Remote Sensing of Environment 278 (2022) 113106

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

UrbanWatch: A 1-meter resolution land cover and land use database for 22 major cities in the United States

Yindan Zhang [a], Gang Chen [a,*], Soe W. Myint [b], Yuyu Zhou [c], Geoffrey J. Hay [d], Jelena Vukomanovic [e,f], Ross K. Meentemeyer [e,g]

[a] Laboratory for Remote Sensing and Environmental Change (LRSEC), Department of Geography and Earth Sciences, University of North Carolina at Charlotte, NC 28223, USA
[b] School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ 85287, USA
[c] Department of Geological and Atmospheric Sciences, Iowa State University, Ames, IA 50011, USA
[d] Department of Geography, University of Calgary, Calgary, AB T2N 1N4, Canada
[e] Center for Geospatial Analytics, North Carolina State University, Raleigh, NC 27695, USA
[f] Department of Parks, Recreation and Tourism Management, North Carolina State University, Raleigh, NC 27695, USA
[g] Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, NC 27695, USA

ARTICLE INFO

Edited by Marie Weiss

Keywords:
Land cover and land use (LCLU)
Very high resolution (VHR)
Urban
Deep learning
UrbanWatch
Open access

ABSTRACT

Very-high-resolution (VHR) land cover and land use (LCLU) is an essential baseline data for understanding fine-scale interactions between humans and the heterogeneous landscapes of urban environments. In this study, we developed a Fine-resolution, Large-area Urban Thematic information Extraction (FLUTE) framework to address multiple challenges facing large-area, high-resolution urban mapping, including the view angle effect, high intraclass and low interclass variation, and multiscale land cover types. FLUTE builds upon a teacher-student deep learning architecture, and includes two new feature extraction modules – Scale-aware Parsing Module (SPM) and View-aware Embedding Module (VEM). Our model was trained with a new benchmark database containing 52.43 million labeled pixels (from 2014 to 2017 NAIP airborne Imagery) to capture diverse LCLU types and spatial patterns. We assessed the credibility of FLUTE by producing a 1-meter resolution database named UrbanWatch for 22 major cities across the conterminous United States. UrbanWatch contains nine LCLU classes – building, road, parking lot, tree canopy, grass/shrub, water, agriculture, barren, and others, with an overall accuracy of 91.52%. We have further made UrbanWatch freely accessible to support urban-related research, urban planning and management, and community outreach efforts: https://urbanwatch.charlotte.edu.

1. Introduction

The past two decades have witnessed tremendous advancements in very-high-resolution (VHR, typically finer than 5 m) Earth observation data acquisition. VHR data that were traditionally collected by airplanes are now also available through satellite sensors or drones (Aasen et al., 2018; Blaschke et al., 2014; Chen et al., 2012). Such a myriad of platforms has fundamentally improved our ability to inform sustainable urban development and smart city practices, which entail an understanding of fine-scale interactions between humans and the heterogeneous landscape (Angelidou et al., 2017; Zhang and Li, 2018). Land cover and land use (LCLU) of geographical entities and patterns identified from VHR data are becoming increasingly ubiquitous for various topics of urban studies, such as building energy consumption (Faroughi et al., 2020; Li et al., 2017), urban development and sprawl (e.g., Blaschke et al., 2014; Huang et al., 2017; Li, et al., 2020; Zhang and Li, 2018), biodiversity conservation (e.g., Chen et al., 2020; Dutta et al., 2020; Godwin et al., 2015; Turner et al., 2003), urban agriculture and gardening (Mathieu et al., 2007; Saha and Eckelman, 2017), social and environmental justice (e.g., Tapiador et al., 2011; Weigand et al., 2019), human thermal comfort and health (e.g., Myint et al., 2015b; Troyo et al., 2008; Whiteman et al., 2019), and humanitarian crisis or disaster response such as the impact of COVID-19 (e.g., Contreras et al., 2016; Giada et al., 2003; Venter et al., 2020).

At the high resolution, traditional pixel-based approaches resulted in significant confusion of spectral signatures among impervious surfaces,

* Corresponding author.
E-mail address: gang.chen@uncc.edu (G. Chen).

https://doi.org/10.1016/j.rse.2022.113106
Received 13 December 2021; Received in revised form 29 April 2022; Accepted 26 May 2022
Available online 2 June 2022

water bodies, and unmanaged soil, leading to unreliable urban LCLU classifications (Myint et al., 2011; Roy et al., 2018; Han et al., 2020). Since the early 2000s, the demands for high-resolution LCLU has driven a paradigm shift from pixel-based to Geographic Object-Based Image Analysis (GEOBIA or OBIA) (Blaschke et al., 2014; Chen et al., 2018; Hay and Castilla, 2008; Kucharczyk et al., 2020). GEOBIA utilizes clusters of similar neighboring pixels as the basic study units (i.e., image-objects) in classification. It takes advantage of state-of-the-art image segmentation to reduce spectral variation in VHR imagery, and it defines unique object-based features to capture contextual information among geographic entities, and GEOBIA semantics to customize classification rule sets (Chen et al., 2018). While the user-driven selection of algorithms, image features, and parameters within a typical GEOBIA framework offer flexibility for practitioners to tweak models and improve classification accuracies, effective human intervention requires knowledge and experience which is particularly challenging for large-area mapping projects. To bypass the intermediate tasks (e.g., defining the optimal scale and selecting the best object-based features) and to improve automation, deep learning and its end-to-end learning structure have recently been introduced to fine-scale LCLU mapping with encouraging performance (Ma et al., 2019; Zhang et al., 2016). Deep learning treats data as a nested hierarchy of concepts and has the ability to capture high-level features in VHR imagery over heterogeneous urban regions similar to how humans interpret imagery (Krizhevsky et al., 2012; Zhao and Du, 2016). While a variety of deep-learning-based algorithms have been developed for the purpose of LCLU mapping (Yuan et al., 2020), their level of success and large-area generalization ability depend on the availability of large high-quality training datasets, high-performance computing, and sophisticated software libraries.

Numerous case studies at the local or regional scale have advanced the science of VHR mapping; however, only a handful of high quality, credible databases are available for identifying fine-scale geographic entities over large areas, i.e., wall-to-wall LCLU maps over multiple cities, states or provinces at a resolution finer than 5 m (see examples in Table 1). The majority of the databases capitalize on two decades of evolution in GEOBIA and have reported accuracies of over 82%, while capturing 5–15 LCLU classes. Although promising, the performance of typical GEOBIA depends on handcrafted (a.k.a., human-designed) image features and classification rule sets, which often requires a priori knowledge of the study area (Cheng et al., 2017). Particularly, the heterogeneous urban landscape reveals more spatial details and higher spectral variation in VHR imagery than in its coarse-to-medium resolution counterparts (Zhang et al., 2020). Optimal parameters, features, or rulesets selected for one city usually require a re-evaluation for another city. Some recent efforts of developing (semi-)automatic parameter or feature optimization approaches or employing cloud computing infrastructure have shown potential for enhancing the classic GEOBIA framework (e.g., Antunes et al., 2019; Torres-Sánchez et al., 2015). We also note that some recent GEOBIA studies focus on capturing urban function using a customized minimum mapping unit (MMU) for

Table 1
List of representative large-area, very-high-resolution (VHR finer than 5 m) land cover and land use databases.

Database name	Spatial resolution	Spatial coverage	Land cover/use classes	Primary data source	Classification Framework	Reported accuracy	Developer, link & data accessibility
Vermont High-Resolution Land Cover	0.5 m	U.S. State of Vermont	8 (tree canopy, grass/shrubs, bare soil, water, buildings, roads, other paved, railroads)	NAIP*, LiDAR	GEOBIA	>90%	Vermont Center for Geographic Information (open access)
C-CAP high resolution land cover	1 m	U.S. coastal regions (4+ states and multiple islands and counties)	16–25 (Impervious surfaces as one single class)	NAIP*	GEOBIA	82–95%	NOAA Office for Coastal Management (open access)
EarthDefine Land Cover	0.6 m	25+ U.S. states (partial or full coverage)	7 (impervious, herbaceous, bare, water, trees, shrubs, Trees over Impervious)	Aerial imagery, LiDAR	Not clearly stated	>95%	EarthDefine LLC (data for sale, few sample scenes available)
Chesapeake Bay Watershed Land Cover	1 m	U.S. Chesapeake Bay watershed (partial coverage of 6 states)	13 (water, forests, roads, non-road impervious, tree canopy over, impervious, etc.)	NAIP*, LiDAR	GEOBIA	82–93%	The Conservation Innovation Center (CIC) (open access)
EnviroAtlas Meter-Scale Land Cover (MULC)	1 m	30 U.S. communities	8 (impervious, tree, shrub, water, grass/herbaceous, soil/barren, agriculture, wetlands)	NAIP*, LiDAR	Pixel-based (some areas) & GEOBIA (the other areas)	88%	United States Environmental Protection Agency (open access; Pilant et al., 2020)
Hi-ULCM	2.1 m	42 China major cities	7 (buildings, grass/shrubs, trees, bare soil, water, roads, other impervious)	Ziyuan-3, vectors from multi-sources	GEOBIA	88.55%	Wuhan University (not openly accessible; Huang et al., 2020)
PKU Urban scape Essential Dataset (PKU-USED)	2.4 m	81 China major cities	12 functional-zones (commercial, residential-1/2/3, institutional, industrial, green, water, transport, woodland, and undeveloped)	Ziyuan-3, GF-6, Google Earth imagery and ArcGIS images	GEOBIA	85.9% (in Beijing)	DoLab, Peking University (10 cities accessible at request; Du et al., 2020)
Urban Atlas	2–4 m	31 countries or country groups in or near Europe	17 urban classes [urban Fabric, green urban area, pastures, forests, water bodies, fast transit roads and associated land, etc.]	SPOT 5&6, Formosat-2, etc.	GEOBIA	>59%***	European Environment Agency (EEA) (open access)
Microsoft high-resolution land cover	1 m	The contiguous U.S.	6 [water, tree canopy/forest, low vegetation/field, barren land, impervious (other), impervious (road)]	NAIP*, Landsat	Pixel-based, deep learning	90–93% (two regions), 87.03% (consistency with NLCD**)	Microsoft (available in 6 states; Robinson et al., 2019)

* NAIP – National Agriculture Imagery Program.
** NLCD – National Land Cover Database.
*** Accuracy tends to be updated using new validation data.

each LCLU, such as 0.25 ha, which is larger than the size of a pixel (e.g., Du et al., 2020; Topaloğlu et al., 2021).

As an alternative, deep learning that has strong model generalization ability can significantly reduce human efforts during model tuning when it comes to large-area applications. For example, in a national scale mapping project over the contiguous U.S., Robinson et al. (2019) compared multiple deep-learning-based models (trained and validated regionally in the northeastern U.S.) and reported the best accuracy of 93% (Table 1). While encouraging, deep learning requires a massive amount of high-quality training data that often restricts its application to local geographical regions. Depending on the size of the study area (e.g., major town, city, metropolis, county, or state), a high-quality (e.g., cloud free) VHR data acquisition campaign could take weeks or years to complete. Additionally, the system or date (time and season) discrepancies in data acquisition could cause inconsistent image quality or scene structure, e.g., disparity in spectral signatures or urban contextual information due to the difference in view angle, and buildings leaning away from the principal point (a.k.a., relief displacement) at varying levels (Jabari et al., 2019). Additionally, geographic objects in urban regions are of multiple scales and are covered by different materials even for the same LCLU class, which further cause high intraclass variance (Safari et al., 2020), e.g., single-family homes versus multi-family properties, and asphalt pavements versus concrete highways. Furthermore, impervious surfaces have been oversimplified in most of large-area LCLU databases (Table 1) by being treated as one single class; however, roads, buildings, and parking lots represent different characteristics of human activities and 3D urban forms. Differentiating between classes that have relatively low interclass variation is essential for informed urban studies and management. Thus far, neither deep learning nor GEOBIA has adequately addressed these large-area mapping challenges. Additionally, most attempts have focused at local scales.

Based on the above considerations, (i) we present a robust *Fine-resolution, Large-area Urban Thematic information Extraction* (FLUTE) framework that capitalizes on state-of-the-art semi-supervised learning and deep learning architectures to leverage numerous high-resolution observations of LCLU. While deep learning is used here, the emphasis of this project is to address multiple challenges in VHR urban LCLU mapping over large geographical areas. (ii) We apply the FLUTE framework to produce a 1-m resolution, high-accuracy LCLU database (subsequently referred to as *UrbanWatch*) for 22 major cities across the conterminous U.S. capturing nine classes – building, road, parking lot, tree canopy, grass/shrub, water, agriculture, barren, and others. (iii) We also develop an online data repository to share the database at no cost to users for supporting urban studies, management and outreach. To facilitate model training, we have constructed a benchmark database that contains over 52.43 million labeled points at the 1-m resolution covering diverse urban LCLU classes and spatial patterns. Here we report the structure of the proposed LCLU mapping framework and describe the generated databases. We also discuss the performance of our framework through an internal assessment among LCLU classes and across regions, as well as an external comparison with several medium- and high-resolution LCLU products.

2. Study areas and data

2.1. Twenty-two U.S. cities

We selected 22 major cities across the conterminous U.S. for the project, including Atlanta, GA; Boston, MA; Charlotte, NC; Chicago, IL; Denver, CO; Dallas, TX; Detroit, MI; Houston, TX; Los Angeles, CA; Miami, FL; Minneapolis, MN; New York City, NY; Philadelphia, PA; Phoenix, AZ; Raleigh, NC; Riverside, CA; San Diego, CA; San Francisco, CA; Seattle, WA; Tampa, FL; St. Louis, MO; and Washington D.C. (Fig. 1). These populous cities are homes to one-tenth of U.S. inhabitants, and they represent diverse built environments and urban spatial patterns in four geographic regions of the country – West, Midwest, South, and Northeast. City boundaries were derived from the U.S. Census Bureau's MAF/TIGER geographic database (Census, 2020). For some megacities (e.g., San Francisco), the rapid development of suburban areas blurs the administrative boundary with their surrounding municipalities. In this case, we have slightly expanded the study area beyond the boundary with the intent to capture heterogeneously distributed geographic entities along the urban-rural gradient.

2.2. NAIP imagery

We acquired 1-m resolution NAIP (National Agriculture Imagery Program) images covering the selected cities from the USGS Earth Explorer data portal (USGS, 2020). Raw NAIP images were taken during the leaf-on seasons by airborne sensors and had four spectral bands, i.e., blue (400–580 nm), green (500–650 nm), red (590–675 nm), and near-

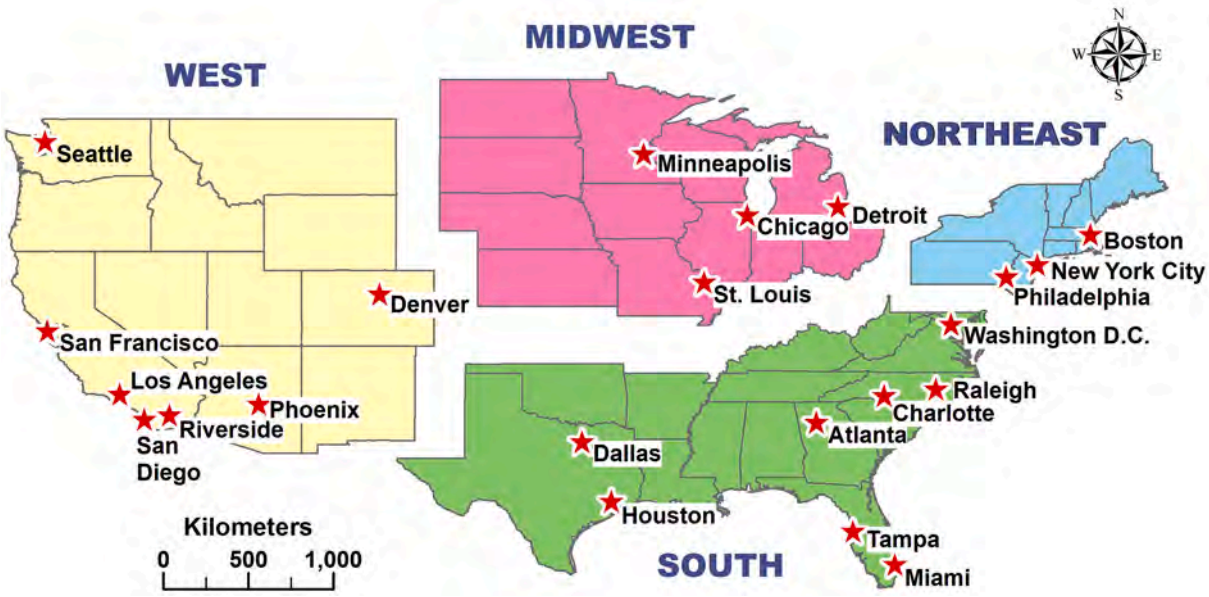

Fig. 1. Twenty-two cities in four geographical zones across the conterminous United States were used to develop the UrbanWatch database.

infrared (675–850) nm. Because NAIP data acquisition is a multi-year program for national coverage, the images for this study were taken during the 2014–17 window. All data have been pre-processed with data quality inspected by the vendor before they were made available to the public. While the specified horizontal positional accuracy is always finer than 6 m, it is typically within +/− 2 m as compared to VHR imagery from Google Maps© (Google LLC, CA), Bing Maps© (Microsoft Corporation, WA), and ArcGIS© base maps (Esri, CA) (Pilant et al., 2020).

3. Methodology

3.1. Classification scheme

Major land cover types in urban areas include impervious surfaces, vegetation, water, and barren. Imagery at very high resolutions offers the possibility of identifying within the four major categories not only land cover but also *land use* through detailed spatial, textural, shape, and contextual information of geographic entities (Chen et al., 2018). In this project, we have defined nine LCLU classes in U.S. urban areas, i.e., building, road, parking lot, tree canopy, grass/shrub, water, agriculture, barren, and others (Fig. 2). Our class scheme builds upon the classic USGS Level 1 classification system (Anderson et al., 1976); however, it has more detailed classes within impervious surface. Given the nature of urban mapping, we divided impervious surface into three higher level classes – building, road, and parking lot – with the intent to facilitate various urban studies that require a nuanced understanding of urban form and function. For instance, building energy consumption (Li et al., 2017), human health due to traffic (Zhang and Batterman, 2013), and

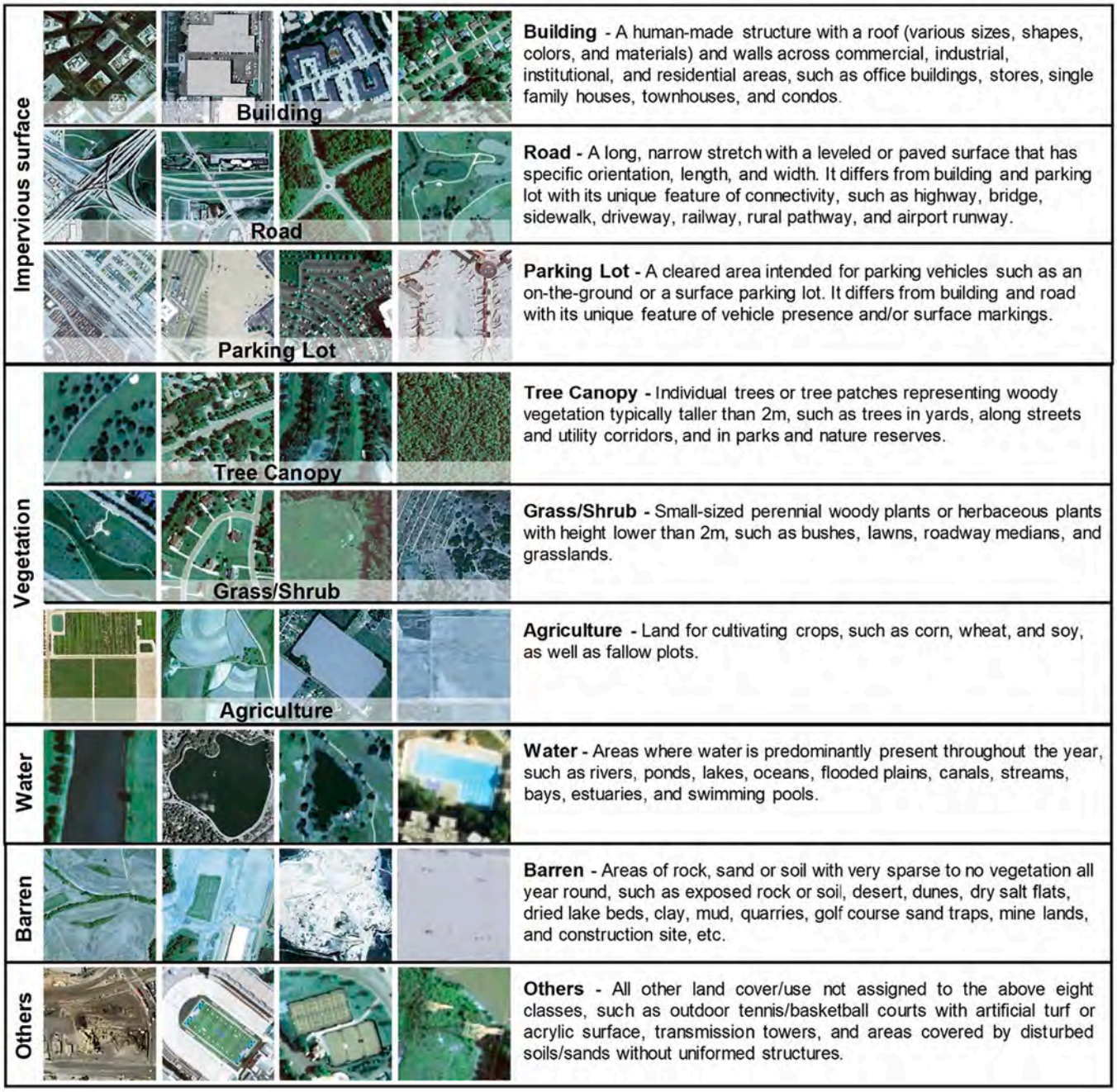

Fig. 2. A nine-class urban classification scheme with diverse geographic patches (in aerial photos) dominated by building, road, parking lot, tree canopy, grass/shrub, agriculture, water, barren, and others.

recharge scheduling of electric vehicles in parking lots (Aghajani and Kalantar, 2017). We further divided vegetation into tree canopy and grass/shrub because these two classes have distinctive 3D structures and offer different ecosystem services (Livesley et al., 2016). We did not separate grass from shrub due to high spectral and spatial similarities, which has posed a similar challenge for other VHR mapping projects (Table 1). While LiDAR offers a promising solution (Pilant et al., 2020), data availability is a concern. In this project, our objective is to rely on one single type of VHR data to develop an operational framework that is more feasible for large-area mapping where full-cover LiDAR or other data layers are not available. Fig. 2 shows diverse urban geographic patches in NAIP true-color imagery dominated by the nine LCLU classes, and our definition of those classes. Compared to most existing VHR classification schemes (Table 1), our scheme has similar or more detailed categories, e.g., for describing impervious surfaces (building, road, and parking lot). As the existing databases have been used in a variety of fields for fine-scale urban studies (see examples and the literature in the first paragraph of Section 1), our scheme of similar or expanded classes has the potential to benefit a broad array of urban topics. We are also aware of several LCLU databases containing more classes (e.g., Urban Atlas, PKU-Used). However, their emphasis is to capture how urban areas function, where the minimum mapping unit (MMU) has one or more LCLU classes that are used in our study (e.g., commercial, residential).

3.2. Benchmark database for framework training

Supervised or semi-supervised learning requires a collection of large volumes of labeled training data representing targets. This is particularly true for deep-learning-based frameworks like the one proposed in our project. However, a challenge for such large-scale LCLU mapping involves high variation in urban design and spatial patterns, while also trying to collect quality training samples, a process that is time consuming and labor intensive. To deal with these challenges, we constructed a labeled urban benchmark database following three criteria: (i) a large image patch size, (ii) diverse patterns of urban neighborhoods with simple individual geographic entities, and (iii) data augmentation.

(i) We defined the size of each patch as 512 × 512 pixels (262,144 pixels) instead of typical small patches and labeled every 1-m pixel, which allowed us to cover multiple geographic objects of varying scales (e.g., streets versus highways) and their complex spatial arrangements in one image scene. This also provided opportunities to extract urban contextual information from different perception fields (see framework description in Section 3.3 for details). The labeled pixels represented the similar LCLU class proportion as that in the studied cities: building: 13.58%, parking lot: 6.67%, road: 17.11%, barren: 1.99%, water: 5.17%; other: 5.95%; agriculture: 5.95%, grass/shrub: 25.77%, and tree canopy: 23.05%. We slightly increased the proportion of some classes (e.g., agriculture, barren) due to their higher spectral and spatial variation than the other ones.

(ii) We manually selected sample patches to represent diverse patterns of neighborhoods along the urban-rural gradient. However, we selected neighborhoods with relatively simple geographic entities that are easy to recognize and label via visual interpretation. Our emphasis on diversity was placed on the spatial interrelations between those objects. Inspired by Gestalt theories of perception (Sternberg, 1980), we applied the strategy with the intent to allow neural networks to learn simple-to-complex landscape patterns and then transfer such knowledge of scene structure to benefit broader-area urban mapping. We note that cities spread over large geographical regions for this project, and some of the cities exhibited unique urban patterns or surface materials. To ensure consistent model performance, we added additional sample patches during model training to help our framework learn new scene structures that were missed during the initial sample selection.

(iii) We applied two typical image augmentation methods to rotate and flip each sample patch (Gidaris et al., 2018). For the 90-degree rotation, we transposed the image and performed an upside-down flip. For the 180-degree rotation, we flipped the image vertically and then horizontally. For the 270-degree rotation, we flipped the image vertically and then transposed the image. Since the sensor view angle effect could dramatically change scene structure in VHR imagery, the two augmentation methods added extra semantic information to reduce cognitive biases between the single-view computer vision and the multi-view remote sensing vision. Here, we did not intend to use these two methods to fully resolve the issues caused by the view angle effect. However, they offer an established function to augment training samples with proven effectiveness in other VHR mapping efforts (e.g., Li et al., 2018; Yu et al., 2017). We have developed a sub-module to tactfully deal with the view angle effect in the proposed framework (see details in Section 3.3.3).

Based on these benchmark criteria, we initially selected 50 sample patches (512 × 512 pixels each) to cover varying types of LCLU along the urban-rural gradient. We then followed a similar approach of integrating transfer learning and manual correction to label all pixels as described in one of our previous efforts (Zhang et al., 2020), where the main idea was to iteratively apply deep learning (ResNet-50; Lu et al., 2018) and manual correction to add and refine all sample patches incrementally. Using the augmentation method, our urban benchmark database was created to contain 200 sample patches (50 × 4) with a total number of over 52.43 million labeled pixels to represent the nine LCLU classes as described in the previous section. Sample patches representing urban, suburban, and rural landscapes are shown in Fig. 3. The benchmark and the validation samples (see Section 3.4.2) were collected and analyzed by a team of three interpreters over a period of six months. To maintain consistency among interpreters, we conducted periodic team meetings to ensure synchronized interpretation of ground objects in the VHR imagery. All interpreters relied on the same reference data sources, including Google Earth sub-meter images and street-view photos.

3.3. Fine-resolution, Large-area Urban Thematic information Extraction (FLUTE) framework

3.3.1. FLUTE overview

The FLUTE framework builds upon semi-supervised learning (Zhu, 2005) to deal with one of the biggest challenges in large-area LCLU mapping, where time consuming and tedious construction of a labeled database is almost never adequate to include all the variability of urban LCLU and their spatial patterns. Here, our semi-supervised learning capitalizes on both small, carefully labeled (Section 3.2) and large, unlabeled data to develop a robust framework and produce satisfactory LCLU results. To effectively exploit the unlabeled data, we have utilized a popular strategy of adding noise (i.e., random perturbations) to the input data, which mimics the way humans interpret objects, i.e., generating consistent classification results even if the input is slightly changed (Tarvainen and Valpola, 2017).

We have further employed the self-ensembling strategy that has recently drawn strong interest with proven efficiency and accuracy (Yu et al., 2019). It differs from classic ensemble predictions (e.g., using multiple neural networks) by operating on one single network (e.g., Laine and Aila, 2017; Rasmus et al., 2015). To do so, our framework includes two sub-networks – a student and a teacher model (Fig. 4). The two models use the same baseline neural network and are identical in network structure. However, they are calibrated with different training data and learn from each other to progressively refine model weights and the accuracy of LCLU mapping over training steps. The development of the baseline neural network capitalized on a state-of-the-art deep

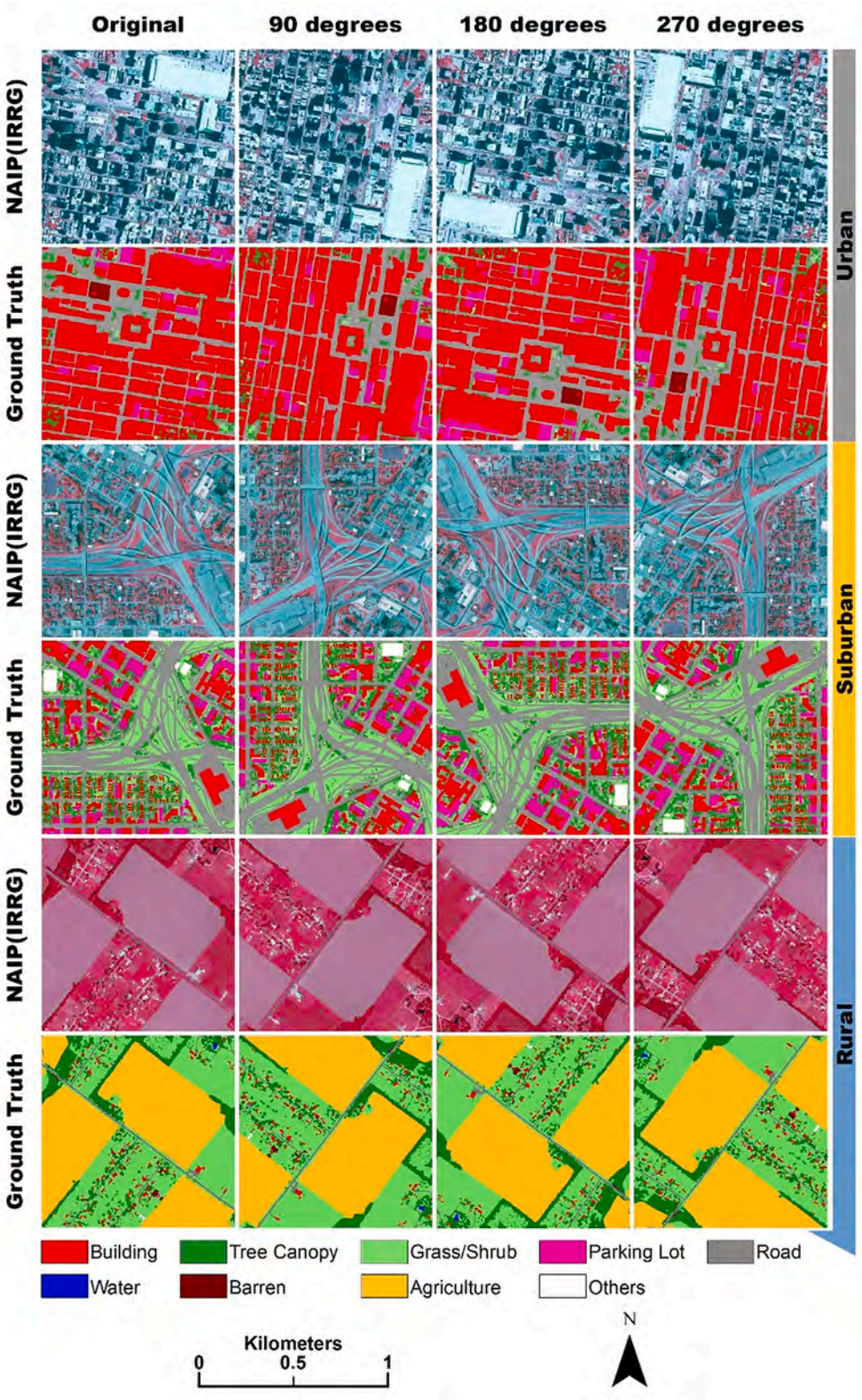

Fig. 3. Three selected sample patches and augmented results (rotated and flipped) represent urban center, suburban and rural landscapes. For each type of landscape, the top images are from NAIP aerial photos (IR-R-G false-color composites), and the bottom ones are from the corresponding labeled database.

learning model; however, we have modified the model by introducing two new submodules to improve FLUTE's ability to extract object features: (i) Scale-aware Parsing Module (SPM) and (ii) View-aware Embedding Module (VEM). The purpose of SPM is to effectively capture urban objects of varying scales, while VEM can help mitigate the multi-view effect on scene structure interpretation. The FLUTE framework details and the baseline neural network with SPM and VEM are described in the two succeeding sections, respectively.

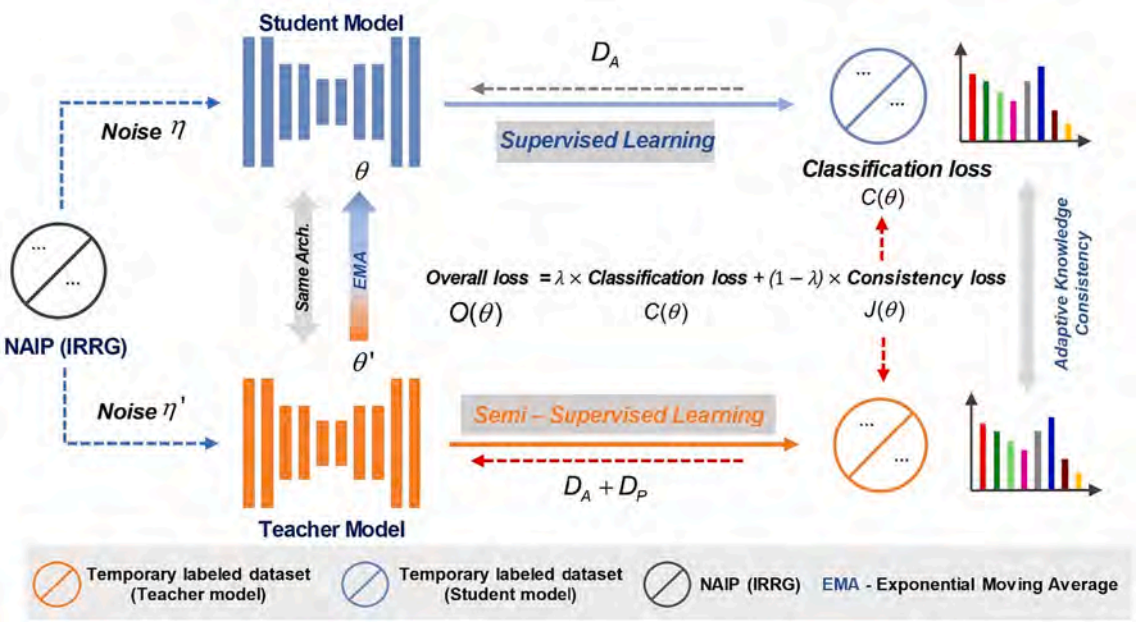

Fig. 4. The pipeline of the FLUTE framework for semi-supervised learning.

3.3.2. FLUTE framework details

We denote the benchmark (labeled) dataset as$D_A = \{(X_i, y_i)\}_{i=1}^{N}$, where X_i represents features extracted from the input imagery for the ith sample pixel, and y_i represents each of the nine LCLU classes for the ith pixel. $N$is the number of labeled pixels. The student model is trained with D_A and generates labels for all the initially unlabeled pixels with the resulting dataset denoted as$D_p = \{(X_k, y_k)\}_{k=1}^{M}$, where M is the number of unlabeled pixels, and X_k,y_k represent the input features and the corresponding label for the kth pixel, respectively. This is followed by training the teacher model with the combined datasets$D_A \cup D_P$. The dataset D_P is then updated with the output of the teacher model.

The optimization of this type of semi-supervised learning requires minimizing the output difference between the teacher and the student model (Goldberger et al., 2003). As shown in Fig. 4, two loss functions play an essential role during backpropagation, including the classification loss function$C(\theta)$ and the consistency loss function $J(\theta)$. Here, $C(\theta)$ is calculated as the standard cross entropy loss between labels predicted by the student model and the labeled input D_A (Laine and Aila, 2016). $J(\theta)$ is the mean squared difference between the predicted outputs of the student and the teacher model, which is defined as follows:

$$J(\theta) = \mathbb{E}_{x,\eta,\eta'}\left[\|f(x,\theta,\eta) - f(x,\theta',\eta')\|^2\right] \quad (1)$$

where weights θ and noiseη are used by the student neural network $f(\cdot)$, and weights θ' and noise η' are for the teacher neural network $f(\cdot)$. Adding noise to the model input has been used in semi-supervised learning for reducing the possibility of overfitting, so the model is not biased towards particular targets (Srivastava et al., 2014). This is especially important for our mapping purpose, because VHR imagery that have been acquired over large areas are from different view angles and often contain pixel-level location errors. Here, random perturbations are purposely added as three types of noise, including random translations and horizontal flips, and Gaussian noise to simulate the variation in the input, and dropout applied to the model structure (Damianou and Lawrence, 2013). The weights θ of the student model are updated in backpropagation by minimizing the overall loss $O(\theta)$, which is an aggregation of $C(\theta)$ and$J(\theta)$:

$$O(\theta) = \lambda C(\theta) + (1-\lambda)J(\theta) \quad (2)$$

Where λ is a ramp-up weighting coefficient that controls the trade-off between the supervised and the unsupervised loss. In any of the following training steps (e.g., step t), the teacher model weights θ_t' are updated by considering the student model's current weights $\theta_t$and the teacher model's previous weightsθ_{t-1}', i.e., averaging model weights over training steps (Tarvainen and Valpola, 2017).

$$\theta_t' = \alpha \cdot \theta_{t-1}' + (1-\alpha)\theta_t \quad (3)$$

During training, the exponential moving average (EMA) weights (Haynes et al., 2012) of the student model are assigned to the teacher model at every step, and the proportion of the weights assigned is controlled by a weighting coefficient α. Here, α is set to 0.97 which allows the student model to assign the optimal proportion of weights to the teacher model at each step, ensuring that the predictions of the two models converge quickly and achieve a high accuracy. With its updated weights, the teacher model is able to update the estimation D_P for the initially unlabeled pixels. By minimizing $J(\theta)$ and $O(\theta)$, the framework again updates the weights for the student model, which is used to further improve the teacher model (Eq. (3)). Through such an iterative process, the teacher and the student models are progressively refined. At the end of the training, the LCLU maps generated from the teacher model are used as the end product.

3.3.3. Scale-aware Parsing Module (SPM) and View-aware Embedding Module (VEM)

The student model and the teacher model have the identical baseline neural network, which builds upon the SegNet (Badrinarayanan et al., 2015) architecture due to its robust performance in semantic segmentation (e.g., Audebert et al., 2018; Jiang et al., 2020; Panboonyuen et al., 2017). The weights of the encoder as the feature extractor were optimized by Inception-v4 (Szegedy et al., 2017) to balance efficiency and performance in VHR mapping based on our previous experience (Zhang et al., 2020). We have further added two submodules – SPM and VEM – for feature extraction and fusion (Fig. 5).

The purpose of SPM is to obtain the optimal LCLU feature representation across scales. Because geographic objects reveal different characteristics at different spatial resolutions, it is challenging to capture sufficient object information from an observation field at a single scale. While a small observation field may miss sufficient context to accurately train a deep network, a large observation field often introduces extra uncertainties, e.g., rugged edges and class errors caused by interclass

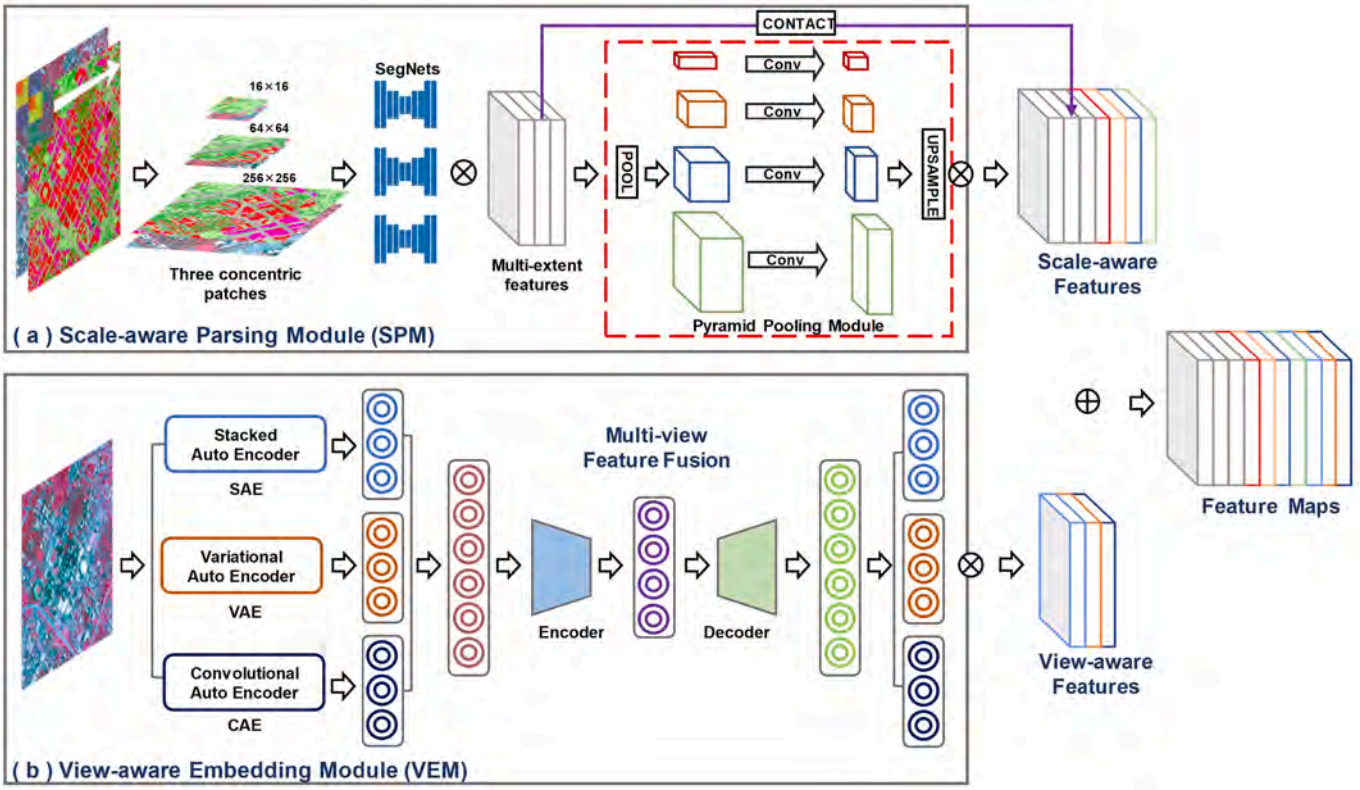

Fig. 5. The anatomical structure of the baseline network: (a) Scale-aware Parsing Module (SPM) is for generating multi-extent features, while (b) View-aware Embedding Module (VEM) is for generating multi-view features.

spectral confusion with increasing intraclass variance (Deepan and Sudha, 2019; Luo et al., 2019; Zhao et al., 2015). Here, SPM extracts multi-scale features, i.e., three concentric patches of 16 × 16, 64 × 64, and 256 × 256 pixels. For each selected patch, its center patch (2 × 2 pixels) is considered as a reference object. We have tested multiple sizes, and these three proved the most effective to understand scene information at varying perception fields across urban LCLU patterns while being permitted by a typical GPU cache. To integrate these features, SPM applies a pyramid pool module (Zhao et al., 2017) to harvest representations from the three perception fields, followed by resampling and concatenating layers to form the final feature representation.

The purpose of VEM is to deal with the view angle effect that has posed a significant challenge for understanding high-resolution image scenes (Fu et al., 2019). A fine-tuned deep learning model trained with features from one view angle is likely to generate an unreliable estimation of LCLU using imagery from a different view angle (Azulay and Weiss, 2018). While we have used data augmentation to generate multi-view training data (Section 3.2), the number of view angles remains limited. To improve the network's generalization ability, VEM contains three unsupervised deep neural networks (DNNs) as multi-view branches that are fed with the same image input. These include (i) stacked autoencoder (SAE; Vincent et al., 2010), (ii) variational autoencoder (VAE; Kulkarni et al., 2015), and (iii) convolutional autoencoder (CAE; Guo et al., 2017). Our method builds on the idea by (Lin et al., 2018) that DNN-based multi-view methods beat the traditional techniques by learning complex nonlinear transformations to obtain powerful multi-view features and exploit effective relationships (i.e., canonical correlation) among multiple views (Andrew et al., 2013; Wang et al., 2015). To integrate the three unique view features provided by DNNs, we designed a joint learning strategy to realize multi-view fusion as inspired by Lin et al. (2018). VEM directly concatenates all the representations and feeds them into a fusion encoder (Zhang et al., 2019). This encoder consists of two fully connected layers and compresses multiple features into a single dense representation. The first fully connected layer realizes the multi-view fusion scheme by imposing an implicit multi-view constraint on a multi-view soft assignment distribution. The second layer imposes an explicit multi-view constraint on a view-specific auxiliary target distribution. Through the two multi-view fusion schemes, multi-view complementary information can be effectively explored in both models during the joint learning process. By doing this, insignificant representations are ignored and VEM captures the latent correlations across views.

3.4. UrbanWatch database

3.4.1. Framework implementation and database development

The FLUTE framework was implemented using the PyTorch library. At the training stage, we trained the model for 1000 epochs using the Adam optimizer with a batch size of 4. The initial learning rate was set to 0.001, and we multiplied it by 0.7 every 5000 steps to reduce the learning rate. These settings were derived from our experimental evaluations based on the default configuration parameters that performed well on problems with sparse gradients (Li et al., 2021; Mehta et al., 2019). To avoid overfitting, we applied dropout to the fully connected layers with a dropout rate of 0.5 except the last layer (Srivastava et al., 2014). The whole training process took approximately 10 h on two GeForce GTX 1080ti GPUs. Directly training an unbalanced dataset often leads to a biased classification result. To address this issue, we adopted the weighted cross-entropy loss function (Panchapagesan et al., 2016) to force FLUTE to focus more on the LCLU classes with fewer samples. We ran the teacher model for fake news detection, and the teacher model started to perform better than the student model after 20 epochs. However, we note that the convergence of the teacher model depends on epoch, batch size, training data size, and the parameter α. We used KL-divergence (Goldberger et al., 2003) as the consistency cost function to tune the model.

At the inference stage, we used sequential patch-wise classification. Each NAIP image was automatically partitioned into non-overlapped patches with an identical size of 512 × 512 pixels, which was chosen to balance framework performance (obtaining consistent receptive fields with those from the benchmark dataset) and processing efficiency. The partitioned patches were sequentially imported to a convolution layer and a SoftMax layer to generate LCLU maps, which were agglomerated to produce the final mapping results for individual cities.

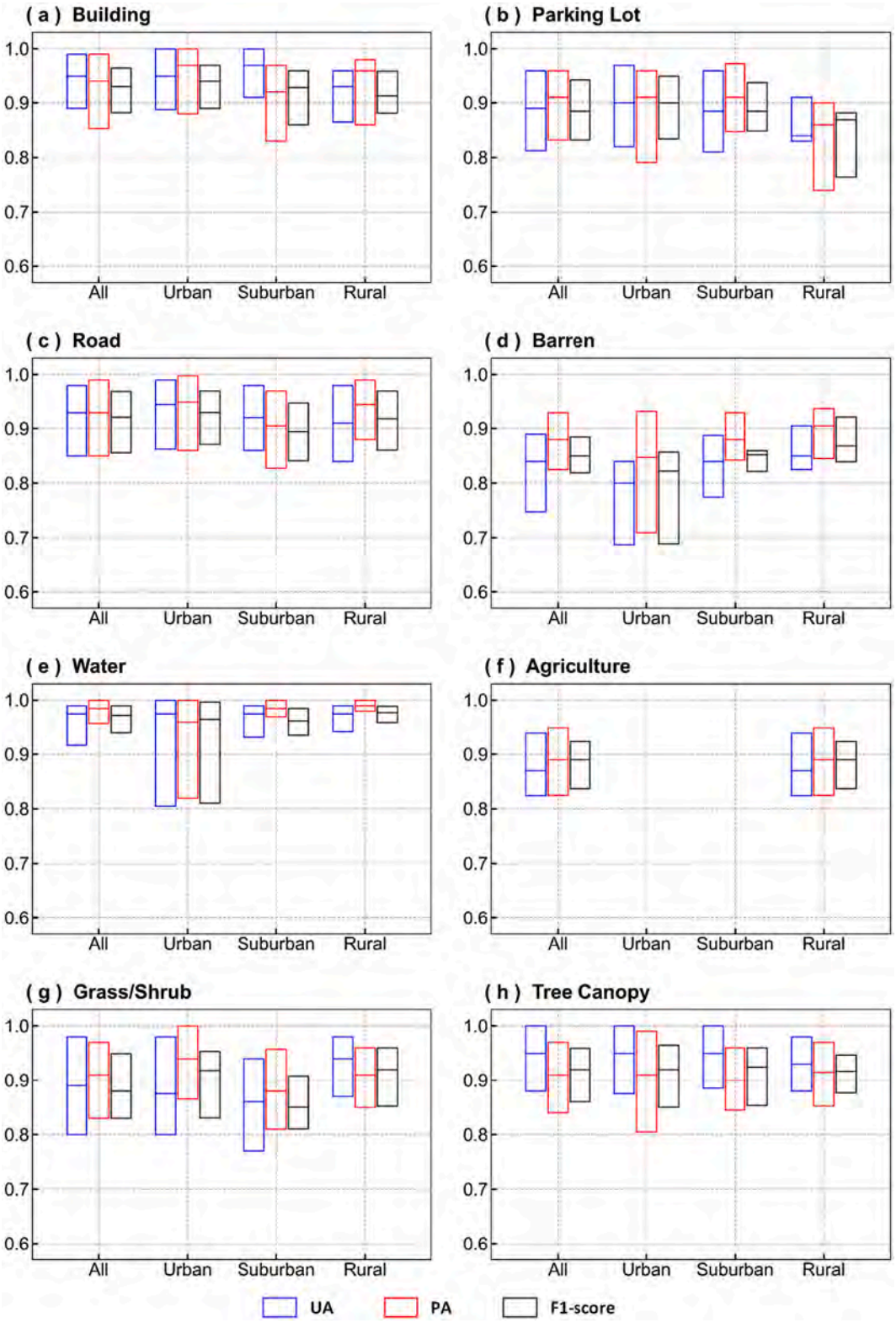

Fig. 6. FLUTE performance for eight LCLU classes - (a) building, (b) parking lot, (c) road, (d) barren, (e) water, (f) agriculture, (g) grass/shrub, (h) tree canopy - in all 22 cities combined and in three main regions (i.e., urban, suburban, and (rural) along the urban-rural gradient. Boxplots show the interquartile range and median values of UA (user' accuracy), PA (producer's accuracy), and F1-score at the city level for each class.

3.4.2. Accuracy assessment

We used a random sampling protocol, i.e., one-stage cluster sampling (Gallego, 2012), to collect 1173 sample patches (10 × 10 pixels for each patch) across the 22 studied cities as validation data from NAIP imagery, and labeled all 117,300 pixels through manual interpretation. Specifically, each city was partitioned into a fixed set of 10 × 10 pixel patches, which were randomly selected. The number of patches depends on city size, i.e., the ratio between the total size of sample patches for a city and city size was kept at 1/100,000. While it is challenging to determine the exact number of validation samples for any study, our 117,300 sample pixels were able to capture diverse urban LCLU and spatial patterns. They represented the similar LCLU class proportion as that in the studied cities. We compared the UrbanWatch database with the validation data using five popular metrics – overall accuracy (OA), user's accuracy (UA), producer's accuracy (PA) (Story and Congalton, 1986), non-site-specific accuracy (NA) (Stehman and Foody, 2019), and F1-score (Goutte and Gaussier, 2005). We have further investigated the agreement in LCLU between UrbanWatch and two groups of databases, including three VHR databases: Microsoft high-resolution land cover (Robinson et al., 2019), Chesapeake Bay Watershed land cover (Chesapeake Conservancy, 2020), and EarthDefine land cover (EarthDefine, 2020) and two medium-resolution databases: Esri 2020 Land Cover (Karra et al., 2021), and National Land Cover Dataset (NLCD; Dewitz, 2019) over the same urban areas.

3.5. Open-access data repository

We have adopted an open standard for data use and distribution, where users have free access to the UrbanWatch database at https://urbanwatch.charlotte.edu. The 1-m maps can be freely used for non-commercial purposes. They have a total size of 211 GB and are categorized for individual cities. To facilitate data downloading, we have further divided each city into multiple grids (approximately 6000 × 7000 m each) that allow users to select the desired grids from a preview image of a city. All maps are in GeoTIFF format and have embedded georeferencing information. The color of each pixel in the maps corresponds to a specific LCLU class (as illustrated in Fig. 3), i.e., red = building, gray = road, purple = parking lot, dark green = tree canopy, light green = grass/shrub, blue = water, yellow = agriculture, dark red = barren, and white = others.

4. Results

4.1. Performance of the FLUTE framework

The FLUTE framework reports an OA of 91.52% based on the validation samples. We calculated OA, F1-score, UA, PA, and NA for nine LCLU classes in each of the 22 mapped cities (see Appendix A). We further divided each city into three regions along the urban-rural gradient to represent three primary urban development intensities: urban center (hereafter urban), suburban, and rural, using percent built-up (PBU) following Chen et al. (2020): rural (PBU ≤ 15%), suburban (15% < PBU ≤ 40%), and urban (PBU > 40%) density. We found consistent performance in the three regions (Fig. 6): urban (OA: 91.51%), suburban (OA: 91.38%), and rural (OA: 91.71%). Please note that almost all agriculture fields existed in the rural areas, and its accuracy was not reported in the urban and suburban regions. We computed standard errors for OA, UA and PU following the formulas provided by Stehman (1997). UA and PU were calculated for individual LCLU classes. Appendix B shows selected results for four cities from four major geographic regions, respectively, including Dallas from the South, New York from the Northeast, Chicago from the Midwest, and Los Angeles from the West. All standard errors are consistently low across regions and across LCLU classes.

For all cities combined and three regions along the urban-rural gradient, we calculated the F1-score for each LCLU class (Fig. 6). The three impervious surface classes – building, road, and parking lot – achieved higher accuracies than several other classes, such as grass/shrub, barren, and others. However, buildings and roads were more accurately identified than parking lots, which sometimes were misclassified as roads due to high spectral and contextual similarities between the two. Although trees tend to have higher structural complexity than low stature grasses or shrubs, they were mapped with consistently higher accuracy along the urban-rural gradient. This was particularly true in suburban areas where lawns and shrubs are prevalent in residential communities. Misclassification often occurred in areas where nearby trees shaded low vegetation. Water bodies were easier to extract in suburban and rural areas with the highest accuracy among all LCLU classes. However, urban centers are occupied by tall buildings, and their self and cast shadows (Zhang et al., 2020) create some challenges in mapping water of dark tones. We further found better performance in mapping barren lands if they are less visited by humans, i.e., increasing accuracy from the urban center to rural regions. Barren lands in the urban center are often linked to construction sites. The fact that their surfaces are mixed with soil and other materials slightly affect accuracy. Overall, the performance of the FLUTE framework is comparable to or better than the state-of-the-art VHR databases based on their reported accuracies (see examples in Table 1). We also note that FLUTE only requires one type of data (i.e., VHR optical imagery) as input while its class scheme is more detailed than those in most existing VHR databases for studying the urban environment.

4.2. Summary and comparison of city-level LCLU in UrbanWatch

We summarized the area of each LCLU class at the city level for all 22 of the cities studied. While municipal boundaries are used here, we are in the process of expanding the database to cover the greater metropolitan area for each city. Fig. 7 shows the percentage of each LCLU, which allows for a straight comparison among cities or among LCLU within a city. Table 2 contains area values for individual classes, facilitating urban assessments where specific values are needed for a quantitative analysis. In general, cities in the Southern U.S. are rich in forest resources. Over 50% of Atlanta and Raleigh are covered by tree canopies. Due to climate variation, cities in other geographic regions of the U.S. have less forest coverage. Please note that we did not use ‘forest’ or ‘tree’ to label tree-related pixels, because forest often refers to large tree patches and tree is three dimensional. Tree canopy is a proper name to represent urban forest that is captured by remote sensing using a synoptic view. The hot desert climate puts Phoenix at the bottom of canopy cover among the 22 cities, although we also noticed the city's effort utilizing lawns and shrubs to improve its greenness. In comparison to grass/shrub, trees remain the major type of vegetation in most of the cities. Impervious surfaces are well represented in large cities. For example, over 50% of Chicago and San Francisco are covered by buildings, roads, and parking lots. We found that buildings are generally comparable to or are slightly less than roads in terms of area size. While parking lot area is significantly less than buildings and roads, they take up 3.22–9.16% of the land surface within a city boundary. Especially in the fast-growing areas, such as Charlotte, Dallas and Denver, parking lots cover relatively large areas (Charlotte: 5.84%; Dallas: 6.78%; Denver: 9.16%). Depending on the location, the 22 cities are highly variant in water coverage. As a seaport city, Seattle has the most significant component of water surface, which is followed by Dallas, Houston, and St. Louis. The other classes – barren, agriculture, and others – are a small portion of the urban LCLU across all cities. There are two exceptions in Denver and Phoenix. Denver still has relatively large amounts of croplands to the northeast, while Phoenix's barren lands are mostly located in sparsely vegetated areas to the north of the city.

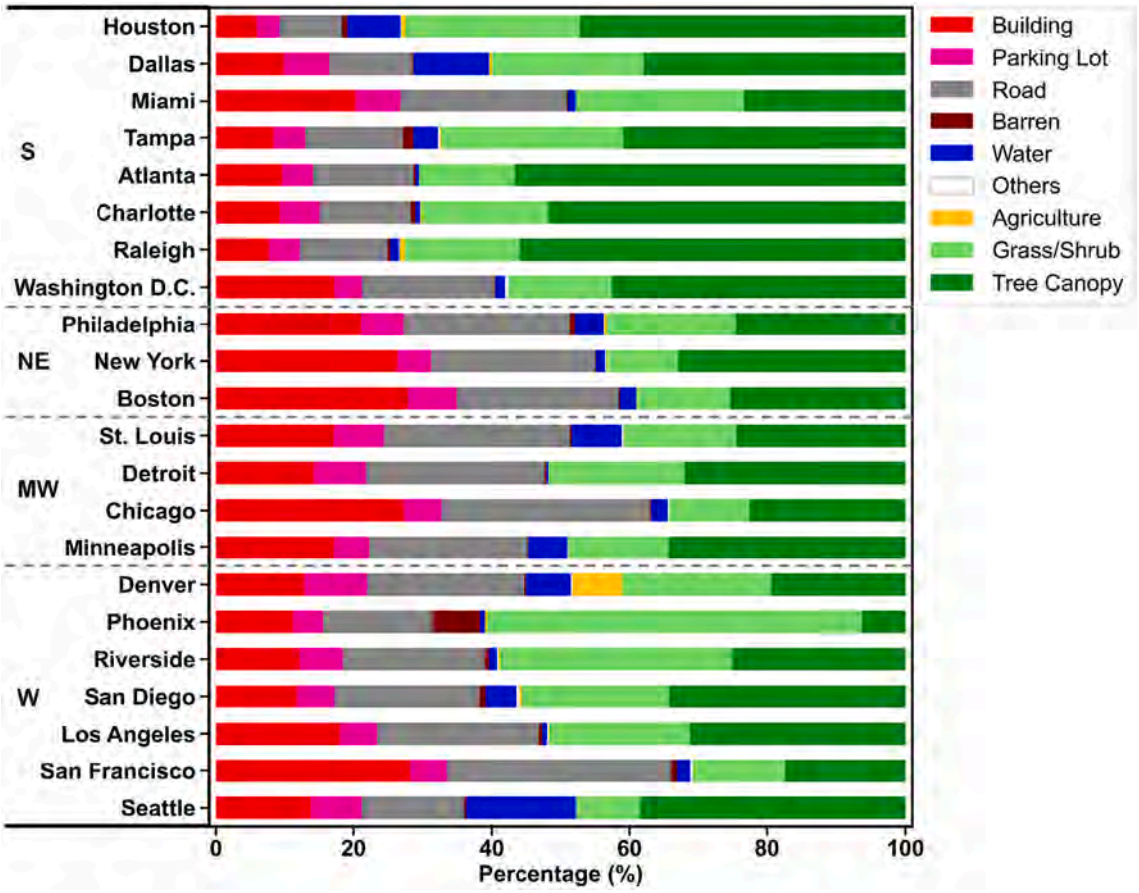

Fig. 7. LCLU compositions of 22 major U.S. cities across four geographic zones.

Table 2
Summary of the area of each LCLU class in 22 U.S. cities (km^2).

City	Building	Road	Parking Lot	Tree Canopy	Grass/Shrub	Water	Barren	Agriculture	Others
Atlanta	33.21	50.50	15.83	196.99	47.68	1.62	1.15	0.03	0.27
Boston	35.61	30.00	9.00	32.36	17.30	3.06	0.33	0.06	0.11
Chicago	163.32	180.66	32.46	134.47	70.54	14.74	0.69	0.03	1.00
Charlotte	71.68	102.39	45.26	401.04	143.13	5.46	4.90	1.04	0.61
Washington D.C.	27.73	30.83	6.49	68.58	24.25	2.25	0.29	0.02	0.79
Dallas	98.03	118.43	67.83	379.47	220.36	109.50	2.27	3.26	0.66
Denver	51.16	91.52	36.67	77.84	86.50	25.96	0.90	29.00	0.62
Detroit	51.44	93.51	26.91	115.42	70.47	1.02	0.82	0.00	0.29
Houston	38.77	59.46	21.02	307.73	164.77	50.30	5.23	4.15	0.65
Los Angeles	237.36	302.69	77.11	350.62	235.81	8.61	7.91	0.94	3.43
Miami	19.19	22.81	6.16	22.07	23.08	1.08	0.20	0.00	0.05
Minneapolis	25.54	34.22	7.54	51.05	21.59	8.45	0.21	0.03	0.10
New York	204.17	184.29	37.56	255.10	80.41	10.08	1.27	0.54	1.31
Philadelphia	76.74	87.79	22.53	89.37	68.39	15.07	3.00	1.27	0.20
Phoenix	81.10	115.56	32.33	45.68	396.83	5.06	50.03	1.67	0.27
Raleigh	35.75	60.23	21.03	262.68	78.64	6.05	1.99	3.37	0.36
Riverside	25.49	43.16	13.30	52.87	71.17	2.45	1.51	0.39	0.44
San Diego	99.08	178.05	47.73	292.11	181.16	37.78	8.01	2.67	4.61
Seattle	35.62	37.98	18.96	99.13	23.87	41.21	0.92	0.02	0.05
San Francisco	34.45	39.61	6.49	21.23	16.43	2.39	1.09	0.03	0.35
St. Louis	29.30	46.23	12.47	41.74	28.13	12.36	0.41	0.08	0.42
Tampa	25.38	43.57	13.97	125.10	80.57	10.69	4.81	0.70	0.99

5. Discussion

5.1. View angle effects on mapping

While the view angle effect offers insights on urban 3D structure, it also causes major challenges in high-resolution LCLU mapping (Matasci et al., 2015). This is especially true in the urban center where landscape surface (including buildings and trees) has high topographic variation, e.g., high-rise buildings blocking the view of neighboring buildings, roads, or vegetation. As a result, a slight change in view angle can lead to high contextual variation, which poses significant challenges for image semantic understanding. The task became even more difficult since our image scenes were acquired at different dates with different solar evaluation angles causing complex cast- and self-shadows (Zhang et al.,

2020).

In this study, we purposely designed a VEM submodule to deal with the view angle effect. To evaluate its performance, we compared the performance of the FLUTE framework with and without VEM. Here, the same training dataset was used to train the two frameworks, which were evaluated with the same validation data. We found that FLUTE with VEM outperformed FLUTE without VEM with noticeable improvement (OA: 91.28% versus 81.15%). This improvement was consistent along the urban-rural gradient (Fig. 8). Similarly, FLUTE with VEM achieved higher accuracies (F1-scores) than FLUTE without VEM in mapping two LCLU classes that are particularly sensitive to the change of view angle – building (F1-score: 90.15% versus 85.62%), and tree canopy (90.30% versus 82.36%, Fig. 8). Our visual interpretation of the UrbanWatch results suggests a relatively robust performance of FLUTE with VEM across various types of urban neighborhoods with development intensity from high to low (Fig. 9). One major contribution of VEM to LCLU mapping is that it mitigates the effect of shadow through a joint learning strategy of integrating three unsupervised DNNs in VEM. Capitalizing on this strategy, VEM was capable of providing comprehensive and unique contextual information across different urban scenes, which can bridge the semantic connection between shadow and non-shadow areas. For example, we found that shadow areas were easily misclassified as water in urban centers and tree canopies without VEM (Fig. 9). In our study, VEM has demonstrated the capacity to accurately extract various LCLU types within shadows. We note that shadows in large tree patches often correspond to tree canopy gaps (e.g., the last row of Fig. 9), which are a valuable indicator of natural or human disturbances. For urban ecological conservation, our framework has the potential to extract such fine-scale disturbances by distinguishing among different types of shadows.

We are aware of recent efforts dealing with the multi-view effect in LCLU classification (e.g., Li et al., 2020a, 2020b; Qiu et al., 2020; Sang et al., 2020), which have primarily focused on improving mapping accuracy by refining single-view features. While promising, such methods may encounter challenges in large-area mapping activities, where images contain high variation in view and/or solar elevation angle. Urban 3D structure varies from one region to another, which adds extra contextual variation encumbering accurate feature extraction. Another group of studies has attempted to integrate data from multiple sources (e.g., Chen et al., 2017; Wang et al., 2017) or multiple view angles for the area of interest (e.g., Huang et al., 2020). The need for multi-source or multi-view data remains a requirement for many cities, especially in developing countries. The FLUTE framework proposed in this study has the potential to map greater geographic regions, because it relies solely on single-date optical imagery to learn multi-view discriminative features while relaxing stringent data needs.

5.2. Effects of intraclass and interclass variation on mapping

The increase in spatial resolution unavoidably leads to an increase in intraclass variation and a decrease in interclass variation, which reduces the ability to retrieve typical LCLU classes in the spectral domain using traditional approaches (Prestele et al., 2016; Wu et al., 2021). In this study, the proposed SPM submodule (Section 3.3.3) in the FLUTE framework provides a unique solution to addressing this challenge. It builds upon and expands existing deep-learning-based strategies in dealing with multi-scale issues, which often focus on either spatial pyramid pooling or the perception field (e.g., Fu et al., 2019; Grippa et al., 2017; Luo et al., 2019; Zhao et al., 2019).

Here, we present examples of 28 sub-regions revealing different intraclass or interclass variations across urban areas and the corresponding mapping results in the UrbanWatch database (Fig. 10). Specifically, building is a representative land cover with multiple sizes across every city, from individual houses to large shopping malls. Over a large geographic region, the materials of building roofs tend to vary significantly from asphalt shingles to clay, metal roof panels, or concrete roofing tiles. Like buildings, roads are different sizes (e.g., small countryside pathways versus big highways), and the surface material could be asphalt, concrete, or gravel. The FLUTE framework did well with high intraclass variation, producing accurate and consistent results across LCLU classes (see F1-scores in Fig. 6 and sample results in Fig. 10). This is attributed to the two strategies used in the framework: (i) the benchmark database covers diverse patterns of urban neighborhoods and allows the network to learn simple-to-complex landscape patterns, and (ii) it integrates multi-extent features from three perception fields which was found to be effective for delineating ground object boundaries. We further note its effectiveness in extracting buildings with dark-tone roof materials and simultaneously surrounded by trees, which may otherwise be misinterpreted as shadow or water.

The challenge for interclass classification was mainly related to roads versus parking lots. They have the same or similar surface material and differences are less about land cover and more about land use. Our framework capitalizes on their subtle variation in shape (elongated roads versus rectangle parking lots), patterns of surface markings or vehicles, and the neighboring LCLU types to distinguish between the two classes. FLUTE proved to have the capacity to extract the border between roads and parking lots (Q, R, S and T in Fig. 10). We are aware of the challenge in mapping high-density residential neighborhoods where parking lots are small and are adjacent to roads (e.g., designated areas for street parking). Although not perfect, our framework shows promising results for dealing with this issue (e.g., sample patch Q_1 in Fig. 10). Tree canopy and grass/shrub are two vegetation classes that often create signature confusion in LCLU classifications. Recent efforts to integrate LiDAR-derived vertical structures provide a viable solution (e.g., Pilant et al., 2020). However, our results, obtained using only

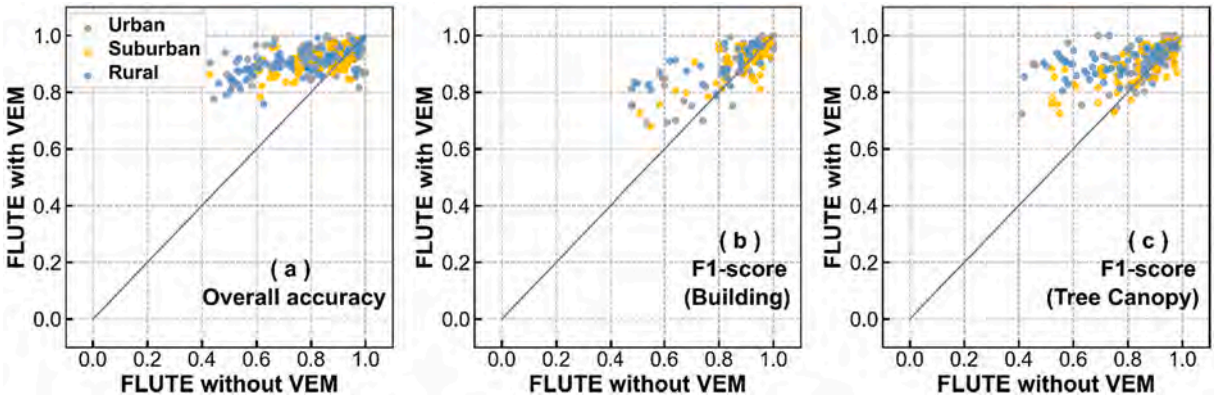

Fig. 8. Comparison of FLUTE performance with and without the VEM submodule using (a) overall accuracy and F1-score for mapping (b) buildings and (c) tree canopies in urban, suburban, and rural regions, respectively.

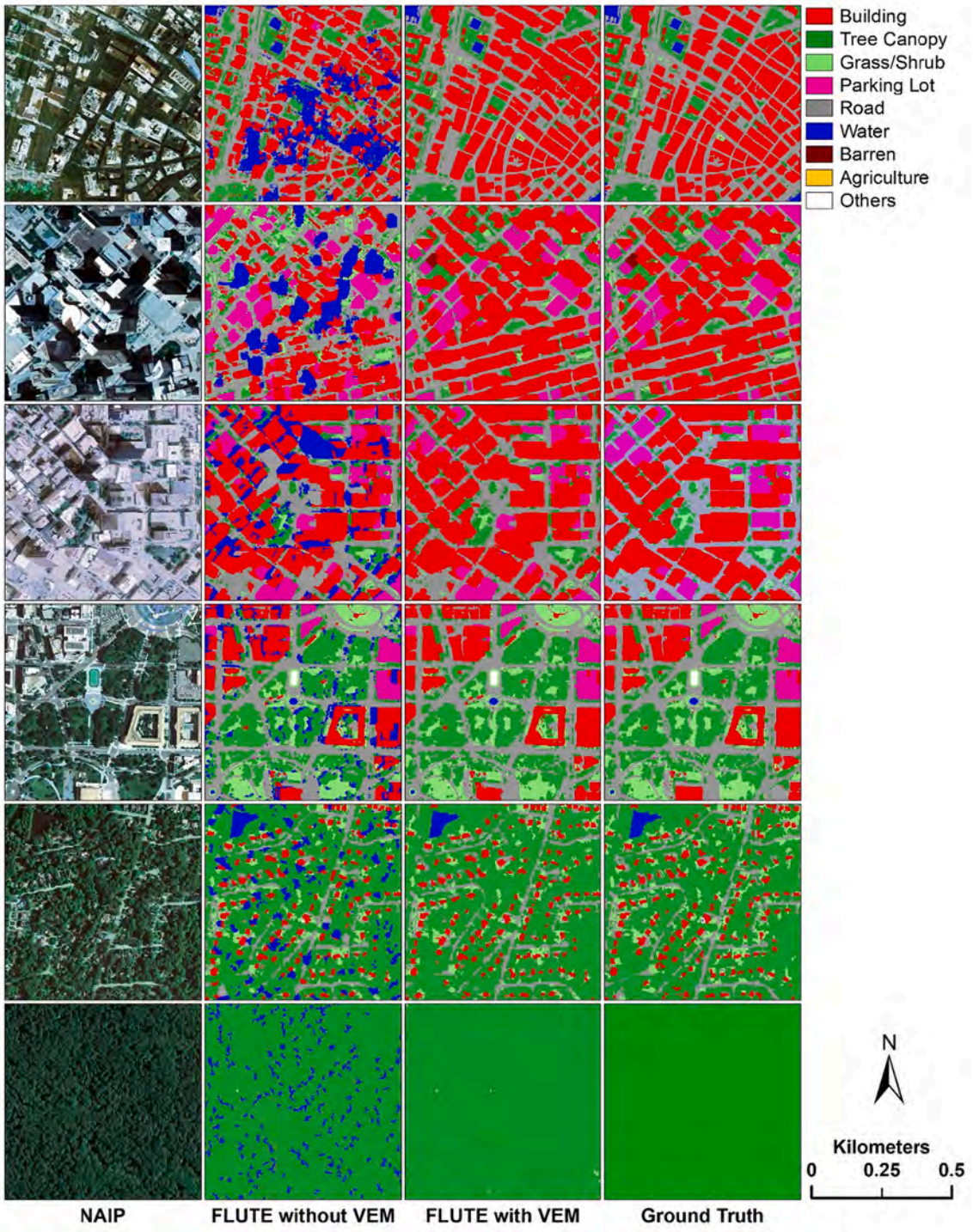

Fig. 9. Six sample areas represent the change in urban development intensity (high at the top and low at the bottom) and the corresponding LCLU maps with and without using VEM. The data from left to right are NAIP (National Agriculture Imagery Program) true-color imagery, LCLU generated by FLUTE without VEM, LCLU generated by FLUTE with VEM, and ground truth, respectively.

optical imagery, achieved a high F1-score (~90%) for both classes. A large number of individual trees (e.g., city-managed street trees from sample patch E in Fig. 10) are readily available in our product, which can facilitate accurate and efficient urban forest management, research, and outreach activities.

5.3. Comparison with state-of-the-art LCLU databases

The comparison between UrbanWatch (1 m) with the coarser resolution databases NLCD (30 m) and Esri (10 m) demonstrates an apparent benefit of capturing detailed urban spatial patterns with high-resolution imagery. Although it is not an apples-to-apples comparison between databases of different spatial resolutions or different geographic coverage (e.g., NLCD maps the entire conterminous United States), UrbanWatch is capable of better distinguishing among various LCLU classes in highly heterogeneous regions. For instance, an urban center is mainly covered by impervious surfaces, and the results from NLCD and Esri are too general to understand the spatial relationships

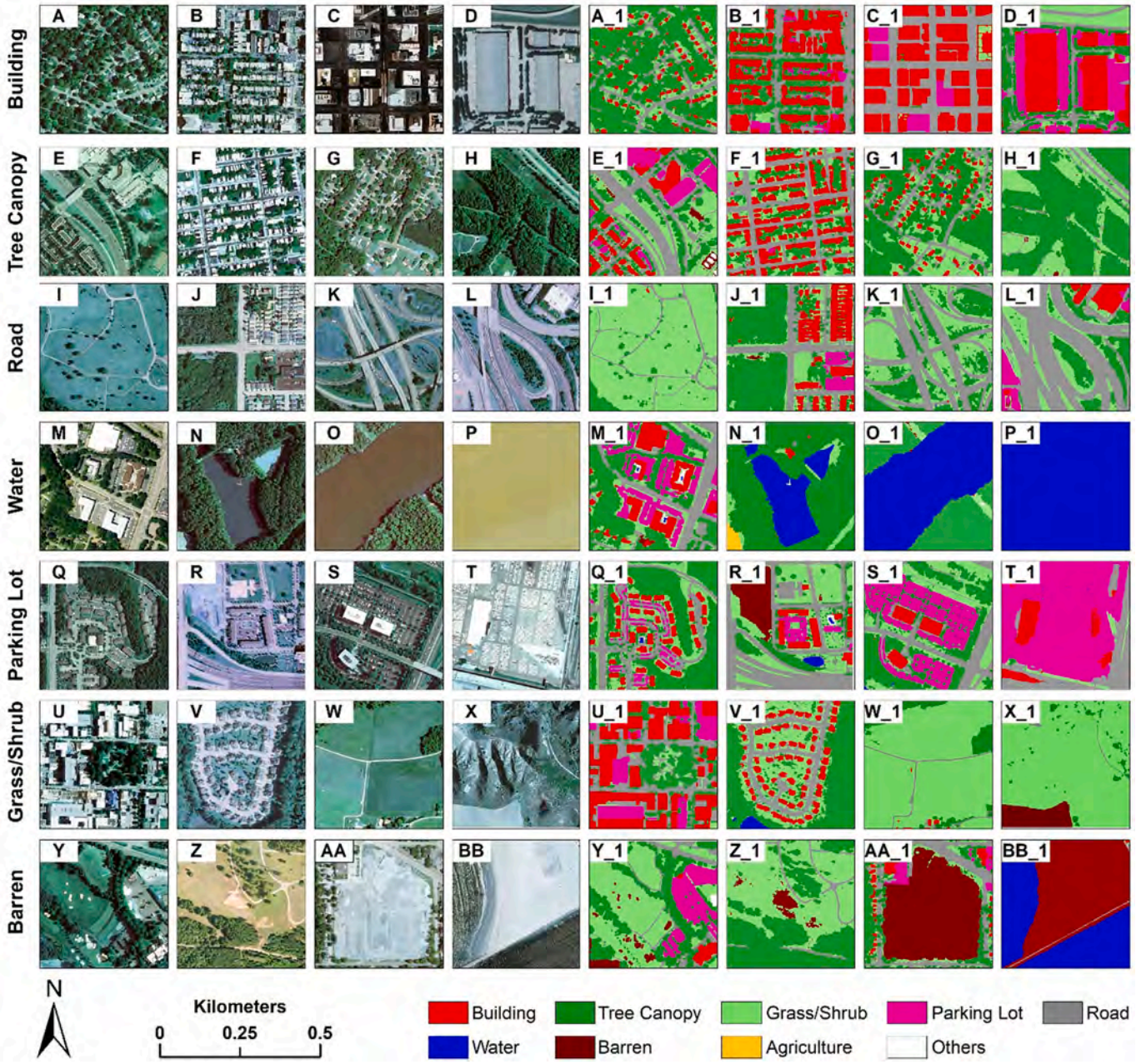

Fig. 10. Examples of various types of urban LCLU (e.g., "A", "B", and "AA") and their corresponding results in UrbanWatch (e.g., "A_1", "B_1", and "AA_1"), demonstrating intraclass or interclass variation.

among roads, buildings, and parking lots (Fig. 11). As buildings and trees are essential components of 3D urban morphology, accurate extraction of these ground objects in UrbanWatch offers the possibility to analyze human thermal comfort and health (Myint et al., 2015a), flooding risk (Mignot et al., 2019), building energy consumption (Li et al., 2017), and the movement of animals (Bierwagen, 2007). Bare ground in many U.S. cities is associated with construction sites. In highly developed areas, it often infers gentrification or urban infill that occur on small land patches. The detection and spatial analysis of these phenomena across large geographic regions can support a range of social, economic, and policy studies (Lees, 2008). Additionally, urban forest inventories track isolated street trees and are managed at the individual tree level. Remote assessment is only achievable with credible, high-resolution LCLU products.

We further compared UrbanWatch with three LCLU databases of the same spatial resolution (1 m), including the Chesapeake Bay Watershed (hereafter CBW), EarthDefine, and Microsoft land cover (see Table 1). EarthDefine and Microsoft have reported an accuracy at or slightly higher than 90%, which is similar to ours. Furthermore, NAIP imagery, CBW, EarthDefine, and Microsoft all used ancillary data, including LiDAR, Landsat, and/or GIS vector layers. For urban regions, EarthDefine has one single class describing impervious surfaces, while both CBW and Microsoft have categorized impervious surfaces into two broad classes – road and non-road (or other). Fig. 12 shows a comparison between UrbanWatch and Microsoft, CBW, and EarthDefine in sample urban areas from three cities – State College, PA, Washington D.C., and Miami, FL. UrbanWatch and Microsoft have revealed generally consistent performance in capturing buildings. However, Microsoft tends to over-estimate buildings by misclassifying some small roads. The contrast between parking lots and buildings in UrbanWatch makes it easier than

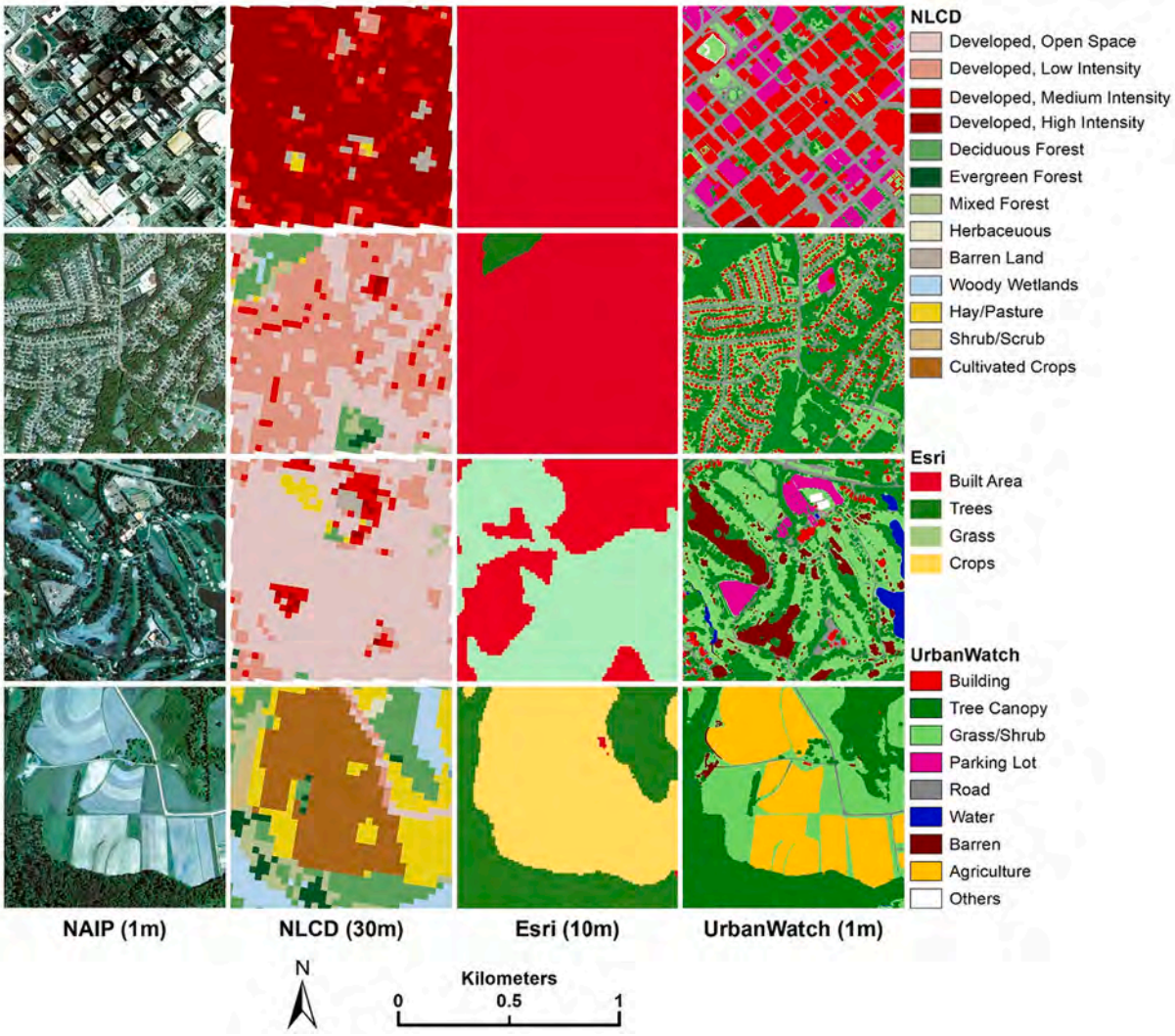

Fig. 11. Comparisons between UrbanWatch (1 m) and two medium-resolution databases NLCD (30 m) and Esri land cover (10 m) in four sample areas along the urban(top)-rural(bottom) gradient.

the Microsoft database to analyze urban 3D form and its effects on urban environments. We note that CBW capitalized on GIS road polygons to distinguish between roads and above-road tree canopies. While UrbanWatch does not have such capacity due to using single-date optical imagery, roads with canopy cover show visually comparable performance as CBW. UrbanWatch is able to distinguish between roads and parking lots that cannot be separated in CBW without ancillary data. Compared to Microsoft and CBW, EarthDefine has an even coarser classification scheme, and all buildings, roads, and parking lots are considered as one single class. We further found that EarthDefine gives slightly more emphasis on forests than UrbanWatch. Especially in the transitional areas when trees meet other LCLU classes at lower heights, EarthDefine is prone to treat these areas as trees. While dealing with the VHR imagery, mixed spectral signatures remain in the scene when tree leaves (with gaps) are overlaid onto other land cover types.

To date, the supervised GEOBIA framework remains popular for creating most of the VHR LCLU databases (e.g., CBW) due to its maturity. Depending on the study area and the classification scheme, the reported accuracy ranges from 82% to over 95% (Table 1). This is comparable with the FLUTE framework built upon semi-supervised learning and deep learning. We note that the computing needs for FLUTE are slightly higher than most of the object-based classifications due to the nature of running deep learning models (e.g., high performance GPU). However, FLUTE has demonstrated high generalization capacity. When applied to geographically broad study areas, FLUTE does not need to re-adjust model parameters or image features, significantly reducing human intervention and the possibility of human biases. Its end-to-end learning structure makes it easier for practitioners to generate accurate LCLU maps over large urban areas.

6. Conclusion

This study aims to support urban research, management, and outreach by advancing our understanding of the fine-scale spatial patterns in urban land cover and land use (LCLU). To achieve this goal, we developed the FLUTE framework to address several challenges that frequently occur in large-area, high-resolution urban mapping, including the view angle effect, high intraclass and low interclass variation, and multiscale land cover. The development of this framework capitalized on recent advances in semi-supervised learning and deep learning with the purpose of enhancing FLUTE's generalization capacity for multi-city mapping while using one type of input data – optical imagery with NIR, R, G and B bands (no ancillary data, such as LiDAR or vector layers). The proposed feature extraction module Scale-aware Parsing Module (SPM) can effectively characterize urban objects of varying scales and estimate LCLU within shadows typically casted by buildings and trees. The View-aware Embedding Module (VEM) mitigates the view angle effect on scene structure interpretation, with a

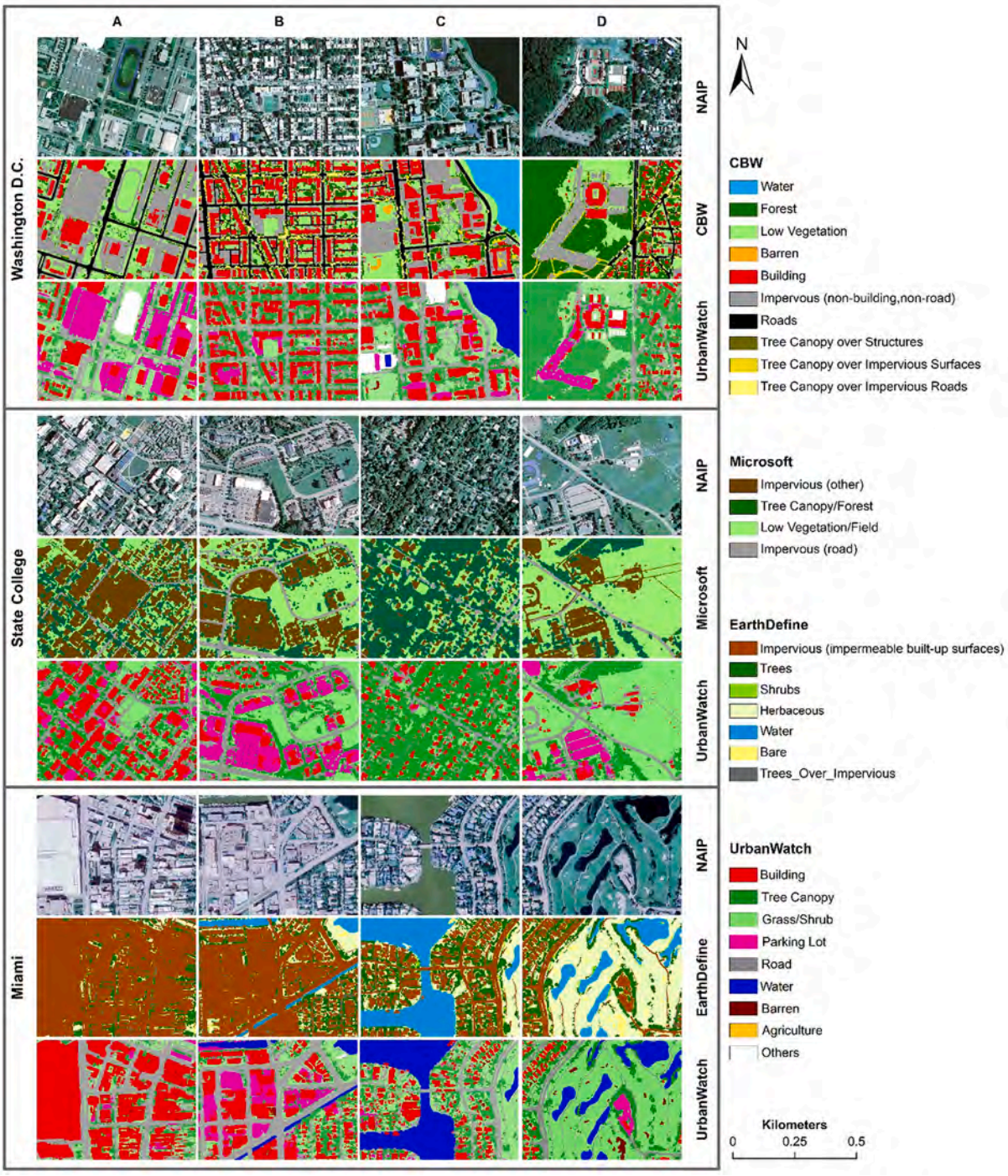

Fig. 12. Comparisons between UrbanWatch and three 1-m LCLU databases: Chesapeake Bay Watershed (CBW) land cover (Washington D.C.), Microsoft high-resolution land cover (State College, PA), and EarthDefine land cover (Miami, FL).

10.13% increase in overall accuracy as compared to the framework without VEM. To facilitate model training, we constructed a new benchmark database protocol which contains 52.43 million labeled pixels to capture diverse LCLU types and spatial patterns. Following training, we successfully applied FLUTE to produce a 1-meter resolution UrbanWatch database for 22 major cities across the conterminous United States. We categorized each city into nine LCLU classes, i.e., building, road, parking lot, tree canopy, grass/shrub, water, agriculture, barren, and others, with an overall accuracy of 91.52%. The credibility of UrbanWatch is also evidenced by a comparison with five other state-of-the-art LCLU databases from medium to high spatial resolution. To benefit evidence-based decision making processes for urban sustainability and the quality of life, we have further adopted an open standard to make UrbanWatch freely accessible at https://urbanwatch.charlotte.edu. The NAIP imagery are available from the USGS Earth Explorer data portal (USGS, 2020). UrbanWatch is an ongoing project, which is built on strong collaborations. We are willing to share the framework with users upon request to achieve mutual benefit, and are in the process of developing new partnerships with other research groups and stakeholders to improve the quality of the database (e.g., more detailed LCLU classes) and expand its coverage to broader geographical regions (e.g., from major cities to smaller cities).

CRediT authorship contribution statement

Yindan Zhang: Methodology, Data curation, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. **Gang Chen:** Conceptualization, Resources, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing. **Soe W. Myint:** Formal analysis, Writing – review & editing. **Yuyu Zhou:** Formal analysis, Writing – review & editing. **Geoffrey J. Hay:** Formal analysis, Writing – review & editing. **Jelena Vukomanovic:** Formal analysis, Writing – review & editing. **Ross K. Meentemeyer:** Formal analysis, Writing – review & editing.

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.

Acknowledgements

This work was supported by the University of North Carolina at Charlotte. The authors are grateful to the editors and four anonymous reviewers for their constructive comments, which greatly helped to improve this paper.

Appendix A

Table A

City-level LCLU mapping accuracy – OA (overall accuracy), F1-score, UA (user's accuracy), PA (producer's accuracy), and NA (non-site-specific accuracy) – for nine LCLU classes in each of the 22 mapped cities.

City	Metric	Building	Parking Lot	Road	Barren	Water	Others	Agriculture	Grass/Shrub	Tree Canopy
Houston (OA: 91.49%)	NA	0.0048	−0.0263	0.0012	0.0245	0.0001	−0.0005	−0.0012	−0.0328	0.0314
	UA	0.9286	0.9100	0.8775	0.8133	0.8736	0.7600	0.8000	0.8970	0.8582
	PA	0.9371	0.8250	0.9013	0.9520	0.9425	0.7200	0.7200	0.8692	0.9918
	F1-score	0.9326	0.8649	0.8898	0.8796	0.9012	0.7398	0.7564	0.8807	0.9206
Dallas (OA: 92.75%)	NA	−0.0058	−0.0017	−0.0073	0.0005	−0.0003	0.0097	0.0022	−0.0034	−0.0112
	UA	0.9376	0.8963	0.9239	0.8517	0.9500	0.7426	0.8775	0.9400	0.9277
	PA	0.9006	0.8850	0.8919	0.8833	0.9200	0.7700	0.8873	0.9109	0.8919
	F1-score	0.9186	0.8902	0.9075	0.8671	0.9342	0.7561	0.8805	0.9258	0.9096
Miami (OA: 89.81%)	NA	0.0147	0.0027	0.0120	0.0009	−0.0014	0.0187	n/a	−0.0158	0.0154
	UA	0.8014	0.8263	0.8067	0.8913	0.9630	0.7400	n/a	0.9100	0.8600
	PA	0.8935	0.8850	0.9300	0.9050	0.9050	0.8200	n/a	0.8500	0.9400
	F1-score	0.8450	0.8547	0.8638	0.8982	0.9330	0.7782	n/a	0.8789	0.8983
Tampa (OA: 89.84%)	NA	−0.0154	−0.0092	0.0120	0.0016	−0.0063	0.0068	n/a	−0.0040	−0.0090
	UA	0.8717	0.9367	0.8759	0.7975	0.9663	0.7736	n/a	0.9077	0.9264
	PA	0.8200	0.9133	0.9029	0.8400	0.9050	0.8047	n/a	0.8192	0.8773
	F1-score	0.8452	0.9247	0.8893	0.8179	0.9347	0.7886	n/a	0.8615	0.9014
Atlanta (OA: 92.42%)	NA	−0.0127	0.0059	0.0115	−0.0069	−0.0072	−0.0106	n/a	−0.0074	0.0043
	UA	0.9244	0.8690	0.8922	0.8700	0.9650	0.8400	n/a	0.9128	0.8984
	PA	0.8744	0.8951	0.9304	0.8400	0.9050	0.7900	n/a	0.8980	0.9184
	F1-score	0.8988	0.8820	0.9106	0.8546	0.9341	0.8141	n/a	0.9057	0.9084
Charlotte (OA: 93.96%)	NA	−0.0023	−0.0082	−0.0108	0.0065	−0.0067	0.0030	0.0015	−0.0097	−0.0109
	UA	0.9500	0.9157	0.9221	0.8867	0.9825	0.8200	0.8642	0.9642	0.9452
	PA	0.9173	0.8899	0.8921	0.9005	0.9450	0.8667	0.8943	0.8943	0.8700
	F1-score	0.9332	0.9027	0.9067	0.8934	0.9635	0.8428	0.8790	0.9281	0.9061
Raleigh (OA: 94.04%)	NA	−0.0064	−0.0102	−0.0073	0.0048	0.0020	0.0115	−0.0023	−0.0057	−0.0048
	UA	0.9686	0.9382	0.9371	0.8625	0.9600	0.8560	0.9395	0.9558	0.9557
	PA	0.9486	0.8773	0.8862	0.8950	0.9800	0.8920	0.8700	0.8883	0.8938
	F1-score	0.9586	0.9066	0.9108	0.8785	0.9697	0.8735	0.9034	0.9207	0.9235
Washington D.C. (OA: 92.96%)	NA	−0.0108	0.0140	−0.0102	0.0092	−0.0103	0.0175	n/a	−0.0388	−0.0214
	UA	0.9400	0.8633	0.9183	0.8200	0.9800	0.7500	n/a	0.9391	0.9580
	PA	0.8753	0.9100	0.8733	0.9200	0.9140	0.8950	n/a	0.8818	0.8769
	F1-score	0.9066	0.8860	0.8951	0.8670	0.9457	0.8161	n/a	0.9094	0.9157
Philadelphia (OA: 92.09%)	NA	−0.0144	0.0109	−0.0085	0.0099	−0.0002	0.0100	n/a	0.0067	−0.0139
	UA	0.9200	0.8865	0.9280	0.7825	0.9500	0.7100	n/a	0.9286	0.9529
	PA	0.8710	0.9136	0.8864	0.8825	0.9354	0.8033	n/a	0.9543	0.8929
	F1-score	0.8946	0.8901	0.9066	0.8295	0.9428	0.7539	n/a	0.9414	0.9220
New York (OA: 90.86%)	NA	−0.0021	0.0063	−0.0028	0.0162	−0.0028	−0.0018	n/a	−0.0020	−0.0115
	UA	0.9464	0.8350	0.9039	0.7850	0.9811	0.7867	n/a	0.8756	0.9707
	PA	0.8943	0.9155	0.8757	0.8567	0.9689	0.7500	n/a	0.8594	0.8900
	F1-score	0.9195	0.8735	0.8897	0.8192	0.9750	0.7679	n/a	0.8673	0.9287
Boston (OA: 90.71%)	NA	−0.0091	−0.0102	0.0115	0.0640	−0.0502	−0.0073	n/a	−0.0146	0.0028
	UA	0.9033	0.9100	0.8038	0.8050	0.9441	0.7967	n/a	0.9770	0.9123
	PA	0.8778	0.8200	0.9300	0.8670	0.9294	0.7157	n/a	0.8630	0.9456
	F1-score	0.8905	0.8624	0.8621	0.8348	0.9366	0.7541	n/a	0.9165	0.9288
St. Louis (OA: 91.69%)	NA	0.0057	−0.0104	0.0039	0.0207	−0.0003	0.0062	n/a	−0.0064	−0.0102
	UA	0.9167	0.9135	0.8546	0.7900	0.9436	0.7157	n/a	0.9330	0.9500
	PA	0.9800	0.8736	0.9338	0.8342	0.9156	0.7662	n/a	0.8746	0.8935
	F1-score	0.9474	0.8930	0.8922	0.8113	0.9295	0.7401	n/a	0.9027	0.9207
Detroit (OA: 92.86%)	NA	0.0011	0.0027	0.0001	0.0011	0.0013	0.0004	n/a	0.0162	−0.0118
	UA	0.9218	0.8886	0.8936	0.7450	0.9600	0.8120	n/a	0.8796	0.9452
	PA	0.9573	0.9303	0.9186	0.8450	1.0000	0.8764	n/a	0.9312	0.8884
	F1-score	0.9392	0.9091	0.9058	0.7918	0.9796	0.8431	n/a	0.9047	0.9158
Chicago (OA: 91.91%)	NA	−0.0061	0.0107	−0.0008	0.0124	0.0002	0.0029	n/a	−0.0117	0.0014
	UA	0.9231	0.8736	0.8778	0.7400	0.9400	0.7864	n/a	0.9488	0.8485

(*continued on next page*)

Table A (*continued*)

City	Metric	Building	Parking Lot	Road	Barren	Water	Others	Agriculture	Grass/Shrub	Tree Canopy
	PA	0.9125	0.9167	0.8633	0.8243	0.9900	0.8800	n/a	0.9067	0.9400
	F1-score	0.9177	0.8945	0.8706	0.7799	0.9645	0.8304	n/a	0.9274	0.8919
Minneapolis (OA: 90.89%)	NA	0.0078	0.0103	0.0068	0.0103	0.0002	0.0214	n/a	−0.0148	−0.0016
	UA	0.9021	0.8788	0.8832	0.7254	0.9250	0.7300	n/a	0.9261	0.9273
	PA	0.9321	0.9064	0.9043	0.8146	0.9500	0.8900	n/a	0.8387	0.9077
	F1-score	0.9167	0.8924	0.8935	0.7673	0.9372	0.8021	n/a	0.8801	0.9174
Denver (OA: 88.96%)	NA	−0.0075	0.0065	−0.0124	0.0027	0.0023	0.0120	0.0134	0.0091	−0.0187
	UA	0.9350	0.8740	0.9213	0.7900	0.9100	0.7858	0.7658	0.8342	0.8480
	PA	0.8533	0.9010	0.8656	0.8000	0.9400	0.8325	0.8825	0.8858	0.7920
	F1-score	0.8923	0.8874	0.8927	0.7952	0.9246	0.8086	0.8206	0.8593	0.8191
Phoenix (OA: 87.97%)	NA	−0.0066	0.0107	0.0013	0.0260	0.0018	0.0103	0.0133	−0.0118	0.0045
	UA	0.8980	0.8667	0.8600	0.7667	0.9100	0.7425	0.8125	0.9567	0.8147
	PA	0.8580	0.9067	0.8929	0.8567	0.9300	0.8425	0.8925	0.8311	0.8800
	F1-score	0.8774	0.8864	0.8762	0.8091	0.9199	0.7894	0.8505	0.8894	0.8460
Riverside (OA: 87.63%)	NA	−0.0072	−0.0246	0.0276	0.0329	−0.0008	0.0127	−0.0024	0.0037	0.0107
	UA	0.9025	0.9800	0.8033	0.7600	0.9880	0.7900	0.8435	0.8000	0.8275
	PA	0.8538	0.7800	0.9425	1.0000	0.9700	0.8600	0.8155	0.8155	0.8717
	F1-score	0.8775	0.8688	0.8675	0.8635	0.9789	0.8236	0.8293	0.8076	0.8490
San Diego (OA: 88.36%)	NA	0.0064	0.0078	−0.0019	0.0184	0.0015	0.0105	−0.0062	0.0091	−0.0103
	UA	0.8886	0.8750	0.9233	0.7000	0.9150	0.7300	0.8568	0.8575	0.8846
	PA	0.9114	0.9167	0.8789	0.8942	0.9432	0.8345	0.8155	0.9163	0.7825
	F1-score	0.8998	0.8955	0.9003	0.7857	0.9289	0.7788	0.8357	0.8859	0.8305
Los Angeles (OA: 90.79%)	NA	−0.0021	0.0061	−0.0009	−0.0100	0.0013	−0.0263	n/a	0.0072	−0.0136
	UA	0.9286	0.8840	0.8792	0.8663	0.9320	0.8974	n/a	0.8522	0.9144
	PA	0.9043	0.9041	0.8671	0.8375	0.9325	0.8100	n/a	0.8926	0.8744
	F1-score	0.9162	0.8938	0.8731	0.8517	0.9321	0.8515	n/a	0.8718	0.8941
San Francisco (OA: 90.95%)	NA	−0.0011	−0.0031	0.0103	0.0108	0.0052	0.0106	n/a	−0.0118	−0.0023
	UA	0.9080	0.8940	0.9019	0.7900	0.9300	0.7676	n/a	0.9186	0.9305
	PA	0.8953	0.8348	0.9304	0.8750	0.9700	0.8233	n/a	0.8673	0.9281
	F1-score	0.9016	0.8634	0.9159	0.8303	0.9496	0.7944	n/a	0.8922	0.9293
Seattle (OA: 87.58%)	NA	−0.0106	0.0102	−0.0308	0.0235	0.0013	0.0089	n/a	0.0168	−0.0109
	UA	0.9067	0.8192	0.9200	0.7250	0.9500	0.7456	n/a	0.8100	0.9033
	PA	0.8533	0.8480	0.7420	0.9300	0.9643	0.8058	n/a	0.8950	0.8183
	F1-score	0.8792	0.8334	0.8214	0.8152	0.9572	0.7743	n/a	0.8504	0.8586

Appendix B

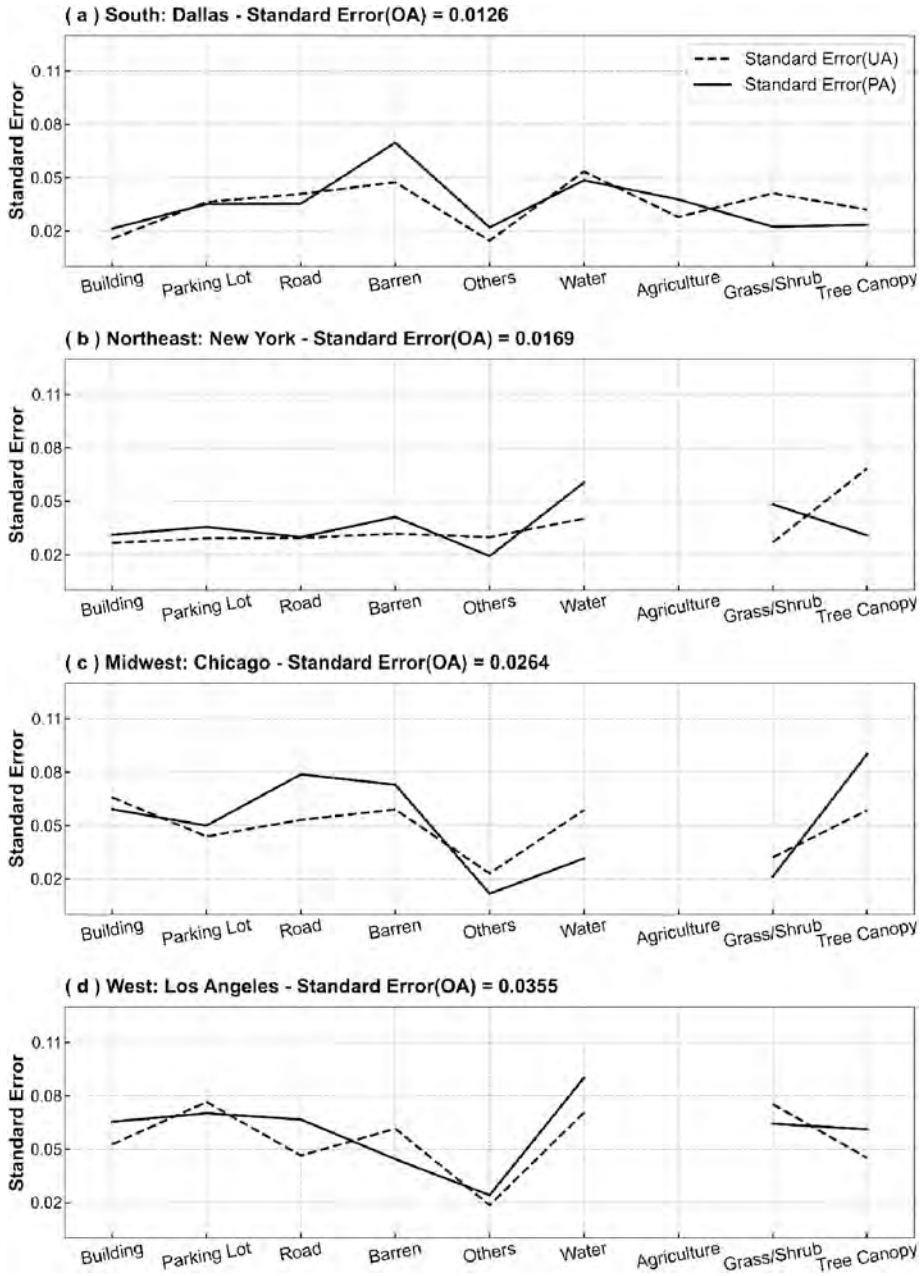

Fig. B. Standard errors for overall accuracy (OA), user's accuracy (UA), and producer's accuracy (PA) for four major cities across U.S. geographic zones - South: Dallas, Northeast: New York, Midwest: Chicago, and West: Los Angeles.

References

Aasen, H., Honkavaara, E., Lucieer, A., Zarco-Tejada, P.J., 2018. Quantitative remote sensing at ultra-high resolution with UAV spectroscopy: a review of sensor technology, measurement procedures, and data correction workflows. Remote Sens. 10, 1091.

Aghajani, S., Kalantar, M., 2017. Operational scheduling of electric vehicles parking lot integrated with renewable generation based on bilevel programming approach. Energy 139, 422–432.

Anderson, J.R., Hardy, E.E., Roach, J.T., Witmer, R.E., 1976. A Land Use and Land Cover Classification System for Use with Remote Sensor Data. Geological Survey Professional Paper 964. United States Department of the Interior, Washington, DC, USA.

Andrew, G., Arora, R., Bilmes, J., Livescu, K., 2013. Deep canonical correlation analysis. In: International Conference on Machine Learning. PMLR, pp. 1247–1255.

Angelidou, M., Psaltoglou, A., Komninos, N., Kakderi, C., Tsarchopoulos, P., Panori, A., 2017. Enhancing sustainable urban development through smart city applications. J. Sci. Technol. Policy Manag. 9, 146–169.

Antunes, R., Blaschke, T., Tiede, D., Bias, E., Costa, G., Happ, P., 2019. Proof of concept of a novel cloud computing approach for object-based remote sensing data analysis and classification. GISci. Remote Sens. 56, 536–553.

Audebert, N., Le Saux, B., Lefèvre, S., 2018. Beyond RGB: very high resolution urban remote sensing with multimodal deep networks. ISPRS J. Photogramm. Remote Sens. 140, 20–32.

Azulay, A., Weiss, Y., 2018. Why do Deep Convolutional Networks Generalize so Poorly to Small Image Transformations? Preprint arXiv:1805.12177.

Badrinarayanan, V., Handa, A., Cipolla, R., 2015. Segnet: A Deep Convolutional Encoder-Decoder Architecture for Robust Semantic Pixel-Wise Labelling arXiv preprint arXiv: 1505.07293.

Bierwagen, B.G., 2007. Connectivity in urbanizing landscapes: the importance of habitat configuration, urban area size, and dispersal. Urban Ecosyst. 10, 29–42.

Blaschke, T., Hay, G.J., Kelly, M., Lang, S., Hofmann, P., Addink, E., Feitosa, R.Q., Van der Meer, F., Van der Werff, H., Van Coillie, F., 2014. Geographic object-based image analysis–towards a new paradigm. ISPRS J. Photogramm. Remote Sens. 87, 180–191.

Census, 2020. U.S. Census Bureau's MAF/TIGER Geographic Database. https://www.census.gov/en.html.

Chen, G., Hay, G.J., Carvalho, L.M., Wulder, M.A., 2012. Object-based change detection. Int. J. Remote Sens. 33, 4434–4457.
Chen, B., Huang, B., Xu, B., 2017. Multi-source remotely sensed data fusion for improving land cover classification. ISPRS J. Photogramm. Remote Sens. 124, 27–39.
Chen, G., Weng, Q., Hay, G.J., He, Y., 2018. Geographic object-based image analysis (GEOBIA): emerging trends and future opportunities. GISci. Remote Sens. 55, 159–182.
Chen, G., Singh, K.K., Lopez, J., Zhou, Y., 2020. Tree canopy cover and carbon density are different proxy indicators for assessing the relationship between forest structure and urban socio-ecological conditions. Ecol. Indic. 113, 106279.
Cheng, G., Han, J., Lu, X., 2017. Remote sensing image scene classification: benchmark and state of the art. Proc. IEEE 105, 1865–1883.
Chesapeake Conservancy, 2020. Chesapeake Bay Watershed Land Cover. https://www.chesapeakeconservancy.org/conservation-innovation-center/high-resolution-data/land-cover-data-project. Last accessced on October 5, 2021.
Contreras, D., Blaschke, T., Tiede, D., Jilge, M., 2016. Monitoring recovery after earthquakes through the integration of remote sensing, GIS, and ground observations: the case of L'Aquila (Italy). Cartogr. Geogr. Inf. Sci. 43, 115–133.
Damianou, A., Lawrence, N.D., 2013. Deep gaussian processes. In: Proc. 16th Int. Conf. Artif. Intell. Statist., pp. 207–215.
Deepan, P., Sudha, L., 2019. Fusion of deep learning models for improving classification accuracy of remote sensing images. Mech Continua Math Sci. 14, 189–201.
Dewitz, J., 2019. National Land Cover Database (NLCD) 2016 Products. US Geological Survey data release.
Du, S., Du, S., Liu, B., Zhang, X., Zheng, Z., 2020. Large-scale urban functional zone mapping by integrating remote sensing images and open social data. GISci. Remote Sens. 57 (3), 411–430.
Dutta, D., Chen, G., Chen, C., Gagné, S.A., Li, C., Rogers, C., Matthews, C., 2020. Detecting plant invasion in urban parks with aerial image time series and residual neural network. Remote Sens. 12 (21), 3493.
EarthDefine, 2020. EarthDefine Land Cover. https://www.earthdefine.com/landcover/. Last accessed on October 5, 2020.
Faroughi, M., Karimimoshaver, M., Aram, F., Solgi, E., Mosavi, A., Nabipour, N., Chau, K.-W., 2020. Computational modeling of land surface temperature using remote sensing data to investigate the spatial arrangement of buildings and energy consumption relationship. Eng. Appl. Comput. Fluid Mech. 14 (1), 254–270.
Fu, Y., Liu, K., Shen, Z., Deng, J., Gan, M., Liu, X., Lu, D., Wang, K., 2019. Mapping impervious surfaces in town–rural transition belts using China's GF-2 imagery and object-based deep CNNs. Remote Sens. 11 (3), 280.
Gallego, F.J., 2012. The efficiency of sampling very high resolution images for area estimation in the European Union. Int. J. Remote Sens. 33, 1868–1880.
Giada, S., De Groeve, T., Ehrlich, D., Soille, P., 2003. Information extraction from very high resolution satellite imagery over Lukole refugee camp, Tanzania. Int. J. Remote Sens. 24, 4251–4266.
Gidaris, S., Singh, P., Komodakis, N., 2018. Unsupervised Representation Learning by Predicting Image Rotations. Preprint arXiv:1803.07728.
Godwin, C., Chen, G., Singh, K.K., 2015. The impact of urban residential development patterns on forest carbon density: an integration of LiDAR, aerial photography and field mensuration. Landsc. Urban Plan. 136, 97–109.
Goldberger, J., Gordon, S., Greenspan, H., 2003. An Efficient Image Similarity Measure Based on Approximations of KL-Divergence Between Two Gaussian Mixtures. ICCV, pp. 487–493.
Goutte, C., Gaussier, E., 2005. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In: European Conference on Information Retrieval. Springer, pp. 345–359.
Grippa, T., Georganos, S., Lennert, M., Vanhuysse, S., Wolff, E., 2017. A local segmentation parameter optimization approach for mapping heterogeneous urban environments using VHR imagery. In: Remote Sensing Technologies and Applications in Urban Environments II 104310G. International Society for Optics and Photonics.
Guo, X., Liu, X., Zhu, E., Yin, J., 2017. Deep clustering with convolutional autoencoders. In: International Conference on Neural Information Processing. Springer, pp. 373–382.
Han, R., Liu, P., Wang, G., Zhang, H., Wu, X., Hong, S.-H., 2020. Advantage of combining OBIA and classifier ensemble method for very high-resolution satellite imagery classification. J. Sensors 2020, 1–15.
Hay, G.J., Castilla, G., 2008. Geographic Object-Based Image Analysis (GEOBIA): A New Name for a New Discipline. Object-Based Image Analysis. Springer, pp. 75–89.
Haynes, D., Corns, S., Venayagamoorthy, G.K., 2012. An exponential moving average algorithm. In: 2012 IEEE Congress on Evolutionary Computation. IEEE, pp. 1–8.
Huang, X., Wen, D., Li, J., Qin, R., 2017. Multi-level monitoring of subtle urban changes for the megacities of China using high-resolution multi-view satellite imagery. Remote Sens. Environ. 196, 56–75.
Huang, X., Wang, Y., Li, J., Chang, X., Cao, Y., Xie, J., Gong, J., 2020. High-resolution urban land-cover mapping and landscape analysis of the 42 major cities in China using ZY-3 satellite images. Sci. Bull. 65 (12), 1039–1048.
Jabari, S., Rezaee, M., Fathollahi, F., Zhang, Y., 2019. Multispectral change detection using multivariate Kullback-Leibler distance. ISPRS J. Photogramm. Remote Sens. 147, 163–177.
Jiang, J., Lyu, C., Liu, S., He, Y., Hao, X., 2020. RWSNet: a semantic segmentation network based on SegNet combined with random walk for remote sensing. Int. J. Remote Sens. 41 (2), 487–505.
Karra, K., Kontgis, C., Statman-Weil, Z., Mazzariello, J.C., Mathis, M., Brumby, S.P., 2021. Global land use / land cover with sentinel 2 and deep learning. In: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, pp. 4704–4707.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Proces. Syst. 25, 1097–1105.
Kucharczyk, M., Hay, G.J., Ghaffarian, S., Hugenholtz, C.H., 2020. Geographic object-based image analysis: a primer and future directions. Remote Sens. 12, 2012.
Kulkarni, T.D., Whitney, W., Kohli, P., Tenenbaum, J.B., 2015. Deep Convolutional Inverse Graphics Network. Preprint arXiv:1503.03167.
Laine, S., Aila, T., 2016. Temporal Ensembling for Semi-Supervised Learning. Preprint arXiv:1610.02242.
Laine, S., Aila, T., 2017. Temporal Ensembling for Semi-Supervised Learning. Mar. Preprint arXiv:1610.02242.
Lees, L., 2008. Gentrification and social mixing: towards an inclusive urban renaissance? Urban Stud. 45 (12), 2449–2470.
Li, W., Zhou, Y., Cetin, K., Eom, J., Wang, Y., Chen, G., Zhang, X., 2017. Modeling urban building energy use: a review of modeling approaches and procedures. Energy 141, 2445–2457.
Li, W., Chen, C., Zhang, M., Li, H., Du, Q., 2018. Data augmentation for hyperspectral image classification with deep CNN. IEEE Geosci. Remote Sens. Lett. 16 (4), 593–597.
Li, H., Dou, X., Tao, C., Wu, Z., Chen, J., Peng, J., Deng, M., Zhao, L., 2020a. RSI-CB: a large-scale remote sensing image classification benchmark using crowdsourced data. Sensors 20, 1594.
Li, X., Zhou, Y., Chen, W., 2020b. An improved urban cellular automata model by using the trend-adjusted neighborhood. Ecol. Process. 9, 1–13.
Li, X., Wen, C., Cao, Q., Du, Y., Fang, Y., 2021. A novel semi-supervised method for airborne LiDAR point cloud classification. ISPRS J. Photogramm. Remote Sens. 180, 117–129.
Lin, B., Xie, Y., Qu, Y., Li, C., Liang, X., 2018. Jointly Deep Multi-View Learning for Clustering Analysis. Preprint arXiv:1808.06220.
Livesley, S., McPherson, E.G., Calfapietra, C., 2016. The urban forest and ecosystem services: impacts on urban water, heat, and pollution cycles at the tree, street, and city scale. J. Environ. Qual. 45 (1), 119–124.
Lu, Z., Jiang, X., Kot, A., 2018. Deep coupled resnet for low-resolution face recognition. IEEE Signal Process. Lett. 25 (4), 526–530.
Luo, N., Wan, T., Hao, H., Lu, Q., 2019. Fusing high-spatial-resolution remotely sensed imagery and OpenStreetMap data for land cover classification over urban areas. Remote Sens. 11 (1), 88.
Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., Johnson, B.A., 2019. Deep learning in remote sensing applications: a meta-analysis and review. ISPRS J. Photogramm. Remote Sens. 152, 166–177.
Matasci, G., Longbotham, N., Pacifici, F., Kanevski, M., Tuia, D., 2015. Understanding angular effects in VHR imagery and their significance for urban land-cover model portability: a study of two multi-angle in-track image sequences. ISPRS J. Photogramm. Remote Sens. 107, 99–111.
Mathieu, R., Freeman, C., Aryal, J., 2007. Mapping private gardens in urban areas using object-oriented techniques and very high-resolution satellite imagery. Landsc. Urban Plan. 81 (3), 179–192.
Mehta, S., Paunwala, C., Vaidya, B., 2019. CNN based traffic sign classification using adam optimizer. In: 2019 International Conference on Intelligent Computing and Control Systems (ICCS). IEEE, pp. 1293–1298.
Mignot, E., Li, X., Dewals, B., 2019. Experimental modelling of urban flooding: a review. J. Hydrol. 568, 334–342.
Myint, S.W., Gober, P., Brazel, A., Grossman-Clarke, S., Weng, Q., 2011. Per-pixel vs. object-based classification of urban land cover extraction using high spatial resolution imagery. Remote Sens. Environ. 115 (5), 1145–1161.
Myint, B.D., Recktenwald, G.W., Sailor, D.J., 2015a. Thermal footprint effect of rooftop urban cooling strategies. Urban Clim. 14, 268–277.
Myint, S.W., Zheng, B., Talen, E., Fan, C., Kaplan, S., Middel, A., Smith, M., Huang, H.-P., Brazel, A., 2015b. Does the spatial arrangement of urban landscape matter? Examples of urban warming and cooling in Phoenix and Las Vegas. Ecosyst. Health Sustain. 1 (4), 1–15.
Panboonyuen, T., Jitkajornwanich, K., Lawawirojwong, S., Srestasathiern, P., Vateekul, P., 2017. Road segmentation of remotely-sensed images using deep convolutional neural networks with landscape metrics and conditional random fields. Remote Sens. 9 (7), 680.
Panchapagesan, S., Sun, M., Khare, A., Matsoukas, S., Mandal, A., Hoffmeister, B., Vitaladevuni, S., 2016. Multi-task learning and weighted cross-entropy for DNN-based keyword spotting. In: Interspeech, pp. 760–764.
Pilant, A., Endres, K., Rosenbaum, D., Gundersen, G., 2020. US EPA EnviroAtlas meter-scale urban land cover (MULC): 1-m pixel land cover class definitions and guidance. Remote Sens. 12 (12), 1909.
Prestele, R., Alexander, P., Rounsevell, M.D., Arneth, A., Calvin, K., Doelman, J., Eitelberg, D.A., Engström, K., Fujimori, S., Hasegawa, T., 2016. Hotspots of uncertainty in land-use and land-cover change projections: a global-scale model comparison. Glob. Chang. Biol. 22 (12), 3967–3983.
Qiu, C., Schmitt, M., Geiss, C., Chen, T.K., Zhu, X.X., 2020. A framework for large-scale mapping of human settlement extent from Sentinel-2 images via fully convolutional neural networks. ISPRS J. Photogramm. Remote Sens. 163, 152–170.
Rasmus, A., Valpola, H., Honkala, M., Berglund, M., Raiko, T., 2015. Semi-Supervised Learning with Ladder Networks. Preprint arXiv:1507.02672.
Robinson, C., Hou, L., Malkin, K., Soobitsky, R., Czawlytko, J., Dilkina, B., Jojic, N., 2019. Large scale high-resolution land cover mapping with multi-resolution data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12726–12735.
Roy, S., More, R., Kimothi, M., Mamatha, S., Vyas, S., Ray, S., 2018. Comparative analysis of object based and pixel based classification for mapping of mango orchards in Sitapur district of Uttar Pradesh. J. Geom. 12, 69–76.

Safari, K., Prasad, S., Labate, D., 2020. A multiscale deep learning approach for high-resolution hyperspectral image classification. IEEE Geosci. Remote Sens. Lett. 18 (1), 167–171.
Saha, M., Eckelman, M.J., 2017. Growing fresh fruits and vegetables in an urban landscape: a geospatial assessment of ground level and rooftop urban agriculture potential in Boston, USA. Landsc. Urban Plan. 165, 130–141.
Sang, Q., Zhuang, Y., Dong, S., Wang, G., Chen, H., 2020. FRF-net: land cover classification from large-scale VHR optical remote sensing images. IEEE Geosci. Remote Sens. Lett. 17 (6), 1057–1061.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.
Stehman, S.V., 1997. Estimating standard errors of accuracy assessment statistics under cluster sampling. Remote Sens. Environ. 60, 258–269.
Stehman, S.V., Foody, G.M., 2019. Key issues in rigorous accuracy assessment of land cover products. Remote Sens. Environ. 231, 111191.
Sternberg, R.J., 1980. Sketch of a componential subtheory of human intelligence. Behav. Brain Sci. 3 (4), 573–584.
Story, M., Congalton, R.G., 1986. Accuracy assessment: a user's perspective. Photogramm. Eng. Remote. Sens. 52 (3), 397–399.
Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A., 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the AAAI Conference on Artificial Intelligence.
Tapiador, F.J., Avelar, S., Tavares-Corrêa, C., Zah, R., 2011. Deriving fine-scale socioeconomic information of urban areas using very high-resolution satellite imagery. Int. J. Remote Sens. 32 (21), 6437–6456.
Tarvainen, A., Valpola, H., 2017. Mean Teachers are Better Role Models: Weight-Averaged Consistency Targets Improve Semi-Supervised Deep Learning Results. Preprint arXiv:1703.01780.
Topaloğlu, R.H., Aksu, G.A., Ghale, Y.A.G., Sertel, E., 2021. High-resolution land use and land cover change analysis using GEOBIA and landscape metrics: a case of Istanbul, Turkey. Geocart. Int. https://doi.org/10.1080/10106049.2021.2012273.
Torres-Sánchez, J., López-Granados, F., Peña, J.M., 2015. An automatic object-based method for optimal thresholding in UAV images: application for vegetation detection in herbaceous crops. Comput. Electron. Agric. 114, 43–52.
Troyo, A., Fuller, D.O., Calderón-Arguedas, O., Beier, J.C., 2008. A geographical sampling method for surveys of mosquito larvae in an urban area using high-resolution satellite imagery. J. Vector Ecol. 33 (1), 1–7.
Turner, W., Spector, S., Gardiner, N., Fladeland, M., Sterling, E., Steininger, M., 2003. Remote sensing for biodiversity science and conservation. Trends Ecol. Evol. 18 (6), 306–314.
USGS, 2020. Earth Explorer. http://earthexplorer.usgs.gov. Last accessed on June 26, 2020.
Venter, Z.S., Barton, D.N., Gundersen, V., Figari, H., Nowell, M., 2020. Urban nature in a time of crisis: recreational use of green space increases during the COVID-19 outbreak in Oslo, Norway. Environ. Res. Lett. 15 (10), 104075.
Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A., Bottou, L., 2010. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. 11, 3371–3408.
Wang, W., Arora, R., Livescu, K., Bilmes, J., 2015. On deep multi-view representation learning. In: International Conference on Machine Learning. PMLR, pp. 1083–1092.
Wang, B., Choi, J., Choi, S., Lee, S., Wu, P., Gao, Y., 2017. Image fusion-based land cover change detection using multi-temporal high-resolution satellite images. Remote Sens. 9 (8), 804.
Weigand, M., Wurm, M., Dech, S., Taubenböck, H., 2019. Remote sensing in environmental justice research—a review. ISPRS Int. J. Geo Inf. 8 (1), 20.
Whiteman, A., Gomez, C., Rovira, J., Chen, G., McMillan, W.O., Loaiza, J., 2019. Aedes mosquito infestation in socioeconomically contrasting neighborhoods of Panama city. EcoHealth 16 (2), 210–221.
Wu, Y., Zhang, P., Wu, J., Li, C., 2021. Object-oriented and deep-learning-based high-resolution mapping from large remote sensing imagery. Can. J. Remote. Sens. 47 (3), 396–412.
Yu, X., Wu, X., Luo, C., Ren, P., 2017. Deep learning in remote sensing scene classification: a data augmentation enhanced convolutional neural network framework. GISci. Remote Sens. 54 (5), 741–758.
Yu, L., Wang, S., Li, X., Fu, C.-W., Heng, P.-A., 2019. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 605–613.
Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J., 2020. Deep learning in environmental remote sensing: achievements and challenges. Remote Sens. Environ. 241, 111716.
Zhang, K., Batterman, S., 2013. Air pollution and health risks due to vehicle traffic. Sci. Total Environ. 450, 307–316.
Zhang, X., Li, H., 2018. Urban resilience and urban sustainability: what we know and what do not know? Cities 72, 141–148.
Zhang, P., Gong, M., Su, L., Liu, J., Li, Z., 2016. Change detection based on deep feature representation and mapping transformation for multi-spatial-resolution remote sensing images. ISPRS J. Photogramm. Remote Sens. 116, 24–41.
Zhang, D., Liu, J., Zhu, H., Liu, Y., Wang, L., Wang, P., Xiong, H., 2019. Job2Vec: Job title benchmarking with collective multi-view representation learning. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 2763–2771.
Zhang, Y., Chen, G., Vukomanovic, J., Singh, K.K., Liu, Y., Holden, S., Meentemeyer, R. K., 2020. Recurrent shadow attention model (RSAM) for shadow removal in high-resolution urban land-cover mapping. Remote Sens. Environ. 247, 111945.
Zhao, W., Du, S., 2016. Spectral–spatial feature extraction for hyperspectral image classification: a dimension reduction and deep learning approach. IEEE Trans. Geosci. Remote Sens. 54 (8), 4544–4554.
Zhao, W., Guo, Z., Yue, J., Zhang, X., Luo, L., 2015. On combining multiscale deep learning features for the classification of hyperspectral remote sensing imagery. Int. J. Remote Sens. 36 (13), 3368–3379.
Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890.
Zhao, W., Bo, Y., Chen, J., Tiede, D., Blaschke, T., Emery, W.J., 2019. Exploring semantic elements for urban scene recognition: deep integration of high-resolution imagery and OpenStreetMap (OSM). ISPRS J. Photogramm. Remote Sens. 151, 237–250.
Zhu, X., 2005. Semi-Supervised Learning with Graphs. Ph.D. Thesis. Department of Psychology, Carnegie Mellon Univ., Pittsburgh, PA.