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

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 mapping 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].

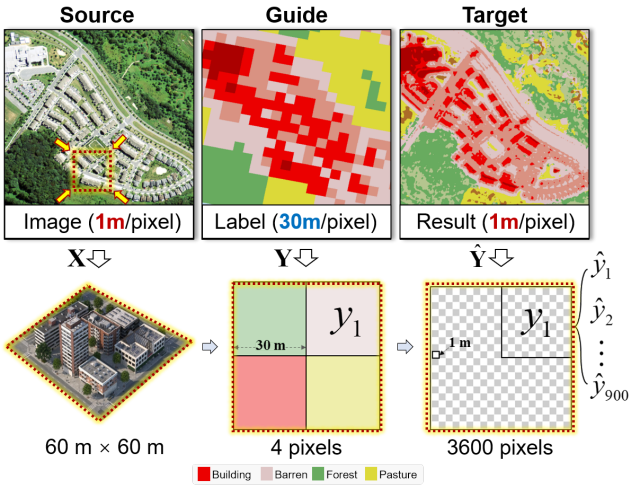

Figure 1. Illustration of resolution mismatched issue in using the HR remote-sensing image (**Source**) and LR historical labels (**Guide**) to generate HR land-cover results (**Target**).

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

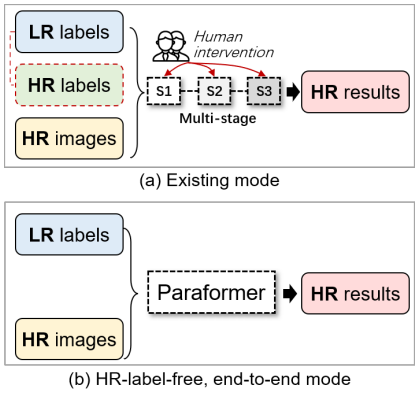

Figure 2. Two modes of large-scale HR land-cover mapping with LR labels. (a) Existing modes either reply on partial HR labels or require non-end-to-end training with human interventions. (b) **Paraformer** aims to form a mode that is HR-label-free and end-to-end trainable.

[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 $\mathbf{X}$. However, in the LR (30 m/pixel) label $\mathbf{Y}$, the area is only labeled by four pixels. To produce the 1-m land-cover result $\hat{\mathbf{Y}}$, a labeled pixel y_1 needs to provide guiding information for 900 target pixels $\{\hat{y}_1, \hat{y}_2 \cdots \hat{y}_{900}\}$, 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

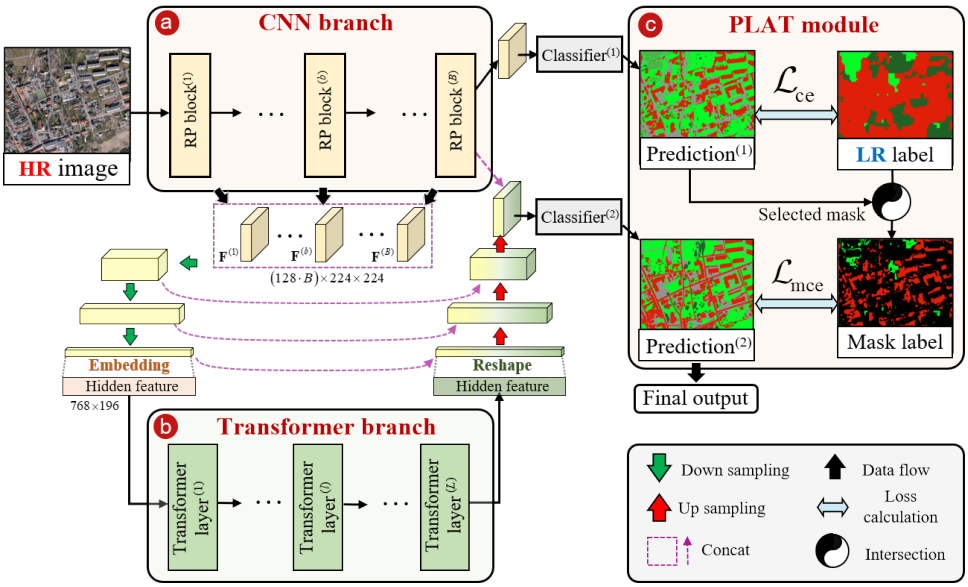

Figure 3. Overall workflow of Paraformer. The framework only takes the HR images and LR labels as training input and includes three components: (a) CNN-based resolution-preserving branch, (b) Transformer-based global-modeling branch, and (c) Pseudo-Label-Assisted Training (PLAT) module.

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-model 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 L2HNets 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, Pafaromer 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 downsampling. 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 1×1, 3×3, and 5×5, 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 downsampling 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 1×1 kernels. The 3×3 and 5×5 kernels capture necessary surrounding information. Furthermore, the multi-scale feature maps are concatenated and reduced to 128 channels for branch light-

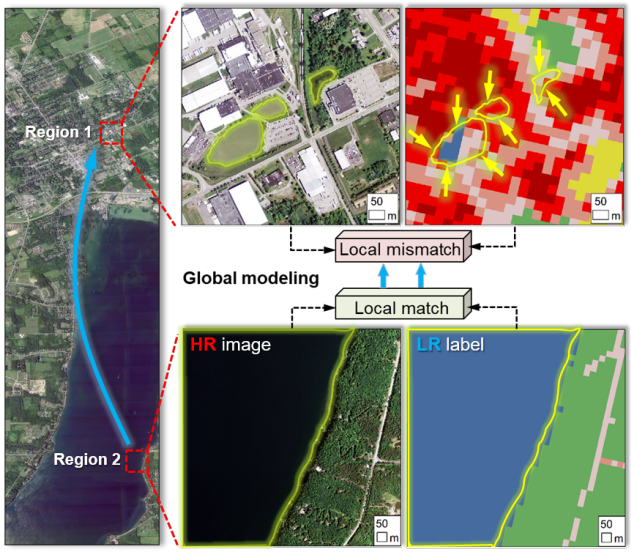

Figure 4. Example of the local mismatch/match in two regions. The edge of water is marked with yellow boundaries. Region 1 shows dispersed lakes around urban areas with unmatched annotation. Region 2 shows a large-scale river with matched annotation.

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 framework 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$^{(1)}$, which is constructed by 3×3 convolution layers, to generate the primal prediction$^{(1)}$ based on the extracted HR feature maps. Then we calculate the Cross-Entropy (CE) loss between prediction$^{(1)}$, represented as $\hat{\mathbf{Y}}'$, and the LR label, represented as $\mathbf{Y}$. 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:

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

As the final output of the framework, prediction$^{(2)}$ is classified from the concatenated feature maps of CNN and Transformer branches, which is represented as $\hat{\mathbf{Y}}''$. During each training iteration, we take the simple but effective **intersection** of prediction$^{(1)}$ 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$^{(2)}$ and mask labels. Formally, the MCE loss is written as:

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

In Eqs. 2, $\mathbf{M}$ 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 $\mathbf{M}(i,j)$ 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:

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

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 × 7500 pixels [37]. The specific data includes:

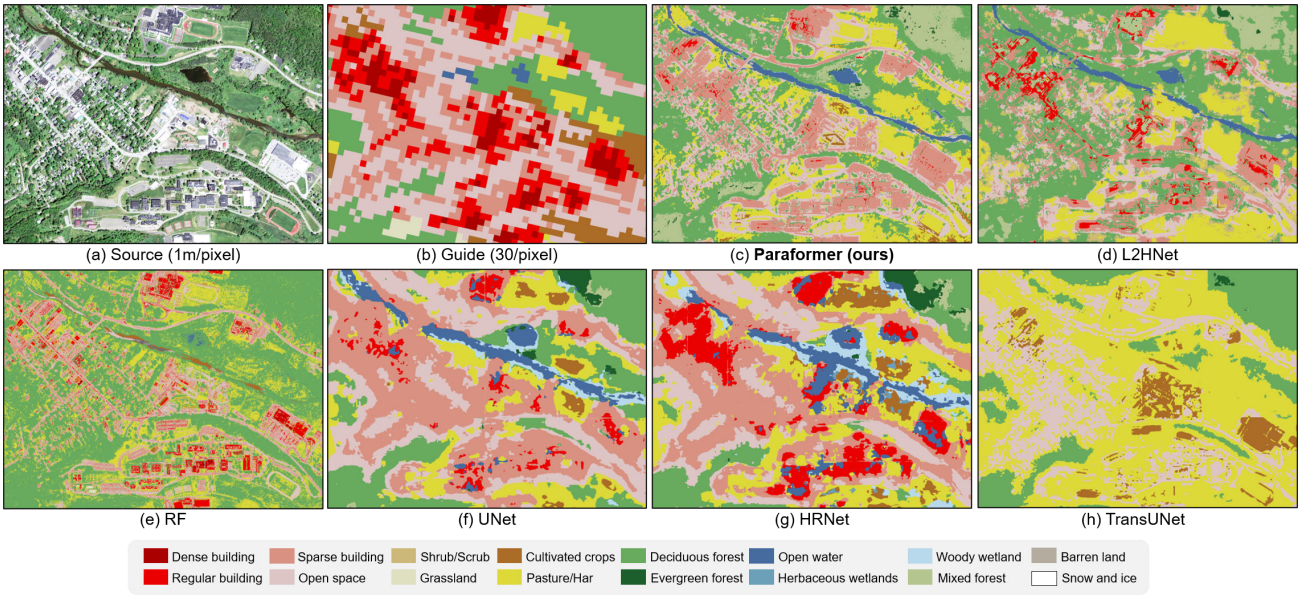

Figure 5. Demonstration of the training data and visual comparisons of the **Paraformer** and other typical methods on the Chesapeake Bay dataset with 16 classes. (a) HR image. (b) LR label. (c) land-cover mapping result of Parafomer. (d–h) land-cover mapping results of five typical methods.

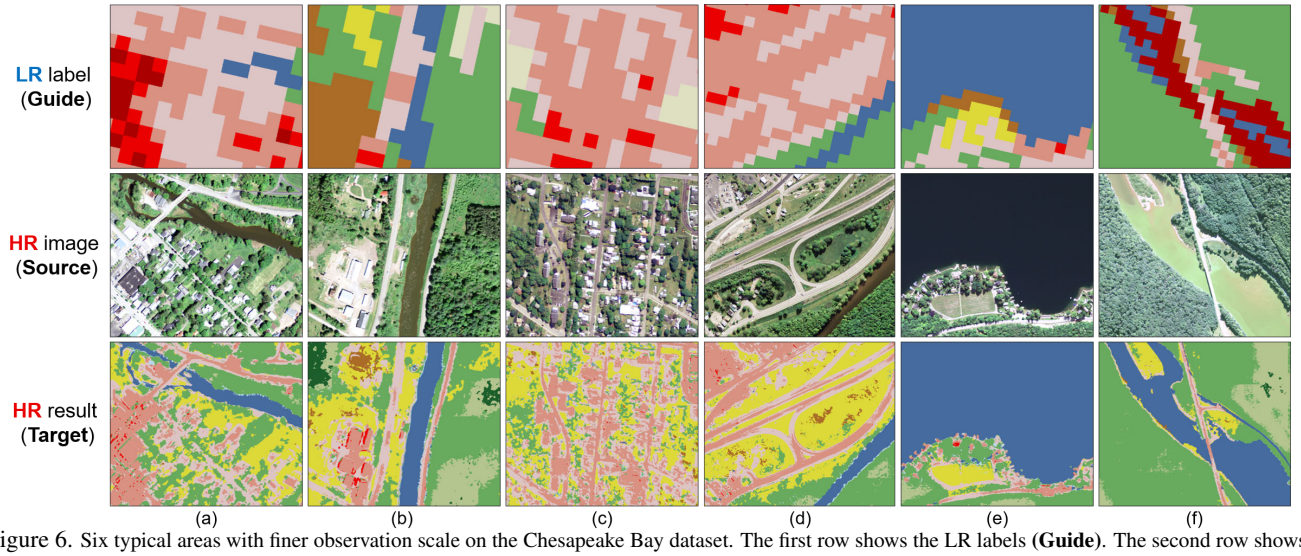

Figure 6. Six typical areas with finer observation scale on the Chesapeake Bay dataset. The first row shows the LR labels **(Guide)**. The second row shows the HR images **(Source)**. Third row shows the HR results **(Target)** produced by **Paraformer**.

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 × 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 224×224 and batch size of 8. The learning rate is set to 0.01 and would decrease by 10%

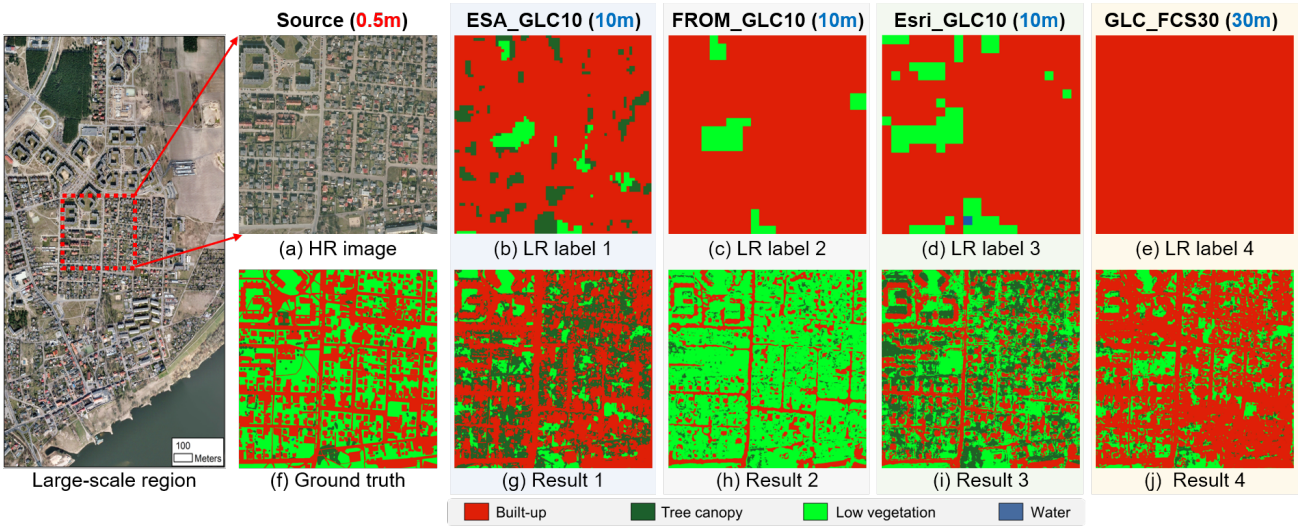

Figure 7. Visual results of **Paraformer** in the Poland dataset. The demonstration area is one of the training pieces sampled from large-scale training regions. (a–e) the training pairs of HR images (0.5 m/pixel) and four types of LR labels including ESA_GLC (10 m/pixel), FROM_GLC (10 m/pixel), Esri_GLC (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×	**Paraformer**	**65.57**	**71.43**	**70.20**	**60.04**	68.01	52.62	**64.65**
	L2HNet [28]	61.77	68.12	65.24	58.52	**69.39**	**55.43**	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	59.10	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	54.66	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]	**56.57**	50.15	38.44	39.36	41.03	43.56	43.66	27.14	21.48
	ESA_GLC10 [44]	**55.19**	52.13	35.58	36.09	38.42	40.05	49.81	28.34	26.97
	Esri_GLC10 [22]	**55.07**	50.78	37.79	38.78	38.50	39.91	46.65	28.18	19.36
120×	GLC_FCS30 [56]	**49.39**	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 hybrid 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

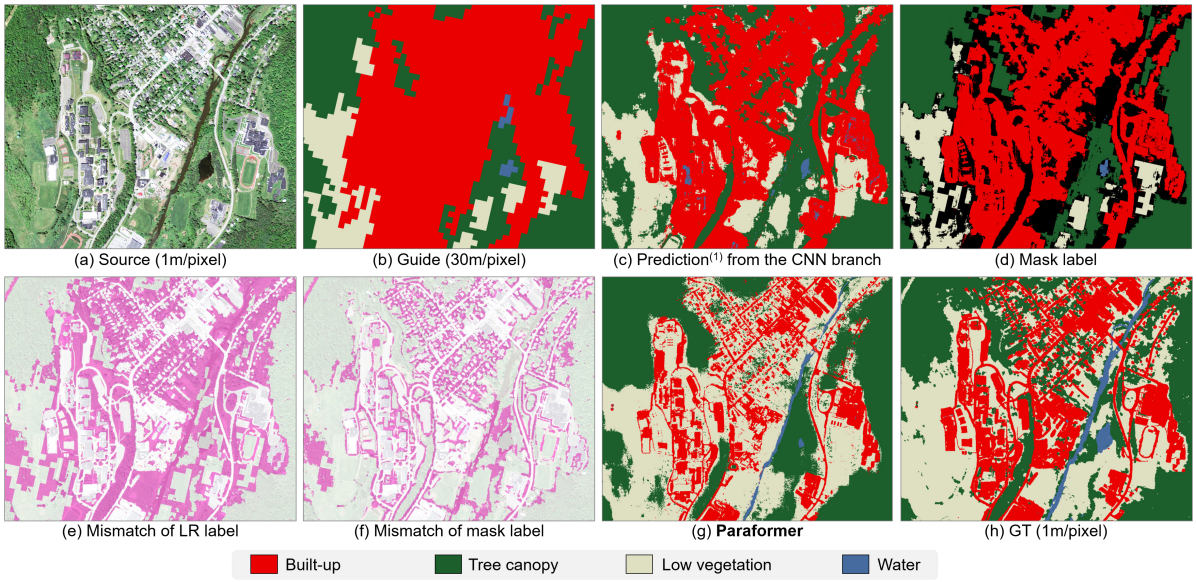

Figure 8. Example of training data and different outputs of Paraformer sampled from the Chesapeake Bay dataset with four unified classes. (a) HR images. (b) LR labels. (c) the primal prediction from the CNN branch. (d) Mask label, as the intersection parts of (b) and (c). The **black areas** are set 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 extract 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 Your	Maryland	Pennsylvania	Virginia	West Virginia	**Average**		
Paraformer	**65.57**	**71.43**	**70.20**	**60.04**	**68.01**	**52.62**	**64.65**	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	62.69	70.39	67.15	58.33	67.47	50.83	62.81	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.

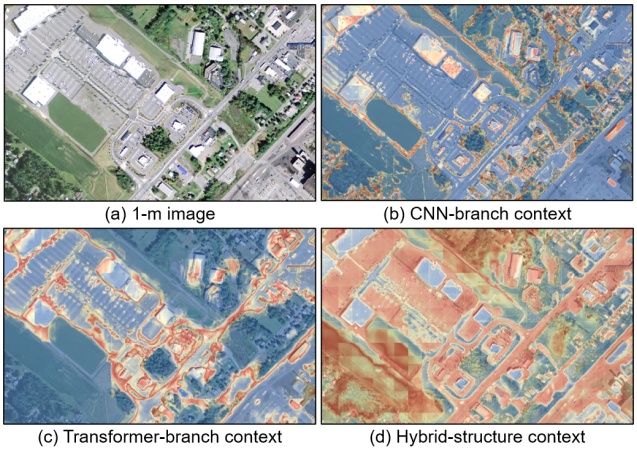

Figure 9. Demonstration of the extracted contexts from the ablation methods. (a) the original HR image. (b) the contexts extracted by the sole CNN branch. (c) the contexts extracted by the sole Transformer branch . (d) the contexts extracted by the CNN-Transformer hybrid backbone.

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).