# Very fine spatial resolution urban land cover mapping using an explicable sub-pixel mapping network based on learnable spatial correlation

Da He[a,b], Qian Shi[a,b,*], Jingqian Xue[a], Peter M. Atkinson[c], Xiaoping Liu[a,b]

[a]School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China
[b]Guangdong Provincial Key Laboratory for Urbanization and Geo-simulation, Guangzhou 510275, China
[c]Faculty of Science and Technology, Lancaster University, Lancaster LA1 4YR, UK

**Abstract:**

Sub-pixel mapping is the prevailing approach for dealing with the mixed pixel effect in urban land use/land cover classification, by reconstructing the sub-pixel-scale distribution inside each mixed-pixel based on spatial autocorrelation. However, 1) traditional spatial autocorrelation is limited to a local window, which cannot model the teleconnection between two locations or objects that are far apart and 2) autocorrelation is based on the idea of "the more proximate, the more similar", which relies on a distance-weight decay parameter and cannot characterize the rich variety of mutual information in spatially heterogenous areas in urban. In this research, we develop and demonstrate a learnable correlation-based sub-pixel mapping (LECOS) method. 1) We use the "mutual retrieval" mechanism of the self-attention operation to model teleconnections that enable more distant locations or objects to be mutually correlated and 2) we design a parameter-free "self-attention in self-attention" operation to learn

* Corresponding author: shixi5@mail.sysu.edu.cn

adaptively the diverse global correlation patterns between pixel and sub-pixel. The learned spatial correlations are then used for reasoning the sub-pixel-scale distribution of each class. We validated our method on the most challenging public datasets of urban scenes, which exhibit considerable spatial heterogeneity with complex structures and broken objects. The learned building-tree, building-road and road-tree correlation patterns contributed most to the sub-pixel reconstruction result of the urban scenes, consistent with *in-situ* reference data. We further explored the model's explicability in a large-area of several metropolises in China, by mapping land cover in these cities at a 2 m very fine spatial resolution using 10 m Sentinel-2 input images, and found that the derived result not only revealed rich urban spatial heterogeneity, but also that the learned correlation was indicative of urban pattern dynamics, suggesting the potential for greater understanding of issues such as urban fairness, accessibility, human exposure and sustainability.

Keywords: land use/land cover classification, urban spatial pattern, sub-pixel mapping, spatial teleconnection, self-attention mechanism

## 1. Introduction

Urban land use/land cover can provide fundamental information for understanding the impact of changes in urban composition on the urban heat island phenomenon (Xia et al., 2022), air pollution exposure (IPCC Climate Change 2013), carbon losses (Foley et al., 2005) and urban sustainability (Stokes and Seto, 2018). Accurate urban land use/land cover monitoring and the corresponding regulation can alleviate the impact of human activities on the climate, land surface and biodiversity (Liu et al., 2020; Liu et al., 2019). However, urban scenes are highly dynamic and heterogenous, which makes the widely used moderate spatial resolution remote sensing images (e.g., Landsat/Sentinel-2) unsuitable for fine-grained monitoring, since the edges of buildings, green parcels and other small-sized urban components will inevitably cut through pixels, leading to a jagged boundary or even disappearance at a coarse spatial resolution, known as the mixed pixel effect. Traditional hard classification methods that assign each pixel to a single label may result in small-sized and elongated objects being missed. Although many high-resolution images are available now, such as World-View series or GF series satellite, they generally have a long revisiting period and narrow width, and are often obscured by clouds with year-round rainfall in southern China. It is difficult to mosaick a complete city-level or provincial-level image without cloud contamination, which is difficult to meet the needs of periodic continuous monitoring.

Researchers have addressed the mixed pixel problem with the spectral unmixing technique. Instead of hard classification, spectral unmixing aims to decompose each mixed pixel into a combination of multiple pure land covers weighted by their proportions (Adam et al., 1995), thereby, estimating the percentage of each land cover within each mixed pixel. Spectral unmixing has been applied widely to derive many land cover fractions, such as fractional forest cover (Rashed et al., 2003; Small, 2003; Xiao and Moody, 2005) and fractional urban impervious cover (Wu and Murray, 2003; Deng et al., 2019). However, the spatial resolution of the fractional images is still coarse, providing no indication of how each pure land cover is located inside each mixed pixel

(Atkinson, 2009).

Sub-pixel mapping (SPM) is an effective method for estimating the location of each pure land cover inside each mixed pixel, which is equivalent to improving the observation resolution. It has been applied widely for the identification of small-sized objects that would otherwise require a very fine resolution, for example, 0.6 m spatial resolution for urban tree reconstruction from 2.4 m QuickBird MS images (Ardila et al., 2011), 2 m resolution for urban tree reconstruction from 10 m Sentinel-2A MSI images (He et al., 2022a), and 0.5 m resolution for individual potato plant reconstruction from 2 m WorldView-2 images (Poudyal, 2013).

The SPM grid divides each mixed pixel into several sub-pixels, and allocates land cover labels (derived from spectral unmixing) to each sub-pixel location to realize a finer land cover map. Considering that the allocation process is ill-posed, a spatial prior is necessary to describe the spatial correlation of each sub-pixel and constrain the SPM solution. According to the formulation of the prior, SPM can be categorized into three types:

1) Spatial dependence prior. This assumes that the adjacent sub-pixel/pixel locations are likely to belong to the same class, and each sub-pixel's label can be predicted according to the abundances of its eight-neighboring mixed-pixels or sub-pixels, such as the sub-pixel swapping model (Atkinson, 2005), spatial attraction model (Mertens et al., 2006), genetic algorithm based SPM (He et al., 2016a), Hopfield neural network based SPM (Su, 2019) and random field based SPM (Kasetkasem et al., 2005).

2) Geostatistical prior. This assumes that the sub-pixel distribution is subject to

empirical geostatistical models that can be regressed with given samples (Boucher and Kyriakidis 2006), such as area-to-point kriging-based SPM (Wang et al., 2015) and spatial distribution pattern-based SRM (Ge et al. 2016).

3) Regularization prior. This assumes that handcrafted filters like the Laplacian model, non-local model or sparse assumption can regularize the sub-pixel distribution, which includes maximum *a posteriori* model-based SPM (Zhong et al., 2015), sparse representation model-based SPM (Song et al., 2019), spectral-spatial fusion-based SPM (Xu et al. 2018), and spatial-temporal-spectral fusion-based SPM (Li et al., 2017). However, the above model-driven spatial priors are essentially based on the idea that "more proximate data points are expected to be more similar", which is not so readily extensible to frequent spatial variation as commonly encountered in urban scenarios.

Deep learning theory is an alternative method that can alleviate the requirement for an empirical spatial prior, focused on "allowing the data to speak for themselves" (Ling et al., 2019; He et al., 2021a). Deep learning approaches assume that the prior can be learned from exemplar pairs of coarse resolution (CR) image and fine resolution (FR) annotation images in actual geographical scenarios. Deep learning-based SPM generally uses the convolution and deconvolution operations to reconstruct an FR result, and a FR ground reference is compared with the estimates and the gradient error is backpropagated to update the network parameters, and learn the underlying spatial correlation. Many prevalent networks were applied for SPM, such as the super-resolution reconstruction network (Ling et al. 2019a; 2019b; Ma et al. 2019; He et al. 2021a; Shang et al. 2020), semantic segmentation network (Arun et al. 2018; Jia et al.

2019), attention mechanism-based network (He et al. 2021b; 2022), graph convolution network (Zhang et al. 2021b) and spatiotemporal fusion network (Chen et al. 2021b). However, the convolutional layer usually has a limited receptive field due to the local connection property (e.g., 3×3 kernel size), which cannot learn the non-local spatial correlation between urban compositions that is essential for sub-pixel location reasoning (Zhang et al., 2019; Verdonck et al., 2017). Besides this, the learning process of the convolution network cannot explain what spatial correlation pattern the network has learned, and how it guides inference of the sub-pixel-scale location.

Based on the above, two challenges follow: 1) all the above spatial correlation approaches are limited to a local region and, thus, cannot exploit the teleconnections that are more characteristic of spatial heterogeneity (e.g., buildings and trees repeatedly occur at a certain distance, exhibiting a correlation between each other remotely (**Fig. 1**)). Although non-local convolution was developed to capture the global context information by measuring each point of the feature map with the weighted average of the other points (Wang et al., 2018), and dilated convolution can also expand the receptive field of the traditional kernels by injecting blank pixels into 3×3 kernels to achieve larger kernel size, while maintaining the calculation efficiency (Fisher et al., 2015). However, they only aim to enlarge the receptive field, and are still limited to convolution framework, i.e., the learning process of the kernel parameters is uncontrollable, which is difficult to explain the meaning of the learned parameters; 2) classical spatial autocorrelation relies on the empirical distance decay assumption, which cannot characterize the various sources of mutual information between urban components, while the implicitly learned correlation in a convolutional network cannot provide a reasonable explanation and, thus, is unreliable.

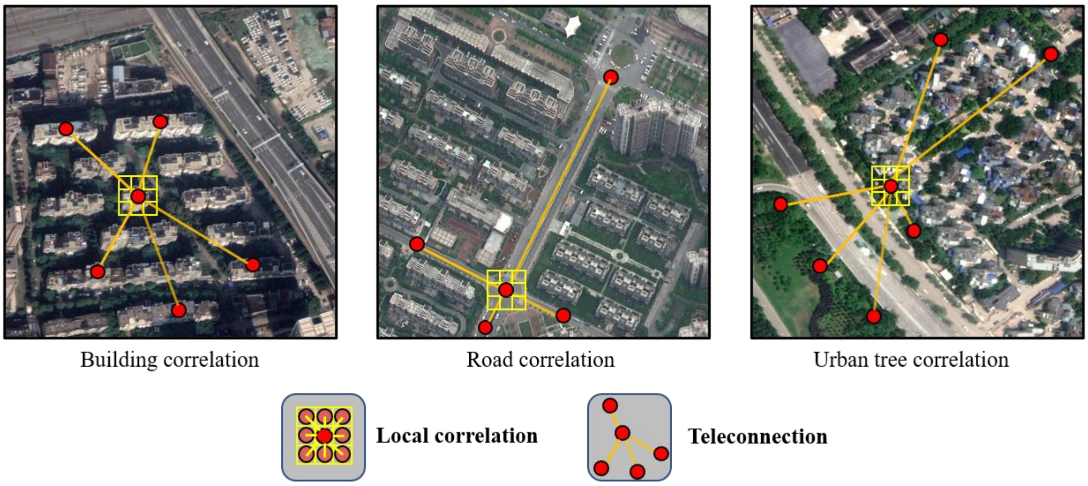


**Fig. 1**. Comparison of local correlation and teleconnection between objects in urban an scenario.

To overcome the above two challenges and reveal urban patterns at high-resolution for planning or decision-making, we combine the learning ability of the data-driven idea with the spatial correlation modeling process, and develop a learnable correlation based sub-pixel mapping network (LECOS). 1) We use the "mutual retrieval" mechanism of the self-attention operation to excavate the diverse contextual correlation patterns, which models the global teleconnection between objects like trees, buildings, and roads that are far apart; 2) The mutual retrieval process is established in the end-to-end network architecture, which enables the spatial correlation patterns to be learned explicitly through data-driven regression; 3) Considering that the typical self-attention operation cannot establish the correlation (between pixels and sub-pixels) that is essential for sub-pixel location inference, we further designed a "self-attention in self-attention (SNS)" operation to enable exploration of the hierarchical correlations between the two scales; 4) Since self-attention operation explicitly learns the spatial dependency between each pixel and sub-pixel in the global scene, so that it can figure out, in a given image patch, which object is related to which one, therefore, the learned spatial correlation is explicable. Besides, the learned correlation was found to be able to quantitatively reflect the urban spatial patterns, such as accessibility of greenspace to impervious area, and was used to evaluate urban environmental planning and resource distribution of several main cities in China.

The aims of this paper were to 1) develop a learnable correlation-based sub-pixel mapping network architecture for reconstructing a fine-grained urban distribution with rich spatial heterogeneity; 2) validate the proposed method on public urban scenario datasets, examining its performance in reconstructing spatial detail and the underlying correlation; 3) evaluate the method over large areas across several main metropolises in China, and reveal our findings on urban spatial patterns as reflected by the correlation between each urban composition at very fine spatial resolution, and validate the findings with multi-source products.

## 2. Method

### *2.1 Foundation of spatial correlation modeling in sub-pixel mapping*

In model-driven SPM methods, given a CR image, the spectral unmixing approach is firstly used to decompose each mixed-pixel into pure land cover classes (i.e., endmember) and estimate their pixel-level proportions. Then, each mixed-pixel is divided into $S \times S$ sub-pixels ($S$ is the scale factor), and a possible label is assigned to each sub-pixel location, based on the spatial correlation among the mixed-pixels and the sub-pixels within the neighborhood. **Fig. 2** demonstrates the specific estimation process with an example in a 3×3 mixed-pixel area.

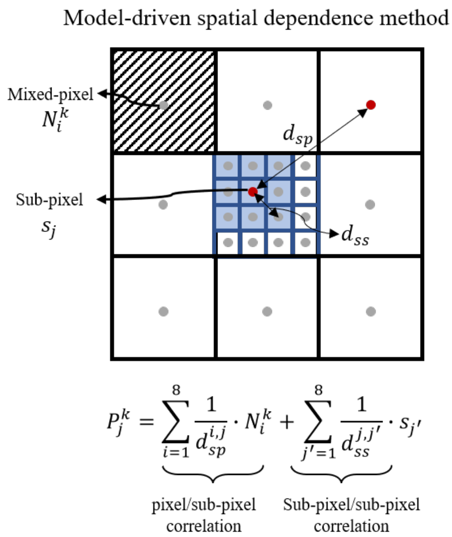


**Fig. 2**. The formulation of the model-driven spatial correlation modeling.

Supposing there are *NC* endmembers, the eight-neighborhood mixed-pixel is denoted as $N_i,\ i = 1 \ldots 8$, and the proportion of the $kth$ endmember in the mixed-pixel $N_i$ is denoted as $N_i^k, k = 1 \ldots NC$. The sub-pixel in the center mixed-pixel area is denoted as $s_j, j = 1 \ldots S^2$. The probability of assigning the $kth$ endmember label to the $jth$ sub-pixel (i.e., $P_j^k$) can be estimated by the proportions of the $kth$ endmember in each mixed-pixel and sub-pixel within the eight-neighborhood, as well as the spatial correlation between them, which can be given as:

$$P_j^k = \sum_{i=1}^{8} \frac{1}{d_{sp}^{i,j}} \cdot N_i^k + \sum_{j'=1}^{8} \frac{1}{d_{ss}^{j,j'}} \cdot s_{j'} \tag{1}$$

where the spatial correlation is usually simplified to the inverse distance weighting in Eq. (1), that is, $d_{sp}^{i,j}$ represents the distance between the $ith$ mixed-pixel and the $jth$

sub-pixel, and $d_{ss}^{j,j'}$ represents the distance between the $jth$ sub-pixel and the $j'th$ sub-pixel in the eight-pixel neighborhood. After deriving all the class probabilities, the class $k$ corresponding to the largest probability $P_j^k$ will be assigned to the $jth$ sub-pixel location.

However, the spatial correlation is variable in practice, and may not be characterized sufficiently by the simple inverse distance assumption. Deep learning is a potential way of learning diverse spatial correlation patterns from exemplar datasets. However, how to model the spatial correlation in an end-to-end network remains an open question.

*2.2 Learnable correlation sub-pixel mapping*

To make full use of the learning ability of the data-driven idea and the correlation modelling ability of the model-driven idea, this research developed a learnable correlation sub-pixel mapping (LECOS) method. Instead of using convolution operation, the LECOS models the global spatial correlation between sub-pixel and mixed-pixel based on the "mutual retrieval" mechanism of the self-attention operation in an end-to-end network structure, to enable the spatial correlation learnable. Furthermore, considering that the typical self-attention operation can only establish the single pixel-level correlation, we further designed a "self-attention in self-attention" (SNS) operation to establish explicitly the spatial correlation rule between pixels and sub-pixels in a feed forward network fashion, which can be iteratively rectified by data-driven, back-propagation learning. Finally, the learned spatial correlation rule is used to infer the class label at each sub-pixel location.

The LECOS consists of the preparation work, encoder part and decoder part (**Fig. 3**). In the preparation work stage, *inductive bias* is designed to divide each mixed-pixel of the input CR image into sub-pixels at the beginning of the network, which analogizes the practice of the model-driven SPM that enables the spatial correlation calculation at each sub-pixel location. Then, *image tokenization* is designed to convert each mixed-pixel and sub-pixel into a 1-dimensional sequential mixed-pixel token and sub-pixel token, which aims to accommodate the self-attention operation. In the encoder part, a 4-stage multi-scale residual structure is designed to consider the large scale-variation of the urban compositions, which analogizes the structure of the typical ResNet (He et al., 2016b). Each stage is composed of a stacked layer of *syntax-dependence builder* and *context-dependence builder* based on the designed SNS operation, which is the main part of the LECOS for learning the spatial correlation rule amongst the mixed-pixels and sub-pixels. In the decoder part, the output features of the 4-stage are reformed from sequential token to a 2-dimensional feature map, and are integrated and recovered to the target spatial resolution. Finally, a classifier is used to estimate the class probability from the integrated feature maps.

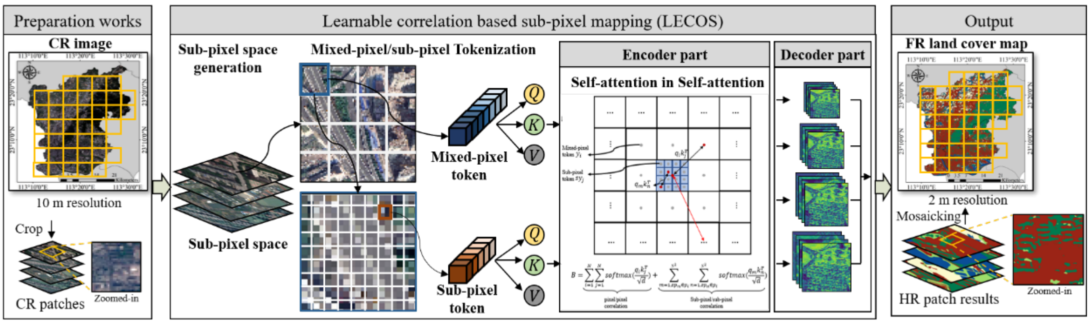

**Fig. 3**. Overall architecture of LECOS.

*2.2.1 Sub-pixel space generation*

Given an CR image $Y \in R^{h \times w \times c}$, it is upsampled by a scale factor $S$ (denoted as $Y' \in R^{H \times W \times c}$) in the first layer of the network to generate the sub-pixel space, which is equivalent to the division of the mixed-pixel into sub-pixels in the classical spatial dependence method (**Fig. 4**). In this way, the sub-pixels can be manipulated explicitly in the subsequent "mutual retrieval" process to explore the potential correlation. This design is different from the conventional learning-based SPM that usually deploys the upsampling in the last layer, which is suitable to the purpose of establishing the correlation at each sub-pixel location.

*2.2.2 Mixed-pixel/sub-pixel tokenization*

Image tokenization is constructed to transform the image into sequential tokens (**Fig. 4**). Specifically, the input image is grid partitioned to patches $p_i \in R^{S \times S \times ch}, i = 1 \ldots N$ with size of $S \times S$ (where $N = H \times W / S \times S$, $ch$ is the channel number), which is equivalent to the area of one mixed-pixel. Each patch is then flattened to a one-dimension vector to generate a mixed-pixel token $y_i \in R^{d \times 1}, i = 1 \ldots N$ (where $d = S \times S \times ch$). Each sub-pixel $s_{i,m} \in R^{ch \times 1}, m = 1 \ldots S^2$ in the patch $p_i$ can be also viewed as a 1-dimensional vector, which forms a sub-pixel token $sy_{i,m} \in R^{ch \times 1}, m = 1 \ldots S^2$.

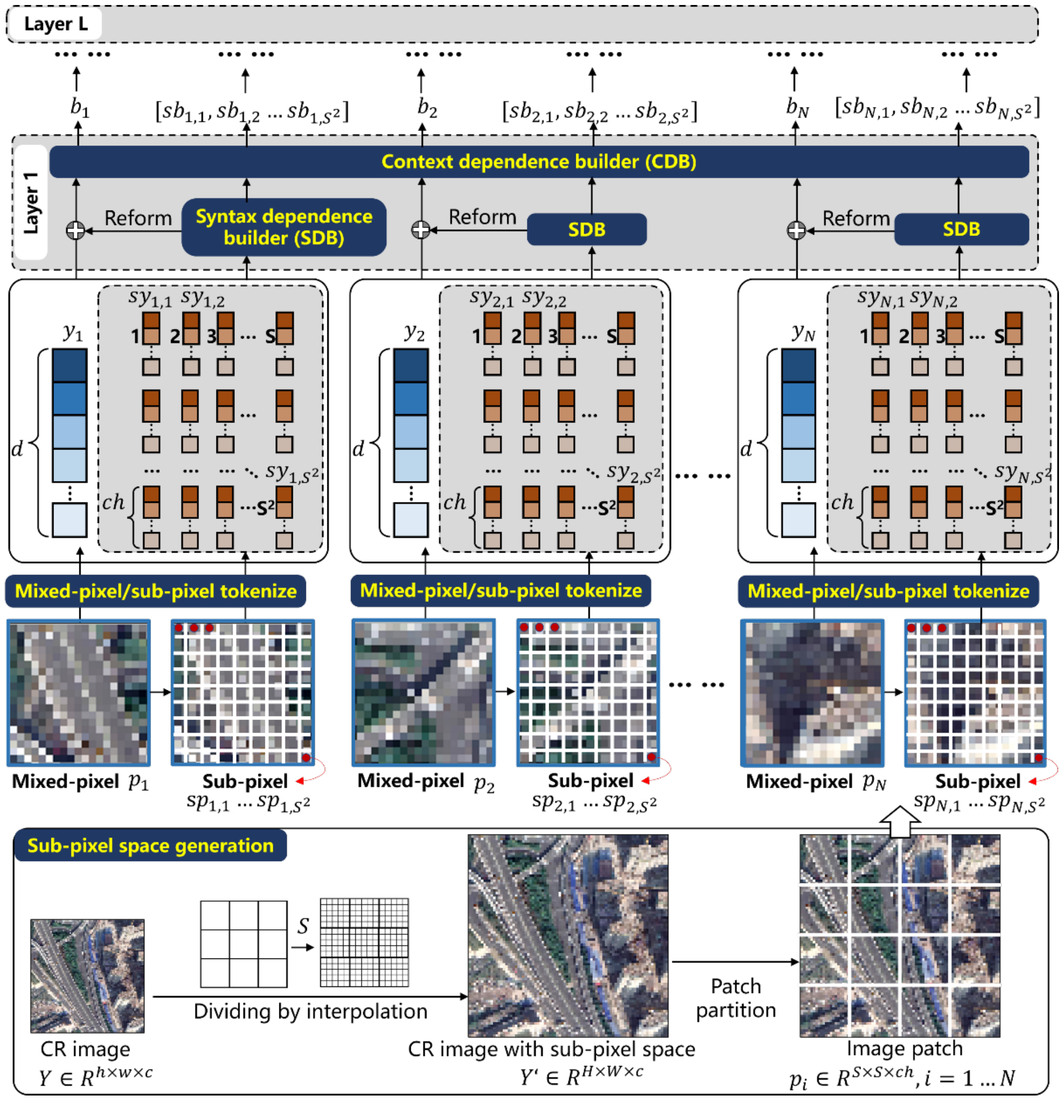


**Fig. 4**. Demonstration of the specific process in the preparation work.

### *2.2.3 Self-attention in self-attention (SNS) operation*

#### *2.2.3.1 Syntax and context in urban scenario*

How to establish the hierarchical correlation amongst mixed-pixels and sub-pixels in the network is key to our method. In this research, we constructed a multi-scale observation framework inspired by the language principle. Specifically, sub-pixels can be regarded as words, which can be aggregated to form a mixed-pixel (sentence) based on syntax. Syntax can indicate the location of the sub-pixels inside each mixed-pixel, such as linear distribution, sparse distribution, etc., according to the morphological shape of the object. On the other hand, mixed-pixels can be aggregated to form a scene (paragraph) based on context, which indicates the distribution rule of the mixed-pixel in the scene, that is, trees located along the road, plane located inside the airport, etc.

An intuitive example of how context and syntax work in the urban scene is provided in **Fig. 5.** In **Fig. 5a**, for example, it is difficult to recognize the highlighted object with its context masked, which could be mistaken for a road. When providing the context information, such as grassland and neatly arranged buildings, it is easy to determine that here is a residential area, and the highlighted object is the pavement. On consideration that the mixed-pixel is too coarse to describe the tiny pavement, the syntax information can be used to determine the sub-pixel location at the place of intersection, according to the morphological shape of each object, and guide the network to reconstruct linearly distributed sub-pixels for the jagged pavement (**Fig. 5b**).

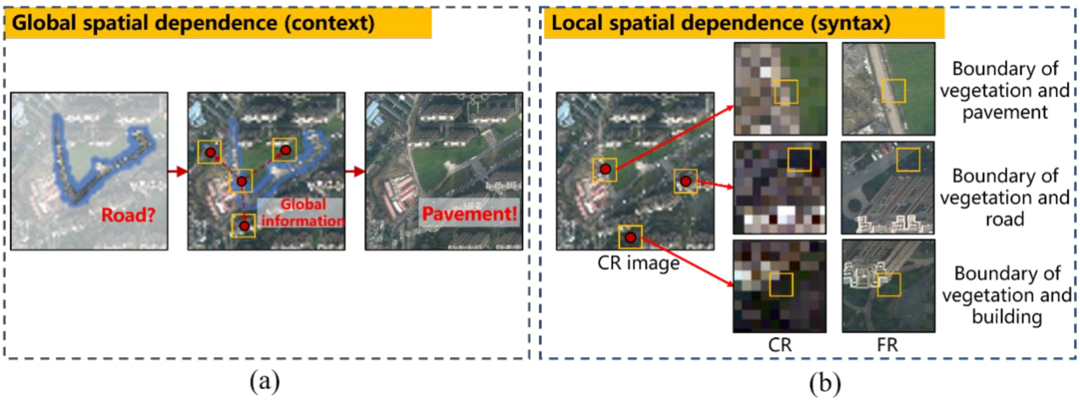


**Fig. 5**. Demonstration of the role of context and syntax in inferring urban compositions.

*2.2.3.2 Syntax- and context-dependence builder*

To learn the syntax and context, respectively, for mixed-pixels and sub-pixels, we designed a syntax-dependence builder (SDB) and context-dependence builder (CDB) (**Fig. 6**), using the self-attention in self-attention (SNS) operation. The SDB aims to learn the syntax between sub-pixels, which represents local features, such as morphological shape, detail edge, etc. After the syntax is learned, the sub-pixels are aggregated by syntax to formulate a complete mixed-pixel (sentence), and the CDB is used to learn the context amongst mixed-pixels. Finally, all the mixed-pixels are aggregated by context to formulate a complete scene.

The self-attention operation is the critical technique to build the relationship amongst mixed-pixel tokens and sub-pixel tokens, inspired by the human vision system that is able to guide attention towards interested objects through “mutual retrieval” between two components within a scene. For example, given an object to be retrieved as a *Query* (e.g., pavement), the pavement-related objects (e.g., grassland, building, etc.) in the scene are taken as *Key*, and the intrinsic features of each object are taken as *Value*. The self-attention operation builds the relationship between *Key* and *Query*, to measure the correlation intensity of each component related to the pavement, and the intensity is used as the attention weight to rescale the *Value* of all the objects, which are finally aggregated to obtain the global feature embedded with correlation information. However, the classical self-attention operation cannot establish the correlation between pixel-level and sub-pixel-level, which inspired us to design the SNS operation to support SDB and CDB.

The architectures of the SDB and the CDB are shown in **Fig. 6**. Taking the first layer of the encoder part as an example, given the mixed-pixel token $y_i \in R^{d \times 1}, i =$

$1 \dots N$ and the sub-pixel token $sy_{i,m} \in R^{ch \times 1}, m = 1 \dots S^2$ derived from image tokenization, the SNS operation is used to investigate the context correlation of the $ith$ token $y_i$ with the $jth$ token $y_j$ ($j = 1 \dots N$), and the syntax correlation of the $mth$ token $sy_{i,m}$ with the $nth$ token $sy_{i,n}$ ($n = 1 \dots S^2$). To enable the sub-pixels to retrieve each other, and to be retrieved, the *Query* ($SQ = \{sq_{i,m}\} \in R^{ch \times S^2}$), *Key* ($SK = \{sk_{i,m}\} \in R^{ch \times S^2}$) and *Value* ($SV = \{sv_{i,m}\} \in R^{ch \times S^2}$) of the sub-pixel token $sy_{i,m}$ within $y_i$, are generated by linear transformation of the $y_i$ and $sy_{i,m}$, respectively:

$$\begin{aligned} sk_{i,m} &= sW^k \cdot sy_{i,m}, \\ sq_{i,m} &= sW^q \cdot sy_{i,m}, \\ sv_{i,m} &= sW^v \cdot sy_{i,m}, \qquad m = 1 \dots S^2 \end{aligned} \tag{2}$$

where $W^q, W^k, W^v \in R^{d \times d}$ , $sW^k, sW^k, sW^k \in R^{c \times c}$ represent the transformation matrices of *Query*, *Key* and *Value*, respectively.

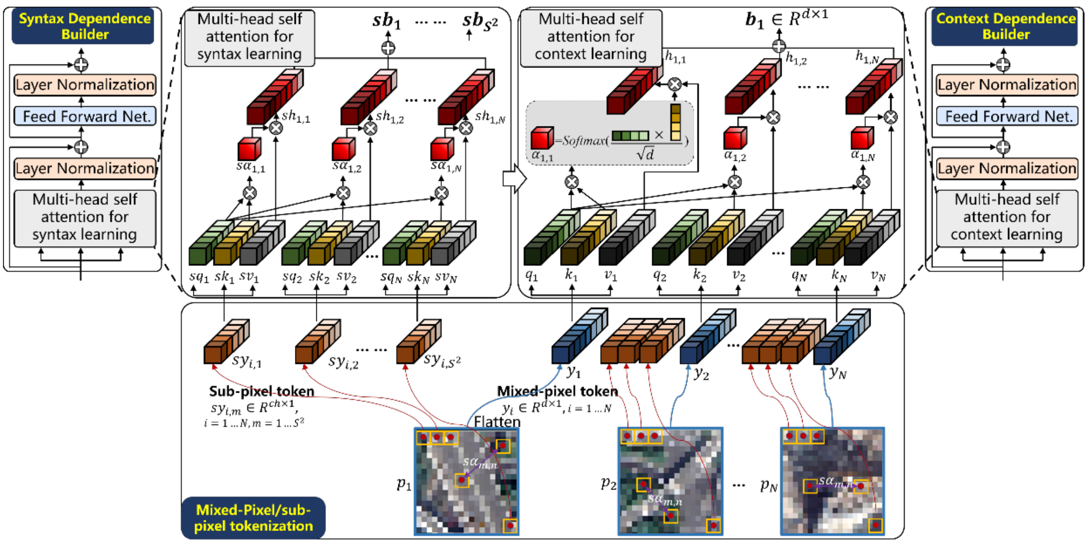


**Fig. 6**. Syntax-dependence and context-dependence builder based on self-attention in self-attention operation.

Firstly, the *Query* of the sub-pixel $sy_{i,m}$ is taken to retrieve the *Key* of the global sub-pixel $sy_n$ to build the syntax correlation $s\alpha_{i,m,n} \in R^{1\times 1}$, which can be given as:

$$s\alpha_{i,m,n} = \sum_{m=1, sy_{i,m}\in y_i}^{S^2} \sum_{n=1, sy_{i,n}\in y_i}^{S^2} softmax(\frac{q_{i,m}k_{i,n}^T}{\sqrt{ch}}) \tag{3}$$

where $\sqrt{ch}$ normalizes the $q_{i,m}k_{i,n}^T$.

The derived syntax correlation is used to rescale the *Value* of each sub-pixel token $sy_{i,n}$ within the $y_i$, and the output syntax feature of $sy_{i,m}$ is obtained by summarizing all the rescaled *Values*:

$$sb_{i,m} = \sum_{n=1}^{S^2} sv_{i,m} \cdot s\alpha_{i,m,n}, m = 1 \ldots S^2 \tag{4}$$

where $sb_{i,m} \in R^{c\times 1}$ is the derived syntax feature of the sub-pixel token $sy_{i,m}$.

The above syntax-dependence modeling process can be given in matrix form:

$$SB_i = \mathbf{SDB}\left(\{sy_{i,m}\}_{m=1\ldots S^2}\right) = SV \cdot SoftMax(SK^T \cdot SQ/\sqrt{ch}) \tag{5}$$

where $SB_i \in R^{ch\times S^2}$ is the aggregation of the sub-pixel token rescaled by the syntax.

Then, the $SB_i$ are flattened to a 1-dimensional vector with size $(ch \times S^2) \times 1$, and propagated to the mixed-pixel token $y_i$ by summarizing:

$$y'_i = y_i + flatten(SB_i) \tag{6}$$

In the same way, the *Query* ($Q = \{q_i\} \in R^{d \times N}$), *Key* ($K = \{k_i\} \in R^{d \times N}$), and *Value* ($V = \{v_i\} \in R^{d \times N}$) of the pixel token $y'_i$ is firstly generated:

$$\begin{aligned} q_i &= W^q \cdot y'_i, \\ k_i &= W^k \cdot y'_i, \\ v_i &= W^v \cdot y'_i, \qquad i = 1 \ldots N \end{aligned} \tag{7}$$

The context correlation $\alpha_{i,j} \in R^{1 \times 1}$ is also derived by the self-attention operation:

$$\alpha_{i,j} = \sum_{i=1}^{N} \sum_{j=1}^{N} softmax(\frac{q_i k_j^T}{\sqrt{d}}) \tag{8}$$

And the output features of the mixed-pixel token $y_i$ that are embedded with the syntax- and context- correlation can be obtained:

$$\hat{B} = \mathbf{CDB}(\{y'_i\}_{i=1 \ldots N}) = V \cdot SoftMax(K^T \cdot Q / \sqrt{d}) \tag{9}$$

where $\hat{B} \in R^{d \times N}$ represents the ensemble of the mixed-pixel token that is organized by context.

Noteworthy is that all the $s\alpha_{i,m,n}$ can be formulated as an attention matrix $SA \in R^{S^2 \times S^2}$, which is the desired syntax-dependence, and all the $\alpha_{i,j}$ can be formulated as a matrix $A \in R^{N \times N}$, which is the desired context-dependence. Both the syntax and context are visualized in the experimental results (Section 4.1), to demonstrate explicitly the learned correlation rule in the urban scenario. From this point of view, the process of CDB and SDB is consistent with the spatial dependence modelling practice of the classical model-driven SPM algorithm, both of which explicitly explore the optimal spatial configuration at the sub-pixel-scale. Furthermore, benefitting from the global “mutual retrieval” ability of SNS operation, the range of the correlation is no longer limited to the eight-neighborhood, but the whole scene, as shown in **Fig. 3**, and the Eq. (1) can be reformed to:

$$\hat{B} = \sum_{i=1}^{N}\sum_{j=1}^{N}(softmax(\frac{q_i k_j^T}{\sqrt{d}}) + \sum_{m=1,sp_m \in p_i}^{S^2}\sum_{n=1,sp_n \in p_i}^{S^2} softmax(\frac{q_m k_n^T}{\sqrt{d}})) \tag{10}$$

To ensure robust training and accelerate convergence, a layer normalization (LayerNorm) operation is adopted to normalize the derived syntax $SB_i$ and context $B_i$ to a Gaussian distribution with zero mean and variance 1 (as shown in **Fig. 6**), given as:

$$LayerNorm(B_i) = g_{ln} \times \frac{\hat{B}_i - \mu(\hat{B}_i)}{\sigma(B_i)} + b_{ln} \tag{11}$$

where $g_{ln}$ is the gain parameter, and $b_{ln}$ is the bias, which are also trainable.

A feed forward network (FFN) is then used for each sequential token feature to enhance the fitting capability, which adopts the multiple layer perceptron (MLP), as was done by Vaswani et al. (2017).

$$\mathrm{FFN}(B_i) = W_f \otimes LayerNorm(\hat{B}_i) + b_f \tag{12}$$

where $W_f$ and $b_f$ represent the $1 \times 1$ kernel weight and bias, and $\otimes$ means the convolution operation.

In summary, the above process can be reorganized by a residual connection strategy:

$$\hat{B} = Y + \mathbf{CBD}(\mathbf{SBD}(Y)) \tag{13}$$

$$B = \hat{B} + \mathrm{FFN}(\hat{B}) \tag{14}$$

Note that the size of the output features $SB$ and $B$ equal the input sub-pixel token and mixed-pixel, that is, $ch \times S^2$ and $d \times N$, which can be used as the input for the next layer of SDB and CDB, and enable the SNS layers to be stacked to form an end-to-end feed forward architecture.

*2.2.4 Encoder and decoder structure*

Considering the large scale-variation of urban compositions, a 4-stage multi-scale structure is designed in the encoder part to learn the hierarchical representations of the syntax and context correlation features (**Fig. 7**), which is analogous to the structure of the typical ResNet (He et al., 2016b). Each stage is composed of stacked layers of SBD and CBD. The patch merging strategy is designed between each stage to generate multi-scale features. After that, the output features of the 4-stage are aggregated in the decoder part, and a *Softmax* classifier is adopted to generate the classification result.

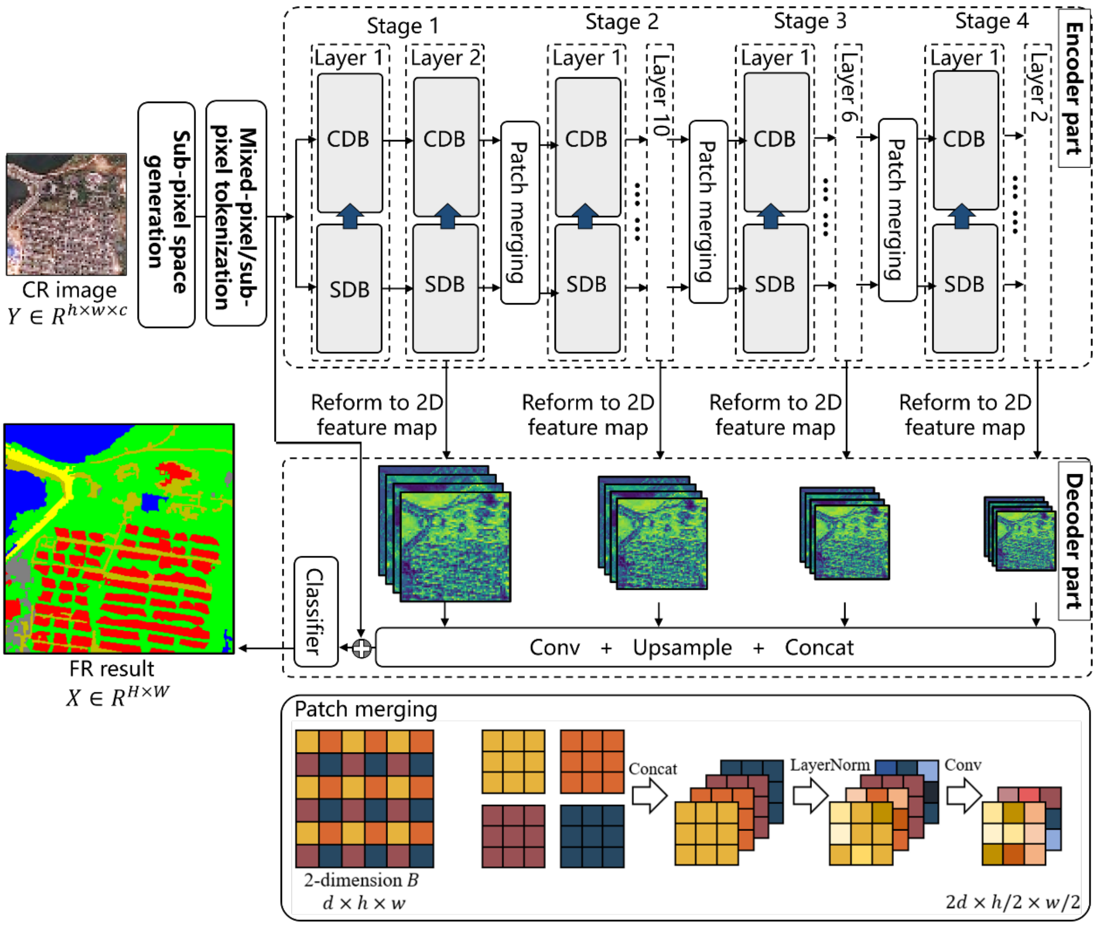


**Fig. 7**. The network architecture of LECOS.

*2.2.4.1 Patch merging*

Patch merging is used to create multi-scale features for each stage. Taking the output feature of $B$ from stage-1 as an example, a reshape operation is firstly applied to resize $B$ from 1-dimension $d \times N$ to 2-dimensions $d \times h \times w$. Then, the 2-dimensional feature maps are split into four equal parts according to **Fig. 7**, and are concatenated along with the channel dimension, equivalent to a downsampling process. Finally, the derived 2-dimensional feature maps are again flattened to a 1-dimensional vector for the next stage. In this way, every time the $B$ pass a stage, the height and width are reduced to half, and the channel dimensions are doubled, that is, $F1 \in R^{H \times W \times C}, F2 \in R^{H/2 \times W/2 \times 2C}, F3 \in R^{H/4 \times W/4 \times 8C}, F4 \in R^{H/8 \times W/8 \times 16C}$. The reason for this design is that the distance between two highly correlated urban compositions is various, which cannot be fully captured by a fixed patch size, and an ablation experiment is also designed to validate the multi-scale architecture (refer to the supplementary file).

*2.2.4.1 Feature integration*

In the decoder part, given the output feature $F1, F2, F3, F4$ from each stage, the pixel-shuffle operation (Shi et al., 2016) is used to upsample them to the target size $H \times W \times C$, and they are integrated by concatenation:

$$[F2', F3', F4'] = Upsample([F2, F3, F4]) \quad (15)$$

$$F_{ms} = Concat([F1, F2', F3', F4']) \quad (16)$$

where $F_{ms} \in R^{H \times W \times 4C}$ represents the integrated features embedded with multi-scale information.

A *SoftMax* classifier is used to classify the feature map $F_{ms}$ to the FR land cover class probability map $P \in R^{H \times W \times NC}$ for $NC$ categories:

$$F = \sigma \cdot (W_c \otimes F_{ms} + b_c) \quad (17)$$

$$P = \sum_{i=1}^{H} \sum_{j=1}^{W} \frac{exp(F_{i,j,k})}{\sum_{k=1}^{NC} exp(F_{i,j,k})} \quad (18)$$

where $F \in R^{H \times W \times NC}$ is the intermediate feature map, and the $F_{i,j,k}$ means the feature value at $(i, j)$ location and $kth$ channel, $W_c$ and $b_c$ are weight and bias, and $\sigma$ is the rectified linear unit (ReLU).

*2.2.5 Loss function*

The loss function aims to measure the differences between the prediction $P$ and the ground reference probability map $Z \in R^{H \times W \times NC}$. The differences are backpropagated to update the network parameter $\theta$ by minimizing the loss function during the training process. Given pairs of $\{Y, Z\}$, the optimal network model can be learned as:

$$\{\hat{P}, \theta\} = arg\,\min_{\theta}\{Loss(P, Z)\} \tag{19}$$

Considering the commonly class unbalanced distribution of urban compositions, the combination of a focal loss function $L_{focal}$ and dice loss function $L_{dice}$ is adopted in our approach (Milletari et al., 2016; Lin et al., 2020). $L_{focal}$ aims to reduce the weight of easy samples and enlarge the weight of hard samples, and $L_{dice}$ aims to maximize the intersection ratio between the prediction and ground reference, given as:

$$L_{focal} = -\alpha(1-P)^{\gamma} \cdot Z \cdot logP - (1-\alpha)P^{\gamma} \cdot (1-Z) \cdot \log(1-P) \tag{20}$$

$$L_{dice} = 1 - \frac{2(P \cdot Z)}{P + Z} \tag{21}$$

$$Loss(P, Z) = (1-\mu) \cdot L_{focal} + \mu \cdot L_{dice} \tag{22}$$

where $\alpha \in [0,1]$ is the balance weight between easy samples and hard samples, and $\mu \in [0,1]$ is the balance weight between $L_{focal}$ and $L_{dice}$.

## 3. Material/Study area

*3.1 Public dataset*

Three public urban scenario datasets were selected to examine the capabilities for detail reconstruction and correlation excavation of our method (**Fig. 8**). Noteworthy is that pairs of CR image and FR annotation images are required for SPM network training. However, due to the difficulty of obtaining matching pairs acquired on exactly the same date, we instead used the semantic segmentation dataset (WHDLD and LoveDA) to ensure that the image and its annotation image were matched, and further downsampled the image by a predefined scale factor to derive a synthetic CR image as the input for the SPM network. Furthermore, a newly published dataset specifically for SPM tasks (i.e., the FLAS dataset) was also adopted, which has matched pairs of CR Sentinel-2 image and FR annotated Google images.

The WHDLD dataset (Shao et al., 2020) consists of pairs of 2 m Google Earth images and the corresponding annotation images, located in the urban area of Wuhan, China. WHDLD contains 4940 image patch pairs with a size of 256×256 pixels and three bands, densely annotated with six land cover classes (i.e., building, pavement, vegetation, bare soil and water). After downsampling by four times the original images, the spatial resolution was converted from 2 m to 8 m and the patch size changed to 64×64 pixels, to serve as the CR image.

The LoveDA dataset (Wang et al., 2021) consists of pairs of 0.3 m spaceborne images and their corresponding annotation images, which are located in the urban and rural areas of Nanjing, Changzhou and Wuhan in July 2016, with a complex background and rich details. LoveDA contains 5987 image patch pairs with a size of $1024\times1024$ pixels and three bands, densely labelled with seven classes (i.e., background, road, water, barren, forest, agriculture and background). The image and annotation of the LoveDA were firstly resampled to 1.2 m spatial resolution ($256\times256$ pixels), and the image was further downsampled to 4.8 m resolution ($64\times64$ pixels) to serve as the CR image. The reason is that 0.3 m and 1.2 m resolution show few scale difference concerning building, road, pavement, green space in this urban scene, reconstructing 1.2 m image to 0.3 m would not reflect the performance of the SPM algorithm. Meanwhile, the scale effect becomes significant when it resampled to 4.8 m resolution, with pixelated boundaries between objects, which can be used to validate the reconstruction ability of SPM.

The FLAS dataset (He et al., 2022b) consists of pairs of 10 m Sentinel-2 images and the corresponding 1 m annotation images annotated on Google Earth image, located in Guangdong, China, in summer 2019. The FLAS dataset was re-cropped to 12785 image patch pairs, in which the Sentinel-2 image has a size of $64\times64$ pixels and seven bands, and the annotation image has a size of $320\times320$ pixels. Training, validation and test sets were split randomly at the ratio of 7:1:2 for the above three datasets (Considering that the self-attention based model requires large amount of training samples, we also examined the effect of various ratio on performance, such as 6:1:3, the results of which can be referred to supplementary file).

Noteworthy is that although the reconstruction scale for the synthetic dataset (WHDLD and LoveDA) is set to 4 in this study, it was sufficiently discussed with an ablation analysis in the supplementary file, which is referred to (Wu et al., 2018).

**Table 1**
Public datasets for validation

| Dataset | CR image | FR label | Scale | CR size | FR size | Number | Purpose |
|---|---|---|---|---|---|---|---|
| WHDLD | Downsampled Google image, 8.0 m/pixel | Annotated Google image, 2.0 m/pixel | 4 | 64×64 | 256×256 | 4940 | Validation on synthetic dataset |
| LoveDA | Downsampled spaceborne image, 4.8 m /pixel | Annotated Spaceborne image, 1.2 m/pixel | 4 | 64×64 | 256×256 | 5987 | Validation on synthetic dataset |
| FLAS | Sentinel-2 image, 10 m /pixel | Annotated Google image, 2.0 m/pixel | 5 | 64×64 | 320×320 | 3461 | Validation on real dataset |

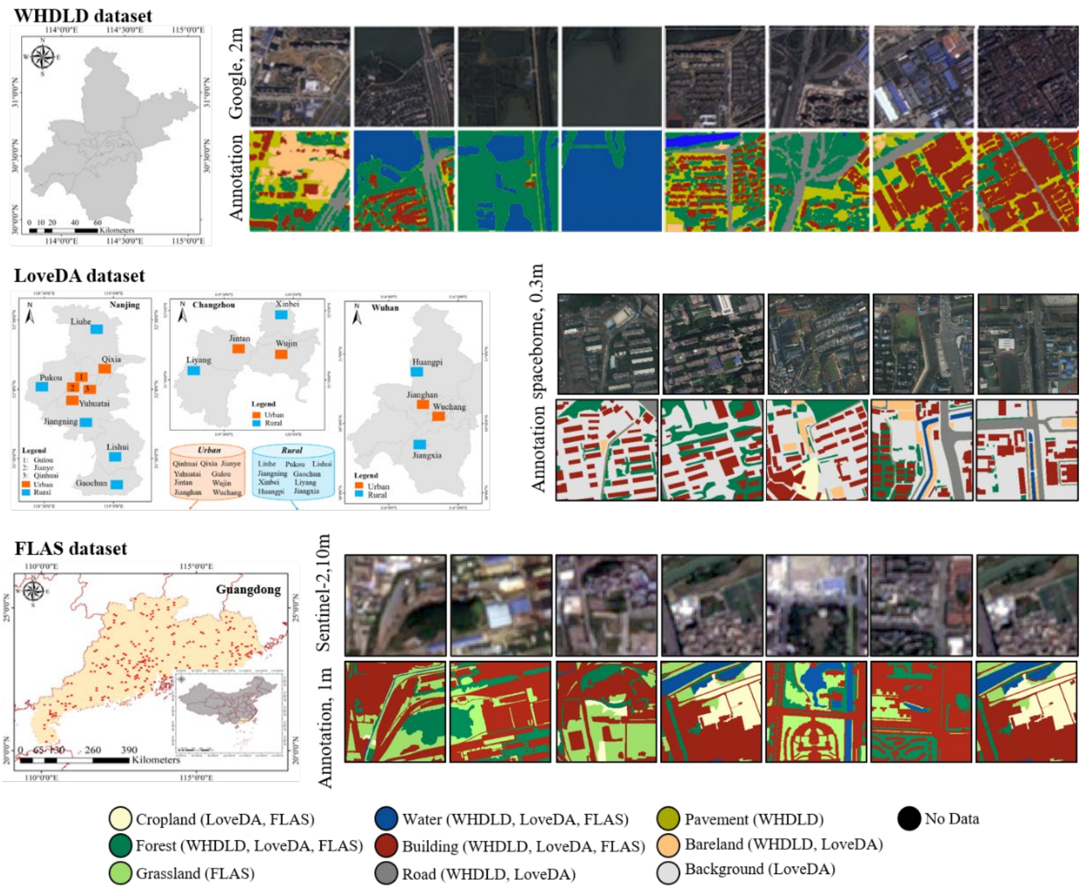


**Fig. 8.** Publicly available urban scenario dataset material for this study. (The dataset figures are cited from Shao et al., 2018, Wang et al., 2021 and He et al., 2022b).

*3.2 Study area*

Instead of validation on the public dataset, we further inspected performance on the downtown areas of the several provincial metropolises in China (**Fig. 9**). These study areas range from north to south, from the temperate monsoon climate to the subtropical monsoon climate, and from a first-tier city to a second-tier city, which have diverse scenes that can inspect the generalization of the LECOS.

Besides, the derived results from the LECOS are compared with the publicly prevailing land cover products, including the 10-m Finer Resolution Observation and Monitoring-Global Land Cover (FROMGLC30) product that was produced by semantic segmentation network and post-processing on Sentinel-2 images (Gong et al., 2019), the 30 -m GlobeLand30 product that was produced by classical machine learning and manual post-processing on Landsat images (Chen et al., 2014), the 10 -m Esri Land Cover product that was produced by semantic segmentation network on Sentinel-2 images (Karra et al. 2021), and the 10-m WorldCover product that was produced by deep learning method on Sentinel-1 and Sentinel-2 image conducted by European Space Agency (ESA) (Zanaga et al., 2021).

Specifically, we used the Sentinel-2A MSI (Multi Spectral Instrument) of these study areas, which can be accessed at [https://developers.google.com/earth-engine/datasets/catalog/ COPERNICUS_S2](https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2). 10 m Red, Green, Blue, NIR, 20 m SWIR-1, SWIR-2, and 60 m Coastal bands were selected, which were uniformly resampled to a 10 m spatial resolution. Cloud area was initially subtracted by a QA60 bitmask band with cloud mask information, and all the data in the growing period of April-September 2020 were composited by the median-mosaic operation to obtain complete coverage.

The pre-trained model trained on the FLAS dataset above was used for 2 m resolution mapping of these study area.

Presently, Sentinel-2 supplies the best available time-series images that have moderate spatial resolution and a large coverage. However, these images are insufficient to distinguish the intricate spatial heterogeneity and dynamics of urban patterns. Therefore, this research attempts to reconstruct the Sentinel-2 images to very fine spatial resolution to reveal finer urban spatial details.

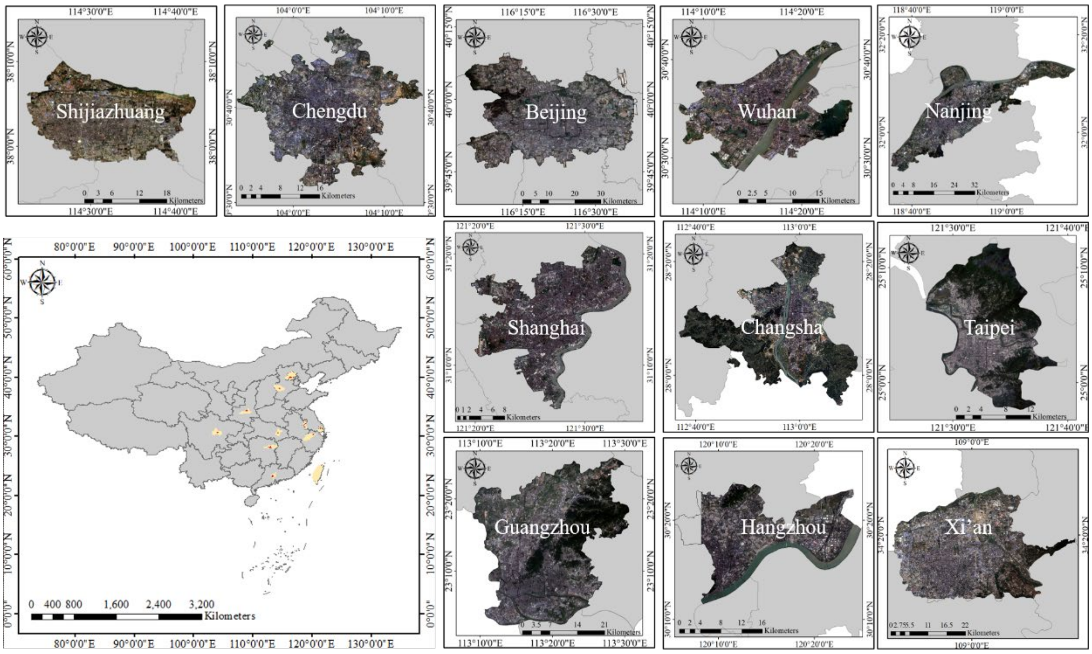

**Fig. 9**. Study area of the selected cities in China.

# 4. Results

## *4.1 Experiment setting*

### *4.1.1 Implementation details*

Firstly, we validated the effectiveness of our proposed method using public dataset material (Section 4.2.1). Considering the advantage of the explicit correlation modeling design of our method, we visualized the underlying attention matrix to inspect the learned syntax correlation patterns among the sub-pixels and the context correlation patterns among the mixed-pixels (Section 4.2.2 and 4.2.3). Besides, we further derived a statistic based on the attention score of each correlation to quantify the connection strength between different urban compositions, which can reflect the characteristic, accessibility and rationality of the urban spatial patterns (Section 4.2.4).

The LECOS model was then applied to the large-scale study areas to examine its practicability (Section 4.3). Specifically, the LECOS model that was well trained on the FLAS datasets was adopted to reconstruct the 10 m resolution Sentinel-2 images of the five metropolises, and generate the 2 m resolution fine-grained urban maps. Besides, based on the quantification strategy of the correlation patterns in Section 4.2.4, we also evaluated the urban spatial pattern of the five cities according to the correlation statistic, which was tested by spatial overlap analysis with the 100 m local climate zone map (Section 4.3).

For the parameter settings, the minibatch size was set to 32, and the Adam optimizer was used with the beta-1 parameter set to 0.9 and the weight decay set to 0.0005, which is recommended in (Liang et al., 2021). The learning rate was initially set to 0.001 and reduced on plateau with the reduction factor set to 0.8 and the patience set to 10, which is also suggested for training self-attention network (Liang et al., 2021). The training epoch was set to 200, the validation performed at each epoch, and we chose the best model from the validation set for evaluation using the test set. All the experiments were implemented in PyTorch 1.9.0 and Python 3.6.13 on a server with three NVIDIA GeForce RTX3090 graphic processing unit (GPU) accelerators (with 24-

GB GPU memory).

### 4.1.2 Comparison algorithm

The classical model-driven SPM algorithms like Spatial Attraction based Sub-pixel Mapping (SASM) (Mertens et al., 2006), the Adaptive MAP model and a winner-takes-all Class Determination strategy for Sub-pixel Mapping (AMCDSM) (Zhong et al. 2015), and the state-of-the-art data-driven SPM algorithms like SPMCNN-ESPCN (He et al., 2021a) are used as benchmark methods for comparison. A comparable experimental strategy is adopted to compare the model-driven and data-driven methods (He et al., 2021a). Further, a pixel-level classification methods are also provided for comparison, including the traditional support vector machine (SVM), and state-of-the-art data-driven semantic segmentation networks including UNet++ (Ronneberger et al., 2015), DeepLabV3+ (Chen et al., 2018), Swin Transformer (Liu et al., 2021). Pixel-level classification results were then downscaled to the target resolution to make them comparable to the SPM result.

### 4.1.3 Evaluation metric

For quantitative assessment of the performance of the compared methods, the overall accuracy (*OA*), *Kappa* coefficient, the mean value of the intersection-over-union (*mIoU*) and the F1-score (*mF1-score*) for each class were used as metrics. In addition, the producer's accuracy (*PA*) was selected to examine the class-wise accuracy.

The specific evaluation procedure was as follows: Firstly, our method was qualitatively and quantitatively evaluated by the test set of each public dataset, and the benchmark SPM algorithms were also applied for comparison (Section 4.2). Then, we deployed the LECOS model on several metropolises in China to inspect its performance in revealing very fine spatial resolution urban distributions across large-areas, and the state-of-the-art land cover products, i.e., 10 m WorldCover, 10 m FROM-GLC10, 10 m Esri Land Cover and 30 m GlobeLand30 were used for comparison (Section 4.3), to inspect that whether the 2 m resolution results from LECOS can significantly improve details of urban pattern compared to the existing products. Besides, we also conducted an ablation analysis to evaluate the contribution of SBD, CBD, the network structure (including the number of layers, the 4-stage structure), and the number of training sample patches, by removing the corresponding design and inspecting the performance change (see in the supplementary file).

### *4.2 Results on public datasets*

#### *4.2.1 Visual and quantitative examination*

Three patches in the test set were selected from each public dataset for visual examination (**Fig. 10**), and quantitative assessment of the comparison algorithms is provided in **Tables 2-4**.

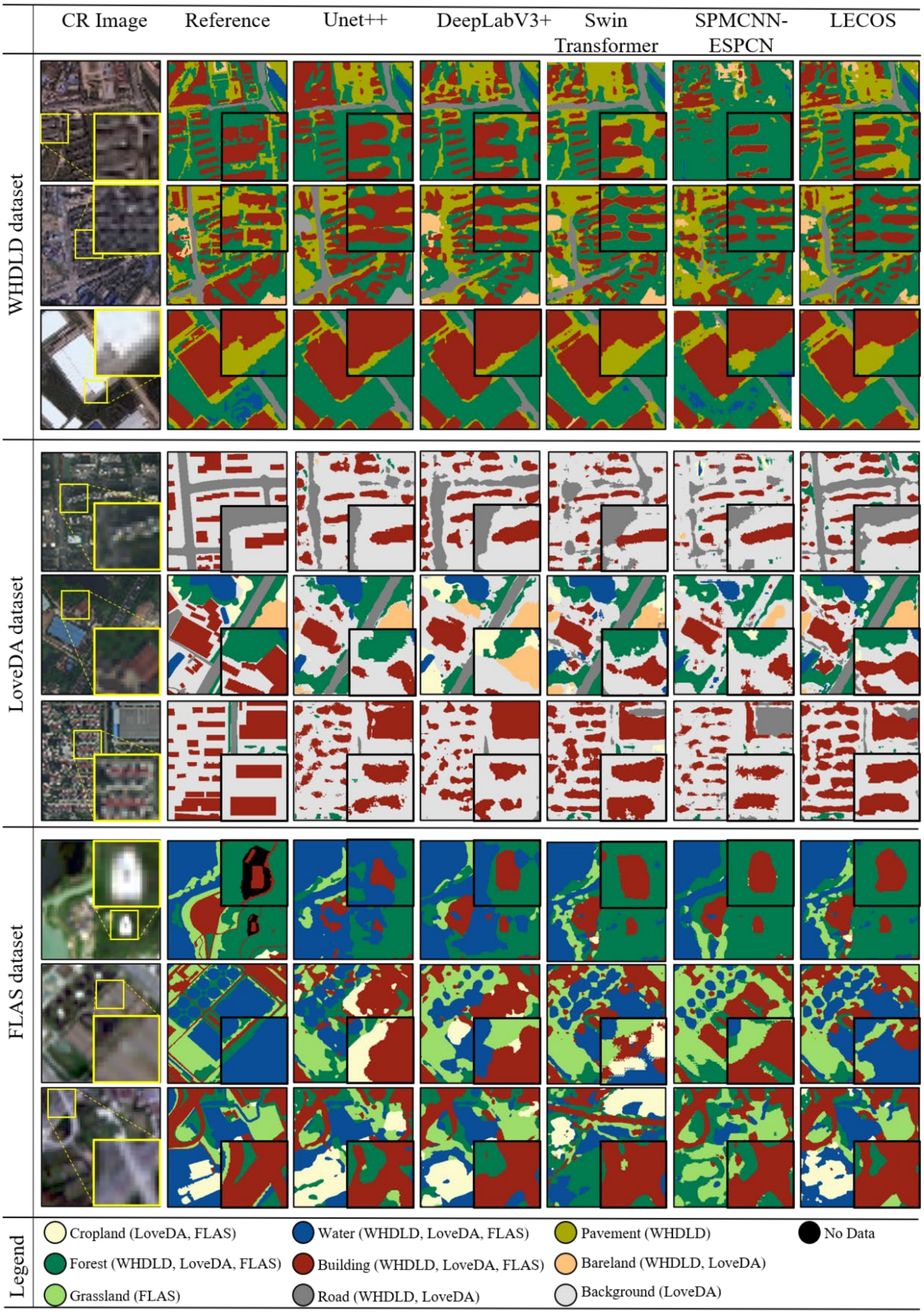


**Fig. 10**. Visual examination of the performance of each algorithm on three public datasets.

For visual inspection relating to the WHDLD dataset, LECOS can reconstruct the detailed compositions and smooth edges well for all the land cover classes in the urban scene, and the outlines of the buildings are closer to the ground reference. Besides, with the aid of the learned building-pavement correlation rule, LECOS can distinguish pavements nearby buildings in the residential district, while the other convolutional-based models that utilize only local information confuse the pavement with roads, and the traditional self-attention based models cannot well reconstruct the spatial morphological shape of the objects, as can be seen in **Fig. 10.**

For the LoveDA dataset, due to the lack of consideration of the correlation between road and green parcels, and road and road, that essentially exists in urban scenes, the roads are broken and the green parcels along the roads are almost lost in the results of the comparison method. In contrast, LECOS can reconstruct the continuous roads and green belts better because the context-dependency between urban components is explicitly learned in the network. Besides, LECOS can better recover the morphological shapes and edge contours of the buildings, because the syntax-dependency between sub-pixels is also modelled explicitly.

For the FLAS dataset, LECOS is able to precisely identify the pond, grass and buildings, because the context dependency between pond and grass, and grass and building are learned. However, the convolutional-based method is prone to misclassify the pond. Noteworthy is that despite the erroneous identification, the traditional SPM method (i.e., SPMCNN-ESPCN) is able to reconstruct a more complete morphological shape of the pond than the other CNN-based method, demonstrating the effectiveness of sub-pixel-scale manipulation. It is also reflected in our method that the shape of the pond is more complete, and the impervious road is more continuous, benefiting from the learned syntax-dependence.

**Table 2**
Quantitative assessment on WHDLD dataset

| | Model-driven | | | Data-driven | | | | Model- and Data-driven |
|---|---|---|---|---|---|---|---|---|
| | SVM | SASM | AMCDSM | UNet++ | DeepLabV3+ | Swin Transformer | SPMCNN-ESPCN | LECOS |
| Building | 0.5144 | 0.1941 | 0.1906 | 0.6208 | 0.6492 | 0.7145 | 0.5857 | **0.7372** |
| Road | 0.0000 | 0.1846 | 0.1843 | 0.5815 | **0.6596** | 0.5731 | 0.1985 | 0.6079 |
| Pavement | 0.4760 | 0.2508 | 0.2513 | 0.5729 | **0.5888** | 0.5381 | 0.4824 | 0.5874 |
| Vegetation | 0.8704 | 0.6879 | 0.7007 | **0.9088** | 0.8766 | 0.8825 | 0.9042 | 0.8987 |
| Bare Soil | 0.2544 | 0.4255 | 0.4269 | 0.3652 | 0.2762 | **0.5203** | 0.3499 | 0.4879 |
| Water | 0.4612 | 0.0907 | 0.0874 | 0.9456 | **0.9654** | 0.9494 | 0.8589 | 0.9562 |
| *OA(%)* | 61.69 | 39.08 | 39.50 | 81.61 | 81.90 | 81.77 | 76.02 | **82.66** |
| *Kappa* | 0.4280 | 0.1205 | 0.1239 | 0.7375 | 0.7355 | 0.7446 | 0.4691 | **0.7562** |
| *mIoU* | 0.3127 | 0.1445 | 0.1529 | 0.5609 | 0.5549 | 0.5664 | 0.4554 | **0.5842** |

**Table 3**
Quantitative assessment on LoveDA dataset

| | Model-driven | | | Data-driven | | | | Model- and Data-driven |
|---|---|---|---|---|---|---|---|---|
| | SVM | SASM | AMCDSM | UNet++ | DeepLabV3+ | Swin Transformer | SPMCNN-ESPCN | LECOS |
| Background | 0.7210 | 0.0504 | 0.0503 | 0.6947 | 0.7361 | 0.7144 | **0.7731** | 0.7262 |
| Building | 0.3965 | 0.0112 | 0.0112 | 0.5986 | 0.5661 | **0.6056** | 0.3845 | 0.6043 |
| Road | 0.0042 | 0.0159 | 0.0157 | 0.4268 | 0.4301 | 0.4903 | 0.2305 | **0.5180** |
| Water | 0.0454 | 0.2484 | 0.2518 | 0.7343 | 0.7006 | 0.7419 | 0.5788 | **0.7549** |
| Barren | 0.0987 | 0.1787 | 0.1767 | 0.5105 | 0.3329 | 0.3412 | 0.1792 | **0.5143** |
| Forest | 0.4359 | 0.7315 | 0.7448 | 0.4771 | 0.4544 | 0.5305 | 0.4199 | 0.5844 |
| Agriculture | 0.1793 | 0.3717 | 0.3730 | 0.5939 | 0.5813 | 0.5503 | 0.3107 | **0.6817** |
| *OA(%)* | 44.14 | 15.44 | 15.57 | 61.75 | 61.26 | 62.33 | 53.13 | **65.90** |
| *Kappa* | 0.1741 | 0.0527 | 0.0536 | 0.4789 | 0.4601 | 0.4842 | 0.3284 | **0.5377** |
| *mIoU* | 0.1706 | 0.0317 | 0.0323 | 0.4355 | 0.4212 | 0.4301 | 0.2877 | **0.4882** |

**Table 4**
Quantitative assessment on FLAS dataset

| | Model-driven | | | Data-driven | | | | Model- and Data-driven |
|---|---|---|---|---|---|---|---|---|
| | SVM | SASM | AMCDSM | UNet++ | DeepLabV3+ | Swin Transformer | SPMCNN-ESPCN | LECOS |
| Cropland | 0.0000 | 0.5558 | 0.6069 | 0.6394 | 0.6762 | 0.6025 | 0.0000 | **0.7050** |
| Tree | 0.7042 | 0.6593 | 0.7139 | 0.7397 | 0.7093 | 0.6349 | **0.7475** | 0.6739 |
| Grass | 0.4188 | 0.1297 | 0.1308 | 0.4472 | 0.4737 | 0.4888 | **0.6051** | 0.4123 |
| Water | 0.8336 | 0.8022 | 0.8270 | 0.7127 | 0.7376 | **0.8542** | 0.6150 | 0.7376 |
| Impervious | 0.8953 | 0.7369 | 0.7876 | 0.8646 | 0.8617 | 0.8791 | 0.7933 | **0.9098** |
| *OA(%)* | 74.11 | 62.49 | 66.88 | 77.02 | 77.18 | 77.25 | 69.40 | **78.80** |
| *Kappa* | 0.5852 | 0.3071 | 0.3479 | 0.6485 | 0.6527 | 0.6462 | 0.5479 | **0.6687** |
| *mIoU* | 0.4432 | 0.2251 | 0.1791 | 0.5373 | 0.5411 | 0.5535 | 0.3981 | **0.5495** |

In terms of quantitative assessment (**Table 2-4**), LECOS achieves the best *OA* and *Kappa* with 82.66% and 0.7562 on the WHDLD dataset, which surpasses the benchmark SPM method (SPMCNN-ESPCN) by 6.64% and 0.2851, and outperforms the benchmark semantic segmentation method (DeepLabv3+) by 0.76% and 0.0187.

Besides, LECOS also performs well in terms of *PA* accuracy on each urban component class that can be hard to identify and reconstruct, such as building, road and barren. The accuracies of these are increased by 0.0057, 0.0879 and 0.0038, respectively, compared to the benchmark on the LoveDA dataset. For the FLAS dataset, our method again achieves the highest accuracy, with *OA* and *Kappa* accuracy of 78.80% and 0.6687, and outperforms the benchmark by 1.62% and 0.0160, respectively. Noteworthy is that the model-driven methods have a low accuracy on the WHDLD and LoveDA dataset. The reason is that the RGB image has insufficient spectral information to derive robust endmember spectrum and abundance images, and the large uncertainty of the abundance images impacts the subsequent SPM process.

*4.2.2 Explicability of the sub-pixel-level syntax-dependence*

A reliable syntax that indicates the sub-pixel location of each urban composition is the key to a successful detail reconstruction. Thus, we visualize and inspect the learned syntax in this section. Specifically, we selected four representative scenes in the WHDLD dataset (correlation patterns of different classes are similar in a given dataset), and focused the evaluation on the zoomed-in area in the CR image within 3×3 mixed-pixels. Our method was used to reconstruct the center mixed-pixel to 16 sub-pixels. Then, the intermediate variable (i.e., the syntax-dependence represented by the attention score matrix $SA \in R^{S^2 \times S^2}$ in stage-1 layer 1, which is defined in Eq. (4)) was visualized in 3-dimensional space (**Fig. 11**). The size of the matrix $SA$ is 16×16, and the z-dimension represent the correlation strength between the two sub-pixels.

Taking **Fig. 11a** as an example, a high attention score occurs between building sub-pixel and vegetation sub-pixel (as the blue and purple area in the 3-D visualization), which means vegetation and building have a close correlation and often appear together in the urban scene, and their distributions are mutually attracted when predicting the sub-pixel locations. Therefore, with the guide of the syntax-dependence indication, the detailed edge and corners of the building can be well restored.

Taking **Fig. 11b** as an example, with rapid urban land development and utilization, most of the bareland in the urban scene is usually derived from deforestation. Therefore, bareland is often accompanied by vegetation, and the bareland sub-pixel exhibits a strong attention score to vegetation sub-pixel. However, the attention score is not symmetrical (i.e., the vegetation seldom pays attention to the bareland), since the vegetation (e.g., greenbelt, horticulture and arboriculture) is not necessarily accompanied by bareland.

Similarly, **Fig. 11c** demonstrates a construction site, where the bareland sub-pixel pays more attention to the building sub-pixel, and in **Fig. 11d** with residential areas near the main street, the building sub-pixel and the road sub-pixel mutually attract each other with high attention score, which meets the needs for convenient settlements.

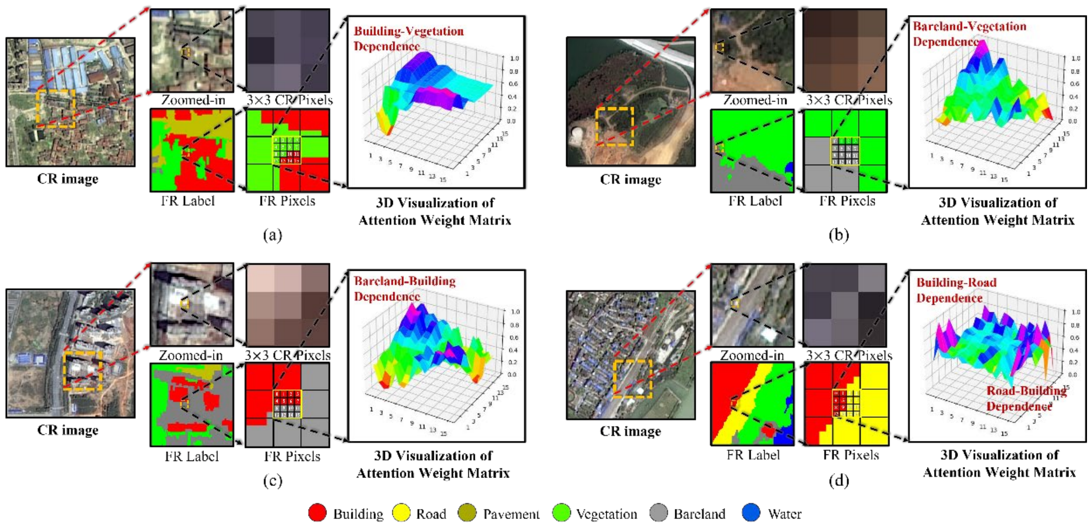


**Fig. 11.** Visualization of the learned syntax-dependence.

*4.2.3 Explicability of the pixel-level context dependence*

Four scenes in the WHDLD dataset are demonstrated for illustration. Taking **Fig. 12a** as an example, we selected the road pixel $i$ (denoted as red point) as the *Query*, and used the *Key* of the residual pixels to generate the context attention score matrix $\{\alpha_{i,j}\}_{j=1\ldots N}$ (defined in Eq. (9)) across different stages throughout the LECOS network, which represents the correlation strength of each pixel to the selected road pixel. In each stage, we randomly selected two attention matrices $\{\alpha_{i,j}\}_{j=1\ldots N}$ at different layers and different heads, one is representative for inter-class correlation and one is for intra-class correlation. Noteworthy is that stage-1 is not visualized, since the context information is not obvious in the shallow layer. Generally, we highlight three highest attention scores in $\{\alpha_{i,j}\}_{j=1\ldots N}$ with directional arrows in each attention score map, to demonstrate the correlation patterns between the red point and the residual pixels. For discrimination, the heat points and arrows with white color notations (i.e., B for building, R for road, V for vegetation, BL for bareland) indicate the intra-class correlation with the red point, while the heat points with orange notation represent the intra-class correlation.

Two phenomena are obvious: 1) we find that the attention tends to focus on the local small land covers in stage-2 or 3 (e.g., road usually pays more attention to the elongated road and scattered vegetation in **Fig. 12a)**, while high attentions are spatially aggregated in stage-4 and tend to focus on the dominant land covers in the scene (e.g., the road pays more attention to buildings), indicating that the context gradually converges towards global perception; 2) Taking stage-4 as an example, the roads are highly correlated with buildings (**Fig. 12a**), the vegetation pays more attention to road and itself in residential areas (**Fig. 12b**), the buildings have a large correlation with the surrounding vegetation (**Fig. 12c**), and the buildings are highly correlated with themselves and bareland in construction sites (**Fig. 12d**).

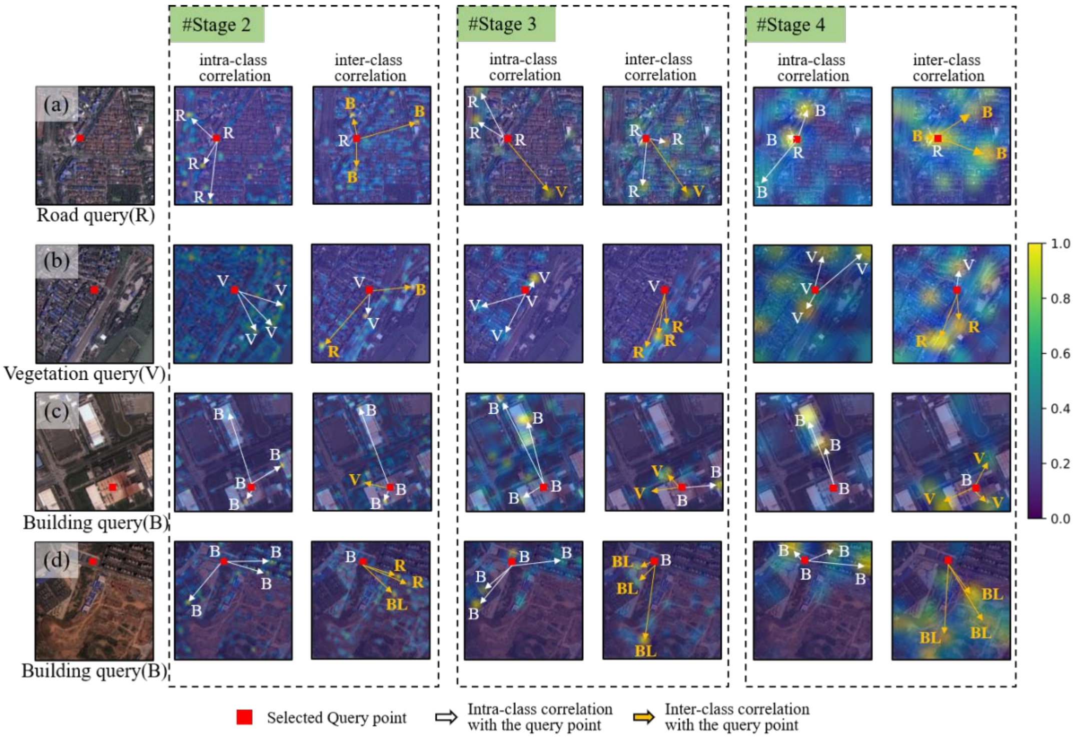


**Fig. 12.** Visualization of the context dependence across different stage.

*4.2.4 Quantitative analysis of the connection strength*

We construct a statistic to quantify the correlation strength for each correlation pattern in the above analysis. Specifically, the output 1-dimensional sequential feature $B \in R^{(4\times4\times C)\times(W/4\times H/4)}$ in stage-1 is reshaped to a 2-dimensional feature with size $H/4 \times W/4 \times 16C$, and the FR labels are resampled from $H \times W$ to $H/4 \times W/4$ to match the size of the reshaped feature $B$, which is used to provide category information for each pixel of $B$. Subsequently, the context-dependence represented by the attention score matrix $A \in R^{(W/4\times H/4)\times(W/4\times H/4)}$ in stage-1, which indicates the correlation between each of the two pixels in $B$, is used for the attention score statistic. For example, when calculating the building-road correlation pattern (i.e., B-R), the attention score between all the building pixels and the road pixels in the matrix $A$, indicated by the FR labels, is summarized. In this way, the correlation patterns can be quantified.

We construct a statistic for the WHDLD dataset. WHDLD has six classes such that there are, in total, 36 correlation patterns between each of the two categories, which are ranked according to total attention score (**Fig. 13**), and which reflect the most contributing and important correlation patterns in the WHDLD urban scenario.

From the ranking list, we find that the intra-class correlation is usually stronger than the inter-class correlation, which is reasonable due to spatial autocorrelation theory where the attraction within the same class is expected to be greater than between different classes. As for inter-class correlations, we visualize the top four patterns in **Fig. 13** (i.e., BL-R, B-V, R-V, B-P), for examining the reliability of the quantification.

We find that: 1) for the BL-R correlation pattern, the bareland in the urban scene is usually represented as a construction site. Thus, a frequent transportation of the construction materials is highly necessary for the construction project, which can explain the frequent occurrence of the BL-R pattern; 2) The high attention sore of the B-V and R-V correlation pattern demonstrates that the building is highly correlated with vegetation, and the road is also tightly correlated with vegetation, indicating the presence of a considerable number of residential scenarios in the WHDLD dataset, and also indicating that the residential horticulture are well organized; 3) The B-P correlation pattern also ranks highly, which means that buildings are also highly correlated with pavements, which meets the convenient transportation requirements of people living in the residential building.

From the above qualitative and quantitative analysis of the syntax- and context-dependence, we find that they are not only explicable, but can also be used as an indicator to understand the complex urban spatial patterns (e.g., estimating the human exposure to greenspace in urban scenes by the B-V and R-V attention score, and estimating the transportation convenience by the B-R and B-P attention score, etc.).

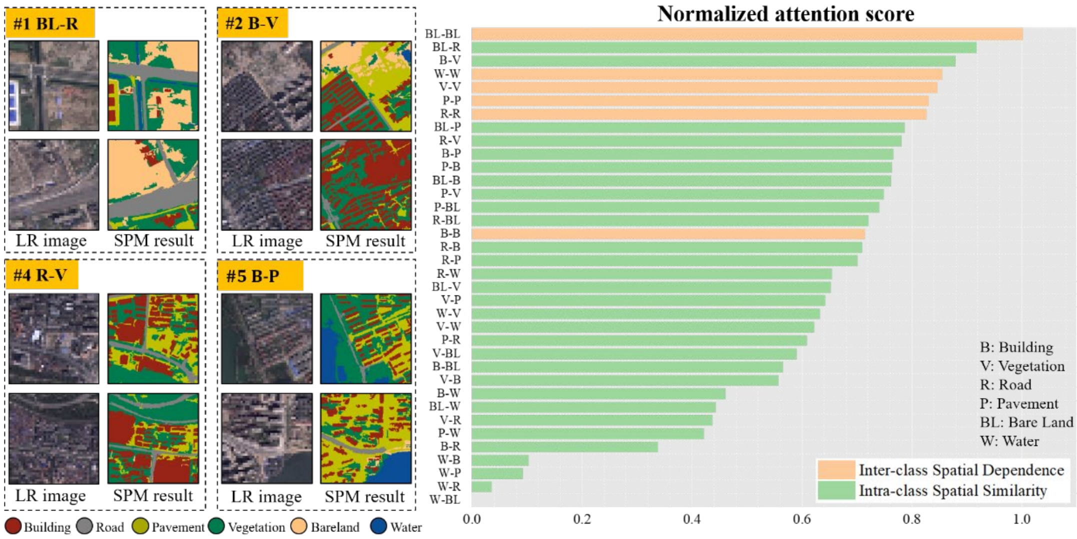


**Fig. 13.** Visualization of the normalized attention score for demonstrating the strength of each correlation pattern.

### *4.3 Results on five metropolises*

#### *4.3.1 Visual and quantitative examination*

The 2 m spatial resolution land cover maps in the downtown area of the five cities are displayed visually in **Fig. 14** (taking Shijiazhuang as an example), as well as the comparison with the public land cover products. A zoomed-in area was selected to inspect the performance in terms of detail reconstruction. Results for other cities are given in Appendix II in the supplementary file.

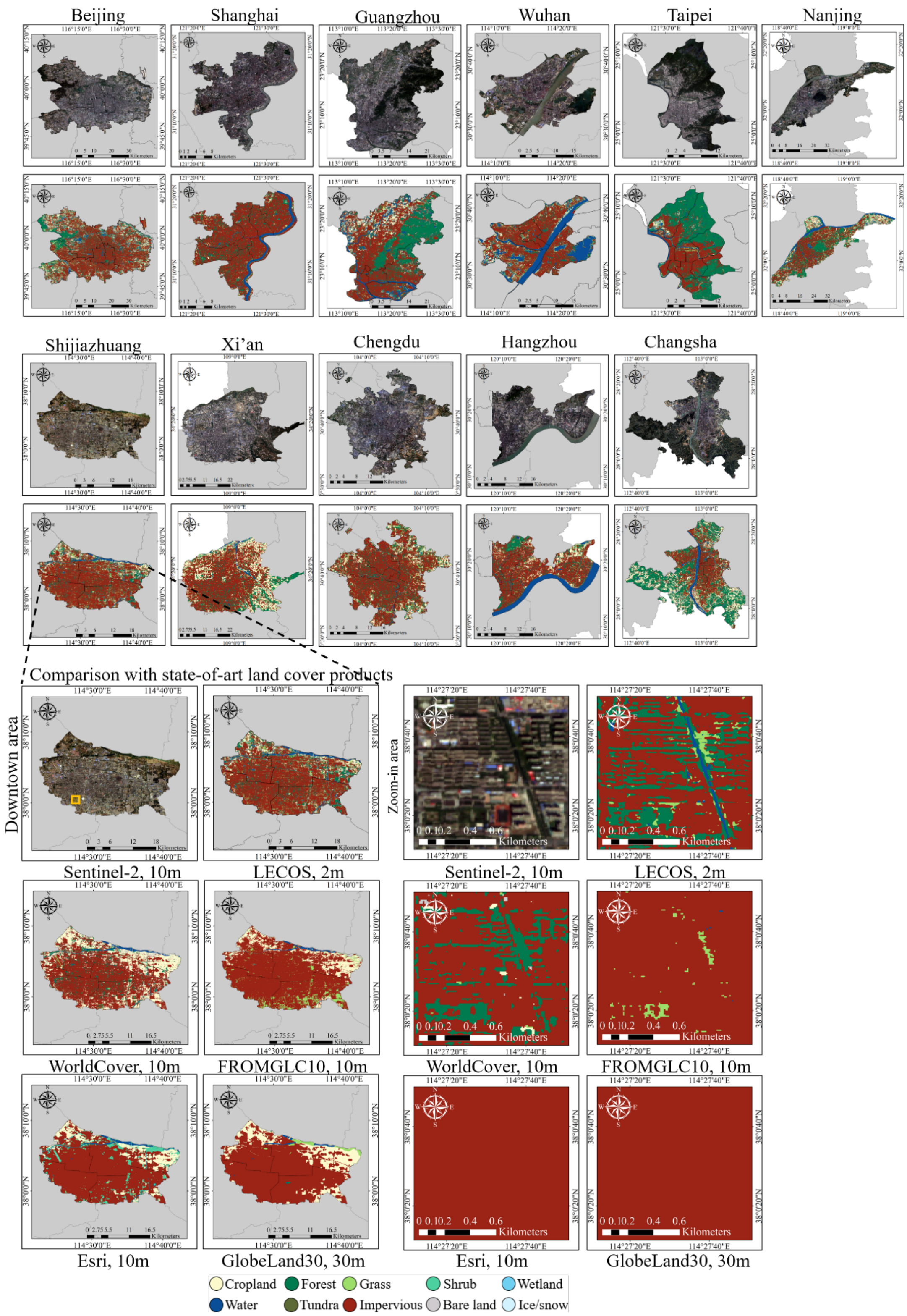


**Fig. 14**. Comparison of our method with state-of-the-art public land cover products.

It is obvious that the detail distribution of the urban compositions such as buildings, urban greenspace and grassland is usually missing in the Esri Land Cover and

GlobeLand30 products. Although detailed compositions can be reconstructed by WorldCover and FROMGLC10, the morphological shape is vague and incomplete with jagged boundaries. In contrast, our method can restore a fine-grained structure into the urban compositions, especially for small-sized or elongated urban greenspace that is usually mixed with impervious surfaces in the original Sentinel-2 images and difficult to detect.

*4.3.2 Connection strength of each correlation pattern*

The normalized attention score statistic of each correlation pattern in the several cities is provided in **Fig. 15**. Specifically, impervious surface was selected as the *Query*, and the other classes were used as the *Key*. Thus, four correlation patterns (i.e., I-CL, I-T, I-G, I-W; I for impervious, CL for cropland, T for Tree, G for Grass and W for water) were inspected.

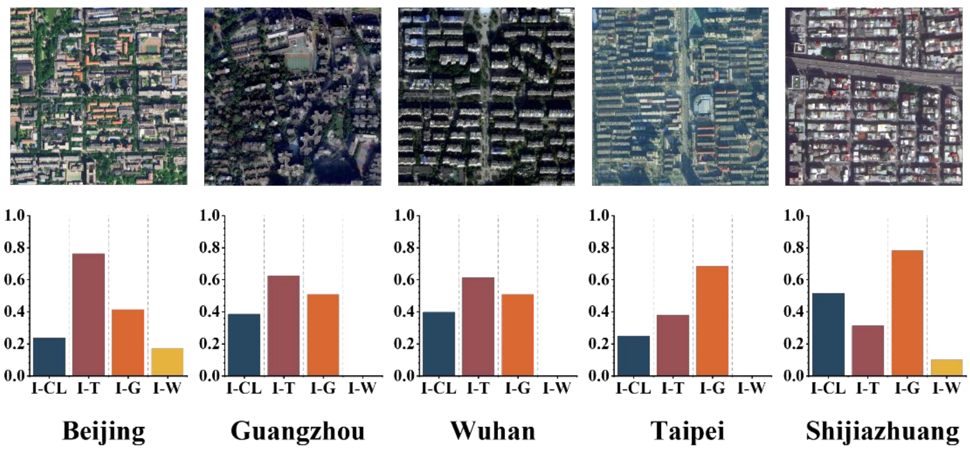


**Fig. 15.** The attention score of each correlation patterns in the five cities.

Focusing on the I-T score representing human exposure to greenspace, it is obvious that Beijing achieves the highest score, which means that the urban greenspace is cross-interweaved with impervious surfaces with highly accessibility. Meanwhile, although Taipei has a larger greenspace area ratio, there is less greenspace interspersed within the impervious surface buildings or roads, resulting in a weaker connection strength and lower exposure to urban greenspace. In this way, we find that the attention score statistic can precisely reflect the urban patterns.

*4.3.3 Overlay with Local Climate Zone (LCZ) dataset*

From the above analysis, Beijing achieves the highest I-T score, which means that more trees are planted within the buildings group. To validate this conclusion and further inspect the effectiveness of the I-T attention score statistic in terms of spatial accessibility, we further consider a spatial overlay analysis between our result and the Local Climate Zone (LCZ) dataset. The LCZ dataset divides impervious surfaces into 10 finer classes, in terms of the density (compact, open, sparse), height (high rise, mid rise, low rise), and material (heavy and lightweight), as shown in **Fig. 16**. We measure the area of the urban trees that fall into the area of each impervious class, which is normalized by dividing the total area of that impervious class. A higher tree ratio in denser classes implies greater spatial accessibility and greenspace exposure. Specifically, the 100 m spatial resolution GLOBAL LCZ MAP (Demuzere et al., 2022) was used. A zoomed-in area in Taipei is provided in **Fig. 16** to spatially demonstrate the overlay of the two products, and the urban tree ratio statistic for each impervious class in the five cities is provided in **Fig. 17**.

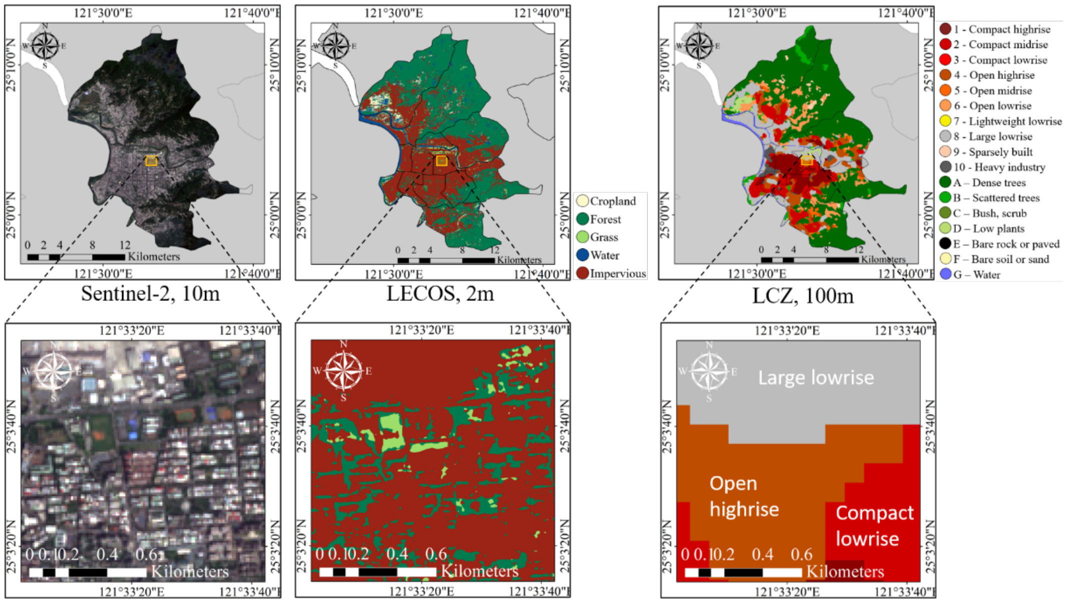


**Fig. 16**. The ratio of trees distributed in each impervious-related category region.

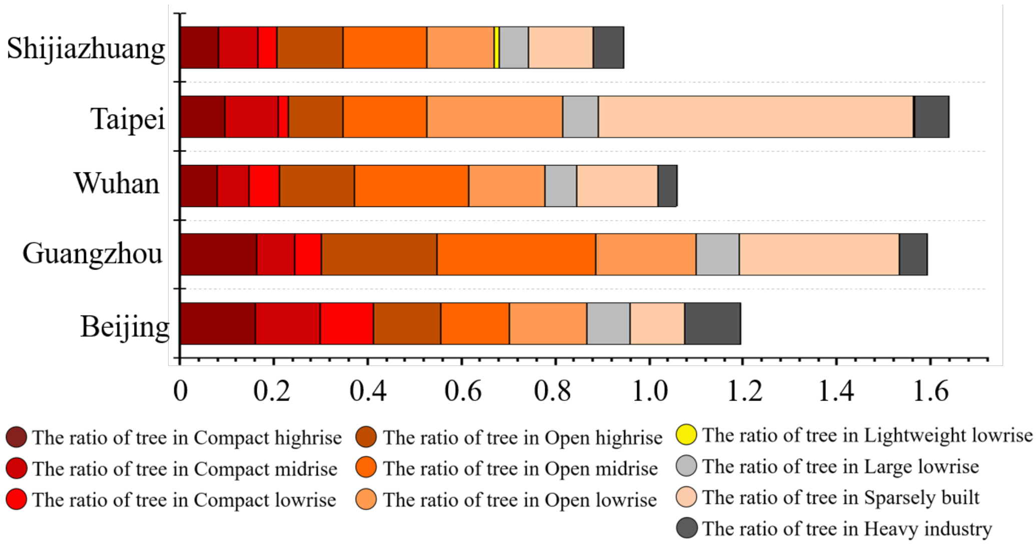


**Fig. 17**. The ratio of trees distributed in each impervious-related category region.

Beijing has the highest greenspace ratio amongst the three compact impervious classes, which is consistent with the I-T score analysis. Specifically, the proportion of greenspace in compact areas is significantly higher than that in other cities, indicating that green exposure and accessibility are greater, which can serve more people and, for example, help to mitigate the urban heat island effect. On the other hand, in Taipei city, a large greenspace ratio is configured in the sparse built area, while the proportion of greenspace in the compact area is small, indicating that the configuration of green vegetation may be sub-optimal for some residents, and resulting in lower greenspace accessibility.

# 5. Discussion

The experimental results demonstrated the superiority of our method, which is due to integrating a data-driven learning procedure with model-driven spatial correlation characterization, which can achieve more fine-grained land cover identification than using either of them alone. Other findings are 1) from a comparison between pixel-level classification and the SPM method, we find that SPM can reconstruct more complete morphological shapes and details owing to its sub-pixel manipulation ability; while the pixel-level classification method commonly achieves higher pixel-level accuracies due to its strong class identification ability. Our method combines the advantages of both (i.e., context for class identification and syntax for sub-pixel location inference) and, thus, achieves greater identification accuracy and detail reconstruction performance; 2) from the visualization of the correlation, the large spatial correlation between two objects usually has long-distance due to the heterogeneity of urban scenes, which demonstrates the necessity of global "mutual retrieval". Besides, datasets with different urban scenes reveal different intensities of the correlation patterns, which demonstrate the explicability of our method; 3) from the ablation experiment, the self-attention in self-attention operation (SNS) can learn more spatial detail features than the typical self-attention.

From the experimental results in the selected cities, we found that 1) the 2 m spatial resolution of our products can effectively reveal a heterogeneous distribution of the urban compositions, compared to the prevailing 10 m or 30 m resolution land cover products, which are more suitable for urban pattern recognition; 2) the traditional green coverage indices cannot reflect human accessibility to greenspace (Chen et al., 2022). For example, Taipei has larger ratio of greenspace than other cities, however, it is distributed in the mountainous area and far away from the residential area, leading to a relatively small correlation between impervious and tree reflected by our learned correlation; 3) an overlay analysis with the local climate zone product also confirms that more greenspace ratio is distributed in compact impervious areas in Beijing than in Taipei, which might be expected to serve more people; 4) the correlation derived from our method is able to consider the global pattern of each composition, which can reflect the cross-interweaved degree of both. Such inferences may find future application in social and environmental analyses of cities, for example, for understanding urban fairness and measuring resource distribution.

There exist several limitations of the proposed method. First, the fixed size of tokenization may adversely affect the learning of the spatial correlation, since the distance between two correlated urban components may vary. Second, the design of interpolating the CR image at the beginning of the network may increase the computational burden, since the size of the feature maps may be increased throughout the network. Third, we make a statistic of the computational efficiency of each algorithm in terms of parameter number (Para) and floating point of per second (FLOPs) (refer to the supplementary file), and find that although the number of parameters in LECOS model is relatively larger than that in other networks, the computational complexity is at the same order of magnitude. Fourth, when evaluating the correlation score of each city, it can only reflect a relative comparison that, e.g., impervious is more correlated with tree than with cropland, or the correlation score of one city is higher than another city, however, it is difficult to give a real correlation value for each city to test our generated correlation score due to the lack of full city ground truth.

## 6. Conclusion

Very fine spatial resolution observation of the urban pattern dynamic is the foundation to reveal urban spatial heterogeneity in cities, and inform judgements and planning with respect to issues such as urban fairness and access to greenspace. However, the current large-area land use/land cover products are usually limited by a moderate spatial resolution (i.e., 10 m or 30 m), which makes them unsuitable for mapping intricate urban components.

The contribution of this research has two folds including algorithm innovation and scientific findings of urban spatial pattern. In aspect of algorithm, the spatial resolution of land use/land cover products were increased, specifically for urban scenarios, by developing a new sub-pixel mapping algorithm, which addresses two challenges in the sub-pixel mapping community: 1) local spatial autocorrelation is usually limited to a fixed window, which ignores global contextual correlation that can inform inference at sub-pixel locations; 2) urban scenarios are complex, which means that the “more proximate, more similar" assumptions of classical spatial autocorrelation are unable to support inference of a heterogeneously distributed sub-pixel pattern. Therefore, how to learn the various correlations amongst urban components that can help sub-pixel location reasoning remains an open question. To this end, we designed an end-to-end contextual spatial correlation learnable network architecture (LECOS) to address directly the drawbacks of the fixed autocorrelation, providing two innovations: 1) a "mutually retrieve" mechanism was designed in an end-to-end network to learn correlation patterns adaptively; 2) a "self-attention in self-attention" operation was designed to explicitly infer the classes of the sub-pixels.

Validation experiments were conducted on three challenging urban datasets, and we found that the LECOS was better able to reconstruct the outlines and edges of urban buildings, roads, trees, etc., with an average *OA* of 82.66%, which significantly outperformed the benchmark. From the ablation analysis, we found that the designed "self-attention in self-attention" operation added greatly to the restoration of spatial details, with an increase in accuracy of 2.14%. From visualization of the learned correlations in LECOS, we found that datasets with different urban scenes reveal different types and intensities of correlation patterns, and all of them were consistent with the *in-situ* reference, demonstrating the explicability of our method.

In aspect of scientific findings, several typical metropolises were selected to examine the practical applicability of the proposed method, and we found that our method not only revealed heterogeneous urban patterns with very fine spatial resolution compared to the prevailing 10 m or 30 m resolution land cover products, but also the statistic of the intermediate attention score was able to reflect the human exposure patterns that are indicative of urban fairness. For example, Beijing was found to have greater urban greenspace accessibility, relative to Taipei. Such inferences may be useful in supporting planning decisions.

In summary, this research provides the first explicable sub-pixel mapping method for reconstructing very fine spatial resolution urban spatial patterns. As such, it represents the first attempt to combine model-driven spatial correlation theory with data-driven learning practice, which makes it possible to characterize the spatial heterogeneity of urban spatial patterns. We hope that LECOS will be used in future to support further sustainable urban development research.

**Acknowledgements**

This work was supported by the National Key R&D Program of China Grant No. 2022YFB3903402, in part by the National Natural Science Foundation of China under Grants 42222106, 61976234, 42201340, Natural Science Foundation of Guangdong Province Grant No. 2020A1515110708.

**Data availability statement**

The executable code that supports the findings of this research are available from the corresponding author upon reasonable request.

**Reference**

Adams, J. B., Sabol, D. E., Kapos, V., Filho, R. A., Roberts, D. A., Smith, M. O., et al.

(1995). Classification of multispectral images based on fractions of endmembers: Application to land-cover change in the Brazilian Amazon. Remote Sens. Environ., 52, 137–154. https://doi.org/10.1016/0034-4257(94)00098-8

Ardila, J. P., Tolpekin, V. A., Bijker, W., Stein, A., 2011. Markov-random-field-based super-resolution mapping for identification of urban trees in VHR images. ISPRS Journal of Photogrammetry and Remote Sensing 66(6): 762-775. https://doi.org/10.1016/j.isprsjprs. 2011.08.002.

Atkinson, P. M. 1997. Mapping sub-pixel boundaries from remotely sensed images. Innovations in GIS IV. London, U.K.; Taylor and Francis, 166–180. https://www.taylorfrancis.com/chapters/edit/10.1201/9781482272956-25/mapping-sub-pixel-boundaries-remotely-sensed-images-peter-atkinson.

Atkinson, P. M. 2005. Sub-pixel target mapping from soft-classified, remotely sensed imagery. Photogramm. Eng. Remote Sensing 71(7): 839-846. https://doi.org/10.14358/PERS.71.7.839.

Chen, B., Wu, S., Song, Y., Webster, C., Xu, B., Gong, P. 2022. Contrasting inequality in human exposure to greenspace between cities of Global North and Global South. Nature commun., 13(1), 1-9. https://doi.org/10.1038/s41467-022-32258-4.

Chen, J., Ban, Y., and Li, S., 2014. China: Open Access to Earth land-cover Map. Nature 514 (7523): 434. https://doi.org/10.1038/514434c.

Demuzere, M., Kittner, J., Martilli, A., Mills, G., Moede, C., Stewart, I. D., van Vliet, J., and Bechtel, B., 2022. A global map of Local Climate Zones to support earth system modelling and urban scale environmental science, Earth Syst. Sci. Data

14(8) 3835-3873. https://doi.org/10.5194/essd-14-3835-2022.

Deng, C., Zhu, Z., 2020. Continuous subpixel monitoring of urban impervious surface using Landsat time series. Remote Sens. Environ., 238, 110929. https://doi.org/10.1016/j.rse.2018.10.011.

Fisher, Y., and Koltun V., 2015. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122.

Foley, J. A. et al., 2005. Global consequences of land use. Science 309, 570–574. https://doi.org/10.1126/science.1111772.

Gong P., et al., 2019. Stable classification with limited sample: transferring a 30-m resolution sample set collected in 2015 to mapping 10-m resolution global land cover in 2017. Sci. Bull., 64, 370-373. https://doi.org/10.1016/j.scib.2019.03.002

He, D., Zhong, Y., Feng, R., Zhang, L., 2016a. Spatial-Temporal Sub-Pixel Mapping Based on Swarm Intelligence Theory. Remote Sens., 8(11), 30. https://doi.org/10.3390/rs8110894.

He, D., Zhong, Y., Wang, X., Zhang, L., 2021a. Deep convolutional neural network framework for subpixel mapping. IEEE Trans. Geosci. Remote Sens., 59(11), 9518–9539. https://doi.org/10.1109/TGRS.2020.3032475.

He, D., Shi, Q., Liu, X., Zhong, Y., Zhang, X., 2021b. Deep Subpixel Mapping Based on Semantic Information Modulated Network for Urban Land Use Mapping. IEEE Trans. Geosci. Remote Sens., 59(12), 10628-10646. https://doi.org/10.1109/TGRS.2021.3050824.

He, D., Shi, Q., Liu, X., Zhong, Y., He, D., Zhong, Y., Feng, R., Zhang, L., 2022a.

Generating 2m fine-scale urban tree cover product over 34 metropolises in China based on deep context-aware sub-pixel mapping network. Int. J. Appl. Earth Obser. Geoinf. 106, 102667. https://doi.org/10.1016/j.jag.2021.102667.

He, D., Shi, Q., Liu, X., Zhong, Y., Xia, G., Zhang, L., 2022b. Generating annual high resolution land cover products for 28 metropolises in China based on a deep super-resolution mapping network using Landsat imagery. GIScience & Remote Sensing, 59(1), 2036-2067. https://doi.org/10.1080/15481603.2022.2142727.

He, K., Zhang, X., Ren, S., Sun, J., 2016b. Deep Residual Learning for Image Recognition." IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 770–778. https://doi.org/10.1109/CVPR.2016.90.

IPCC Climate Change 2013: The Physical Science Basis (eds Stocker, T. F. et al.) (Cambridge Univ. Press, 2013).

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. IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Brussels, Belgium, 4704–4707. https://doi.org/10.1109/IGARSS47720.2021.9553499

Kasetkasem, T., Arora, M. K., Varshney, P. K., 2005. Super-resolution land cover mapping using a Markov random field based approach. Remote Sens. Environ. 96(3-4): 302-314. https://doi.org/10.1016/j.rse.2005.02.006.

Li, X., Ling, F., Foody, G. M., Ge, Y., Zhang, Y., Du, Y., 2017. Generating a series of fine spatial and temporal resolution land cover maps by fusing coarse spatial resolution remotely sensed images and fine spatial resolution land cover maps.

Remote Sens. Environ. 196: 293-311. https://doi.org/10.1016/j.rse.2017. 05.011.

Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., & Timofte, R. (2021). SwinIR: Image restoration using swin transformer. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 1833-1844). https://arxiv.org/abs/2108.10257.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollar, P., 2020. Focal loss for dense object detection. IEEE Trans. Pattern Anal. Mach. Intell. 42 (2), 318–327. https://doi.org/10.1109/TPAMI.2018.2858826.

Liu, X., Pei, F., Wen, Y., Li, X., Wang, S., et al., 2019. Global urban expansion offsets climate-driven increases in terrestrial net primary productivity. Nature Communication 10: 5558. https://doi.org/10.1038/s41467-019-13462-1.

Liu, X., Huang, Y., Xu, X., Li, X., Li, X., et al., 2020. High-spatiotemporal-resolution mapping of global urban change from 1985 to 2015. Nature Sustainability 3: 564–570. https://doi.org/10.1038/s41893-020-0521-x.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision, 10012-10022.

Mertens, K. C., Baets, B. D., Verbeke, L. P. C., Wulf, R. D., 2006. A sub-pixel mapping algorithm based on sub-pixel/pixel spatial attraction models. Int. J. Remote Sens. 27 (15): 3293-3310. https://doi.org/10.1080/01431160500497127.

Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. IEEE International

Conference on 3D Vision (3DV), 565-571. https://doi.org/10.1109/3DV.2016.79

Poudyal, A., 2013. Spatial Statistics and Super Resolution Mapping for Precision Agriculture Using VHR Satellite Imagery [M]. Netherland: University of Twente.

Rashed, T., Weeks, J. R., Roberts, D., Rogan, J., Powell, R., 2003. Measuring the physical composition of urban morphology using multiple endmember spectral mixture models. Photogramm. Eng. Remote Sensing, 69, 1011–1020. https://doi.org/10.14358/PERS.69.9.1011

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, Springer, Cham, 234-241.

Shao, Z., Yang, K., Zhou, W., 2018. Performance Evaluation of Single-Label and Multi-Label Remote Sensing Image Retrieval Using a Dense Labeling Dataset. Remote Sens., 10(6), 964. https://doi.org/10.3390/rs10060964

Small, C., 2003. High spatial resolution spectral mixture analysis of urban reflectance. Remote Sens. Environ., 88, 170– 186. https://doi.org/10.1016/j.rse.2003.04.008

Song, M., Zhong, Y., Ma, A., Feng, R., 2019. Multiobjective Sparse Subpixel Mapping for Remote Sensing Imagery. IEEE Trans. Geosci. Remote Sens. 57(7), 4490-4508. https://doi.org/10.1109/TGRS.2019.2891354.

Stokes, E. C., Seto, K., 2018. Characterizing and measuring urban landscapes for sustainability. Environ. Res. Lett. 14, 045002. https://doi.org/10.1088/ 1748-9326/aafab8

Su, Y., 2019. Integrating a scale-invariant feature of fractal geometry into the Hopfield neural network for super-resolution mapping. Int. J. Remote Sens. 40(23): 8933-8954. https://doi.org/10.1080/01431161.2019.1624865.

Verdonck, M. L., Okujeni, A., van der Linden, S., Demuzere, M., De Wulf, R., Van Coillie, F., 2017. Influence of neighbourhood information on ‘Local Climate Zone’mapping in heterogeneous cities. Int. J. Appl. Earth. Obs. Geoinf., 62, 102-113. https://doi.org/10.1016/j.jag.2017.05.017

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., et al., 2017. Attention is all you need. Advances in neural information processing systems, 30.

Wang, Q., Shi, W., Atkinson, P.M., Zhao, Y., 2015. Downscaling modis images with area-to-point regression kriging. Remote Sens. Environ. 166, 191–204. https://doi.org/10.1016/j.rse.2015.06.003.

Wang, J., Zheng, Z., Ma, A., Lu, X., Zhong, Y., 2021. A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation. Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. https:// doi.org/ 10.48550/arXiv.2110.08733

Wang, X., Ross G., Abhinav G., and He, K., 2018. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7794-7803.

Wu, C., Murray, A. T., 2003. Estimating impervious surface distribution by spectral mixture analysis. Remote Sens. Environ., 84, 493–505. https://doi.org/10. 1016/S0034-4257(02)00136-0

Wu, S., Ren, J., Chen, Z., Jin, W., Liu, X., Li, H., . . . Guo, W., 2018. Influence of reconstruction scale, spatial resolution and pixel spatial relationships on the sub-pixel mapping accuracy of a double-calculated spatial attraction model. Remote Sens. Environ., 210, 345-361. https://doi.org/https://doi.org/10.1016/j.rse.2018.03.015

Xia, H., Chen, Y., Song, C., Li, J., Quan, J., Zhou, G., 2022. Analysis of surface urban heat islands based on local climate zones via spatiotemporally enhanced land surface temperature. Remote Sens. Environ., 273, 112972. https://doi.org/10.1016/j.rse.2022.112972

Xiao, J., Moody, A., 2005. A comparison of methods for estimating fractional green vegetation cover within a desert-to-upland transition zone in central New Mexico, USA. Remote Sens. Environ., 98(2-3), 237-250. https://doi.org/10.1016/j.rse.2005.07.011

Xu, X., Tong, X., Plaza, A., Li, J., Zhong, Y., Xie, H., et al., 2018. A New Spectral-Spatial Sub-Pixel Mapping Model for Remotely Sensed Hyperspectral Imagery. IEEE Trans. Geosci. Remote Sens. 56(11): 6763-6778. https://doi.org/10.1109/TGRS.2018.2842748.

Zanaga, D., Van De Kerchove, R., De Keersmaecker, W., Souverijns, N., Brockmann, et al., 2021. ESA WorldCover 10 m 2020 v100. https://doi.org/10.5281/zenodo.5571936.

Zhang, G., Ghamisi, P., Zhu, X., 2019. Fusion of heterogeneous earth observation data for the classification of local climate zones. IEEE Trans. Geosci. Remote Sens. 57 (10), 7623–7642. https://doi.org/10.1109/TGRS.2019.2914967

Zhong, Y., Wu, Y., Xu, X., Zhang, L., 2015. An Adaptive Subpixel Mapping Method Based on MAP Model and Class Determination Strategy for Hyperspectral Remote Sensing Imagery. IEEE Trans. Geosci. Remote Sens. 53 (3): 1411-1426. https://doi.org/10.1109/TGRS.2014.2340734.