

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

Zhuohong Li<sup>1\*</sup>, Wei He<sup>1\*</sup>, Jiepan Li<sup>1</sup>, Fangxiao Lu<sup>1</sup>, Hongyan Zhang<sup>1,2†</sup>

<sup>1</sup>Wuhan University <sup>2</sup>China University of Geosciences

{ashelee, weihe1990, jiepanli, fangxiaolu}@whu.edu.cn, zhanghongyan@cug.edu.cn

## Abstract

*Large-scale high-resolution (HR) land-cover mapping is a vital task to survey the Earth’s surface and resolve many challenges facing humanity. However, it is still a non-trivial task hindered by complex ground details, various landforms, and the scarcity of accurate training labels over a wide-span geographic area. In this paper, we propose an efficient, weakly supervised framework (Paraformer) to guide large-scale HR land-cover mapping with easy-access historical land-cover data of low resolution (LR). Specifically, existing land-cover mapping approaches reveal the dominance of CNNs in preserving local ground details but still suffer from insufficient global modeling in various landforms. Therefore, we design a parallel CNN-Transformer feature extractor in Paraformer, consisting of a downsampling-free CNN branch and a Transformer branch, to jointly capture local and global contextual information. Besides, facing the spatial mismatch of training data, a pseudo-label-assisted training (PLAT) module is adopted to reasonably refine LR labels for weakly supervised semantic segmentation of HR images. Experiments on two large-scale datasets demonstrate the superiority of Paraformer over other state-of-the-art methods for automatically updating HR land-cover maps from LR historical labels.*

## 1. Introduction

Land-cover mapping is a semantic segmentation task that gives each pixel of remote-sensing images a land-cover class such as “cropland” or “building” [14]. The land-cover data should be continuously updated since nature and human activities frequently change the landscape [37]. As sensors and satellites developed, massive high-resolution (HR) remote-sensing images ( $\leq 1$  meter/pixel) could be easily obtained [1]. Rapid large-scale HR land-cover map-

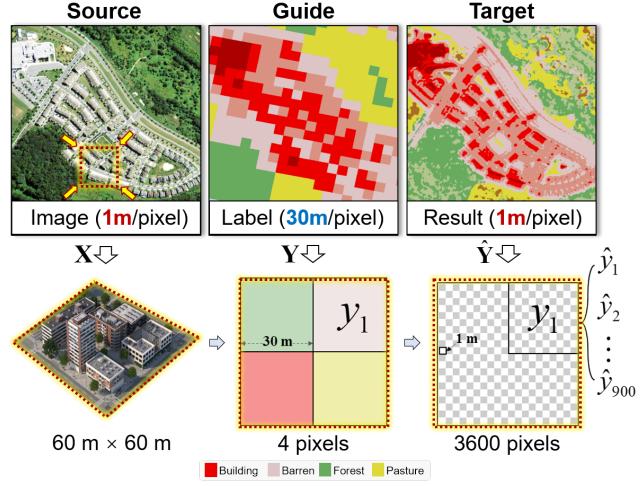


Figure 1. Illustration of resolution mismatched issue in using the HR remote-sensing image (**Source**) and LR historical labels (**Guide**) to generate HR land-cover results (**Target**).

ping is even more critical to facilitate downstream applications as the up-to-date HR land-cover data can accurately describe the land surface [21, 27, 55]. However, the complex ground details reflected by HR images and various landforms over wide-span areas still challenge the periodic updating of large-scale HR land-cover maps [28].

The advanced methods for HR land-cover mapping have been dominated by the convolutional neural network (CNN) for many years. Although CNN-based models can finely capture local details for semantic segmentation of HR images, the intrinsic locality of convolution operations still limits their implementation in various landforms across larger areas [2]. Recently, Transformer has achieved tremendous success in semantic segmentation [5, 18, 34] and large-scale applications of Earth observation [11, 41, 48]. It adopts multi-head self-attention mechanisms to model global contexts but struggles in the representation of local details due to the shortage of low-level features [10, 48]. Besides, current methods with either CNN or Transformer structures generally rely on sufficient exact training labels by adopting a fully supervised strategy

\*Indicates equal contribution. †Corresponding author. The code and data are released at <https://github.com/LiZhuoHong/Paraformer>

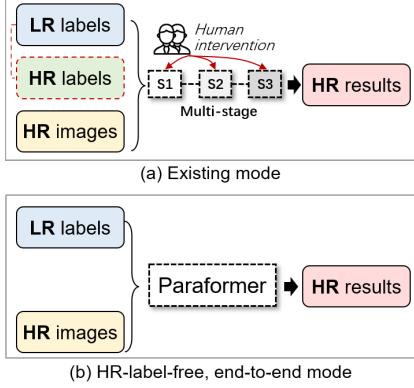


Figure 2. Two modes of large-scale HR land-cover mapping with LR labels. (a) Existing modes either rely on partial HR labels or require non-end-to-end training with human interventions. (b) **Paraformer** aims to form a mode that is HR-label-free and end-to-end trainable.

[20, 32, 39]. However, creating accurate HR land-cover labels for large-scale geographic areas is extremely time-consuming and laborious [6, 37].

Fortunately, many low-resolution (LR) land-cover data with large coverage have already emerged in the past decades [9, 22, 44, 56]. Utilizing these LR historical land-cover data as alternative guidance is a way to alleviate the scarcity of HR labels [29]. Nevertheless, the unmatched training pairs of HR images and inexact LR labels posed a challenge for fully supervised methods. Moreover, due to the different applied scenarios, existing weakly supervised semantic segmentation methods for natural scenes (e.g., learning from bounding box or image-level labels) are not applicable in handling the challenge as well [15, 23, 24, 57].

Distinctively, the incorrect samples of LR land-cover labels are brought by satellites in different spatial resolutions during Earth observation. As shown in Figure 1, the objects in a  $60m \times 60m$  area can be clearly observed from the HR (1 m/pixel) image X. However, in the LR (30 m/pixel) label Y, the area is only labeled by four pixels. To produce the 1-m land-cover result  $\hat{Y}$ , a labeled pixel  $y_1$  needs to provide guiding information for 900 target pixels  $\{\hat{y}_1, \hat{y}_2 \dots \hat{y}_{900}\}$ , which raises a serious geospatial mismatch. How to reasonably exploit LR labels as the only guidance for semantic segmentation of large-scale HR satellite images is a particular problem shared in the fields of Earth observation and computer vision [28, 31, 37]. By summarizing the state-of-the-art methods of exploiting LR labels for large-scale HR land-cover mapping, there are still two main problems:

1. *For the wide-span application areas, existing feature extractors are difficult to jointly capture local details from HR images and model global contexts in various landforms at once [29, 54].*
2. *For the mismatch of training pairs, existing pipelines, as shown in Figure 2 (a), either still rely on partial HR labels or require non-end-to-end optimization with human interventions [12, 27].*

To resolve these problems, as shown in Figure 2 (b), we propose the Paraformer as an HR-label-free, end-to-end framework to guide large-scale HR land-cover mapping with LR land-cover labels. Specifically, Paraformer parallelly hybrids a downsampling-free CNN branch with a Transformer branch to jointly capture local and global contexts from the large-scale HR images and adopts a pseudo-label-assisted training (PLAT) module to dig up reliable information from LR labels for framework training.

The main contributions of this study are summarized as follows: **(a)** We introduce an efficient, weakly supervised Paraformer to facilitate large-scale HR land-cover mapping by getting rid of the well-annotated HR labels and human interventions during framework training; **(b)** a downsampling-free CNN branch is parallelly hybridized with a Transformer branch to capture features with both high spatial resolution and deep-level representation. The structure aims to globally adapt large-scale, various landforms and locally preserve HR ground details; **(c)** the PLAT module iteratively intersects primal predictions and LR labels to constantly refine labeled samples for guiding the framework training. It provides a concise way to update large-scale HR land-cover maps from LR historical data.

## 2. Related Work

**Land-cover mapping approach:** In the early stage, pixel-to-pixel classification methods, such as decision tree [19], random forest [7], and support vector machine [40], were popular in the land-cover mapping of multi-spectral LR images. However, these methods generally ignore contextual information and have fragmented results in HR cases, as optical HR images contain abundant spatial details but limited spectral features [29]. With the development of data-driven semantic segmentation, many CNN-based models were widely used in land-cover mapping of HR images [37, 52, 53]. Besides, as an alternative architecture, Transformer shows great power in capturing global contexts with sequence-to-sequence modeling [3, 10, 30] and demonstrates outstanding performance in many large-scale applications of Earth observation, such as building extraction [25, 41], road detection [11], and land-object classification [47]. Besides, many works developed new ways by saving labor to produce finer labels with the Segment Anything Model (SAM) [35, 50]. However, sufficient exact training labels are the foundation for large-scale applications of both CNN- and Transformer-based methods. The scarcity of HR labels still impedes these fully supervised approaches from large-scale HR land-cover mapping.

**Land-cover labeled data:** Creating large-scale HR labels via manual and semi-manual annotations is extremely time-consuming and expensive [17, 36]. Therefore, exiting HR land-cover data is generally limited to small scales. E.g., the LoveDA dataset contains 0.3-m land-cover data, covering

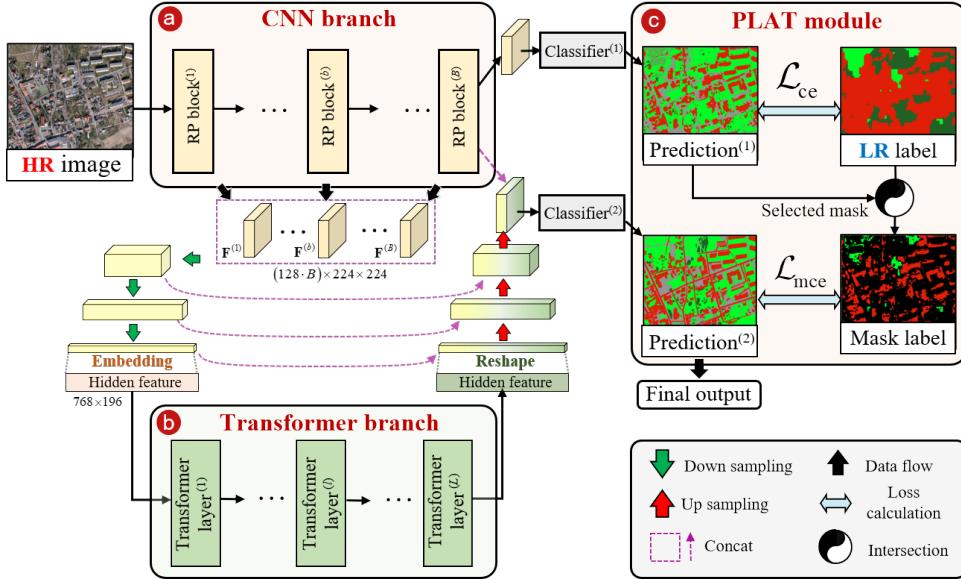


Figure 3. Overall workflow of Paraformer. The framework only takes the HR images and LR labels as training input and includes three components: (a) CNN-based resolution-preserving branch, (b) Transformer-based global-modeling branch, and (c) Pseudo-Label-Assisted Training (PLAT) module.

536.15  $km^2$  of China [46]. The Agri-vision dataset contained 0.1-m labeled data, covering 560  $km^2$  of the USA [13]. In the contract, the LR land-cover data generally has a larger coverage. E.g., the United States Geological Survey cyclically updates 30-m land-cover data covering the whole USA [49]. The European Space Agency (ESA) has updated an annual 10-m global land-cover data since 2020 [44]. These LR data can be seen as an alternative label source for guiding large-scale HR land-cover mapping. However, massive inexactly labeled samples still hinder them from being practicable.

**Strategies for LR historical label mining:** To alleviate the scarcity of accurate labels in large-scale HR land-cover mapping, many studies have made efforts to mine reliable information from LR labels. E.g., a label super-resolution network was designed to constrain the inexact parts of LR labels by using the statistical distribution inferred from HR labels [31, 37]. A multi-stage framework, named WESUP, was built for 10-m land-cover mapping with 30-m labels [12]. In WESUP, multi-models were trained to refine clean samples from LR labels. Similarly, the winner approach of the 2021 IEEE GRSS Data Fusion Contest (DFC) deployed a shallow CNN to refine the 30-m labels, and then multi-models were trained with pseudo-labels to create the 1-m land-cover map of Maryland, USA [27]. Moreover, a low-to-high network (L2HNet) was proposed to select confident parts of LR labels via weakly supervised loss functions [28]. To produce 1-m land-cover maps across China with available 10-m labels, seven L2HNet models were selectively trained to adapt wide-span geographic areas [29].

Different from these approaches that either still rely on partial HR labels or require human interventions, Paraformer

is designed as an HR-label-free end-to-end framework to facilitate large-scale HR land-cover mapping.

### 3. Methodology

To jointly capture local and global contexts and reasonably exploit LR labels for large-scale HR land-cover mapping, Paraformer combines parallel CNN and Transformer branches with a PLAT module. In this section, the three components are introduced sequentially.

#### 3.1. CNN-based resolution-preserving branch

As a basic feature extractor of Paraformer and also the main structure of previous L2HNet V1 [28], the CNN branch is designed to capture local contexts from HR images and preserve the spatial details by preventing feature down-sampling. As shown in Figure 3 (a), the CNN branch is constructed by five serially connected resolution-preserving (RP) blocks. Each RP block contains parallel convolution layers with the sizes of  $1 \times 1$ ,  $3 \times 3$ , and  $5 \times 5$ , whose steps are set to 1 for feature size maintaining. Partly similar to the inception module [42], the channel numbers of different scales' layers in each block are inversely proportional to their kernel sizes, which are set to 128, 64, and 32. Based on the setting, the RP blocks can capture features with a proper receptive field instead of down-sampling the feature maps with a deep encoder-decoder pattern. The serial blocks aim at sufficiently preserving the spatial resolution of features by using the majority of  $1 \times 1$  kernels. The  $3 \times 3$  and  $5 \times 5$  kernels capture necessary surrounding information. Furthermore, the multi-scale feature maps are concatenated and reduced to 128 channels for branch light-

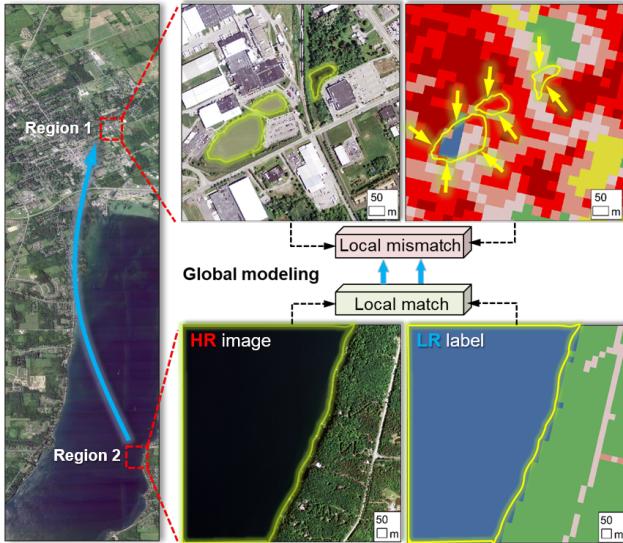


Figure 4. Example of the local mismatch/match in two regions. The edge of water is marked with yellow boundaries. Region 1 shows dispersed lakes around urban areas with unmatched annotation. Region 2 shows a large-scale river with matched annotation.

ening. Besides, a shortcut connection is adopted between blocks for residual learning and detail preserving.

### 3.2. Transformer-based global-modeling branch

The ground objects with the same land-cover class may have distinctive attributes in HR images and are differently annotated in LR labels. Figure 4 shows typical cases of lakes and rivers located in different areas. By considering that the CNN branch with intrinsic locality hinders the adaptation of various landforms over large-scale areas, we further hybrid the CNN branch with a Transformer branch which aims at capturing global contexts and building long-range support among dispersed geographic areas. As shown in Figure 3 (b), the Transformer branch contains 12 transformer layers. Each layer includes layer normalization, multi-head self-attention, and multi-layer perception. The feature maps extracted by each RP block are concatenated and inputted to the Transformer branch. Specifically, the extracted features from the CNN branch are downsampled and embedded in a hidden feature layer. And then the Transformer branch encodes the dense feature patches to capture global contexts. Subsequently, the encoded features are constantly upsampled to the size of HR images and classified to the final results. During the upsampling process, the outputted features of each stage are concatenated with the pre-encoded features, which bring massive local contextual information to the final feature maps.

### 3.3. Pseudo-Label-Assisted Training module

To reasonably guide the large-scale HR land-cover mapping with weak LR labels, as shown in Figure 3 (c), a weakly supervised PLAT module is adopted to optimize the frame-

work training. The PLAT module aims to screen out uncertain samples and dig up reliable information from the LR labels. Specifically, the two parts of the PLAT module are explained as follows. For the CNN branch, we use classifier<sup>(1)</sup>, which is constructed by  $3 \times 3$  convolution layers, to generate the primal prediction<sup>(1)</sup> based on the extracted HR feature maps. Then we calculate the Cross-Entropy (CE) loss between prediction<sup>(1)</sup>, represented as  $\hat{Y}'$ , and the LR label, represented as  $Y$ . Formally, by regarding  $H$ ,  $W$ , and  $L$  as the height, weight, and land-cover class of the patch, the CE loss of the CNN branch is written as:

$$\mathcal{L}_{ce}(Y, \hat{Y}') = \frac{\sum_{i=0}^W \sum_{j=0}^H \left[ \sum_{l=1}^L y_{ij}^{(l)} \log(\hat{y}_{ij}^{(l)}) \right]}{H \times W}. \quad (1)$$

As the final output of the framework, prediction<sup>(2)</sup> is classified from the concatenated feature maps of CNN and Transformer branches, which is represented as  $\hat{Y}''$ . During each training iteration, we take the simple but effective **intersection** of prediction<sup>(1)</sup> and LR label to generate mask labels. Specifically, the inconsistent samples in mask labels are set as void values to remove them from loss calculations. Moreover, since predictions of the CNN branch contain HR textual information that is highly consistent with the images, the mask labels also outline fine edges and retain stable labeled samples. Finally, the proposed Mask-Cross-Entropy (MCE) loss is calculated between prediction<sup>(2)</sup> and mask labels. Formally, the MCE loss is written as:

$$\mathcal{L}_{mce}(M \cdot Y, \hat{Y}'') = \frac{\sum_{i=0}^W \sum_{j=0}^H \left[ \sum_{l=1}^L y_{ij}^{(l)} m_{ij} \log(\hat{y}_{ij}''^{(l)}) \right]}{\text{Sum}(M(i, j) = 1)}. \quad (2)$$

In Eqs. 2,  $M$  is the **intersected** mask with the size of  $H \times W$ .  $m_{ij}, i \in [0, H], j \in [0, W]$  is the element of  $M(i, j)$  which can be simply represented as:

$$m_{ij} = \begin{cases} 1 & | Y_{ij} = Y'_{ij} \\ 0 & | Y_{ij} \neq Y'_{ij} \end{cases}. \quad (3)$$

The total loss of the Paraformer is the combination of two branches' losses, which is written as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{ce} + \mathcal{L}_{mce}. \quad (4)$$

## 4. Experiments

### 4.1. Study areas and using data

To comprehensively evaluate Paraformer on various landforms and different LR labels, the experiments are conducted on two large-scale datasets.

**The Chesapeake Bay dataset** is sampled from the largest estuary in the USA and organized into 732 non-overlapping tiles, where each tile has a size of  $6000 \times 7500$  pixels [37]. The specific data includes:

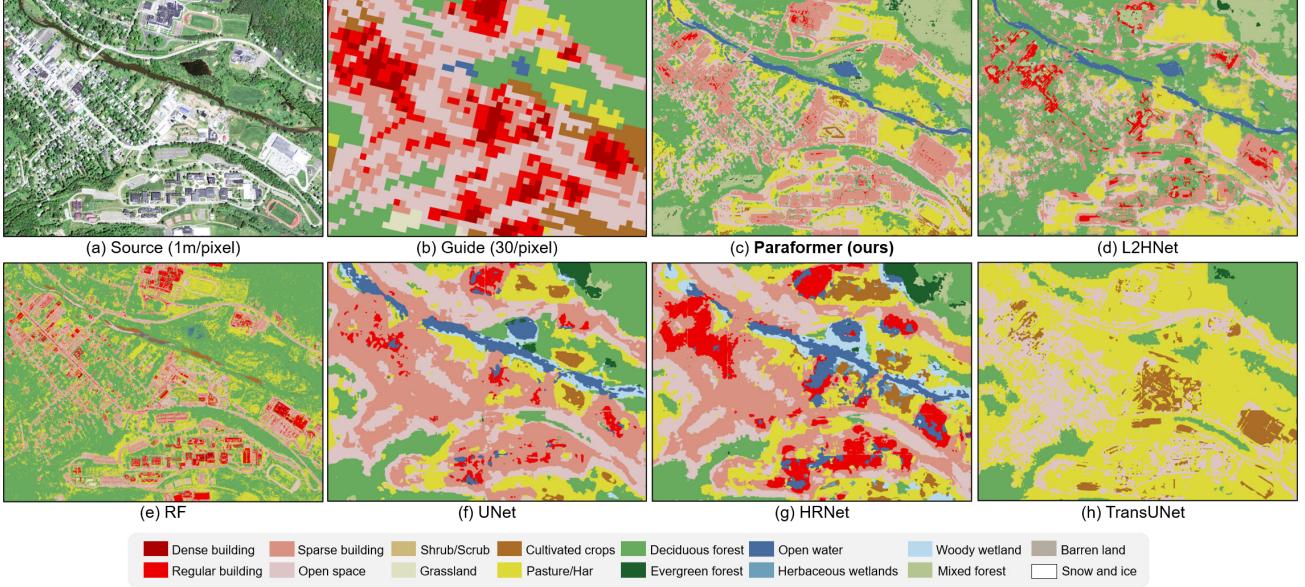


Figure 5. Demonstration of the training data and visual comparisons of the **Paraformer** and other typical methods on the Chesapeake Bay dataset with 16 classes. (a) HR image. (b) LR label. (c) land-cover mapping result of Paraformer. (d–h) land-cover mapping results of five typical methods.

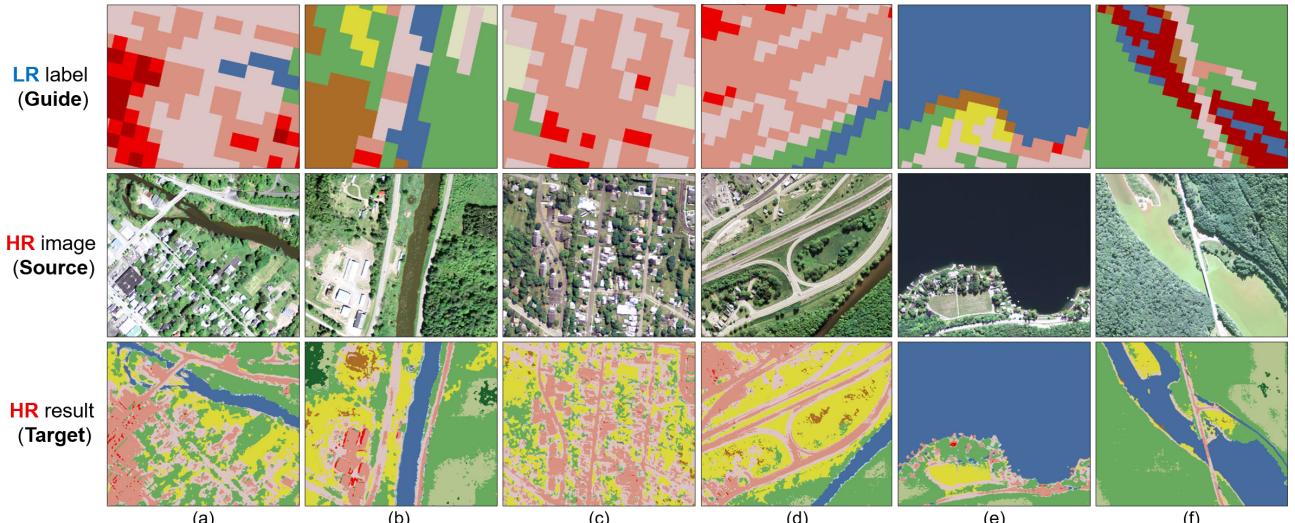


Figure 6. Six typical areas with finer observation scale on the Chesapeake Bay dataset. The first row shows the LR labels (**Guide**). The second row shows the HR images (**Source**). Third row shows the HR results (**Target**) produced by **Paraformer**.

1. *The HR images (1 m/pixel)* are from the U.S. Department of Agriculture’s National Agriculture Imagery Program (NAIP). The images contained four bands of red, green, blue, and near-infrared [33].
2. *The LR historical labels (30 m/pixel)* are from the USGS’s National Land Cover Database (NLCD) [49], including 16 land-cover classes.
3. *The ground truths (1 m/pixel)* are from the Chesapeake Bay Conservancy Land Cover (CCLC) project.

**The Poland dataset** contains 14 provinces of Poland and is organized into 403 non-overlapping tiles, where each tile has a size of  $1024 \times 1024$  pixels. The specific data includes:

1. *The HR images (0.25m and 0.5 m/pixel)* are from the

LandCover.ai [4] dataset. The images contained three bands of red, green, and blue.

2. *The LR historical labels* are collected from three types of 10-m land-cover data and one type of 30-m data, which are named FROM\_GLC10 [9], ESA\_GLC10 [44], ESRI\_GLC10 [22], and GLC\_FCS30 [56].
3. *The HR ground truths* are from the OpenEarthMap [51] dataset with seven land-cover classes.

## 4.2. Implementation Detail and Metrics

In the experiments, all methods only take LR land-cover data as training labels. Paraformer is trained by the AdamW optimizer with a patch size of  $224 \times 224$  and batch size of 8. The learning rate is set to 0.01 and would decrease by 10%

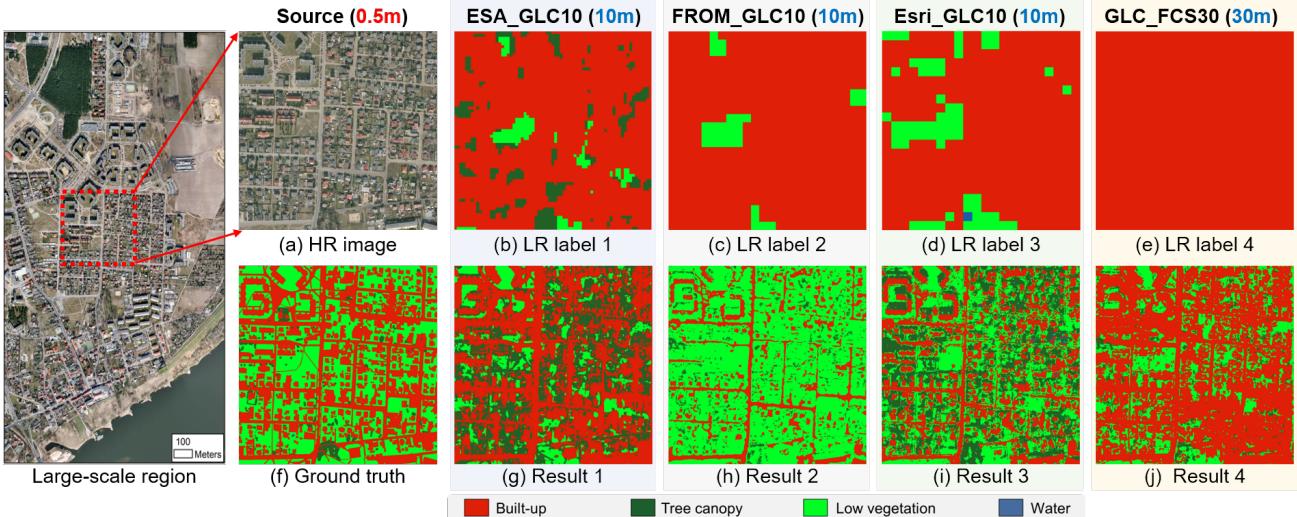


Figure 7. Visual results of **Paraformer** in the Poland dataset. The demonstration area is one of the training pieces sampled from large-scale training regions. (a–e) the training pairs of HR images (0.5 m/pixel) and four types of LR labels including **ESA\_GLC** (10 m/pixel), **FROM\_GLC** (10 m/pixel), **Esri\_GLC** (10 m/pixel), and **GLC\_FCS30** (30 m/pixel). (f–g) the ground truth (0.5 m/pixel) and the mapping results of Paraformer with different LR labels.

Resolution gap	Method	mIoU (%) of six states in the Chesapeake Bay watershed						
		Delaware	New York	Maryland	Pennsylvania	Virginia	West Virginia	Average
30×	<b>Paraformer</b>	<b>65.57</b>	<b>71.43</b>	<b>70.20</b>	<b>60.04</b>	68.01	52.62	<b>64.65</b>
	L2HNet [28]	61.77	68.12	65.24	58.52	<b>69.39</b>	<b>55.43</b>	63.08
	TransUNet [10]	53.15	60.53	60.42	51.08	66.21	47.52	56.49
	ConViT [18]	55.26	60.71	61.58	53.94	59.80	49.11	56.73
	CoAtNet [16]	56.89	62.83	61.25	53.57	65.67	51.34	58.59
	MobileViT[34]	58.03	61.32	61.84	55.53	57.04	48.64	57.07
	EfficientViT[5]	53.72	61.28	59.48	51.38	57.34	48.76	55.33
	UNetFormer[48]	58.85	65.11	61.34	<b>59.10</b>	60.84	47.20	58.74
	DC-Swin[47]	59.65	65.99	58.60	58.06	64.11	48.15	59.09
	UNet [38]	54.16	58.79	56.42	53.21	57.34	46.11	54.34
	HRNet [45]	52.11	56.21	50.76	50.03	57.48	45.42	52.00
	LinkNet [8]	58.27	62.05	52.96	52.11	48.71	48.93	53.84
	SkipFCN [26]	60.97	64.83	59.44	55.37	64.72	<b>54.66</b>	60.00
	SSDA [43]	57.91	61.54	54.85	51.71	57.71	47.15	55.15
	RF [7]	59.35	55.03	55.26	51.07	52.29	54.36	54.56

Table 1. The quantitative comparison of the Paraformer and other methods on six states of the Chesapeake Bay watershed. All methods were trained with the 1-m images and 30-m labels. The mIoU (%) of different methods was calculated between their results and the 1-m ground truth.

Max gap	LR label	mIoU (%) of different methods								
		Paraformer (ours)	L2HNet [28]	TransUNet [10]	ConViT [18]	MobileViT [34]	DC-Swin [47]	HRNet [45]	SkipFCN [26]	RF [7]
40×	FROM_GLC10 [9]	<b>56.57</b>	50.15	38.44	39.36	41.03	43.56	43.66	27.14	21.48
	ESA_GLC10 [44]	<b>55.19</b>	52.13	35.58	36.09	38.42	40.05	49.81	28.34	26.97
	Esri_GLC10 [22]	<b>55.07</b>	50.78	37.79	38.78	38.50	39.91	46.65	28.18	19.36
120×	GLC_FCS30 [56]	<b>49.39</b>	43.62	26.20	29.16	29.57	30.14	41.46	23.67	17.02

Table 2. The quantitative comparison on the Poland dataset. The mIoU (%) of the Paraformer and other methods that trained with three types of 10-m labels (i.e., **FROM\_GLC10**, **ESA\_GLC10**, and **Esri\_GLC10**) and one type of 30-m label (i.e., **GLC\_FCS30**) are demonstrated.

when the loss stopped dropping over eight epochs. The metric of mean intersection over union (mIoU) is calculated between the results and the HR ground truths after their land-cover classes are unified into four base classes. The compared methods include: Random Forest (RF) is a pixel-to-pixel method widely used in large-scale land-cover mapping [7]. TransUNet [10], ConViT [18], CoAtNet [16], MobileViT [34], and EfficientViT [5] are CNN-Transformer hy-

brid methods for semantic segmentation. UNetformer [48] and DC-Swin [47] are dedicated CNN-Transformer methods for remote-sensing images. UNet [38], HRNet [45], and LinkNet [8] are typical CNN-based semantic segmentation methods which are widely adopted in HR land-cover mapping [37, 52, 53]. SkipFCN [26] and SSDA [43] are shallow CNN-based methods for updating 1-m land-cover change maps from 30-m labels, which won first and second place

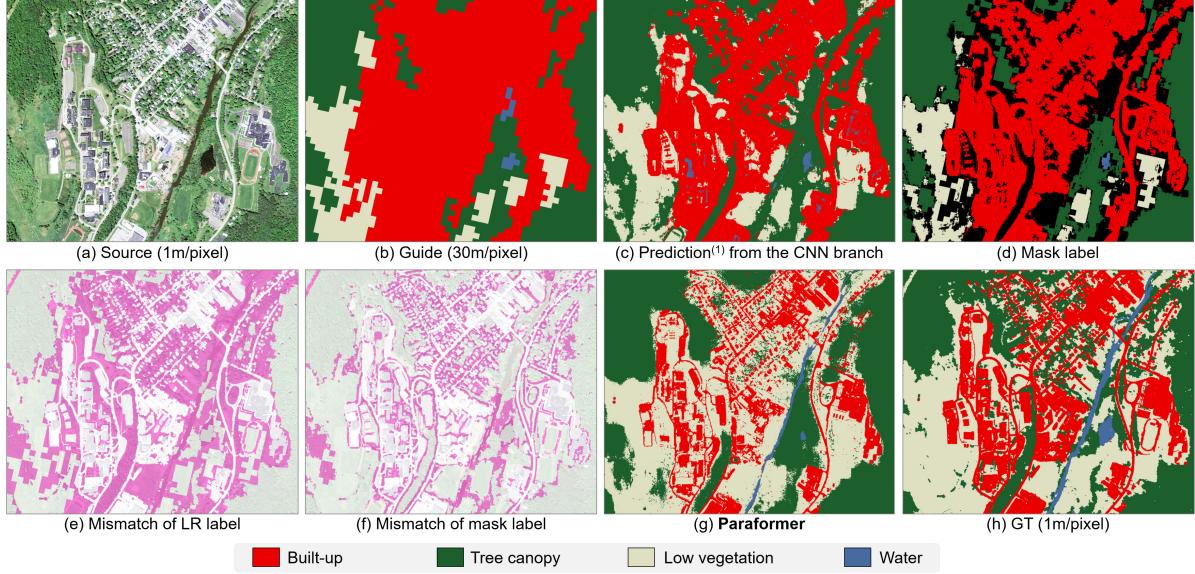


Figure 8. Example of training data and different outputs of Paraformer sampled from the Chesapeake Bay dataset with four unified classes. (a) HR images. (b) LR labels. (c) the primal prediction from the CNN branch. (d) Mask label, as the intersection parts of (b) and (c). The **black areas** are set to void without supervised information. (e–f) the incorrect samples (with **pink** color) of LR label and mask label. (g) the final results of Paraformer. (h) HR ground truth.

in the 2021 IEEE GRSS DFC [27]. L2HNet is a state-of-the-art method designed for weakly supervised land-cover mapping [28].

#### 4.3. Comparison Results

**Comparison on the Chesapeake Bay dataset:** Table 1 and Figure 5 show the comparisons on the Chesapeake Bay dataset. From the quantitative results, Paraformer shows superiority in the states of Delaware, New York, Maryland, and Pennsylvania. The L2HNet shows better results in Virginia and West Virginia. On average, Paraformer has the most accurate HR land-cover mapping results over the entire area, with a mIoU of 64.65%. As shown in Figure 5 (c), the visual result of Paraformer is more consistent with the HR image compared with other methods. Unlike the fully supervised semantic segmentation task, the unmatched training pairs can cause serious misguidedness during the model training. E.g., as the rough results shown in Figure 5 (f) and (g), UNet and HRNet over-downsample the features and encourage results to fit LR labels instead of being consistent with the HR images. Furthermore, quantitative results reveal that UNet, LinkNet, and HRNet have insufficient performance, with mIoU of 54.34%, 53.84%, and 52.00%. Although the compared CNN-Transformer methods (e.g., TransUNet) combine local and global contextual information, the structure does not focus on preserving the feature resolution or dealing with the geospatial mismatch. As a result, TransUNet shows a weak performance in visual results, shown in Figure 5 (h), and has a mIoU of 56.49%. Furthermore, SkipFCN, SSDA, and RF use small receptive fields or pixel-to-pixel strategies to ex-

tract features with fine land details. However, due to the lack of deep-level feature representation and global contextual information, SkipFCN, SSDA, and RF obtain a mIoU of 59.99%, 55.15%, and 54.56%, respectively. As an example shown in Figure 5 (e), RF finely predicts ground details but incorrectly classifies rivers, lakes, and pastures. To further demonstrate the effect of Paraformer on different landscapes, we sample six typical areas in Figure 6. The visual results indicate that the complex ground details among various landforms of HR land-cover maps can be well updated from the LR historical land-cover labels.

**Comparison on the Poland dataset:** In the experiments with the Poland dataset, all methods were used to produce 0.25/0.5-m land-cover maps of 14 provinces of Poland by exploiting four LR labels separately. These LR labels include 10-m FROM\_GLC10, ESA\_GLC10, Esri\_GLC10, and 30-m GLC\_FCS30. As shown in Table 2, Paraformer is compared with eight representative methods (i.e., weakly supervised, CNN-Transformer, CNN-based, pixel-to-pixel approaches) in a more extreme geospatial mismatch. Compared with the state-of-the-art method, the Paraformer has an increase in mIoU of 6.42%, 3.06%, and 4.29% in exploiting 10-m labels. By resolving 30-m labels with a max resolution gap of  $120 \times$ , Paraformer has a mIoU of 49.39% with an increase of 5.77% compared with L2HNet. The typical CNN-based method has an average mIoU of 46.71% among the 10-m cases and 41.46% in the 30-m case. Skip\_FCN and RF have the lowest mIoU among all methods, which shows the difficulty of dealing with extremely unmatched situations. Moreover, the quantitative results of Paraformer shown in the four cases reveal that the proposed frame-

Ablation method	mIoU (%) of six states in the Chesapeake Bay watershed							Params	FLOPs
	Delaware	New York	Maryland	Pennsylvania	Virginia	West Virginia	Average		
Paraformer	<b>65.57</b>	<b>71.43</b>	<b>70.20</b>	<b>60.04</b>	<b>68.01</b>	<b>52.62</b>	<b>64.65</b>	109.4M	141.3G
Sole CNN branch	59.57	67.87	64.30	53.86	65.26	50.01	60.15	4.5M	56.1G
Sole Transformer branch	53.15	60.53	60.42	51.08	66.22	47.52	56.49	96.9M	83.3G
Hybrid without PLAT	<u>62.69</u>	<u>70.39</u>	<u>67.15</u>	<u>58.33</u>	<u>67.47</u>	<u>50.83</u>	<u>62.81</u>	109.4M	141.3G

Table 3. The ablation results of the Paraformer on six states of the Chesapeake Bay watershed. The sole CNN branch, sole Transformer branch, and Hybrid without PLAT aim to investigate the contribution of the CNN branch, Transformer branch, and PLAT module, respectively.

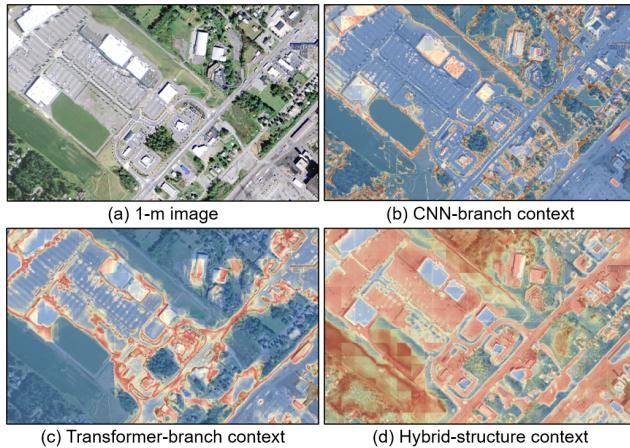


Figure 9. Demonstration of the extracted contexts from the ablation methods. (a) the original HR image. (b) the contexts extracted by the sole CNN branch. (c) the contexts extracted by the sole Transformer branch . (d) the contexts extracted by the CNN-Transformer hybrid backbone.

work obtains stable results from different LR labels. Figure 7 shows the visual results of Paraformer among four cases. With the parallel CNN-Transformer structure and PLAT module, Paraformer is able to refine the clear ground details (e.g., vegetation and roads) even if they are roughly labeled in local areas. In general, Paraformer shows the potential to robustly update large-scale HR land-cover maps from available LR historical labels.

#### 4.4. Ablation experiments

In this section, ablation experiments were conducted on the Chesapeake Bay dataset to evaluate different components of Paraformer. Each ablation in Table 3 is explained as follows: (1) the sole CNN branch is dependently trained by calculating CE loss with LR labels; (2) the sole Transformer branch embeds HR images instead of features from the CNN branch and calculates CE loss with LR labels; (3) the hybrid structure without PLAT directly calculates CE loss with the LR labels.

By ablating the PLAT module, the results obtained an average mIoU of 62.81%, which indicates a 1.84% decrease compared with the 64.65% of Paraformer. By ablating the CNN and Transformer branches, the results of the sole CNN branch obtained a mIoU of 60.15% and had a 4.5% decrease. Results of the sole Transformer branch obtained the lowest mIoU of 56.49% and had the most obvious decrease (8.16%). Figure 8 shows different outputs of Paraformer,

where the inexact LR labels are gradually refined during framework training. The final result shown in Figure 8 (g) indicates both fine ground details and accurate land-cover patterns that are consistent with the ground truth. Moreover, Figure 9 shows the visualized contexts captured by the CNN branch, Transformer branch, and hybrid structure. Figure 9 (b) indicates that the CNN branch mostly focuses on capturing local details (e.g., the edges of roads, single houses, and shrubs). Figure 9 (c) indicates that the Transformer branch captures the feature in object scale, focusing on intact land objects of building areas and parking spots. The hybrid structure shows a strong response to the obvious objects with both fine edges and intact areas.

In general, the ablation results demonstrate two findings: (1) The PLAT module can stably optimize the framework training and reasonably exploit the LR labels during the large-scale HR land-cover mapping process. (2) The parallel CNN and Transformer branches are indispensable parts of the framework, which construct a more robust feature extractor to bridge local and global contextual information.

## 5. Conclusion

In this paper, a weakly supervised CNN-Transformer framework, Paraformer, is proposed to update large-scale HR land-cover maps in an HR-label-free, end-to-end manner. Experiments on two datasets show that Paraformer outperforms other approaches in guiding semantic segmentation of large-scale HR remote-sensing images with easy-access LR land-cover data. Further analysis reveals that the Paraformer can robustly adapt various landforms of wide-span areas and stably exploit different LR labels in producing accurate HR land-cover maps. The ablation studies demonstrate the effectiveness of the parallel CNN-Transformer structure and the PLAT module. Moreover, intermediate results of each training process and visualized contexts of each branch are demonstrated to transparently explain the components of Paraformer. In general, the proposed Paraformer has the potential to become an effective method for facilitating large-scale HR land-cover mapping.

## Acknowledgments

This work has been supported by the National Key Research and Development Program of China (grant no. 2022YFB3903605) and the National Natural Science Foundation of China (grant no.42071322).

## References

- [1] Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sensing of Environment*, 237:111322, 2020. [1](#)
- [2] Cross-spatiotemporal land-cover classification from vhr remote sensing images with deep learning based domain adaptation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 191:105–128, 2022. [1](#)
- [3] Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190:196–214, 2022. [2](#)
- [4] Adrian Boguszewski, Dominik Batorski, Natalia Ziembajkowska, Tomasz Dziedzic, and Anna Zambrzycka. Land-cover. ai: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1102–1110, 2021. [5](#)
- [5] Han Cai, Junyan Li, Muyan Hu, Chuang Gan, and Song Han. Efficientvit: Lightweight multi-scale attention for high-resolution dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17302–17313, 2023. [1, 6](#)
- [6] Yinxia Cao and Xin Huang. A coarse-to-fine weakly supervised learning method for green plastic cover segmentation using high-resolution remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 188:157–176, 2022. [2](#)
- [7] Jonathan Cheung-Wai Chan and Desiré Paelinckx. Evaluation of random forest and adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing of Environment*, 112(6):2999–3011, 2008. [2, 6](#)
- [8] Abhishek Chaurasia and Eugenio Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE visual communications and image processing (VCIP)*, pages 1–4. IEEE, 2017. [6](#)
- [9] Bin Chen, B Xu, Z Zhu, C Yuan, H Ping Suen, J Guo, N Xu, W Li, Y Zhao, JJSB Yang, et al. 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):3, 2019. [2, 5, 6](#)
- [10] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. [1, 2, 6](#)
- [11] Keyan Chen, Zhengxia Zou, and Zhenwei Shi. Building extraction from remote sensing images with sparse token transformers. *Remote Sensing*, 13(21):4441, 2021. [1, 2](#)
- [12] Yujia Chen, Guo Zhang, Hao Cui, Xue Li, Shasha Hou, Jin-hao Ma, Zhijiang Li, Haifeng Li, and Huabin Wang. A novel weakly supervised semantic segmentation framework to improve the resolution of land cover product. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196:73–92, 2023. [2, 3](#)
- [13] Mang Tik Chiu, Xingqian Xu, Yunchao Wei, Zilong Huang, Alexander G Schwing, Robert Brunner, Hrant Khachatrian, Hovnatan Karapetyan, Ivan Dozier, Greg Rose, et al. Agriculture-vision: A large aerial image database for agricultural pattern analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2828–2838, 2020. [3](#)
- [14] J. Cihlar. Land cover mapping of large areas from satellites: Status and research priorities. *International Journal of Remote Sensing*, 21(6-7):1093–1114, 2000. [1](#)
- [15] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. [2](#)
- [16] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in neural information processing systems*, 34:3965–3977, 2021. [6](#)
- [17] Runmin Dong, Weizhen Fang, Haohuan Fu, Lin Gan, Jie Wang, and Peng Gong. High-resolution land cover mapping through learning with noise correction. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2021. [2](#)
- [18] Stéphane d’Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, pages 2286–2296. PMLR, 2021. [1, 6](#)
- [19] Mark A Friedl and Carla E Brodley. Decision tree classification of land cover from remotely sensed data. *Remote sensing of environment*, 61(3):399–409, 1997. [2](#)
- [20] Raffaele Gaetano, Dino Ienco, Kenji Ose, and Remi Cresson. A two-branch cnn architecture for land cover classification of pan and ms imagery. *Remote Sensing*, 10(11):1746, 2018. [2](#)
- [21] Nicolas Girard, Dmitriy Smirnov, Justin Solomon, and Yuliya Tarabalka. Polygonal building extraction by frame field learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5891–5900, 2021. [1](#)
- [22] Krishna Karra, Caitlin Kontgis, Zoe Statman-Weil, Joseph C Mazzariello, Mark Mathis, and Steven P Brumby. Global land use/land cover with sentinel 2 and deep learning. In *2021 IEEE international geoscience and remote sensing symposium IGARSS*, pages 4704–4707. IEEE, 2021. [2, 5, 6](#)
- [23] Jungbeom Lee, Seong Joon Oh, Sangdoo Yun, Junsuk Choe, Eunji Kim, and Sungroh Yoon. Weakly supervised semantic segmentation using out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16897–16906, 2022. [2](#)
- [24] Jing Li, Junsong Fan, and Zhaoxiang Zhang. Towards noiseless object contours for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16856–16865, 2022. [2](#)
- [25] Jiepan Li, Wei He, Weinan Cao, Liangpei Zhang, and Hongyan Zhang. Uanet: An uncertainty-aware network for building extraction from remote sensing images. *IEEE*

- Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024. 2
- [26] Zhuohong Li, Fangxiao Lu, Hongyan Zhang, Guangyi Yang, and Liangpei Zhang. Change cross-detection based on label improvements and multi-model fusion for multi-temporal remote sensing images. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 2054–2057. IEEE, 2021. 6
- [27] Zhuohong Li, Fangxiao Lu, Hongyan Zhang, Lilin Tu, Jiayi Li, Xin Huang, Caleb Robinson, Nikolay Malkin, Nebojsa Jojic, Pedram Ghamisi, et al. The outcome of the 2021 ieee grss data fusion contest—track msd: Multitemporal semantic change detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:1643–1655, 2022. 1, 2, 3, 7
- [28] Zhuohong Li, Hongyan Zhang, Fangxiao Lu, Ruoyao Xue, Guangyi Yang, and Liangpei Zhang. Breaking the resolution barrier: A low-to-high network for large-scale high-resolution land-cover mapping using low-resolution labels. *ISPRS Journal of Photogrammetry and Remote Sensing*, 192:244–267, 2022. 1, 2, 3, 6, 7
- [29] Zhuohong Li, Wei He, Mofan Cheng, Jingxin Hu, Guangyi Yang, and Hongyan Zhang. Sinolc-1: the first 1 m resolution national-scale land-cover map of china created with a deep learning framework and open-access data. *Earth System Science Data*, 15(11):4749–4780, 2023. 2, 3
- [30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021. 2
- [31] Kolya Malkin, Caleb Robinson, Le Hou, Rachel Soobitsky, Jacob Czawlytko, Dimitris Samaras, Joel Saltz, Lucas Joppa, and Nebojsa Jojic. Label super-resolution networks. In *International Conference on Learning Representations*, 2018. 2, 3
- [32] Diego Marcos, Michele Volpi, Benjamin Kellenberger, and Devis Tuia. Land cover mapping at very high resolution with rotation equivariant cnns: Towards small yet accurate models. *ISPRS journal of photogrammetry and remote sensing*, 145:96–107, 2018. 2
- [33] Aaron E Maxwell, Timothy A Warner, Brian C Vanderbilt, and Christopher A Ramezan. Land cover classification and feature extraction from national agriculture imagery program (naip) orthoimagery: A review. *Photogrammetric Engineering & Remote Sensing*, 83(11):737–747, 2017. 5
- [34] Sachin Mehta and Mohammad Rastegari. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, 2021. 1, 6
- [35] Lucas Prado Osco, Qiusheng Wu, Eduardo Lopes de Lemos, Wesley Nunes Gonçalves, Ana Paula Marques Ramos, Jonathan Li, and José Marcato Junior. The segment anything model (sam) for remote sensing applications: From zero to one shot. *International Journal of Applied Earth Observation and Geoinformation*, 124:103540, 2023. 2
- [36] Bruce Pengra, Jordan Long, Devendra Dahal, Stephen V Stehman, and Thomas R Loveland. A global reference database from very high resolution commercial satellite data and methodology for application to landsat derived 30 m continuous field tree cover data. *Remote sensing of environment*, 165:234–248, 2015. 2
- [37] Caleb Robinson, Le Hou, Kolya Malkin, Rachel Soobitsky, Jacob Czawlytko, Bistra Dilkina, and Nebojsa Jojic. Large scale high-resolution land cover mapping with multi-resolution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12726–12735, 2019. 1, 2, 3, 4, 6
- [38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 6
- [39] Elif Sertel, Burak Ekim, Paria Ettehadi Osgouei, and M Erdem Kabadayi. Land use and land cover mapping using deep learning based segmentation approaches and vhr worldview-3 images. *Remote Sensing*, 14(18):4558, 2022. 2
- [40] Dee Shi and Xiaojun Yang. Support vector machines for land cover mapping from remote sensor imagery. *Monitoring and Modeling of Global Changes: A Geomatics Perspective*, pages 265–279, 2015. 2
- [41] Zhongyu Sun, Wangping Zhou, Chen Ding, and Min Xia. Multi-resolution transformer network for building and road segmentation of remote sensing image. *ISPRS International Journal of Geo-Information*, 11(3):165, 2022. 1, 2
- [42] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3
- [43] Lilin Tu, Jiayi Li, and Xin Huang. High-resolution land cover change detection using low-resolution labels via a semi-supervised deep learning approach-2021 ieee data fusion contest track msd. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 2058–2061. IEEE, 2021. 6
- [44] Ruben Van De Kerchove, Daniele Zanaga, Wanda Keersmaecker, Niels Souverijns, Jan Wevers, Carsten Brockmann, Alex Grosu, Audrey Paccini, Oliver Cartus, Maurizio Santoro, et al. Esa worldcover: Global land cover mapping at 10 m resolution for 2020 based on sentinel-1 and 2 data. In *AGU Fall Meeting Abstracts*, pages GC45I–0915, 2021. 2, 3, 5, 6
- [45] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 6
- [46] Junjue Wang, Zhuo Zheng, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 3

- [47] Libo Wang, Rui Li, Chenxi Duan, Ce Zhang, Xiaoliang Meng, and Shenghui Fang. A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. [2](#), [6](#)
- [48] Libo Wang, Rui Li, Ce Zhang, Shenghui Fang, Chenxi Duan, Xiaoliang Meng, and Peter M Atkinson. Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190:196–214, 2022. [1](#), [6](#)
- [49] James Wickham, Stephen V Stehman, Daniel G Sorenson, Leila Gass, and Jon A Dewitz. Thematic accuracy assessment of the nlcld 2016 land cover for the conterminous united states. *Remote Sensing of Environment*, 257:112357, 2021. [3](#), [5](#)
- [50] Qiusheng Wu and Lucas Prado Osco. samgeo: A python package for segmenting geospatial data with the segment anything model (sam). *Journal of Open Source Software*, 8(89):5663, 2023. [2](#)
- [51] Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. Openearthmap: A benchmark dataset for global high-resolution land cover mapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6254–6264, 2023. [5](#)
- [52] Jie Xie, Leyuan Fang, Bob Zhang, Jocelyn Chanussot, and Shutao Li. Super resolution guided deep network for land cover classification from remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2021. [2](#), [6](#)
- [53] Yue Xu, Jianya Gong, Xin Huang, Xiangyun Hu, Jiayi Li, Qiang Li, and Min Peng. Luojia-hssr: A high spatial-spectral resolution remote sensing dataset for land-cover classification with a new 3d-hrnet. *Geo-spatial Information Science*, pages 1–13, 2022. [2](#), [6](#)
- [54] Naoto Yokoya, Pedram Ghamisi, Ronny Hänsch, and Michael Schmitt. 2020 ieee grss data fusion contest: Global land cover mapping with weak supervision [technical committees]. *IEEE Geoscience and Remote Sensing Magazine*, 8(1):154–157, 2020. [2](#)
- [55] Hongyan Zhang, Wenbin Liu, and Liangpei Zhang. Seamless and automated rapeseed mapping for large cloudy regions using time-series optical satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 184:45–62, 2022. [1](#)
- [56] Xiao Zhang, Liangyun Liu, Xidong Chen, Yuan Gao, Shuai Xie, and Jun Mi. Glc\_fcs30: Global land-cover product with fine classification system at 30 m using time-series landsat imagery. *Earth System Science Data*, 13(6):2753–2776, 2021. [2](#), [5](#), [6](#)
- [57] Tianfei Zhou, Meijie Zhang, Fang Zhao, and Jianwu Li. Regional semantic contrast and aggregation for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4299–4309, 2022. [2](#)

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

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<sup>1</sup>. 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  $3 \times 3$  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  $W_{C_1, C_2}^{n \times n}$  (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:

$$\mathbf{M}^{(b)} = \text{concat} \left[ \begin{array}{l} bn(\mathbf{I}^{(b)} * W_{C_I, C_I}^{1 \times 1}), \\ bn(\mathbf{I}^{(b)} * W_{C_I, \frac{C_I}{2}}^{3 \times 3}), \\ bn(\mathbf{I}^{(b)} * W_{C_I, \frac{C_I}{4}}^{5 \times 5}) \end{array} \right]. \quad (\text{S1})$$

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  $1 \times 1$  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:

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

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  $1 \times 1$  convolutional layer  $W_{C_I, L}^{1 \times 1}$ , 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).

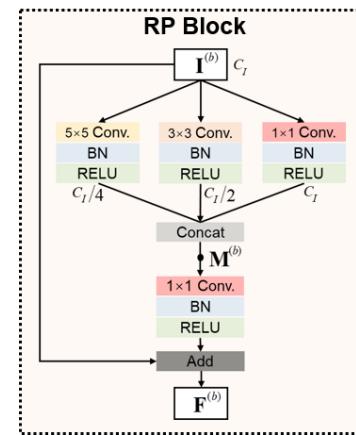


Figure S1. An illustration of an RP block. The input map  $\mathbf{I}^{(b)}$  is sampled by three convolutional layers with sizes of  $1 \times 1$ ,  $3 \times 3$ , and  $5 \times 5$ , and the convolution kernels in each layer are set to the proportion of  $\omega$  for preventing feature resolution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

<sup>1</sup><https://doi.org/10.1016/j.isprsjprs.2022.08.008>

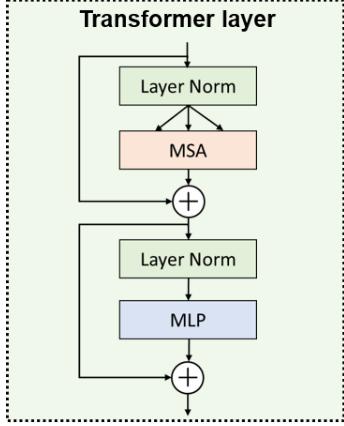


Figure S2. An illustration of a Transformer layer. The layer includes layer normalization (Layer Norm), multi-head self-attention (MSA), and multi-layer perception (MLP).

## 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 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<sup>2</sup>, 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

<sup>2</sup><https://lila.science/datasets/chesapeakelandcover>

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, Łódzkie, Lubuskie, Dolnośląskie, 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).
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.

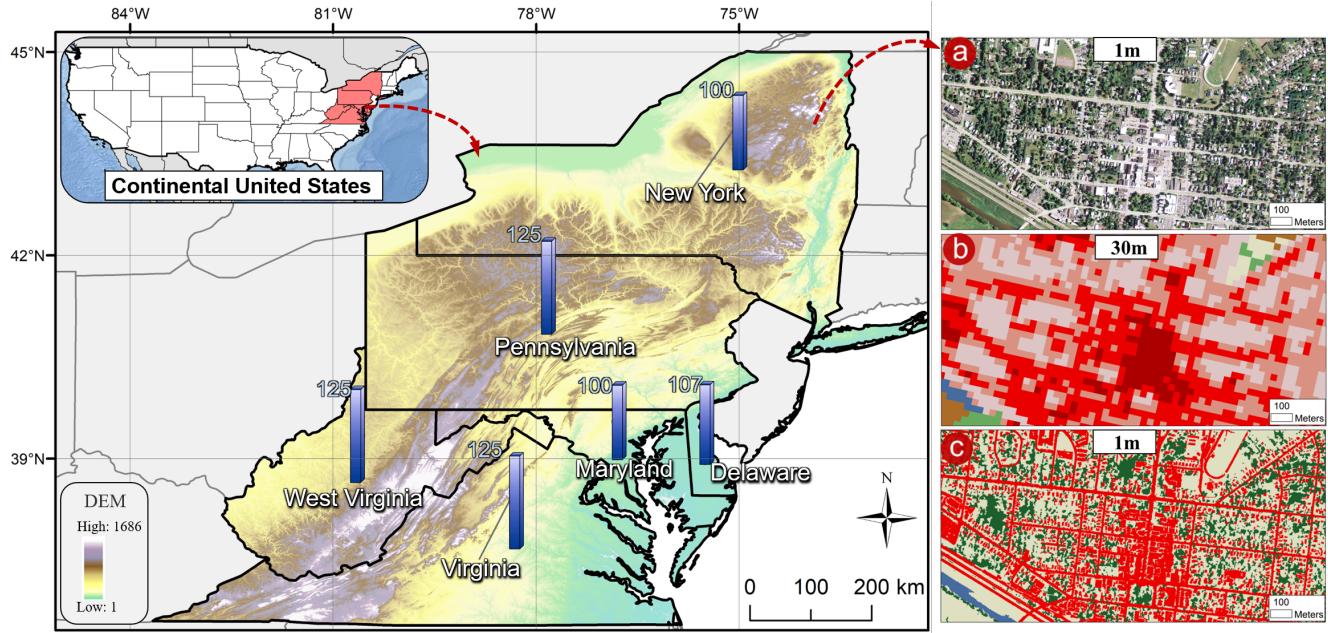


Figure S3. The Chesapeake Bay dataset covers six states of the USA, including the data sources of (a) The 1-m NAIP imagery, (b) The 30-m NLCD labels, and (c) The 1-m ground truth. The blue columns show the number of tiles.

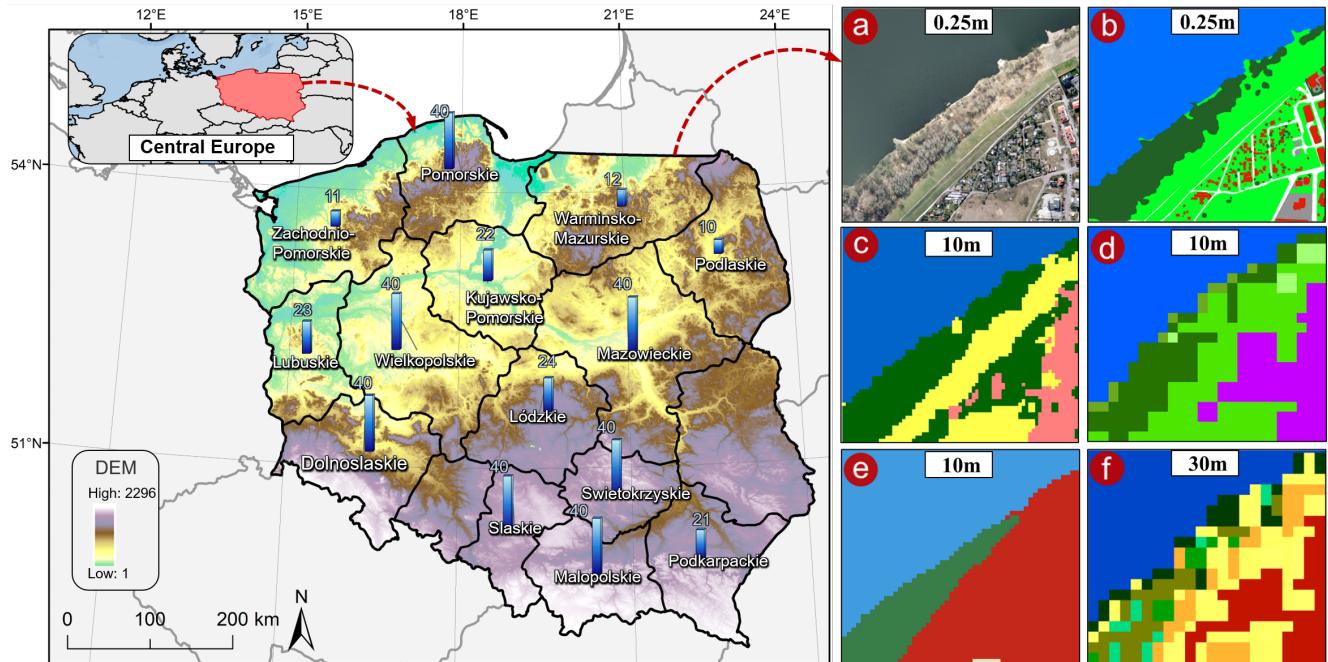


Figure S4. The Poland dataset covers 14 provinces of the country, including the data sources of (a) The 0.25-m imagery, (b) The 0.25-m ground truth, (c) The 10-m FROM\_GLC10, (d) The 10-m ESA\_GLC10, (e) The 10-m ESRI\_GLC10, and (f) The 30-m GLC\_FCS30. The blue columns show the number of tiles.

## 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

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

Name	NLCD	CCLC	
Affiliation	USGS, USA	Chesapeake Conservancy, USA	Target classes
Resolution	30 meters	1 meter	
	Developed open space	Roads	
	Developed low c	Building	
	Developed medium	Barren	<span style="color:red">■</span> Built-up
	Developed high		
	Deciduous forest		
	Evergreen forest	Tree canopy	<span style="color:green">■</span> Tree canopy
	Mixed forest		
	Woody wetland		
Class	Barren land		
	Shrub/Scrub		
	Grassland		
	Pasture/Har	Low vegetation	<span style="color:tan">■</span> Low vegetation
	Cultivated crops		
	Herbaceous wetlands		
	Open water	Water	<span style="color:blue">■</span> 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	
	<span style="color:green">■</span> Forest	<span style="color:green">■</span> Trees	<span style="color:green">■</span> Trees	<span style="color:green">■</span> Deciduous broadleaved forest <span style="color:tan">■</span> Open deciduous broadleaved forest <span style="color:darkgreen">■</span> Evergreen needle-leaved forest <span style="color:tan">■</span> Mixed leaf forest	Tree	<span style="color:green">■</span> Tree canopy
Class	<span style="color:green">■</span> Shrubland	<span style="color:tan">■</span> Scrub/Shrub	<span style="color:tan">■</span> Shrubland	<span style="color:tan">■</span> Orchard <span style="color:tan">■</span> Sparse shrubland <span style="color:orange">■</span> Grassland <span style="color:tan">■</span> Herbaceous cover	Rangeland	<span style="color:green">■</span> Low vegetation
	<span style="color:green">■</span> Grassland	<span style="color:green">■</span> Grassland	<span style="color:tan">■</span> Grassland	<span style="color:tan">■</span> Rainfed cropland <span style="color:cyan">■</span> Irrigated cropland	Agriculture land	
	<span style="color:lightgreen">■</span> Cropland	<span style="color:yellow">■</span> Crops	<span style="color:violet">■</span> Cropland		Building Road Developed space	<span style="color:red">■</span> Built-up
	<span style="color:magenta">■</span> Impervious area	<span style="color:red">■</span> Built Area	<span style="color:red">■</span> Built-up	<span style="color:red">■</span> Impervious surfaces	Water	<span style="color:blue">■</span> 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.

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  $\mathcal{L}_{ce}$  and  $\mathcal{L}_{mce}$  during framework training. The two training losses are stable to decrease in six states of 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.

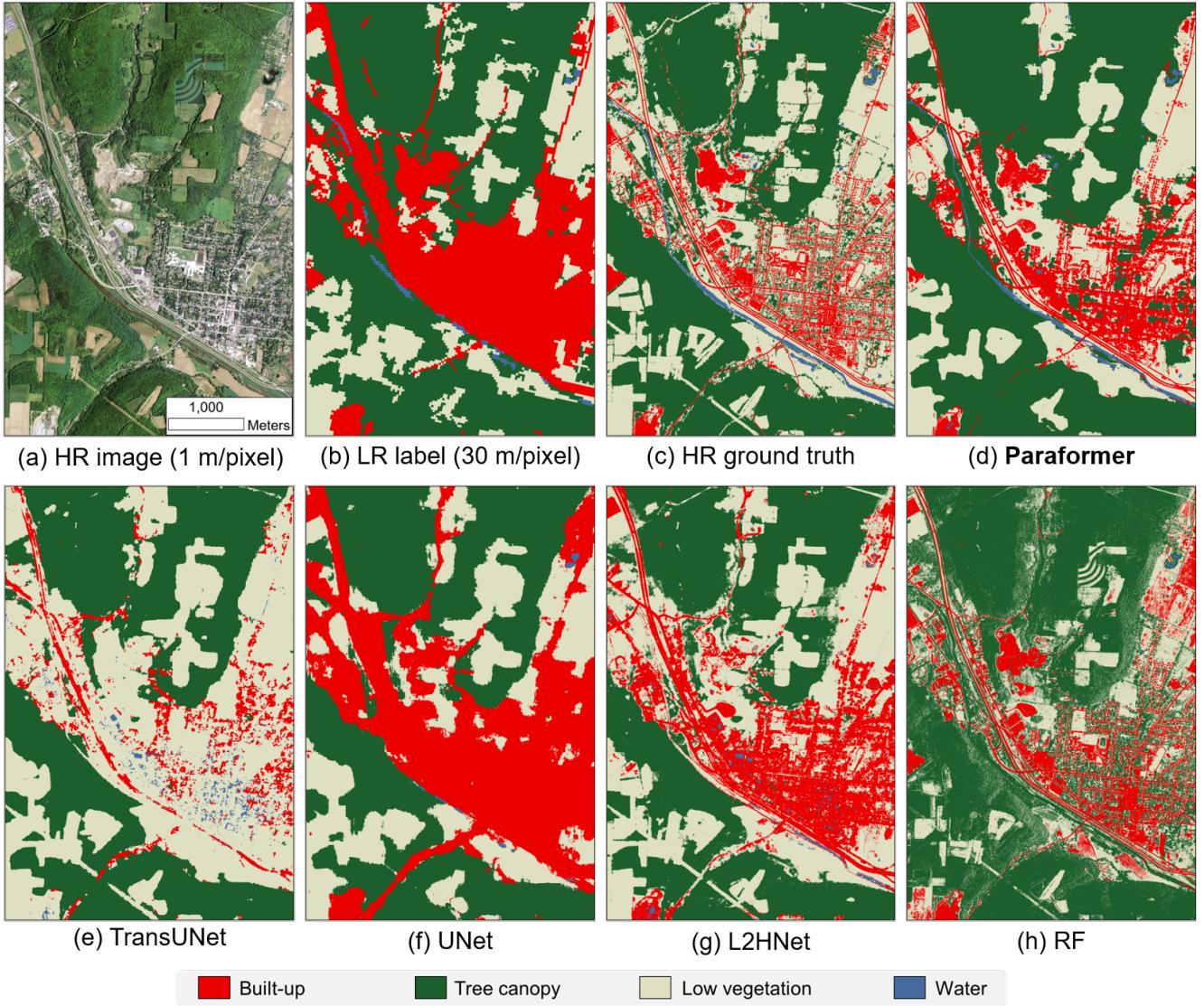


Figure S5. Demonstration of the training data and visual comparisons of the **Paraformer** and other typical methods on the Chesapeake Bay dataset with four unified classes. (a) HR image. (b) LR label. (c) HR ground truth. (d) land-cover mapping result of Paraformer. (e–h) land-cover mapping results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

Based on our previous work on SinoLC-1<sup>3</sup> (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.

<sup>3</sup><https://doi.org/10.5194/essd-15-4749-2023>

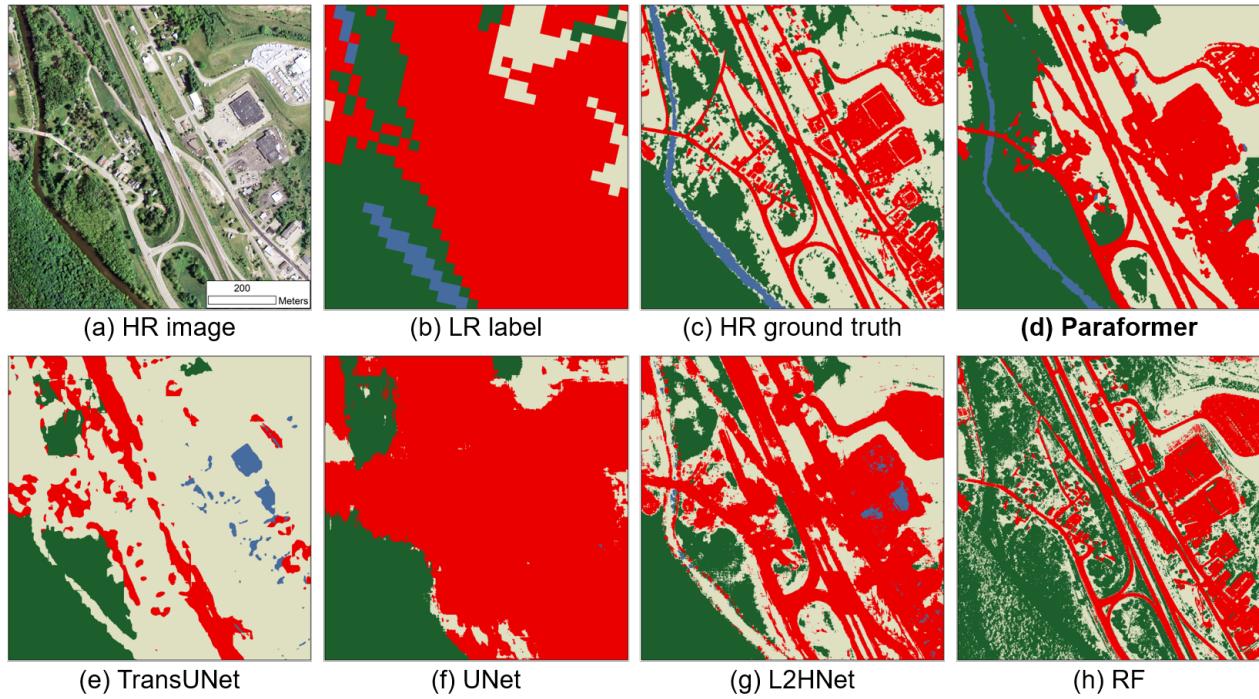


Figure S6. Sample A of the training data and visual comparisons of the **Paraformer** and other typical methods on the Chesapeake Bay dataset with four unified classes. (a) HR image. (b) LR label. (c) HR ground truth. (d) land-cover mapping result of Paraformer. (e-h) land-cover mapping results of four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

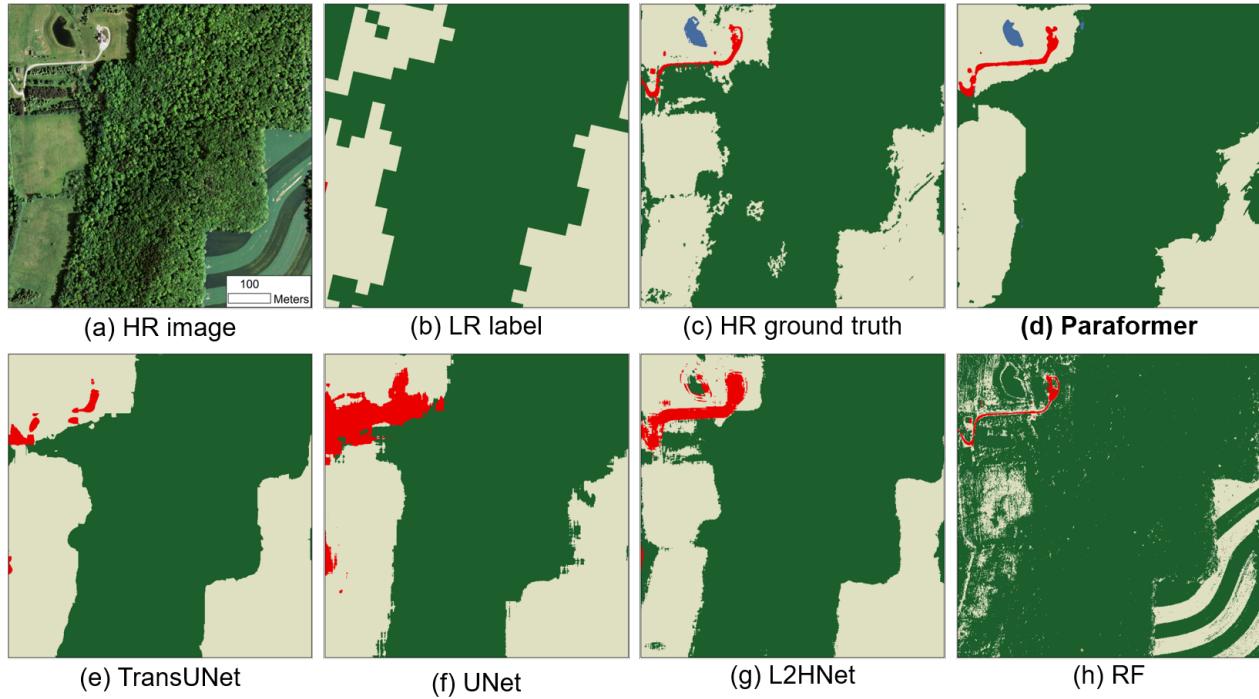


Figure S7. Sample B of the training data and visual comparisons of the **Paraformer** and other typical methods on the Chesapeake Bay dataset with four unified classes. (a) HR image. (b) LR label. (c) HR ground truth. (d) land-cover mapping result of Paraformer. (e-h) land-cover mapping results of four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

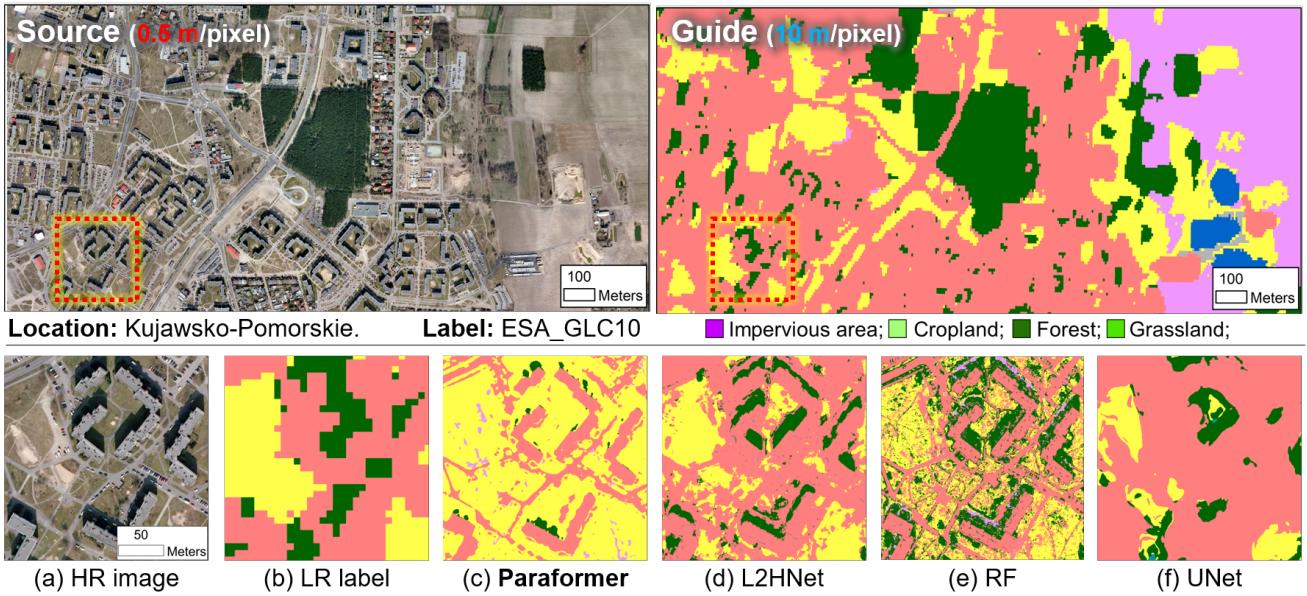


Figure S8. The visual results of Poland dataset with 10-m ESA\_GLC10 training labels. (a) The 0.5-m image, (b) The 10-m label sampled from the ESA\_GLC10. (c) Result of Paraformer. (d) Result of L2HNet. (e) Result of RF. (f) Result of UNet.

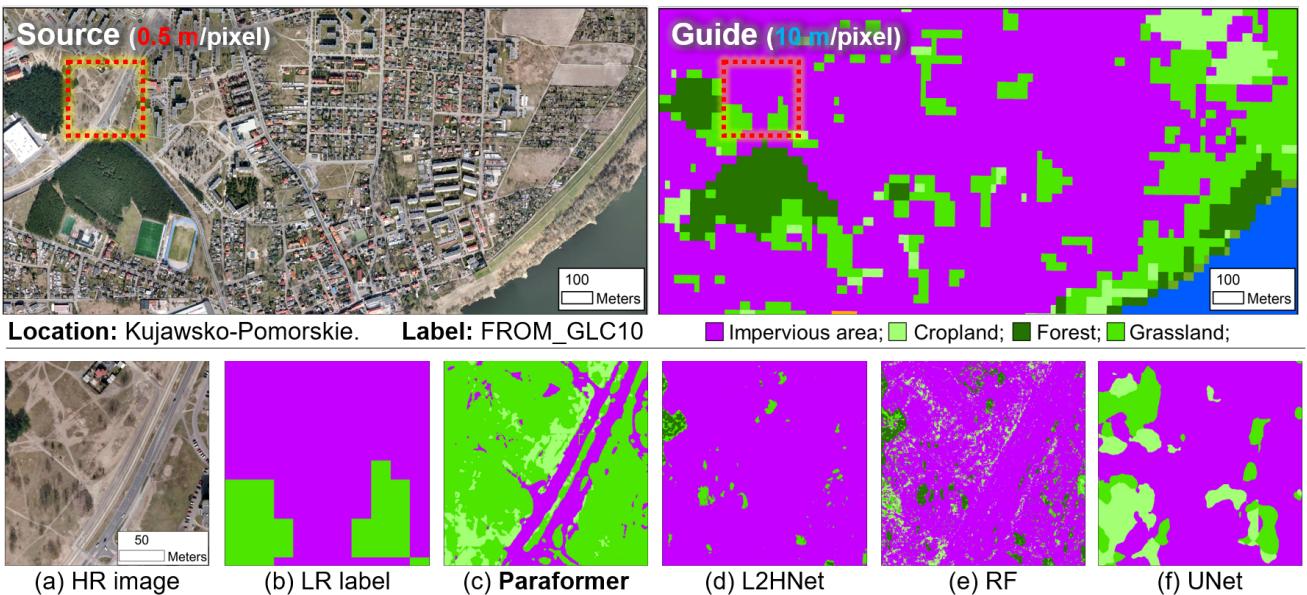


Figure S9. The visual results of Poland dataset with 10-m FROM\_GLC10 training labels. (a) The 0.5-m image, (b) The 10-m label sampled from the FROM\_GLC10. (c) Result of Paraformer. (d) Result of L2HNet. (e) Result of RF. (f) Result of UNet.

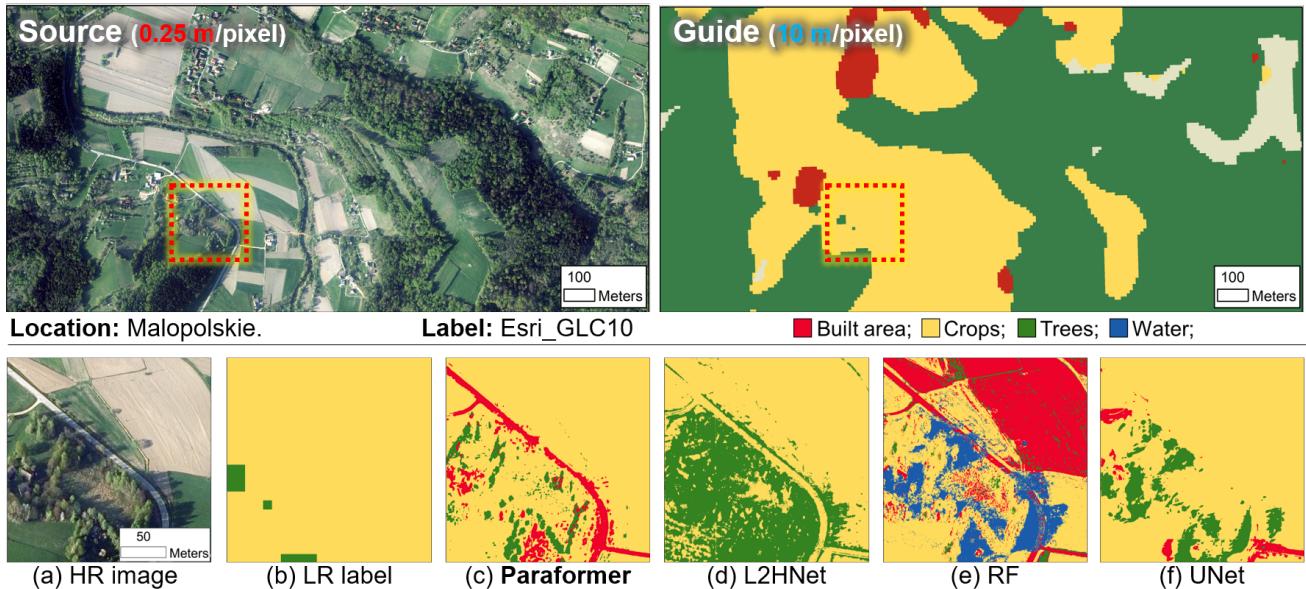


Figure S10. The visual results of Poland dataset with 10-m Esri\_GLC10 training labels. (a) The 0.25-m image, (b) The 10-m label sampled from the Esri\_GLC10. (c) Result of Paraformer. (d) Result of L2HNet. (e) Result of RF. (f) Result of UNet.

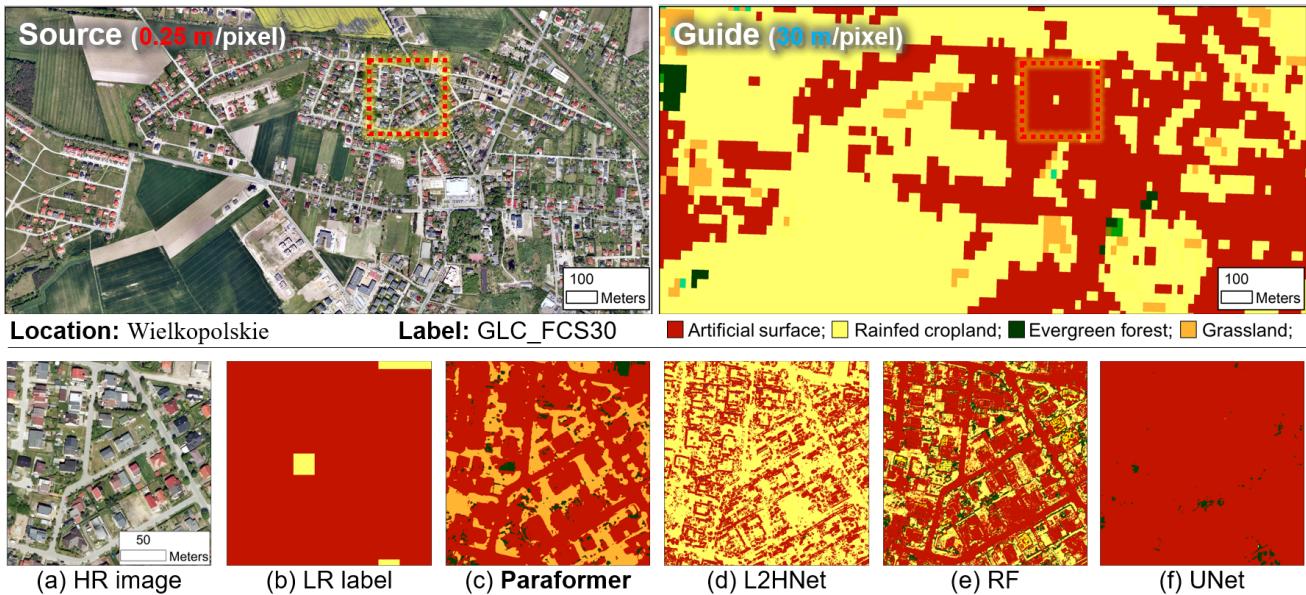


Figure S11. The visual results of Poland dataset with 30-m GLC\_FCS30 training labels. (a) The 0.5-m image, (b) The 10-m label sampled from the GLC\_FCS30. (c) Result of Paraformer. (d) Result of L2HNet. (e) Result of RF. (f) Result of UNet.

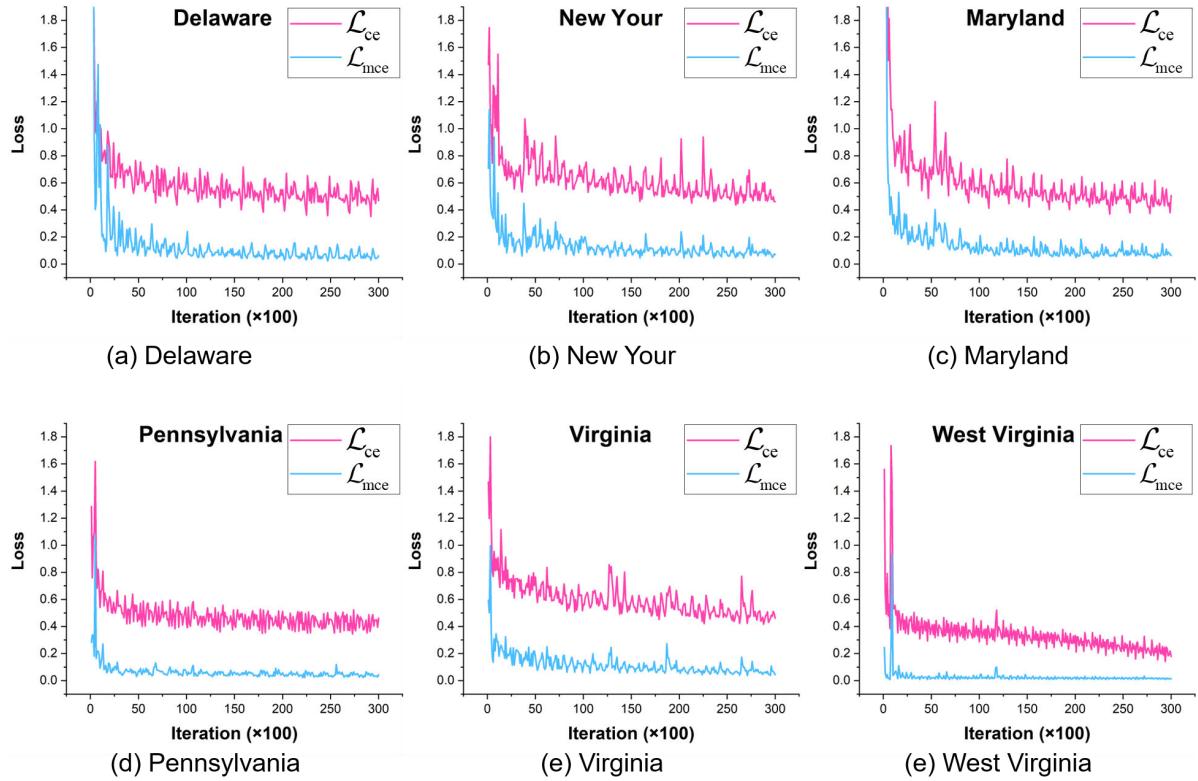


Figure S12. Demonstration of the loss functions  $\mathcal{L}_{ce}$  and  $\mathcal{L}_{mce}$  during framework training. Sub-figures (a)–(e) demonstrate the training process in six states of the Chesapeake Bay dataset.

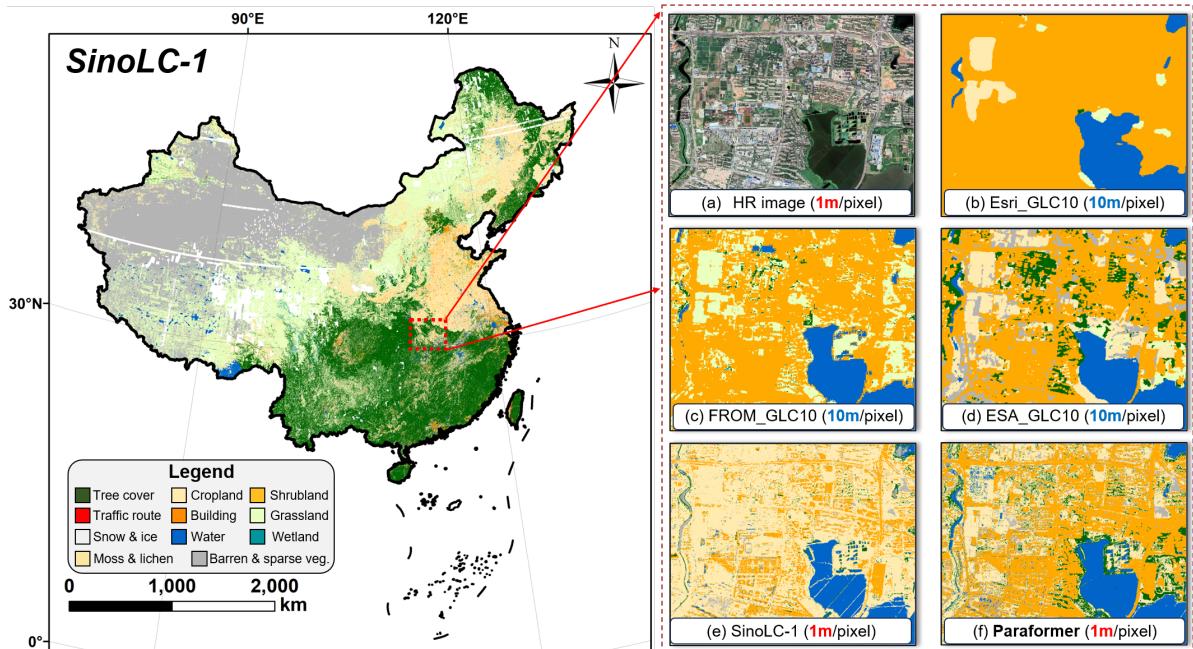


Figure S13. Demonstration of the supplementary experiments of SinoLC-1 dataset. The visual results are sampled from Wuhan, China. The Paraformer is used to update the 1-m land-cover map in the whole of Wuhan City, reporting an OA of 74.98%.