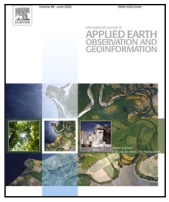
International Journal of Applied Earth Observation and Geoinformation 119 (2023) 103323

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag



Integrating satellite and street-level images for local climate zone mapping

Rui Cao [a,b,c,*], Cai Liao [a,c,d], Qing Li [e], Wei Tu [f], Rui Zhu [g], Nianxue Luo [d], Guoping Qiu [h], Wenzhong Shi [a,b,*]

[a] Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong Special Administrative Region of China
[b] Otto Poon Charitable Foundation Smart Cities Research Institute, The Hong Kong Polytechnic University, Hong Kong Special Administrative Region of China
[c] The Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen, China
[d] School of Geodesy and Geomatics, Wuhan University, Wuhan, China
[e] Pengcheng Laboratory, Shenzhen, China
[f] Guangdong Key Laboratory of Urban Informatics & School of Architecture and Urban Planning, Shenzhen University, Shenzhen, China
[g] Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore
[h] School of Computer Science, University of Nottingham, Nottingham, UK

ARTICLE INFO

Keywords:
Local climate zone (LCZ)
Climate change
Remote sensing
Street view images
Data fusion
GeoAI

ABSTRACT

Timely and accurate local climate zone (LCZ) classification maps are valuable for urban climate studies. The integration of remote sensing and street-level images is promising to produce high-quality LCZ maps, since the former can efficiently capture the information of landscapes on a large-scale while the latter include ground-level details. However, due to their significant differences in spatial distributions and capture views, as well as existing sampling issues of street-level images, how to fuse them effectively is challenging and remains an uncharted research area. To address these issues and fill the gap, this study proposes an effective method to integrate satellite and street-level images for LCZ mapping. Additionally, a simple yet effective street-level image sampling method is proposed. Extensive experiments have been performed and the results demonstrate the effectiveness of the proposed data fusion method and also confirm the usefulness of fusing street-level images with satellite images in enhancing the performance of LCZ mapping. Moreover, the proposed sampling method can increase data representativeness and avoid data redundancy, thus significantly reducing the number of required images while retaining high classification accuracy. To the best of our knowledge, this study is the first attempt to integrate cross-view satellite and street-level images for LCZ mapping. The study and proposed methods can contribute to the development of multi-source data fusion for LCZ map production and further benefit urban climatic research.

1. Introduction

Urbanization and climate change are the most important topics in the 21st century, which are critical for sustainable development and are parts of the United Nations' sustainable development goals (SDGs), i.e., SDG 11 (sustainable cities and communities) and SDG 13 (climate action) (United Nations, 2015). With rapid urbanization, more than half of the world's population now lives in cities and the proportion is expected to increase to 68% by 2050 (United Nations, 2018). Urban areas therefore become particularly exposed and vulnerable to the potential risks and disasters caused by climate change (Wamsler et al., 2013).

The local climate zones (LCZ) is a classification scheme initially proposed for urban heat island research, which characterizes the landscape of urban morphology and function as well as land cover based mainly on properties of surface structure and surface cover (Stewart and Oke, 2012). There are 17 LCZ types, including 10 built types (type 1–10) and 7 land cover types (type A–G), as illustrated in Fig. 1. As demonstrated by many climatic studies, the LCZ classification scheme is effective for climatic modeling and can serve as a universal standard for communication among the climatic research community (Xue et al., 2020). Rapid urbanization significantly changes the urban landscapes in a short period of time, and it is significantly important to keep the data up to date for more accurate climatic modeling in tackling degenerating climate change. Thus, timely and accurate LCZ mapping becomes a critical prerequisite towards high-quality weather and climatic modeling, which is essential for scientific climate change research and climatic-responsive design.

There are mainly three kinds of methods for LCZ mapping, i.e., in-situ measurement (Thomas et al., 2014), GIS-based methods (Wang

* Corresponding authors.
E-mail addresses: rucao@polyu.edu.hk (R. Cao), lswzshi@polyu.edu.hk (W. Shi).

https://doi.org/10.1016/j.jag.2023.103323
Received 15 January 2023; Received in revised form 16 April 2023; Accepted 18 April 2023
Available online 3 May 2023

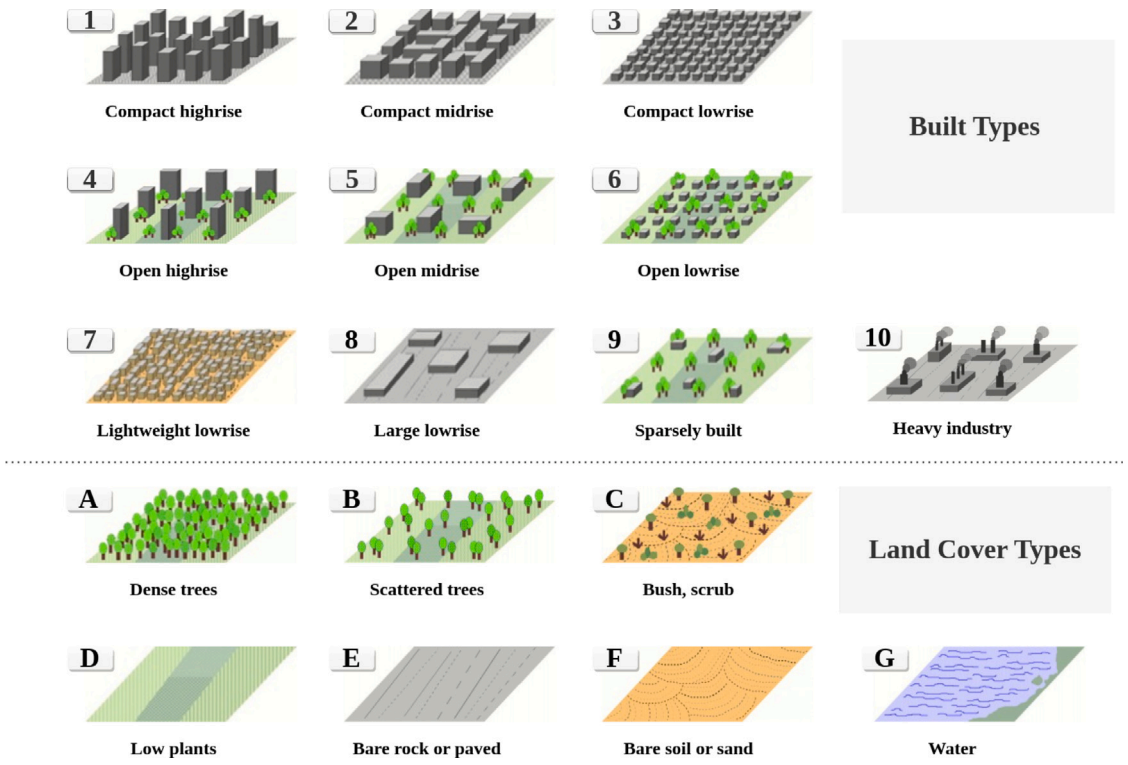


Fig. 1. Local climate zone classification scheme, including 10 built types (1–10) and 7 land cover types (A–G) (Stewart and Oke, 2012).

et al., 2018a; Zheng et al., 2018), and remote sensing-based methods (Bechtel et al., 2015; Yoo et al., 2019; Liu and Shi, 2020; Zhu et al., 2020). The in-situ measurement-based and GIS-based methods can directly calculate the key indicators of LCZ classification standards, such as sky view factor, building height/width aspect ratio, building surface fraction, etc., and then categorize those zones into different LCZ types with rules associated with LCZ type definitions. However, in-situ measurement methods are usually labor-intensive and time-consuming and thus unscalable; while GIS-based methods are data-intensive and require complete and accurate urban GIS data, which however are usually not complete or available to the public, especially for developing and undeveloped countries and regions. Due to the wide availability of remote sensing images, e.g., Landsat and Sentinel satellite images, remote sensing-based methods have attracted increasing attention (Bechtel et al., 2015; Yoo et al., 2019; Liu and Shi, 2020; Zhu et al., 2020).

However, remote sensing images lack ground-level details such as three-dimensional building structures which are critical for effective LCZ type recognition (Ren et al., 2019). Street-level images have shown to be useful in providing such information, with growing accessibility and spatial coverage, they have been widely exploited for various applications (Kang et al., 2020; Biljecki and Ito, 2021). Pioneer research has also shown that street view images (SVI) alone are useful for LCZ classification in image-level by providing more ground-level details of urban environment (Xu et al., 2019; Ignatius et al., 2022). It is thus promising to exploit street-level images to enhance LCZ classification.

Both satellite images and street-level images have their advantages and limitations. It is thus promising to integrate them to complement each other to help enhance the LCZ mapping results. However, there are still challenges in fusing them for LCZ mapping. Firstly, satellite images and street-level images have very different spatial distributions; street-level images have limited coverage over space, which are sparsely distributed along roads, while the LCZ mapping needs continuous coverage of the space (satellite images can well suit the need). Secondly, the two kinds of image data are collected from very different perspectives, with satellite images captured from nadir view and street-level images captured from horizontal view, it is thus non-trivial to integrate them effectively. Thirdly, due to the data source provider and special distribution pattern, the availability of street-level images is limited, how to effectively sample street-level images to increase data representativeness and avoid data redundancy as well as reduce collection cost is rarely explored but crucial for practical usage. Due to the aforementioned challenges, fusing satellite and street-level images for LCZ mapping still remains an uncharted area of research.

To address these challenges, in this paper, we propose an effective cross-view image fusion method for LCZ mapping and an effective street view images sampling method to avoid data redundancy. To evaluate the proposed methods, extensive experiments have been performed in Hong Kong, a representative high-density city with complex and diverse landscapes where street-level images can provide additional details that satellite images lack. The major contributions of this article are summarized as follows:

- To the best of our knowledge, this study is the first attempt to integrate cross-view satellite and street-level images for LCZ mapping. We propose an effective method and framework to integrate the cross-view images with different capture views and spatial distributions, for LCZ mapping, which outperforms the results of using either data source alone.
- To address the large collection cost and data redundancy of street-level images and meanwhile retain classification performance, we propose an effective method to sample street view images along road networks under the hexagonal constraint, which ensures sufficient spatial coverage and representativeness as well as classification performance, significantly reducing the required number of street-level images as input and simultaneously achieving high classification accuracy.

- Extensive experiments have been conducted to evaluate the effectiveness of the proposed cross-view image fusion method for LCZ mapping, which demonstrates the usefulness of using street-level images to augment LCZ mapping performance. Additionally, experiments have also been performed to validate the efficacy of the proposed sampling method.

The rest of the paper is organized as follows. We review the related works of LCZ mapping and remote sensing and street view image fusion in Section 2. Section 3 introduces the study area and data. Section 4 elaborates the proposed method and framework of cross-view image fusion for LCZ mapping, and the effective street-level image sampling method. Furthermore, extensive experiments are conducted to examine the effectiveness of the proposed methods in Section 5. In Section 6, we discuss some important issues. Finally, Section 7 concludes the paper.

2. Related work

2.1. Local climate zone mapping

Currently, the LCZ mapping methods can be mainly categorized into three types according to the data used, i.e., in-situ measurement-based methods, GIS-based methods, and remote sensing-based methods. In-situ measurement-based methods rely on professional instruments to collect LCZ-related parameters (such as sky view factor, building height-to-width aspect ratio, building surface fraction, impervious surface fraction, etc.) in the field, which are then used to classify LCZ types (Thomas et al., 2014). These methods are usually labor-intensive and time-consuming, and thus are not easily scalable for large-area LCZ mapping. GIS-based methods can directly calculate the key indicators of LCZ classification scheme based on GIS data of urban morphology and building information, and then categorize those zones into different LCZ types with rules associated with LCZ type definitions (Wang et al., 2018a; Zheng et al., 2018). These methods can usually achieve high classification accuracy; however, they are data-intensive and require complete and accurate urban GIS data that are usually not complete or available to the general public, especially for developing and undeveloped countries and regions.

Benefiting from the advances in geospatial technologies, growing accessibility to remote sensing data equips us with abundant data in sensing our urban environment, which provides us with high-quality data sources for automated LCZ mapping, such as Landsat and Sentinel satellite images. Therefore, remote sensing-based methods have attracted increasing attention and are widely used (Bechtel et al., 2015; Yoo et al., 2019; Liu and Shi, 2020; Zhu et al., 2020). The World Urban Database and Access Portal Tools (WUDAPT) initiative (Bechtel et al., 2015) proposes a standard workflow for LCZ mapping using open-access Landsat imagery, which first resamples Landsat imagery to 100 m spatial resolution, and then train Random Forests (RF) model from collected imagery and reference data for LCZ mapping. The method is widely used, however, it ignores the context of surrounding environment. To address this issue, scene-based image classification is used (Liu and Shi, 2020; Zhu et al., 2020), which divides the study area into grids and then classifies the image patches enclosed by grids instead of image pixels. This method can include more information from surrounding environment which is critical for LCZ-related information acquisition.

Although the expansion of study unit can include more information, due to the inherent limitation of capturing view, remote sensing images fall short in providing ground-level details such as 3D building information which is important for accurate LCZ type recognition (Ren et al., 2019). Street-level images, however, have shown to be able to offer such information (Biljecki and Ito, 2021; Yan and Huang, 2022). Previous work has shown that street view images are useful for LCZ classification by providing more ground-level details of urban environment (Xu et al., 2019). However, this research is conducted in image-level and the classification results are presented as sparse points in space, which is not sufficient for practical use. To take full advantage of both remote sensing and street view images for practical use, we propose an effective machine learning-based method to fuse them for grid-based LCZ mapping, which can take into account of the ground-level details by fusing street view images and produce a more accurate and practical LCZ map for downstream applications such as climate modeling.

2.2. Integration of remote sensing and street view images

Remote sensing data has the advantage of large-scale coverage and are increasingly accessible. However, due to inherent limitations, they usually can only capture the physical attributes from the top view (Cao et al., 2018, 2020; Chen et al., 2022b). On the other hand, street view images are captured in the horizontal view and can capture more ground-level details such as 3D urban structure, building facade, and tree volume. With their growing accessibility, street view images are widely used for various applications (Biljecki and Ito, 2021; Zhou et al., 2021). However, street-level images also have their limitations, with limited spatial and temporal coverage and resolution.

Due to the complementary characteristics of remote sensing and street view images, it is promising to fuse them and related research has attracted increasing attentions. According to the study object, related works can be categorized into two types of tasks. The first type focuses on specific spatial object, the most studied are buildings. Remote sensing images can capture the information of building roofs from top view, while the street view images can capture the information of building facades from horizontal view. Integrating remote sensing and street view images can provide multiple viewpoints of buildings and thus include more complete information for many applications, such as building type classification (Hoffmann et al., 2019), building vulnerability assessment (Xing et al., 2023), and building damage evaluation (Khajwal et al., 2023). In this type of task, deep learning models are usually used and the number of street view images corresponded to a remote sensing image is fixed, which is suitable for end-to-end classification approaches.

Another type of task focuses on the land, in the research, satellite or aerial images are integrated with street view images to estimate the characteristics of the land, including land use classification (Cao et al., 2018; Cao and Qiu, 2018), urban village detection (Chen et al., 2022a), urban forestry assessment (Barbierato et al., 2020), commercial activeness evaluation (Wang et al., 2018b), etc. In these studies, both traditional machine learning models and deep learning methods are used. For example, Cao et al. (2018) and Chen et al. (2022a) use end-to-end deep neural networks to fuse aerial and street view images which can achieve high accuracy, however, they do not consider the problem of variable numbers of street view points within one land parcel. Besides, deep neural networks are computing-intensive, Chen et al. (2022a) only randomly samples one street view point in each parcel to relieve computational burden. While other studies that use traditional machine learning methods show the good balance between achieved accuracy and computational cost in fusing remote sensing and street-level images and can be easily extended to large areas (Wang et al., 2018b; Barbierato et al., 2020). We can see that previous studies usually ignore the issues of variable numbers of street view images across space as well as the sampling strategy, and deep learning-based methods require a high computational costs. These issues limit the practical use. To address these issues, in this study, we leverage computational efficient machine learning-based methods and efficient SVI sampling strategies to provide an efficient and effective solution to LCZ mapping for practical use.

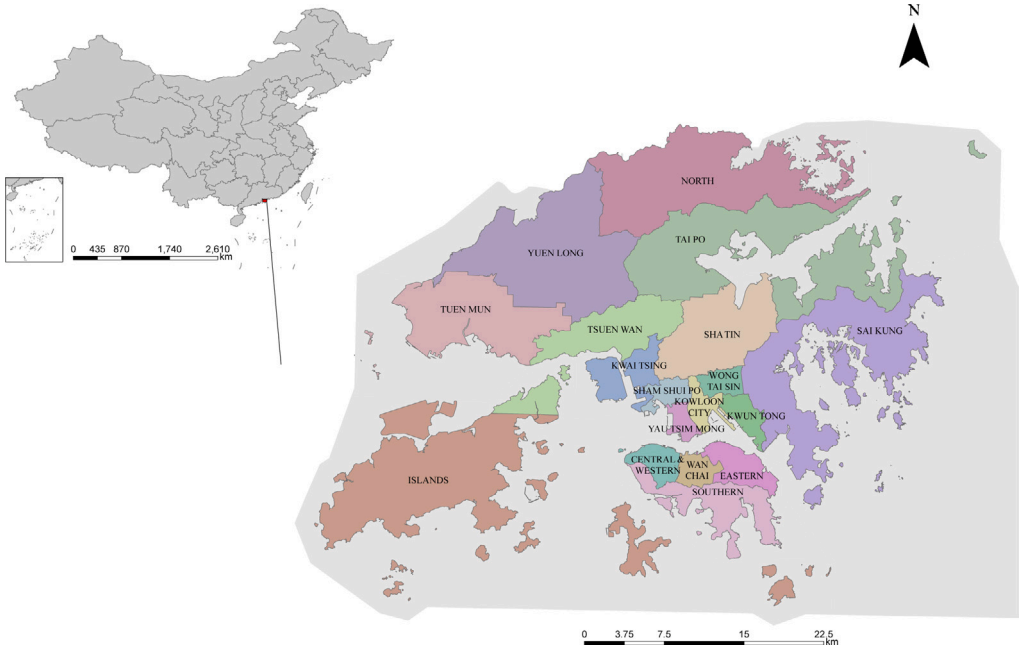


Fig. 2. Study area of Hong Kong, with 18 districts surrounded by large marine water area (indicated by the light gray area).

3. Study area and data

Hong Kong is chosen as the study area, as shown in Fig. 2. Located on the southeast coast of China, Hong Kong has 18 districts, with a total land area of about 1104 km^2 and over 7.5 million population, which is one of the most densely populated cities in the world. Hong Kong is hilly and mountainous, within which about three-quarters of the land is a mountainous area and is reserved as country parks. Only less than 25% of the land is allowed for urban development. Due to the limited land, over 90% of Hong Kong residents live in highrise buildings. The limited developed land and high population density shape the urban morphology of Hong Kong's downtown areas. The complex landscape and high-density built-up areas make it challenging for effective LCZ mapping.

3.1. Remote sensing imagery (RSI)

Cloud-free Sentinel-2 L2A satellite imagery captured in 2020 of the study area is preprocessed and downloaded from Google Earth Engine (Gorelick et al., 2017). Specifically, the imagery with cloud coverage less than 3% is selected and the clouds are masked out. Firstly, we filtered the Google Earth Engine Sentinel-2 L2A datasets based on the date range of the year 2020 and cloud pixel percentage of less than 3% to obtain cloud-free images of 2020. Next, we reduced the obtained image collection to a single multispectral imagery by calculating the median value of each pixel in each band. Finally, the imagery was clipped according to Hong Kong's administrative boundaries. Following Zhu et al. (2020), 10 bands (B2, B3, B4, B5, B6, B7, B8, B8A, B11, B12) are used, with R-G-B-NIR bands of 10 m spatial resolution and Red Edge 1,2,3,4 bands and SWIR 1,2 bands of 20 m resolution.

3.2. Street view images (SVI)

The street view images are downloaded from Google Street View API. Using the proposed sampling methods introduced in Section 4, 69,957 sampling points along road networks were obtained. After the query process, only 32,622 sampling points are available via the Street View Image request API,[1] with the following key parameters:

- *size*: size specifies the output size of the image in pixels. We set it to '640×640' pixels, which is the maximum resolution provided by the API.
- *heading*: heading indicates the compass heading of the camera. Since we want to utilize the comprehensive 360° view of the panorama, we request 4 images for each location with headings of 0, 90, 180, and 270 degrees, respectively.
- *fov*: fov determines the horizontal field-of-view of the image, we use the default 90 degrees.

Examples of Google SVIs of different LCZ types are shown in Fig. 3.

3.3. Support GIS datasets

To define the boundary of the study area, the Hong Kong administrative boundary data is downloaded from the government's open data portal. In addition, since the street view images are usually captured by cars while driving, to generate the sampling points for acquiring SVIs, we downloaded the road networks of Hong Kong from the OpenStreetMap (OSM) via the Python package OSMnx (Boeing, 2017).

3.4. Reference dataset

WUDAPT (Bechtel et al., 2015) is a global initiative of tools to create LCZ maps for cities using a standard methodology and workflow. Based on this platform, crowd-sourcing data have been collected and formed a rich LCZ database. Our reference data of Hong Kong are organized from the contributing datasets from WUDAPT with high accuracy, which are shown in Fig. 4, containing polygons with different LCZ labels. We can see that the labels are not uniformly sampled, including polygons of different types with noticeable difference in area. For example, class

[1] https://developers.google.com/maps/documentation/streetview/request-streetview

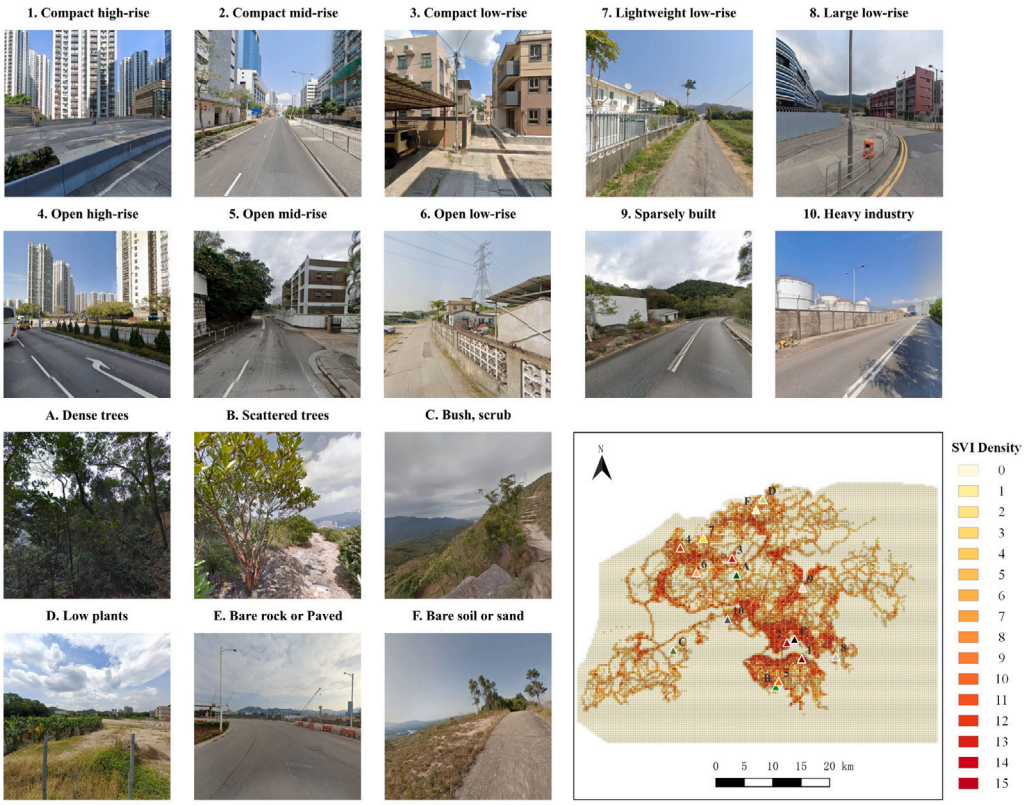


Fig. 3. Spatial distribution of Google Street View images in Hong Kong, with examples of different LCZ types. The map on the bottom-right shows the density of SVIs as well as the locations of corresponding SVIs presented.

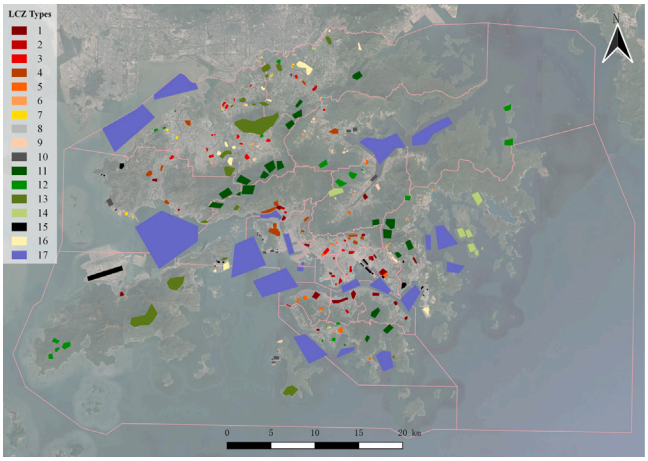


Fig. 4. Reference data of different LCZ types in Hong Kong.

G (*water*) type accounts for large areas since they are more easily recognized. This results in imbalanced labeled samples in the following experiments.

4. Methodology

To integrate RSI and SVI for LCZ mapping, we propose a four-step method, as illustrated in Fig. 5. First, we divide the study area into uniform grids, which can cover the whole study area and serve as basic mapping units for LCZ mapping. Second, we segment the RSIs into image patches to fit in with the mapping units and extract features from them. Third, we sample SVIs along road networks, extract features from SVIs, and further map SVIs to land. Finally, based on the mapping units, we combine the extracted features from RSI and SVI together for LCZ classification and mapping. The details of each step are elaborated in the following subsections.

4.1. Mapping units generation

The local climate zones are regarded as uniform units with similar climatic characteristics and are normally larger than 100m^2. Thus, we divide the study area into 320×320 m^2 spatial grids across the Hong Kong administrative boundary. The size of 320×320 m^2 is widely adopted in literature (Zheng et al., 2018; Liu and Shi, 2020; Zhu et al., 2020) due to the appropriate size for LCZ definition and also suit satellite image resolution in extracting sufficient information.

4.2. RSI feature extraction

For each mapping unit, we can generate a 10-band multispectral image patch with the size of 32×32 pixels based on the grid-level mapping units generated. Spectral features are extracted by computing the mean, variance, maximum, minimum, skewness, and kurtosis of the digital numbers of all the pixels of each band for the image patches, which results in 60-dimensional feature vector for each grid. In addition, we use the deep neural network model ResNet-50 trained on the So2Sat LCZ42 dataset (Zhu et al., 2020) to extract the semantic features from the image patches, which results in 2048-dimensional feature vector for each mapping unit. The dataset consists of about half a million co-registered Sentinel-1 and Sentinel-2 remote sensing image patches, as well as the corresponding LCZ labels annotated by domain experts following a rigorous labeling workflow and evaluation process. We then concatenate the spectral and semantic features to represent RSI image patches.

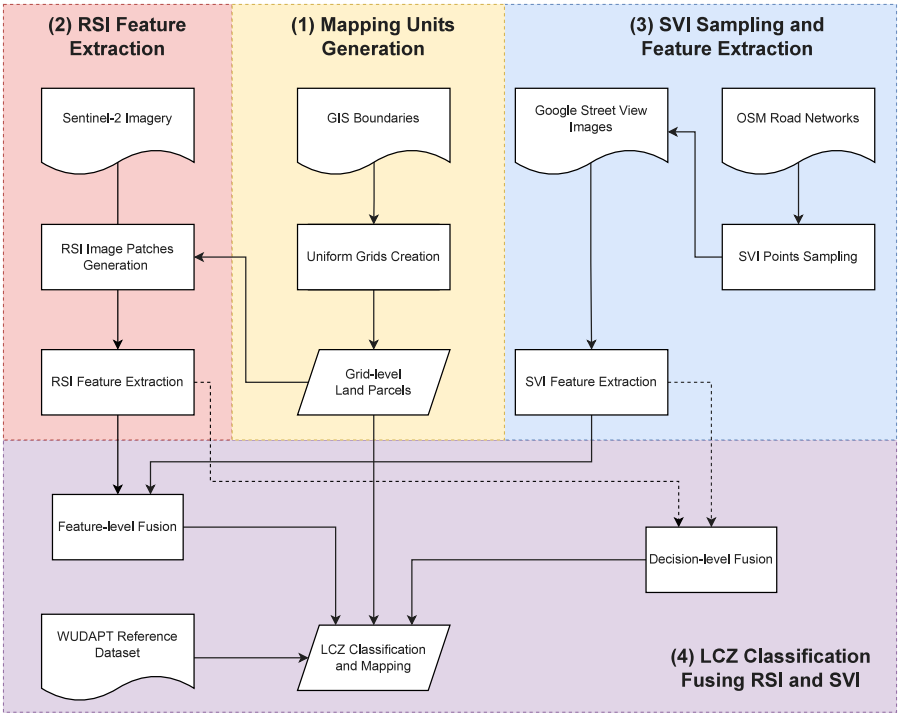


Fig. 5. Overview of the proposed method to fuse remote sensing imagery (RSI) and street view images (SVI) for LCZ mapping, including four steps: (1) mapping units generation, (2) RSI feature extraction, (3) SVI sampling and feature extraction, (4) LCZ classification and mapping fusing RSI and SVI.

4.3. SVI sampling and feature extraction

4.3.1. SVI sampling

The street view images are downloaded from the Google Street View API, which has a limitation on data download in terms of both time and economical cost. To take full advantage of the SVIs without incurring unnecessary costs, we propose an effective way to sample street view images along road networks under the constraints of hexagon units, which can ensure efficiency as well as coverage. The workflow of the sampling method is shown in Fig. 6, which includes three major steps. Firstly, we generate dense sampling points along the OSM road networks with an interval of 10 m (road ends are also included). Secondly, we generate the tessellated grid of hexagons (with a radius of 50 m) across the study area to cover all the mapping units. Finally, based on the generated sampling points and hexagons, we compare the two kinds of objects and ensure that each hexagon only contains one sampling point to reduce the required numbers of SVIs. If there are multiple points within a hexagon, only one will be left which is the closest to the center of the hexagon. It should also be noted that, as can be seen in Fig. 6, each mapping unit (uniform grid) covers multiple hexagons, which means that each mapping unit can contain multiple sampling points.

The proposed method is based on the assumption that the SVIs at a location can capture the horizontal landscape of nearby 50 m. This setting is also in line with the function setting in Google Street View API, in which the default radius to search for the nearest SVI is 50 m, centered on the given geographical location.

4.3.2. SVI feature extraction

To take full advantage of the SVIs, semantic features are extracted through two approaches. The first kind of features are extracted by transfer learning from pretrained deep convolutional neural network model. Specifically, semantic features of SVIs are extracted by Places-CNN, a convolutional neural network used for ground-level scene recognition, and the model is trained on the Places365 dataset (Zhou et al., 2018) which is a large scene-centric image dataset with more than 10 million images of indoor and outdoor scenes labeled with diverse scene semantic categories, including both urban and nature scenes. The features extracted from this model can be distinctive to distinguish street-level images of over 300 categories.

The second kind of features are the statistical features of semantic categories after recognizing the objects presented in street view images. Specifically, the state-of-the-art semantic segmentation deep neural network DeepLab-v3+ (Chen et al., 2018) trained on Cityscapes dataset (Cordts et al., 2016), which focuses on the semantic understanding of urban street scenes, has been exploited to recognize the categories of all the pixels of the SVIs. The model can classify the pixels of SVIs into 19 classes, including road, sidewalk, building, wall, fence, pole, traffic light, traffic sign, vegetation, terrain, sky, person, rider, car, truck, bus, train, motorcycle, bicycle. Then, we can compute the number of pixels of different categories and use the distribution of categorical pixel numbers as feature to represent SVIs.

Since each sampled location has four images facing different directions (i.e., 0, 90, 180, 270 degrees), the extracted features from the four images are concatenated together to represent the location. We also empirically validated that the combination of all the four images outperform using single image in Section 6.

4.4. LCZ classification and mapping fusing RSI and SVI

To exploit both satellite and street view images, we propose two strategies to fuse the two kinds of data, one is in the feature level, and the other is in the decision level.

4.4.1. Feature-level fusion

For feature-level fusion, we concatenate the features extracted from RSI and SVI together, which are denoted as F_r and F_s, respectively. For each mapping unit, the SVI features are firstly aggregated into a fixed feature vector by applying the permutation-invariant aggregation function $\mathrm{aggregate}(\cdot)$, such as mean, max pooling, or bag-of-features (BoF). For those mapping units without SVIs, we use features with the

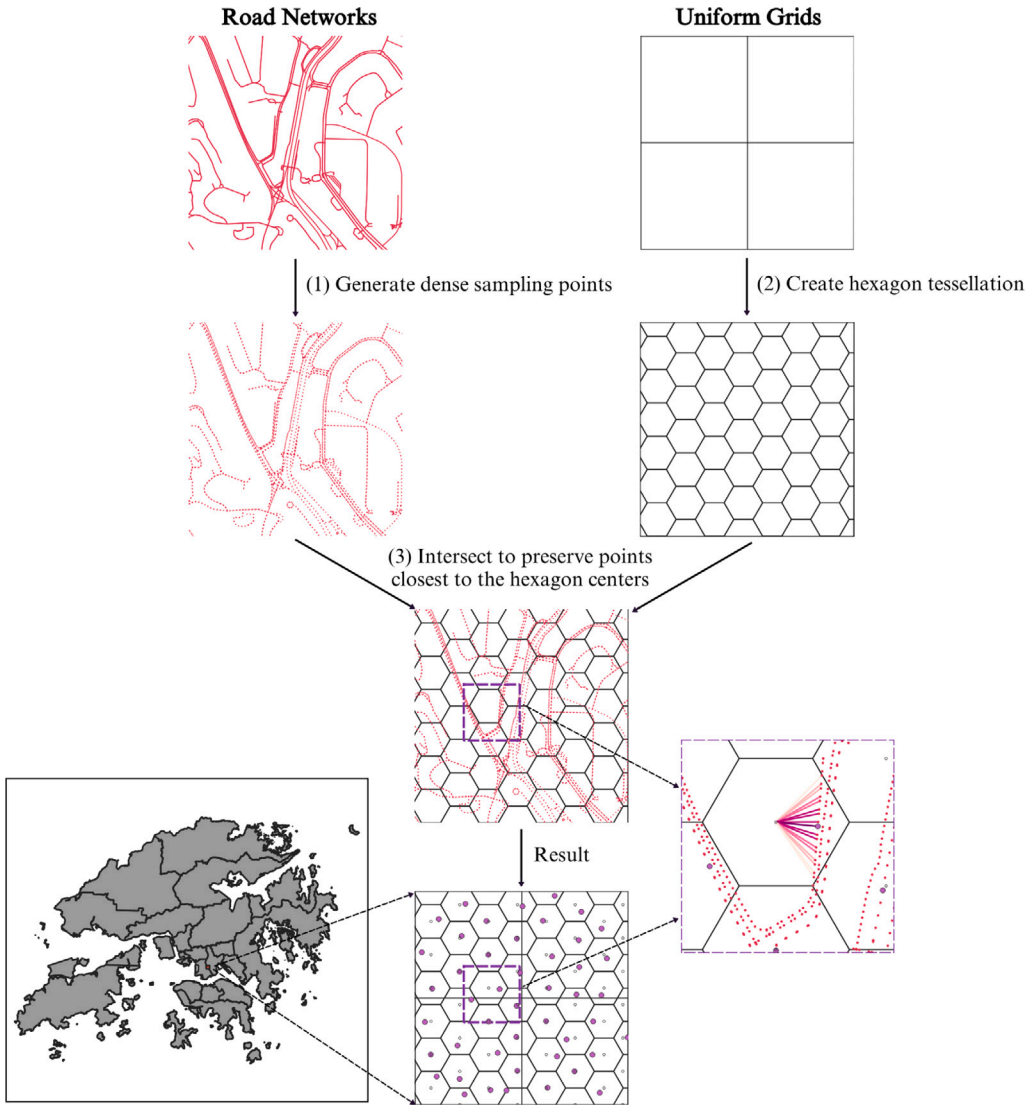


Fig. 6. Workflow of the proposed SVI sampling method, including: (1) generate dense sampling points along road networks with an interval of 10 m, (2) generate tessellated grid of hexagons across the study area to cover all the mapping units, (3) intersect the generated sampling points and hexagons to preserve only the points closest to the hexagon centers.

same size filled with the mean of all the SVI features. The obtained feature vector F_u of a mapping unit u can be formulated as follows:

$$F_u = \text{concat}\left(F_r, \text{aggregate}\left(\{F_s^i\}_{i=1}^n\right)\right) \quad (1)$$

where F_r is the corresponding RSI feature of the mapping unit u, n denotes the number of SVI points within u, and F_s^i is the feature of one of the SVI points i. Then, based on the concatenated features $\{F_u\}$ and reference dataset, we use classification models (such as XGBoost) to train classifiers and perform classification for final LCZ mapping. To illustrate this process, the details of feature-level data fusion for LCZ mapping are presented in Algorithm 1.

4.4.2. Decision-level fusion

For decision-level fusion, we first use two separate classification models (e.g., XGBoost) to classify all the RSIs and SVIs based on the features extracted from RSI and SVI respectively, obtaining the predicted probability distributions of all the classes $\hat{p}_r$ and $\hat{p}_s^i$, respectively. Then, for those mapping units with both RSI and SVI, we will firstly obtain the mean of all the predicted probability from SVI, and then normalize it to make it a unit vector; then we will calculate the sum of the probability distributions and make the prediction as the class $\hat{t}$ with the largest probability:

$$\hat{t} = \arg\max_t \left(\hat{p}_r + \text{normalize}\left(\frac{\sum_{i=1}^n \hat{p}_s^i}{n}\right)\right)_t, t = 1, 2, \ldots, 17 \quad (2)$$

where $\hat{p}_r$ is the predicted probability distribution of LCZ classes obtained from RSI in the mapping unit, n denotes the number of SVI points within the mapping unit, and $\hat{p}_s^i$ is the predicted probability distribution of LCZ classes obtained from SVI point i. While for those units with only RSIs, the predicted classes will solely base on RSIs. To illustrate this process, the details of decision-level data fusion for LCZ mapping are presented in Algorithm 2.

5. Experiments and results

5.1. Experimental setup

5.1.1. Evaluation metrics

In the experiments, following previous research (Bechtel et al., 2017, 2020; Zhu et al., 2020), the overall accuracy (OA), average accuracy (AA), weighted accuracy (WA), Kappa coefficient, and F1 scores are used as evaluation metrics for performance assessment. Besides, the

Algorithm 1 Feature-level cross-view image fusion for LCZ classification

Input: RSI feature set $\mathcal{F}_r = \{F_r^i\}_{i=1}^N$, SVI feature set $\mathcal{F}_s = \{F_s^i\}_{i=1}^M$, mapping unit set $\mathcal{U} = \{u_i, t_i | t_i \in \mathcal{L} \cup \{void\}\}_{i=1}^N$ ($\mathcal{L} = \{LCZ_j\}_{j=1}^{17}$)
Input: Classification model f_Θ (Θ denotes learnable parameters)
Output: Predicted classes of mapping units $\{u_i, \hat{t}_i\}_{i=1}^N$

```
for u_i ∈ U do                                   ▷ Feature-level data fusion
    n_i = GetNumOfSVI(u_i)
    if n_i > 0 then
        F_u^i = concat(F_r^i, aggregate({F_s^j}_{j=1}^{n_i}))  (as in Eq. (1))
    else
        F_u^i = concat(F_r^i, (Σ_{j=1}^M F_s^j)/M)
    end if
end for
D_train = {F_u^i, t_i | (u_i, t_i) ∈ U ∧ t_i ∈ L}          ▷ Model training
f_Θ̂ = argmin_Θ (CrossValidation(D_train, f_Θ))
for u_i ∈ U do                                   ▷ Model evaluation for LCZ mapping
    t̂_i = f_Θ̂(F_u^i)
end for
```

Algorithm 2 Decision-level cross-view image fusion for LCZ classification

Input: RSI feature set $\mathcal{F}_r = \{F_r^i\}_{i=1}^N$, SVI feature set $\mathcal{F}_s = \{F_s^i\}_{i=1}^M$, mapping unit set $\mathcal{U} = \{u_i, t_i | t_i \in \mathcal{L} \cup \{void\}\}_{i=1}^N$ ($\mathcal{L} = \{LCZ_j\}_{j=1}^{17}$)
Input: Classification model f_Θ^r, f_Θ^s (Θ denotes learnable parameters)
Output: Predicted classes of mapping units $\{u_i, \hat{t}_i\}_{i=1}^N$

```
D^r_train = {F_r^i, t_r^i | (u_i, t_r^i) ∈ U ∧ t_r^i ∈ L}      ▷ Model training
f^r_Θ̂ = argmin_Θ (CrossValidation(D^r_train, f^r_Θ))
D^s_train = {F_s^i, t_s^i | (u_i, t_s^i) ∈ U ∧ t_s^i ∈ L}      ▷ Model training
f^s_Θ̂ = argmin_Θ (CrossValidation(D^s_train, f^s_Θ))
for u_i ∈ U do                   ▷ Decision-level data fusion for LCZ mapping
    p̂_r^i = f^r_Θ̂(F_r^i)
    n_i = GetNumOfSVI(u_i)
    if n_i > 0 then
        for j ≤ n_i do
            p̂_s^j = f^s_Θ̂(F_s^j)
        end for
        t̂_i = argmax_t (p̂_r^i + normalize((Σ_{j=1}^{n_i} p̂_s^j)/n_i))_t , t ∈ L  (as in Eq. (2))
    else
        t̂_i = argmax_t (p̂_r^i)_t , t ∈ L
    end if
end for
```

overall accuracy of urban types (OA_{urb}) (type 1–10), and overall accuracy of natural types (OA_{nat}) (type A–G), are also leveraged to measure the performances between built-up and non-built-up regions (Yoo et al., 2019; Qiu et al., 2020). The average values of 5-fold cross-validation results of the evaluation metrics are reported.

5.1.2. Dataset settings

The WUDAPT (Bechtel et al., 2015) reference data are originally polygons with labels of 17 LCZ types, as shown in Fig. 4. The generated mapping units (as described in Section 4) are labeled as the LCZ types based on the rasterized polygonal labels. Satellite image patches within the labeled units are assigned with the same labels. SVIs are represented as points in space, and thus the ones within labeled polygons will be assigned with the same labels as the polygons where they locate.

Then, for RSIs, a dataset consisting of 2555 labeled grids are obtained, each grid is represented by a feature vector and has a counterpart label. Then, we split this whole dataset into training and testing sets with a 5-fold cross-validation using the stratified sampling method.

Table 1
Overall classification results using different data sources. The best results are highlighted in bold. (Note that SVI has smaller numbers of training and testing samples due to limited spatial coverage.)

Method	OA	WA	AA	Kappa	Avg F1
SVI	0.6274	0.8978	0.4550	0.5817	0.4436
RSI	0.7926	0.9424	0.4510	0.7032	0.4476
Feature-level Fusion	0.8129	**0.9527**	**0.5054**	0.7319	**0.5025**
Decision-level Fusion	**0.8157**	0.9513	0.5013	**0.7363**	0.4973

For SVIs, due to limited spatial coverage, we get a whole dataset consisting of 1888 labeled sample points with 4 images per location. The labeled points fall into less than 600 grids, which means that each grid contains variable number of sampling points, ranging from 0 to 15. The data split follows RSIs in space.

5.1.3. Model settings

For the feature extracting ResNet-50 model of RSIs, we load the model weights from the pretrained ResNet-50 model of So2Sat LCZ42 dataset (Zhu et al., 2020). Then we freeze the model parameters, modify the output dimension of the fully-connected layer to 2048-d and initialize the weights and bias of the fully-connected layer to an identity matrix and zero respectively. For the feature extracting PlacesCNN (with ResNet-50 as backbone) model of SVIs, we load the model weights from the pretrained ResNet-50 model of Place365 dataset (Zhou et al., 2018). Then we freeze the model parameters, modify the output dimension of the fully-connected layer to 2048-d and initialize the weights and bias of the fully-connected layer to an identity matrix and zero respectively.

For the classification task, we use the Random Forests and Support Vector Machine (SVM) of the scikit-learn package with the default hyperparameters. We also set the class weight as balanced which will adjust the learning objective taking into account the number of samples in each category. For XGBoost, we use the xgboost package with default settings and also set class balanced settings.

5.2. LCZ classification and mapping results

5.2.1. Overall classification results

The overall classification results of using different data sources are presented in Table 1. It can be seen that the SVIs can achieve a relatively good performance with OA over 62%. In contrast, RSI can achieve an accuracy of more than 79%. The fusion of RSI and SVI can significantly improve the classification performance in terms of all the evaluation metrics. The improvement is particularly noticeable on class-wise metrics including AA and Average F1 score, with about 4%–5% increase for decision-level and feature-level fusion respectively, which indicates that the addition of SVIs can help improve the classification of some difficult-to-recognize categories.

5.2.2. Comparison between urban and natural types

The classification results of built-up and non-built-up LCZ types are presented in Table 2. We can find that OA values lie in the middle of the OA_{urb} and OA_{nat}. The classification results of natural types (type A–G) are significantly higher than that of the urban types (type 1–10), which implies that urban built-up landscapes are more complex than the natural ones and thus are more difficult to recognize. SVI has relatively much more accurate recognition accuracy on urban categories than that of RSI, and thus their fusion with RSI can further boost the performance in OA_{urb} significantly.

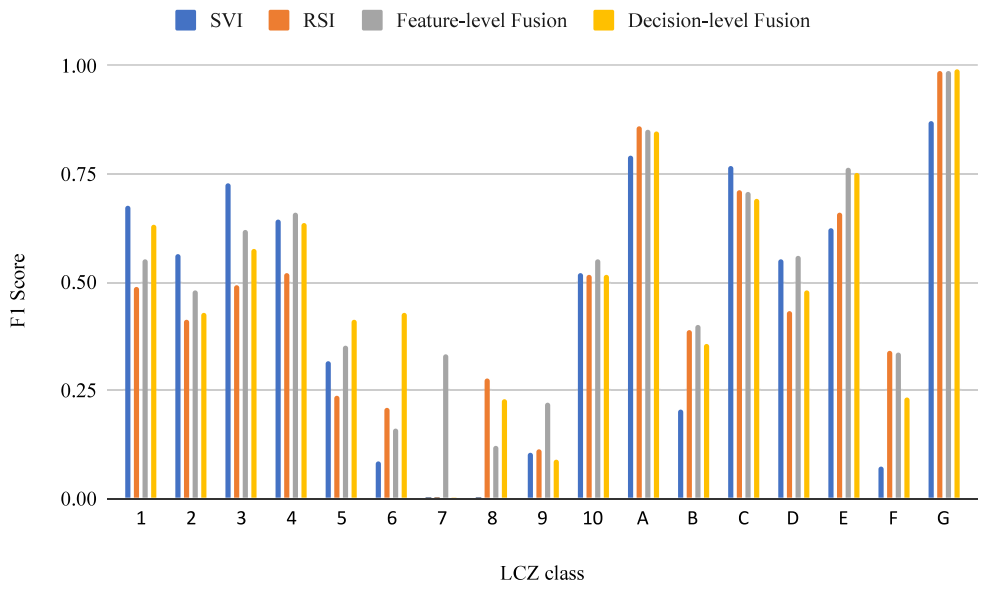


Fig. 7. Bar chart of per-class F1 scores using different data sources.

Table 2
Classification results of urban (built-up) and natural (non-built-up) categories. The best results are highlighted in bold. (Note that SVI has smaller numbers of training and testing samples due to limited spatial coverage.)

Method	OA	OA_{urb}	OA_{nat}
SVI	0.6274	0.5800	0.6842
RSI	0.7926	0.4083	0.8669
Feature-level Fusion	0.8129	0.4880	**0.8758**
Decision-level Fusion	**0.8157**	**0.5267**	0.8716

5.2.3. Per-class classification results

To further investigate into class-wise classification performance, the per-class F1 scores are presented in Fig. 7. As can be seen, most classes witness performance boost in fusing both RSIs and SVIs compared with using RSIs alone. We can see that the performance of class G (*water*) nearly all over 0.9, with RSI and data fusion over 0.98, significantly outperforms all the other classes. Besides, all the methods also achieve high F1 scores (all over 0.79) on class A (*dense trees*). Urban types of class 1 (*compact highrise*), class 3 (*compact lowrise*), class 4 (*open highrise*) and class 10 (*heavy industry*) can achieve relatively good performance with F1 scores of about 0.5 for all the methods; while natural types of class C (*bush, scrub*) and class E (*bare rock or paved*) can also achieve good performance with F1 scores all over 62% for all the methods. For class 7 (*lightweight lowrise*), only feature-level fusion can recognize them. This shows the difficulty of distinguishing this category because the classes are relatively rare in Hong Kong, which is in line with previous research in Hong Kong using official GIS-based method (Wang et al., 2018a). In addition, class 8 (*large lowrise*), class 9 (*sparsely built*), class B (*scattered trees*), and class F (*bare soil or sand*) are also relatively difficult to distinguish with significantly lower F1 scores. It is also of note that SVIs perform significantly better than RSIs on urban types of class 1 (*compact highrise*), class 2 (*compact midrise*), class 3 (*compact lowrise*), class 4 (*open highrise*), class 5 (*open midrise*), and natural types of class C (*bush, scrub*) and class D (*low plants*), which implies that street-level images can provide more information about built-up regions and ground-level details.

In order to figure out the classification relationships between different categories. The normalized confusion matrices of the classification results using different data sources are presented in Fig. 8. As can be seen, in general, SVI and RSI based results differs due to different data sources. While for RSI, and the fusion of RSI and SVI, the results are basically consistent, and the additional SVI data can help further boost the performance for most classes.

5.2.4. LCZ mapping results

The LCZ mapping results are presented in Fig. 9. As can be seen from Fig. 9(b), the mapping results of SVIs do not cover the study area completely, since SVIs are sparsely distributed along road networks. We can also notice the density of road networks in Hong Kong from the results. For the mapping results of RSIs (Fig. 9(c)) and fusing RSI and SVI (Fig. 9(d)), the results are visually similar, and are basically in line with the mapping results from previous research results of Hong Kong (Wang et al., 2018a; Zheng et al., 2018; Liu and Shi, 2020). In addition, the fusion of RSI and SVI (Fig. 9(d)) can help improve some details compared with using RSI alone (Fig. 9(c)), such as small islands and the airport. These demonstrate the reliability of our results and superiority of fusing SVI data.

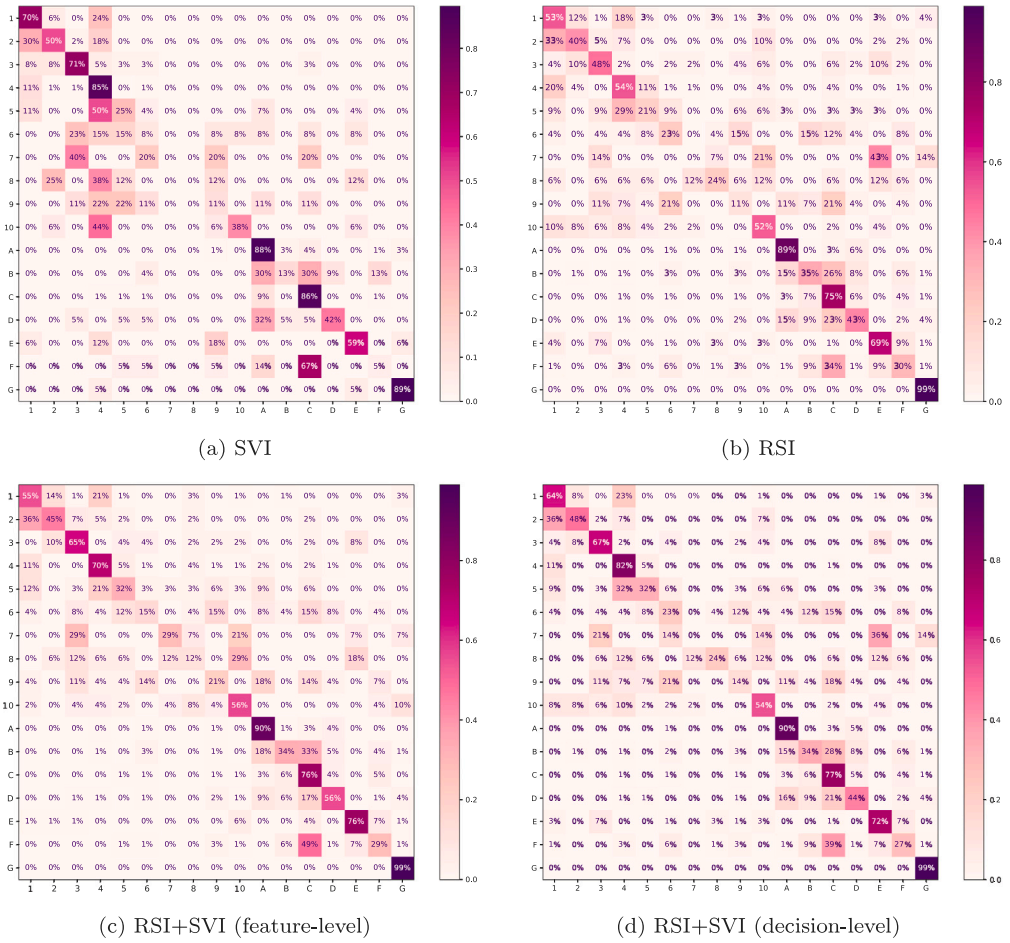
5.3. Evaluation on SVI sampling strategy

Most previous research on SVI adopts the equidistant sampling strategy along roads with a fixed interval (Zhang et al., 2019). However, this is costing in terms of time and money for large scale (such as city-scale or even region-scale) SVI research, and there are possible data redundancy. To address these issues, we propose an effective SVI sampling strategy to accommodate coverage, efficiency, and performance simultaneously.

5.3.1. Evaluation on sampling efficiency

The requested numbers of SVI points and images of using different sampling methods are presented in Table 3. In our study area, Hong Kong, the OSM road networks contain a total number of 326,182 road segments. With the strategy of equidistant sampling along the roads with an interval of 10 m, 2,799,079 SVI points are generated (which means there will be over 8 points generated for each road segment on average), resulting in over 11 million images to request for download via Google Street View API. This is quite inefficient and time- and money-costing.

With the proposed sampling method, we reduce the candidate SVI points into 69,957 points, the number of which is only 2.5% of the number of candidate points using the naive equidistant sampling method (2,799,079). Finally, we request for 69,957 SVI points via Google API, and only retrieve 32,622 points with valid Google SVIs. The sampling



(a) SVI (b) RSI

(c) RSI+SVI (feature-level) (d) RSI+SVI (decision-level)

Fig. 8. Normalized confusion matrices of the classification using different data sources: (a) SVI only, (b) RSI only, (c) fusing RSI and SVI in feature level, and (d) fusing RSI and SVI in decision level.

Table 3
SVI sampling results using different strategies (Equid: equidistant sampling, hex: hexagonal constraint). The best results are highlighted in bold.

Method	#Point	#Image
Equid-10 m	2799079	11196316
Equid-20 m	1651832	6607328
Equid-30 m	1275272	5101088
Equid-50 m	982074	3928296
Equid-10 m + hex (Ours)	**69957**	**279828**

result in the *Yau Tsim Mong* district is presented in Fig. 10. The district is one of the busiest downtown areas in Hong Kong, with highly dense road networks. We can see that our proposed sampling method can significantly reduce the number of sampled SVI points while remain sufficient coverage and representativeness of all the mapping units.

5.3.2. Evaluation on classification and mapping performance

To further evaluate the impact of the proposed sampling method on LCZ classification performance, we conduct experiments on using SVIs sampled by different methods, the number of sampled SVI points, number of SVIs, and corresponding classification performance are shown in Table 4.

It can be seen that the proposed sampling method only results in 1888 SVI points and 7552 images in the labeled regions, which are less than 1/18 that of naive equidistant sampling (34,477 points and 137,908 images). This is a huge reduction in requested SVI numbers and can significantly reduce the cost for SVI data collection and processing. While for the classification performance, although the results of using SVI data alone decrease to a reasonable degree, the final performances after fusing RSI and SVI present consistent trends using different sampling methods, and our proposed method can achieve very competitive performances over naive equidistant sampling, with OA and Kappa coefficient improved over 2%–3% and AA and average F1 improved over 5% compared to only using RSI. This also implies that there are redundancy in SVI and RSI data and good sampling strategy can help alleviate the redundancy while retain competitive classification performance. The results demonstrate the effectiveness of the proposal SVI sampling strategy in remaining competitive classification performance with much less cost for SVI requirement.

To further evaluate the impact of the proposed sampling method on LCZ mapping performance, five representative districts in the downtown of Hong Kong with highly dense road networks are selected, i.e., *Yau Tsim Mong*, *Kowloon City*, *Central and Western*, *Wan Chai*, and *Eastern* (refer to Fig. 2). For fair comparison, for each sampling method, the training and testing samples are collected based on the same sampling strategy. The final mapping results are shown in Fig. 11. We can see that the SVI-based mapping using equidistant sampling results in relatively more smooth and cluttered mapping results, with large areas of the same LCZ types compared with our proposed method. This may be caused by the cluttered sampled SVIs near the boundaries of mapping units, and thus capturing the information of nearby units. For the mapping results fusing RSI and SVI, we can see that the results are relatively consistent and are visually similar to the result of using RSI alone. The results further demonstrate the effectiveness of the proposed SVI sampling method in retaining reliable and consistent LCZ mapping results.

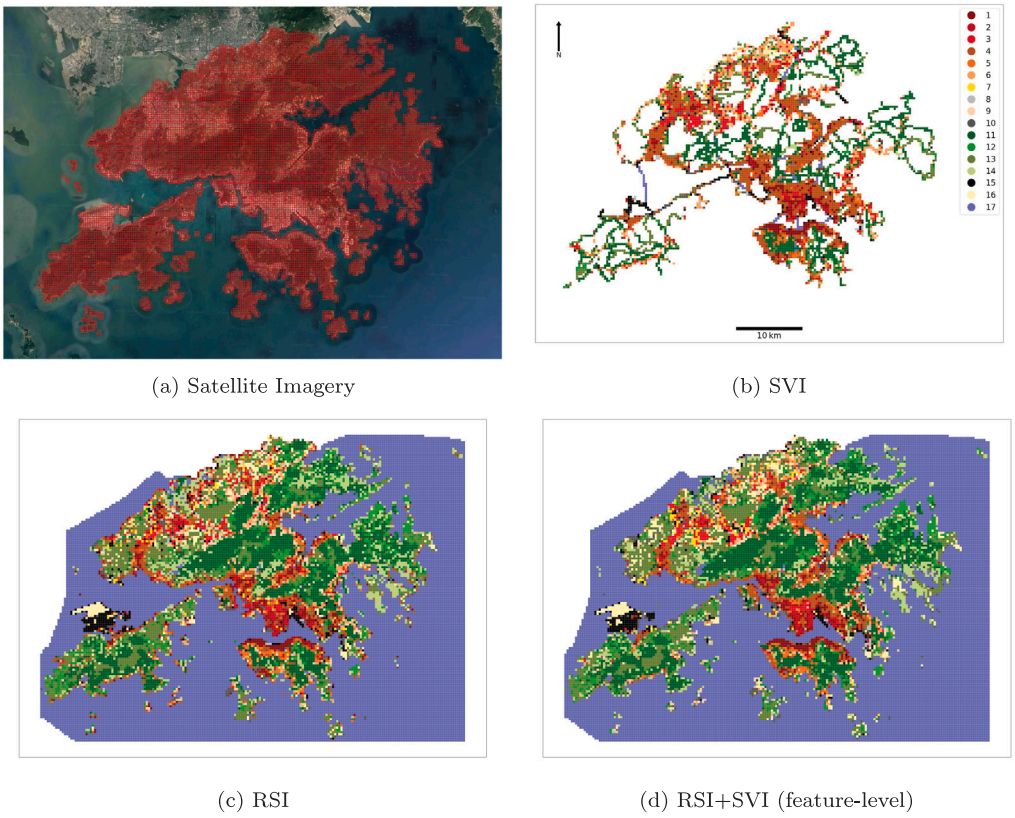


Fig. 9. LCZ mapping results in the study area using different data sources. (a) Sentinel-2 satellite imagery (RGB composite); mapping results based on (b) SVI only, (c) RSI only, (d) fusing RSI and SVI in feature level.

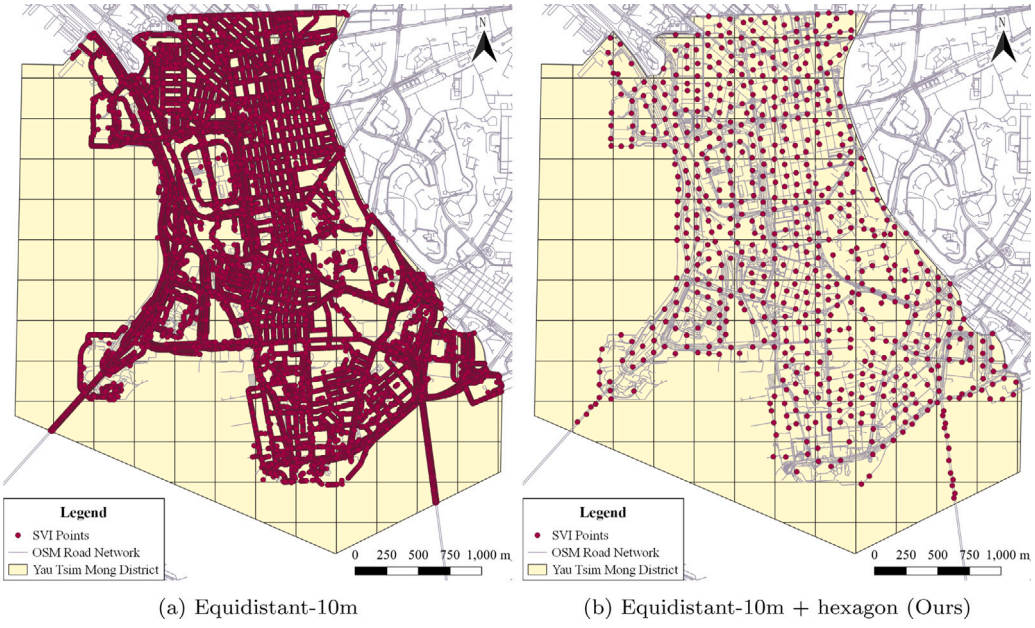
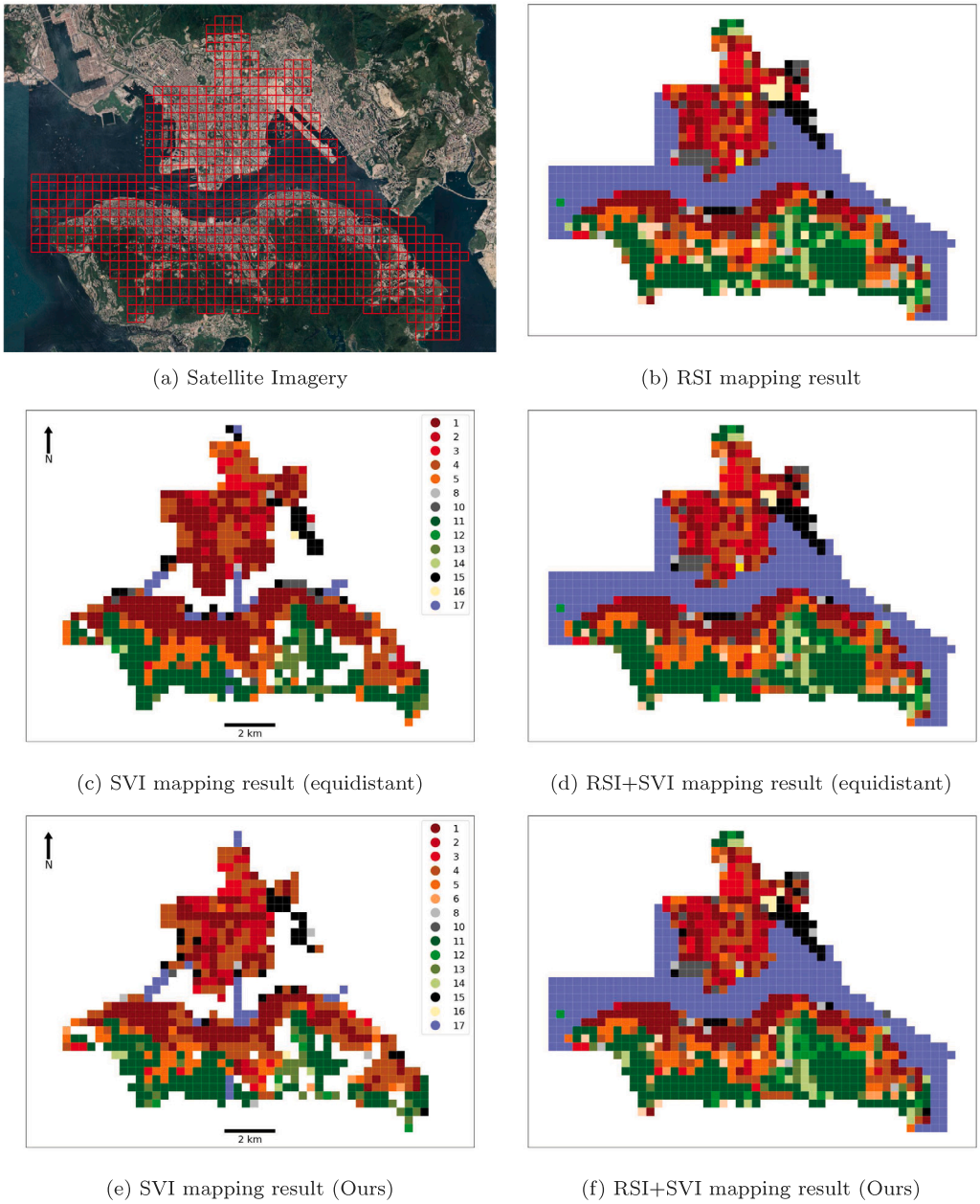


Fig. 10. Spatial distribution of SVI sampling points using different sampling methods. (a) Equidistant sampling with 10 m interval, (b) Equidistant sampling with 10 m interval and the proposed hexagonal constraint.

Table 4
Classification results using different SVI sampling strategies (Equid: equidistant sampling, hex: hexagonal constraint). The best results are highlighted in bold.

Sampling	Method	OA	WA	AA	Kappa	Avg F1
	RSI	0.7926	0.9424	0.4510	0.7032	0.4476
Equid-10m	SVI	0.7332	0.9238	0.5626	0.7022	0.5532
#Point: 34477	Feature-level fusion	0.8231	**0.9551**	0.5217	0.7466	0.5226
#Image: 137908	Decision-level fusion	**0.8247**	0.9543	**0.5283**	**0.7492**	**0.5245**
Equid-10 m + hex	SVI	0.6274	0.8978	0.4550	0.5817	0.4436
#Point: 1888	Feature-level fusion	0.8129	**0.9527**	**0.5054**	0.7319	**0.5025**
#Image: 7552	Decision-level fusion	**0.8157**	0.9513	0.5013	**0.7363**	0.4973



(a) Satellite Imagery (b) RSI mapping result

(c) SVI mapping result (equidistant) (d) RSI+SVI mapping result (equidistant)

(e) SVI mapping result (Ours) (f) RSI+SVI mapping result (Ours)

Fig. 11. LCZ mapping results in the selected five districts using different sampling methods. (a) Sentinel-2 satellite imagery (RGB composite); mapping results based on (b) RSI only, (c) SVI only (equidistant sampling), (d) fusing RSI and SVI in feature level (equidistant sampling), (e) SVI only (equidistant sampling + hexagonal constraint), (f) fusing RSI and SVI in feature level (equidistant sampling + hexagonal constrain).

Table 5
Classification results of SVI with different heading directions. The best results are highlighted in bold.

Heading	OA	WA	AA	Kappa	Avg F1
0°	0.5747	0.8734	0.4127	0.5168	0.4197
90°	0.5826	0.8747	0.4254	0.5265	0.4282
180°	0.5948	0.8832	0.4386	0.5408	0.4420
270°	0.5821	0.8812	0.4168	0.5261	0.4193
All	**0.6472**	**0.9054**	**0.4918**	**0.6003**	**0.5008**

6. Discussion

6.1. Evaluation on SVI classification

Accurate image-level SVI classification results are the basis for following LCZ mapping in space. To evaluate the varying effects on SVI classification performance, we have examined the effects of SVI heading directions and different data split settings and evaluation methods in assessing LCZ classification results in the following subsections.

6.1.1. Evaluation on the effect of heading directions

SVIs with different heading directions will capture different views. To evaluate the effect of including different directions for LCZ classification, we have examined the results of using single-heading SVIs, i.e., 0, 90, 180, and 270 degrees, respectively; and combining SVIs with all the headings together. The results are presented in Table 5.

We can see that the overall classification results of SVIs captured from different heading directions are relatively similar for all the evaluation metrics. The combination of all the images from different heading directions can significantly boost the results on all the evaluation metrics. Therefore, all the images are exploited in our experiments.

6.1.2. Evaluation on different data splits and evaluation levels

SVIs are represented as sparse points in space. To derive LCZ maps, the sparse points need to be aggregated into land parcels. There are different data split approaches to separate training and testing data for SVIs. The first approach is to use stratified sampling at image level, and the second approach is to follow the split settings as the mapping units, which will take into consideration of the spatial distribution of SVIs. There are also different perspectives to evaluate SVI-based classification, including image level and parcel level. Image-level evaluation regards street view images as individuals, while parcel-level evaluation firstly aggregates the information of images into land parcels and then evaluate the measures on the mapping units.

To evaluate the effects of different data splits and evaluation levels, we compare the results of different settings, as shown in Table 6. As can be seen, for stratified sampling, the overall classification performances of image-level and parcel-level evaluations are similar, and the aggregation of SVI points into land parcels does not result in significant changes. While when following the data split of mapping units, the performances between image and parcel levels vary significantly, with OAs differ in over 2% and AAs and average F1 scores differ in 3%. Moreover, when comparing at the same level of evaluation, the results of using stratified sampling outperform that of following mapping units noticeably. These results implies the importance of both data splits and evaluation levels.

6.2. Evaluation on different features

There are different ways to extract features from SVIs and RSIs, such as the methods introduced in Section 4. To investigate the effectiveness of different features, we conduct ablation studies on using different features of SVI and RSI, the results are presented in Tables 7 and 8, respectively. We can see that the results of using different features vary noticeably. For SVI features, the semantic features extracted by PlacesCNN obtain the best results, significantly outperforming the features using semantic segmentation network DeepLab-v3. The fusion of the features from PlacesCNN and DeepLab does not improve the results. While for RSI, the spectral features beat the deep features extracted by ResNet50, and the combination of spectral features and deep features does not improve the results. Therefore, in our experiments, we use the PlacesCNN features to represent SVIs while use the spectral features to represent RSI patches.

6.3. Evaluation on different aggregation methods for fusion

When fusing RSI and SVI, a key step is to aggregate the information provided by variable numbers of SVIs within mapping units. There are different aggregation methods, including average, max pooling, BoF, etc. To investigate the effectiveness of different aggregation methods, we conduct experiments on the above three methods, and the results are presented in Table 9. It can be seen that the three different methods achieve relatively consistent results, with the average method achieving the best performance. Therefore, we adopt the average aggregation method in our feature-level fusion experiments in this study.

6.4. Evaluation on different classifiers

To demonstrate the generalization of the improvement of fusing SVIs with RSIs for LCZ mapping, besides XGBoost, we experiment on other widely used classification models, including Random Forests and Support Vector Machine. The results of using different classifiers are shown in Table 10.

As can be seen, in general, compared with using only RSI, the classification performances have been enhanced noticeably when fusing SVI with RSI in all the evaluation metrics. This demonstrates the usefulness of SVIs in help improve the LCZ classification performance. Furthermore, we can find that XGBoost model can outperform the RF model noticeably in most metrics and the SVM model generally has lower performance than the other tree-based classification models.

6.5. Practical applications and limitations

Street-level images contain useful ground-level information which remote sensing images lack and thus it is promising to integrate them for LCZ mapping as well as more general urban thermal environment and urban climate studies. The major contributions of SVIs in LCZ-related studies can be summarized as follows.

Firstly, due to the inherent limitation of remote sensing images, the accuracy of LCZ classification using remote sensing imagery alone may be limited, especially when it comes to the LCZ categories with building information (e.g. LCZ types 1–3), resulting in LCZ classification results that are not practically applicable. These problems are particularly evident in the study of urban LCZs in China, and building height information is significant in improving the accuracy of the identification of these categories (Ren et al., 2019). Usually, DSM data and 3D building data contain the above information, but these data are often not available in many areas. Street-level images, on the other hand, is currently available in a wide range of areas and can be collected efficiently when necessary, making it more efficient than traditional field work. Therefore, street view image data can provide easily accessible ground details, and the fusion of remote sensing and street view images can significantly improve LCZ classification accuracy (as demonstrated by the experimental results in our study), while high accuracy LCZ data can support more accurate decision making. In addition, interpretability is also an important factor in decision making, and street view images can provide more intuitive and easy-to-understand data support during case study.

Secondly, LCZ is mainly used for studies related to urban thermal environment and urban climate, such as urban heat island effect (Stewart and Oke, 2012; Zhu et al., 2022); in practical studies, street view images can obtain relevant information that is difficult to obtain from

Table 6
Classification results of SVI with different training/testing data splits and different evaluation levels.

Data split	Evaluation level	OA	WA	AA	Kappa	Avg F1
Stratified sampling	Image-level	0.6472	0.9054	0.4918	0.6003	0.5008
	Parcel-level	0.6459	0.9074	0.4961	0.6018	0.5042
Follow mapping units	Image-level	0.6026	0.8905	0.4250	0.5492	0.4161
	Parcel-level	0.6274	0.8978	0.4550	0.5817	0.4436

Table 7
Classification results of SVI using different features. The best results are highlighted in bold.

Method	OA	WA	AA	Kappa	Avg F1
PlacesCNN	**0.6274**	**0.8978**	**0.4550**	**0.5817**	**0.4436**
DeepLab	0.5340	0.8754	0.3598	0.4762	0.3414
PlacesCNN+DeepLab	0.6236	0.8962	0.4542	0.5784	0.4320

Table 8
Classification results of RSI using different features. The best results are highlighted in bold.

Method	OA	WA	AA	Kappa	Avg F1
Spectral	**0.7926**	0.9424	**0.4510**	**0.7032**	**0.4476**
ResNet50	0.7800	0.9418	0.4314	0.6849	0.4267
Spectral+ResNet50	0.7863	**0.9441**	0.4502	0.6938	0.4465

Table 9
Classification results using different SVI aggregation methods for feature-level data fusion. The best results are highlighted in bold.

Method	OA	WA	AA	Kappa	Avg F1
Average	**0.8129**	**0.9527**	**0.5054**	**0.7319**	**0.5025**
Max pooling	0.8027	0.9459	0.4899	0.7172	0.4913
BoF	0.7992	0.9456	0.4731	0.7124	0.4752

remote sensing images, including: 3D building information (e.g., building height and vertical spatial layout); building facade texture and material; height and volume of trees and vegetation; uneven road and terrain information; etc. These kinds of information is crucial for modeling the urban climate and thermal environment. Future research can extract the relevant information from street-level images in a more refined and targeted manner to serve research related to the urban thermal environment and urban climate, as well as derived research topics such as urban energy and solar cities (Zhu et al., 2023).

Although street-level image data has many advantages and can provide a wealth of information, and the integration of remote sensing and street-level imagery is promising to combine useful information to solve more problems; there are still some limitations in the street view image data and in this study. Firstly, at the data level, there are inherent limitations in the street view image data, with limited spatial and temporal coverage and resolution. Street view images are sparsely distributed along roads thus limiting its spatial coverage and resolution. In addition, the update of SVIs is difficult to control which is determined by the street view service providers. The street scenes may change as time goes by and the update frequency is highly uncertain across different regions. All these limitations on spatial and temporal coverage have significantly limited the availability of SVIs. Studies on how to make full use of limited SVIs are worth of further efforts. Secondly, the main study area of this study is in Hong Kong, and the study area can be extended in the future to further figure out whether the method of fusing remote sensing and street view image data is equally effective for other areas and what differences exist. In addition, this study mainly adopts the traditional machine learning method to achieve a balance between accuracy and cost, which may limit the potential of street view images in terms of accuracy. In the future, alternative data fusion methods can be explored on the basis of the current method, such as using more advanced deep learning methods and fusion strategies to further improve the accuracy of data fusion.

7. Conclusion

Timely and accurate LCZ classification maps are important for urban climate research. While remote sensing images have shown to be useful for LCZ mapping, it often falls short in providing crucial ground-level details that are essential for accurate classification. Street-level images offer an alternative perspective that can fill this gap. In this study, we propose an effective method to integrate satellite and street-level images for LCZ mapping, which can make full use of them to improve the LCZ classification and mapping performance. In addition, we propose a simple yet effective sampling method to significantly reduce the number of required street-level images while maintaining high LCZ mapping performance. Extensive experiments have been carried out and the results demonstrate the effectiveness of the proposed cross-view data fusion method and the efficacy of the proposed street view image sampling method. The study has illuminated the potential value of integrating street-level images to enhance LCZ mapping and can further benefit urban climatic studies.

In the future, there are several directions worth further investigation. Firstly, without considering computational cost, alternative data fusion methods can be further explored to make full use of street view images and further improve the LCZ mapping accuracy by using state-of-the-art deep learning-based methods. Secondly, figuring out how street-level images can help augment the LCZ mapping results can help us better understand the contribution of SVIs and further refine the strategies in exploiting the data. Finally, the current study only focuses on one city, it is important to further investigate the LCZ mapping results in more regions and compare the results to evaluate the generalization performance of the proposed methods.

Declaration of competing interest

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.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 42101472 and 42071360, the Hong Kong Polytechnic University Start-Up, Hong Kong SAR of China under Grant BD41, and the Microsoft AI for Earth Grant, USA.

Table 10
Classification results of using different classifiers. The best results are highlighted in bold.

Classifier	Method	OA	WA	AA	Kappa	Avg F1
XGBoost	SVI	0.6274	0.8978	0.4550	0.5817	0.4436
	RSI	0.7926	0.9424	0.4510	0.7032	0.4476
	Feature-level Fusion	0.8129	**0.9527**	**0.5054**	0.7319	**0.5025**
	Decision-level Fusion	**0.8157**	0.9513	0.5013	**0.7363**	0.4973
RF	SVI	0.5440	0.8677	0.3275	0.4789	0.3032
	RSI	0.7961	0.9431	0.4373	0.7072	0.4354
	Feature-level Fusion	0.8020	0.9448	0.4375	0.7152	0.4243
	Decision-level Fusion	**0.8074**	**0.9482**	**0.4578**	**0.7235**	**0.4528**
SVM	SVI	0.5883	0.8937	0.4081	0.5368	0.3788
	RSI	0.4180	0.7316	0.2323	0.2842	0.1913
	Feature-level Fusion	**0.4959**	**0.8321**	**0.3347**	**0.3653**	**0.3034**
	Decision-level Fusion	0.4830	0.7631	0.3212	0.3558	0.2722

Appendix. List of abbreviations

Abbreviation	Explanation
AA	average accuracy
BoF	bag-of-features
LCZ	local climate zone
OA	overall accuracy
OA_{nat}	overall accuracy of natural types
OA_{urb}	overall accuracy of urban types
OSM	OpenStreetMap
RF	Random Forests
RSI	remote sensing imagery
SDGs	sustainable development goals
SVI	street view image
SVM	Support Vector Machine
WA	weighted accuracy
WUDAPT	World Urban Database and Access Portal Tools

References

Barbierato, E., Bernetti, I., Capecchi, I., Saragosa, C., 2020. Integrating remote sensing and street view images to quantify urban forest ecosystem services. Remote Sens. 12 (2), 329. http://dx.doi.org/10.3390/rs12020329.

Bechtel, B., Alexander, P.J., Böhner, J., Ching, J., Conrad, O., Feddema, J., Mills, G., See, L., Stewart, I., 2015. Mapping local climate zones for a worldwide database of the form and function of cities. ISPRS Int. J. Geo-Inf. 4 (1), 199–219.

Bechtel, B., Demuzere, M., Sismanidis, P., Fenner, D., Brousse, O., Beck, C., Van Coillie, F., Conrad, O., Keramitsoglou, I., Middel, A., Mills, G., Niyogi, D., Otto, M., See, L., Verdonck, M.L., 2017. Quality of crowdsourced data on urban morphology—The human influence experiment (HUMINEX). Urban Sci. 1 (2), 15.

Bechtel, B., Demuzere, M., Stewart, I.D., 2020. A weighted accuracy measure for land cover mapping: Comment on Johnson et al. Local climate zone (LCZ) map accuracy assessments should account for land cover physical characteristics that affect the local thermal environment. Remote Sens. 2019, 11, 2420. Remote Sens. 12 (11), 1769.

Biljecki, F., Ito, K., 2021. Street view imagery in urban analytics and GIS: A review. Landsc. Urban Plan. 215, 104217.

Boeing, G., 2017. OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. Comput. Environ. Urban Syst. 65, 126–139.

Cao, R., Qiu, G., 2018. Urban land use classification based on aerial and ground images. In: 2018 International Conference on Content-Based Multimedia Indexing. CBMI 2018, la Rochelle, France, September 4–6, 2018, pp. 1–6.

Cao, R., Tu, W., Yang, C., Li, Q., Liu, J., Zhu, J., Zhang, Q., Li, Q., Qiu, G., 2020. Deep learning-based remote and social sensing data fusion for urban region function recognition. ISPRS J. Photogramm. Remote Sens. 163, 82–97.

Cao, R., Zhu, J., Tu, W., Li, Q., Cao, J., Liu, B., Zhang, Q., Qiu, G., 2018. Integrating aerial and street view images for urban land use classification. Remote Sens. 10 (10), 1553.

Chen, B., Feng, Q., Niu, B., Yan, F., Gao, B., Yang, J., Gong, J., Liu, J., 2022a. Multi-modal fusion of satellite and street-view images for urban village classification based on a dual-branch deep neural network. Int. J. Appl. Earth Obs. Geoinf. 109, 102794. http://dx.doi.org/10.1016/j.jag.2022.102794.

Chen, D., Tu, W., Cao, R., Zhang, Y., He, B., Wang, C., Shi, T., Li, Q., 2022b. A hierarchical approach for fine-grained urban villages recognition fusing remote and social sensing data. Int. J. Appl. Earth Obs. Geoinf. 106, 102661.

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

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. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27.

Hoffmann, E.J., Wang, Y., Werner, M., Kang, J., Zhu, X.X., 2019. Model fusion for building type classification from aerial and street view images. Remote Sens. 11 (11), 1259. http://dx.doi.org/10.3390/rs11111259.

Ignatius, M., Xu, R., Hou, Y., Liang, X., Zhao, T., Chen, S., Wong, N.H., Biljecki, F., 2022. Local climate zones: Lessons from Singapore and potential improvement with street view imagery. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. X-4-W2-2022. Copernicus GmbH, pp. 121–128.

Kang, Y., Zhang, F., Gao, S., Lin, H., Liu, Y., 2020. A review of urban physical environment sensing using street view imagery in public health studies. Ann. GIS 1–15.

Khajwal, A.B., Cheng, C.S., Noshadravan, A., 2023. Post-disaster damage classification based on deep multi-view image fusion. Comput.-Aided Civ. Infrastruct. Eng. 38 (4), 528–544. http://dx.doi.org/10.1111/mice.12890.

Liu, S., Shi, Q., 2020. Local climate zone mapping as remote sensing scene classification using deep learning: A case study of metropolitan China. ISPRS J. Photogramm. Remote Sens. 164, 229–242.

Qiu, C., Tong, X., Schmitt, M., Bechtel, B., Zhu, X.X., 2020. Multilevel Feature Fusion-Based CNN for local climate zone classification from sentinel-2 images: Benchmark results on the So2Sat LCZ42 Dataset. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 2793–2806.

Ren, C., Cai, M., Li, X., Zhang, L., Wang, R., Xu, Y., Ng, E., 2019. Assessment of local climate zone classification maps of cities in China and feasible refinements. Sci. Rep. 9 (1), 18848. http://dx.doi.org/10.1038/s41598-019-55444-9.

Stewart, I.D., Oke, T.R., 2012. Local climate zones for urban temperature studies. Bull. Am. Meteorol. Soc. 93 (12), 1879–1900.

Thomas, G., Sherin, A.P., Ansar, S., Zachariah, E.J., 2014. Analysis of urban heat island in Kochi, India, using a modified local climate zone classification. Procedia Environ. Sci. 21, 3–13.

United Nations, 2015. Transforming our world: The 2030 agenda for sustainable development. General Assembley 70 Session.

United Nations, 2018. World urbanization prospects: The 2018 revision.

Wamsler, C., Brink, E., Rivera, C., 2013. Planning for climate change in urban areas: from theory to practice. J. Clean. Prod. 50, 68–81.

Wang, R., Ren, C., Xu, Y., Lau, K.K.L., Shi, Y., 2018a. Mapping the local climate zones of urban areas by GIS-based and WUDAPT methods: A case study of Hong Kong. Urban Clim. 24, 567–576.

Wang, W., Yang, S., He, Z., Wang, M., Zhang, J., Zhang, W., 2018b. Urban perception of commercial activeness from satellite images and streetscapes. In: Companion Proceedings of the the Web Conference 2018. WWW '18, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pp. 647–654. http://dx.doi.org/10.1145/3184558.3186581.

Xing, Z., Yang, S., Zan, X., Dong, X., Yao, Y., Liu, Z., Zhang, X., 2023. Flood vulnerability assessment of urban buildings based on integrating high-resolution remote sensing and street view images. Sustainable Cities Soc. 92, 104467. http://dx.doi.org/10.1016/j.scs.2023.104467.

Xu, G., Zhu, X., Tapper, N., Bechtel, B., 2019. Urban climate zone classification using convolutional neural network and ground-level images. Prog. Phys. Geogr. Earth Environ. 43 (3), 410–424.

Xue, J., You, R., Liu, W., Chen, C., Lai, D., 2020. Applications of local climate zone classification scheme to improve urban sustainability: A bibliometric review. Sustainability 12 (19), 8083.

Yan, Y., Huang, B., 2022. Estimation of building height using a single street view image via deep neural networks. ISPRS J. Photogramm. Remote Sens. 192, 83–98.

Yoo, C., Han, D., Im, J., Bechtel, B., 2019. Comparison between convolutional neural networks and random forest for local climate zone classification in Mega Urban Areas using Landsat images. ISPRS J. Photogramm. Remote Sens. 16.

Zhang, F., Wu, L., Zhu, D., Liu, Y., 2019. Social sensing from street-level imagery: A case study in learning spatio-temporal urban mobility patterns. ISPRS J. Photogramm. Remote Sens. 153, 48–58.

Zheng, Y., Ren, C., Xu, Y., Wang, R., Ho, J., Lau, K., Ng, E., 2018. GIS-based mapping of local climate zone in the high-density city of Hong Kong. Urban Clim. 24, 419–448.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A., 2018. Places: A 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. 40 (6), 1452–1464.

Zhou, W., Liu, J., Lei, J., Yu, L., Hwang, J.N., 2021. GMNet: graded-feature multilabel-learning network for RGB-thermal urban scene semantic segmentation. IEEE Trans. Image Process. 30, 7790–7802.

Zhu, R., Dong, X., Wong, M.S., 2022. Estimation of the urban heat island effect in a reformed urban district: A scenario-based study in Hong Kong. Sustainability 14 (8), 4409. http://dx.doi.org/10.3390/su14084409.

Zhu, X.X., Hu, J., Qiu, C., Shi, Y., Kang, J., Mou, L., Bagheri, H., Haberle, M., Hua, Y., Huang, R., Hughes, L., Li, H., Sun, Y., Zhang, G., Han, S., Schmitt, M., Wang, Y., 2020. So2Sat LCZ42: A benchmark data set for the classification of global local climate zones. IEEE Geosci. Remote Sens. Mag. 8 (3), 76–89.

Zhu, R., Kwan, M.P., Perera, A.T.D., Fan, H., Yang, B., Chen, B., Chen, M., Qian, Z., Zhang, H., Zhang, X., Yang, J., Santi, P., Ratti, C., Li, W., Yan, J., 2023. GIScience can facilitate the development of solar cities for energy transition. Adv. Appl. Energy 10, 100129. http://dx.doi.org/10.1016/j.adapen.2023.100129.