model DeepFill-v2 (Yu et al., 2019), a free-form image inpainting method with gated convolution, to generate alternative contents for blank areas in a visually realistic and semantically correct manner. Figure 3.1d introduces the simplified overall network structure of DeepFill-v2. For this neural network, the input data is divided into two channels: RGB Channel and Mask Channel. The architecture of the model consists of a two-stage generator and a discriminator. The initial stage of the generator consists of a coarse network that produces a coarse output. The second stage is a two-branch refinement network with contextual attention that produces a refined result, which can significantly enhance the image quality and repair results' fidelity. Gated convolution dramatically improves performance when the mask pictures have arbitrary shapes and the inputs are conditionally free-from, such as in the sparse sketch (Yu et al., 2019). Thus, the model is able to synthesize a new image structure on a blank image in a learning-based manner, using the surrounding image features as a reference to generate reliable estimates.

3.3   Experiments and results

This part describes the production of the SVBFI dataset, the training of the GAN model, and the quality evaluation of the generated images.

3.3.1 Street-level images classification

For the training and testing sets, there were 2,700 pictures each. Each class was given 750 images, which accounted for 0.83 percent of the total training set. 450 testing images accounted for just 0.17 percent of the total training set. Several state-of-the-art CNN models are introduced, named InceptionNet_v4 (Szegedy et al., 2016), XceptionNet (Chollet, 2017), EfficientNet (Bódis-Szomorú et al., 2017), and ResNeSt (H. Zhang et al., 2020), by fine-tuning all the convolutional layers with benchmark datasets. Figure 3.5 depicts the normalized confusion matrix for the trained CNNs as determined by the test data. One way to measure classification accuracy was to use the matrix value, which represents the percentage of samples from one category that was correctly classified into another. The F_1 score is utilized to evaluate model performance, which was generated using the equations below:

$$F_1 = 2 \cdot \frac{p \cdot r}{p + r} \tag{3.2}$$

where p is precision and r is recall. After calculating the F_1 scores of the four networks, the classification performance of ResNeSt performs better than the other networks. For the class of building facades, ResNeSt achieved the highest F_1 score with 0.87. Therefore, the trained ResNeSt model was selected for the upcoming extraction of unoccluded facade images.
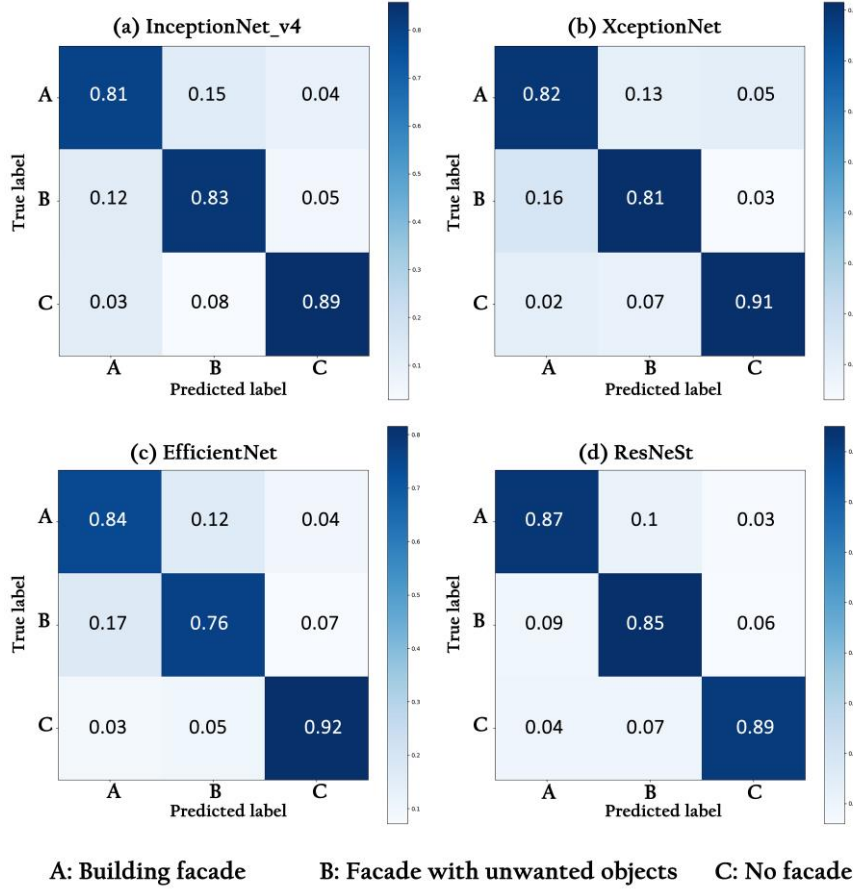
Figure 3.5. The test results of normalized confusion matrices associated with the four networks. InceptionNet_v4 (Top-left), XceptionNet (Top-right), EfficientNet (Bottom-left) and ResNeSt (Bottom-right).

More than 300,000 street view images were downloaded from Google Street View, with each image measuring 680 × 512 pixels and containing images of unobstructed facades, facades with unwanted objects, and no facades. The ResNeSt model had been pre-trained in the previous step. The SVBFI datasets of 9,000 unoccluded facade images are obtained by filtering the street view images. The SVBFI is used as the training set for image inpainting GAN.

3.3.2 Image inpainting model training

GANs, in general, are made up of a generator and a discriminator, which compete with each other to produce images that are constantly optimized and semantically similar to the ground-truth image. Recently developed spectral normalization (Miyato

et al., 2018) was used to stabilize the GANs training further. The SN-GAN is a utilized default fast approximation algorithm for spectral normalization. To discriminate if the input was real or fake, the hinge loss is used as the objective function for the generator $\mathcal{L}_G$ and discriminator $\mathcal{L}_{D^{sn}}$.

$$\mathcal{L}_G = -\mathbb{E}_{z\sim\mathbb{P}_z(z)}[D^{sn}(G(z))] \tag{3.3}$$

$$\mathcal{L}_{D^{sn}} = \mathbb{E}_{x\sim\mathbb{P}_{data}(x)}[\text{Re}\,LU(\mathbb{1}-D^{sn}(x)) + \mathbb{E}_{z\sim\mathbb{P}_z(z)}[\text{ReL}\,U(\mathbb{1}+D^{sn}(G(z)))] \tag{3.4}$$

where D^{sn} represents spectral-normalized discriminator, $G(z)$ is an image inpainting network that takes incomplete image z. In the training process, the datasets were trained for 300 epochs, which iterated 216,000 steps. Figure 3.6 shows the loss of generator $\mathcal{L}_G$ and discriminator $\mathcal{L}_{D^{sn}}$ in this model. The loss of the generator was decreasing, and the discriminator loss was increasing. As the generator and discriminator reach equilibrium, the overall performance of the work steadily improves.
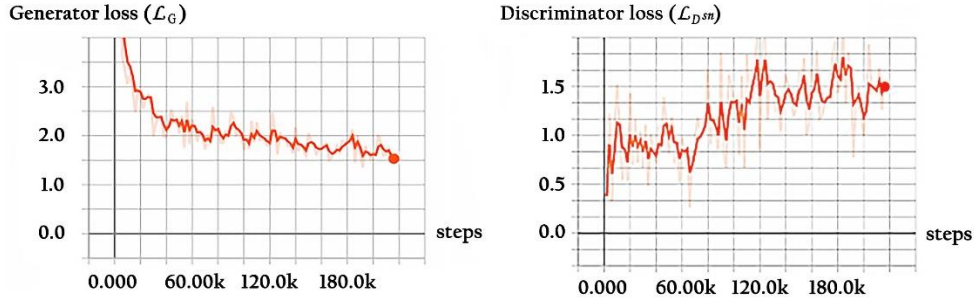
Figure 3.6. Training loss of the GAN model. Left: generator loss; Right: discriminator loss.

3.3.3   *Testing and qualitative comparisons*

Figure 3.7 depicts a street-level test example of automatic object removal with facade smearing. Both the proposed method and the previous example-based image smearing method were introduced into the experiment. The two synthetic images are compared with ground truth images. The ROI of each project was covered by a mask full of contextual concerns, and no post-processing was performed to ensure fairness.

| Original Image | Input | Output by the proposed method | Output by exemplar-based method | Groundtruth |
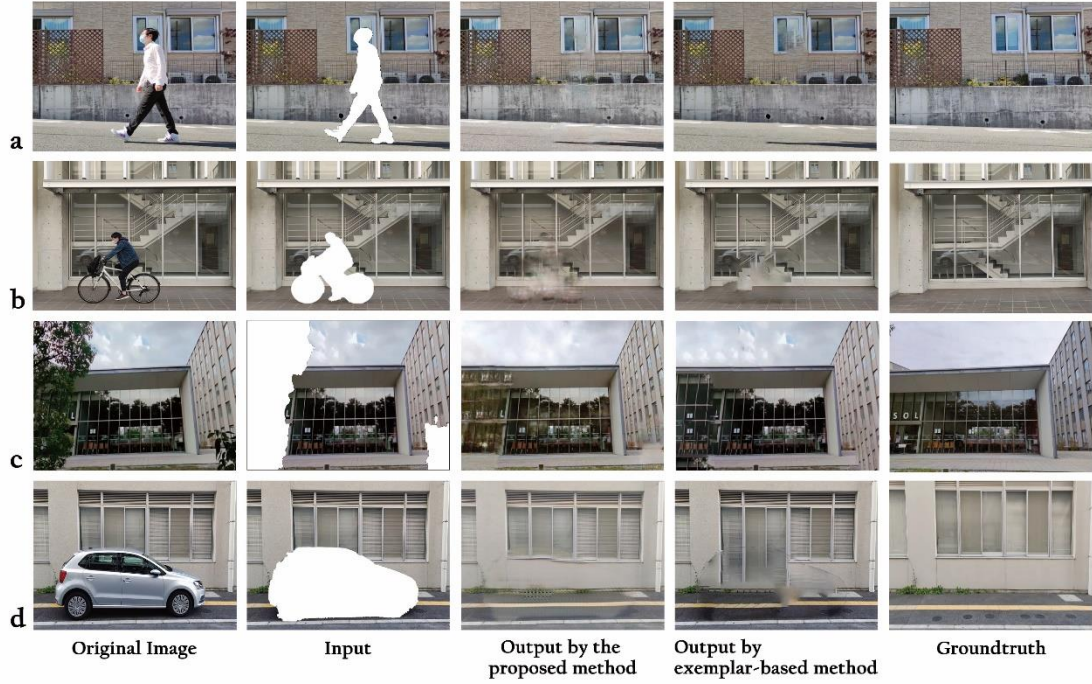
Figure 3.7. Test examples of automatic object removal and facade completion with several classes. (a) People, (b) cyclist, (c) vegetation, (d) car.

As shown in Figure 3.7a and Figure 3.7d, DeepFill-v2 using the SVBFI dataset performs well visually, with the synthesized parts matching the color of the surrounding texture. Figure 3.7b shows that the model can fill in the facades with transparent and reflective materials in the input image, but it blurs some details. Figure 3.7c shows that the proposed model accurately contours the building in the input image to the actual situation. Figure 3.7d shows that the proposed method is able to perform well in recovering regular building components, such as rectangular windows, in the case of facades with complex backgrounds. The results show that the GAN method learned from massive data can effectively consider the image semantics and outperforms the exemplar-based methods in complex scenes.

3.3.4   *Validation and quantitative comparisons*

Two widely used full-reference IQA metrics, PSNR and IFC, are used for the quality assessment of generated images based on visual perception. On SVBFI, 900 images are used to test the proposed model against the exemplar-based approach.

Unoccluded facade images are used as ground truth, and mask images superimposed are used as input images, as shown in Figure 3.8 input images. The occluded objects, including people, cyclists, trees, and cars, are used as mask shapes. These unwanted objects overlap the mask area on the ground in the street-level images to simulate obstructions in the actual street. The masking ratio is allocated from 0 to 50% of the image size. Figure 3.8 shows the validation example of the proposed method and the exemplar-based method at different mask ratios from 0-10% to 40-50%.

Figure 3.8. Validation examples of automatic object removal and facade completion.

Figures 3.9 and 3.10 show the generated image quality based on the PSNR and IFC. The quantitative comparisons of full-reference metrics indicate that the proposed method achieves better results than the exemplar-based method. The proposed model improves PSNR by 2.26 dB and IFC by 0.061 over the fill-by-replication method across the entire mask range. Although the GAN-based approach using the proposed tailored dataset is marginally superior to the filling through copying method for mask ratios of 40-50%, with PSNR improving by 1.56 dB and IFC improving by 0.042 in mean value, it is worth noting.
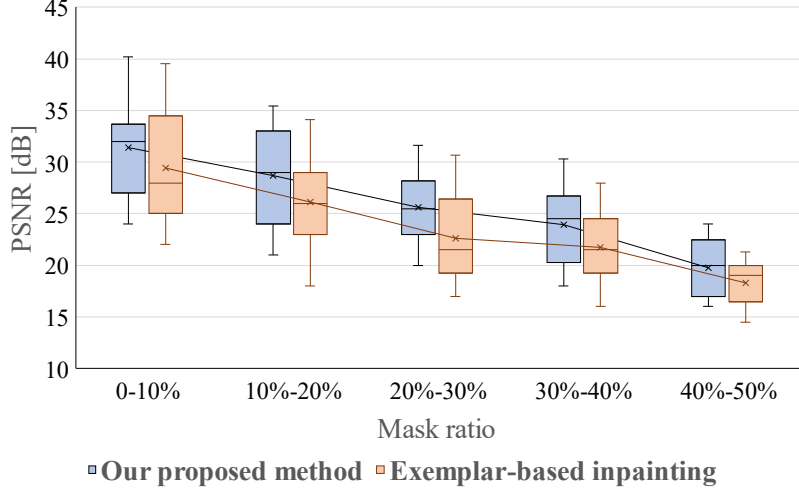
Figure 3.9. The PSNR results with the proposed method and exemplar-based method of different mask ratios on the validation data.
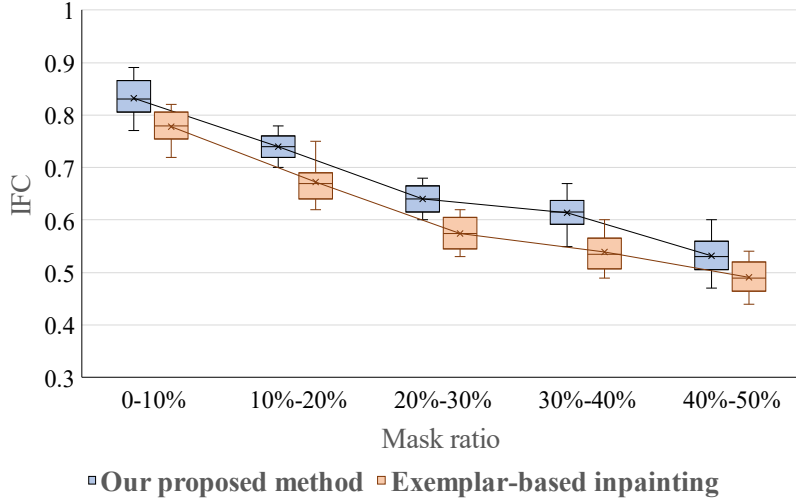
Figure 3.10. The IFC results with the proposed method and the exemplar-based method of various mask ratios.

3.4   Discussion

3.4.1   Advantages

The system can perform object removal tasks for 2D images in different street scenes. Eliminating obstacles in front of buildings can help improve data completeness when extracting facade information using street view images and computer vision techniques. The method balances the high quality of the generated images with the detail of the textures and performs better for complex scenes compared to the exemplar-

based method. In addition, a dataset SVBFI is built for learning-based obstructed facade painting from street view images. The proposed dataset is more focused and consumes less computation for training than current open-source datasets such as Places2. The proposed method is more practical than previous methods for field simulation (Kido et al., 2020) and does not require background facade information in advance. The previous observation-based method requires a pre-taken background scene of the image background, which can be used as a reference to directly replace the foreground obstacle. The inpainting-based approach used in this paper uses the texture and patch information of the source image to fill in the detected regions. Therefore, the method in this paper only needs to train a GAN-based inpainting model to handle obstacle elimination, which is more convenient and cost-effective than the observation method because it saves the trouble of shooting background information in the field. An image-based method can help stakeholders visualize the redevelopment project and eliminate unnecessary elements. The proposed strategy is quick to implement, lightweight to deploy, and applicable to a wide range of situations, making it an excellent starting point for further research.

3.4.2 Limitations

Currently, the system uses semantic segmentation, where two visually overlapping objects are difficult to segment separately. For example, if two cars visually overlap, they will be eliminated together. This shortcoming can be solved by using instance segmentation. In addition, this method does not automatically detect shadows or remove them in the subsequent inpainting. An example is shown in Figure 3.7a, where the shadows of pedestrians are not removed from the picture. This deficiency can be addressed by labeling the shadows of objects in the training set of the semantic segmentation model.

3.5   Summary of this Chapter

In this Chapter, an image-based approach for cityscape visualization is presented. Unwanted objects on the street level scene are automatically detected without prior background information. The visual removal of these objects is accomplished by smearing the ROI. To this end, a semantic segmentation model was introduced to detect the ROI of obstructions. Then, the SVBFI dataset was used for training the GAN-based image inpainting model. The comparison experiments show this approach performs better than the exemplar-based inpainting method. The on-site validation results proved the effectiveness of the proposed method. By automatically removing unwanted objects and filling in obstructed building facades by replication and modeling (J. Zhang et al., 2021a), this approach improves the degradation of information acquisition from buildings due to obscuration. By eliminating redundant objects and using only images instead of 3D models, urban landscapes can be simulated and visualized.

Chapter 4. Instance segmentation of building facades based on city digital twin datasets

4.1   Overview of facade instance segmentation using synthetic data

The semantic enrichment of facades can be used for building information modeling in construction management and architecture design (S. Cai et al., 2019; Xue et al., 2021). Large-scale automated measurement of building facades using semantic segmentation can provide data support both for retrofits and energy analyses of buildings (M. Dai et al., 2021; M. Deng et al., 2019). Using deep learning in real-time visualization of demolished building facades can be used to enhance stakeholder engagement and design assistance (Kikuchi et al., 2021; J. Zhang et al., 2021a). However, most previous studies have involved the semantic segmentation of building facades, and it is difficult to extract the instance information for connected building facades one by one using semantic segmentation. In contrast, the annotation task for the instance segmentation requires both classification at the pixel level and the identification of different instances of the same class (Ghiasi et al., 2021). It is challenging to collect large-scale annotated datasets for the segmentation of individual building facades (M. Dai et al., 2021).

Supervised machine learning methods generally perform well in instance segmentation tasks. Even the best autoencoders, visual descriptors, and discriminative machine learning techniques cannot obtain reliable results without a properly annotated dataset containing sufficient diversity. Data annotation is a laborious, manual task that

requires precise correction as it is prone to errors. The quality and quantity of the dataset largely determine how well the instance segmentation model performs. Therefore, a large and diverse annotated dataset is necessary for performing the instance segmentation of building facades with high accuracy.

With the development of the DCNNs, instance segmentation tasks have yielded significant performance gains (Z. Cai & Vasconcelos, 2019). On the one hand, researchers expected to use massive amounts of accurately annotated data for DCNN training, and on the other hand, they often struggled with the expensive costs associated with all that data. Photorealistic synthetic data have received increased attention as a means for addressing these issues owing to the possibility of automatically generating a vast number of high-quality images with diverse annotations (Tremblay et al., 2018).

Synthesizing instance-labeled datasets of building facades from 3D city models for DCNN training is a promising method for reducing labeling costs and improving model performance. However, when the application scenarios become complex, the synthetic images of virtual urban environments have difficulty accurately representing the original features in the physical world, such as object materials and ambient lighting, and their misrepresentation can lead to problems with dataset shifts (Quiñonero-Candela et al., 2009). Recently, digital twins (DTs) have been proposed as a possibility for bridging the gap between synthetic and real-world data.

The DT paradigm is an information construct comprising a physical asset, its corresponding digital asset, and the data connection to them (J. Liu et al., 2021). It has recently been applied to urban systems to produce models called city digital twins (CDTs) (Fan et al., 2021; Shahat et al., 2021). Typically, a digital asset in a CDT with a high level of detail (LOD) is a copy of its counterpart in the physical world and accurately reflects real-world information. Using synthetic datasets from CDTs rather than from fictional cities (which have no real-world counterparts) as the training sets could be promising for improving the accuracy of the instance segmentation model.

The anticipated challenges include synthetic datasets generation and training DCNN-based instance segmentation models. For the former, several 3D virtual city-based methods have been proposed for producing synthetic datasets of urban features, but these virtual data are criticized for their realism. Digital assets from the physical world have closer to real materials and are expected to be ideal for rendering high-fidelity synthetic images for the training of DCNN models. Figure 4.1 shows a comparison training process of manual labeling and synthetic datasets for the facade instance segmentation. This study is expected to substitute manual labeling by using an auto-generation system based on CDT to create training sets for DCNN models.
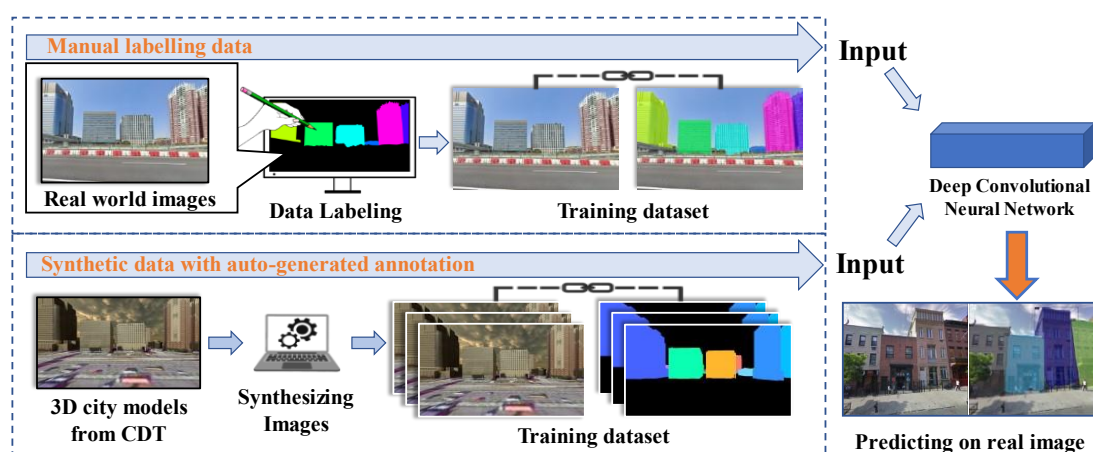
Figure 4.1. A comparison of manually annotated datasets and automatically generated synthetic datasets. (The conventional method requires hand-made labeling of images to produce the training set, while the proposed system can automatically create synthetic data with instance annotations by using digital assets of CDT.)

For the latter, the instance segmentation images generated by 3D models are not directly usable and need to be converted to an annotation format with object class and mask polygon for DCNN-based model training and evaluation. Besides, suboptimal accuracy of semantic segmentation on real-world images has been seen in previous studies when only using synthetic datasets as training sets. This study introduced a

hybrid dataset to solve this issue and verified the performance of the synthetic dataset generated by CDT for facades instance segmentation in multiple real cities.

This study attempts to develop an auto-generation method of synthetic facade datasets for training instance segmentation models using high-quality digital assets from CDT. The proposed system can produce synthetic street view images with auto-annotated individual facade instances, including mask polygon and class information, for DCNN training. In addition, a hybrid dataset, consisting of a variable proportion of synthetic and real data, is built to train the DCNN model to compare the results of three training sets (synthetic only, real-world only, and hybrid) on the instance segmentation accuracy of building facades, which can show the contribution of synthetic data in improving the DCNN model performance and reducing the annotation cost. Furthermore, this Chapter validated the pre-trained model using the proposed datasets in multiple cities to demonstrate the effectiveness and transferability of the research framework. The quantitative and qualitative results indicate that the proposed method can produce cost-effective synthetic data of building facades for training supervised instance segmentation models and can potentially be used to extract and integrate facade instance information in built environments.

4.2   Method and material

This section presents the process and evaluation methodology for the proposed datasets (Figure 4.2). First, the 3D city model was downloaded from a city information modeling platform and was imported into the Unity game engine for asset management. The virtual camera and atmospheric effects were set up for rendering the 3D model, and synthetic street images and facade annotations were produced. Second, synthetic datasets (CDT-based and virtual-based), real-world datasets, and hybrid datasets are built for training instance segmentation models. The synthetic dataset was converted so that it would be available for DCNN-based model training, and the real dataset was collected from street-view images and manually labeled. Several state-of-the-art

instance segmentation models were selected. Finally, three assessment aspects are considered: precision, size, and the number of detections. Six corresponding COCO metrics are then introduced to evaluate the pre-trained models on real-world street-level images using the proposed dataset.
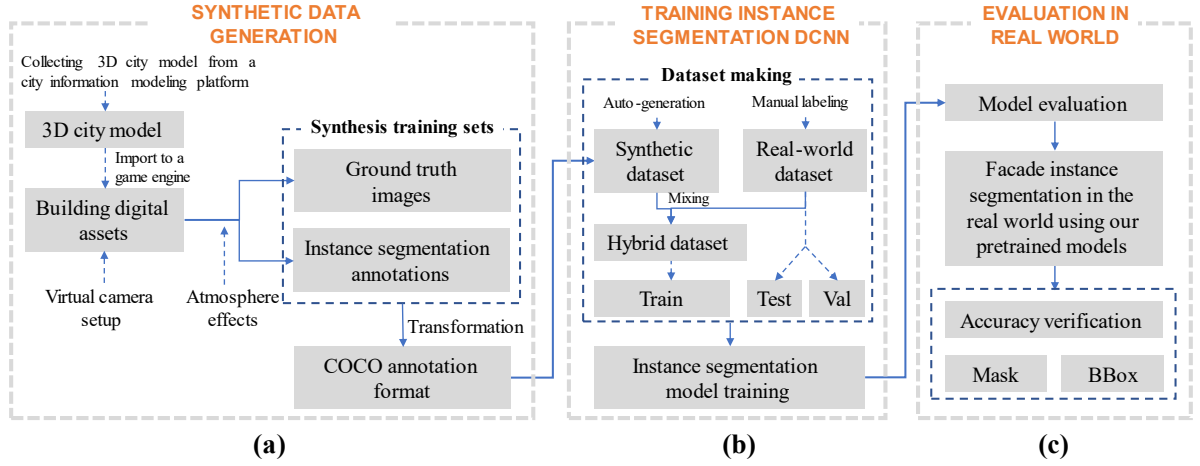
Figure 4.2. Workflow for the study: (a) the synthetic data generation process, (b) training DCNN-based instance segmentation, and (c) evaluation using real-world imagery.

4.2.1 Study areas and datasets

The selected study area is Koto City, which is located in the eastern part of Tokyo. The data used in this study included a 3D city model, synthetic images (CDT and virtual) with facade annotations, and real-world street-level images with facade annotations, as shown in Table 4.1. The 3D city model was obtained from the PLATEAU platform. The synthetic datasets for building instance training were auto-generated using the proposed method. Natural street-view images were extracted from Google Street View (GSV), and facade annotations were obtained using manual labeling.

Table 4.1. Datasets description

Datasets	Data source	Description
3D city model	PLATEAU	This platform provides digital assets of buildings to the public for research or commercial purposes. It covers most cities in Japan and contains massive LOD1 and LOD2 building models. The LOD2 building models used in this study are textured, and their geometries are created by emulating the corresponding real-world buildings.
CDT synthetic	Auto-generated	Synthetic images are digital copies of street-view

dataset	by the system	images in the real world and have building facade annotations.
Virtual synthetic dataset	Auto-generated by the system	Synthetic data are generated from a fictional 3D city model that includes virtual street views and annotations of fictitious facades. It is used for comparison with the CDT data.
Real-world dataset	Google Street View and manually labeling	Real-world images of street views with building annotations.

Figure 4.3a shows a well-developed area of Tokyo as it is represented in PLATEAU. It is a LOD1 building model that covers 23 wards in Tokyo with a total area of about 627 square kilometers. The pink regions are 3D building models with LOD2, with a total area of 6.72 square kilometers. Figure 4.3b shows the building digital assets with LOD2 downloaded for this study. To increase the diversity of the data sample, digital assets are selected for different types of buildings: residential, commercial, office, industrial, and transportation. In addition, buildings from a diverse range of sizes, including low-rise, mid-rise, high-rise, and large urban complexes, were captured in the study area.
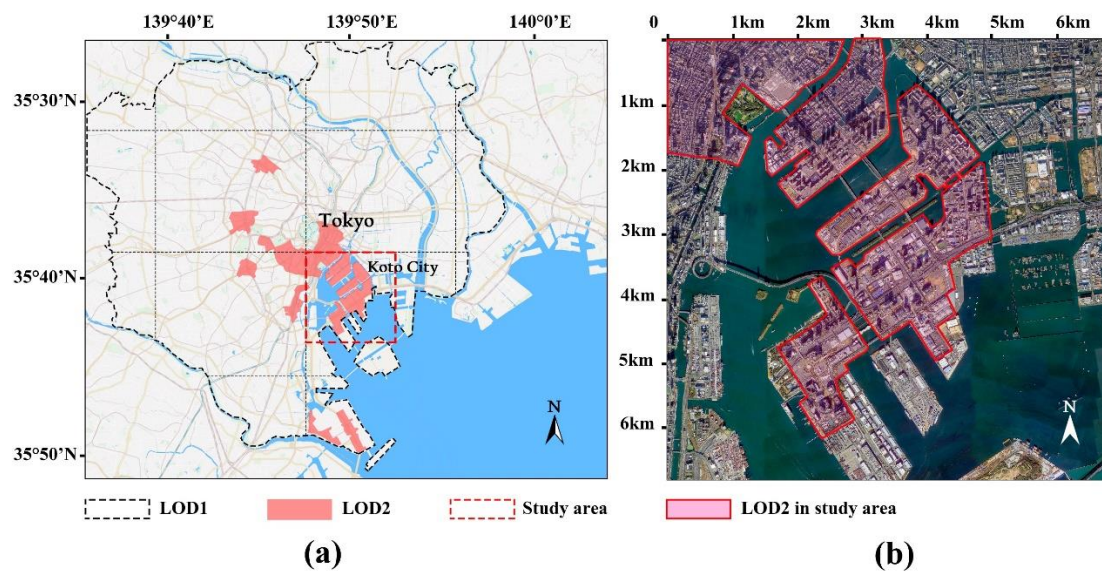
Figure 4.3. The 3D city model from PLATEAU. (a) The built-up area of PLATEAU in Tokyo and (b) the study area: Koto-ku, Tokyo, Japan.

4.2.2   *Automatic generation of synthetic data*

4.2.2.1. *3D city model downloaded and pre-processed*

Figure 4.4 shows an aerial view of the 3D city model in the study area. In total, it contains 413 building models with LOD2. The selected 3D city model can be downloaded for free from the Project PLATUEU database (n.d.) in the CityGML 2.0 and Filmbox formats, the latter of which was used for this study. Since the 3D model includes geographic information, the satellite and topographic map can be loaded in to match the virtual world. In addition, common elements from the urban environment of the actual image are arbitrarily placed into the 3D model to increase its realism, including greenery, overpasses, and vehicles.

Distortion is an important parameter for texture mapping of the 3D model. The textures for the CDT model were taken from the physical camera and underwent rigorous distortion correction before being placed on the model surface, as close as possible to the real building facade. Figure 4.5 shows an example of the CDT models that were used and a distortion-corrected mapping of a building facade next to its real-world counterpart.

This study used an open-source package from Unity called Unity Perception (Borkman et al., 2021), which can help speed up and simplify the process of generating labeled synthetic datasets. In the virtual city model, each building has an object ID that was created by PLATUEU, and they can be automatically tagged with the category BUILDING. Consequently, the virtual camera in Unity can automatically recognize different objects of the same category and generate instance annotations.
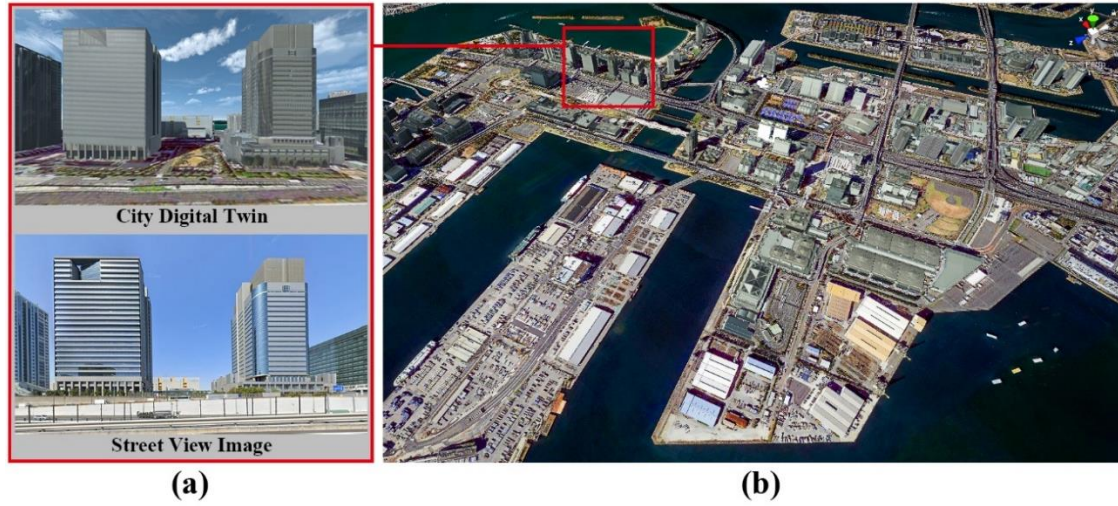
Figure 4.4. 3D city model of the study area. (a) An example of CDT with its real-world street views (Wangan-doro Avenue, Tokyo; March 2021; latitude: 35.6283, longitude: 139.7782). (b) Aerial view of city digital twin.
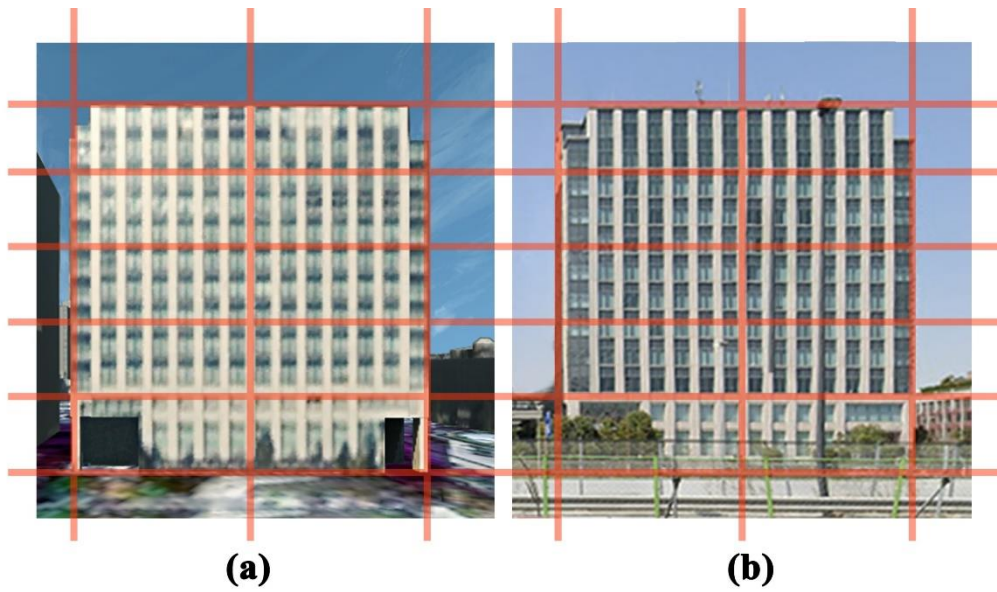
Figure 4.5. Distortion correction of a CDT model texture mapping. (a) The CDT texture is corrected for distortion before it is placed on the model surface. (b) Real-world building facade (Wangan-doro Avenue, Tokyo; March 2021; latitude: 35.6279, longitude: 139.7785).

4.2.2.2. *Virtual camera setup in the game engine*

In the virtual environment, the camera on top of a car is set to acquire the building images, and the height of the camera is limited to between 1.5 m and 2.5 m above the ground. The acquisition platform consisted of one multi-camera made up of four

monocular cameras linked by a common center, with the orientation changing every 90 degrees, as shown in Figure 4.6. All of the cameras have a horizontal field of view (FOV) angle of 100 degrees and a vertical FOV angle of 79 degrees. The vehicle moved through the 3D city and interacted dynamically with the buildings within it. This interaction allowed us to collect building images at different horizontal angles. This collection was intended to provide data that could be used with the spatio-temporal constraints of the objects.
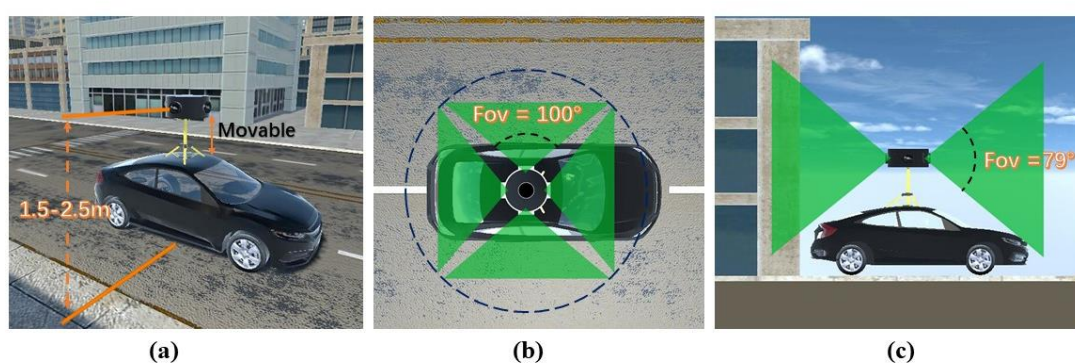
Figure 4.6. Virtual car setup used for data acquisition. One virtual multi-cameras with four perspectives are used. The horizontal and vertical view angles are 100 degrees and 79 degrees.

4.2.2.3. 3D city model rendering

Unity has the ability to adjust lighting and global illumination, allowing it to be used as a rendering tool for 3D city models. First, various atmospheric effects are used to increase the realism and diversity of the virtual scenes for data augmentation. Second, a vehicle with an attached multi-camera collected street-level images from four directions in the 3D city model and automatically generated instance annotations for the buildings.

(1) Details of the Unity renderer

The Lit Shader from Unity's Universal Render Pipeline (URP) is used. Lit Shader is provided by Unity and uses the Bidirectional Reflectance Distribution Function

(BRDF) model to easily create realistic materials (Doppioslash, 2018). Direct and ambient lighting is turned on, and shadows were rendered. Baking a city-scale model to derive the ambient occlusion requires a large amount of data and computational resources, which affects the efficiency of generating synthetic data. As an alternative, post-processing methods are used, such as anti-aliasing and adjusting the exposure and white balance, to improve the realism of the rendered images.

(2) Atmospheric effects for data augmentation

The same scene can vary significantly under different atmospheric parameters, such as solar zenith angle, sky tone, and cloud density. Four atmospheric effects are used for the city model in Unity to enhance the diversity of the synthetic data, including rendering the scene in sunny conditions, cloudy conditions, and during the evening. These are shown in Figure 4.7.
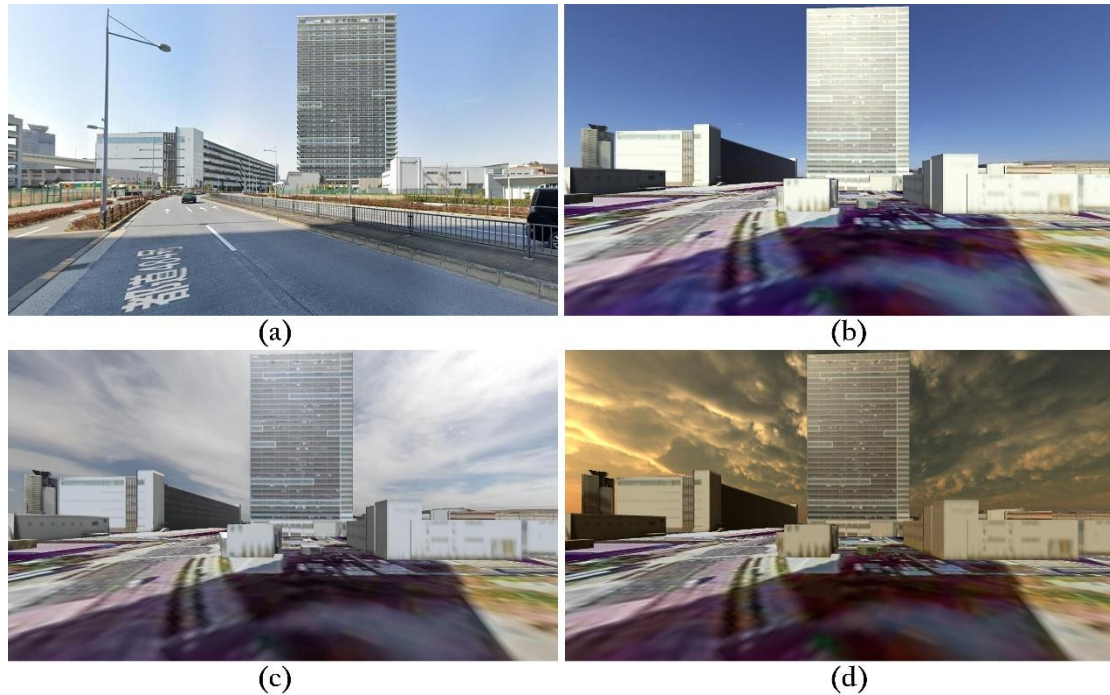
Figure 4.7. Real street-view image (latitude 35.6351; longitude 139.7829) and rendering images of the CDT with different atmospheric conditions. (a) Real street view, (b) synthetic image with sunny conditions, (c) synthetic image with cloudy conditions, and (d) synthetic image during the evening.

(3) Automatic generation of the ground-truth and facade annotations

The post-processing capabilities of Unity can render virtual scenes and automatically obtain synthetic data, including RGB images of street views and the instance segmentation of building facades. An example of a single shot is shown in Figure 4.8. It shows the four views that were captured by the multi-camera system with the corresponding instance segmentation masks of the building facades.
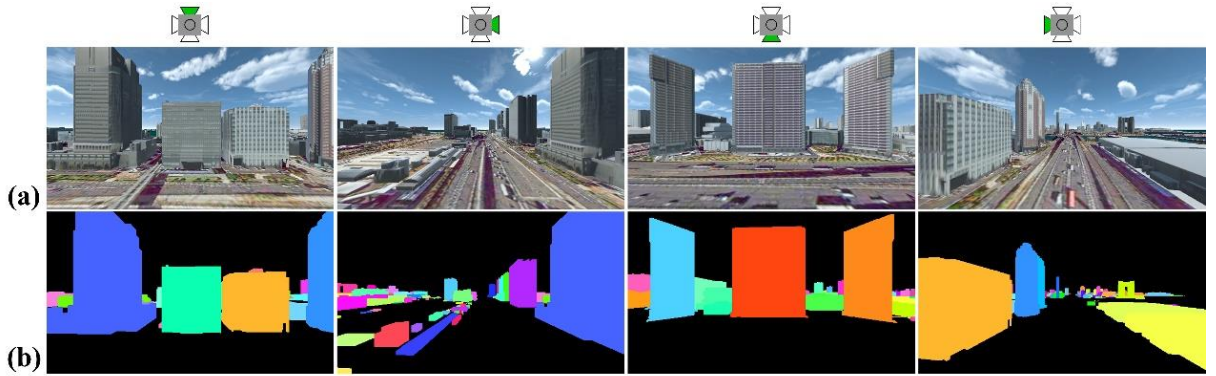
Figure 4.8. Four views of a single shot captured by the multi-camera system for the CDT synthetic data (the coordinates of the real-world counterpart are latitude 35.6284, longitude 139.7784): (a) synthetic street views and (b) corresponding instance segmentation masks.

4.2.3   Training instance segmentation

4.2.3.1. Facade instance annotations making

Four datasets with ground-truth images and instance annotations are built for training the building instance segmentation, including a synthetic dataset (CDT and virtual), a real-world dataset, and a hybrid dataset. Previous studies have demonstrated that using synthetic data alone for semantic segmentation tasks involving real images is unsatisfactory (Ikeno et al., 2021; B. Sun & Saenko, 2014; Vazquez et al., 2013). Alternatively, training a model on a large number of synthetic images and then fine-tuning it on a reduced number of real-world ones yields better results (Ros, Sellart, et al., 2016). In this study, a hybrid dataset is created and used as the training set to

demonstrate the effectiveness of CDT synthetic data in improving the instance segmentation results of building facades for real-world images. The hybrid dataset, which is called the Hybrid collection of Synthetic and Real-world Building Facade Images and Annotations (HSRBFIA), can construct subsets of synthetic and real-world data with variable proportions of each.

(1) Synthetic dataset transformed to the COCO annotation format

Instance segmentation comes with additional complexity in the form of label and annotation formats, requiring a unique value for each element in the sample image during the training process. The data format generated by Unity cannot be used directly to train instance segmentation algorithms. Most instance segmentation algorithms follow the COCO annotation format. Therefore, a format conversion open-source tool is developed (Mortyzhang, 2021/2022b) that converts the data format generated in the previous step to the COCO annotation format. The conversion procedure produces the categories and annotations fields from synthetic mask images, and since this study has only the category BUILDING. The annotations are an array of multiple annotation instances (Lin et al., 2014), as shown in Table 4.2.

Table 4.2. Fields split by instance annotations

```
Annotation {
    "id": int,
    "image_id": int,
    "category_id": int,
    "segmentation": Run length encoding (RLE) or polygon,
    "area": float,
    "Bbox": [x, y, width, height],
    "iscrowd": 0 or 1,
}
```

"id" and "image_id" represent the serial number of the image. "category_id" points to the category of the tag. If iscrowd=0, the segmentation is in polygon format, and if

iscrowd=1, the segmentation is in Run Length Encoding (RLE) format. "area" is the area of encoded masks, which is the labeled area. "Bbox" is the bounding box of the detection object. The coordinates of the upper-left corner of the rectangular box and its length and width are provided in the form of an array.

(2) Real-world dataset from GSV
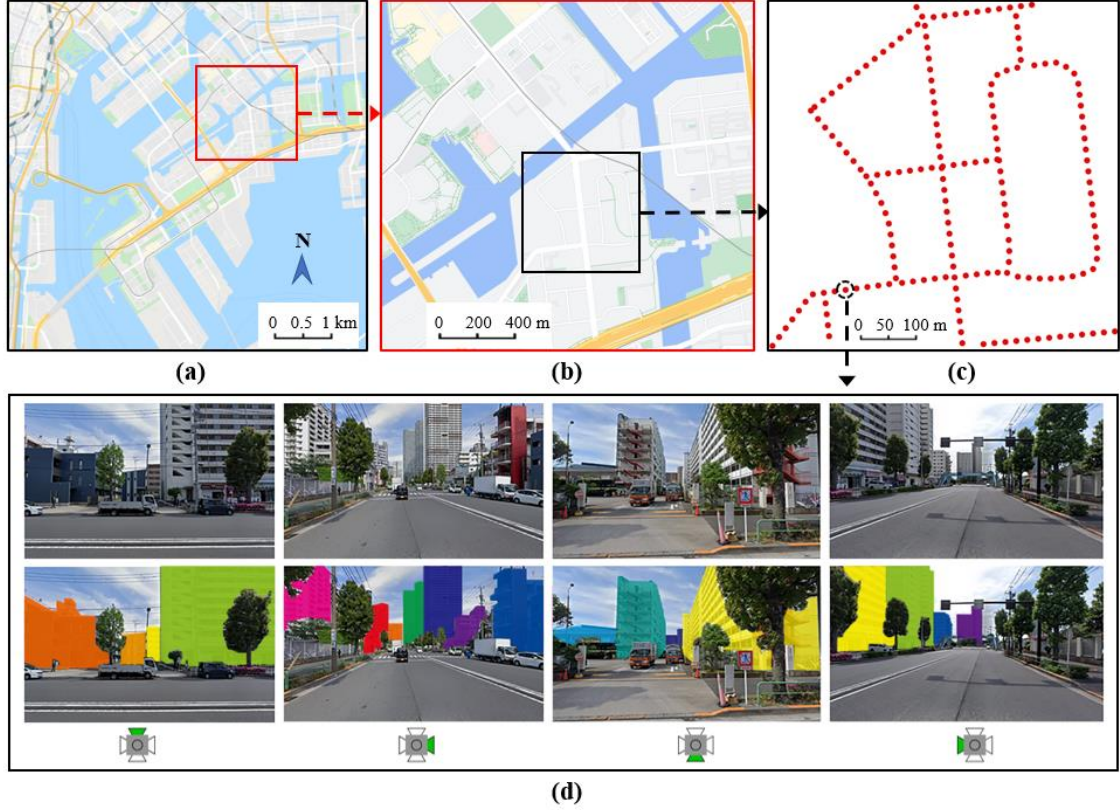
Figure 4.9. Workflow for collecting real-world, street-level images and building annotations. (a) Study area in OSM, (b) example of a randomly selected area with a road network, (c) sampling point locations along with the road networks, and (d) street-level images with building instance annotations that were manually applied.

To compare the differences between the real-world data and synthetic data in supervised instance segmentation, the real street-view data need to be from the same area as the virtual city scene. In this study, the Google Maps API is used to obtain real street-level images of the study area. First, the road network was traversed in OSM (OpenStreetMap, 2021), and sampling points were taken at 20-m intervals to validate scenario differences (J. Zhang et al., 2021b). Figure 9c shows the sampled points on a

road network, where randomly selected areas have been zoomed in. Then, Google Maps API is used to obtain street-view images in four different directions at each sampling point, with the point's latitude and longitude coordinates. The street-view images were set to 1280 × 800, and the vertical FOV was set to 79 degrees (same as in the synthetic dataset). Then, the street-view images were selected as the training set, validation set, and test set for the building instance segmentation. Finally, the LabelMe tool is used (Russell et al., 2008) to manually label the facade instance annotations in the street-view images.

(3) Mixing CDT synthetic data and real data into the hybrid dataset

The proposed HSRBFIA dataset contains a mixture of building images and facade instance annotations from the CDT synthetic data and the real-world data, with 2,000 real and 2,000 CDT synthetic images. A scalable hybrid subset, HSRBFIA-x, can be constructed from HSRBFIA. HSRBFIA-x is used for training, 400 real images from HSRBFIA are used for testing, and 400 real images from HSRBFIA are used for validation. In the subset HSRBFIA-x, the total number of images is 1200, and the proportion of them that are real and CDT synthetic images is calculated according to

$$a = x\% \times 1200$$

$$b = (100 - x)\% \times 1200,$$

(4.1)

where a represents the number of real images, b the number of CDT synthetic images, and x is the percentage of real data in HSRBFIA-x. For example, HSRBFIA-40 indicates that the portion of real-world data is 40%, which means the hybrid dataset comprises 480 real images and 720 CDT synthetic images.

(4) Baseline strategy for generating virtual synthetic data

A baseline strategy is presented for generating virtual synthetic data and comparing its generation performance with that of the proposed CDT dataset to definitively show the improvement of the CDT synthetic data as an enhanced training

set. Fictitious LOD2 city models are chosen for the generation system. The LOD2 city models contained 283 buildings with a total area of 2.37 square kilometers. These buildings were of different types and included residential, commercial, office, and industrial buildings. The virtual synthetic data were generated by the same automatic system as the CDT previously, and all of the rendering parameters were set identically to those in the CDT to ensure fairness. Figure 4.10 shows an example of the virtual synthetic data completed with all four views and the building facade annotations. The virtual models were textured in high quality but had no counterparts in the real world.
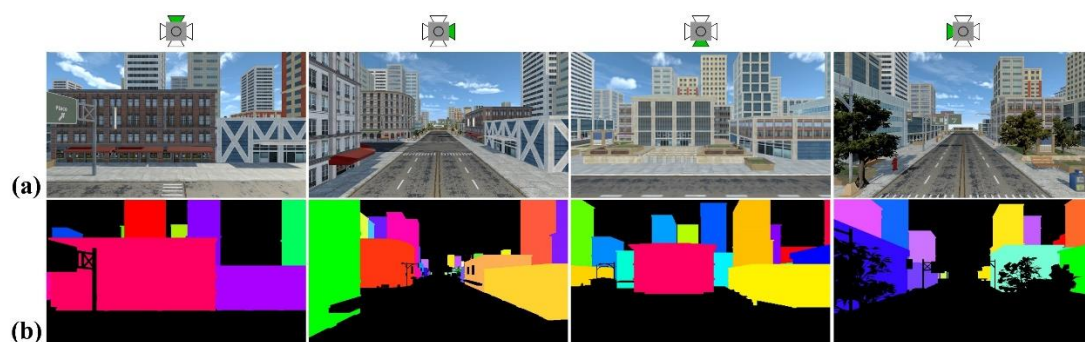
Figure 4.10. Four views from a single shot captured by the multi-camera system for virtual synthetic data (no real-world counterparts): (a) virtual street views and (b) corresponding instance segmentation masks.

4.2.3.2. *State-of-the-art instance segmentation models*

Instance segmentation combines the functionalities of both semantic segmentation and object detection to classify different labels and separate instances of objects belonging to the same class. Current instance segmentation techniques typically have the following four frameworks (Hafiz & Bhat, 2020): the classification of mask proposals (Girshick, 2015), detection followed by segmentation (He et al., 2017), labeling pixels followed by clustering (Bai & Urtasun, 2017), and dense sliding window methods (X. Chen et al., 2019). This study selected three state-of-the-art models: Mask R-CNN (He et al., 2017), YOLACT (Bolya et al., 2019), and BlendMask (H. Chen et al., 2020). After testing them with the proposed dataset, the model with the highest accuracy can be recommended.

The backbone structure has a significant impact on the instance segmentation model. Due to computational power, it can vary depending on the desired performance, training speed, and limitations. The most used backbone structures are ResNet (He et al., 2016) and its variants (S. Xie et al., 2017) combined with the Feature Pyramid Network (FPN) (Lin et al., 2017). For object instance segmentation tasks in complex scenes, increasing the number of convolutional network layers often produces improved accuracy (Carvalho et al., 2020).

The same training protocol is adopted for every selected model: (1) 36,000 iterations, optimizing tracking validation loss to a convergence point to avoid overfitting; (2) four pictures per batch (it is worth noting that training with hybrid datasets is conducted with a mixed batch of two real and two synthetic pictures.); and (3) the Adam optimizer starting with a learning rate of 0.001 that is reduced to 0.0002 after 10k iterations.

4.2.4   Accuracy analysis

Accuracy analysis of the instance segmentation model allows for insight into its applicability in the real world. Specifically, analyzing the accuracy of the HSRBFIA dataset in testing real-world street-level images allows us to evaluate the performance of synthetic data in augmented DCNN training. The three primary metrics used to assess the instance segmentation performance are precision, recall, and the intersection over union (IoU), and the equations are given by

$$precision = \frac{TP}{TP+FP}, \tag{4.2}$$

$$recall = \frac{TP}{TP+FN}, \tag{4.3}$$

$$IoU = \frac{TP}{TP+FP+FN}. \tag{4.4}$$

For a given category, a true positive (TP) is the number of correctly identified positive pixels, a false positive (FP) is the number of pixels that are mistakenly classified, and a false negative (FN) is the number of pixels that are not classified as belonging to this category but should have been. While precision and recall provide great insight into the data, the threshold cutoffs are equally crucial for evaluating instance segmentation models. The IoU of the bounding boxes is considered when calculating the threshold value.

There are many standard COCO metrics for evaluating the object detection and segmentation performance of instance segmentation tasks with different considerations (Lin et al., 2014). For building instance segmentation with street-level scenes, three aspects are important: (1) the detected precision, (2) the size of the detections, and (3) the number of detections in each image. This study chose several metrics according to specific requests, including (a) the average precision (AP), (b) AP_{50}, (c) AP_{75}, (d) AP_{medium}, (e) AP_{large}, and (f) the Average Recall (AR) with ten maximum detections (AR_{10}). For the detected precision, the AP uses the mean value from 10 IoU thresholds, starting at 0.5 and going up to 0.95 with steps of size 0.05 (0.50: 0.05: 0.95), and AP_{50} represents the calculation under an IoU threshold of 0.50. Likewise, AP_{75} is a stricter metric and represents the calculation under an IoU threshold of 0.75. The closer the AP is to 1, the better the authenticity of the instance segmentation model will be. As for the size of detections, a 1280 × 800 pixels street-level image contains buildings with a variety of scales. Tiny buildings are rare in our datasets, and thus two categories are selected, AP_{medium} (32^2 pixels < detection area < 96^2 pixels) and AP_{large} (detection area > 96^2 pixels), for consideration while excluding AP_{small} (detection area < 32^2 pixels) [21]. When trying to determine the number of detections in each image, the AR is used because it takes the maximum number of detections into consideration. Since the maximum quantity of buildings for a single street-view image in the dataset is 10, AR_{10} is the appropriate evaluation metric.

4.3 Experiments and results

This section describes the time cost on synthetic and real image annotations and the experiments to verify the accuracy of facade instance segmentation on real street-level imagery using the HSRBFIA dataset.

The software and hardware environment configurations used to develop and execute all the experiments are listed in Tables 4.3 and 4.4.

Table 4.3. Software and libraries.

Software	Details
Operating System	Ubuntu 16.04 64 bit
Programming language	Python
Deep learning framework	PyTorch
Dependent library	Torch, Torchvision, CUDA, PIL etc.
Game engine	Unity 2020.3.12f1 with universal render pipeline
Labeling tool	LabelMe

Table 4.4. Hardware.

Content	Appellation
CPU	Intel Core i7-9700 @3.00GHz
RAM	DDR4-2666 16GB $\times$ 2
GPU	NVIDIA GeForce RTX 2070 SUPER 8GB $\times$ 2
Graphics tablet for manual labeling	Wacom Intuos Pro

4.3.1 Time cost results of data annotation

To calculate the time needed for annotating CDT synthetic data, the proposed system is used with a 3D city model of Tokyo to automatically generate 2,000 synthetic street-level images and facade annotations. Then, the total time is recorded and calculated.

To calculate the time needed for labeling real data, four graduate students are invited from aged 23-28 years with architectural design backgrounds to manually label

real data that were randomly selected. The same 100 images (size 1280 × 800 pixels), labeling software (LabelMe), and labeling device (graphics tablet, Wacom Intuo Pro) were used. Then, the total time that they spent was recorded, and the average labeling time per image was calculated. Table 4.5 compares the time required per image to annotate the synthetic and real-world datasets. It is worth noting that the annotation time for the real data is approximately 2,050 times greater than that of the automatically generated synthetic data.

Table 4.5. Time consumption of synthetic and real datasets for each image.

Dataset	Contents	Labeling method	Time cost per image (s)
Synthetic	Virtual street views and building instance annotations	Automated	0.12
Real-world	Natural street views and building instance annotations	Manual	246

4.3.2   *Accuracy verification of the proposed datasets*

This study aims to show the potential of the auto-generated synthetic data in improving DCNN-based instance segmentation models trained using real-world imagery. Four experiments are presented that used CDT synthetic data to test this. The first experiment used a baseline strategy for virtual synthetic data and compared its performance with that of the proposed CDT synthetic dataset using several DCNN-based instance segmentation models. The second experiment selected 100 real-world images from HSRBFIA as the training set and then performed extended training using the proposed CDT synthetic data and virtual synthetic data. The third experiment used the same number of training images from HSRBFIA-x. The fourth used a pre-trained model from the HSRBFIA-x dataset to perform facade instance segmentation for street-view images from multiple cities not located in Japan, then validated its accuracy.

4.3.2.1. *Comparison with virtual synthetic data using several instance segmentation models*

The same amount of virtual synthetic, CDT synthetic, and real-world data as the training set are selected for several instance segmentation models (Mask R-CNN with SpineNet-96, Mask R-CNN with ResNet-101-FPN, BlendMask with ResNet-101-FPN, and YOLACT with ResNet-101-FPN), 1200 images each. The pre-trained models were then tested on 400 real-world images, and the corresponding AP values were calculated separately. As shown in Table 4.6, the AP values (both mask and bounding box) for all models show that using the proposed CDT synthetic data as the training set for the instance segmentation of real-world images leads to better performance than when the virtual data are used. However, there is still a performance gap compared with when only real-world data are used. According to the overall accuracies shown in Table 4.6, Mask R-CNN with a SpineNet-96 backbone performs the best, with only the mask AP for virtual synthetic data being inferior to BlendMask with ResNet-101-FPN (0.309 compared with 0.314). In the following experiments, Mask R-CNN is used exclusively with the SpineNet-96 backbone as the training model.

Table 4.6. AP values for the instance segmentation using different datasets when training several state-of-the-art models

	Type	Mask R-CNN (SpineNet-96)	Mask R-CNN (ResNet-101-FPN)	BlendMask (ResNet-101-FPN)	YOLACT (ResNet-101-FPN)
Virtual synthetic data only (baseline)	mask AP	0.309	0.282	0.314	0.227
	bbox AP	0.332	0.314	0.329	0.263
CDT synthetic data only (the proposed)	mask AP	0.415	0.377	0.409	0.312
	bbox AP	0.433	0.405	0.431	0.341
Real-world data only	mask AP	0.591	0.537	0.587	0.432
	bbox AP	0.632	0.559	0.624	0.513

Bbox refers to bounding boxes.

4.3.2.2. Comparing results for training using real data with two types of synthetic

data extensions

The second experiment compared the precision results of the COCO metrics. Only 100 real images were used for training, and the training set was then extended using virtual synthetic images and the proposed CDT synthetic images. The pre-trained models were all tested on the initial batch of 100 real images, and the results for this are shown in Table 4.7. Comparing the results for the two synthetic extensions, the inclusion of CDT synthetic data provides a greater improvement to the accuracies of mask segmentation and bounding box detection than the virtual one when using the same amount of data. AP_{medium} is the metric that benefited most when the training set was extended to include 1000 synthetic data elements, and this was true for both types of synthetic data. The accuracy of AP_{medium} increased by 17.7% (Mask) and 19.4% (Bounding box) relative to the baseline when the virtual data were included and by 24.8% (Mask) and 23.7% (Bounding box) when the CDT data were included.

Table 4.7. Results from training facade instance segmentation on real-world images only and from extending the training sets with virtual synthetic and CDT synthetic images. The improvements, as compared with the baseline (training only with real data), are highlighted in bold.

Training sets (number of images)	Type	AP	AP_{50}	AP_{75}	AP_{medium}	AP_{large}	AR_{10}
100 (R)	Mask	0.366	0.574	0.398	0.124	0.428	0.431
	Bbox	0.382	0.587	0.399	0.139	0.437	0.425
100 (R) + 100 ($S_{virtual}$)	Mask	0.392 (2.6%)	0.629 (5.5%)	0.436 (3.8%)	0.177 (5.3%)	0.454 (2.6%)	0.462 (3.1%)
	Bbox	0.411 (2.9%)	0.647 (6.0%)	0.442 (4.3%)	0.211 (7.2%)	0.454 (1.7%)	0.438 (1.3%)
100 (R) + 100 (S_{CDT})	Mask	0.398 (3.2%)	0.638 (6.4%)	0.442 (4.4%)	0.181 (5.7%)	0.461 (3.3%)	0.466 (3.5%)
	Bbox	0.415 (3.3%)	0.652 (6.5%)	0.449 (5.0%)	0.227 (8.8%)	0.455 (1.8%)	0.441 (1.6%)
100 (R) + 500 ($S_{virtual}$)	Mask	0.430 (6.4%)	0.692 (11.8%)	0.521 (12.3%)	0.259 (13.5%)	0.522 (9.4%)	0.517 (8.6%)
	Bbox	0.453 (7.1%)	0.680 (9.3%)	0.531 (13.2%)	0.308 (16.9%)	0.514 (7.7%)	0.526 (10.1%)
100 (R) +	Mask	0.473	0.705	0.538	0.302	0.543	0.548

|							
500 (S_{CDT})		(10.7%)	(13.1%)	(14.0%)	(17.8%)	(11.5%)	(11.7%)
	Bbox	0.485	0.696	0.543	0.334	0.531	0.551
		(10.3%)	(10.9%)	(14.4%)	(19.5%)	(9.4%)	(12.6%)
100 (R) +	Mask	0.483	0.702	0.550	0.301	0.539	0.546
1,000		(11.7%)	(12.8%)	(15.2%)	(17.7%)	(11.1%)	(11.5%)
($S_{virtual}$)	Bbox	0.493	0.720	0.575	0.333	0.555	0.554
		(11.1%)	(13.3%)	(17.6%)	(19.4%)	(11.8%)	(12.9%)
100 (R) +	Mask	0.511	0.728	0.604	0.372	0.576	0.579
1,000 (S_{CDT})		(14.5%)	(15.4%)	(20.6%)	(24.8%)	(14.8%)	(14.8%)
	Bbox	0.535	0.743	0.613	0.376	0.581	0.576
		(15.3%)	(15.6%)	(21.4%)	(23.7%)	(14.4%)	(15.1%)

Figure 4.11. Qualitative results for training real datasets only and for extending them with the two types of synthetic datasets (CDT and virtual).

Figure 4.11 shows qualitative results that demonstrate how using the two types of synthetic data during training improves the ability of the system to recognize individual building facades in realistic scenarios. The results obtained using only the 100 real images as the training set were not ideal, but with the inclusion of CDT or virtual synthetic data, the ability to perform detection and mask segmentation of building facades was improved, especially for small targets. When it comes to performing instance segmentation for real images that have smaller structures and partial facades, a pre-trained model using a dataset that has been augmented with CDT data can obtain better results than one that has been augmented with virtual synthetic data.

4.3.2.3. *Comparison of results for different ratios of HSRBFIA-x*

For the training sets, HSRBFIA-x is selected with different ratios of CDT synthetic data and real data. The results for the COCO metrics are listed in Figure 4.12 and were obtained by using 400 natural street-view images from the downtown areas of Tokyo. In Figure 4.12, the horizontal axis denotes the proportions of real data that were used in the HSRBFIA-x datasets, and the vertical axis gives the values for the COCO metrics. Observing the overall trend exhibited by all of the line graphs, the testing results using synthetic data alone for the training set are the worst among the COCO metrics. In addition, as the proportion of real-world data in the HSRBFIA-x dataset is increased, the metric precision first grows substantially, then becomes gradual, and finally stagnates. Taking the AP as an example, when the proportion of real data in the training set reaches 60%, the bounding box detection result is 0.611, and the mask segmentation result is 0.577, achieving 96.7% and 97.6% of the results using 100% real data for the training set (Figure 4.12a). For the growth of the metric precision after switching from 100% synthetic data to 100% real data, the improvement is minimal for mask segmentation and bounding box detection when AP_{50} is used in the analysis, coming out to be 16.1% and 16.2%, respectively (Figure 4.12b). However, the difference is significant for AP_{medium}. The accuracy of the mask segmentation and bounding box detection increased by 26.2% and 25.1%, respectively (Figure 12d).

Figure 4.13 shows the instance segmentation of several building types using HSRBFIA-x datasets with varying ratios of synthetic to real data for the training sets. The results from this qualitative analysis are similar overall to the results of the quantitative analysis shown in Figure 4.12. In street-level images, some buildings are located far away from the camera. As a result, they appear small in the images and tend to vanish during the down-sampling process (Figure 4.13c).
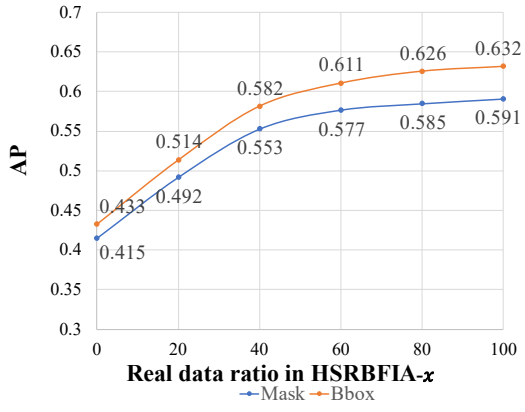
(a) AP results on real data
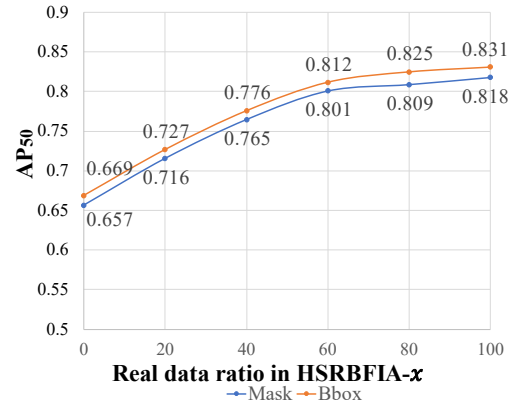
(b) AP$_{50}$ results on real data
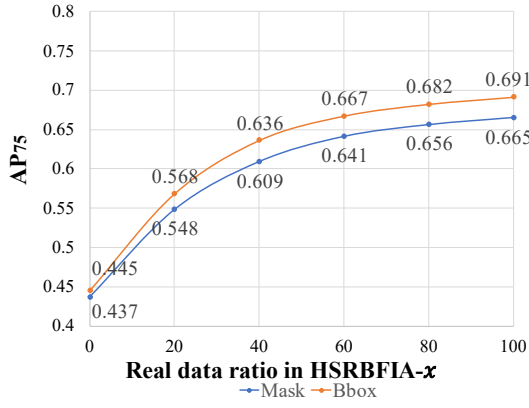
(c) AP$_{75}$ results on real data
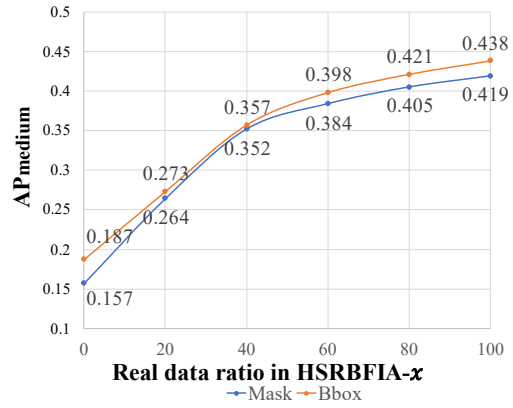
(d) AP$_{medium}$ results on real data
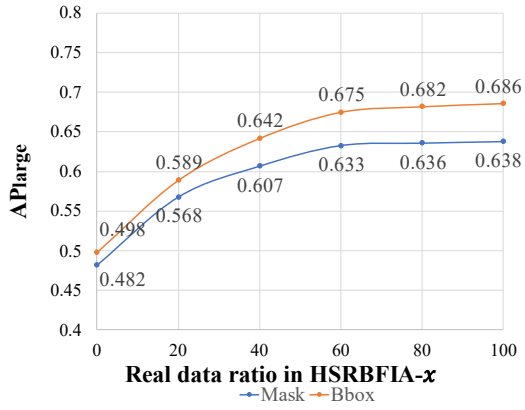
(e) AP$_{large}$ results on real data
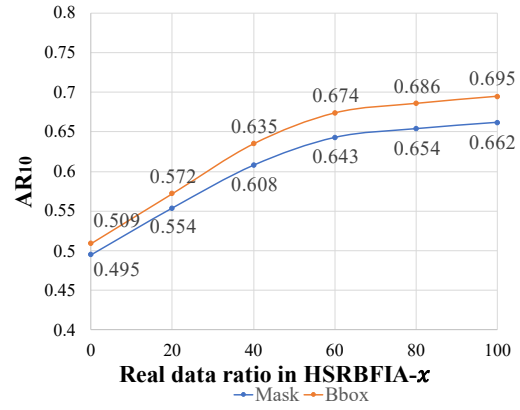
(f) AR$_{10}$ results on real data

Figure 4.12. Comparison of the results for COCO metrics precision on real images. HSRBFIA-x datasets with differing ratios of real data were used for the training set: (a) AP, (b) AP$_{50}$, (c) AP$_{75}$, (d) AP$_{medium}$, (e) AP$_{large}$, and (f) AR$_{10}$.

Figure 4.13. Qualitative results for different building categories from training HSRBFIA-x datasets with different ratios of synthetic to real data: (a) traditional Japanese houses, (b) mu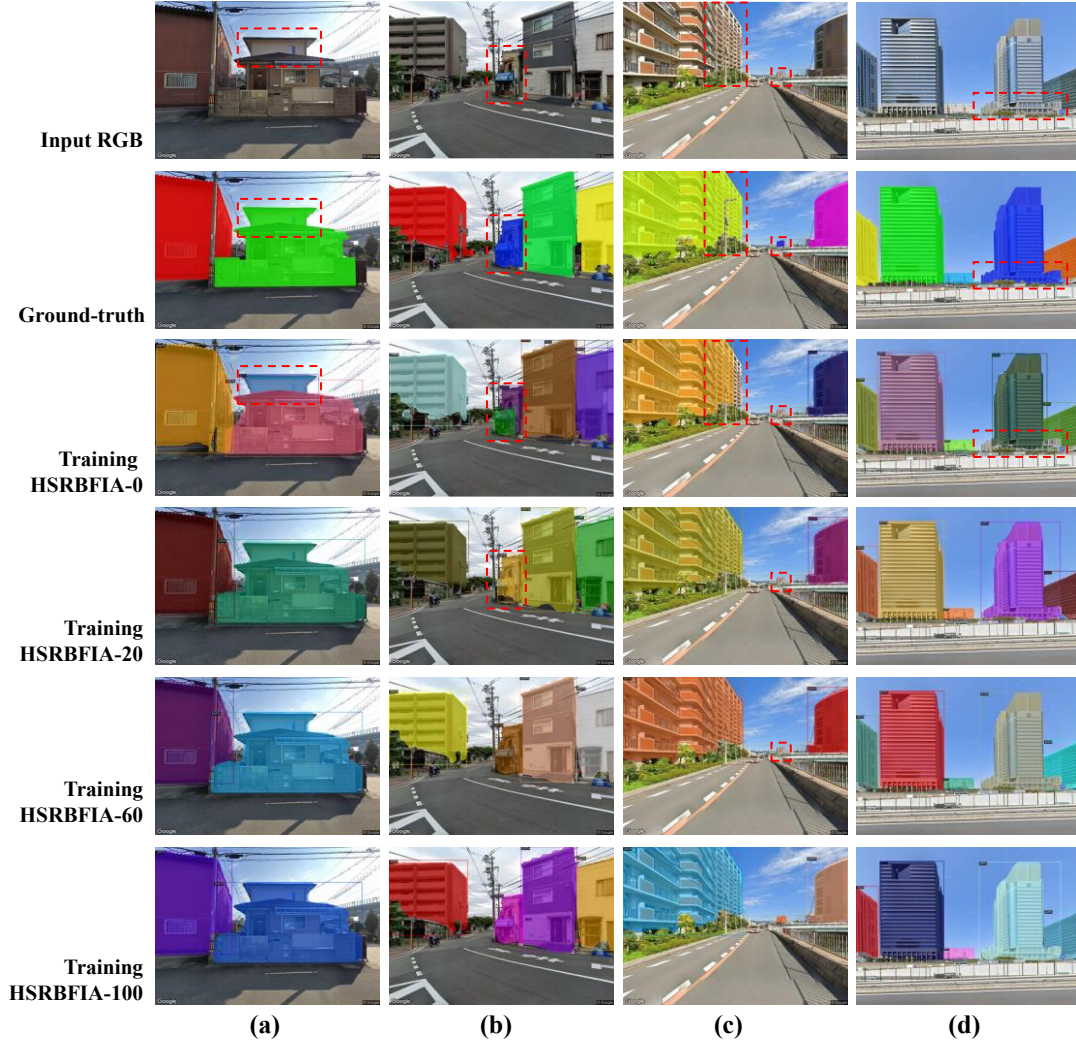lti-story residential buildings, (c) apartments, and (d) public high-rise buildings. (The red dashed rectangles highlight parts of the natural street-level images that are prone to failure during facade instance segmentation.)

4.3.2.4. Verification on other cities

As a comparison, the HSRBFIA-0, HSRBFIA-60, and HSRBFIA-100 that are employed in Section 4.4.2.3 are re-used as the training set, and a total of 400 street view images from four cities, including Osaka, Japan; Los Angeles (L.A.), US; New York City (NYC), US; and Shanghai, China, were used as the test set. The test images are downloaded through Street View services, covering a wide range of building types and sizes, including residential, office, commercial, and industrial, as well as low-rise, high-

rise, and complex buildings. Table 4.8 lists the training set information and COCO metrics results for facade instance segmentation evaluated by the pre-trained model on real-world images in multiple cities. The worst segmentation is obtained with entirely synthetic data (HSRBFIA-0) as the training set. The accuracy of 60% real data (HSRBFIA-60) is close to 100% real data (HSRBFIA-100) as the training set. The results verified in these cities are similar to those presented in Section 4.4.2.3 for Tokyo and show that the proposed training dataset HSRBFIA-x generalizes well to street views in different cities without further fine-tuning.

Table 4.8. COCO metrics precision of facade instance segmentation with training the proposed dataset HSRBFIA-x in multiple cities.

Cities	Type	HSRBFIA-0			HSRBFIA-60			HSRBFIA-100		
		AP	AP_{50}	AP_m	AP	AP_{50}	AP_m	AP	AP_{50}	AP_m
Osaka	Mask	0.412	0.641	0.152	0.569	0.791	0.378	0.584	0.812	0.411
	Bbox	0.437	0.663	0.181	0.607	0.803	0.393	0.616	0.825	0.423
L.A.	Mask	0.403	0.607	0.139	0.541	0.762	0.362	0.569	0.798	0.395
	Bbox	0.411	0.631	0.172	0.583	0.774	0.381	0.583	0.809	0.404
NYC	Mask	0.388	0.572	0.121	0.544	0.727	0.331	0.536	0.751	0.351
	Bbox	0.402	0.603	0.148	0.583	0.741	0.348	0.551	0.767	0.357
Shanghai	Mask	0.371	0.538	0.102	0.527	0.701	0.316	0.517	0.718	0.328
	Bbox	0.377	0.553	0.124	0.553	0.712	0.323	0.543	0.735	0.335

The training set of 400 street view images includes 100 images of Osaka, Japan; 100 images of Los Angeles (L. A.), US; 100 images of New York City (NYC), US; and 100 images of Shanghai, China.

In natural street scenes, the residential architectural styles in American and Chinese cities differ significantly from those in Japan, but most public building styles are similar. Figure 4.14 shows the qualitative results of the building segmentation of different building types for each city, and the red dashed rectangle is used to highlight some parts of the street view images that are easy to fail in facade instance segmentation, such as buildings far from the camera and complicated facade compositions. It was observed that training on synthetic data generated by CDT was sufficient to recognize

low-rise, high-rise, and integrated buildings. The combination of real and synthetic data (HSRBFIA-60) then yielded high precision results for non-constructed items (Figure 4.14a), small buildings (Figure 4.14b and 4.14c), and even complex facades (Figure 4.14d).
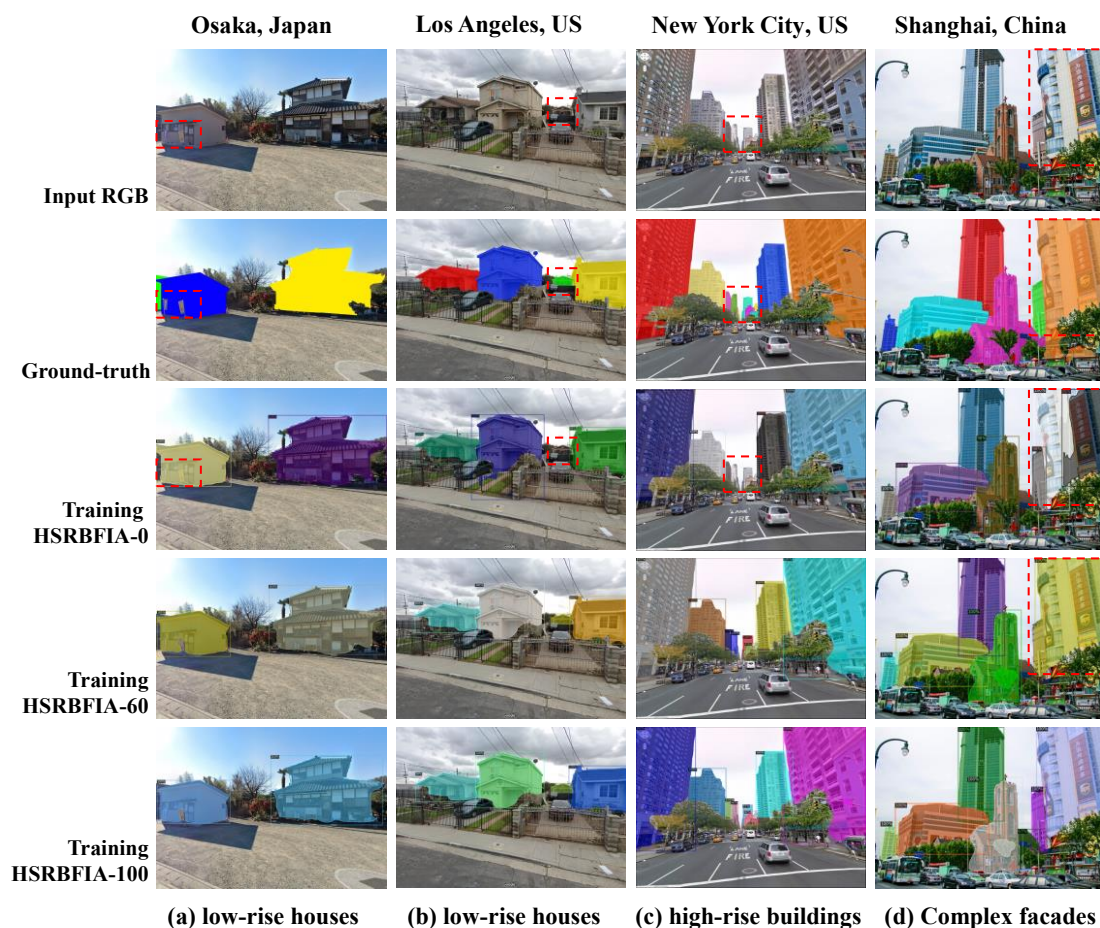

| | Osaka, Japan | Los Angeles, US | New York City, US | Shanghai, China |

Input RGB

Ground-truth

Training HSRBFIA-0

Training HSRBFIA-60

Training HSRBFIA-100

(a) low-rise houses   **(b) low-rise houses**   **(c) high-rise buildings**   **(d) Complex facades**

Figure 4.14. Qualitative results for different types and sizes of buildings with training different synthetic-real ratios of HSRBFIA-x datasets. (a) Low-rise houses in Osaka, Japan; (b) low-rise houses in Los Angeles, US; (c) high-rise houses in New York City, US; (d) Complex facades in Shanghai, China. (The red dashed rectangles on the images highlight some parts of the street view images that are easy to fail in facade instance segmentation.)

4.4   Discussion

4.4.1   *Automatic generation of instance annotation for building facades based on CDT*

This Chapter investigated the possibility of using auto-generated synthetic datasets from CDT to boost the DCNN-based instance segmentation accuracy on real street-level imagery. The results demonstrated that models trained by adding a fraction of real data to synthetic datasets could obtain results comparable to models built with real datasets. Furthermore, using synthetic data for extending the training set on real data can improve the segmentation accuracy. These findings are significant because they offer the possibility that the synthetic data from CDT can be used as an alternative to real data for training supervised learning-based models, which will significantly slash the cost of data annotation.

A hybrid dataset HSRBFIA on building facade images and annotations are built for improving the facade instance segmentation on real images. Several large-scale collections of synthetic datasets in the virtual city have previously been used for semantic segmentation of streetscape elements (Ros, Sellart, et al., 2016; Saleh et al., 2018). However, the effectiveness of synthetic data generated from real-world copies for training DCNNs could not be concluded in these studies. Since they all use datasets created from virtual cities (no correspondence with the natural world) as training sets, the texture gap between virtuality and reality is more likely to cause domain shift problems and lead to poor performance (Tremblay et al., 2018). In addition, instance segmentation is more complex than semantic segmentation when annotating synthetic datasets, distinguishing both semantics and instances. Compared to previous studies based on virtual cities, the building digital assets with high LODs from CDT have high-quality textures that emulate the physical counterparts in the virtual space, producing more realistic synthetic data.

4.4.2   *Effective use of CDT data for street-level facade instance segmentation*

From the results discussed in Section 4.4.2.1, it appears that CDT synthetic data outperform virtual data when synthetic data are used for the training set and real images for the test set. This may be because the 3D model textures in the CDT closely resemble those of the real building images in the test set. Also, the feature map extracted by the CNN model from a CDT synthetic image is similar to that extracted from the real image.

From the experimental results discussed in Section 4.4.2.2, extending the real-world training set by adding CDT and virtual synthetic data can improve the instance segmentation accuracy. This result is similar in some ways to previous work on the semantic segmentation of urban scenes based on fictional cities (Prakash et al., 2019; Ros, Sellart, et al., 2016; Saleh et al., 2018), but the proposed CDT synthetic data outperform the virtual data when used in an enhanced training set for instance segmentation tasks, and works for the street-level building category. Despite this benefit, the value of AP_{medium} was the lowest among all of the results for the COCO metrics, indicating that the detection of smaller buildings remains a challenge, particularly when only a small fraction of the real data in the training sets account for them. This could be due to the fact that the semantic information for small objects appears in the shallower feature maps, and their details may vanish entirely as the network gets deeper. In addition, real images have more variation in texture, shape, and color compared with synthetic images. When the fraction of real images in a dataset is limited, it is difficult to transfer learned weights to the synthetic dataset.

From the results in Section 4.4.2.3, it appears that extending the proposed synthetic data into the real-world training set can improve the instance segmentation accuracy. This result is partly similar to previous reports on semantic segmentation of urban scenes based on virtual city data (Prakash et al., 2019; Ros, Sellart, et al., 2016; Saleh et al., 2018), but the attempt is a complement to instance segmentation task using synthetic data of CDT as the training set, and it works in the street-level building

category. Despite the benefit, the value of AP_{medium} is the lowest among the metrics results of all experiments, which indicates that the small building images are challenging to be detected, particularly when the real data in training sets account for a small fraction. This could be due to the following reasons: the semantic information of small objects appears in the shallower feature maps, and their details may vanish entirely as the network gets deeper. In addition, real images have higher variation in texture, shape, and color than synthetic images. When the dataset has a limited fraction of real images, the learned weights are difficult to be transferred to the synthetic dataset.

The test results for multiple cities can show the robustness and transferability of the proposed HSRBFIA. The instance segmentation results for street-level images of Osaka, which has similar architectural styles as Tokyo, showed the best results in the experimental cities. This means that the prediction results can be satisfactory based on resembling feature distributions. In the test results of two US cities, the pre-trained model works well for low-rise residential and high-rise buildings with modern style. This could be because the buildings in the test streetscape are non-dense and clearly separated, showing the ability of the proposed method to handle simple scenes in other cities. In contrast, the building instance segmentation results for street-level images of Shanghai under three training sets (synthetic dataset, HSRBFIA-60, real-world dataset) are relatively less accurate than the other three cities. This may be because the test images are heavily sourced from public buildings in high-density urban areas, and the buildings have appearance gaps from the training set. It is difficult to obtain satisfactory results when the detection features of the training and test set are widely divergent, even using a large amount of synthetic data and then fine-tuning it with real images.

4.4.3   Limitations

The holistic goal of this research is to implement an automatic, scalable, and high-fidelity synthetic data generation system for urban scenes. The system will largely contribute towards reducing manual labeling costs involving built environment data for

supervised machine learning. The proposed approach has two limitations that need to be pursued in future work. One is to synthesize virtual data with realism by enhancing the rendering CDT model while auto-generating annotations of various elements. The other is to improve the efficient use of CDT data to train DCNN-based models.

Given the nature of current CDTs 3D reconstruction, with LiDAR data and visible-light photography capture, photorealistic virtual images can be rendered using fine-grained 3D models with subdivision materials. Furthermore, new rendering techniques, such as physical-based rendering, can be integrated into the system to improve illumination effectiveness, bringing the lighting in virtual data rendering closer to the natural environment.

Domain shifts and the loss of small buildings in the down-sampling process are the main issues that impact the use of synthetic data for training instance segmentation models in urban scenes. By systematically investigating the mechanisms at play, the efficiency of synthetic data utilization can be optimized. It has been shown that domain adaptation can transfer the knowledge learned by machine learning models in the source domain (synthetic data) to the target domain (real data), which could be incorporated into the method. Moreover, the recent emergence of deep learning-based methods for small target detection will also be considered in further work.

4.5   Summary of this Chapter

The extraction of building facade data is integral to the construction of information infrastructure. Compared with semantic segmentation, instance segmentation can distinguish individual facades when acquiring and analyzing building information. However, collecting and labeling a large amount of data from the real world for DCNN training to perform accurate instance segmentation of building facades is a labor-intensive process. This Chapter developed a system that can auto-generate synthetic datasets from a CDT for the instance segmentation of building facades. The digital assets of buildings are used in an area of Tokyo as an example. The proposed system

can produce synthetic images of street views from multiple viewpoints under different atmospheric effects. The system can also generate pixel-level instance annotation for synthetic building facades. The general conclusions that can be drawn are as follows.

- Conventional methods for labeling data rely on manual labor. The Chapter attempts to substitute the manual labeling process with an automated generation system to create CDT synthetic data for training DCNNs. The proposed method takes about 1/2,050 of the time that it takes to manually annotate each image, which can significantly reduce the cost required to annotate data.

- By comparing the DCNN training results for real, synthetic, and hybrid datasets, extending the training set with the proposed synthetic data can improve the accuracy of facade instance segmentation on real pictures. A baseline strategy is introduced to show that, at the same LOD and rendering settings, enhancements using CDT synthetic data are better than ones using virtual synthetic data. Specifically, the segmentation accuracy is boosted significantly when a certain fraction of real data is loaded into the CDT synthetic datasets, to the point where its performance becomes competitive with what is seen when 100% real data are used. This indicates that the proposed synthetic dataset has the potential to replace the real imagery in the training set.

- Verification for multiple other cities demonstrated the transferability of the proposed framework. The proposed dataset can obtain promising prediction results for most modern architectural styles. However, the segmentation accuracy needs to be improved for environments that have characteristic architectural styles or high-density streets.

- This study generates synthetic datasets based on a CDT, which effectively utilizes city information modeling and digital assets. As CDTs are further

developed and refined, the research framework can be applied to other elements in the urban environment, which will allow them to enrich their semantic information in the further development of digital infrastructure.

Chapter 5. The large-scale approach for extracting data on multiple elements of building facades

5.1 Overview of building facade information extraction at a large scale

Facade data extraction, or facade parsing, is an important problem in computer vision. The building facade elements are classified, segmented, and 3D reconstructed using computer vision techniques, and then the building facade data can be recorded according to rules (Martinović et al., 2012). Textualization, editorialization, and semanticization of building facade data can store information more efficiently. Large-scale facade data can support urban issues such as urban building energy models (Ferrando et al., 2020), building retrofits (Al-Habaibeh et al., 2021), urban renewal (Zheng et al., 2014), and urban vitality studies (Mouratidis & Poortinga, 2020). Therefore, building facade data collection and digital management have become an important part of developing a smart city.

Generating facade data from spatial data is a key tool to address these challenges. Some geographic open databases or platforms have already achieved remarkable results. For example, OSM covers the plan outline, area, and height of buildings, and Google Earth can observe the mesh models of the world's major cities. However, these urban databases or platforms still have the problem of data integration. In terms of data type, researchers still need to obtain the data necessary for specific urban analysis tasks manually and are not available from open platforms. For example, the dominant color of the facade required for urban color design needs to be measured by professionals on-site (Zhong et al., 2021). Building facade material for building renovation (Piccardo et

al., 2020). Window-to-wall ratio data for building energy and lighting calculations are not automatically available from existing urban databases or platforms (Szcześniak et al., 2022). Non-automatic and semi-automatic methods of data collection are still dominant. If urban building data is to be digitally managed effectively, automated spatial information collection methods must be developed. It will be a significant challenge for industry and academia to develop an automatic framework to efficiently collect and integrate multiple building data, build databases, and use these data effectively for urban management and analysis.

It is necessary to collect, record and analyze data from the surrounding area before specifying the project for implementation. In traditional approaches, work relies on professional surveyors to take on-site measurements of the project, and manual-based workflows can lead to heavy workloads in large-scale remediation projects. Novel workflows have emerged in recent years that combine computer vision techniques with open cityscape datasets. These studies have yielded promising results in parsing building facades. Automating the rapid construction and continuous updating of large-scale building facade databases will help designers and managers control the project throughout the building cycle. This study aims to realize the large-scale automated acquisition of existing building facade data. A toolbox was developed to assist architectural and urban design with current urban development issues. Empowering traditional workflows with digital technology can improve the efficiency of data collection, the reliability of analysis results, and the refinement of management.

This Chapter will integrate the current state-of-the-art technical means to collect multidimensional information on urban facades. Typical facade parsing tasks such as facade color calculation, building function classification, and window-wall semantic segmentation are used as examples to reveal the possibility of large-scale data extraction of urban facades using street view images and deep learning. This chapter is organized as follows. First, data collection, pre-processing, and facade parsing methods are proposed. Then, the proposed methods are integrated with multiple facade data

collection tasks. The experiments were conducted on a street in Osaka, Japan. Finally, the discussion section shows the advantages of the proposed method over the conventional methods. The potential application value and the limitations of the method are presented.

This Chapter integrates the techniques of Chapters 3 and 4 for removing unwanted objects in front of buildings and extracting individual facades of connected buildings. After data pre-processing, facade information is collected, including dominant color calculation of building facades, building functional classification, and semantic segmentation. A street of length 500m in Osaka, Japan, is used as an example to construct a database.

5.2   Methods

This section will describe the technology integration workflow for data extraction of building facades, facade data collection, pre-processing of street view images, and data mining.

5.2.1   *Technology integration workflow for data extraction of building facades*

This Chapter attempts to develop an end-to-end integrated multitasking framework by collecting street-level images at a large scale to extract completed building facades. State-of-the-art computer vision techniques are then used to identify information of the individual, including the facade dominant color, the building functions, and the window-wall semantics. The workflow of this proposed method is as follows.

First, the obscured parts of the building facade will be completed using the GAN-based method. Then, the buildings are extracted one by one using the proposed instance segmentation pre-trained model in Chapter 4. Finally, the information on each building facade will be extracted. Figure 5.1 shows a street-level image including only one

building. The system can remove the trees and cyclists in front of the building. After that, the building facade data are extracted, the facade dominant color is calculated (7.5Y6.5/3.2 in Munsell color system), the building functions (Public service) are automatically identified, and the window-wall semantics are segmented. Figure 5.2 shows a street-level image that includes several buildings that visually overlap. The system can remove the tree, cars, and pedestrians in front of the building. The individual building can be separated by instance segmentation. After that, the individual building facade data are extracted, including window-wall semantics of a single facade, the facade dominant color is calculated (10B7.5/1 in Munsell color system), and the building functions (Public service) are automatically identified.
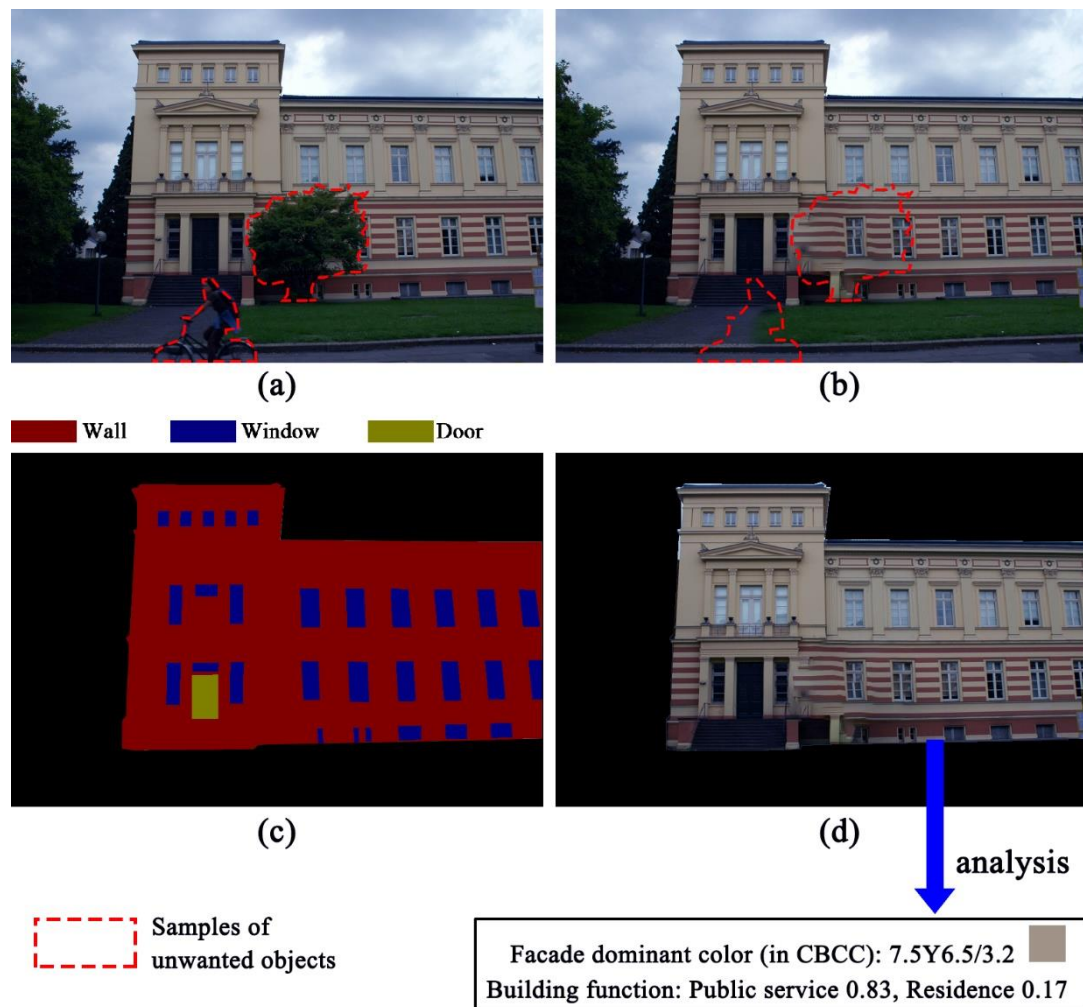
Figure 5.1. Workflow for extracting multiple data in a street view image with one building facade. (a) Data acquisition: original street view image, (b) data pre-processing: complete building facade after color calibration and removal of unwanted objects, (c) data mining: window and wall

semantic extraction in building facade, (d) data mining: facade information extraction, including facade dominant color 7.5Y6.5/3.2 in Munsell color system and building function public service 0.83 confidence.
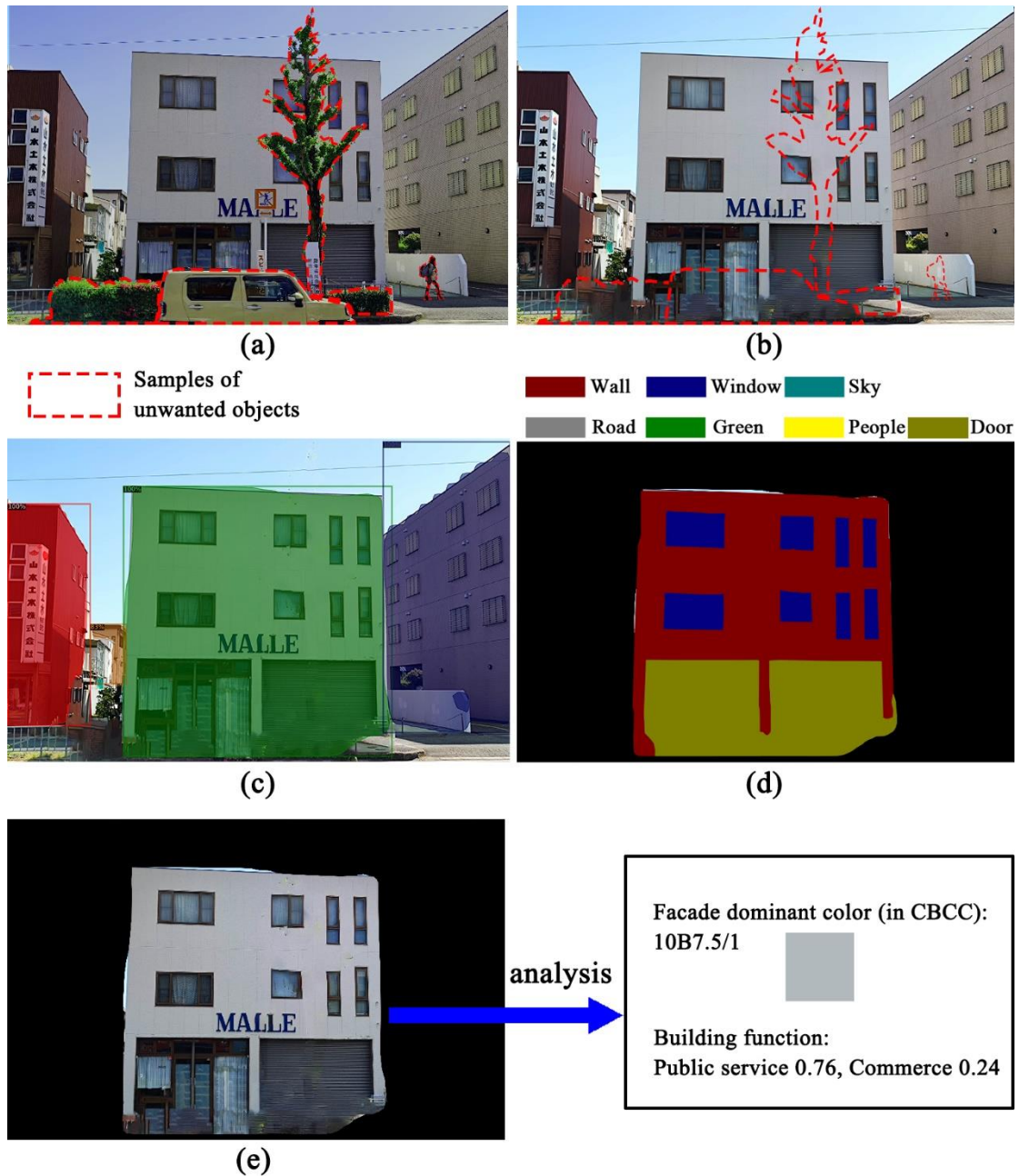
Figure 5.2. Workflow for extracting multiple data in a street image with several building facades visually overlapping. (a) Data acquisition: original street view image, (b) Data pre-processing: complete building facade after color calibration and removal of unwanted objects, (c) Data pre-processing: instance segmentation of building facade, (d) Data mining: semantic segmentation of a single facade, (e) Data mining: facade information extraction, including facade dominant color 10B7.5/1 in Munsell color system and building function public service 0.76 confidence.

5.2.2 *Data acquisition*

The images can be extracted from the street view platform to provide comprehensive coverage of urban streets in street view photos. First, urban road networks with geography coordinate information were chosen and obtained from OSM. The road networks were then simplified into lines. Next, the sampling points with geographical coordinate information can be obtained and shown in spatial distribution. However, it is worth noting that not all sampled points in the street view service have corresponding street view images. Lastly, to obtain the building facade, two pictures (including left and right) are downloaded perpendicular to the road from the street view service (the viewing angle is 90 degrees, the horizontal angle is 0 degrees, image size is 800 × 500 pixels) for each sampling point (as shown in Figure 5.3).

Figure 5.3. Street-level imagery collection at an urban road coordinate.

5.2.3 *Data pre-processing*

5.2.3.1. *Image color calibration*

The color stimulus is significantly influenced by the ambient light. The captured item will appear bluish if the color temperature of the sunlight is cold. A warm temperature light source, on the other hand, will cause the object to appear reddish (Jechow et al., 2020). Since the saturation and brightness of street view photographs are modified by weather and time, the analysis premise is to eliminate the variation caused by ambient light. Previous research has demonstrated that HSV (Hue, Saturation, and Value) color spaces perform better in color calibration than RGB (Red, Green, and

Blue) channels (Mazzeo et al., 2011). Therefore, the collected images are converted to HSV color space. The AWB method was used for the saturation calibration of street view images (Lam et al., 2008). In addition, the AEC of the digital photographs method proposed by Yuan et al. (2012) was introduced to adjust overexposed and overly dark street view images. Figure 5.4 shows the calibration demo by AWB and AEC.

（a）Street view image   （b）After color calibration
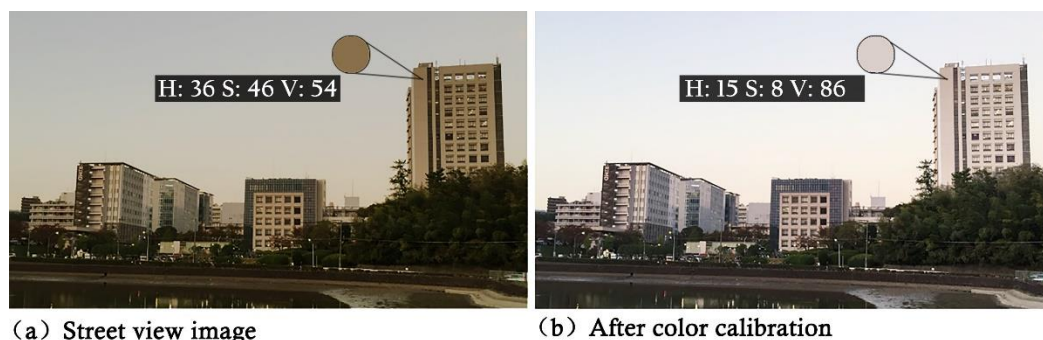
Figure 5.4. A color calibration demo. (a) Ground truth of street view image; (b) color calibration image.

5.2.3.2. *Obstructed facade completion*

In urban environments, there is extensive foreground occlusion of building facades. As described in Chapter 3 of the supplemental methods for the obscured facade, to obtain complete information about the building facade, it is necessary to supplement the obscured part of the facade with rationalities. There are many methods in previous research, which are not repeated here, and can be viewed in section 2.2 of the literature review, and the method was introduced in Chapter 3. This Chapter uses a GAN-based and data-driven inpainting model, DeepFill-v2, for image inpainting, and a custom facade dataset is proposed. Figure 5.5 shows that the proposed system automatically detects people and trees in the foreground of a facade and reasonably filled based on context and data learning. The complete building facade will be fed into the next building information extraction and analysis step.

| (a) Obstructed objects in streets | (b) Unwanted objects removal |
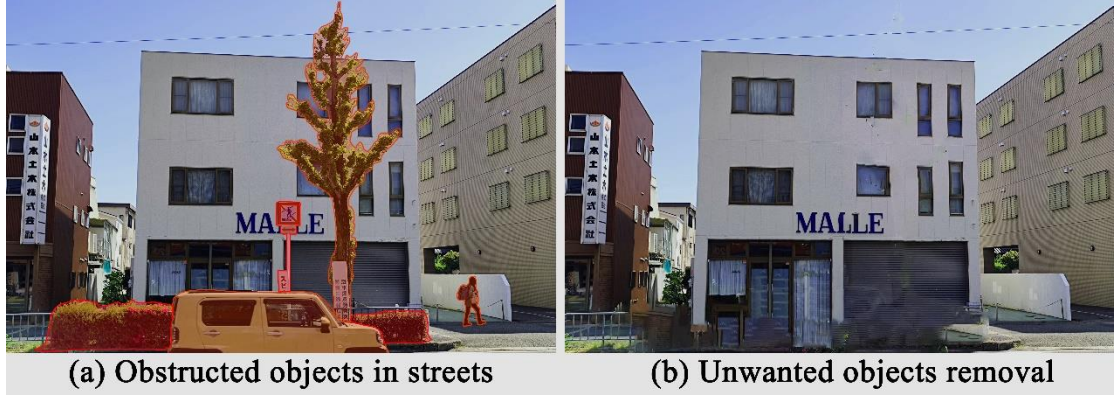
Figure 5.5. An example where people, car, and trees in the foreground of facades are automatically detected by the proposed system and reasonably filled based on context and data learning. (a) Obstructed objects in streets, (b) unwanted objected removal.

5.2.3.3. *Facade instance segmentation*

When multiple buildings are connected or visually overlapping, the study tries to analyze the information of each building and requires the use of instance segmentation. Many methods are proposed in previous research, which is not repeated here, and can be viewed in section 2.3.1 of the literature review. The facade instance segmentation datasets used in this chapter are presented in Chapter 4. Figure 5.6 shows the facade extraction that uses semantic segmentation and instance segmentation, and the instance segmentation can extract building facade information one by one when multiple buildings are connected in a single image.

(a) Ground truth   (b) Facades semantic segmentation   (c) Facades instance segmentation

Figure 5.6. Compared to facade extraction methods that use semantic segmentation, instance segmentation can extract building facade information one by one when multiple buildings are connected in a single image. (a) Ground truth, (b) facade semantics segmentation, (c) facade instance segmentation.

5.2.4   Data mining

5.2.4.1.   facade dominant color calculation

There are thousands of color values in an image, and it is difficult to define the dominant color without merging colors. Therefore, extracting the dominant color of the urban facade requires a standard color card for integrating the colors in the image to the standard color. Since the use of color in architectural design and building decoration should conform to standard color codes in different countries, this study chooses China Building Color Chart (CBCC)-1026 as the standard color (the CBCC-1026 selects 1,026 commonly used architectural colors from the complete CBCC library) that can cover most building colors in urban facades. The specific HSV information of CBCC-1026 can be found in the online color chart (Architectural standard color chart, 2020.). Then, the raw color data of street view images is merged to the standard color chart by calculating the HSV value of the street view color and replacing them with the closest architectural standard color (in terms of the Euclidean distance). In the HSV color space model, the three-dimensional coordinate (x, y, z) of the color point (H, S, V) was defined according to Equation. (5.1):

$$\begin{cases} x = r \cdot v \cdot s \cos h \\ y = r \cdot v \cdot s \sin h \\ z = L(1 - v) \end{cases}, \qquad\qquad (5.1)$$

where r is the radius of the bottom circle, and L is the height, and taking r and L to the integer 100 for the convenience of later analysis. (h, s, v) is the HSV value of the image color. After calculating and merging the distance to the standard color, all colors on the street view images will be converted to the architectural standard color chart. Then, the color proportion from each street view picture can be counted. Although color dominance can be established in several aspects, such as the strength of hue, the sharpness of vision, contrast, and perception of saturation, G. A. Agoston (2013)

suggested that the two most critical factors affecting the dominant color of the picture are the color proportion and the saturation contrast. Therefore, the following is the approach to dominant color selection in this study. (1) The dominant color should be the largest part of the building facade; (2) when the color proportions are equal in a street view picture, the color with high saturation is the dominant color. The facade dominant color open-source tool can be found in (Mortyzhang, 2020/2022).

5.2.4.2. *Multi-Label Classification of Building Function*

From the perspective of the facade in urban streets, there are four main types of building functions in the city proper (Tardioli et al., 2018), including residence (R), commercial service (B), public service (A), and other facilities (O). To effectively classify the types of buildings, a DCNN-based model is conducted to automatically classify the building functions of the study areas. In the previous research, single-label methods have typically been used to classify building classes, with each photo corresponding to only one label (Kang et al., 2018). However, the single-label method cannot accurately separate the street view pictures of several building functions, resulting in inaccurate experimental results. To solve this problem, a multi-label image classification method is used to identify multiple building categories in street view images.

To train the multi-label building classifier, the semantically segmented building images was used firstly to build the corresponding street-view benchmark dataset that contains 4,965 images from 4 basic categories: residential, commercial services, public services, and other facilities. Meanwhile, images with more than one label are classified as mixed services. The ground-truth labels of the training data are from the OSM, and Table 5.1 contains descriptions of the different building function classes. There are around 3,500 single-label images and 1,500 multi-label images in these training images, as shown in Figures 5.7 and Figure 5.8. These street-level images were divided into a training set (75%) and a testing set (25%). It is worth noting that all test images are not

retrieved from a single city and are different from those utilized for training. To augment the training data, 720 × 450 pixels from the original 800 × 500 pixels are randomly selected, and the cropped images are flipped horizontally. Then, several state-of-the-art CNN-based models are trained, including DenseNet (Iandola et al., 2014), EfficientNet (Koonce, 2021), InceptionNet_v4 (Szegedy et al., 2017), and ResNeSt (K.-L. Chen et al., 2021), and demonstrated the corresponding classification performances. To improve the learning rate, these models are trained for 100 epochs and decayed the learning rate by a factor of 0.1 every 25 epochs. Each training batch contained a total of 64 images. Other not mentioned values are default. The building functional classification open-source tool is available online (Mortyzhang, 2021/2022a).

Table 5.1. Description of building class in the city.

Building classifications	Description
Residential (R)	Buildings are for people living, including villas, apartments, and dormitories.
Commercial service (B)	Buildings allow people to engage in various business activities, including retail, shopping malls, markets, hotels, restaurants, and entertainment facilities.
Public services (A)	Buildings allow people to carry out various public activities, including office, education, health, culture, transportation, and tourism buildings.
Other facilities (O)	Buildings or structures that appear in urban areas other than the above three.
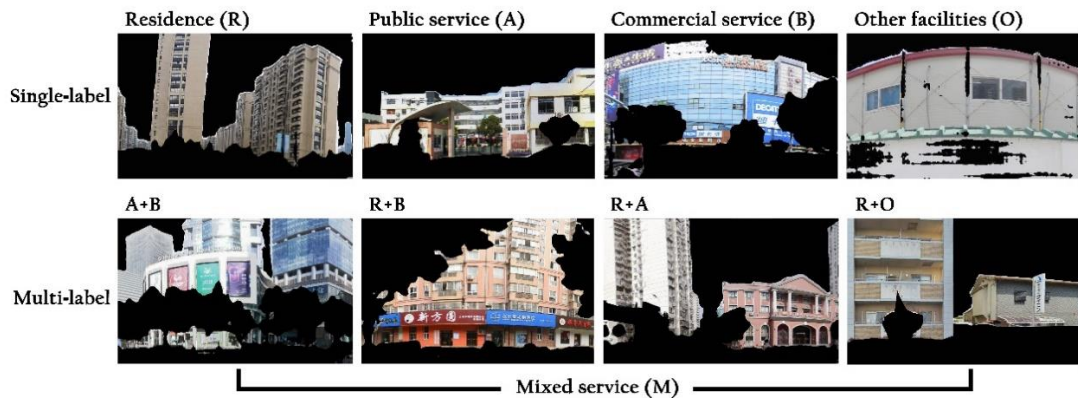
Figure 5.7. The first row is a single-label category, from left to right: Residential, Public, Commerce, and Other Facilities. The second row is a multi-label category, from left to right:

public services and commercial, residential and commercial services, residential and public services, and residential and other facilities. The classification benchmarks have 4,965 street view images with four labels.

Number of Street view images (N=4965)
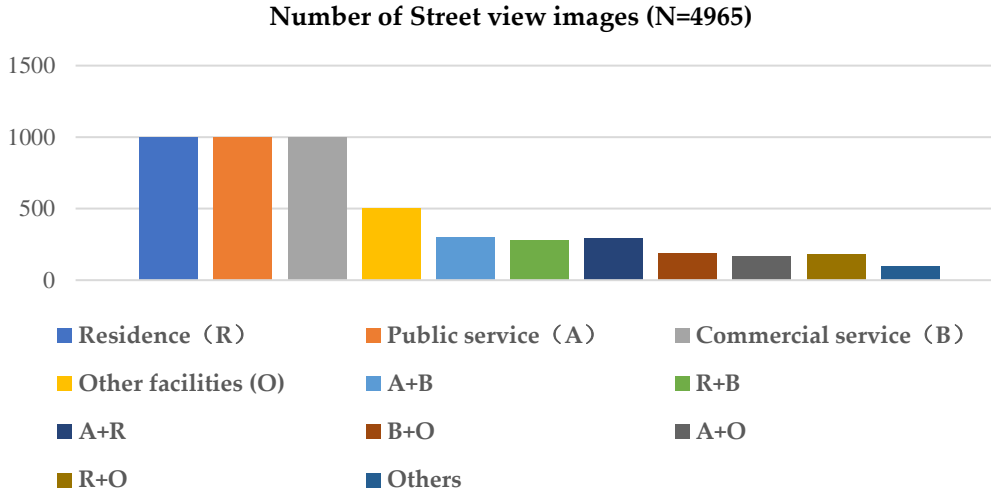
Figure 5.8. The number of training set images for each building category.

5.2.4.3. Semantic segmentation of windows and walls

(1) The categorical semantic segmentation algorithm

The fully convolutional network (FCN) is an early semantic segmentation model that recovers the class to which each pixel belongs from abstract features (Long et al., 2015). The classification pattern of FCN can be extended from image-level to pixel-level compared to traditional methods. CNN-based classification models typically map images to feature vectors by running the convolutional layer output through a fully connected neural network to generate a vector output (X. Li et al., 2019). However, FCNs use convolutional deconvolution layers instead of fully connected layers, and the resolution of the feature map is reduced throughout the feature extraction process (J. Dai et al., 2016). As a result, the downsampling rates (the ratio of the input image resolution to the output feature map resolution) becomes a concern (Tang et al., 2019). Redundant spatial resolution reduction will cause the target object to vanish, whereas

insufficient resolution reduction may result in a model with insufficient translational invariance (M. Dai et al., 2021).

The U-Net model is an FCN-based semantic segmentation model originally developed for medical images (Siddique et al., 2021). The original U-Net consists of an encoder with a standard CNN architecture and an asymmetric decoder that recovers the spatial resolution of the feature maps. Skip connections concatenate feature maps from shrinking paths before doubling the number of feature channels and symmetric feature maps in the expanding path. The symmetric U-Net architecture is more advantageous in handling the facade images with many small objects (Siddique et al., 2021). In addition, building appearances are fixed in structure (windows are located in walls) and not particularly rich in semantic information (the building usually consists of walls, windows, doors, roofs, balconies, etc.). This situation is similar in medical images where U-Net has been found effective (such as human brain structures with fixed positions). Many empirical studies (Du et al., 2020; Esser et al., 2018; Siddique et al., 2021) have shown that skip connection, and U-shaped structures of U-Net can obtain pleasing segmentation results for fixed semantic information.

It has been shown that the upgraded version of U-Net, U-Net++, achieves 75.5% mIoU performance on the Cityscapes val dataset (Zhou et al., 2018), but DeepLabv3+ can achieve 79.6% (L.-C. Chen et al., 2018). Although the Cityscapes val dataset is not based on the wall and window segmentation task, the segmented objects are in the same building environment. DeepLabv3+ can integrate two advantages: one is the spatial pyramidal pooling that encodes multi-scale contextual information, and the other is the encoder-decoder structure that captures clear edge by gradually recovering spatial information. This work will compare the accuracy results of U-Net++ and DeepLabv3+ in the tasks of segmenting walls (commonly referred to as buildings in traditional computer vision datasets) and windows. The model that obtains better accuracy will be recommended.

(2) Semantic dataset for facade parsing

Commonly used open-source facade parsing datasets include eTRIMs (Korc & Förstner, 2009), ECP2011 (Teboul et al., 2011), Graz2012 (Riemenschneider et al., 2012), and CMP2013 datasets (Tyleček & Šára, 2013). The eTRIMs dataset is well diversified, it is built on multi-view images of many European cities, but it has only 60 annotated images. The ECP2011 dataset contains 104 annotated images of Paris in seven categories, including balconies, rooftops, stores, sky, doors, walls, and windows. The Graz2012 dataset consists of 50 images from Germany and Austria. This dataset has only four categories: door, window, wall and sky. The disadvantage of these datasets is that they do not perform well for the training set of images of building facades with widely varying urban styles. The CMP2013 dataset is larger. It has 378 basic images and 228 extended images from around the world. The dataset has a variety of building styles with 12 categories, including wall, molding, cornice, column, window, door, bay window, sash, balcony, store, trim, and background. However, the CMP2013 is a relatively simple dataset of scenes with few foreground occlusions. In general, existing methods based on these publicly available datasets do not perform well in practical applications.

Several recent studies have complemented and enhanced the publicly available datasets. Femiani et al. (2018) built a facade dataset based on street view images in diverse cities. The dataset had only spherical facade photographs in frontal view. It still had limitations related to calibration and single view. LabelMeFacade (Kong & Fan, 2021) extended eTIRMs based on the LabelMe database (Russell et al., 2008) to contain 945 polygonal images with annotations. However, the eTRIMs and LabelMeFacade datasets include buildings, cars, doors, sidewalks, roads, sky, vegetation, and windows. The resolution of these dataset images is below 2k, and small objects such as small windows in the images are faintly represented. In addition, the illumination of the image varies very little. In reality, a considerable portion of the facade images are in a low illumination state. A facade labeled high resolution dataset containing 500 street view

images with 2,048 × 1,152 pixels is proposed, which is higher than any previous publicly available datasets. The dataset contains nine categories, including, sky, wall, window, tree, sign, car, roof, door, and pedestrian. The photos in the dataset are from different weather, and the lighting variations enrich the generalization of the dataset. Figure 5.9 shows examples of facade images in previous datasets and the proposed dataset.
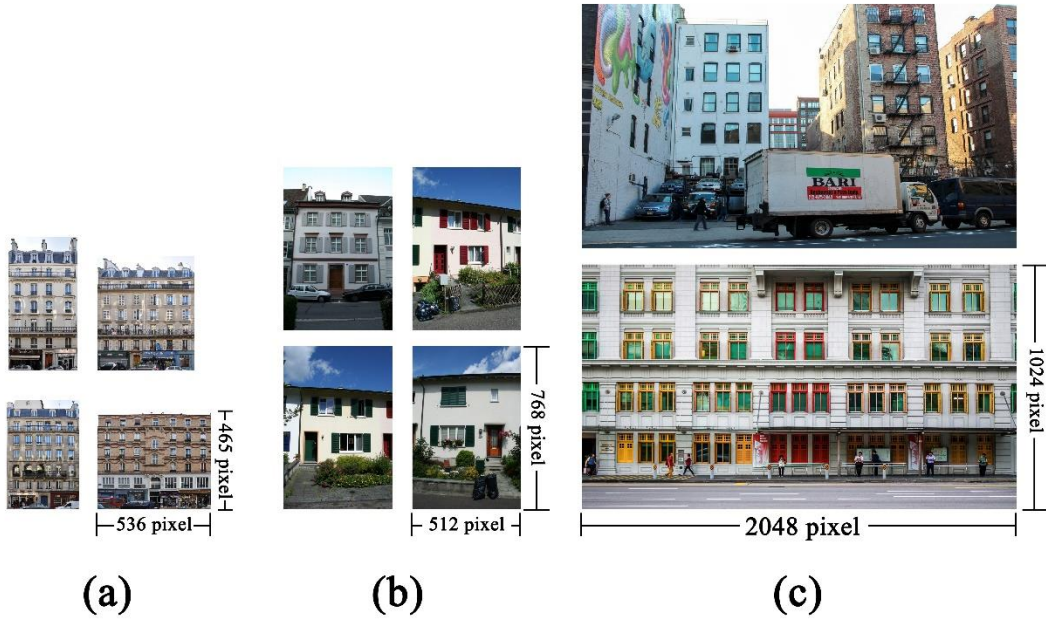
Figure 5.9. Examples of facade images in previous datasets and the proposed dataset. (a) ECP, an open-source facade dataset with the front view buildings. (b) eTRIMS, an open-source facade dataset without complicated obstacles. (c) Proposed datasets, high resolution (2048 × 1152) with diverse scenes.

5.3   Results

5.3.1   *Accuracy verification of facade color calculation*

Two materials (MAT. 1 is ceramic tiles, and MAT. 2 is veneer brick) are firstly selected with standard HSV information. Then, a digital camera is used to take ortho-projected photographs of the materials at six ambient color temperatures. Next, the AWB and AEC methods were used to conduct color calibration of the photos, and the corrected HSV values of the two materials can be obtained. Table 5.2 lists sample
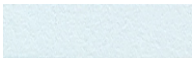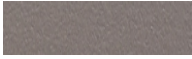
materials, the digital camera specifications, and the software used for the experiments. Finally, the shortest Euclidean distance between the standard HSV color value and the image color can be used to calculate the color deviation ΔE, and Equation 5.2 is as follows:

$$\Delta E = sqrt((x_n - x_s)^2 + (y_n - y_s)^2 + (z_n - z_s)^2) \tag{5.2}$$

where the HSV spatial coordinates can be calculated as (x_n, y_n, z_n) according to Equation (1), and (x_s, y_s, z_s) is the standard color HSV coordinate.

Figure 5.10 depicts the color deviation of the two materials in digital photos before and after color calibration at several ambient color temperatures. The results indicate that the introduced color calibration methods can significantly reduce the color deviation of digital images when the color temperature is warm or cold.

Table 5.2. Materials, apparatus, and software.

Materials			
ID	Facade Material Name	Facade Color Samples	Standard HSV Value
MAT. 1	Ceramic tiles		H:198, S: 8%, V: 96%
MAT. 2	Veneer brick		H: 16, S: 11%, V: 51%
Apparatus/Product			
Digital camera/Canon EOS 60D			
Software/Contents			
Photoshop CS4: An image processing software developed by Adobe, used to obtain the HSV value of the image color.			
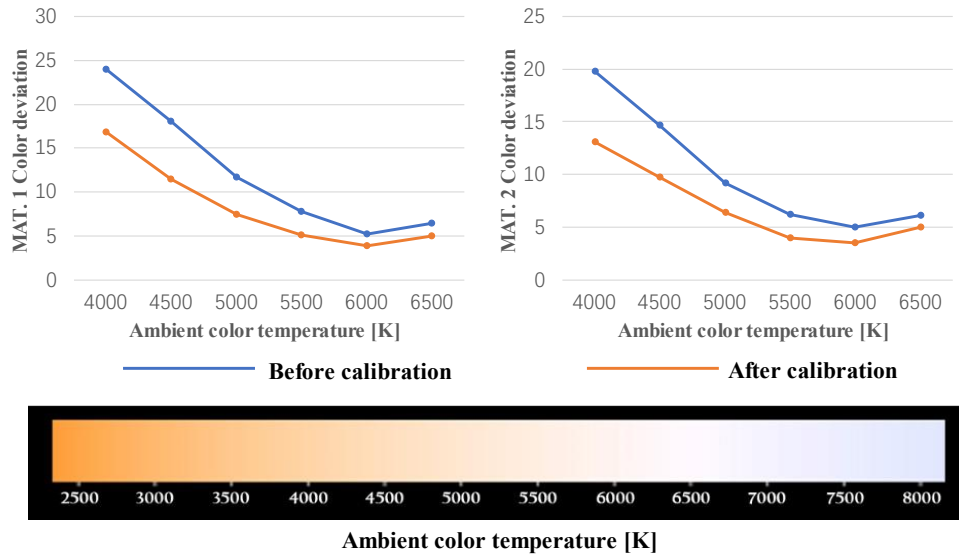
Figure 5.10. Color deviation of two materials in several color temperatures before and after photo color calibration.

The proposed methodology is validated in terms of color measurement based on 200 field survey images of street views randomly extracted from the three Chinese cities (Shanghai, Nanjing, and Hefei). The comparisons between the field survey and the proposed measurement method are shown in Figure 5.11. For color measurement validation, the architectural standard color card was first visually compared with the surveyed facade. The color code closest to the investigated object was recorded as the ground truth. Then, the HSV value of the measured color of the surveyed building facade was obtained. Finally, the color deviation between the measured color and ground truth was calculated for each field survey sample, and the range of color deviation was counted. The histogram of color deviation is shown in Figure 5.12, and more than 67% of the color deviation is lower than 20.

(Lat, Lon)	(32.0627, 118.7520)	(31.9841, 118.7225)	(31.9863, 118.7246)	(31.2334, 121.4753)
Street view				
Ground truth by field survey				
Façade color and Munsell color code	7.5Y9/3.6 (H: 40, S: 9%, V: 75%)	6.9R5.5/4.8 (H:4, S: 35%, V: 68%)	6.9PB6/9.2 (H:220, S: 43%, V: 82%)	7.5Y9/3.6 (H: 35, S: 15%, V: 93%)
Building function	Bank	School	Hotel	Residence
Measurement and classification by our method				
Façade color and Munsell color code	7.5Y7.5/2 (H: 34, S: 14%, V: 93%)	8.1R6/3.6 (H:7, S: 26%, V: 69%)	6.9PB7/4 (H: 212, S: 20%, V: 76%)	7.5Y9/1.2 (H: 44, S: 12%, V: 95%)
Building classification	Commercial service (B)	Public service (A)	Commercial service (B)	Residence (R)

Figure 5.11. The proposed measurement method results and the field survey data.
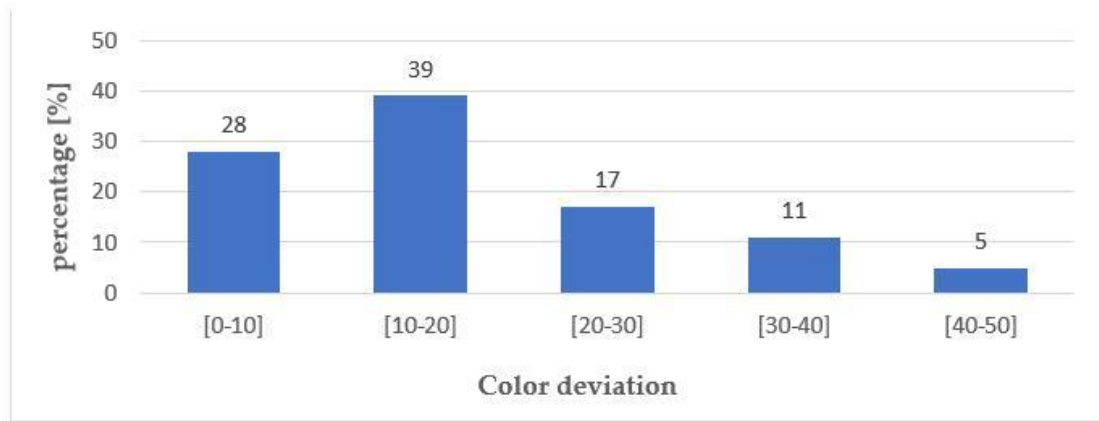
Figure 5.12. After color calibration, the distribution of the dominant color deviation was 28% for samples less than 10 and 67% for samples less than 20.

5.3.2 *Classification accuracy of building functions*

As shown in Figure 5.13 and Table 5.3, the four areas under the curve (AUC) of the trained DCNN model were evaluated through the test data. AUC is defined as the area enclosed by the coordinate axis under the receiver operating characteristic (ROC) curve. Since the maximum value of x and y after normalization is 1, and the ROC curve is generally above the line y=x, the AUC takes values in the range of 0.5 and 1. The closer the AUC is to 1.0, the higher the authenticity of the detection method. When it is equal to 0.5, the authenticity is the lowest and has no application value. As shown in

the results, the overall classification performance of EfficientNet was worse than the other networks. For the accuracy of commercial service and public service classification, ResNeSt performed better than the other three. For the class of residence (R), InceptionNet-v4 achieved the highest AUC value. After comparison, the trained ResNeSt model was selected, which has the highest overall accuracy among the four models, for the following generation of building functional classification maps.

For classification validation, the proposed method and ground truth were compared to the results of the classification of building functions. The overall building functional classification accuracy is 86.5%, as shown in Table 5.4. Most categories exceeded 85% accuracy, except for the residential type. These results are similar to the classification accuracy in Figure 5.13 and show that the prediction results by the trained ResNeSt achieve consistency with the verification results of the field investigation data.
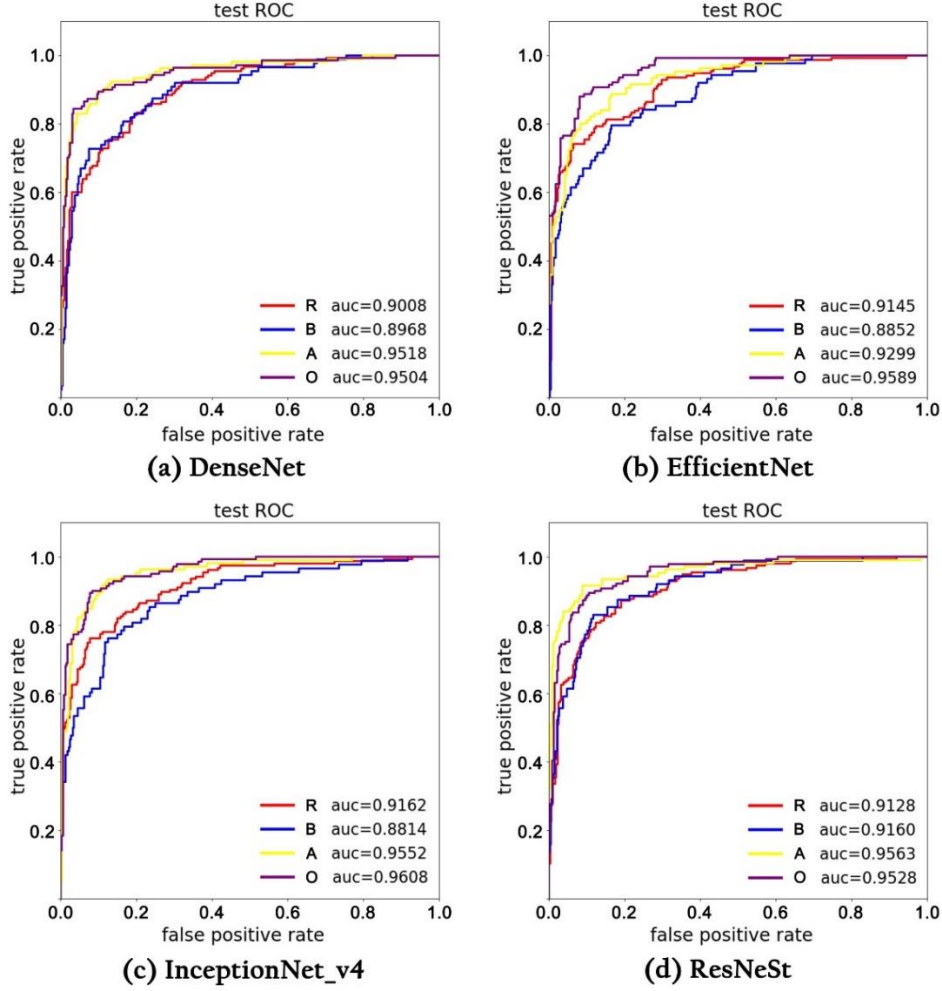
Figure 5.13. The AUC of the trained models including (a) DenseNet, (b) EfficientNet, (c) InceptionNet-v4, and (d) ResNeSt. The red line indicates the AUC of residence, the blue line B indicates the AUC of commerce, the yellow line A indicates the AUC of public service, and the purple line O represents the AUC of other facilities.

Table 5.3. Multi-label classification performance of all the trained networks.

Type	DenseNet	EfficientNet	InceptionNet-v4	ResNeSt
Residence (R)	0.9008	0.9145	**0.9162**	0.9148
Commercial service (B)	0.8968	0.8852	0.8814	**0.9160**
Public service (A)	0.9518	0.9299	0.9552	**0.9563**
Other facilities (O)	0.9504	0.9589	**0.9608**	0.9528
Overall	0.9249	0.9221	0.9284	**0.9349**

Bold values represent the highest output achieved among all the listed DCNNs.

Table 5.4. Building classification accuracy for the 200 sampled images.

Type	R	B	A	O	R + A	B + A
Number of samples	46	42	38	30	20	24
Subclass accuracy	84.8%	88.1%	89.5%	86.7%	85%	87.5%
Overall accuracy				86.5%		

5.3.3   Accuracy for analysis of wall and window segmentation

This study uses PyTorch, an open-source machine learning framework, and tests the pre-trained model on street-level images using the proposed dataset. The training sets are 400 images, and the testing sets are 100 images. The advantages of U-Net++ and DeepLabv3+ for the facade parsing task were described previously, and the two models were chosen to be trained separately on the proposed dataset, and then their segmentation performance was compared. During model training, a data enhancement technique was used: small rotations were applied at random to 60% of the data. In addition, a 10% color adjustment was applied at random to 60% of the training set to generate more training images.

The performance of the segmentation model was evaluated quantitatively and qualitatively. The quantitative evaluation includes precision, recall, and IoU to indicate the performance of the model. The IoU measures the overlap between positive predictions and positive samples. The detection results of small and large objects can be visually observed by Qualitative results. Besides, the difference in performance between the two models can be observed by human vision.

Table 5.5 gives the segmentation results of the two with the training model on the test set. Looking at the IoU metrics, the DeepLabv3+ performed better in the wall category. The qualitative analysis shows the same overall results. Figure 5.14 shows the segmentation examples for wall and window using DeepLabv3+ and U-Net++. Overall, the DeepLabv3+ shows better performance in dealing with boundaries and large objects

(such as walls), and a little improvement from U-Net++ in dealing with small objects (such as windows).

Table 5.5. Wall and windows segmentation performance using U-Net++ and DeepLabv3+.

	Model	Precision	Recall	IoU
Wall	U-Net++	88.35%	89.53%	83.61%
	DeepLabv3+	92.35%	90.87%	86.86%
Window	U-Net++	89.10%	87.19%	82.19%
	DeepLabv3+	90.67%	87.57%	83.88%

| Wall | Window | Sky | Road | Green | People | Door |

(a) Ground truth   (b) Prediction by DeepLabv3+   (c) Prediction by U-Net++
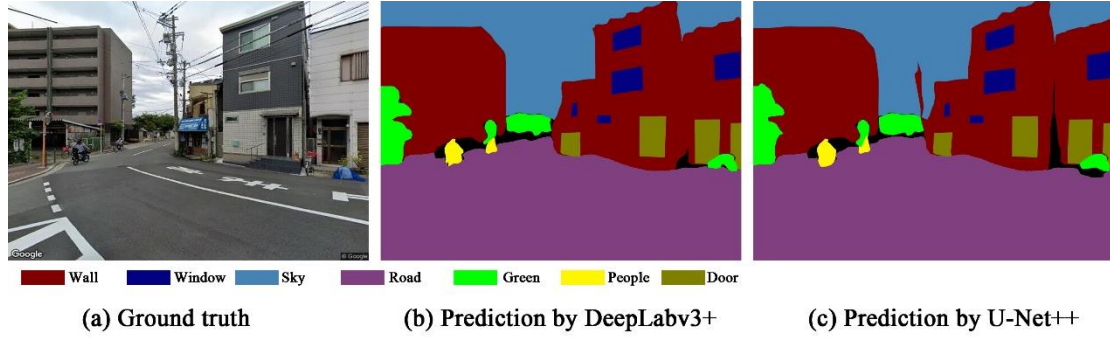
Figure 5.14. The segmentation examples for wall and window using DeepLabv3+ and U-Net++. (a) Ground truth, (b) prediction by DeepLabv3+, and (c) prediction by U-Net++.

5.3.4   *Automatic extraction of building facade results*

The urban facade database can be constructed by integrating the proposed methods. The construction process of the database can be divided into three steps. The first is data acquisition. The sampling points in the centerline of the city road can be acquired from urban geo-databases (like OSM). The street view images on both sides of the sampling points will be downloaded from Street View Service. The second is data pre-processing. The developed system will detect building facades in the street view image and remove the obstructions in front of the buildings. The separate building facade will be extracted based on the pre-trained instance segmentation model. The third is data mining. Building facades will be numbered, and each facade's dominant color, function, and window wall semantics will be counted by the proposed method. A 500m long

urban facade database was constructed for a street in Osaka. Figure 5.15 shows the location of the study area, and the street sampling points are set at 50m intervals. Figure 5.16 shows the details of the urban facade database, including the sampling point ID; the left and right along the direction of the street car; the coordinates of the sampling point; the street view images; the pictures after the unwanted object removal, and the facade instance segmentation; the number of individual facades; the dominant color of each facade (based on the Munsell color system); the function of the building (A for public service, B for commerce service, R for residence, O for other facilities); window-wall semantic segmentation of the facade.
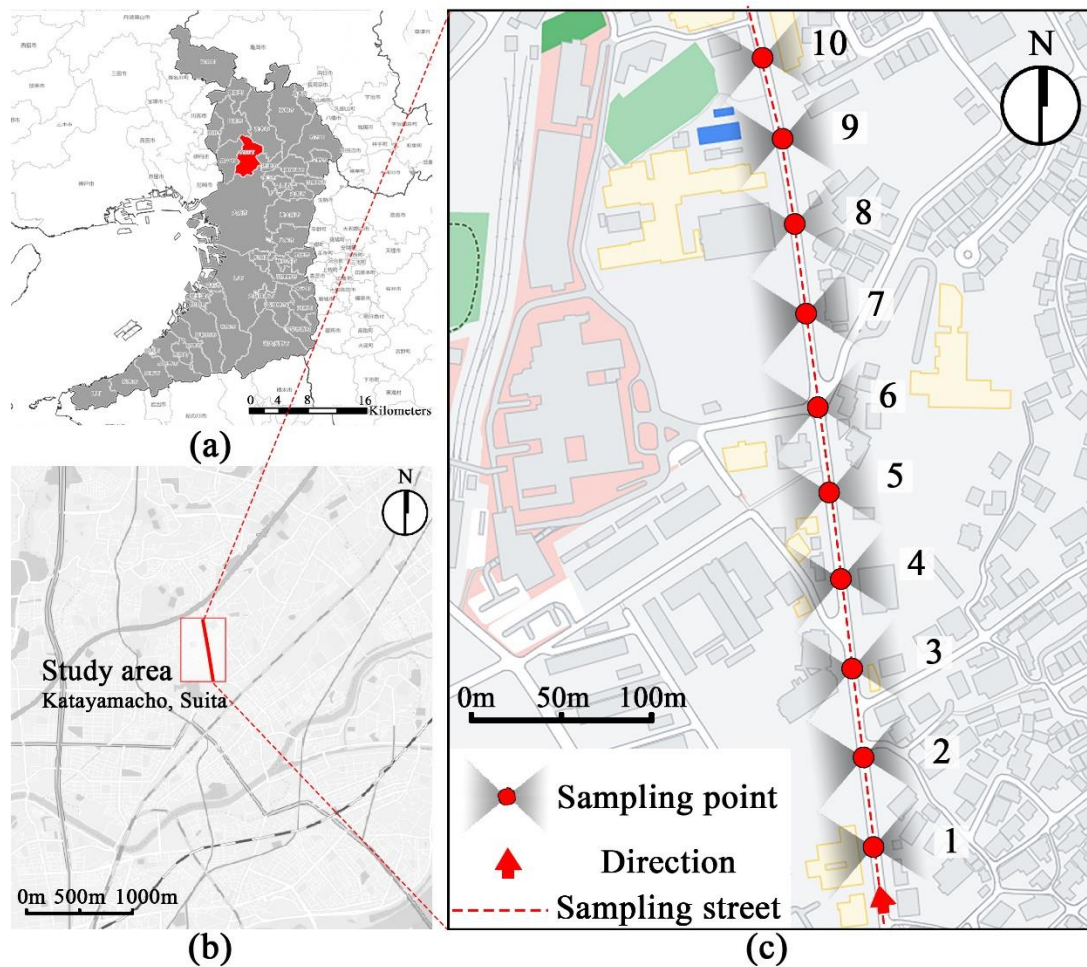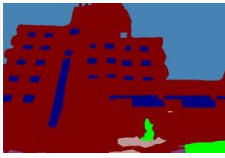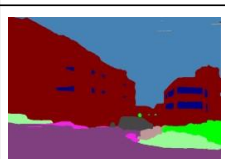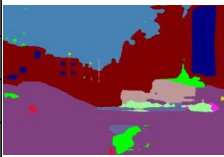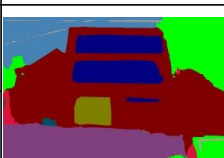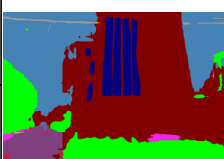
Figure 5.15. Study area, (a) Osaka Prefecture region, (b) A case study street in Suita, Osaka, (c) Ten sampling points are selected on a 500m-long street, and the street-level images are acquired from Google Street View service on the left and right sides of each sampling point along the street direction.

| ID | Dir | Coordinates | | Street view images | Unwanted objects removal and instance segmentation | FID | Dominant color | | Func-tion | Wall and windows segmentation |
		Lat	Lng							
1	Left	34.766 96	35.520 01			1	N7.25		A	
	Right					1	7.5GY9/1		O	
2	Left	34.767 18	135.51 993			N/A	N/A		N/A	N/A
	Right					2	8.1YR5.5/4		R	
							9.4YR7.5/5		O	
3	Left	34.767 85	135.51 983			2	8.1R5.5/1		R	
							N6.75		R	
	Right					1	8.8R5/1.6		A	
4	Left	34.768 41	135.51 977			2	3.8Y6/2		A	
							N7		R	
	Right					2	0.6YR7/2		R	
							1.3Y6.5/2.4		A	

ID	Dir	Coordinates		Street view images	Unwanted objects removal and instance segmentation	FID	Dominant color		Function	Wall and windows segmentation
		Lat	Lng							
5	Left	34.768 55	135.51 976			4	2.5Y4/1.6		R	
							2.5Y6/1.2		A	
							0.6GY5.5/1		R	
							1.9Y5.5/1.2		A	
	Right					4	3.8GY4/3.6		R	
							1.3GY5/2.4		R	
							0.6GY7/1		A	
							8.8P8.5/1		A	
6	Left	34.768 95	135.51 966			N/A	N/A		N/A	N/A
	Right					2	8.8R5.5/1.6		R	
							8.8YR6/5		A	
7	Left	34.769 03	135.51 950			N/A	N/A		N/A	N/A
	Right					2	8R4.5/1.4		R	
							10Y3.5/1.8		B	
8	Left	34.769 46	135.51 958			2	3.8Y6.5/1		R	
							N5.25		A	
	Right					2	3YR3/1		R	
							10B6.5/1		R	

(Continued on next page)

ID	Dir	Coordinates		Street view images	Unwanted objects removal and instance segmentation	FID	Dominant color	Func-tion	Wall and windows segmentation
		Lat	Lng						
9	Left	34.769 95	135.51 951			N/A	N/A	N/A	N/A
	Right					1	7.5GY9/1	R	
10	Left	34.771 15	135.51 927			N/A	N/A	N/A	N/A
	Right					1	5.6YR7.5/2	R	

Figure 5.16. An example facade database is constructed for a 500m street in Suita, Osaka. The facade database includes the sampling point ID; the left and right along the direction of the street car; the coordinates of the sampling point; the street view images; the pictures after the unwanted object removal and the facade instance segmentation; the number of individual facades; the dominant color of each facade (based on the Munsell color system); the function of the building (A for public service, B for commerce service, R for residence, O for other facilities); window-wall semantic segmentation of the facade. N/A means no facade.

5.4   Discussion

5.4.1   Comparison with conventional methods

For the facade color measurements, the previous methods developed by Lu et al. (2010) and Nguyen & Teller (2017) are computationally expensive in terms of facade color measurement and building function statistics, based mainly on field studies, and with low expansibility. These methods require a significant amount of manual measurement data, including on-site streetscape images and questionnaires, and are restricted to neighborhood-scale studies. In contrast to the field survey-based method, the proposed deep learning-based data processing method can analyze large amounts

of data with high accuracy and is more cost-effective in measuring the facade color corresponding to the building function classification. The proposed method can quantitatively analyze the color distribution at different building functions to support evidence-based urban analytics and design rather than simply qualitative descriptions.

The proposed method can accurately parse facades in street-level complex scenes for the segmentation of walls and windows. Compared to previous methods (Gadde et al., 2016; H. Liu et al., 2020; Ma et al., 2020; Teboul et al., 2012), this study's contribution is to customize a high-resolution facade parsing dataset for complex scenes. The new dataset contains wall annotation based on individual buildings. A larger dataset of street-level facades with multiple views, foreground occlusions, various lighting conditions, and complex facade backgrounds is included in the proposed dataset.

Compared to facade extraction methods that use semantic segmentation (J. Zhang et al., 2021b), the proposed method using instance segmentation can extract building facade information one by one when multiple buildings are connected in a single image. The conventional method treats the information of all buildings in a picture as a whole and is unable to parse the building monolithically. The proposed method is more accurate and overcomes the previous problem of not being able to parse individual facades information (such as individual facade dominant color or individual building function) in connected buildings (visual adjacency or overlay).

5.4.2   *Potential applications*

This study attempts to construct a quantitative research method for the city-scale measurement of facade data, including color, functions, walls, and windows. After testing, the technique demonstrated its viability and convenience in initial investigations of urban design and city modeling, implying potential application as an augmented tool for designers to establish objective decision bias and enable a data-driven strategy. Given the method's benefits, it could be used to discover discordant architectural colors in particular functional areas, assess the color planning of the built

environment, and provide foundation color details for urban design implementation, thus facilitating a feedback process. For example, the new and old facade color has a noticeable difference because of the pace of construction and business distribution. This study provides city managers with a clear understanding of street-level facade colors with building classification to realize the optimal balanced development of the new buildings and traditions. In addition, quantitative measurement and classification provide empirical value for intelligent design guidelines in various areas, such as residential, commercial, and public services. By analyzing the color and function of the city, the authorities could explore the color tendencies of functional buildings in different cities. Then propose urban planning solutions with their own identity. This process helps avoid the drawbacks of stylistic homogenization induced by the prevalence of functionalism. It is expected to help improve the color quality of the urban built environment, especially in further exploring the visual environment design, to better support urban renewal in the post-urbanization period.

The workflow proposed in this study can help create a portrait of the building at the city scale and move to the next modeling step, such as urban building energy modeling and 3D modeling reconstruction, which is crucial for building retrofitting solutions and semantic enrichment of BIM. Other important building indicators, such as the windows-to-wall ratio, which is essential for assessing the building energy performance, can also be calculated from a semantic segmentation model integrated with orthogonal transformation. In addition, the method may be more convenient and economical than traditional methods, as it is characterized by ease of implementation and does not rely on intensive physical labor.

5.4.3 Limitations

For the facade color measurements, the intense sunlight will impact the quality of street view images, affecting the color calculation based on the introduced method; an example is shown in Figure 5.17a. The color calibration of street view images can

improve the calculation results. However, with the current color calibration methods, it is difficult to obtain the actual color of the building facade for some overexposed and overly dark street view images. In this way, the low image quality has a negative impact on the accuracy of the color measurement, classification, and segmentation tasks of the buildings.

According to the building classification results of the four classes, some residential areas are relatively more difficult to identify than other classes, owing to the fact that residential areas in older towns are highly mixed in function. Commercial services often exist on the ground floor of residences, and few individual houses are on the streets of these study cities, causing the classification accuracy of some residential buildings to be lower than other classes. As shown in Figure 5.17b, the building in the street view image is predicted to be a mixed service. Last, there are a few manual tagging errors from OSM users in the training set of the classification model, especially for similar facade features. As shown in Figure 5.17c, the building in the street view photo tends to be a residential apartment, while the label from the OSM user is a hotel. Detecting multiple labels for building functions is possible. By recreating the dataset for multi-label image classification training, automatic recognition of multiple building functions (more than two) for classification in a single street view image will be achieved.
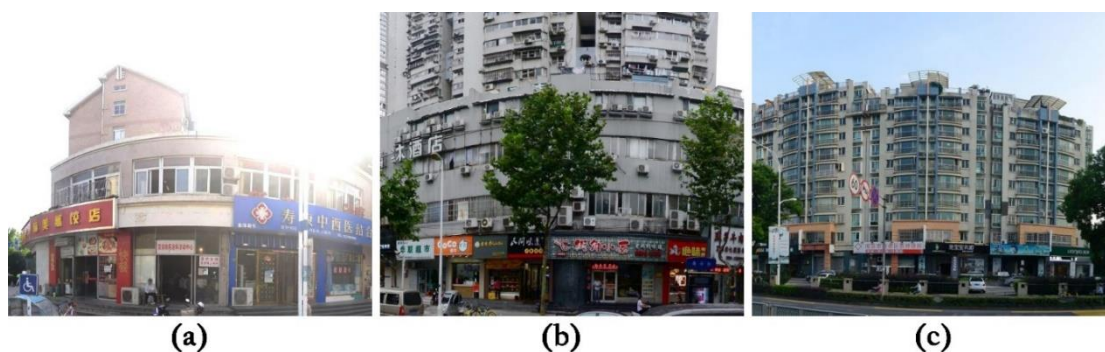
Figure 5.17. Some observations from street view images illustrate the limitations of color measurements and functional classification. (a) Color deviations persist in the overexposed street view image despite color calibration. (b) It is difficult to identify a residential building with commercial service. (c) The building in the street view photo is an apartment, whereas the label from the OSM user is a hotel.

For the walls and windows segmentation, the current limitation is that the proposed building facade dataset has not been validated on a large range of streetscape images under complex weather and crowded street, which is important for practical engineering-oriented applications. Since some windows are small targets in street view images, they are prone to lose information when down-sampling during CNN-based model training, resulting in suboptimal segmentation accuracy for small targets. The purpose of this research is oriented to the update of buildings and the semantic understanding of building information models, so the accuracy of the results is required to be high. The segmentation task for walls and windows of building facades is characterized by the presence of both large and small targets in the picture, the diversity of object angles and forms, and the state-of-the-art semantic model used in this study is a generic design that does not achieve the most desirable segmentation accuracy results. Furthermore, when glass reflects the sky or street objects, the segmentation model tends to identify the reflection of window glass as other objects, which reduces its overall quantitative performance (as shown in Figure 5.18). In future work, segmentation accuracy will be further improved by tailored datasets and algorithm enhancements. The semantic information of the building facade in the street image will be distortion corrected to obtain a usable segmentation result.

Figure 5.18. Some observations from street view images illustrate the limitations of facade segmentation. (a) The window glass reflects trees, and (b) the window glass reflects buildings will reduce the segmentation accuracy of walls and windows.

The GAN-based method eliminates the unwanted objects in front of the building and completes them, which can solve the interference caused by obstacles to the building elevation information extraction and improve the accuracy of the acquired facade information. The proposed method currently generates prediction results that will be slightly inconsistent with the real situation. In future research, the proposed method will improve the accuracy of the algorithm, and the generated textures will be consistent with the ground truth as much as possible. Many full reference metrics have been proposed, but it is difficult to assess the gap between the generated images and ground truth due to the unavailability of ground truth. The current criteria for evaluating the generated image are two-fold: (1) based on human visual perception, and (2) truthfulness of the content. First, after the GAN-based obstructed facade completion, the goodness of the generated image can be judged by the experience of human visual perception, such as color and semantic consistency, gray scale, or similarity. Second, the current method cannot assess the authenticity of the generated image content. As an example, the window size on the infill cannot be compared with the ground truth. Therefore, calculating the window-to-wall ratio or the semantic segmentation of windows is impossible to determine whether the generated images are accurate compared to the real situation, leading to uncertain applications to the window-to-wall ratio problems. This problem might be solved by synthesizing the street view. Unwanted objects, such as trees, cars, people, etc., are superimposed on the complete building facade image. The ground truth image, the obscured facade image with masks, and generated image by inpainting the mask can be obtained.

5.5   Summary of this Chapter

This study proposed an automatic approach for facade color measurements, building functional classification, and window-walls segmentation at a large scale by applying state-of-the-art deep learning methods and street-level images. A pre-processing data method for facade color measurement was developed in two steps:

image color calibration and obstructed facades completion. A tailored dataset of street view images is built to train a multi-label classifier for building functions, including residential, public, commercial, and other facilities. Finally, a tailored dataset of building facades is built for training semantic segmentation models for walls and windows.

The proposed methods measure facade color, classify building functions, and segment walls and windows using street-level images. The accuracy of the proposed method was verified by field surveys. The results show that the proposed methods have satisfactory accuracy, with a color deviation of less than 20 for more than 67% of the measured data and overall accuracy of 86.50% for the building functional classification. The IoUs for semantic segmentation of walls and windows using DeepLabv3+ are 86.86% and 83.88%, respectively. The proposed method can automatically collect basic building information to support data for urban renewal. This work aims to quickly access inventory data of existing buildings to aid in the application of large-scale city information modeling and building renewal.

Chapter 6. Conclusions

6.1   Summary

This study presents an automated workflow for extracting multiple types of building facade data based on deep learning and geo-tagged street view imagery, assigning measured data to individual buildings in urban areas. The main body of the study is divided into three parts. Firstly, this study develops an unwanted object elimination system that can obtain complete building facades to improve the fidelity of building facade information based on street-level pictures. Secondly, a facade instance segmentation method based on CDT synthetic dataset is proposed, which has two benefits: one is to solve the segmentation problem of adjacent buildings. The other is that the automatically generated synthetic dataset dramatically reduces the cost of data annotation. Thirdly, an integrated multitasking facade data extraction method is proposed. Building information, including building dominant color, building functional classification, and wall and window semantics, will be automatically counted. For dataset making, the author proposes a publicly available facade dataset that can be used for obstacle removal in streets, DCNN-based facade instance segmentation models, and window-wall semantic segmentation. The proposed method has been validated in several cities, and the results prove its effectiveness.

6.2   Research contributions

This study proposes a workflow for large-scale acquisition and quantitative analysis of multiple data of building facades based on street view images. The method can overcome the interference of obstacles in the street to the facade data acquisition. In addition, it is difficult to obtain individual information on connected buildings based on the previous methods because of the lack of diverse facade instance annotations. The

proposed CDT synthetic dataset HSRBFIA in this study can be effectively used for facade instance segmentation for real images, revealing the potential of the proposed synthetic dataset to replace real data. Further, multiple building facade data types, such as facade dominant color and window-wall semantics, that are not recorded in existing urban geographic databases (such as OSM) are measured and analyzed. Overall, the contribution of this study consists of three parts, which are concluded as follows.

1) An obstructed facade completion method was proposed. As a result, unwanted objects in the street, including people, greenery, and cars, can be removed. In addition, a dataset called SVBFI was tailored for DCNN-based facade inpainting with unoccluded facade images, mask images, and semantic segmentation labels. Eliminating obstacles in front of the building can effectively improve the loss of information about the building facade obtained through street view images.

2) An automatic generation system is proposed to create CDT synthetic data for training facade instance segmentation. This approach takes about 1/2,050 of the time that it takes to manually annotate each image, which can significantly reduce the cost required to annotate data. The segmentation accuracy is boosted significantly when a certain fraction of real data is loaded into the CDT synthetic datasets, to the point where its performance becomes competitive with what is seen when 100% real data is used. Verification for multiple other cities demonstrated the transferability of the proposed framework. CDT synthetic dataset can obtain promising prediction results for most modern architectural styles. In addition, this method can effectively achieve monolithic data extraction of connected buildings.

3) The extraction methods of multiple building facade data based on geo-tagged street view imagery and deep learnings were presented. A pre-processing data method for facade color measurement was developed in three steps: image

color calibration, obstructed facades completion, and facade instance segmentation. A tailored dataset of street view images is built to train a multi-label classifier for building functions, including residential, public, commercial, and other facilities. A tailored dataset of high-resolution facade images is built for training semantic segmentation models for walls and windows.

The tools developed by this study have been made open access and are listed as follows.

- A format conversion tool was in Chapter 4: from synthetic data to COCO format for training deep learning-based instance segmentation models: https://github.com/Mortyzhang/Mask2polygon_tool

- A facade dominant color calculation tool was used in Chapter 5: https://github.com/Mortyzhang/Facade-color-calculation-based-on-colorcard

- A building function classification tool using street view images in Chapter 5: https://github.com/Mortyzhang/Nanjing-street-view-datasets-and-classification-tasks

6.3   Limitations and future work

The study's limitations can be divided into two aspects, one is the customized model architecture, and the other is the bridging application. For the former, the general CNN-based models in this study have not been targeted to improve the building facade parsing task, so the accuracy needs to be further improved. For example, since most windows are rectangular, controlled algorithm improvements based on a priori knowledge can significantly improve segmentation accuracy for parsing exterior windows. For the latter, this study can measure the facade data of buildings (facade dominant color, building functions, and window-wall semantics) at a large scale. However, the variety is limited, such as precise facade geometry data (like building

height, building perimeter, building volume, etc.) cannot be covered. Therefore, it needs to be supplemented with data from other open-source urban geodatabases (like OSM and PLATEAU). This makes applying the current method directly in urban building energy modeling or building retrofitting tasks difficult.
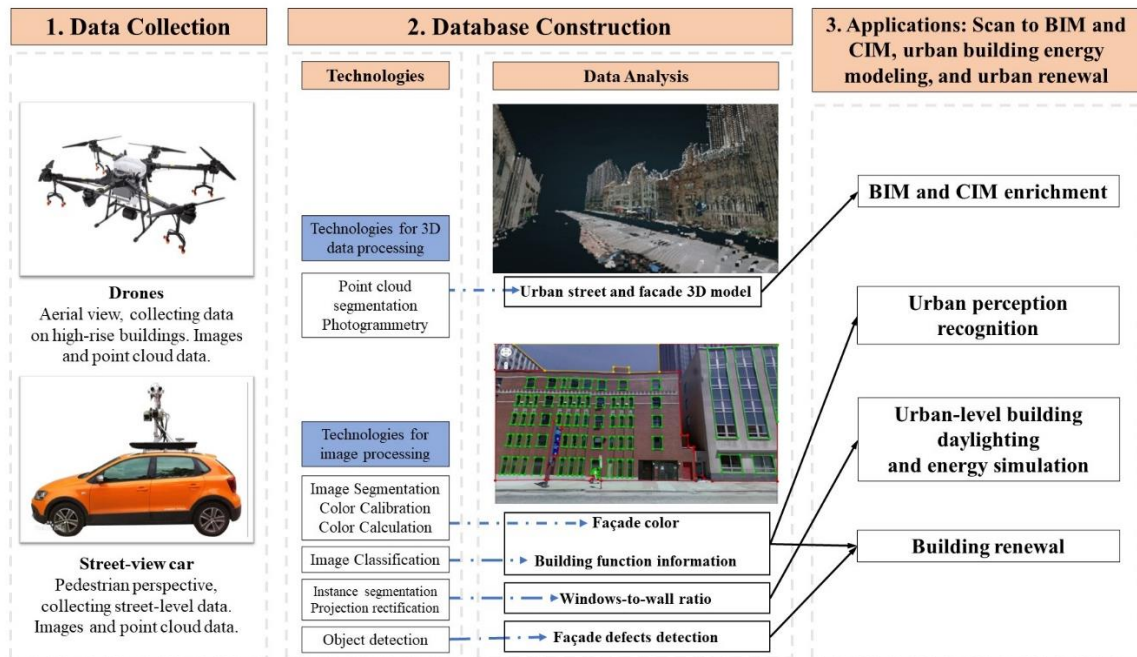
Figure 6.1. The research work for the future can be divided into the following three directions. (1) Supplementary 3D point cloud data collection; (2) construction of the urban database; (3) and practical application-oriented data analysis.

The overall goal of the future work is to implement an automated, scalable and comprehensive building facade analysis system to enable efficient measurement of urban building facades with multi-source data types. This system will largely help transform existing building analysis methods, reduce data limitations, and increase efficiency for developing urban facade databases. Figure 6.1 shows the research work for the future. It can be divided into the following three steps. (1) Data collection supplementary. 3D point cloud data needs to be collected for measuring facade geometries. Buildings need to be monolithic in the 3D reconstruction model, rather than treating all objects as a single mesh. (2) Construction of the database. All the collected

data needs to be integrated into one database to address urban development issues in an integrated manner. More building facade data analysis will be implemented based on databases. For example, defects detection of building facades, especially glass curtain walls and other vulnerable parts. Thermal and hyperspectral images are used to determine the building's thermal properties and material type, respectively. It is possible to identify building materials using spectral characteristics. Similarly, thermal maps of the building's facade can be used to assess the presence of different components' thermal bridges separately. (3) Practical application-oriented data analysis. Combined with computer vision techniques and multispectral cityscape pictures, localized building features will directly help automate the extraction of current existing building data for use by local government authorities or other stakeholders. With the development of the facade data extraction systems and evaluation of the proposed method by city-scale data, it is clear that the prospect of applying this research to urban development issues, such as building information modeling and city information modeling enrichment (Xue et al., 2021), urban perception recognition (Larkin et al., 2021), urban-level building daylighting and energy simulation (Szcześniak et al., 2022), and building renewal (Zheng et al., 2017).

References

Aheleroff, S., Xu, X., Zhong, R. Y., & Lu, Y. (2021). Digital twin as a service (DTaaS) in industry 4.0: An architecture reference model. *Advanced Engineering Informatics*, *47*, 101225. https://doi.org/10.1016/j.aei.2020.101225

Al-Habaibeh, A., Sen, A., & Chilton, J. (2021). Evaluation tool for the thermal performance of retrofitted buildings using an integrated approach of deep learning artificial neural networks and infrared thermography. *Energy and Built Environment*, *2*(4), 345–365. https://doi.org/10.1016/j.enbenv.2020.06.004

Anguelov, D., Dulong, C., Filip, D., Frueh, C., Lafon, S., Lyon, R., Ogale, A., Vincent, L., & Weaver, J. (2010). Google street view: Capturing the world at street level. *Computer*, *43*(6), 32–38.

Architectural standard color chart. (n.d.). *https://www.colortell.com/colorbook/?callbook=a8*.

Bai, M., & Urtasun, R. (2017). Deep watershed transform for instance segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5221–5229.

Baidu Street View service. (2022). *Panoramic Static Map API | Baidu Map API SDK*. https://lbsyun.baidu.com/index.php?title=viewstatic

Bescos, B., Neira, J., Siegwart, R., & Cadena, C. (2019). Empty cities: Image inpainting for a dynamic-object-invariant space. *2019 International Conference on Robotics and Automation (ICRA)*, 5460–5466.

Biljecki, F., Ledoux, H., & Stoter, J. (2016). GENERATION OF MULTI-LOD 3D CITY MODELS IN CITYGML WITH THE PROCEDURAL MODELLING ENGINE RANDOM3DCITY. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *IV-4/W1*, 51–59.

https://doi.org/10/gcc634

Bódis-Szomorú, A., Riemenschneider, H., & Van Gool, L. (2017). Efficient edge-aware surface mesh reconstruction for urban scenes. *Computer Vision and Image Understanding*, *157*, 3–24. https://doi.org/10.1016/j.cviu.2016.06.002

Bolya, D., Zhou, C., Xiao, F., & Lee, Y. J. (2019). Yolact: Real-time instance segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9157–9166.

Boodi, A., Beddiar, K., Amirat, Y., & Benbouzid, M. (2022). Building Thermal-Network Models: A Comparative Analysis, Recommendations, and Perspectives. *Energies*, *15*(4), 1328. https://doi.org/10.3390/en15041328

Borkman, S., Crespi, A., Dhakad, S., Ganguly, S., Hogins, J., Jhang, Y.-C., Kamalzadeh, M., Li, B., Leal, S., Parisi, P., Romero, C., Smith, W., Thaman, A., Warren, S., & Yadav, N. (2021). Unity Perception: Generate Synthetic Data for Computer Vision. *ArXiv:2107.04259 [Cs]*. http://arxiv.org/abs/2107.04259

Botvinick, M., Ritter, S., Wang, J. X., Kurth-Nelson, Z., Blundell, C., & Hassabis, D. (2019). Reinforcement learning, fast and slow. *Trends in Cognitive Sciences*, *23*(5), 408–422. https://doi.org/10.1016/j.tics.2019.02.006

Cai, S., Ma, Z., Skibniewski, M. J., & Bao, S. (2019). Construction automation and robotics for high-rise buildings over the past decades: A comprehensive review. *Advanced Engineering Informatics*, *42*, 100989. https://doi.org/10.1016/j.aei.2019.100989

Cai, Z., & Vasconcelos, N. (2019). Cascade r-cnn: High quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Campbell, A., Both, A., & Sun, Q. C. (2019). Detecting and mapping traffic signs from Google Street View images using deep learning and GIS. *Computers, Environment and Urban Systems*, *77*, 101350. https://doi.org/10.1016/j.compenvurbsys.2019.101350

Cao, Y., Zhang, X., Fu, Y., Lu, Z., & Shen, X. (2020). Urban spatial growth modeling using logistic regression and cellular automata: A case study of Hangzhou. *Ecological Indicators*, *113*, 106200. https://doi.org/10.1016/j.ecolind.2020.106200

Carvalho, O. L. F. de, de Carvalho Júnior, O. A., Albuquerque, A. O. de, Bem, P. P. de, Silva, C. R., Ferreira, P. H. G., Moura, R. dos S. de, Gomes, R. A. T., Guimarães, R. F., & Borges, D. L. (2020). Instance Segmentation for Large, Multi-Channel Remote Sensing Imagery Using Mask-RCNN and a Mosaicking Approach. *Remote Sensing*, *13*(1), 39. https://doi.org/10.3390/rs13010039

Chen, H., Sun, K., Tian, Z., Shen, C., Huang, Y., & Yan, Y. (2020). Blendmask: Top-down meets bottom-up for instance segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8573–8581.

Chen, K.-L., Lee, C.-H., Garudadri, H., & Rao, B. D. (2021). ResNEsts and DenseNEsts: Block-based DNN Models with Improved Representation Guarantees. *Advances in Neural Information Processing Systems*, *34*, 3413–3424.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of the European Conference on Computer Vision (ECCV)*, 801–818.

Chen, M.-C., Ho, T.-P., & Jan, C.-G. (2006). A system dynamics model of sustainable urban development: Assessing air purification policies at Taipei City. *Asian Pacific Planning Review*, *4*(1), 29–52.

Chen, X., Girshick, R., He, K., & Dollár, P. (2019). Tensormask: A foundation for dense object segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2061–2069.

Chen, X.-W., & Lin, X. (2014). Big data deep learning: Challenges and perspectives.

IEEE Access, *2*, 514–525. https://doi.org/10.1109/ACCESS.2014.2325029

Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1800–1807. https://doi.org/10/gfxgtm

Chun, B.-S., & Kim, H.-Y. (2010). Analysis of urban heat island effect using information from 3-dimensional city model (3DCM). *Spatial Information Research*, *18*(4), 1–11.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The Cityscapes Dataset for Semantic Urban Scene Understanding. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3213–3223. https://doi.org/10.1109/CVPR.2016.350

Criminisi, A., Pérez, P., & Toyama, K. (2004). Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on Image Processing*, *13*(9), 1200–1212. https://doi.org/10/fbk8ft

Dai, D., Riemenschneider, H., Schmitt, G., & Van, L. (2013). Example-Based Facade Texture Synthesis. *2013 IEEE International Conference on Computer Vision*, 1065–1072. https://doi.org/10/ghhnkc

Dai, J., He, K., Li, Y., Ren, S., & Sun, J. (2016). Instance-sensitive fully convolutional networks. *European Conference on Computer Vision*, 534–549.

Dai, M., Ward, W. O. C., Meyers, G., Densley Tingley, D., & Mayfield, M. (2021). Residential building facade segmentation in the urban environment. *Building and Environment*, *199*, 107921. https://doi.org/10.1016/j.buildenv.2021.107921

Degaev, E., & Barkhi, R. (2019). Integrated assessment of contractor's building production culture during facade repair. *Journal of Physics: Conference Series*, *1425*(1), 012066. https://doi.org/10.1088/1742-6596/1425/1/012066

Deng, M., Gan, V. J. L., Tan, Y., Joneja, A., & Cheng, J. C. P. (2019). Automatic generation of fabrication drawings for façade mullions and transoms through BIM models. *Advanced Engineering Informatics*, *42*, 100964.

https://doi.org/10.1016/j.aei.2019.100964

Deng, T., Zhang, K., & Shen, Z.-J. M. (2021). A systematic review of a digital twin city: A new pattern of urban governance toward smart cities. *Journal of Management Science and Engineering*, *6*(2), 125–134. https://doi.org/10.1016/j.jmse.2021.03.003

Doppioslash, C. (2018). *Physically Based Shader Development for Unity 2017*. Springer.

Doyle, S. (2019). Siblings make sense of smart cities. *Engineering & Technology*, *14*(1), 42–45. https://doi.org/10.1049/et.2019.0103

Du, G., Cao, X., Liang, J., Chen, X., & Zhan, Y. (2020). Medical image segmentation based on u-net: A review. *Journal of Imaging Science and Technology*, *64*(2), 20508–1. https://doi.org/10.2352/J.ImagingSci.Technol.2020.64.2.020508

Elharrouss, O., Almaadeed, N., Al-Maadeed, S., & Akbari, Y. (2020). Image inpainting: A review. *Neural Processing Letters*, *51*(2), 2007–2028. https://doi.org/10.1007/s11063-019-10163-0

Esser, P., Sutter, E., & Ommer, B. (2018). A variational u-net for conditional appearance and shape generation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8857–8866.

Fan, C., Zhang, C., Yahja, A., & Mostafavi, A. (2021). Disaster City Digital Twin: A vision for integrating artificial and human intelligence for disaster management. *International Journal of Information Management*, *56*, 102049. https://doi.org/10.1016/j.ijinfomgt.2019.102049

Femiani, J., Para, W. R., Mitra, N. J., & Wonka, P. (2018). Facade Segmentation in the Wild. *CoRR*, *abs/1805.08634*. http://arxiv.org/abs/1805.08634

Ferrando, M., & Causone, F. (2020). An overview of urban building energy modelling (UBEM) tools. *Building Simulation 2019*, *16*, 3452–3459. https://doi.org/10.26868/25222708.2019.210632

Ferrando, M., Causone, F., Hong, T., & Chen, Y. (2020). Urban building energy

modeling (UBEM) tools: A state-of-the-art review of bottom-up physics-based approaches. *Sustainable Cities and Society*, *62*, 102408. https://doi.org/10.1016/j.scs.2020.102408

Final UK greenhouse gas emissions national statistics: 1990 to 2019. (n.d.). GOV.UK. Retrieved April 21, 2022, from https://www.gov.uk/government/statistics/final-uk-greenhouse-gas-emissions-national-statistics-1990-to-2019

Fukuyama, M. (2018). Society 5.0: Aiming for a new human-centered society. *Japan Spotlight*, *1*, 47–50.

Gadde, R., Marlet, R., & Paragios, N. (2016). Learning Grammars for Architecture-Specific Facade Parsing. *International Journal of Computer Vision*, *117*(3), 290–316. https://doi.org/10.1007/s11263-016-0887-4

Gao, Q., Shen, X., & Niu, W. (2020). Large-scale synthetic urban dataset for aerial scene understanding. *IEEE Access*, *8*, 42131–42140.

Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.-Y., Cubuk, E. D., Le, Q. V., & Zoph, B. (2021). Simple copy-paste is a strong data augmentation method for instance segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2918–2928. https://doi.org/10.1109/CVPR46437.2021.00294

Girshick, R. (2015). Fast r-cnn. *Proceedings of the IEEE International Conference on Computer Vision*, 1440–1448. https://doi.org/10.1109/ICCV.2015.169

Google Street View service. (2022). *Street View Static API overview*. Google Developers. https://developers.google.com/maps/documentation/streetview/overview

Hafiz, A. M., & Bhat, G. M. (2020). A survey on instance segmentation: State of the art. *International Journal of Multimedia Information Retrieval*, *9*(3), 171–189. https://doi.org/10.1007/s13735-020-00195-x

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. *Proceedings of the IEEE International Conference on Computer Vision*, 2961–2969.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity mappings in deep residual networks. *European Conference on Computer Vision*, 630–645.

Hu, M. (2020). Life-cycle environmental assessment of energy-retrofit strategies on a campus scale. *Building Research & Information*, *48*(6), 659–680. https://doi.org/10.1080/09613218.2019.1691486

Iandola, F., Moskewicz, M., Karayev, S., Girshick, R., Darrell, T., & Keutzer, K. (2014). Densenet: Implementing efficient convnet descriptor pyramids. *ArXiv Preprint ArXiv:1404.1869*.

Ikeno, K., Fukuda, T., & Yabuki, N. (2021). An enhanced 3D model and generative adversarial network for automated generation of horizontal building mask images and cloudless aerial photographs. *Advanced Engineering Informatics*, *50*, 101380. https://doi.org/10.1016/j.aei.2021.101380

Istenič, K., Gracias, N., Arnaubec, A., Escartín, J., & Garcia, R. (2020). Automatic scale estimation of structure from motion based 3D models using laser scalers in underwater scenarios. *ISPRS Journal of Photogrammetry and Remote Sensing*, *159*, 13–25. https://doi.org/10.1016/j.isprsjprs.2019.10.007

Jechow, A., Kyba, C. C., & Hölker, F. (2020). Mapping the brightness and color of urban to rural skyglow with all-sky photometry. *Journal of Quantitative Spectroscopy and Radiative Transfer*, *250*, 106988. https://doi.org/10/ghztcd

Kang, J., Körner, M., Wang, Y., Taubenböck, H., & Zhu, X. X. (2018). Building instance classification using street view images. *ISPRS Journal of Photogrammetry and Remote Sensing*, *145*, 44–59. https://doi.org/10/gddhfh

Kheiri, F. (2018). A review on optimization methods applied in energy-efficient building geometry and envelope design. *Renewable and Sustainable Energy Reviews*, *92*, 897–920. https://doi.org/10.1016/j.rser.2018.04.080

Kido, D., Fukuda, T., & Yabuki, N. (2020). Diminished reality system with real-time object detection using deep learning for onsite landscape simulation during redevelopment. *Environmental Modelling & Software*, 104759.

Kido, D., Fukuda, T., & Yabuki, N. (2021). Assessing future landscapes using enhanced mixed reality with semantic segmentation by deep learning. *Advanced Engineering Informatics*, *48*, 101281. https://doi.org/10/gkgn3g

Kikuchi, T., Fukuda, T., & Yabuki, N. (2021). Automatic Diminished Reality-Based Virtual Demolition Method using Semantic Segmentation and Generative Adversarial Network for Landscape Assessment. *Proceedings of the 39th eCAADe Conference*, *2*, 529–538.

Kong, G., & Fan, H. (2021). Enhanced Facade Parsing for Street-Level Images Using Convolutional Neural Networks. *IEEE Transactions on Geoscience and Remote Sensing*, *59*(12), 10519–10531. https://doi.org/10.1109/TGRS.2020.3035878

Koonce, B. (2021). EfficientNet. In *Convolutional Neural Networks with Swift for Tensorflow* (pp. 109–123). Springer.

Korc, F., & Förstner, W. (2009). ETRIMS Image Database for interpreting images of man-made scenes. *Dept. of Photogrammetry, University of Bonn, Tech. Rep. TR-IGG-P-2009-01*.

Lam, E. Y., Fung, G. S., & Lukac, R. (2008). Automatic white balancing in digital photography. *Single-Sensor Imaging: Methods and Applications for Digital Cameras*, 267–294. https://doi.org/10/c32v36

Larkin, A., Gu, X., Chen, L., & Hystad, P. (2021). Predicting perceptions of the built environment using GIS, satellite and street view image approaches. *Landscape and Urban Planning*, *216*, 104257. https://doi.org/10.1016/j.landurbplan.2021.104257

Lateef, F., & Ruichek, Y. (2019). Survey on semantic segmentation using deep learning techniques. *Neurocomputing*, *338*, 321–348. https://doi.org/10/gfwf5v

Li, W., Wang, F.-D., & Xia, G.-S. (2020). A geometry-attentional network for ALS point cloud classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *164*, 26–40. https://doi.org/10.1016/j.isprsjprs.2020.03.016

Li, X., Ding, M., & Pižurica, A. (2019). Deep feature fusion via two-stream

convolutional neural network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, *58*(4), 2615–2629. https://doi.org/10.1109/TGRS.2019.2952758

Li, X., Zhang, C., Li, W., Ricard, R., Meng, Q., & Zhang, W. (2015). Assessing street-level urban greenery using Google Street View and a modified green view index. *Urban Forestry & Urban Greening*, *14*(3), 675–685. https://doi.org/10.1016/j.ufug.2015.06.006

Li, Z., & Snavely, N. (2018). CGIntrinsics: Better Intrinsic Image Decomposition Through Physically-Based Rendering. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer Vision – ECCV 2018* (Vol. 11207, pp. 381–399). Springer International Publishing. https://doi.org/10.1007/978-3-030-01219-9_23

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2117–2125.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. *European Conference on Computer Vision*, 740–755.

Liu, H., Xu, Y., Zhang, J., Zhu, J., Li, Y., & Hoi, S. C. H. (2020). DeepFacade: A Deep Learning Approach to Facade Parsing With Symmetric Loss. *IEEE Transactions on Multimedia*, *22*(12), 3153–3165. https://doi.org/10.1109/TMM.2020.2971431

Liu, J., Cao, X., Zhou, H., Li, L., Liu, X., Zhao, P., & Dong, J. (2021). A digital twin-driven approach towards traceability and dynamic control for processing quality. *Advanced Engineering Informatics*, *50*, 101395. https://doi.org/10.1016/j.aei.2021.101395

Liu, L., Zhang, X., Wan, X., Zhou, S., & Gao, Z. (2022). Digital twin-driven surface roughness prediction and process parameter adaptive optimization. *Advanced*

Engineering Informatics, *51*, 101470. https://doi.org/10.1016/j.aei.2021.101470

Liu, X., Wang, X., Wright, G., Cheng, J. C., Li, X., & Liu, R. (2017). A state-of-the-art review on the integration of Building Information Modeling (BIM) and Geographic Information System (GIS). *ISPRS International Journal of Geo-Information*, *6*(2), 53. https://doi.org/10.3390/ijgi6020053

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., & Bachem, O. (2019). Challenging common assumptions in the unsupervised learning of disentangled representations. *International Conference on Machine Learning*, 4114–4124.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440.

Lu, X., Yin, J., Ding, Y., & Chen, P. (2010). Investigation and quantitative analysis of urban color: A case study of shennan avenue in shenzhen. *City Planning Review*, *12*, 88–92.

Ma, W., Ma, W., Xu, S., & Zha, H. (2020). Pyramid ALKNet for semantic parsing of building facade image. *IEEE Geoscience and Remote Sensing Letters*, *18*(6), 1009–1013. https://doi.org/10.1109/LGRS.2020.2993451

Marchand, K. A., Earl, W. R., Davis, C. E., Conrath, E. J., & Hadjioannou, M. (2018). Extending building facade performance requirements for blast: Hazard and injury assessment investigations. *Structures Congress 2018: Bridges, Transportation Structures, and Nonbuilding Structures*, 509–517. https://doi.org/10.1061/9780784481332.046

Martinez, A., & Choi, J.-H. (2017). Exploring the potential use of building facade information to estimate energy performance. *Sustainable Cities and Society*, *35*, 511–521. https://doi.org/10.1016/j.scs.2017.07.022

Martinović, A., Mathias, M., Weissenberg, J., & Gool, L. V. (2012). A three-layered

approach to facade parsing. *European Conference on Computer Vision*, 416–429.

Mazzeo, P. L., Giove, L., Moramarco, G. M., Spagnolo, P., & Leo, M. (2011). HSV and RGB color histograms comparing for objects tracking among non overlapping FOVs, using CBTF. *2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 498–503. https://doi.org/10/d9285z

Meijer, A., & Bolívar, M. P. R. (2016). Governing the smart city: A review of the literature on smart urban governance. *International Review of Administrative Sciences*, *82*(2), 392–408. https://doi.org/10.1177/0020852314564308

Ministry of Land, Infrastructure, Transport and Tourism of Japan. (n.d.). *Project PLATEAU*. Retrieved July 14, 2021, from https://www.geospatial.jp/ckan/dataset/plateau

Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. *ArXiv Preprint ArXiv:1802.05957*.

Mori, S., Ikeda, S., & Saito, H. (2017). A survey of diminished reality: Techniques for visually concealing, eliminating, and seeing through real objects. *IPSJ Transactions on Computer Vision and Applications*, *9*(1), 17. https://doi.org/10.1186/s41074-017-0028-1

Mortyzhang. (2022a). *The SOTA model for street view classification* [Python]. https://github.com/Mortyzhang/Nanjing-street-view-datasets-and-classification-tasks (Original work published 2021)

Mortyzhang. (2022). *Mortyzhang/Facade-color-calculation-based-on-colorcard* [Python]. https://github.com/Mortyzhang/Facade-color-calculation-based-on-colorcard (Original work published 2020)

Mortyzhang. (2022b). *A format conversion tool: From mask to COCO format for instance semgmentation training* [Jupyter Notebook]. https://github.com/Mortyzhang/Mask2polygon_tool/blob/dbbac7f1d1326c17a

bf1bd3abab8a7123da3ca2e/README.md (Original work published 2021)

Mouratidis, K., & Poortinga, W. (2020). Built environment, urban vitality and social cohesion: Do vibrant neighborhoods foster strong communities? *Landscape and Urban Planning*, *204*, 103951. https://doi.org/10.1016/j.landurbplan.2020.103951

Nathan Mundhenk, T., Ho, D., & Chen, B. Y. (2018). Improvements to context based self-supervised learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9339–9348.

Nguyen, L., & Teller, J. (2017). Color in the urban environment: A user-oriented protocol for chromatic characterization and the development of a parametric typology. *Color Research & Application*, *42*(1), 131–142. https://doi.org/10/f9x84b

Nitsche, M., & Maureen, M. (2004). Play it again sam: Film performance, virtual environments and game engines. *Visions in Performance: The Impact of Digital Technologies, Carver G., Beardon C.,(Eds.). Swets & Zeitlinger*, *4*.

Niu, X., & Qin, S. (2021). Integrating crowd-/service-sourcing into digital twin for advanced manufacturing service innovation. *Advanced Engineering Informatics*, *50*, 101422. https://doi.org/10.1016/j.aei.2021.101422

NVIDIA Omniverse Replicator. (2022). *https://developer.nvidia.com/nvidia-omniverse-platform*. Accessed: 2022-03-07.

OpenStreetMap. (2021). *Available online: https://www.openstreetmap.org*. (Accessed on 31 July 2021).

Öztürk, A. E., & Erçelebi, E. (2021). Real UAV-Bird Image Classification Using CNN with a Synthetic Dataset. *Applied Sciences*, *11*(9), 3863. https://doi.org/10.3390/app11093863

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2536–2544.

Piccardo, C., Dodoo, A., Gustavsson, L., & Tettey, U. (2020). Retrofitting with different building materials: Life-cycle primary energy implications. *Energy*, *192*, 116648. https://doi.org/10.1016/j.energy.2019.116648

PLATEAU. (2022). Plateau. https://www.mlit.go.jp/plateau/

Poucin, F., Kraus, A., & Simon, M. (2021). Boosting Instance Segmentation with Synthetic Data: A study to overcome the limits of real world data sets. *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 945–953. https://doi.org/10.1109/ICCVW54120.2021.00110

Prakash, A., Boochoon, S., Brophy, M., Acuna, D., Cameracci, E., State, G., Shapira, O., & Birchfield, S. (2019). Structured domain randomization: Bridging the reality gap by context-aware synthetic data. *2019 International Conference on Robotics and Automation (ICRA)*, 7249–7255. https://doi.org/10.1109/ICRA.2019.8794443

Quiñonero-Candela, J., Sugiyama, M., Lawrence, N. D., & Schwaighofer, A. (2009). *Dataset shift in machine learning*. Mit Press.

Rahmani, K., & Mayer, H. (2018). HIGH QUALITY FACADE SEGMENTATION BASED ON STRUCTURED RANDOM FOREST, REGION PROPOSAL NETWORK AND RECTANGULAR FITTING. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, *4*(2).

Riemenschneider, H., Krispel, U., Thaller, W., Donoser, M., Havemann, S., Fellner, D., & Bischof, H. (2012). Irregular lattices for complex shape grammar facade parsing. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 1640–1647. https://doi.org/10.1109/CVPR.2012.6247857

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., & Lopez, A. M. (2016). The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3234–3243. https://doi.org/10.1109/CVPR.2016.352

Ros, G., Stent, S., Alcantarilla, P. F., & Watanabe, T. (2016). Training constrained deconvolutional networks for road scene semantic segmentation. *ArXiv Preprint ArXiv:1604.01545*.

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision*, *77*(1–3), 157–173.

Saleh, F. S., Aliakbarian, M. S., Salzmann, M., Petersson, L., & Alvarez, J. M. (2018). Effective use of synthetic data for urban scene semantic segmentation. *Proceedings of the European Conference on Computer Vision (ECCV)*, 84–100.

Sara, U., Akter, M., & Uddin, M. S. (2019). Image quality assessment through FSIM, SSIM, MSE and PSNR—a comparative study. *Journal of Computer and Communications*, *7*(3), 8–18. https://doi.org/10/ggr3gf

Schumacher, A., Nemeth, T., & Sihn, W. (2019). Roadmapping towards industrial digitalization based on an Industry 4.0 maturity model for manufacturing enterprises. *Procedia Cirp*, *79*, 409–414. https://doi.org/10.1016/j.procir.2019.02.110

Schwarz, M., Milan, A., Periyasamy, A. S., & Behnke, S. (2018). RGB-D object detection and semantic segmentation for autonomous manipulation in clutter. *The International Journal of Robotics Research*, *37*(4–5), 437–451. https://doi.org/10/gdkm7q

Settles, B. (2009). *Active learning literature survey*.

Shahat, E., Hyun, C. T., & Yeom, C. (2021). City Digital Twin Potentials: A Review and Research Agenda. *Sustainability*, *13*(6), 3386. https://doi.org/10.3390/su13063386

Sheikh, H. R., Bovik, A. C., & Veciana, G. de. (2005). An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on Image Processing*, *14*(12), 2117–2128. https://doi.org/10/dc6cn3

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for

deep learning. *Journal of Big Data*, *6*(1), 1–48. https://doi.org/10/ggb3hw

Siddique, N., Paheding, S., Elkin, C. P., & Devabhaktuni, V. (2021). U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access*. https://doi.org/10.1109/ACCESS.2021.3086020

Sorokin, A., & Forsyth, D. (2008). Utility data annotation with amazon mechanical turk. *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 1–8. https://doi.org/10.1109/CVPRW.2008.4562953

Sultana, M., & Storch, I. (2021). Suitability of open digital species records for assessing biodiversity patterns in cities: A case study using avian records. *Journal of Urban Ecology*, *7*(1), juab014. https://doi.org/10.1093/jue/juab014

Sun, B., & Saenko, K. (2014). From Virtual to Reality: Fast Adaptation of Virtual Object Detectors to Real Domains. *BMVC*, *1*(2), 3. https://doi.org/10.1109/TPAMI.2013.163

Sun, Y., Liu, M., & Meng, M. Q.-H. (2017). Improving RGB-D SLAM in dynamic environments: A motion removal approach. *Robotics and Autonomous Systems*, *89*, 110–122. https://doi.org/10/f9n37m

Sun, Y., Zhu, H., Zhuang, F., Gu, J., & He, Q. (2018). Exploring the urban region-of-interest through the analysis of online map search queries. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2269–2278. https://doi.org/10/ghj6mp

Szcześniak, J. T., Ang, Y. Q., Letellier-Duchesne, S., & Reinhart, C. F. (2022). A method for using street view imagery to auto-extract window-to-wall ratios and its relevance for urban-level daylighting and energy simulations. *Building and Environment*, *207*, 108108. https://doi.org/10.1016/j.buildenv.2021.108108

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1).

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the

Inception Architecture for Computer Vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818–2826. https://doi.org/10/gftjd7

Tang, T., Yang, J., Du, B., & Tang, L. (2019). Down-Sampling Based Rate Control for Mobile Screen Video Coding. *IEEE Access*, *7*, 139560–139570. https://doi.org/10.1109/ACCESS.2019.2943887

Tardioli, G., Kerrigan, R., Oates, M., O'Donnell, J., & Finn, D. P. (2018). Identification of representative buildings and building groups in urban datasets using a novel pre-processing, classification, clustering and predictive modelling approach. *Building and Environment*, *140*, 90–106. https://doi.org/10/gdxmhn

Teboul, O., Kokkinos, I., Simon, L., Koutsourakis, P., & Paragios, N. (2011). Shape grammar parsing via reinforcement learning. *CVPR 2011*, 2273–2280. https://doi.org/10.1109/CVPR.2011.5995319

Teboul, O., Kokkinos, I., Simon, L., Koutsourakis, P., & Paragios, N. (2012). Parsing facades with shape grammars and reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(7), 1744–1756. https://doi.org/10.1109/TPAMI.2012.252

Toda, Y., Okura, F., Ito, J., Okada, S., Kinoshita, T., Tsuji, H., & Saisho, D. (2020). Training instance segmentation neural network with synthetic datasets for crop seed phenotyping. *Communications Biology*, *3*(1), 173. https://doi.org/10.1038/s42003-020-0905-5

Tremblay, J., Prakash, A., Acuna, D., Brophy, M., Jampani, V., Anil, C., To, T., Cameracci, E., Boochoon, S., & Birchfield, S. (2018). Training deep networks with synthetic data: Bridging the reality gap by domain randomization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 969–977.

Tyleček, R., & Šára, R. (2013). Spatial Pattern Templates for Recognition of Objects with Regular Structure. In J. Weickert, M. Hein, & B. Schiele (Eds.), *Pattern*

Recognition (Vol. 8142, pp. 364–374). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-40602-7_39

UK housing: Fit for the future? (2019). *Climate Change Committee*. https://www.theccc.org.uk/publication/uk-housing-fit-for-the-future/

Valada, A., Radwan, N., & Burgard, W. (2018). Incorporating semantic and geometric priors in deep pose regression. *Workshop on Learning and Inference in Robotics: Integrating Structure, Priors and Models at Robotics: Science and Systems (RSS)*, *1*, 3.

van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, *109*(2), 373–440. https://doi.org/10.1007/s10994-019-05855-6

Vazquez, D., Lopez, A. M., Marin, J., Ponsa, D., & Geronimo, D. (2013). Virtual and real world adaptation for pedestrian detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(4), 797–809. https://doi.org/10.1109/TPAMI.2013.163

Virtual Singapore. (2022). https://www.nrf.gov.sg/programmes/virtual-singapore

Wang, B., Yin, C., Luo, H., Cheng, J. C., & Wang, Q. (2021). Fully automated generation of parametric BIM for MEP scenes based on terrestrial laser scanning data. *Automation in Construction*, *125*, 103615. https://doi.org/10.1016/j.autcon.2021.103615

Wang, G., Cao, Y., & Zhang, Y. (2022). Digital twin-driven clamping force control for thin-walled parts. *Advanced Engineering Informatics*, *51*, 101468. https://doi.org/10.1016/j.aei.2021.101468

Wang, J., Zhang, L., & Gou, A. (2021). Study of the color characteristics of residential buildings in Shanghai. *Color Research & Application*, *46*(1), 240–257. https://doi.org/10.1002/col.22565

Wang, M., & Deng, W. (2018). Deep visual domain adaptation: A survey. *Neurocomputing*, *312*, 135–153. https://doi.org/10.1016/j.neucom.2018.05.083

Wang, Y., Chen, Q., Zhu, Q., Liu, L., Li, C., & Zheng, D. (2019). A survey of mobile laser scanning applications and key techniques over urban areas. *Remote Sensing*, *11*(13), 1540. https://doi.org/10.3390/rs11131540

Wang, Y., Ma, Y., Zhu, A., Zhao, H., & Liao, L. (2018). Accurate facade feature extraction method for buildings from three-dimensional point cloud data considering structural information. *ISPRS Journal of Photogrammetry and Remote Sensing*, *139*, 146–153. https://doi.org/10.1016/j.isprsjprs.2017.11.015

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1492–1500.

Xie, X., Chen, J., Li, Y., Shen, L., Ma, K., & Zheng, Y. (2020). Instance-aware self-supervised learning for nuclei segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 341–350.

Xue, F., Wu, L., & Lu, W. (2021). Semantic enrichment of building and city information models: A ten-year review. *Advanced Engineering Informatics*, *47*, 101245. https://doi.org/10.1016/j.aei.2020.101245

Yi, Z., Tang, Q., Azizi, S., Jang, D., & Xu, Z. (2020). Contextual Residual Aggregation for Ultra High-Resolution Image Inpainting. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7505–7514. https://doi.org/10/gg9969

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., & Huang, T. S. (2019). Free-form image inpainting with gated convolution. *Proceedings of the IEEE International Conference on Computer Vision*, 4471–4480.

Yuan, L., & Sun, J. (2012). Automatic exposure correction of consumer photographs. *European Conference on Computer Vision*, 771–785.

Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., Sun, Y., He, T., Mueller, J., & Manmatha, R. (2020). Resnest: Split-attention networks. *ArXiv Preprint ArXiv:2004.08955*.

Zhang, J., Fukuda, T., & Yabuki, N. (2021a). Automatic Object Removal With Obstructed Façades Completion Using Semantic Segmentation and Generative Adversarial Inpainting. *IEEE Access*, *9*, 117486–117495. https://doi.org/10.1109/ACCESS.2021.3106124

Zhang, J., Fukuda, T., & Yabuki, N. (2021b). Development of a City-Scale Approach for Façade Color Measurement with Building Functional Classification Using Deep Learning and Street View Images. *ISPRS International Journal of Geo-Information*, *10*(8), 551. https://doi.org/10.3390/ijgi10080551

Zhang, N., Ji, H., Liu, L., & Wang, G. (2019). Exemplar-based image inpainting using angle-aware patch matching. *EURASIP Journal on Image and Video Processing*, *2019*(1), 1–13. https://doi.org/10/ggb2p9

Zhang, Y., David, P., Foroosh, H., & Gong, B. (2019). A curriculum domain adaptation approach to the semantic segmentation of urban scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(8), 1823–1841. https://doi.org/10/ggv89g

Zheng, H. W., Shen, G. Q., Song, Y., Sun, B., & Hong, J. (2017). Neighborhood sustainability in urban renewal: An assessment framework. *Environment and Planning B: Urban Analytics and City Science*, *44*(5), 903–924. https://doi.org/10/ggr3h4

Zheng, H. W., Shen, G. Q., & Wang, H. (2014). A review of recent studies on sustainable urban renewal. *Habitat International*, *41*, 272–279. https://doi.org/10.1016/j.habitatint.2013.08.006

Zhong, T., Ye, C., Wang, Z., Tang, G., Zhang, W., & Ye, Y. (2021). City-Scale Mapping of Urban Façade Color Using Street-View Imagery. *Remote Sensing*, *13*(8), 1591. https://doi.org/10/gjvr6m

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2018). UNet++: A Nested U-Net Architecture for Medical Image Segmentation. *ArXiv:1807.10165 [Cs, Eess, Stat]*. http://arxiv.org/abs/1807.10165

Zhu, J., Wright, G., Wang, J., & Wang, X. (2018). A critical review of the integration of geographic information system and building information modelling at the data level. *ISPRS International Journal of Geo-Information*, *7*(2), 66. https://doi.org/10.3390/ijgi7020066