Supplementary Material - Learning without Exact Guidance: Updating Large-scale High-resolution Land Cover Maps from Low-resolution Historical Labels

Zhuohong Li[1*], Wei He[1*], Jiepan Li[1], Fangxiao Lu[1], Hongyan Zhang[1,2†]
[1]Wuhan University [2]China University of Geosciences
{ashelee, weihe1990, jiepanli, fangxiaolu}@whu.edu.cn, zhanghongyan@cug.edu.cn

In this supplementary, we provide a detailed description of the proposed framework and dataset organization. More experimental results are also presented. These three parts are demonstrated sequentially.

A . Details of Paraformer

In the proposed Paraformer, a robust feature extractor parallel hybrids a downsampling-free CNN branch with a Transformer branch. To demonstrate the structures of CNN and Transformer branches more clearly, Figures S1 and S2 show the basic units of CNN and Transformer branches.

In this section, we focus on illustrating the basic units of the CNN branch in detail. The resolution preserving (RP) block shown in Figure S1 was firstly proposed in our previous work: L2HNet[1]. Here, we use $\mathbf{I}^{(b)}$, $\mathbf{M}^{(b)}$, and $\mathbf{F}^{(b)}$ to denote the input, middle, and fusion feature maps of the b-th block. Specifically, the input feature map of the first block is generated by a 3×3 convolution input layer with four input channels (i.e., the R-G-B-NIR bands of the images) and C_I output channels. Therefore, the input feature map of the first block can be expressed as $\mathbf{I}^{(1)} \in \mathbb{R}^{N \times C_I \times H_I \times W_I}$, where N represents the batch size and $C_I \times H_I \times W_I$ represents the channels, height, and width of the map, respectively. For the operation symbols, we represent a one-stride $(n \times n)$ convolutional layer with C_1 input channels and C_2 output channels as $W^{n\times n}_{C_1,C_2}$ (with padding when $n = 3, 5$). In addition, the batch normalization layer with the rectified linear unit (ReLU) function is simply denoted by $bn(\cdot)$, and $*$ represents the convolution operator. Based on this, the multi-scale feature fusion process from $\mathbf{I}^{(b)}$ to $\mathbf{M}^{(b)}$ can be described as:

$$\mathbf{M}^{(b)} = \text{concat}\begin{bmatrix} bn(\mathbf{I}^{(b)} * W^{1\times 1}_{C_I,C_I}), \\ bn(\mathbf{I}^{(b)} * W^{3\times 3}_{C_I,\frac{C_I}{2}}), \\ bn(\mathbf{I}^{(b)} * W^{5\times 5}_{C_I,\frac{C_I}{4}}) \end{bmatrix}. \tag{S1}$$

As shown in Eq. (S1), the kernel numbers of the multi-scale convolutional layers are set to $\omega = \{\sqrt{2^{(1-n)}}\}_{n=1,3,5}$, which is inversely proportional to their kernel sizes.

Subsequently, we adopt a 1×1 convolutional layer after the concatenation of the multi-scale layers to reduce the dimensions of $\mathbf{M}^{(b)}$ from $C_I\,(1 + 1/2 + 1/4)$ to C_I, thus keeping the blocks lightweight. In addition, to maintain the shallow features and put residual learning into effect, a shortcut connection is adopted from $\mathbf{I}^{(b)}$ to $\mathbf{F}^{(b)}$. As a result, the final $\mathbf{F}^{(b)}$ can be described as:

$$\mathbf{F}^{(b)} = bn(\mathbf{M}^{(b)} * W^{1\times 1}_{C_I(1+1/2+1/4),C_I}) + \mathbf{I}^{(b)}. \tag{S2}$$

From Eqs. (S1)–(S2), $\mathbf{F}^{(b)}$ is a multi-scale fusion feature map with the same size, channels, and resolution as $\mathbf{I}^{(b)}$. Based on the structures, the RP block synchronously combines the multi-scale fusion attributes and residual learning ability to appropriately prevent the feature resolution reduction caused by the over-downsampling. Furthermore, after the feature fusion of several RP blocks, the predictions and corresponding CP maps are generated through a classifier that is constructed by a SoftMax function and a 1×1 convolutional layer $W^{1\times 1}_{C_I,L}$, where $C_I = 128$ is the channel numbers maintained in the entire backbone and L is the output channel determined by the number of land-cover categories.

Moreover, the basic unit of the Transformer branch is shown in Figure S2, which includes a layer normalization (Layer Norm), multi-head self-attention (MSA), and multi-layer perception (MLP).

B . Details of Study area and using data

In this section, we demonstrate the details of two large-scale datasets. Figures S3 and S4 show the location, coverage, and data samples of the Chesapeake Bay dataset and the Poland dataset. Tables S1 and S2 show the land-cover class unifying relations between the LR labels and HR ground truths.

The Chesapeake Bay dataset: The Chesapeake Bay, as the largest estuary in the USA, is about 320 kilometers long from its northern headwaters in the Susquehanna River to its outlet in the Atlantic Ocean. The Chesapeake Bay watershed covers about 160,000 km^2 areas of the surrounding drainage basin. It includes six administrative states of the

[1]https://doi.org/10.1016/j.isprsjprs.2022.08.008

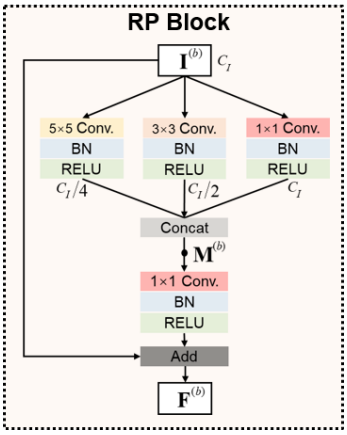

Figure S1. An illustration of an RP block. The input map $\mathbf{I}^{(b)}$ is sampled by three convolutional layers with sizes of 1×1, 3×3, and 5×5, and the convolution kernels in each layer are set to the proportion of ω for preventing feature resolution reduction caused by the over downsampling.

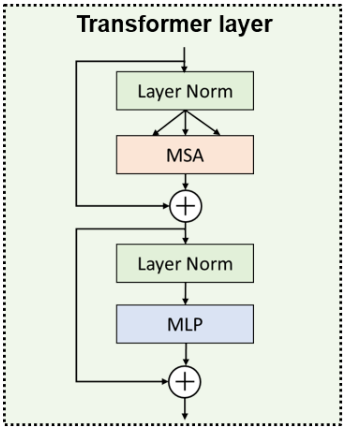

Figure S2. An illustration of a Transformer layer. The layer includes layer normalization (Layer Norm), multi-head self-attention (MSA), and multi-layer perception (MLP).

USA which are New York, Pennsylvania, Delaware, Maryland, Virginia, and West Virginia. The Chesapeake Bay watershed contains various landforms with abundant ecological communities and diverse flora which brings challenges for large-scale high-resolution (HR) land-cover mapping. The Chesapeake Bay dataset, grouped by Microsoft[2], contains 1-meter resolution images and a 30-meter resolution land-cover product as the training data pairs and also contains a 1-meter resolution ground reference for assessment. Figure S3 illustrates the location, Digital Elevation Model (DEM), numbers of the tiles, and data samples of the Chesapeake Bay dataset. In more detail, the data sources are shown as follows:

1. The HR remote sensing images with 1-meter resolution were captured by the airborne platform of the U.S. Department of Agriculture's National Agriculture Imagery Program (NAIP). The images contained four bands of red, green, blue, and near-infrared.

2. The rough historical land-cover products with 30-meter resolution were collected from the National Land Cover Database of the United States Geological Survey (USGS). The NLCD data contains 16 land-cover types and is utilized as the labels during the training process of the proposed Paraformer framework.

3. The HR ground references with 1-meter resolution were obtained from the Chesapeake Bay Conservancy Land Cover (CCLC) project. The CCLC data were interpreted based on the 1-meter NAIP imagery and LiDAR data containing six land-cover types. In this paper, the CCLC data were only used as the ground reference for quantitative and qualitative assessment and were not involved in the framework training or optimization process.

The Poland dataset: The Republic of Poland has a territory traversing the Central European Plain and extends from Baltic Sea in the north to the Sudeten and Carpathian Mountains in the south. Topographically, with the flat, long sea lie and the hilly, mountainous terrain, the landscape of Poland is characterized by diverse landforms, river systems, and ecosystems. The Poland dataset contained 14 Provinces of Poland which included the Provinces of Pomorskie, Lódzkie, Lubuskie, Dolnoslaskie, and so on. The Poland dataset contains 0.25-meter resolution images, three kinds of 10-meter resolution land-cover products, and a 30-meter resolution land-cover product to construct the training data pairs with different combinations. Figure S4 demonstrated the location, DEM, numbers of the tiles, and data samples of the Poland dataset. In more detail, the data sources are shown as follows:

1. The HR remote sensing images with 0.25-meter and 0.5-meter resolution were collected from the Land-Cover.ai dataset where the image sources are from the public geodetic resource used in the Land Parcel Identification System (LPIS). The images contained three bands of red, green, and blue.

2. The rough historical labeled data with 10-meter resolution were collected from three types of global land-cover products which were (1) The FROM_GLC10 provided by the Tsinghua University, (2) The ESA_WorldCover v100 provided by the European Space Agency (ESA), and (3) The ESRI 10-meter global land cover (abbreviated as ESRI_GLC10) provided by the ESRI Inc. and IO Inc. The 30-meter resolution labeled data were collected from the 30-meter global land-cover product GLC_FCS30 provided by the Chinese Academy of Sciences (CAS).

[2]https://lila.science/datasets/chesapeakelandcover

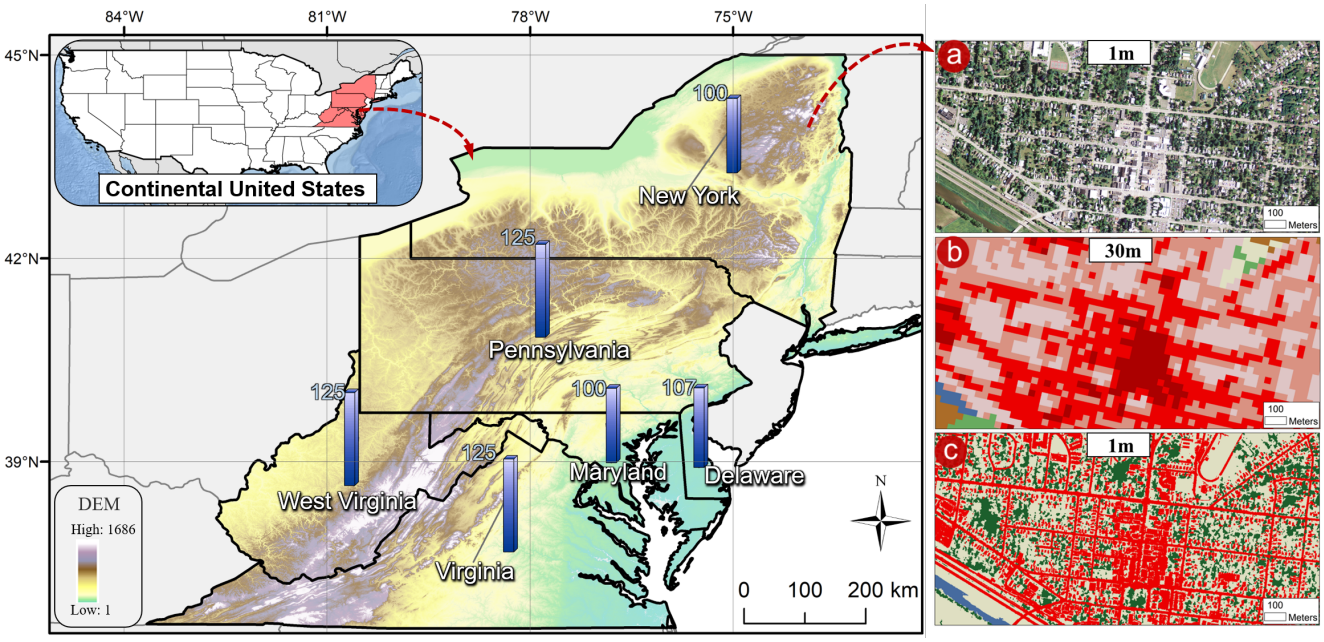

Figure S3. The Chesapeake Bay dataset covers six states of the USA, including the data sources of (a) The 1-m NAIP imagery, (b) The 30-m NLCD labels, and (c) The 1-m ground truth. The blue columns show the number of tiles.

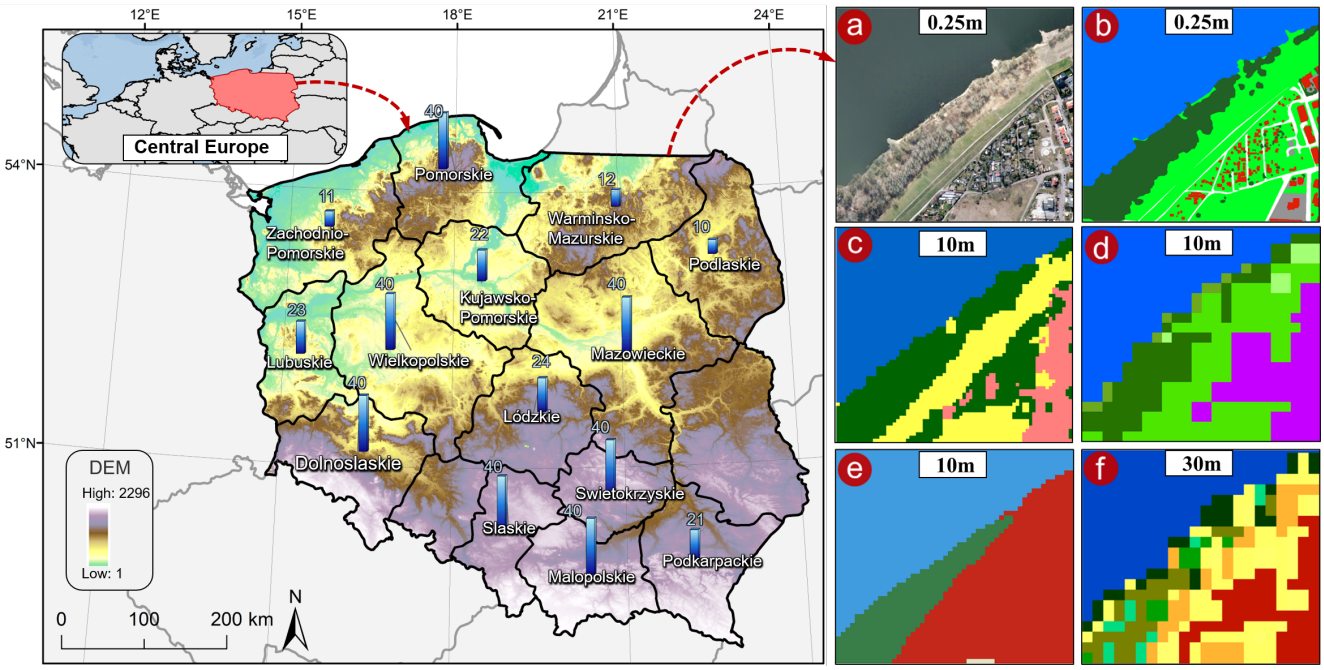

Figure S4. The Poland dataset covers 14 provinces of the country, including the data sources of (a) The 0.25-m imagery, (b) The 0.25-m ground truth, (c) The 10-m FROM_GLC10, (d) The 10-m ESA_GLC10, (e) The 10-m ESRI_GLC10, and (f) The 30-m GLC_FCS30. The blue columns show the number of tiles.

3. The HR ground references were obtained from the OpenEarthMap dataset provided by the University of Tokyo. The ground references were interpreted based on the 0.25-meter and 0.5-meter resolution LPIS imagery and contained five land-cover types.

C . Supplementary experiment results

To comprehensively demonstrate the performance of Paraformer, we sequentially illustrate supplementary experiment results as follows:

Visual results of the Chesapeake Bay dataset: Figures

Name	NLCD	CCLC	
Affiliation	USGS, USA	Chesapeake Conservancy, USA	**Target classes**
Resolution	30 meters	1 meter	
Class	Developed open space Developed low c Developed medium Developed high	Roads Building Barren	Built-up
	Deciduous forest Evergreen forest Mixed forest Woody wetland	Tree canopy	Tree canopy
	Barren land Shrub/Scrub Grassland Pasture/Har Cultivated crops Herbaceous wetlands	Low vegetation	Low vegetation
	Open water	Water	Water
Note:	USGS= United States Geological Survey;		

Table S1. Land-cover class unifying relations between the LR labels (NLCD) and HR ground truths. The first column shows the legends of LR labels. The last column shows the target classes for accuracy assessment and their colors shown in the visual results.

Name	FROM_GLC10	ESRI_GLC10	ESA_GLC10	GLC_FCS30	OpenEarthMap	
Affiliation	THU, China	Esri&IO, USA	ESA, Europe	CAS, China	UTokyo, Japan	**Target classes**
Resolution	10 meters	10 meters	10 meters	30 meters	0.25/0.5 meter	
Class	Forest	Trees	Trees	Deciduous broadleaved forest Open deciduous broadleaved forest Evergreen needle-leaved forest Mixed leaf forest	Tree	Tree canopy
	Shrubland	Scrub/Shrub	Shrubland	Orchard Sparse shrubland	Rangeland	
	Grassland	Grassland	Grassland	Grassland Herbaceous cover		Low vegetation
	Cropland	Crops	Cropland	Rainfed cropland Irrigated cropland	Agriculture land	
	Impervious area	Built Area	Built-up	Impervious surfaces	Building Road Developed space	Built-up
	Water body	Water	Open water	Water body	Water	Water
Note:	THU=Tsinghua University; ESRI=ESRI Inc.; IO=IO Inc.; ESA=European Space Agency; CAS=Chinese Academy of Science; UTokyo=The University of Tokyo					

Table S2. Land-cover class unifying relations among four types of LR labels and HR ground truths. The 1–4 column shows the legends of LR labels. The last column shows the target classes for accuracy assessment and their colors shown in the visual results.

S5–S7 demonstrate one large-scale and two small-scale visual comparisons between Paraformer and four typical methods. From these visual results, the Paraformer is able to update accurate HR land-cover maps from the HR images source and LR label guidance. TransUNet shows clear urban patterns but underestimates the built-up areas. UNet, as a typical CNN-based encoder-decoder framework, has a rough result consistent with the LR labels. L2HNet, as the state-of-the-art method for updating HR land-cover results from LR labels, shows an accurate edge of land objects but still has incorrect fragments in the results. RF, as a pixel-to-pixel learning method, has the finest edges but lacks of contextual information learning, which causes insufficient results overall (underestimating the water and low vegetation).

Visual results of the Poland dataset: Figures S8–S11 show the visual comparison between Paraformer and the other three typical methods which are trained with different LR land-cover labels. From the visual results, the Paraformer is able to refine a clear land-cover pattern from different types of LR land-cover labels. Even though some of the classes in the demonstration patches are not contained, Paraformer can jointly capture the local and global contexts and produce HR results that are consistent with the HR images.

Further discussion: In this part, we demonstrate more details of the loss fluctuation and supplementary large-scale experiments in China. Figure S12 shows the loss functions of $\mathcal{L}_{\text{ce}}$ and $\mathcal{L}_{\text{mce}}$ during framework training. The two training losses are stable to decrease in six states of

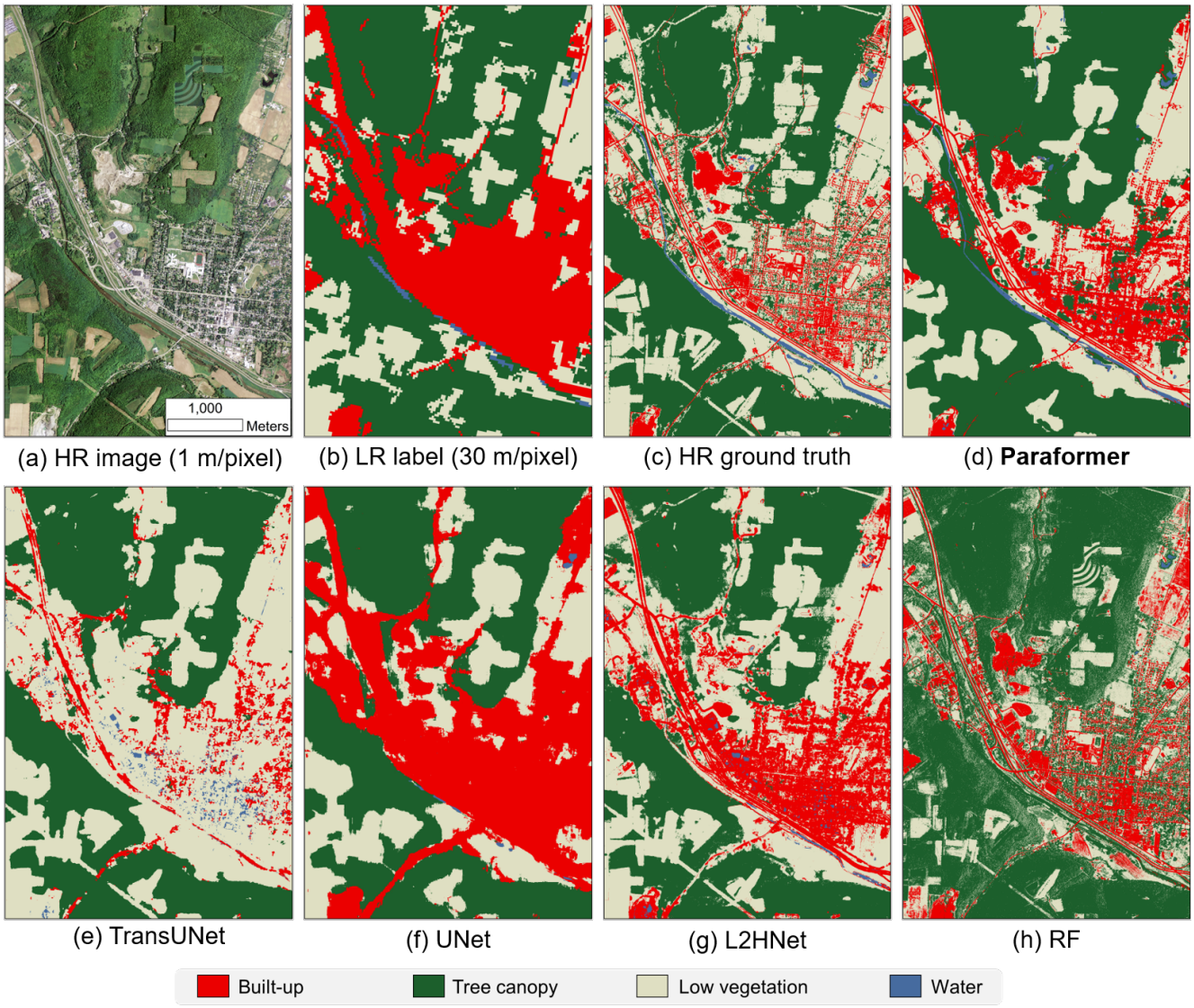

Figure S5. Demonstration of the training data and visual comparisons of the **Paraformer** and other typical methods on the Chesapeake Bay dataset with four unified classes. (a) HR image. (b) LR label. (c) HR ground truth. (d) land-cover mapping result of Parafomer. (e–h) land-cover mapping results of four typical methods.

the Chesapeake Bay dataset. This further indicates the robustness of the pseudo-label-assisted training (PLAT) module in learning from inexact LR labels. To further discuss the applicability of Paraformer, we conduct large-scale experiments in the whole of Wuhan City, China. Based on our previous work on SinoLC-1[3] (i.e., the first 1-m land-cover map of China), we regard the intersected results of three 10-m land-cover products (ESA_GLC10, Esri_GLC10, and FROM_GLC10) as the LR training labels of 1-m Google Earth images. As shown in Fig. S13 (a), the 1-m Google Earth image reveals clear land details. Fig. S13 (b–d) demonstrates three types of 10-m land-cover products. Compared with the original 1-m SinoLC-1 shown in Fig. S13 (e), the Paraformer is able to refine a more accurate urban pattern shown in Fig. S13 (f). For the whole of Wuhan City, the reported overall accuracy (OA) of SinoLC-1 is 72.40%. The updated results of the proposed Paraformer reach 74.98% with a 2.58% improvement.

[3]https://doi.org/10.5194/essd-15-4749-2023

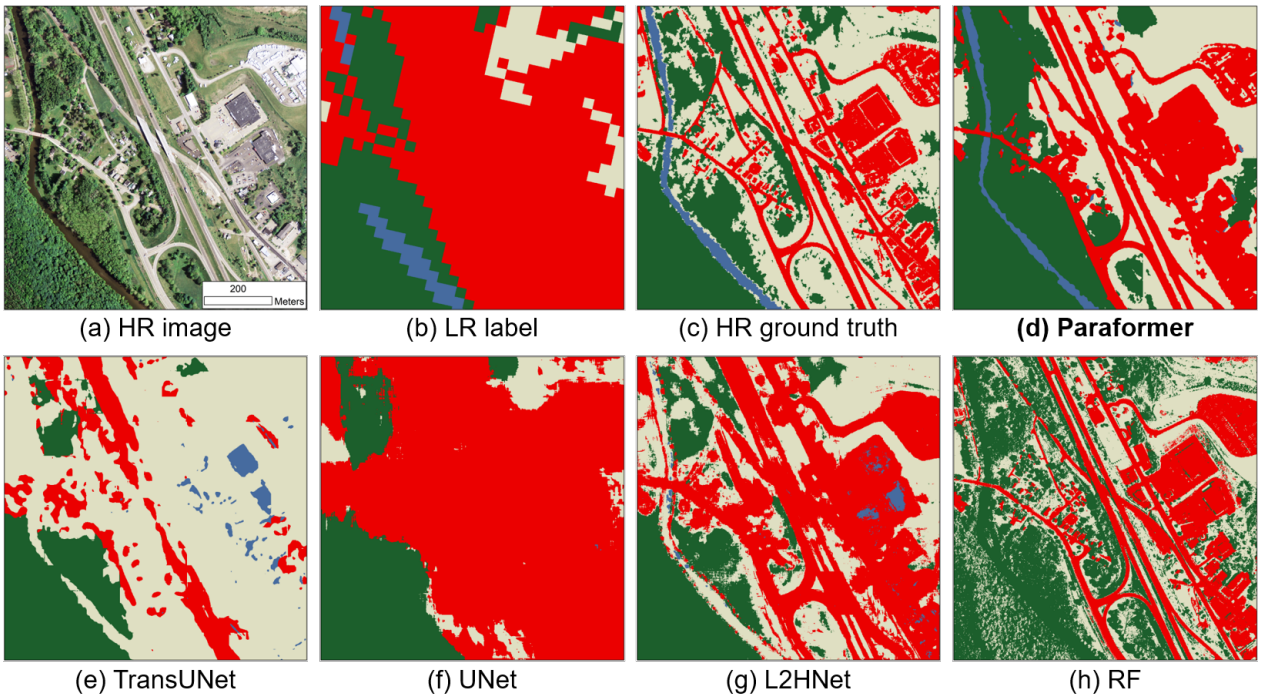

Figure S6. Sample A of the training data and visual comparisons of the **Paraformer** and other typical methods on the Chesapeake Bay dataset with four unified classes. (a) HR image. (b) LR label. (c) HR ground truth. (d) land-cover mapping result of Parafomer. (e–h) land-cover mapping results of four typical methods.

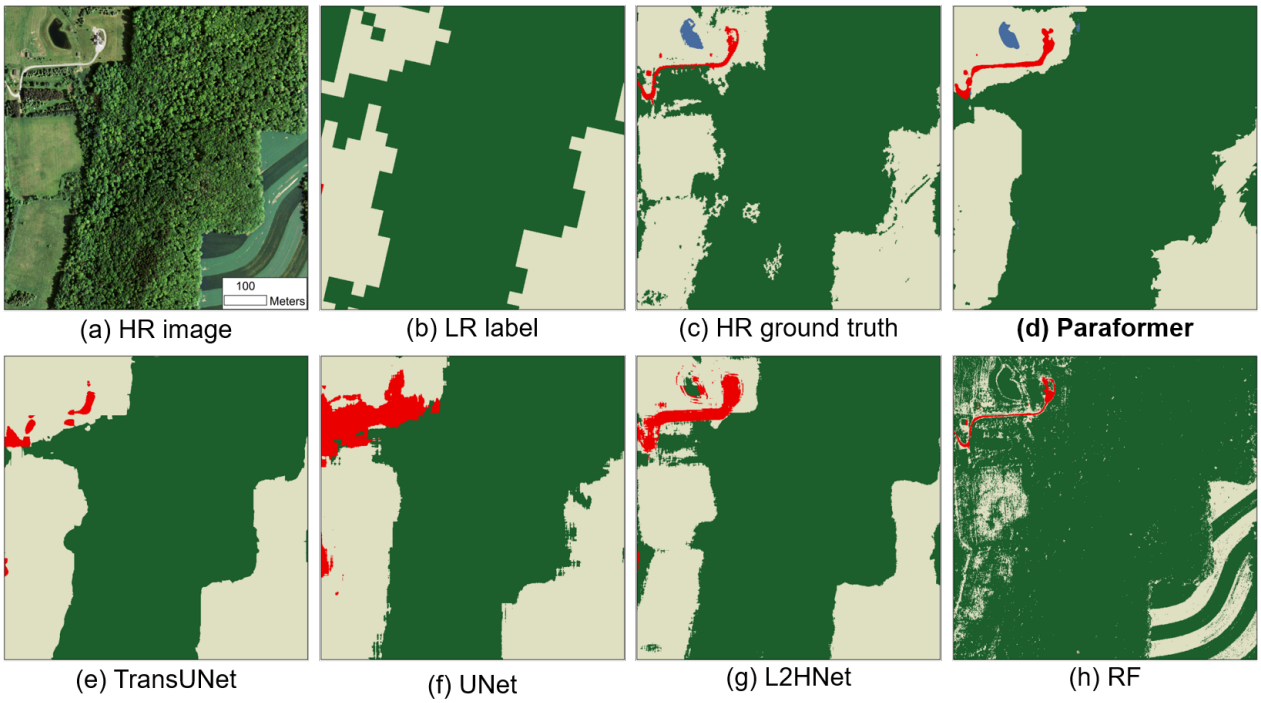

Figure S7. Sample B of the training data and visual comparisons of the **Paraformer** and other typical methods on the Chesapeake Bay dataset with four unified classes. (a) HR image. (b) LR label. (c) HR ground truth. (d) land-cover mapping result of Parafomer. (e–h) land-cover mapping results of four typical methods.

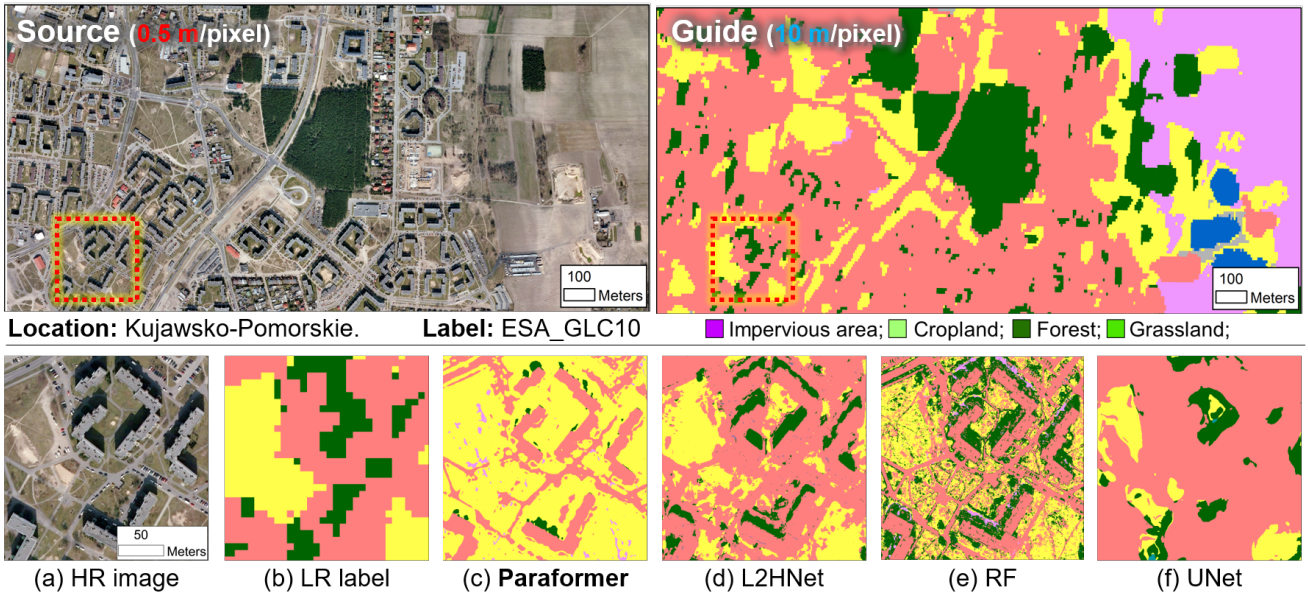

Figure S8. The visual results of Poland dataset with 10-m ESA_GLC10 training labels. (a) The 0.5-m image, (b) The 10-m label sampled from the ESA_GLC10. (c) Result of Paraformer. (d) Result of L2HNet. (e) Result of RF. (f) Result of UNet.

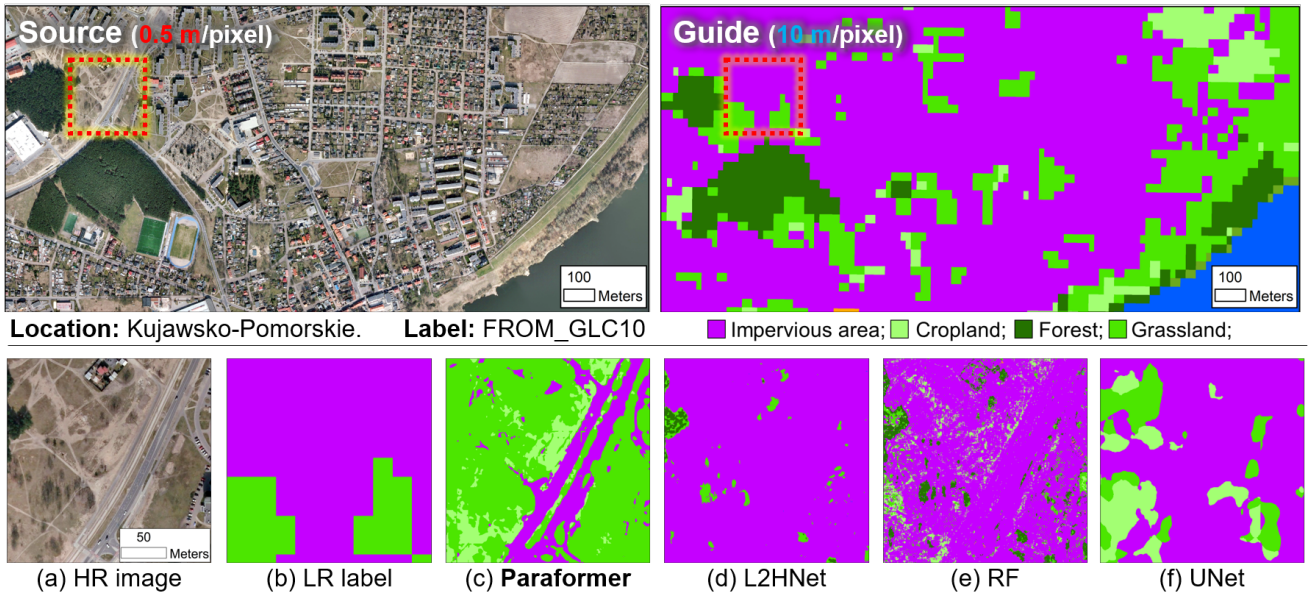

Figure S9. The visual results of Poland dataset with 10-m FROM_GLC10 training labels. (a) The 0.5-m image, (b) The 10-m label sampled from the FROM_GLC10. (c) Result of Paraformer. (d) Result of L2HNet. (e) Result of RF. (f) Result of UNet.

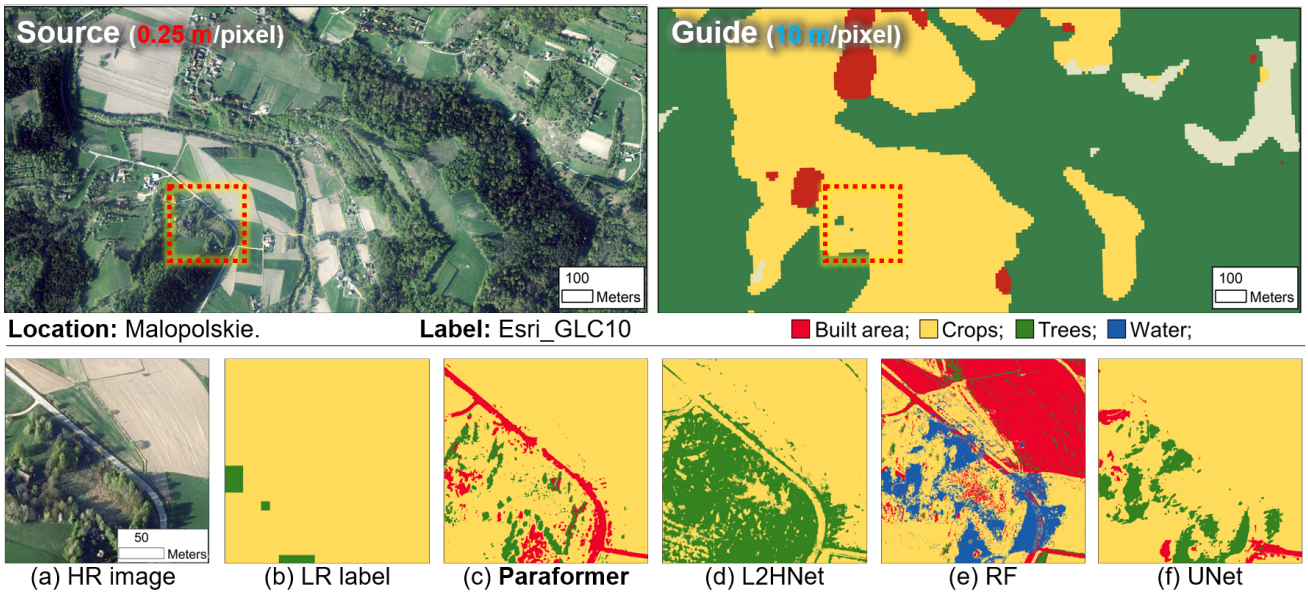

Figure S10. The visual results of Poland dataset with 10-m Esri_GLC10 training labels. (a) The 0.25-m image, (b) The 10-m label sampled from the Esri_GLC10. (c) Result of Paraformer. (d) Result of L2HNet. (e) Result of RF. (f) Result of UNet.

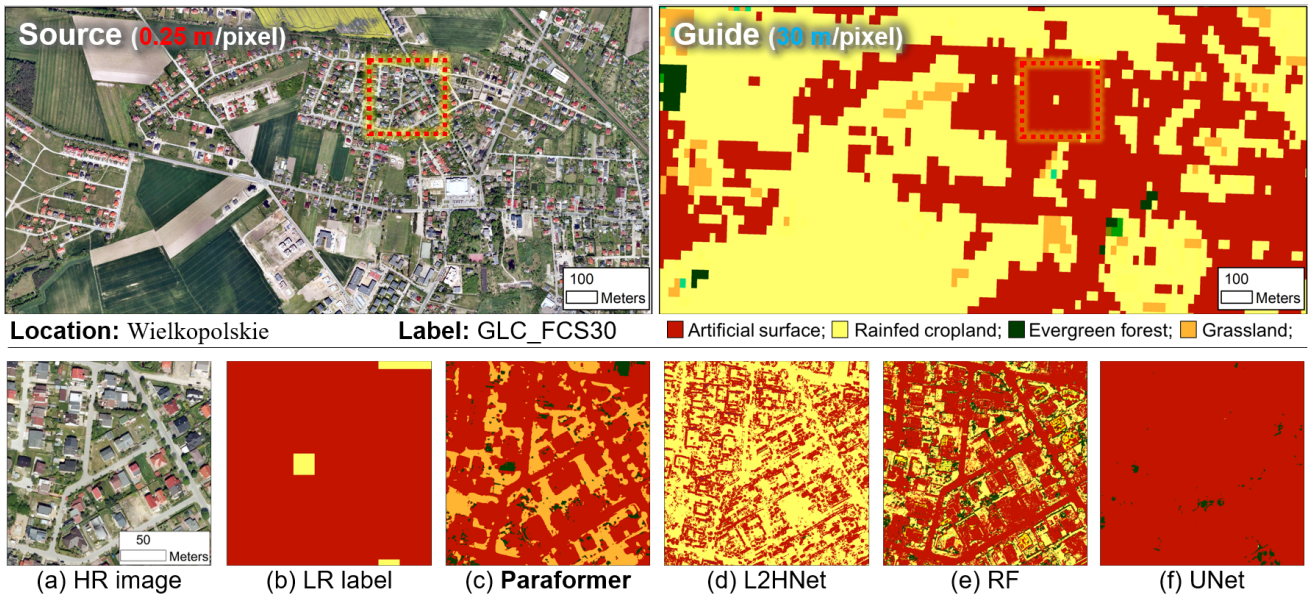

Figure S11. The visual results of Poland dataset with 30-m GLC_FCS30 training labels. (a) The 0.5-m image, (b) The 10-m label sampled from the GLC_FCS30. (c) Result of Paraformer. (d) Result of L2HNet. (e) Result of RF. (f) Result of UNet.

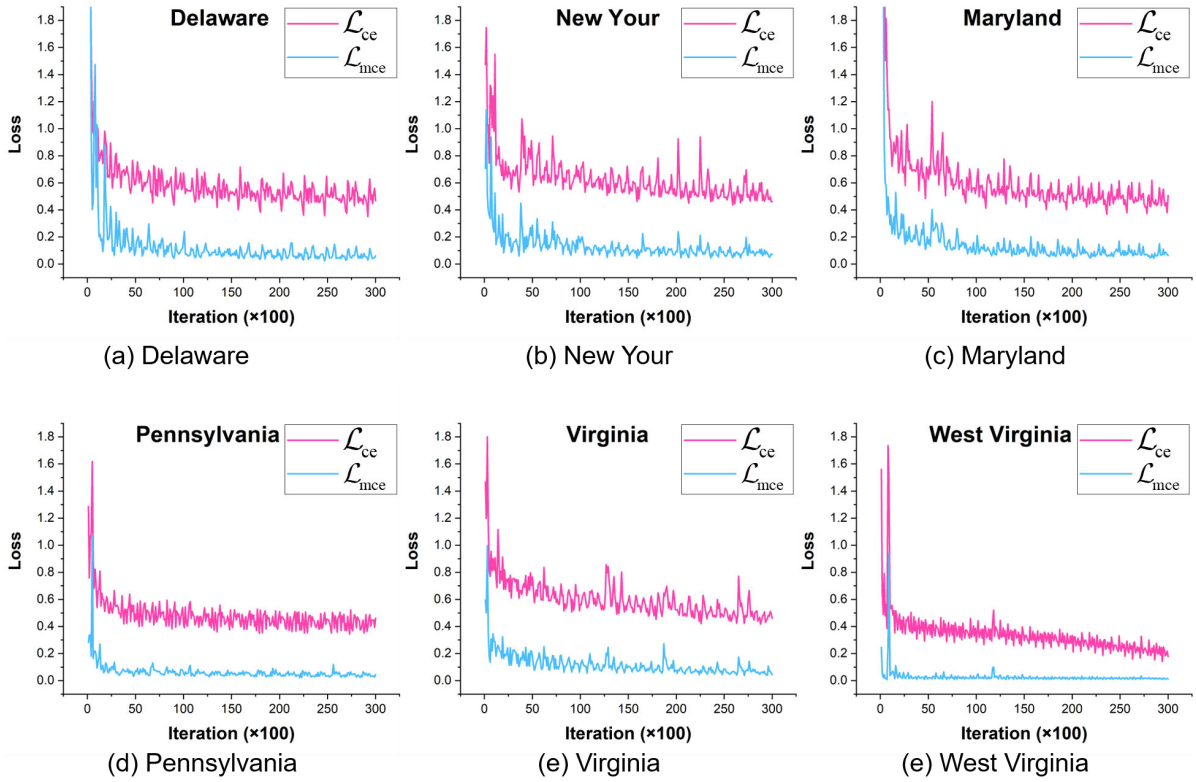

Figure S12. Demonstration of the loss functions $\mathcal{L}_{ce}$ and $\mathcal{L}_{mce}$ during framework training. Sub-figures (a)–(e) demonstrate the training process in six states of the Chesapeake Bay dataset.

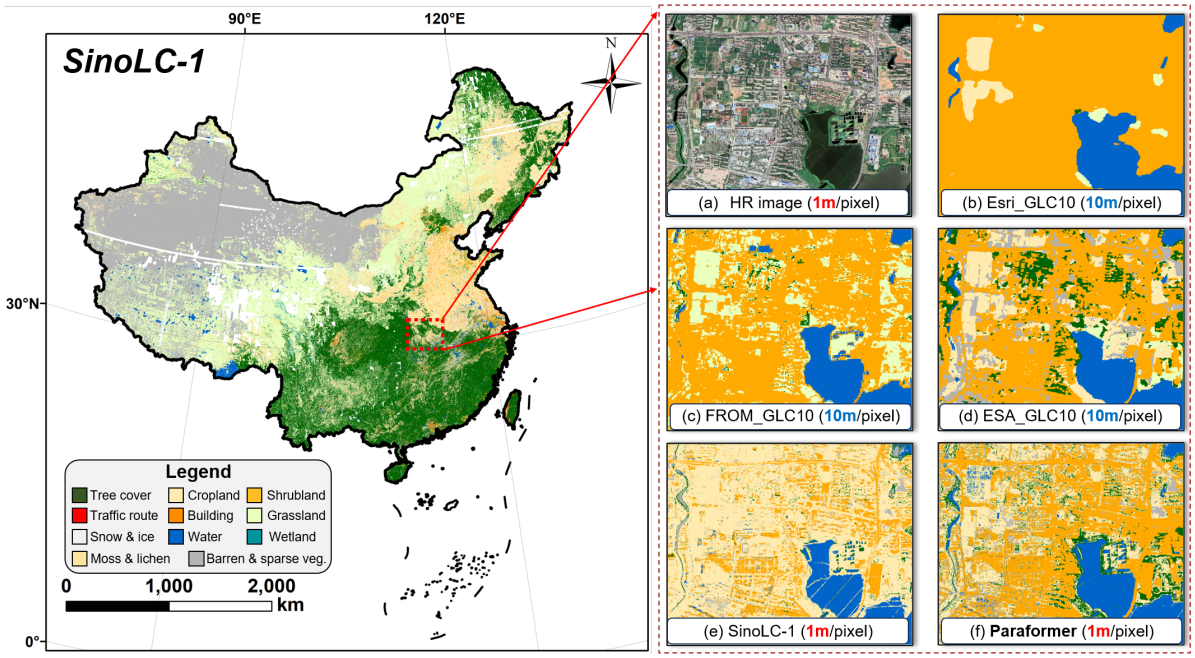

Figure S13. Demonstration of the supplementary experiments of SinoLC-1 dataset. The visual results are sampled from Wuhan, China. The Paraformer is used to update the 1-m land-cover map in the whole of Wuhan City, reporting an OA of 74.98%.