CIM-WV: A 2D semantic segmentation dataset of rich window view contents in high-rise, high-density Hong Kong based on photorealistic city information models

Maosu Li[1], Anthony G. O. Yeh[1] and Fan Xue[2*]

Abstract

Large-scale assessment of window views is demanded for precise housing valuation and quantified evidence for improving the built environment, especially in high-rise, high-density cities. However, the absence of a semantic segmentation dataset of window views forbids an accurate pixel-level assessment. This paper presents a City Information Model (CIM)-generated Window View (CIM-WV) dataset comprising 2,000 annotated images collected in the high-rise, high-density urban areas of Hong Kong. The CIM-WV includes seven semantic labels, i.e., building, sky, vegetation, road, waterbody, vehicle, and terrain. Experimental results of training a well-known deep learning (DL) model, DeepLab V3+, on CIM-WV, achieved a high performance (per-class Intersection over Union (IoU) ≥ 86.23%) on segmenting major landscape elements, i.e., building, sky, vegetation, and waterbody, and consistently outperformed the transfer learning on a popular real-world street view dataset, Cityscapes. The DeepLab V3+ model trained on CIM-WV was robust (mIoU ≥ 72.09%) in Hong Kong Island and Kowloon Peninsula, and enhanced the semantic segmentation accuracy of real-world and Google Earth CIM-generated window view images. The contribution of this paper is three-fold. CIM-WV is the first public CIM-generated photorealistic window view dataset with rich semantics. Secondly, comparative analysis shows a more accurate window view assessment using DL from CIM-WV than deep transfer learning from ground-level views. Last, for urban researchers and practitioners, our publicly accessible DL models trained on CIM-WV enable novel multi-source window view-based urban applications including precise real estate valuation, improvement of built environment, and window view-related urban analytics.

Keywords Image dataset, Window view, City Information Models, High-rise buildings, High-density cities, Semantic segmentation, Deep learning

*Correspondence:
Fan Xue
xuef@hku.hk
[1] Department of Urban Planning and Design, The University of Hong Kong, Pokfulam, Hong Kong SAR, China
[2] Department of Real Estate and Construction, The University of Hong Kong, Pokfulam, Hong Kong SAR, China

1 Introduction

Assessment of multi-angle urban views is significant informatics for comprehensively examining the development of urban environment (Shi et al., 2022a, 2022b). For example, researchers assess overhead urban views via remote sensing imagery for monitoring the change in land use (Wang et al., 2022), greenery exposure (Chen et al., 2022), and building morphologies (Liao et al., 2023). Ground-level urban views from street view images are exploited for urban accessibility, vitality, and sustainability (Biljecki & Ito, 2021; He et al., 2022; Yang et al., 2021).

Li *et al. Urban Informatics*   (2024) 3:12

Page 2 of 19

Different from overhead and ground-level urban views, window views of varying heights can serve as a supplement to the urban view hub by vertically examining the urban environment (Li et al., 2021, 2022). As a new angle of urban views, window views depict the neighborhood-built environment that urban dwellers observe long-term from their residential and working places.

Window views have shown high socio-economic values and impacts on multiple urban applications. For example, the assessment results can support precise housing valuation and selection regarding scenic window views, e.g., sea and greenery views (Baranzini & Schaerer, 2011; Jim & Chen, 2009), prioritization of the neighborhood-built environment improvement (Li et al., 2023b, 2023c), and optimization of acquisition of nature view and daylight for architectural design (Laovisutthichai et al., 2021; Zhou & Xue, 2023). In addition, quantified results of window views can guide policymakers of government and planning agencies on quantitative regulation on minimum visual greenery exposure for residences and workplaces (Fisher-Gewirtzman, 2018; HKTPB, 2010). Last, the urban-scale assessment results of window views relate to urban issues, e.g., health, safety, and environmental justice (Helbich et al., 2019; Kuo & Sullivan, 2001) especially in high-rise, high-density urban areas, leading to a new wave of urban analytics for healthy and sustainable urban development.

Assessment of window views has been conducted through manual and simulation methods. Initially, researchers in the fields of physiology, psychology, and urban health manually label the window views for computing window view indicators, e.g., window view proportion and structure (Ulrich, 1984; Stamps III 2005). However, the manual collection limits the large-scale window view assessment. Recently, high-quality simulated window view images generated on photorealistic City Information Models (CIMs) have enabled a large-scale assessment (Li & Samuelson, 2020; Li et al., 2022). Particularly, Li et al. (2022) assessed proportions of view features e.g., greenery, waterbody, sky, and construction for urban-scale windows through DeepLab V3+ (Chen et al., 2018) trained on the real-world street view dataset, Cityscapes (Cordts et al., 2016). However, the photorealistic CIM-generated window views of varying heights are different from ground-level street views captured from the real urban landscape. The absence of annotated photorealistic window view images fails to support an accurate pixel-level semantic segmentation of window views. Thus, it is significant to present an annotated window view image dataset for advancing an accurate urban-scale window view assessment.

This paper presents a photorealistic CIM-generated Window View (CIM-WV) dataset for 2D semantic segmentation of multi-level urban scenes in high-rise, high-density cities. The CIM-WV comprises 2,000 photorealistic window view images with seven semantic labels, i.e., building, sky, vegetation, road, waterbody, vehicle, and terrain. Window view images of CIM-WV were generated from high-rise, high-density urban areas of Hong Kong Island and the Kowloon Peninsula. Each window view image with 900×900 pixels is arranged in two layers: photorealistic image and semantic segmentation mask.

The contribution of the study is three-fold.

i. First, it presents the first public CIM-generated photorealistic window view image dataset with rich semantics. The annotated photorealistic window view imagery supplements the existing semantic segmentation datasets of multi-angle urban views.
ii. Thereafter, we provide a comprehensive evaluation of CIM-WV, including a baseline using DeepLab V3+, a comparative analysis of view segmentation using CIM-WV and Cityscapes, and robustness and transferability analyses of the trained DeepLab V3+ models. Experimental results confirm a more accurate window view assessment using deep learning from CIM-WV than deep transfer learning from ground-level views. The robust DeepLab V3+ model in Hong Kong enhances the semantic segmentation accuracy of real-world and Google Earth CIM-generated window view images.
iii. The publicly accessible deep learning models trained on CIM-WV enable multi-source window view-based applications including precise real estate valuation, improvement of built environment, and window view-related urban analytics.

The remainder of this paper is arranged as follows. The related work in the literature is reviewed in Sect. 2. The descriptions of CIM-WV including the specification, characteristics, and evaluation process are represented in Sect. 3. Section 4 describes the experimental settings and results. The discussion and conclusion are presented in Sects. 5 and 6, respectively.

2 Related work
2.1 Semantic segmentation datasets of urban views
Semantic segmentation datasets for urban views are developed in an imbalanced way. There exist numerous semantic segmentation datasets of overhead and street views in the fields of remote sensing, computer vision, and urban studies. For example, satellite and aerial images with pixel-level semantic annotations such as DeepGlobe (Demir et al., 2018), LoveDA (Wang et al., 2021), and SkyScapes (Azimi et al., 2019) are provided

Li *et al. Urban Informatics*   (2024) 3:12

Page 3 of 19

for landscape element extraction (Zhou et al., 2018), land cover mapping (Wang et al., 2022), and urban management (Liao et al., 2023). In addition, researchers also annotated synthetic and real-world street view image datasets such as SYNTHIA (Ros et al., 2016), Cityscapes (Cordts et al., 2016), and BDD100K (Yu et al., 2020) for ground-level urban scene understanding. The high accessibility of diversified street view datasets has triggered numerous urban studies, e.g., autonomous driving and navigation (Chen et al., 2017) and urban analytics (Xue et al., 2021; Yang et al., 2021). However, the annotated window view image dataset is less available despite the socio-economic values of window views in the real estate market (Jim & Chen, 2009), landscape management (Li et al., 2023b), and urban planning and design (Laovisutthichai et al., 2021) especially in high-rise, high-density cities.

2.2 Current automatic assessment of window view

Automatic assessment methods of window views are constantly evolving, taking full advantage of the increasingly high-resolution CIMs. Particularly, with the high-resolution 3D photorealistic CIMs and deep transfer learning-based semantic segmentation techniques, automatic quantification of urban-scale window views has been enabled for real estate management (Li et al., 2021), architectural design (Laovisutthichai et al., 2021; Li & Samuelson, 2020), and urban planning of greenery space (Li et al., 2023b, 2023c). Deep learning models trained on street view imagery, e.g., Cityscapes were transductively applied to segment the CIM-generated window view images at the urban scale (Li et al., 2021, 2022). However, the current assessment methods fail to support an accurate pixel-level window view assessment. First, ground-level street views centered around roads and vehicles cannot accurately represent window views of varying heights. Thereafter, urban views of the real urban landscape are different from photorealistic views with distortions and inconsistent resolutions. Consequently, transfer learning-based semantic segmentation leads to inaccurate detection of photorealistic window view features. Thus, it is significant to construct a semantic segmentation dataset especially for photorealistic window views to advance an accurate pixel-level window view assessment.

3 The presented window view dataset
3.1 Dataset specification
3.1.1 Overview of CIM-WV
CIM-WV comprises 2,000 photorealistic window view images sampled from 203 buildings in Hong Kong as shown in Fig. 1a. The target areas in Wan Chai and Yau Tsim Mong Districts are within the top high-rise,

high-density zones according to the Hong Kong Planning Standards and Guidelines (HKPlanD, 2018). In addition, the target areas with a long land reclamation history ranging from the 1850s to 2010s, offer us the most representative and diverse views of the built environment of urban Hong Kong (HKCEDD, 2019). Specifically, Fig. 1b shows window images with varying locations and orientations collected on both sides of Victoria Harbor to represent diversified landscape elements, encompassing sea, multi-style buildings, vegetation, roads, etc. Figures 1c and 1d show window view images of CIM-WV with varied elevations to represent the multi-level urban environment, e.g., street and sky scenes in Hong Kong Island and Kowloon Peninsula, respectively.

Window view images of CIM-WV were generated from publicly available photorealistic CIMs (HKPlanD, 2019a) shared by the Planning Department, Government of Hong Kong SAR. The photorealistic CIMs registered in the Hong Kong 1980 Grid coordinate system (EPSG:2326) were reconstructed from more than 340,000 aerial images through photogrammetry, covering nearly the whole urban areas of Hong Kong Island and Kowloon Peninsula (HKPlanD, 2019b). Figures 2b and a show a photorealistic window view image of CIM-WV vividly representing a landscape scene captured in the real world. The photorealistic window view image was horizontally captured from the center of the window site regardless of the window configuration, such as window shape and frames. The window view image of CIM-WV contains 900×900 pixels, and the field of view is 60 degrees to represent the vision of a normal person. Each window view image was named by combining the IDs of its affiliated building, facade, 3D coordinates (*lon, lat, elevation*), and heading.

CIM-WV includes seven semantic labels, i.e., building, sky, vegetation, road, waterbody, vehicle, and terrain. For example, Table 1 shows building denotes construction with roofs and walls. The sky is the space above the landscape where the sun and clouds appear. Image pixels representing trees, bushes, and grass are labeled as vegetation. The seven semantic labels represent the most commonly seen city objects in Hong Kong. Figure 2c shows a typical segmentation mask with all seven semantic labels for a photorealistic window view as shown in Fig. 2b.

3.1.2 Creation process of CIM-WV
We created the CIM-WV through a semi-automatic method. The inputs were 2D building footprints with building height information and 3D photorealistic CIMs, shared by the Hong Kong Lands Department (HKLandsD, 2014) and Hong Kong Planning Department (HKPlanD, 2019a), respectively. And the output
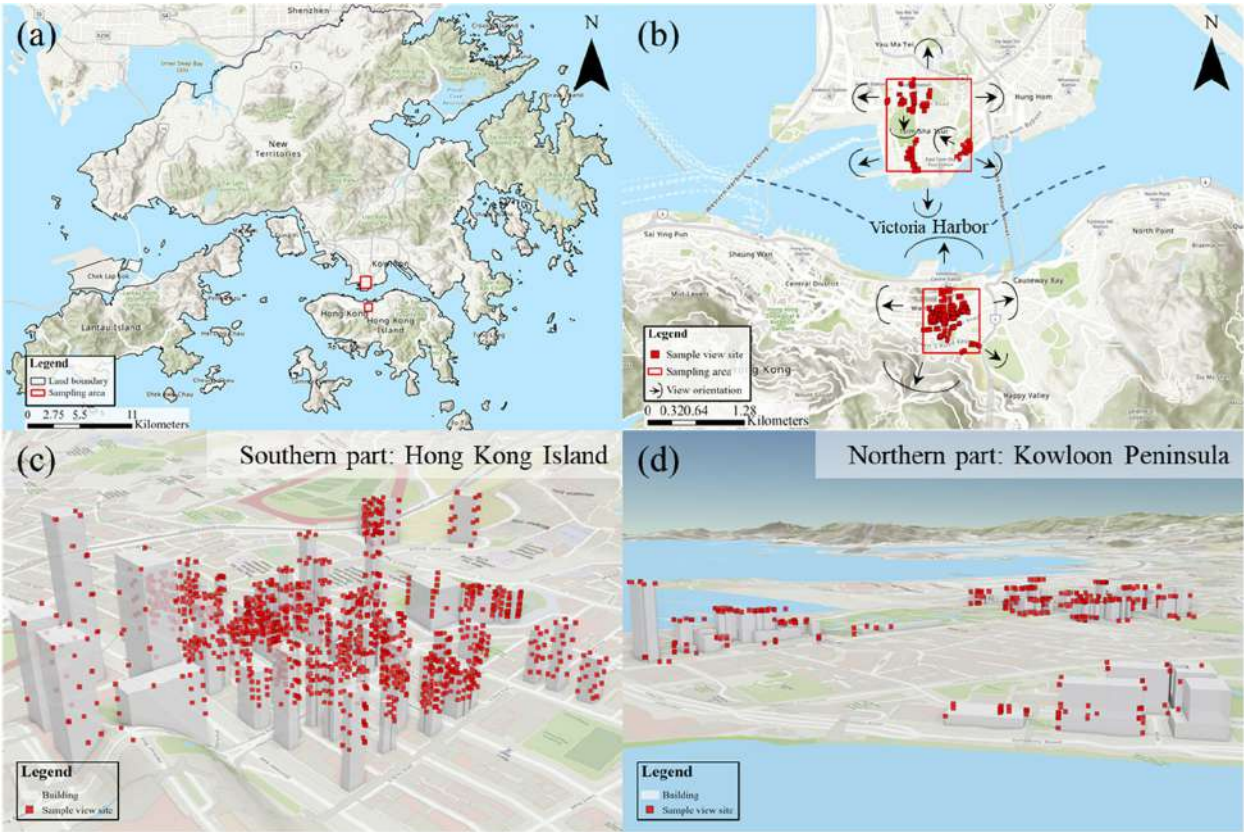
Li *et al. Urban Informatics*   (2024) 3:12

Page 4 of 19

Fig. 1 Collection area of CIM-WV in Hong Kong. (**a**) Location, (**b**) two sampling urban areas, and 2,000 window view sites in (**c**) the Hong Kong Island, and (**d**) the Kowloon Peninsula
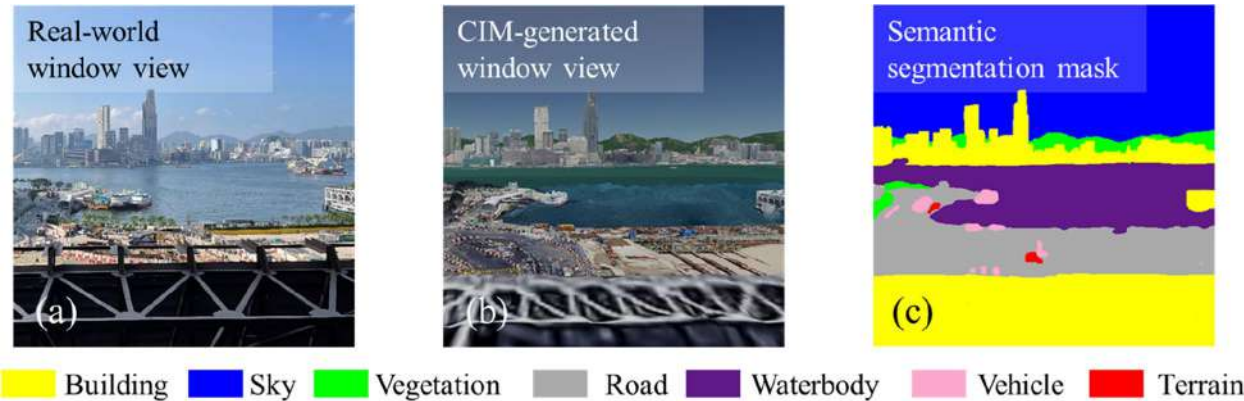

Fig. 2 An example window view. (**a**)Window view photo captured from the real world, (**b**) photorealistic window view image of CIM-WV, and (**c**) the semantic segmentation mask

was the CIM-WV. A three-step workflow for creating CIM-WV comprises batch generation of window view images, cleansing of window view images, and semantic annotation.

The first step is batch generation of window view images. Making full use of the 3D geo-visualization platform, Cesium (Cesium GS 2022), we followed Li et al.'s (2022) method to place a virtual camera on each window

Li *et al. Urban Informatics*   (2024) 3:12

Page 5 of 19

Table 1 Description of seven semantic labels of CIM-WV

Type	Description
Building	Construction with roofs and walls
Sky	The space above the landscape where the sun and clouds appear
Vegetation	Plants, e.g., trees, bushes, and grass
Road	Constructed paths without covers, e.g., streets and walk-ways
Waterbody	Areas of water, e.g., pools, ponds, lakes, rivers, and sea
Vehicle	Car, ship, vessel, etc
Terrain	Unconstructed or rugged land areas, e.g., bared earth

site to capture the outside photorealistic view as shown in Fig. 3b. Specifically, Fig. 3a shows the location parameters of the virtual camera in the WGS-84 coordinate system (EPSG: 4326), i.e., *lon*, *lat*, and *elevation* were set according to the window site's 3D coordinates. Rotation parameters of the virtual camera i.e., *heading*, *pitch*, and *roll*, were set to the heading value of the window site, 0, and 0, respectively. And the field of view of the camera was set to 60 degrees to generate the window view images of CIM-WV. 3D coordinates and heading value of the window site were computed from geometric and height attributes of shared building footprints. To sample window views of diversified landscape scenes of varying heights cost-effectively, we followed Li et al.'s (2022) method to generate window views by setting intervals of view sites at 20 m and 5 m for large and small building facades, respectively. Last, we named and saved each window view image in the database. Figure 3c shows the window view image is named by the combination of its four IDs, i.e., unique ID, affiliated building ID, affiliated facade ID, and view ID, and location and rotation parameters of the virtual camera to generate the view.

The next step is cleansing of window view images. We first removed incorrectly modeled window view images by detecting fragmented CIM faces. Photorealistic CIMs in this study are surface modeling of the real urban landscape. The placement of the virtual camera towards the interior of photorealistic CIMs can lead to incorrect representation of window views, which often co-occur with visible fragmented CIM faces. Thereafter, we manually selected window view images to comprehensively represent scenes of the multi-level urban environment. Figure 4a shows the varied elevations of window views were distributed among low (0–30 m), middle (30–60 m), and high (≥60 m) regions. Figure 4b shows similar quantities of window views facing the east, west, south, and north. Figure 4c shows the distribution of seven kinds of landscape elements of CIM-WV after the manual selection. Generally, buildings and sky dominate the window view images of CIM-WV.

The last step is semantic annotation. The annotation of a photorealistic window view image leads to a semantic segmentation mask. Figure 5 shows annotations of the seven landscape elements listed in Table 1 in four typical window views, i.e., pure building view, street view, mountain view, and sea view. To ensure a high-quality ground truth, the authors annotated 100 samples of representative window views at pixel level, and then supervised professional annotators hired online (Alibaba, 2023) to annotate the remaining 1,900 window views.

3.2 The characteristics of CIM-WV
3.2.1 CIM-generated window view images
CIM-WV is the first window view image dataset generated on photorealistic CIMs. On the other hand, the quality of photorealistic CIMs leads to limitations of the data quality of CIM-WV. First, there exist inconsistent colors of window view images in the Hong Kong Island and Kowloon Peninsula due to discontinuous collection dates of CIMs. For example, Fig. 6 shows the holistically inconsistent color settings, e.g., brightness and color contrast of window view images in Hong Kong Island and Kowloon Peninsula. The color inconsistency of window view images of CIM-WV may enhance the robustness of the trained deep learning models for segmenting multi-source CIM-generated window view images.

In addition, the window view images of CIM-WV own three kinds of representation defects inherited from low-resolution photorealistic CIMs. First, bottom-level window view images tend to be more distorted than upper-level ones. For example, Fig. 7a shows more distortions exist at the bottom parts of buildings from the photorealistic CIMs. In addition, close-range window views are more blurred than distant ones. Figure 7b shows the limited resolution of CIM textures leads to blurred close-range views. Last, complex landscape surfaces are more distorted than simple ones. Figure 7c shows complex building surfaces in the red rectangles are more distorted than ones in the green rectangles at similar view distance and elevation.

3.2.2 Representation of the multi-level urban environment
CIM-WV represents the multi-level urban environment of high-rise, high-density areas. Window view images of CIM-WV depict diversified urban scenes of Hong Kong at different locations, elevations, and orientations. Figure 8a shows the window views facing the streets, buildings, and open space at the same elevation and orientation but from different buildings. Figure 8b shows three window views at changed elevations of the same building facade. Figure 8c shows the varied window views of multiple orientations of the same building.
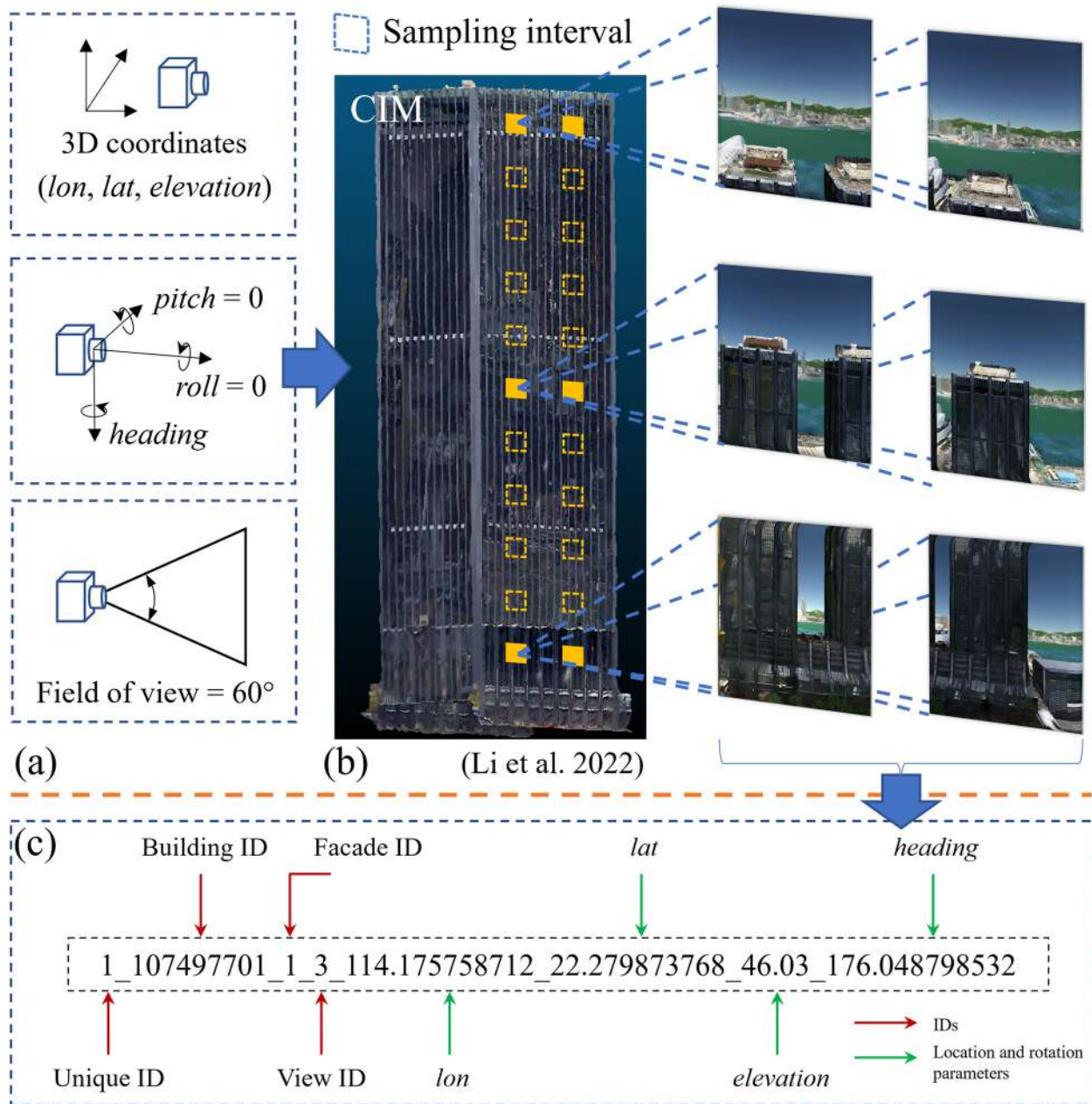
Fig. 3 Batch generation of window view images. (**a**) Camera settings based on window site information, (**b**) image generation process (Li et al., 2022), and (**c**) naming rule of the window view images of CIM-WV

3.3 Evaluation of CIM-WV

We provide an evaluation on assessing semantic segmentation of window views using CIM-WV. First, a baseline of a popular deep learning model, DeepLab V3+on CIM-WV was provided regarding three backbones, i.e., ResNet, Dilated ResNet (DRN), and Xception, and two optional values of the output stride (OS) of the model, i.e., 8 and 16. Thereafter, we compared the segmentation accuracy of DeepLab V3+trained on the proposed CIM-WV with the one trained on Cityscapes transductively used in (Li et al., 2022). Next, we validated the robustness of trained DeepLab V3+models in multiple areas of Hong Kong, i.e., the test sets of CIM-WV in the Hong Kong Island and Kowloon Peninsula, and non-CIM-WV images in the western Hong Kong Island. Last, the transferability of the trained DeepLab V3+model was validated for multi-source window view images, e.g., the real-world images in Hong Kong and Google Earth
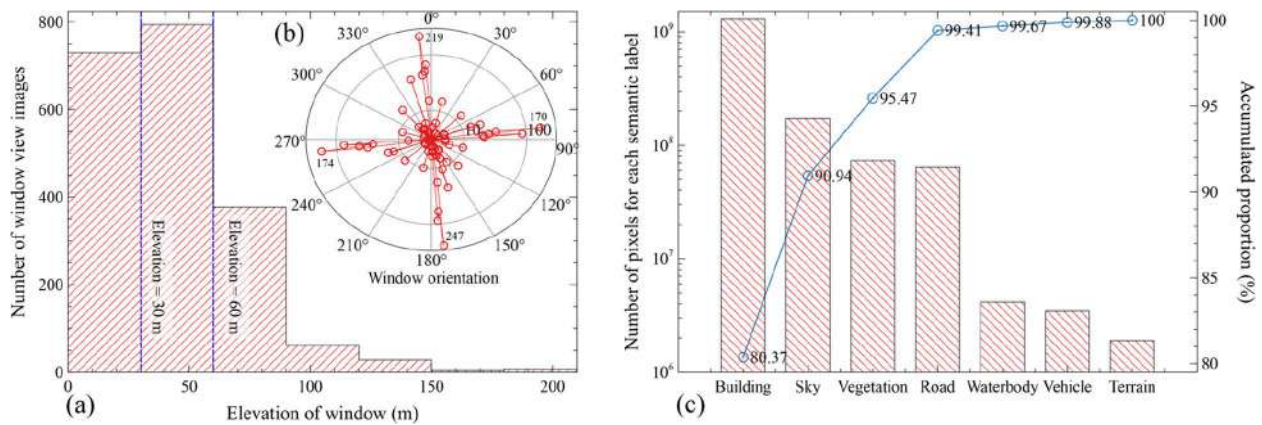
Li *et al. Urban Informatics*   (2024) 3:12

Page 7 of 19

Fig. 4 Quantity distributions of window sites by (**a**) elevation and (**b**) heading and (**c**) quantity distribution of semantic labels of CIM-WV
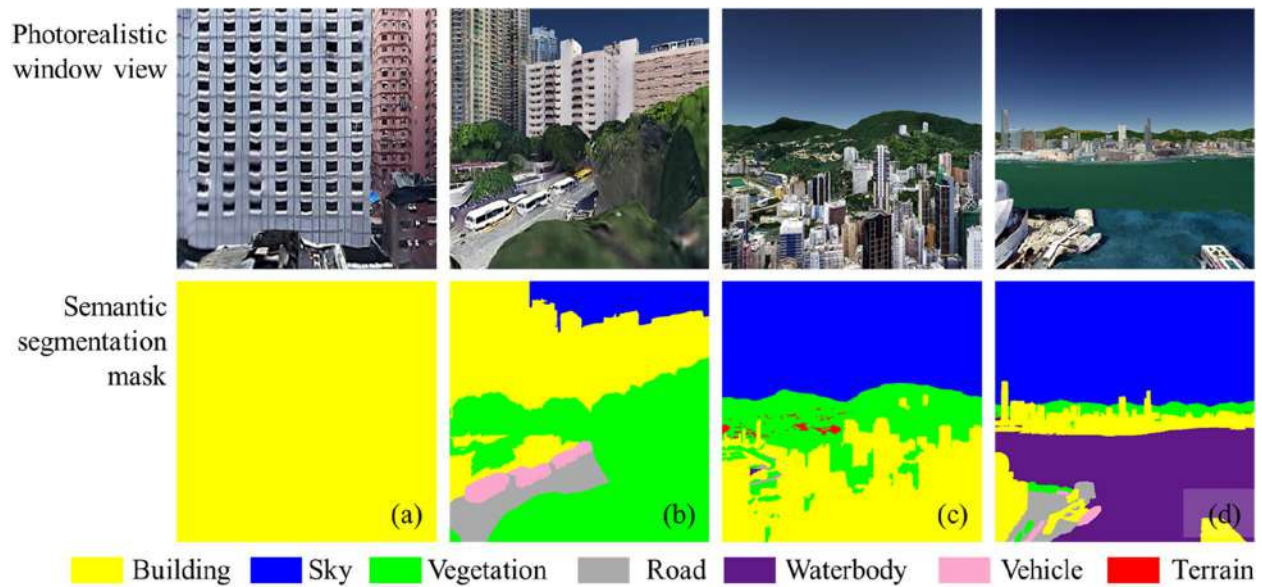

Fig. 5 Typical window views of CIM-WV with semantic annotations. (**a**) building view, (**b**) street view, (**c**) mountain view, and (**d**) sea view

CIM-generated images in two other typical high-rise, high-density cities, i.e., New York and Singapore.

For the first two analyses, we selected 1,400 window view images of CIM-WV as the training set, 300 window view images for validation, and the last 300 window view images as the test set. The goal of the control of dataset splits was similar quantity distributions of the seven annotated view elements in the training, validation, and test sets. First, we clustered the window view images by their quantity distributions of seven annotated view elements in Hong Kong Island and the Kowloon Peninsula, respectively. Then, window view images in each cluster were randomly assigned into the training, validation, and test sets according to the ratio, 70:15:15. Last, we

examined each set of CIM-WV to ensure a similar and balanced quantity distribution of seven annotated view elements. Figure 9a shows a similar quantity distribution of seven annotated view elements in three sets. The numbers of pixels especially for waterbody, vehicle, and terrain in validation and test sets were non-zero and almost equal. Meanwhile, window view images of Hong Kong Island and the Kowloon Peninsula were distributed into the training, validation, and test sets with a similar ratio as shown in Fig. 9b.

For the third analysis, 60 more window view images from the western Hong Kong Island, more than 3.05 km away from the target areas of CIM-WV, were manually annotated to test the robustness of the trained DeepLab
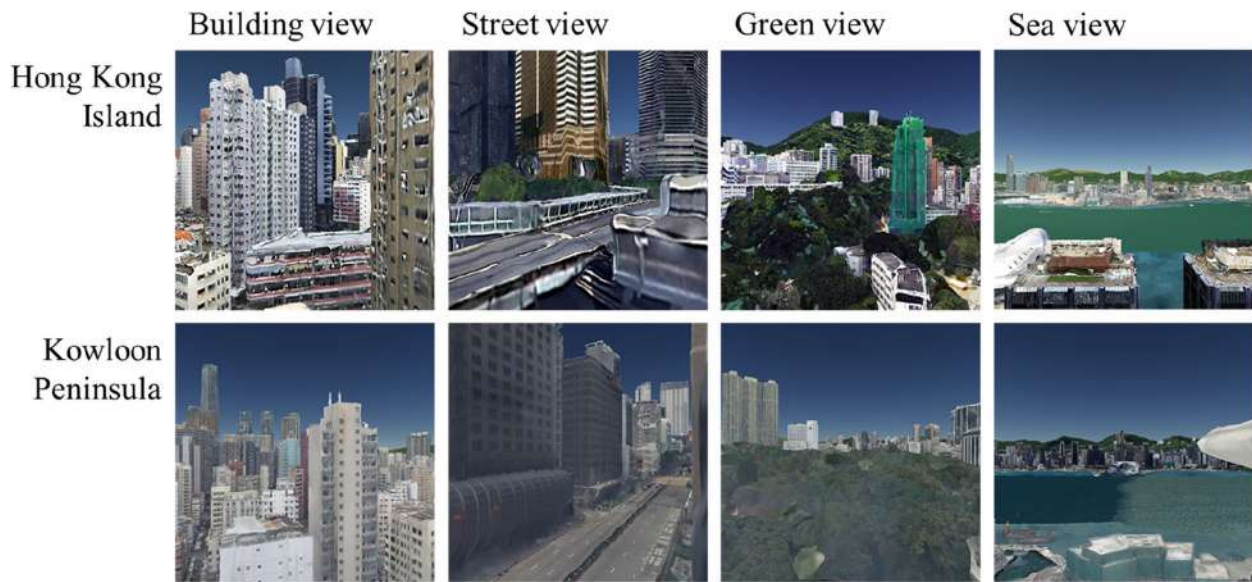
Li *et al. Urban Informatics*   (2024) 3:12

Page 8 of 19

Fig. 6 Inconsistent colors of window view images collected in the Hong Kong Island and the Kowloon Peninsula
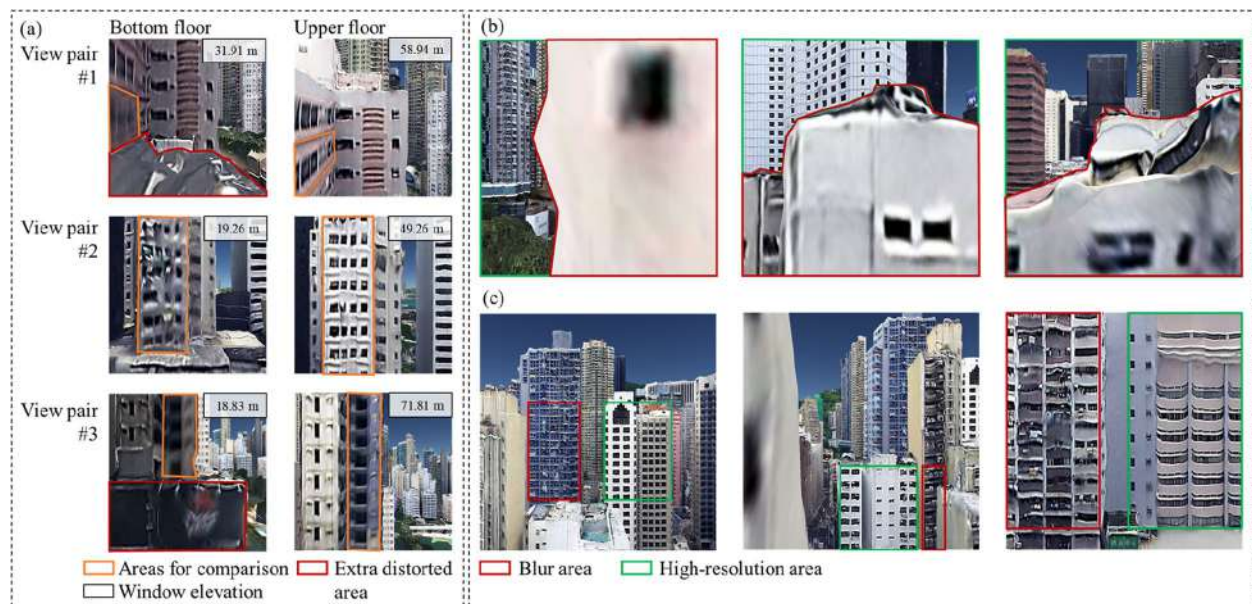

Fig. 7 Three kinds of low-resolution representation of CIM-WV. (**a**) Comparison of distortions of window view elements at bottom and upper floors, (**b**) comparison of blurs of close-range and distant views, and (**c**) comparison of distortions of window views against complex and simple building surfaces

V3 + models. Last, we annotated 30 real-world window view images in Hong Kong and 30 Google Earth CIM-generated photorealistic window view images from each of another two cities, i.e., New York and Singapore to initially test the transferability of the trained DeepLab V3 + models. To compare the segmentation performances, we trained a DeepLab V3 + model on multi-source window view images only and meanwhile finetuned another DeepLab V3 + model trained on CIM-WV by feeding multi-source window view images.

Three indicators including Overall Accuracy (OA), mean class Accuracy (mAcc), and mean Intersection over Union (mIoU) were used to evaluate the performance of six trained DeepLab V3 + models. OA reports
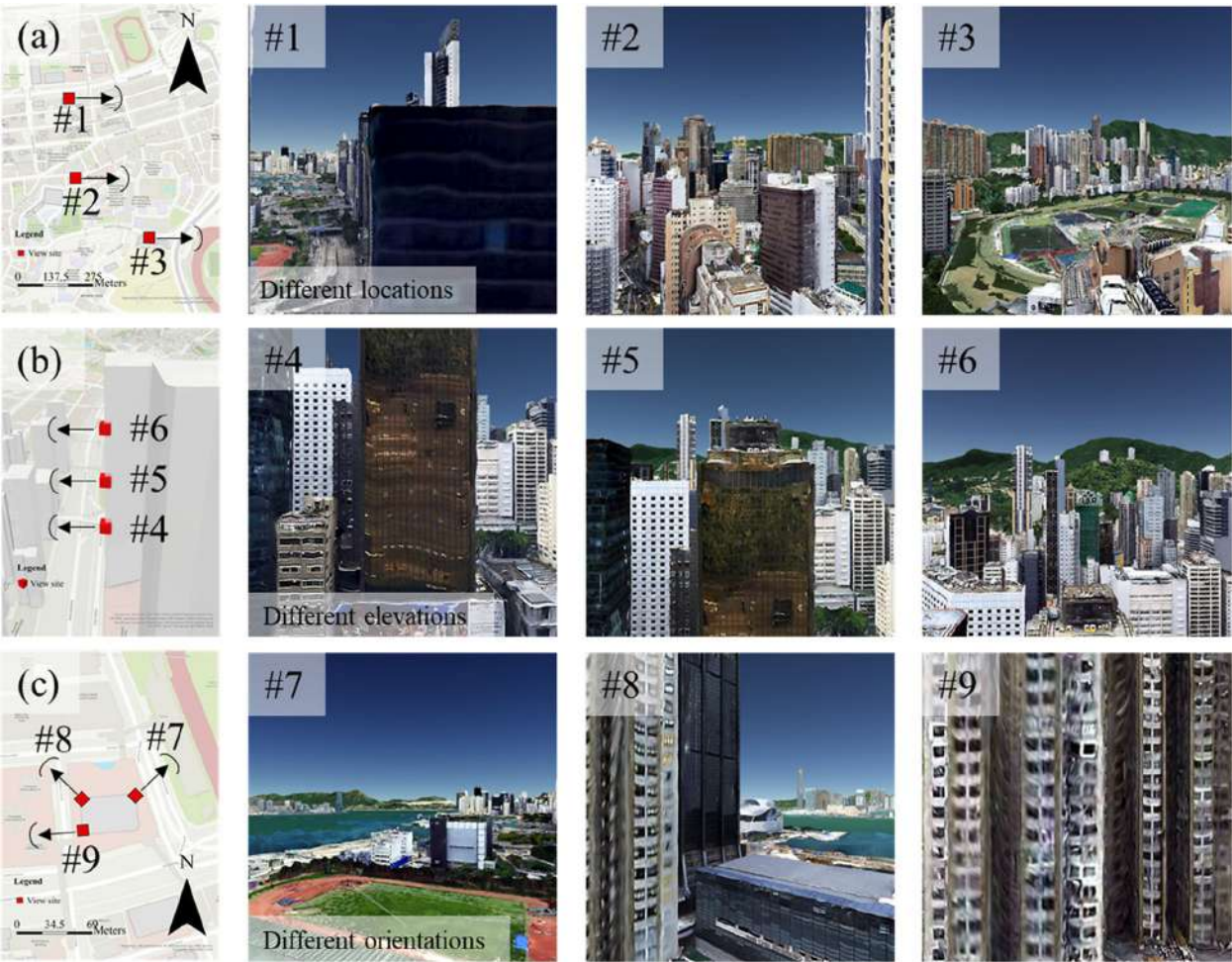
Fig. 8 Window view images of CIM-WV representing diversified urban scenes of Hong Kong at different (**a**) locations, (**b**) elevations, and (**c**) orientations
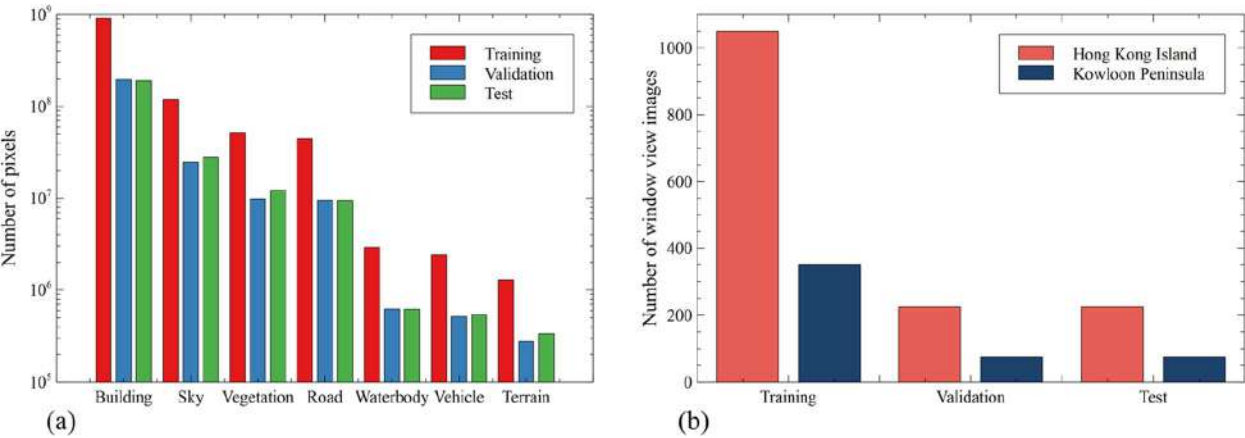

Fig. 9 Quantity distributions of window view elements in training, validation, and test sets by (**a**) semantic labels and (**b**) locations

the percentage of total pixels that are correctly classified, whereas mAcc represents the average percentage of pixels that are correctly classified in each semantic class *l*. mIoU is the average of the IoUs of the set of seven semantic classes *L*, which indicates the average magnitude of the detection confusion between each semantic label *l*. We compute OA, mAcc, and mIoU using Eqs. 1, 2, and 3, respectively.

$$OA = \sum_{L} |P_{ll}| / \sum_{L} |P_l|, \tag{1}$$

$$mAcc = \sum_{L} (|P_{ll}| / |P_l|)/7, \tag{2}$$

$$mIoU = \sum_{L} (IoU_l) / 7, \text{ where } IoU_l = |P_{ll}| / (\sum_{o} |P_{lo}| + \sum_{o} |P_{ol}| - |P_{ll}|), \tag{3}$$

Table 2 Controlled hypermeters of six models

Parameter	Value
Batch size	4
Loss function	Cross entropy
Optimizer	Stochastic gradient descent
Learning rate	0.005
Scheduler mode	Polynomial
Momentum factor	0.9
Weight decay	5e-4

where P_{ll} is the set of pixels belonging to the class *l* and predicted as *l*, P_l is the whole set of pixels manually assigned the label *l* as ground truth, P_{lo} is the set of pixels belonging to the class *l* but wrongly predicted into other classes *o*, and P_{ol} is the set of pixels belonging to other classes *o* but wrongly predicted into the class *l*, and $|\cdot|$ is the operator indicating the number of pixels, e.g., in the set P_{ll} and P_l. Three indicators, i.e., OA, mAcc, and mIoU all range between 0 and 1. The high values of the metrics indicate an accurate semantic segmentation, where OA measures pixel-level performance and mIoU emphasizes the performance at the class level.

4 Experimental tests
4.1 Experimental settings
The experiments were implemented on a high-performance computing cluster with 7 servers, each of which owns dual Intel Xeon 6226R (16 core) CPUs, 384GB RAM, 4×NVIDIA V100 (32GB) SXM2 GPUs, and a CentOS 8 system. Specifically, each DeepLab V3+model with specific settings described in Sect. 3.3 was trained on assigned 16 core CPUs, 64GB RAM, and one NVIDIA V100 (32GB) SXM2 GPU. All six models were trained with the environment of PyTorch (ver. 1.10) and Python (ver. 3.7). We finetuned seven hypermeters, i.e., the batch size, loss function, optimizer, learning rate, scheduler mode, momentum factor, and weight decay to compare the model performances. Table 2 lists the finally controlled values of the seven hypermeters to achieve holistically optimal results of six models we could acquire. The early-stop method was applied to avoid overfitting and we saved the checkpoint with minimal validation loss for comparison. Last, the four experimental tests were implemented in the same development environment.

4.2 Results
4.2.1 Baseline of CIM-WV via DeepLab V3+
Table 3 lists the performances of six variants of DeepLab V3+on the test set of CIM-WV. Overall, the six trained models achieved similar performances on segmentation, with OA, mAcc, and mIoU consistently equal to or greater than 97.49%, 87.96%, and 76.55%, respectively. DeepLab V3+with the backbone, Xception, and OS at 8 achieved the highest mAcc and mIoU at 91.17% and 77.93%, respectively, while DeepLab V3+with the backbone, DRN, and OS at 16 achieved the highest value of OA at 97.80%. By contrast, the trained model with the backbone, ResNet, and OS at 16 performed poorly with the lowest OA, mAcc, and mIoU.

For all six models, landscape elements in the window view, i.e., buildings, sky, vegetation, and waterbody were mostly detected, with per-IoUs beyond 83.21%. Figure 10a shows a typical window view image with the four well-segmented landscape elements. By contrast, roads, vehicles, and terrain were poorly segmented, with per-IoUs lower than 72.93%. There existed three reasons for the low-performance segmentation of roads, vehicles, and terrain. First, there existed incorrect recognitions of close-range roads in the window view as parts of nearby buildings, and close-range flat building roofs as roads, which led to low per-class IoUs of the road (per-class IoUs ≤ 72.93%). For example, the high-performance model (Backbone = Xception, OS = 8) detected the close-range flat building roofs as roads due to similar textures as shown in Fig. 10b. In addition, considerable false negative and false positive detections of vehicles caused by incomplete vehicle representation in limited pixels of the window views led to a low-performance segmentation (per-class IoUs ≤ 48.31%). For example, Fig. 10c shows the profiles of small vehicles on the building podium were not recognized. Last, there existed confusion among

Table 3 Performances of DeepLab V3 + with different backbones and values of OS

Backbone	OS (px)	Overall metric (%)				Per-class IoU (%)						
		Training efficiency t (s) / epoch	OA	mAcc	mIoU	Building	Sky	Vegetation	Road	Waterbody	Vehicle	Terrain
ResNet	8	269.152	97.59	89.22	77.08	97.48	<u>98.64</u>	<u>83.21</u>	70.78	<u>86.48</u>	47.44	55.54
ResNet	16	101.311	<u>97.49</u>	<u>87.96</u>	<u>76.55</u>	<u>97.35</u>	98.73	84.37	<u>67.19</u>	89.01	46.13	53.06
Xception	8	302.781	97.77	**91.17**	**77.93**	**97.69**	98.83	85.35	71.75	87.64	**48.31**	55.96
Xception	16	116.438	97.68	90.74	77.22	97.51	98.73	**86.23**	70.78	89.00	46.19	<u>52.09</u>
DRN	8	185.567	97.74	88.74	77.64	97.58	98.89	83.89	**72.93**	**89.35**	<u>44.21</u>	**56.62**
DRN	16	187.102	**97.80**	88.90	77.23	97.68	**98.90**	84.83	71.82	88.94	45.44	52.97

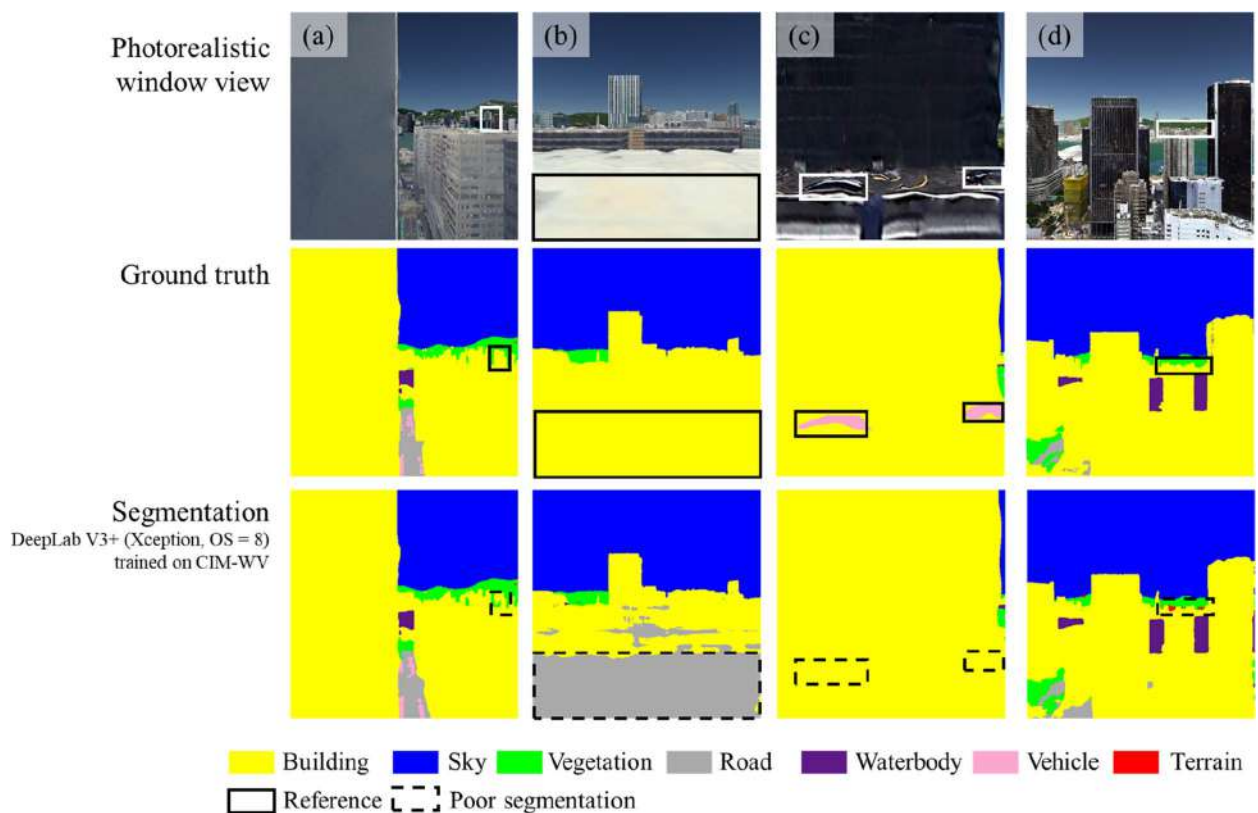The highest and lowest values in each column are in bold and underlined, respectively

Fig. 10 Semantic segmentation results of four example images in the test set using the DeepLab V3 + (Backbone = Xception, OS = 8) trained on CIM-WV. High-performance segmentation of (**a**) buildings, sky, vegetation, and waterbody, and poor segmentations of (**b**) roads, (**c**) vehicles, and (**d**) terrain

terrain, vegetation, and buildings especially at the distant landscape layers of the window views due to their close proximity, similar textures, and limited pixels of representation as shown in Fig. 10d.

In addition, we also trained six variants of DeepLab V3 + models on the sets of Hong Kong Island and Kowloon Peninsula, respectively, and then tested their performances in the local areas. Table 4 shows six DeepLab V3 + models trained on the Hong Kong Island set achieved a consistently high performance (mIoU ≥ 75.54%) in the Island area, while the models trained on the Kowloon Peninsula set achieved mIoUs above 71.81% in Kowloon as shown in Table 5. Window view elements including the building, sky, and waterbody were highly detected (per-class IoUs ≥ 86.47%) by both branches of models trained on the Hong Kong Island set and the Kowloon Peninsula set, whereas roads, vehicles, and terrain were poorly detected with

per-class IoUs ≤ 73.42%. Similar reasons for the poor performance in detecting roads, vehicles, and terrain are i) confusion between roads and buildings, ii) considerable false negative and false positive detection errors of incomplete small vehicles, and iii) confusion among terrain, vegetation, and buildings especially in the distant landscape layers of the window views due to similar textures. In addition, differently, models trained on the Hong Kong Island set segmented vegetation and terrain more accurately, but poorly for vehicles than models trained on the Kowloon Peninsula set, as shown in Tables 4 and 5. Possible reasons include the inconsistent model quality of landscape elements. For example, the modeling quality of vegetation and terrain in Hong Kong Island is the higher of the two, whereas the modeled vehicles in Kowloon Peninsula are more complete and less distorted.

4.2.2 Comparative analysis of photorealistic window view segmentation using CIM-WV and Cityscapes

Table 6 lists the evaluation results of DeepLab V3 + (Backbone = Xception, OS = 16) trained on ImageNet (Deng et al., 2009) and Cityscapes transductively used in Li et al.'s method (2022), and on ImageNet and our proposed CIM-WV, respectively. Evaluation results showed that the DeepLab V3 + model trained on Cityscapes poorly detected all the seven window view elements, with mIoU at 34.14%. Waterbody was fully incorrectly detected as roads and terrain due to the absence of the label in Cityscapes as shown in Fig. 11a. The per-class IoUs of the road, vehicle, and terrain were low (per-class IoUs ≤ 11.98%) due to the significant difference between multi-level window views and ground-level street views. For example, regarding window view as street view, the trained model segmented close-range building facades

Table 4 Performances of six DeepLab V3 + models trained on the Hong Kong Island set

Backbone	OS (px)	Overall metric (%)			Per-class IoU (%)						
		OA	mAcc	mIoU	Building	Sky	Vegetation	Road	Waterbody	Vehicle	Terrain
ResNet	8	<u>97.79</u>	<u>87.07</u>	76.14	<u>97.79</u>	97.92	84.66	<u>68.69</u>	<u>88.04</u>	<u>35.13</u>	**60.73**
ResNet	16	97.81	88.73	75.70	97.95	<u>97.27</u>	84.24	70.19	90.21	36.23	53.79
Xception	8	**97.99**	**89.41**	77.06	97.95	98.53	**85.46**	**73.42**	90.49	37.00	56.53
Xception	16	97.85	89.14	<u>75.54</u>	97.87	98.18	<u>82.94</u>	73.16	89.78	36.04	<u>50.78</u>
DRN	8	97.82	88.90	76.48	97.81	**98.53**	84.33	70.69	90.73	35.41	57.83
DRN	16	97.97	88.79	**77.20**	**98.00**	98.52	85.04	72.22	**91.41**	**37.88**	57.35

The highest and lowest values in each column are in bold and underlined, respectively

Table 5 Performances of six DeepLab V3 + models trained on the Kowloon Peninsula set

Backbone	OS (px)	Overall metric (%)			Per-class IoU (%)						
		OA	mAcc	mIoU	Building	Sky	Vegetation	Road	Waterbody	Vehicle	Terrain
ResNet	8	96.61	**85.49**	72.33	95.99	98.79	69.73	69.46	89.28	62.04	21.01
ResNet	16	<u>96.49</u>	83.96	<u>71.81</u>	<u>95.93</u>	98.75	71.36	<u>64.03</u>	**89.85**	61.37	21.37
Xception	8	**96.86**	85.44	72.54	**96.42**	98.59	72.64	70.47	88.50	<u>55.06</u>	**26.13**
Xception	16	96.81	83.90	71.83	96.20	<u>98.41</u>	**73.73**	69.58	<u>86.47</u>	58.64	19.77
DRN	8	96.75	83.68	72.50	96.10	98.94	<u>68.85</u>	**72.04**	87.38	**65.75**	18.46
DRN	16	96.77	<u>82.66</u>	**73.42**	96.21	**98.95**	70.32	69.65	89.79	64.26	<u>14.76</u>

The highest and lowest values in each column are in bold and underlined, respectively

Table 6 Performances of DeepLab V3 + with the backbone, Xception, and OS = 16 (highest value in each column is in bold)

Training dataset	Overall metric (%)			Per-class IoU (%)						
	OA	mAcc	mIoU	Building	Sky	Vegetation	Road	Waterbody	Vehicle	Terrain
ImageNet + Cityscapes	59.18	50.98[a]	34.14[a]	53.43	94.25	40.90	11.98	0.00	0.28	3.97
ImageNet + CIM-WV	**97.60**	**91.48**	**76.78**	**97.49**	**98.67**	**85.23**	**70.02**	**89.95**	**47.53**	**48.56**

[a] The "Waterbody" class excluded due to no training label in Cityscapes

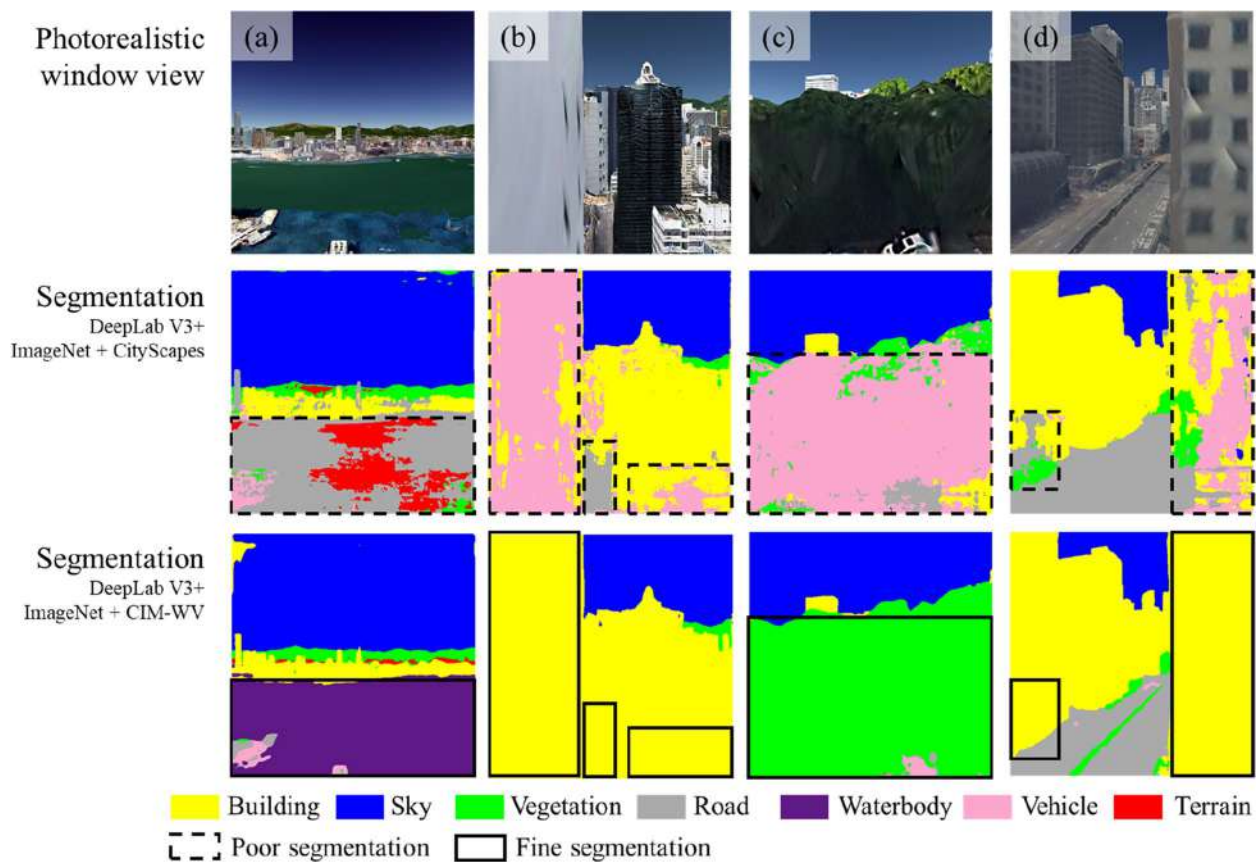Li *et al. Urban Informatics*   (2024) 3:12

Page 13 of 19

Fig. 11 Comparison of semantic segmentation results of DeepLab V3 + trained on Cityscapes and CIM-WV. (**a**) Sea view, (**b**) building view, (**c**) greenery view, and (**d**) street view

as vehicles and roads as shown in Fig. 11b. There existed confusion between flat terrain and sea surface as shown in Fig. 11a. In addition, the model trained on Cityscapes exhibited only half the performance in segmenting vegetation and buildings compared to the model trained on CIM-WV. This disparity can be attributed to the discrepancies between CIM-generated images and real-world view images. For example, close-range CIM-generated vegetation view was incorrectly recognized as vehicles as shown in Fig. 11c. Close-range building facades with low-resolution textures were incorrectly segmented as vehicles, roads, and vegetation as shown in Figs. 11b and d. By contrast, DeepLab V3 + trained on CIM-WV achieved higher performance, with mIoU at 76.78% as shown in Table 6.

4.2.3 Robustness of trained DeepLab V3 + models in the study area

Table 7 compares the performances of the best Deep-Lab V3 + model (Backbone = Xception, OS = 8) trained

Table 7 Robustness of the DeepLab V3 + model (Backbone = Xception, OS = 8) trained on CIM-WV on the test sets of Hong Kong Island (CIM-WV: Hong Kong), Kowloon Peninsula (CIM-WV: Kowloon), and another 60 images in the western Hong Kong Island

Training set	Test set	Overall metric (%)			Per-class IoU (%)						
		OA	mAcc	mIoU	Building	Sky	Vegetation	Road	Waterbody	Vehicle	Terrain
CIM-WV	CIM-WV: Hong Kong	**98.07**	87.94	**76.00**	**98.05**	<u>98.70</u>	86.07	**72.10**	87.00	<u>32.02</u>	**58.08**
	CIM-WV: Kowloon	<u>96.87</u>	**90.90**	75.78	96.43	99.06	<u>83.64</u>	71.16	**88.59**	**60.30**	<u>31.30</u>
	Non CIM-WV: Another 60 images in the western Hong Kong Island	97.16	<u>84.52</u>	<u>72.09</u>	<u>96.30</u>	**99.08**	**93.43**	<u>58.01</u>	84.99	35.42	37.37

The highest and lowest values in each column are in bold and underlined, respectively

on CIM-WV on the test sets of Hong Kong Island and Kowloon Peninsula, and another 60 diversified photorealistic window view images in the western Hong Kong Island. Overall, the model achieved similar performance on the test sets of Hong Kong Island and Kowloon Peninsula with mIoU above 75.78%. The mIoU of the trained model slightly dropped to 72.09% for 60 other photorealistic window view images in unseen areas of Hong Kong. A possible reason is the fluctuated low performance of the model on detecting non-dominated landscape elements, e.g., roads, vehicles, and terrain. For example, there existed confusion between building podiums and roads as shown in Fig. 12b, incorrect detection of unevenly modeled building roofs as vehicles as shown in Fig. 12c, and confusion between vegetation and terrain (see Fig. 12d) at the distant layer of the window view. By contrast, similar and high per-class IoUs were achieved for dominated landscape elements in the window view, i.e., building, sky, vegetation, and waterbody as shown in Table 7. Figure 12a shows a typical well-segmented window view image with the four landscape elements. The similarly high values of OAs also reflected the holistically consistent segmentation performance of the trained

DeepLab V3+ (Backbone=Xception, OS=8) for different areas of Hong Kong.

Table 8 lists the performances of the best DeepLab V3+ model (Backbone=DRN, OS=16) trained on the Hong Kong Island set only on test sets of Hong Kong Island and Kowloon Peninsula. The trained model achieved high performance in segmenting photorealistic window view images in Hong Kong Island with mIoU at 77.20%. By contrast, the performance of the model on the test set of Kowloon Peninsula slumped to 57.11%. There existed a similar performance decrease in Table 9 for the best model (Backbone=DRN, OS=16) trained on the Kowloon Peninsula set on segmenting window view images of the Hong Kong Island set. The mIoU of the trained DeepLab V3+ declined from 73.42% to 43.17%. The significant performance differences also reflected the diverse style representations, e.g., brightness, color contrast, and modeling difference of CIM-WV as mentioned in Sect. 3.2.1.

4.2.4 Transferability of trained DeepLab V3 + models for learning multi-source view images

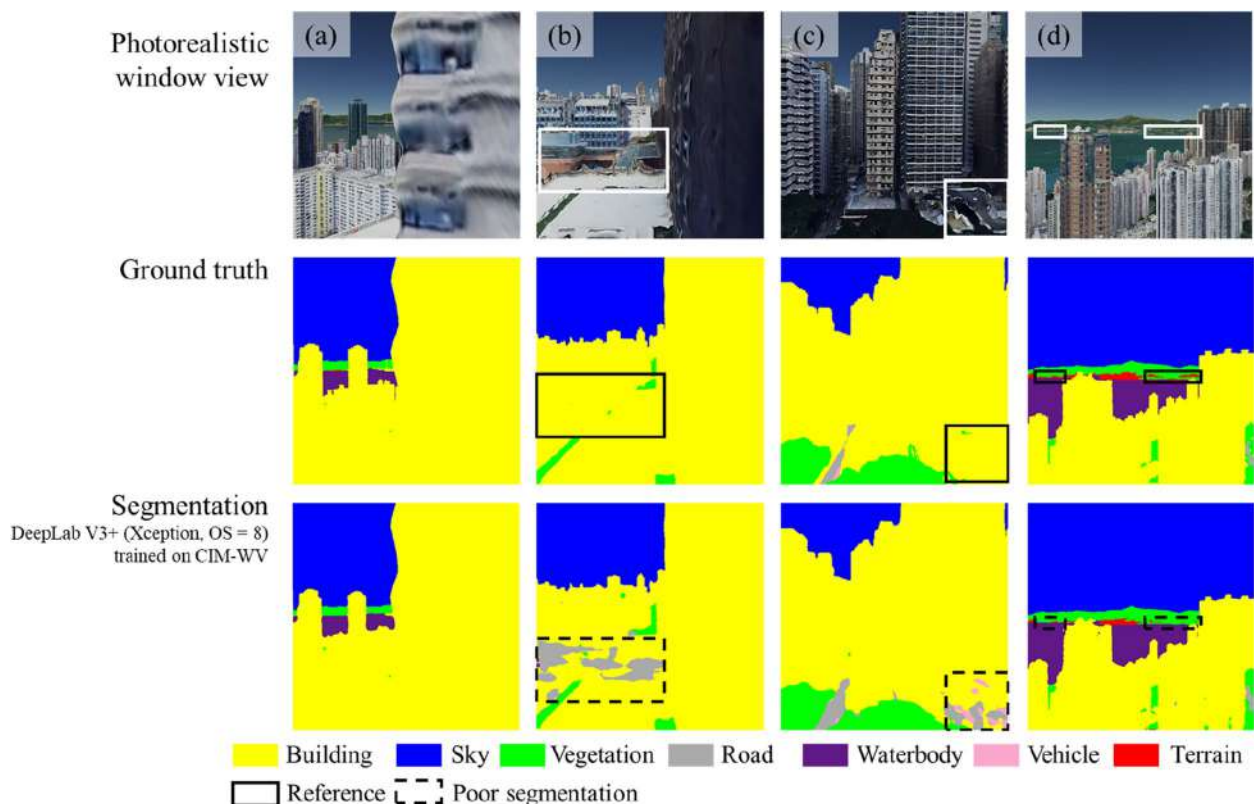Table 10 lists the performance of two DeepLab V3+ models trained on real-world window view images.

Fig. 12 Semantic segmentation results of four example images in the western Hong Kong Island using the DeepLab V3+ (Backbone=Xception, OS=8) trained on CIM-WV. Similar high-performance detection of (**a**) buildings, sky, vegetation, and waterbody, and poor segmentations of (**b**) roads, (**c**) vehicles, and (**d**) terrain

Table 8 Performances of the DeepLab V3 + model (Backbone = DRN, OS = 16) trained on the Hong Kong Island set only on the test set of CIM-WV

Training set	Test set	Overall metric (%)			Per-class IoU (%)						
		OA	mAcc	mIoU	Building	Sky	Vegetation	Road	Waterbody	Vehicle	Terrain
CIM-WV: Hong Kong	CIM-WV: Hong Kong	**97.97**	**88.79**	**77.20**	**98.00**	98.52	**85.04**	**72.22**	**91.41**	**37.88**	**57.35**
	CIM-WV: Kowloon	94.88	63.87	57.11	94.06	**98.80**	61.79	52.86	75.25	12.57	4.41

The highest value in each column is in bold

Table 9 Performances of the DeepLab V3 + model (Backbone = DRN, OS = 16) trained on the Kowloon Peninsula set only on the test set of CIM-WV

Training set	Test set	Overall metric (%)			Per-class IoU (%)						
		OA	mAcc	mIoU	Building	Sky	Vegetation	Road	Waterbody	Vehicle	Terrain
CIM-WV: Kowloon	CIM-WV: Kowloon	**96.77**	**82.66**	**73.42**	**96.21**	**98.95**	**70.32**	**69.65**	**89.79**	**64.26**	**24.76**
	CIM-WV: Hong Kong	88.65	63.21	43.17	87.24	95.58	49.69	21.74	40.97	3.02	3.94

The highest value in each column is in bold

Table 10 Performance improvement in predicting real-world window view images using CIM-WV

Training dataset	Overall metric (%)			Per-class IoU (%)						
	OA	mAcc	mIoU	Building	Sky	Vegetation	Road	Waterbody	Vehicle	Terrain
Real-world window view images	81.86	52.28	33.97	75.58	73.47	66.88	19.96	0.12	1.79	0.00
CIM-WV + real-world window view images	**95.15**	**61.87**	**52.22**	**93.66**	**90.34**	**89.69**	**45.41**	**41.97**	**4.47**	0.00

The highest value in each column is in bold

With currently limited annotated real-world window view images, the finetuned model achieved a higher performance (mIoU = 52.22%) than the model trained from scratch (mIoU = 33.97%). Specifically, buildings and sky were more finely detected as shown in Figs. 13a, b, and d. Table 10 shows a high improvement of per-class IoUs (≥ 22.81%) of vegetation, roads, and waterbody. Example semantic segmentation results are shown in Figs. 13 b, d, and c, respectively. A possible reason is the reutilization of low-level representations from the DeepLab V3 + model trained on CIM-WV to improve the performance comprehensively.

Similar performance improvements were also observed on models trained on CIM-WV and finetuned by feeding Google Earth CIM-generated window view images from New York and Singapore. With limited annotated Google Earth CIM-generated images, Fig. 14 shows a holistically consistent performance improvement for all seven landscape elements by OA, mAcc, mIoU, and per-class IoUs. Specifically, the performances of segmenting small-volume landscape elements, i.e., roads were significantly improved (Increase of per-class IoUs ≥ 28.82%) in both

New York and Singapore sets, as shown in Fig. 14. Roads were mostly detected from buildings in both New York and Singapore as shown in Figs. 15a, b, and d. In addition, the finetuned model outperformed in detecting vehicles from buildings in New York as shown in Fig. 15a, and in detecting waterbody in Singapore, as shown in Fig. 15c.

5 Discussion

5.1 Significance and contribution

High-quality window views, e.g., sea view, sky view, and greenery view are valued by urban dwellers, especially in high-rise, high-density cities. The narrow living space and crowded cityscapes further amplify the benefits of high-quality window views to human physical and mental health. Large-scale quantified window view indicators can bring high socio-economic values, e.g., supporting precise housing valuation and selection and prioritization of improvement of the built environment. The correlations between quantified window view indicators and human perception and physical and mental health may further bring quantified evidence for reshaping the multi-level urban environment, e.g., the
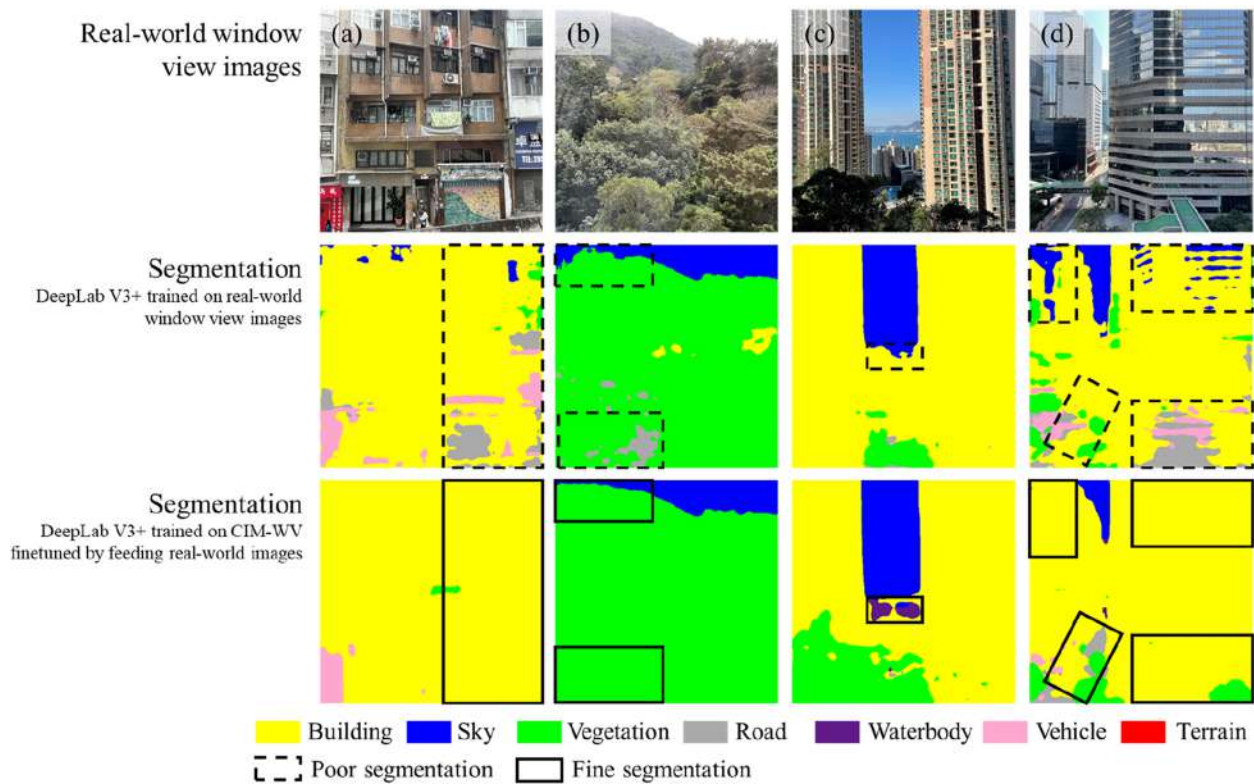
Fig. 13 Comparison of semantic segmentation results of DeepLab V3 + trained on real-world window view images and trained on CIM-WV and finetuned by feeding real-world view images. Improved detection of (**a**) buildings, (**b**) vegetation, (**c**) waterbody, (**d**) sky, and roads

quantified evidence on minimum window-level nature exposure for decreasing the depression and crime rates.

However, traditional manual methods, e.g., onsite assessment, are limited to small-scale experiments. Although CIM-generated window views have opened up opportunities for assessing urban-scale window views, current automatic methods based on deep transfer learning fail to accurately segment CIM-generated window view images. The models trained on other urban view datasets, e.g., street views in the real world cannot accurately assess CIM-generated window view images of varying heights. Thus, a semantic segmentation dataset of rich window view contents is significant for advancing an accurate pixel-level window view assessment.

This paper presents the first publicly accessible window view image dataset with rich semantic annotations. To our best knowledge, CIM-WV is the first dataset of CIM-generated window view images for advancing the urban-scale window view assessment. The CIM-WV supplements the existing semantic segmentation datasets of the multi-angle urban view hub including satellite and street views. Experimental results confirmed a more accurate window view assessment using

deep learning from CIM-WV than deep transfer learning from ground-level views. The DeepLab V3 + model trained on CIM-WV was robust (mIoU ≥ 72.09%) in Hong Kong and enhanced the semantic segmentation accuracy of real-world and Google Earth CIM-generated window view images in multiple cities. Last, for urban researchers and practitioners, our publicly accessible deep learning models trained on CIM-WV enable novel multi-source window view-based urban applications, including precise real estate valuation, improvement of built environment, and window view-related urban analytics.

In addition, the findings in this paper may also inspire researchers to study or regenerate urban views of any viewports from photorealistic CIMs for unlocking potential socio-economic values. Possible examples include quantifying window views of specific groups, such as the elderly and disabled, as well as revitalizing street views along pedestrian walkways and bicycle lanes for cyclists, rather than focusing solely on central lanes for cars. In addition, the proposed CIM-WV can also facilitate the projection-based 3D semantic segmentation of photorealistic CIM (Li et al., 2023a).
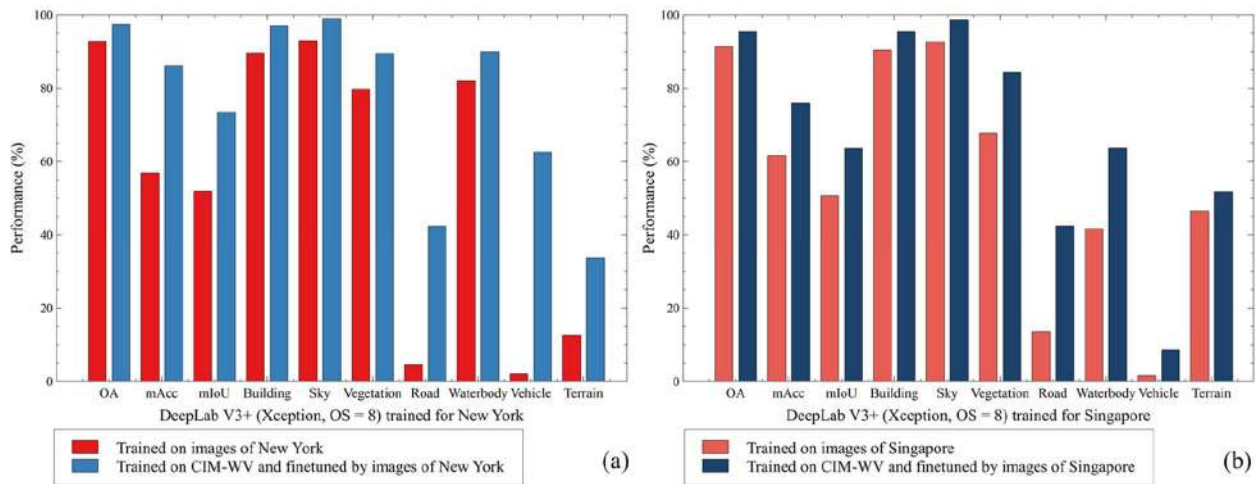
Fig. 14 Performance improvement in segmenting window view images in (**a**) New York and (**b**) Singapore by using the DeepLab V3 + (Backbone = Xception, OS = 8) trained on CIM-WV
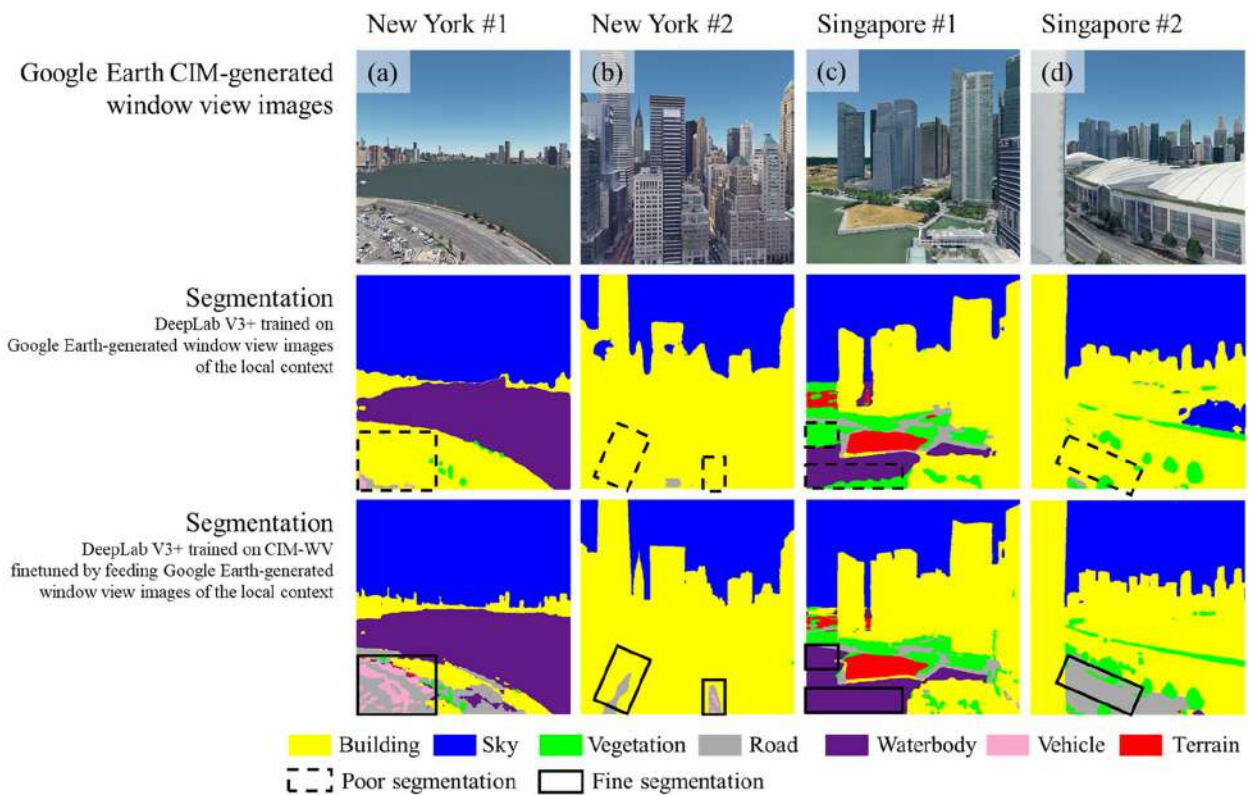

Fig. 15 Comparison of example window view images segmented by DeepLab V3 + models trained on images of the local text and trained on CIM-WV and finetuned by feeding images of the local context, respectively. Four window view images in (**a**)-(**b**) New York and (**c**)-(**d**) Singapore

5.2 Limitations and future work

This study has several limitations. First, the number, collection areas, and styles of window view images of CIM-WV are limited. Our work encourages more interest and annotated datasets for large-scale window view assessment across high-rise, high-density cities. Thereafter, the study lacks a quantitative judgment of the image quality of CIM-WV against real-world view images. Clustering of CIM-generated window views by image quality may further improve the semantic segmentation accuracy. Our future

research directions include a comparison of patterns of window view images across different cities, quality assessment of CIM-generated urban view images for improvement of segmentation accuracy, and interpretability analysis of deep learning models using architecture, engineering, and construction knowledge (Liang & Xue, 2023).

6 Conclusion

Urban-scale assessment of window views plays a significant role in advancing precise housing valuation and selection and prioritization of improvement of the built environment, especially in high-rise, high-density cities. However, existing automatic assessment methods fail to precisely assess window views due to the deep transfer learning from the other urban views, e.g., street views. The absence of publicly accessible annotated photorealistic window view image datasets has hindered accurate pixel-level semantic segmentation.

This paper presents a City Information Model-generated Window View image dataset (CIM-WV) with rich semantic annotations. CIM-WV comprises 2,000 window view images containing 1.62 billion pixels. Window view images of CIM-WV were collected in high-rise, high-density urban areas of Hong Kong with seven semantic labels, i.e., building, sky, vegetation, road, waterbody, vehicle, and terrain. We provided a comprehensive evaluation of CIM-WV, including a baseline of assessment of seven window view elements using DeepLab V3+, a comparative analysis of view segmentation using CIM-WV and Cityscapes, and robustness and transferability analyses of the trained DeepLab V3+models for multi-source window view images in different high-rise, high-density cities. Experimental results confirmed a more accurate window view assessment using deep learning from CIM-WV than deep transfer learning from street views. The robust DeepLab V3+model in Hong Kong enhances the semantic segmentation accuracy of real-world and Google Earth CIM-generated window view images.

The proposed CIM-WV pushes the boundary of semantic segmentation of the multi-angle urban view hub beyond ground-level street views and overhead-level satellite views. To our best knowledge, it is the first annotated window view image dataset generated on photorealistic CIMs. We make the CIM-WV dataset and trained DeepLab models publicly accessible for researchers to advance future vertical urban view applications. Our future work includes examining patterns of window view images across cities, quality assessment of simulated urban view images, and interpretability analysis of deep learning for improvement of segmentation accuracy.

Authors' contributions
ML: Conceptualization, Data Collection and Processing, Methodology, Software, Validation, Charting, Writing – Original Draft Preparation, Funding Acquisition; AGOY: Supervision, Writing – Review and Editing, Funding Acquisition; FX: Conceptualization, Writing – Review and Editing, Funding Acquisition.

Availability of data and materials
Published in online data repository: https://doi.org/10.25442/hku.24647487.

Declarations

Consent for publication
All authors have read and agreed to the submitted version of the manuscript.

Competing interests
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References
Alibaba. (2023). Taobao. Hangzhou: Alibaba Group. Retrieved from https://ai.taobao.com/

Azimi, S. M., Henry, C., Sommer, L., Schumann, A. & Vig, E. (2019). Skyscapes fine-grained semantic understanding of aerial scenes. Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 7393–7403). IEEE. https://doi.org/10.1109/ICCV.2019.00749

Baranzini, A., & Schaerer, C. (2011). A sight for sore eyes: Assessing the value of view and land use in the housing market. *Journal of Housing Economics,20*(3), 191–199. https://doi.org/10.1016/j.jhe.2011.06.001

Biljecki, F., & Ito, K. (2021). Street view imagery in urban analytics and GIS: A review. *Landscape and Urban Planning,215*, 104217. https://doi.org/10.1016/j.landurbplan.2021.104217

Cesium GS. (2022). The Cesium Platform. Philadelphia, USA: Cesium GS, Inc. Retrieved from https://cesium.com/platform/

Chen, X., Ma, H., Wan, J., Li, B. & Xia, T. (2017). Multi-view 3d object detection network for autonomous driving. IEEE Conference on Computer Vision and Pattern Recognition (pp. 1907–1915). IEEE. https://doi.org/10.1109/CVPR.2017.691

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) (pp. 801–818). Springer. https://doi.org/10.1007/978-3-030-01234-2_49

Chen, B., Tu, Y., Wu, S., Song, Y., Jin, Y., Webster, C., Xu, B., & Gong, P. (2022). Beyond green environments: Multi-scale difference in human exposure to greenspace in China. *Environment International,166*, 107348. https://doi.org/10.1016/j.envint.2022.107348

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. Proceedings of the IEEE

Li *et al. Urban Informatics*   (2024) 3:12

Page 19 of 19

Conference on Computer Vision and Pattern Recognition (pp. 3213–3223). IEEE. https://doi.org/10.1109/CVPR.2016.350

Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D. & Raskar, R. (2018). Deepglobe 2018: A challenge to parse the earth through satellite images. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 172–181). IEEE. https://doi.org/10.1109/CVPRW.2018.00031

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K. & Li, F.-F. (2009). Imagenet: A large-scale hierarchical image database. IEEE Conference on Computer Vision and Pattern Recognition (pp. 248–255). IEEE. https://doi.org/10.1109/CVPR.2009.5206848

Fisher-Gewirtzman, D. (2018). Integrating 'weighted views' to quantitative 3D visibility analysis as a predictive tool for perception of space. *Environment and Planning b: Urban Analytics and City Science,45*(2), 345–366. https://doi.org/10.1177/0265813516676486

He, D., Miao, J., Lu, Y. S., & Liu, Y. (2022). Urban greenery mitigates the negative effect of urban density on older adults' life satisfaction: Evidence from Shanghai. *China. Cities,124*, 103607. https://doi.org/10.1016/j.cities.2022.103607

Helbich, M., Yao, Y., Liu, Y., Zhang, J., Liu, P., & Wang, R. (2019). Using deep learning to examine street view green and blue spaces and their associations with geriatric depression in Beijing, China. *Environment International,126*, 107–111. https://doi.org/10.1016/j.envint.2019.02.013

HKCEDD. (2019). Role of Reclamation in Hong Kong Development. Hong Kong: Civil Engineering and Development Department, Government of Hong Kong SAR. Retrieved from https://www.cedd.gov.hk/filemanager/eng/content_954/Info_Sheet3.pdf

HKLandsD. (2014). *iB1000 Digital Topographic Map*. Lands Department, Government of Hong Kong SAR.

HKPlanD. (2018). Hong Kong Planning Standards and Guidelines. Hong Kong: Planning department, Hong Kong SAR. https://www.pland.gov.hk/pland_en/tech_doc/hkpsg/full/pdf/ch2.pdf

HKPlanD. (2019b). 3D Photo-realistic Model Data Specification. Hong Kong: Planning Department, Government of Hong Kong SAR. Retrieved from https://www.pland.gov.hk/pland_en/info_serv/3D_models/3D_Photo_realistic_Model_Specification.pdf

HKPlanD. (2019a). 3D Photo-realistic Model. Hong Kong: Planning Department, Government of Hong Kong SAR. Retrieved from https://www.pland.gov.hk/pland_en/info_serv/3D_models/download.htm

HKTPB. (2010). Guidelines on submission of visual impact assessment for planning applications to the Town Planning Board. Hong Kong: Town Planning Board. https://www.info.gov.hk/tpb/en/forms/Guidelines/TPB_PG_41.pdf

Jim, C. Y., & Chen, W. Y. (2009). Value of scenic views: Hedonic assessment of private housing in Hong Kong. *Landscape and Urban Planning,91*(4), 226–234. https://doi.org/10.1016/j.landurbplan.2009.01.009

Kuo, F. E., & Sullivan, W. C. (2001). Environment and crime in the inner city: Does vegetation reduce crime? *Environment and Behavior,33*(3), 343–367. https://doi.org/10.1177/0013916501333002

Laovisutthichai, V., Li, M., Xue, F., Lu, W., Tam, K. & Yeh, A. G. (2021). CIM-enabled quantitative view assessment in architectural design and space planning. 38th International Symposium on Automation and Robotics in Construction (ISARC 2021). Dubai. https://doi.org/10.22260/ISARC2021/0011

Li, M., Xue, F., Yeh, A. G. & Lu, W. (2021). Classification of photo-realistic 3D window views in a high-density city: The case of Hong Kong. Proceedings of the 25th International Symposium on Advancement of Construction Management and Real Estate (pp. 1339–1350). Wuhan: Springer, Singapore. doi:https://doi.org/10.1007/978-981-16-3587-8_91

Li, M., Xue, F. & Yeh, A. G. (2023c). Efficient Assessment of Window Views in High-Rise, High-Density Urban Areas Using 3D Color City Information Models. Proceedings of the 18th International Conference on Computational Urban Planning and Urban Management (pp. 1-11). Montreal: OSF.

Li, M., Wu, Y., Yeh, A. G. & Xue, F. (2023a). HRHD-HK: A benchmark dataset of high-rise and high-density urban scenes for 3D semantic segmentation of photogrammetric point clouds. 2023 IEEE International Conference on Image Processing (pp. 1–5). IEEE, in press. https://doi.org/10.48550/arXiv.2307.07976

Li, M., Xue, F., Wu, Y., & Yeh, A. G. (2022). A room with a view: Automated assessment of window views for high-rise high-density areas using City Information Models and transfer deep learning. *Landscape and Urban Planning,226*, 104505. https://doi.org/10.1016/j.landurbplan.2022.104505

Li, M., Xue, F., & Yeh, A. G. (2023b). Bi-objective analytics of 3D visual-physical nature exposures in high-rise high-density cities for landscape and urban planning. *Landscape and Urban Planning,233*, 104714. https://doi.org/10.1016/j.landurbplan.2023.104714

Li, W., & Samuelson, H. (2020). A new method for visualizing and evaluating views in architectural design. *Developments in the Built Environment,1*, 100005. https://doi.org/10.1016/j.dibe.2020.100005

Liang, D., & Xue, F. (2023). Integrating automated machine learning and interpretability analysis in architecture, engineering and construction industry: A case of identifying failure modes of reinforced concrete shear walls. *Computers in Industry,147*, 103883. https://doi.org/10.1016/j.compind.2023.103883

Liao, C., Hu, H., Yuan, X., Li, H., Liu, C., Liu, C., Fu, G., Ding, Y., & Zhu, Q. (2023). BCE-Net: Reliable building footprints change extraction based on historical map and up-to-date images using contrastive learning. *ISPRS Journal of Photogrammetry and Remote Sensing,201*, 138–152. https://doi.org/10.1016/j.isprsjprs.2023.05.011

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) (pp. 3234–3243). Las Vegas: IEEE. https://doi.org/10.1109/CVPR.2016.352

Shi, W., Batty, M., Goodchild, M., & Li, Q. (2022a). The digital transformation of cities. *Urban Informatics,1*(1), 1. https://doi.org/10.1007/s44212-022-00005-1

Shi, W., Goodchild, M., Batty, M., Li, Q., Liu, X., & Zhang, A. (2022b). *Prospective for Urban Informatics. Urban Informatics,1*(1), 2. https://doi.org/10.1007/s44212-022-00006-0

Stamps, A. E., III. (2005). Enclosure and safety in urbanscapes. *Environment and Behavior,37*(1), 102–133. https://doi.org/10.1177/0013916504266806

Ulrich, R. S. (1984). View through a window may influence recovery from surgery. *Science,224*(4647), 420–421. https://doi.org/10.1126/science.6143402

Wang, J., Zheng, Z., Ma, A., Lu, X. & Zhong, Y. (2021). LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation. Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (pp. 1-12). Virtual: Curran Associates, Inc. https://doi.org/10.5281/zenodo.5706578

Wang, J., Ma, A., Zhong, Y., Zheng, Z., & Zhang, L. (2022). Cross-sensor domain adaptation for high spatial resolution urban land-cover mapping: From airborne to spaceborne imagery. *Remote Sensing of Environment,277*, 113058. https://doi.org/10.1016/j.rse.2022.113058

Xue, F., Li, X., Lu, W., Webster, C. J., Chen, Z., & Lin, L. (2021). Big data-driven pedestrian analytics: Unsupervised clustering and relational query based on Tencent Street View photographs. *ISPRS International Journal of Geo-Information,10*(8), 561. https://doi.org/10.3390/ijgi10080561

Yang, L., Ao, Y., Ke, J., Lu, Y., & Liang, Y. (2021). To walk or not to walk? Examining non-linear effects of streetscape greenery on walking propensity of older adults. *Journal of Transport Geography,94*, 103099. https://doi.org/10.1016/j.jtrangeo.2021.103099

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V. & Darrell, T. (2020). Bdd100k: A diverse driving dataset for heterogeneous multitask learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2636–2645). Seattle: IEEE. https://doi.org/10.1109/CVPR42600.2020.00271

Zhou, L., Zhang, C. & Wu, M. (2018). D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 182–186). Salt Lake City: IEEE. https://doi.org/10.1109/CVPRW.2018.00034

Zhou, Q., & Xue, F. (2023). Pushing the boundaries of Modular-integrated Construction: A symmetric skeleton grammar-based multi-objective optimization of passive design for energy savings and daylight autonomy. *Energy and Buildings,296*, 113417. https://doi.org/10.1016/j.enbuild.2023.113417

Publisher's Note