This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2025.3609767

REST: Holistic Learning for End-to-End Semantic Segmentation of Whole-Scene Remote Sensing Imagery

Wei Chen, Lorenzo Bruzzone, *Fellow, IEEE,* Bo Dang, Yuan Gao, Youming Deng, Jin-Gang Yu, Liangqi Yuan, *Student Member, IEEE,* and Yansheng Li, *Senior Member, IEEE*

Abstract—Semantic segmentation of remote sensing imagery (RSI) is a fundamental task that aims at assigning a category label to each pixel. To pursue precise segmentation with one or more fine-grained categories, semantic segmentation often requires holistic segmentation of whole-scene RSI (WRI), which is normally characterized by a large size. However, conventional deep learning methods struggle to handle holistic segmentation of WRI due to the memory limitations of the graphics processing unit (GPU), thus requiring to adopt suboptimal strategies such as cropping or fusion, which result in performance degradation. Here, we introduce the **R**obust **E**nd-to-end semantic **S**egmentation architecture for whole-scene remo**T**e sensing imagery (REST). REST is the first intrinsically end-to-end framework for truly holistic segmentation of WRI, supporting a wide range of encoders and decoders in a plug-and-play fashion. It enables seamless integration with mainstream semantic segmentation methods, and even more advanced foundation models. Specifically, we propose a novel spatial parallel interaction mechanism (SPIM) within REST to overcome GPU memory constraints and achieve global context awareness. Unlike traditional parallel methods, SPIM enables REST to process a WRI effectively and efficiently by combining parallel computation with a divide-and-conquer strategy. Both theoretical analysis and experiments demonstrate that REST attains near-linear throughput scalability as additional GPUs are employed. Extensive experiments demonstrate that REST consistently outperforms existing cropping-based and fusion-based methods across a variety of scenarios, ranging from single-class to multi-class segmentation, from multispectral to hyperspectral imagery, and from satellite to drone platforms. The robustness and versatility of REST are expected to offer a promising solution for the holistic segmentation of WRI, with the potential for further extension to large-size medical imagery segmentation. The source code will be released at https://weichenrs.github.io/REST.

Index Terms—Holistic segmentation, whole-scene remote sensing imagery, spatial parallelism, semantic segmentation, deep learning.

1 Introduction

Remote sensing plays a crucial role in studying and understanding the causes and effects of global environmental changes [1], [2], [3] while also in promoting sustainable development [4], [5]. Many application tasks heavily rely on the capability to perform automatic semantic segmentation of remote sensing imagery (RSI) [6], [7], [8]. So far, although deep learning-based semantic segmentation of RSI has made much progress in recent years [9], [10], [11],

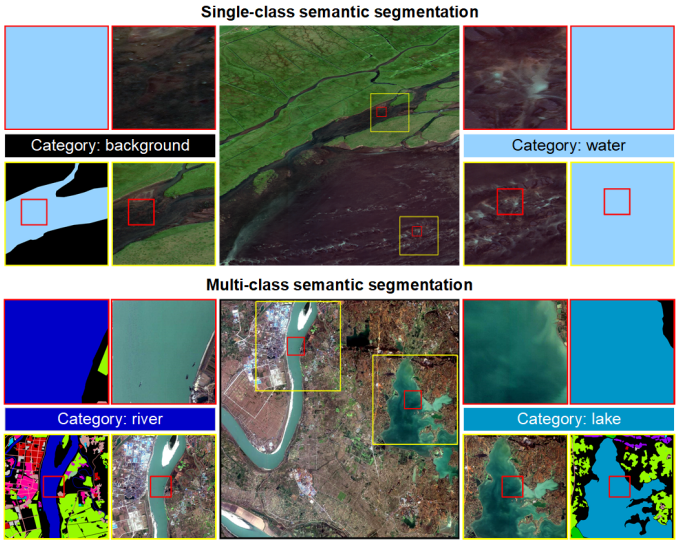


Fig. 1. **The necessity of holistic segmentation**. When relying solely on small-size tiles (red boxes), effectively segmenting geographical elements of interest is challenging in both single-class and multi-class semantic segmentation, due to intra-class variation and inter-class confusion. By expanding the field of view to larger tiles (yellow boxes) or even the WRIs, however, models can holistically leverage contextual information to achieve more precise and meaningful segmentation.

This work was supported by the National Key Research and Development Program of China under Grant 2024YFB3909001, and the National Natural Science Foundation of China under Grant 42371321. (Corresponding authors: Yansheng Li)

Wei Chen, Bo Dang, and Yansheng Li are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: weichenrs@whu.edu.cn; bodang@whu.edu.cn; yansheng.li@whu.edu.cn).
Lorenzo Bruzzone is with the Department of Information Engineering and Computer Science, University of Trento, Trento 38123, Italy (e-mail: lorenzo.bruzzone@unitn.it).
Yuan Gao is with the School of Artificial Intelligence, Wuhan University, Wuhan 430072, China (e-mail: Ethan.Y.Gao@gmail.com).
Youming Deng is with the Department of Computer Science, Cornell University, Ithaca 14853, NY, USA (e-mail: ymdeng@cs.cornell.edu).
Jin-Gang Yu is with the School of Automation Science and Engineering, South China University of Technology, Guangzhou 510641, China (e-mail: jingangyu@scut.edu.cn).
Liangqi Yuan is with the College of Engineering, Purdue University, West Lafayette 47907, IN, USA (e-mail: liangqiy@purdue.edu).

it is often limited to handling small-size image tiles due to graphics processing unit (GPU) memory limitations [12]. In particular, the recently developed remote sensing foundation models (RSFMs) [13], [14], [15], while demonstrating superior capabilities compared to traditional deep learning-based methods [16], [17], [18], are usually constrained to processing small image tiles due to their large number of parameters. Significantly, the limited information available within small-size image tiles poses a major challenge for the semantic segmentation of RSI [19].

As demonstrated in Fig. 1, when only a relatively small-size image tile (i.e., the red box) is provided, the limited receptive field in the figure makes it difficult to determine the specific category, particularly when the tile consists of many homogeneous pixels. This difficulty is further amplified when attempting to distinguish between fine-grained semantic categories that exhibit visual similarity in a small-size image tile, such as river and lake. In contrast, given a large-size image tile (i.e., the yellow box) or even the whole-scene RSI (WRI), we can holistically leverage more comprehensive contextual information, such as the structural patterns of geographical elements and the relationships between geographical elements, to facilitate meaningful segmentation. This enhanced contextual understanding naturally reveals the necessity of holistic segmentation of WRI (HSW), which is a promising research direction. Nevertheless, the GPU memory consumption grows near-linearly with image size, limiting the capability of common deep learning-based methods to achieve HSW.

In literature, several pioneering works have recognized the importance of HSW and proposed preliminary methods to address this challenge. For example, some methods leverage global contextual information by exploiting and fusing global-local complementary information from downsampled WRI and cropped image tiles [20], [21], [22]. However, the indirect downsampling and fusing results in suboptimal performance without efficiency improvement. Additionally, although similar problems have been investigated in related fields, such as whole-slide image (WSI) recognition in medical image analysis [23], [24], [25] and long-context extension in large language models [26], [27], [28], it is non-trivial to adapt these methods to HSW due to the significant task differences. Overall, existing deep learning methods fail to enable end-to-end HSW.

In this paper, we present the **R**obust **E**nd-to-end semantic **S**egmentation architecture for whole-scene remo**T**e sensing imagery (REST) to directly achieve HSW. In contrast to existing methods [20], [21], [22], which focus on either cropping or the fusion of global and local context, our REST aims to address deep learning-based HSW through an intrinsic end-to-end solution. To tackle the critical memory limitations of individual GPUs, REST distributes the computational load of WRI across multiple GPUs. Based on this, a spatial parallel interaction mechanism (SPIM) is designed to facilitate information exchange among multiple GPUs, ensuring comprehensive exploration and utilization of global context. The computation in REST is performed in parallel without significantly compromising efficiency, and REST demonstrates scalable throughput (i.e., processable image size) with an increasing number of GPUs, thereby effectively enabling HSW. Furthermore, REST naturally supports advanced semantic segmentation methods with various encoders and decoder in a plug-and-play fashion, such as the emerging RSFMs [13], [14], [15].

Specifically, REST starts by dividing the WRI into multiple spatial regions, each processed by a single encoder on separate GPUs, as illustrated in Fig. 2a and b. Initially, those separate encoders operate independently, extracting local spatial features corresponding to their assigned regions. These features are then aggregated and enhanced via SPIM, where all-to-all communication facilitates information exchange across encoders from separate GPUs along both spatial and channel dimensions. Each GPU shares the entire features of its corresponding local region with others and receives partial-channel features from all remaining local features on other GPUs, covering the entire spatial region of the WRI. This exchange enables comprehensive feature interaction across GPUs. After that, local features are concatenated and fused using multi-head self-attention to capture global context. The globally enriched features are redistributed back to their respective GPUs based on the original spatial regions via another round of all-to-all communication. As a result, each GPU retains its local features while integrating global contextual information, enhancing the overall representation of the WRI. The enhanced features are then passed to the decoders on separate GPUs to perform final predictions. Due to SPIM's ability to extract and leverage global context, spatial integrity is maintained among these separate predictions, ensuring deep learning-based HSW. It should be emphasized that REST can be regarded as a novel feature-level distributed parallel training method. While existing parallelization approaches (e.g., data parallelism, model parallelism, pipeline parallelism, tensor parallelism, and sequence parallelism) fail to support deep learning-based HSW, REST enables HSW with lossless global interaction across the spatial dimension.

To verify the effectiveness of our proposed REST architecture, we conduct experiments across diverse semantic segmentation tasks, as shown in Fig. 2c. Experiments demonstrate that REST substantially outperforms state-of-the-art (SOTA) methods for HSW. In summary, our key contributions are three-fold:

- REST, a fundamentally new end-to-end framework for holistic whole-scene remote sensing segmentation, is presented. REST processes entire scenes in a single pass without cropping or stitching, integrates seamlessly with diverse encoder–decoder backbones in a plug-and-play fashion. Furthermore, REST can leverage the powerful representations of RSFMs for further gains.
- We propose SPIM, an innovative spatial parallel interaction mechanism that dynamically partitions whole-scene feature maps across multiple GPUs and aggregate global context information effectively. SPIM overcomes single-GPU memory limitations and delivers near-linear throughput scalability as additional GPUs are added.
- Extensive experiments on four WRI benchmarks, a medium-size RSI dataset, and three medical imagery segmentation tasks demonstrate that REST consistently outperforms state-of-the-art methods across a wide range of scenarios. Notably, REST achieves strong performance on large-size medical imagery, highlighting its scalability and generalization across domains.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2025.3609767

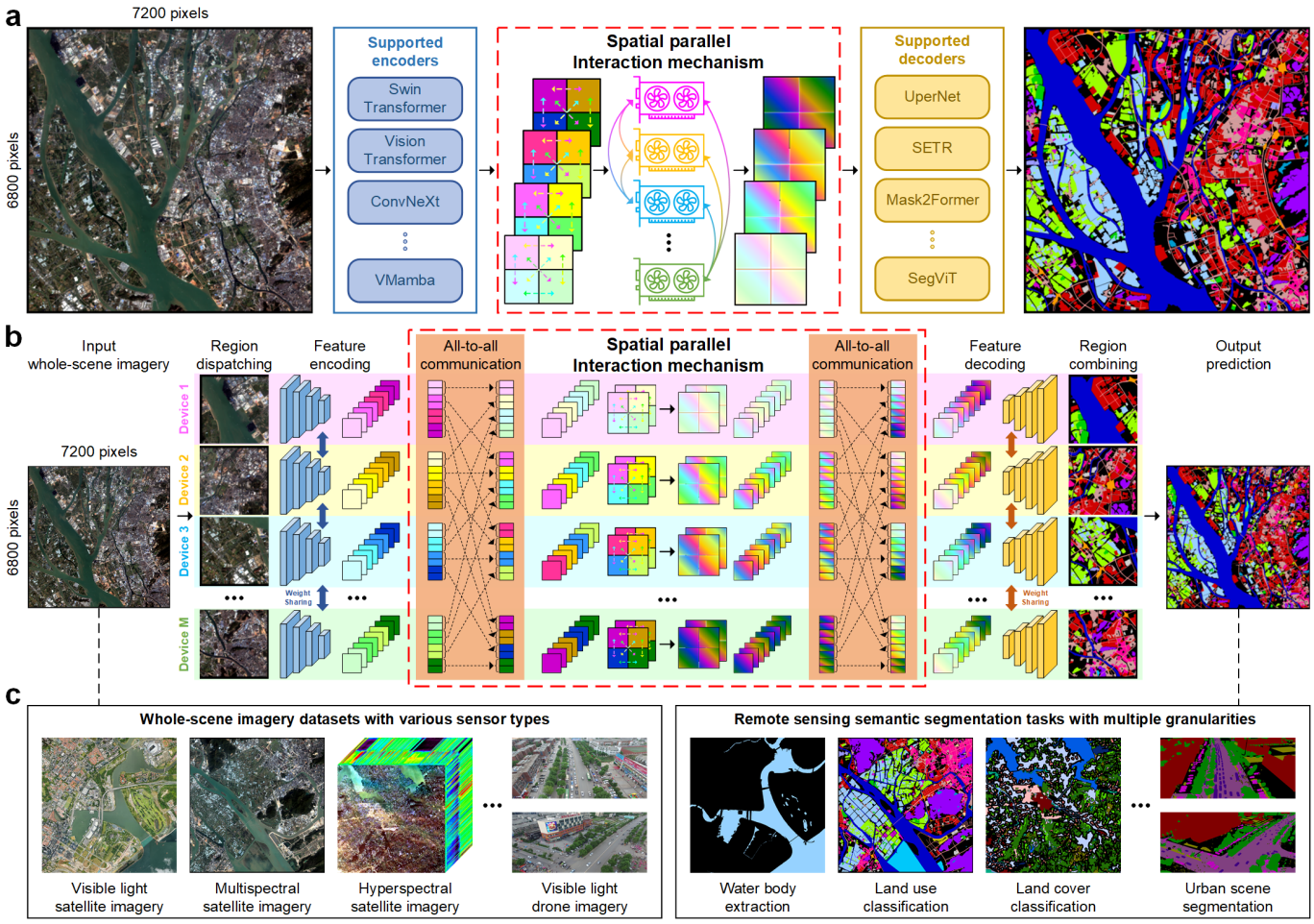


Fig. 2. **Overview of the Robust End-to-end semantic Segmentation architecture for whole-scene remoTe sensing imagery (REST).** (a) REST is built on an encoder, a decoder, and a novel SPIM module bridging them to capture and exploit global contextual information. Note that REST can support various encoders and decoders. (b) The input WRI is first divided based on the number of available GPUs, with each region distributed to its corresponding GPU for feature encoding. Within our SPIM, the encoded features undergo all-to-all communication and multi-head attention calculations, interacting with features from other GPUs to extract global context. These globally enriched features are then redistributed back to their respective GPUs through an inverse communication operation for decoding. Finally, the processed features are combined to yield a holistic prediction of the WRI. Despite performing feature encoding and decoding operations independently on each GPU, REST effectively mines the global spatial context of the entire scene within SPIM. (c) REST can accommodate different data types and handle diverse semantic segmentation tasks, underscoring its versatility.

The remainder of this paper is organized as follows: Section 2 reviews the work related to this paper. Section 3 details the proposed REST framework. Section 4 presents and analyses experimental results. Section 5 discusses REST's advantages and characteristics. Section 6 concludes the paper with future directions.

2 Related Work

2.1 Semantic Segmentation

Semantic segmentation aims to assign a category label to each pixel in an image. The advancement of deep learning has revolutionized semantic segmentation, with convolutional neural networks (CNNs) and Transformer architectures achieving unprecedented performance through hierarchical feature learning. However, when processing WRIs, GPU memory limitations pose a significant challenge. To address this issue, existing techniques primarily adopt two approaches: cropping and fusion.

2.1.1 *Cropping-based methods*

Cropping-based methods divide WRI into local patches to alleviate memory constraints [29], [30], [31]. Early approaches like PSPNet [32] introduce spatial pyramid pooling modules to enhance the contextual awareness of local patches through multi-scale feature fusion. DeepLab-v3+ [33] employs an encoder-decoder architecture with depthwise separable convolutions, optimizing computational efficiency while preserving details. HRNet [34] maintains continuous high-resolution representations from low to high-level feature streams through parallel multi-resolution subnetworks, significantly improving the segmentation accuracy. With the rise of Transformer architectures [35], [36], SegFormer [37] proposes a hierarchical encoder without position encoding, combined with a lightweight decoder, providing flexible support for dynamic cropping inputs. Although these cropping-based methods effectively reduce memory usage, they introduce artifacts like boundary discontinuities due to the cropping process.

Recently, RSFMs have significantly advanced the performance of semantic segmentation [13], [14], [15]. RingMo [16] introduces a framework that employs generative self-supervised learning on a large dataset of two million images, excelling in detecting small objects in complex scenes. GFM [17] proposes an efficient method for building geospatial foundation models through continual pretraining from ImageNet [38] on a diverse dataset. Satlas [14] provides a large-scale dataset with extensive annotations to enhance deep learning models' performance on geospatial tasks. ScaleMAE [15] focuses on learning multiscale representations via masking and reconstruction at different scales. SatMAE++ [18] extends Transformer-based pretraining to incorporate multiscale information from various sensors, with upsampling for reconstruction. SkySense develops a multimodal foundation model using a unified self-supervised learning framework to interpret diverse imagery, including optical and synthetic aperture radar data [13]. While these RSFMs achieve improved performance, their massive parameter sizes impose high demands on GPU memory, further limiting their application in WRI.

2.1.2 Fusion-based methods

To mitigate the information fragmentation in cropping strategies, researchers have developed fusion-based semantic segmentation methods that integrate local details with global contexts [12], [20], [21], [22]. GLNet [12] proposes an interactive architecture with global and local branches. The global branch processes downsampled images to extract semantic contexts, while the local branch focuses on original resolution patches to capture details. LCF-ALE [20] further introduces a local-aware context correlation mechanism, utilizing graph convolution to model the spatial topology of adjacent patches, significantly improving segmentation consistency in complex texture areas. MagNet [21] adopts a progressive coarse-to-fine segmentation strategy, generating a low-resolution semantic map to capture global structures and refining edge accuracy through deformable convolution alignment with high-resolution features. ISDNet proposes a relation-aware fusion module that combines shallow texture features with deep semantic features, achieving adaptive weighted integration through channel attention mechanisms. These approaches leverage innovative feature interaction mechanisms and hierarchical multi-scale modeling to enhance global semantic consistency while preserving localized details [22]. However, their practical implementation faces two key challenges: downsampling introduces information degradation and noise propagation, and intensive data I/O interactions result in longer training times that hinder the scalability of these methods.

In summary, while existing methods partially mitigate GPU memory consumption and contextual preservation challenges, neither cropping-based nor fusion-based approaches fundamentally achieve end-to-end HSW.

2.2 Long-sequence Processing

The computational demands of WRI, stemming from its large size, align fundamentally with challenges in long-sequence processing. Key strategies emerge from two domains: whole-slide image recognition addresses massive inputs through tiling or multi-resolution analysis while preserving spatial relationships, and long-context extension overcomes sequence length limits via memory-efficient attention and distributed processing. Both employ hierarchical processing to balance local details and global context, which is potentially adaptable to WRI, while requiring adaptation to its unique spatial semantics.

2.2.1 Whole-slide Image Recognition

WSIs are crucial in computational pathology for their unparalleled ability to capture detailed tissue structures, yet their huge size presents formidable computational challenges. To tackle these, several methods have been designed with unique strategies [23], [24], [25], [39], [40]. Streaming CNN [23] uses a tiling approach, dividing large WSIs into smaller tiles to manage memory efficiently, enabling analysis of images that exceed typical memory limits while retaining essential details. Building on this, Enhanced Streaming CNN [39] enhances efficiency with optimizations that reduce memory usage further and boost processing speed, ideal for time-sensitive clinical applications. Conversely, Prov-GigaPath [24] employs a Transformer model to handle extensive contextual data across numerous tiles, capturing long-range dependencies vital for understanding complex tissue patterns in WSIs. Whole-Slide Training [25] takes a direct route, processing entire WSIs without tiling by leveraging advanced memory management across multiple GPUs, preserving image integrity for potentially superior diagnostic outcomes. Meanwhile, STAMP offers a flexible toolkit with multi-resolution analysis, allowing adjustment of detail levels to meet varied analysis needs of WSIs, supporting both detailed and broad perspectives [40]. These methods address the size and complexity of WSIs, advancing computational pathology by enabling more effective and efficient diagnostic tools.

2.2.2 Long-context Extension

Processing and effectively utilizing long sequences poses a significant challenge for deep learning models, especially Transformer-based large language models, where the quadratic complexity of self-attention traditionally limits context length. To address this, researchers have developed novel model architectures and advanced training techniques [26], [27], [28], [41]. Early efforts focused on architectural modifications. Transformer-XL [26], for instance, introduces segment-level recurrence and state reuse to enable information flow beyond fixed-length segments, effectively extending the contextual horizon. Approaches like Longformer [27] employ sparse attention mechanisms to reduce the computational cost, thereby allowing models to handle much longer documents directly. Complementary to extending context, scaling the training process for the resulting super-large models is crucial, with research exploring efficient distributed training techniques, potentially using 2D parallelism, to make training feasible [41]. DeepSpeed-Ulysses tackles memory and scalability barriers in long-sequence Transformer training through sequential parallelism and optimized communication, enabling linear scaling. Its modular design supports diverse attention mechanisms and integrates with ZeRO-3 for joint model and context scaling, circumventing quadratic complexity [28]. These advances

address both the technical challenge of processing longer sequences and the fundamental requirement for models to utilize information across extended contexts.

In general, although methods for whole-slide image recognition and long-context extension provide valuable insights into long-sequence processing, their significant differences from the HSW task make adapting these methods to HSW a non-trivial challenge.

3 METHODS

3.1 Overview

To clarify how REST works, we detail its steps in Algorithm 1. As shown in Fig. 2a and b, REST allows for the efficient handling of large-size images by distributing the computational load of WRI and leveraging the communication across multiple GPUs, overcoming the limitation of global context awareness and enhancing the model's ability to capture long-range dependency. Assuming we have M GPUs, the objective function of REST can be defined as follows:

$$\begin{aligned} \min_{\theta} \mathcal{L}(\theta) &= \sum_{i=1}^{M} \mathcal{L}^{(i)}(\theta), \\ \mathcal{L}^{(i)}(\theta) &= \mathbb{E}_{(\boldsymbol{X}^{(i)}, \boldsymbol{Y}^{(i)}) \sim \mathcal{D}} \left[\ell(\theta; \boldsymbol{X}^{(i)}, \boldsymbol{Y}^{(i)}) \right], \end{aligned} \tag{1}$$

where $\theta = \{\theta_{\text{enc}}, \theta_{\text{spim}}, \theta_{\text{dec}}\}$, and θ_{enc}, θ_{spim}, and θ_{dec} represent the parameters of the encoder, SPIM, and decoder in REST, respectively. $(\boldsymbol{X}^{(i)}, \boldsymbol{Y}^{(i)})$ denotes the WRI-label pair in the dataset $\mathcal{D}$. Given the objective function, we describe the REST algorithm step-by-step in the following.

3.2 Region Dispatching

Let $\boldsymbol{X} \in \mathbb{R}^{H \times W \times C}$ denote the input WRI, where H,W represent the height and width of the WRI respectively, and C denotes the number of channels. The input image is dispatched along its height and width into M different, non-overlapping regions:

$$\phi : \mathbb{R}^{H \times W \times C} \to \mathbb{R}^{M \times \left(\frac{H \times W}{M} \times C\right)}, \tag{2}$$

$$\left[\boldsymbol{X}^{(1)}; \boldsymbol{X}^{(2)}; \ldots; \boldsymbol{X}^{(M)}\right] = \phi(\boldsymbol{X}), \quad \boldsymbol{X}^{(i)} \in \mathbb{R}^{\frac{H \times W}{M} \times C}, \tag{3}$$

where ϕ denotes the dispatching operation, and $\boldsymbol{X}^{(i)}$ represents the dispatched region of $\boldsymbol{X}$ on GPU i. Concretely, we first compute $\log_2(M)$ and then divide this quantity into two parts to determine the vertical and horizontal partition counts:

$$sp_h = 2^{\lfloor \frac{1}{2} \log_2(M) \rfloor}, \tag{4}$$

$$sp_w = 2^{\log_2(M) - \lfloor \frac{1}{2} \log_2(M) \rfloor}, \tag{5}$$

such that $sp_h \times sp_w = M$, where $\lfloor \cdot \rfloor$ denotes the floor operation. Each GPU process, identified by a global rank index $rank \in \{0, \ldots, M-1\}$, then computes its assigned subregion indices as

$$ind_w = rank \bmod sp_h, \tag{6}$$

$$ind_h = \lfloor \tfrac{rank}{sp_h} \rfloor. \tag{7}$$

We partition $\mathbf{X}$ into sp_h horizontal strips and select the ind_h-th one, then further partition that strip into sp_w vertical blocks and select the ind_w-th block. This procedure yields M subregion pairs $\{\mathbf{X}^{(i)}\}_{i=1}^{M}$:

$$\boldsymbol{X}^{(i)} \in \mathbb{R}^{\frac{H}{sp_h} \times \frac{W}{sp_w} \times C} \cong \mathbb{R}^{\frac{H \times W}{M} \times C}, \tag{8}$$

which are then processed independently on the M GPUs. In practice, the choice of M can be flexibly adjusted according to the size of WRI and the number of available GPUs, typically set to a power of two.

Algorithm 1 Algorithm of the proposed REST

Input: Encoder model θ^1_{enc} initialized from pretrained model; Randomly initialized SPIM θ^1_{spim} and decoder model θ^1_{dec}; WRI dataset $\boldsymbol{D}$; number of available GPUs M; number of total iterations n_{iter}; learning rate η_θ
Output: Converged encoder model θ_{enc}, SPIM θ_{spim}, and decoder models θ_{dec}; predicted segmentation results $\boldsymbol{P}$
for $t = 1, \ldots, n_{iter}$ **do**
  if not converged **then**
    Sample WRI $\boldsymbol{X}$ and ground truth $\boldsymbol{Y}$ from $\boldsymbol{D}$
    $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \ldots, \boldsymbol{X}^{(M)} \leftarrow \phi(\boldsymbol{X})$ ▷ Eq. 3
    $\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)}, \ldots, \boldsymbol{Y}^{(M)} \leftarrow \phi(\boldsymbol{Y})$
    for $i = 1, \ldots, M$ **in parallel do**
      $\boldsymbol{F}^{(i)} \leftarrow \theta^t_{\text{enc}}(\boldsymbol{X}^{(i)})$ ▷ Eq. 5
      for $j = 1, \ldots, M$ **do**
        GPU i sends $\boldsymbol{F}^{(i)}_j$ to all other GPUs
        GPU i receives $\boldsymbol{F}^{(j)}_i$ from all other GPUs
        $\boldsymbol{G}^{(i)}_j \leftarrow \boldsymbol{F}^{(j)}_i$ ▷ Eq. 6
      end for
      $\boldsymbol{G}^{(i)} \leftarrow \theta^t_{\text{spim}}\left([\boldsymbol{G}^{(i)}_1, \boldsymbol{G}^{(i)}_2, \ldots, \boldsymbol{G}^{(i)}_M]\right)$ ▷ Eq. 8
      for $k = 1, \ldots, M$ **do**
        GPU i sends $\boldsymbol{G}^{(i)}_k$ to all other GPUs
        GPU i receives $\boldsymbol{G}^{(k)}_i$ from all other GPUs
        $\boldsymbol{F}^{(i)}_k \leftarrow \boldsymbol{G}^{(k)}_i$ ▷ Eq. 9
      end for
      $\boldsymbol{P}^{(i)} \leftarrow \theta^t_{\text{dec}}(\boldsymbol{F}^{(i)})$ ▷ Eq. 11
    end for
    $\mathcal{L} \leftarrow \sum_{i=1}^{M}(\mathcal{L}_i(\boldsymbol{P}^{(i)}, \boldsymbol{Y}^{(i)}))$
    $\theta^{t+1} \leftarrow \theta^t - \eta_\theta \nabla \mathcal{L}(\theta^t)$
  end if
end for

3.3 Feature Encoding

After the dispatching, $\boldsymbol{X}^{(i)}$ is encoded into features on its corresponding GPU i:

$$\theta_{\text{enc}} : \mathbb{R}^{\frac{H \times W}{M} \times C} \to \mathbb{R}^{(N \times D)}, \tag{9}$$

$$\boldsymbol{F}^{(i)} = \theta_{\text{enc}}(\boldsymbol{X}^{(i)}), \quad \boldsymbol{F}^{(i)} \in \mathbb{R}^{N \times D}, \tag{10}$$

where $\boldsymbol{F}^{(i)}$ denotes the encoded features, D denotes the number of features, and $N = \frac{H \times W}{M}$. The encoder network model θ_{enc} normally contains several network layers, such as convolutional, attention, and mamba layers. Without loss of generality, we can assume that $\boldsymbol{F}^{(i)}$ and $\boldsymbol{X}^{(i)}$ are consistent in spatial dimension, unaffected by downsampling in the encoder. Note that the encoder typically outputs features from multiple stages for subsequent operations, but we only present the features from a single stage here for simplicity. This feature encoding process yields features containing only information from the corresponding region.

3.4 Spatial Parallel Interaction

To enhance the spatial interaction across the entire WRI, we design a novel spatial parallel interaction mechanism utilizing the all-to-all communication operation. Specifically, the encoded features $\boldsymbol{F}^{(i)}$ are first split into M sub-features (i.e., attention head) along the channel dimension, with each attention head maintaining a sub-feature $\boldsymbol{F}_j^{(i)} \in \mathbb{R}^{N\times\frac{D}{M}}$. The number of attention heads is kept equal to the number of GPUs to ensure a one-to-one mapping between attention heads and compute devices. Features on different GPUs interact with each other in the all-to-all communication. This can be represented as

$$\boldsymbol{G}_j^{(i)} = \boldsymbol{F}_i^{(j)}, \quad \forall i,j \in \{1,2,\ldots,M\}, \tag{11}$$

where each GPU i sends its j-th sub-feature in $\boldsymbol{F}^{(i)}$ to GPU j and receives the corresponding sub-features from the other GPUs. $\boldsymbol{G}_j^{(i)}$ denotes the sub-feature that GPU i receives from GPU j on attention head i after the communication. This process can also be regarded as a transpose operation between the spatial dimension and the channel dimension from a global perspective. Then, features on the same GPU i are aggregated by

$$\theta_{\text{spim}} : \mathbb{R}^{N\times D} \rightarrow \mathbb{R}^{(N\times M)\times\frac{D}{M}} \rightarrow \mathbb{R}^{N\times D}, \tag{12}$$

$$\boldsymbol{G}^{(i)} = \theta_{\text{spim}}\left([\boldsymbol{G}_1^{(i)}, \boldsymbol{G}_2^{(i)}, \ldots, \boldsymbol{G}_M^{(i)}]\right), \boldsymbol{G}_j^{(i)} \in \mathbb{R}^{N\times\frac{D}{M}}, \tag{13}$$

where θ_{spim} is composed of multiple self-attention layers, and $\boldsymbol{G}^{(i)} \in \mathbb{R}^{(N\times M)\times\frac{D}{M}} \cong \mathbb{R}^{H\times W\times\frac{D}{M}}$. Thus, the computation in SPIM is distributed across M GPUs according to the channel dimension instead of the spatial dimension, with each GPU responsible for features of $\frac{D}{M}$ channels across the entire $H \times W$ spatial region, maintaining the spatial integrity. SPIM ensures a comprehensive and seamless exchange of features, enabling the model to extract global context information. With SPIM, REST achieves near-linear throughput-scalability in handling WRI. After that, the features are sent back to their corresponding GPUs through the all-to-all communication:

$$\boldsymbol{F}_i^{(j)} = \boldsymbol{G}_j^{(i)}, \quad \forall i,j \in \{1,2,\ldots,M\}, \tag{14}$$

3.5 Feature Decoding

The extracted features from region i are fed into the decoder, which processes them to generate predictions $\boldsymbol{P}^{(i)}$:

$$\theta_{\text{dec}} : \mathbb{R}^{(N\times D)} \rightarrow \mathbb{R}^{\frac{H\times W}{M}\times K}, \tag{15}$$

$$\boldsymbol{P}^{(i)} = \theta_{\text{dec}}(\boldsymbol{F}^{(i)}), \quad \boldsymbol{P}^{(i)} \in \mathbb{R}^{\frac{H\times W}{M}\times K}, \tag{16}$$

where K represents the number of categories. Predictions $\boldsymbol{P}^{(i)}$ and corresponding ground truth labels $\boldsymbol{Y}^{(i)}$ are used to compute the loss and optimize the whole model $\theta = \{\theta_{\text{enc}}, \theta_{\text{spim}}, \theta_{\text{dec}}\}$. In line with prevailing methods, we employ the UPerNet [42] architecture as our decoder, which takes multiscale features as input to handle targets of different scales better. Notably, REST can also be adapted to many other decoder architectures, as illustrated in Fig. 10d and Supplementary Figure S2.

3.6 Region Combining

During the testing stage, predictions on each GPU are gathered on GPU 1 to generate the whole prediction $\boldsymbol{P}$.

$$\psi : \mathbb{R}^{M\times(\frac{H\times W}{M}\times K)} \rightarrow \mathbb{R}^{H\times W\times K}, \tag{17}$$

$$\boldsymbol{P} = \psi(\left[\boldsymbol{P}^{(1)}; \boldsymbol{P}^{(2)}; \ldots; \boldsymbol{P}^{(M)}\right]), \quad \boldsymbol{P} \in \mathbb{R}^{H\times W\times K}, \tag{18}$$

where ψ denotes the combining operation. It is noted that the proposed REST is model-agnostic, which can adapt to different types of encoder and decoder network models, including CNN [43], Transformer [36], as well as the latest Mamba [44]. It is worth noting that, since REST distributes the storage and computational load across GPUs, it achieves better parallelism and much higher computational efficiency than global-local fusion-based methods. At the same time, it maintains high precision.

3.7 A General Framework for Holistic Segmentation

Our REST framework delivers a general holistic segmentation pipeline that readily supports both current and future backbones (e.g. the emerging RSFMs). Rather than merely sharding input tiles across GPUs [45] or splitting network layers [41], we partition the image into independent regions and employ our SPIM for full cross-region feature exchange and unified attention. This strategy eliminates boundary artifacts, consolidates global context, achieves near-linear scaling as GPUs are added, and supports genuine plug-and-play integration with a wide range of segmentation architectures.

In practical implementation, REST requires fewer computational resources during inference than in the training phase, and it can be deployed on a reduced number of GPUs or within smaller GPU memory, while still preserving the benefits of global feature fusion. Users maintain full flexibility in selecting input resolution and region partitioning, and may incorporate their preferred encoder–decoder architectures or leverage pretrained weights obtained from specific datasets with minimal adaptation overhead. This combination of efficiency and adaptability makes REST a simple but effective solution for HSW.

4 Experiments and Analysis

4.1 Experimental Setup

4.1.1 Datasets for HSW

To validate the effectiveness of REST, we specifically select four representative HSW datasets for experimentation including: the GLH-Water dataset [49], with an image size of 12,800×12,800 and three spectral bands; the Five-Billion-Pixels dataset [50], with an image size of 6,800×7,200 and four spectral bands; the UAVid dataset [51], with an image size of about 4,096×2,160 and three spectral bands; and the WHU-OHS dataset [52], with an image size of around 6,272×6,272 and twenty-four spectral bands. Details about the datasets can be found in Supplementary Section 8.

TABLE 1
Performance comparison of REST and different methods on various whole-scene remote sensing image segmentation tasks. The numbers in parentheses (e.g., +4.22) at the bottom of the table represent the accuracy improvements achieved by the REST method compared to the cropping-based baseline methods (e.g., 82.24→86.46).

Methods			GLH-Water (12,800×12,800×3)		Five-Billion-Pixels (6,800×7,200×4)		WHU-OHS (6,272×6,272×24)		UAVid (4,096×2,160×3)	
Type	Encoder	Decoder	IoU	$F1$	$mIoU$	OA	$mIoU$	OA	$mIoU$	$mAcc$
Cropping-based methods	ResNet-101 [46]	PSPNet [32]	77.32	87.21	64.48	89.02	20.77	45.72	69.14	78.93
	ResNet-101 [46]	DeepLab-v3+ [33]	78.54	87.98	64.01	89.00	18.14	45.26	69.80	78.70
	HRNet-48 [34]	FCN [47]	76.27	86.53	64.84	88.95	17.76	46.89	69.50	78.80
	HRNet-48 [34]	OCRNet [29]	78.77	88.13	63.91	89.33	21.60	<u>48.47</u>	70.19	79.63
	ResNet-101 [46]	PointRend [30]	73.59	84.79	60.93	88.16	20.22	47.16	68.34	78.19
	STDC-1446 [31]	STDC [31]	75.82	86.25	50.94	85.70	17.23	41.13	63.52	73.29
	MIT-B5 [37]	SegFormer [37]	82.77	90.57	65.73	89.30	18.19	44.24	71.27	80.78
	ConvNeXt-v2-Large [43]	UPerNet [42]	82.24	90.25	59.72	88.13	13.26	32.68	68.38	77.74
	VMamba-Base [44]	UPerNet [42]	83.12	90.78	68.00	90.52	20.40	46.84	<u>72.00</u>	<u>81.68</u>
	Swin-Large-SkySense [13]	UPerNet [42]	86.64	92.84	<u>69.68</u>	90.51	20.33	46.04	70.20	79.13
Fusion-based methods	ResNet-50 [46]	GLNet [12]	-	-	44.73	-	-	-	-	-
	VGG-16 [48]	LCF-ALE [20]	74.92	85.66	48.28	-	-	-	-	-
	ResNet-50 [46]	MagNet [21]	62.77	-	44.20	-	-	-	-	-
	ResNet-18 [46]	ISDNet [22]	53.04	-	21.98	-	-	-	-	-
Our REST (holistic segmentation methods)	ConvNeXt-v2-Large [43]	UPerNet [42]	86.46 (+4.22)	92.74 (+2.49)	61.57 (+1.85)	88.50 (+0.37)	13.54 (+0.28)	34.60 (+1.92)	69.95 (+1.57)	79.08 (+1.34)
	VMamba-Base [44]	UPerNet [42]	<u>89.08</u> (+5.96)	<u>94.09</u> (+3.31)	69.41 (+1.41)	<u>91.96</u> (+1.44)	<u>21.97</u> (+1.57)	47.78 (+0.94)	**73.16** (+1.16)	**82.30** (+0.62)
	Swin-Large-SkySense [13]	UPerNet [42]	**89.15** (+2.51)	**94.26** (+1.42)	**72.95** (+3.27)	**92.78** (+2.27)	**22.81** (+2.48)	**49.26** (+3.22)	71.77 (+1.57)	80.54 (+1.41)

Bold value and <u>underlined</u> value indicate the best and the second-best performance, respectively.

4.1.2 Evaluation Metrics

To access the performance, we use metrics including overall accuracy (OA), mean intersection over union (mIoU), F1-score, and mean accuracy (mAcc). For the GLH-Water [49] dataset, we report only the IoU metric for the positive class (i.e., water), excluding the background class. Additionally, we report the F1-score for GLH-Water because it provides a comprehensive reflection of both precision and recall. For the remaining three multi-class segmentation datasets (Five-Billion-Pixels [50], WHU-OHS [52], and UAVid [51]), we provide the IoU metrics for each category, along with mIoU and OA to reflect overall segmentation performance. Specifically, for UAVid, we report mAcc instead of OA to align with the results from Ringmo-Aerial [53].

4.1.3 Architectural Configuration

We equip REST with three distinct types of typical encoders, including a CNN (ConvNeXt-V2-Large [43]), a state space model (VMamba-Base [44]), and a Transformer (Swin-Large [36]). Specifically, we utilize the SkySense-pretrained model for Swin-Large [13], and ImageNet-pretrained models for the other two (see Supplementary Section 6 and Supplementary Figure S5 for details). The prevalent UPerNet [42] is selected as the decoder due to its powerful capabilities. It should be noted that we also conduct experiments on more encoders and decoders, as shown in Fig. 10d and Supplementary Figure S2.

4.1.4 Implementation Details

REST is developed using the mmsegmentation framework. During the training and testing phases, the required number of NVIDIA A100 GPUs ranges from 1 to 16, depending on the chosen sizes during training and testing. It is important to note that REST imposes no restrictions on the number of GPUs that can be used. With a sufficient number of GPUs, REST can scale up the throughput to handle images with larger sizes. The comparison models are trained using the same random seed and the default parameters, initialized with ImageNet-pretrained models. For our REST, the ConvNeXt and VMamba models are initialized using ImageNet-pretrained models [38] from corresponding repositories. For the Swin model, we use SkySense-pretrained models [13] for weight initialization. The model parameters are tuned with the AdamW optimizer, based on the minimization of cross-entropy. We use a two-stage-learning strategy to improve the training efficiency of our REST. Specifically, we first train models on cropped image tiles, use the trained weights to initialize the REST, and then continue training on WRIs. The REST models are trained on WRIs for 10,000 iterations. All comparison models are trained for 80,000 iterations on cropped image tiles according to the official parameter settings in the mmsegmentation framework.

4.1.5 Baseline Methods

We compare REST with SOTA semantic segmentation methods. These methods can be categorized into two groups: cropping-based methods and fusion-based methods. For cropping-based methods [29], [30], [31], [32], [33], [34], [37], the WRI is divided into smaller tiles for training and testing. These methods often fall short in effectively introducing global context information. We also implement specific models for water body extraction [54], [55], UAV image

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2025.3609767

TABLE 2
Performance comparison of REST and different methods on selected categories of Five-Billion-Pixels datasets. The $IoU(\%)$ value of each category is presented. The complete table of 24 categories is provided in Supplementary Table S17. The abbreviations and corresponding full names of categories are provided in the Supplementary Table S18.

Type	Encoder	Decoder	Pond	Stad	Rive	Park	Rail	Road	Lake	Snow	Average
Cropping-based methods	ResNet-101 [46]	PSPNet [32]	35.6	46.4	74.4	50.8	45.5	73.8	82.6	66.8	64.48
	ResNet-101 [46]	DeepLab-v3+ [33]	34.0	44.6	75.2	52.3	48.6	74.2	83.0	63.7	64.01
	HRNet-48 [34]	FCN [47]	29.6	48.4	73.1	49.8	48.4	74.9	80.3	74.8	64.84
	HRNet-48 [34]	OCRNet [29]	36.9	7.8	80.2	25.7	46.7	71.4	79.6	57.9	63.91
	ResNet-101 [46]	PointRend [30]	32.4	38.8	74.1	23.5	41.7	70.6	76.7	41.5	60.93
	STDC-1446 [31]	STDC [31]	30.1	0.0	70.9	16.1	32.3	62.7	80.0	20.3	50.94
	MIT-B5 [37]	SegFormer [37]	<u>37.2</u>	0.0	80.1	12.7	38.9	64.5	80.7	9.1	65.73
	ConvNeXt-v2-Large [43]	UPerNet [42]	28.2	44.6	71.7	42.5	41.0	73.6	80.4	34.3	59.72
	VMamba-Base [44]	UPerNet [42]	34.3	50.6	80.4	47.3	54.9	74.3	85.5	69.1	68.00
	Swin-Large-SkySense [13]	UPerNet [42]	31.6	<u>52.0</u>	73.2	<u>52.4</u>	<u>57.5</u>	<u>75.9</u>	81.7	77.8	<u>69.68</u>
Fusion-based methods	ResNet-50 [46]	GLNet [12]	23.7	27.3	61.0	4.7	26.0	47.4	77.5	0.0	44.73
	VGG-16 [48]	LCF-ALE [20]	24.0	42.3	48.2	25.3	28.4	42.0	71.2	28.7	48.28
	ResNet-50 [46]	MagNet [21]	20.2	0.0	63.1	0.0	23.1	44.4	78.8	8.7	44.20
	ResNet-18 [46]	ISDNet [22]	0.2	0.0	14.6	0.0	0.0	54.2	56.8	0.0	21.98
Our REST (holistic segmentation methods)	ConvNeXt-v2-Large [43]	UPerNet [42]	27.2	47.4	70.8	42.3	40.0	74.3	79.9	43.5	61.57
	VMamba-Base [44]	UPerNet [42]	34.0	47.4	<u>87.3</u>	45.5	51.6	75.2	<u>88.8</u>	<u>79.2</u>	69.41
	Swin-Large-SkySense [13]	UPerNet [42]	**41.5**	**55.6**	**90.3**	**54.5**	**59.4**	**77.7**	**90.5**	**80.9**	**72.95**

Bold value and <u>underlined</u> value indicate the best and the second-best performance, respectively.

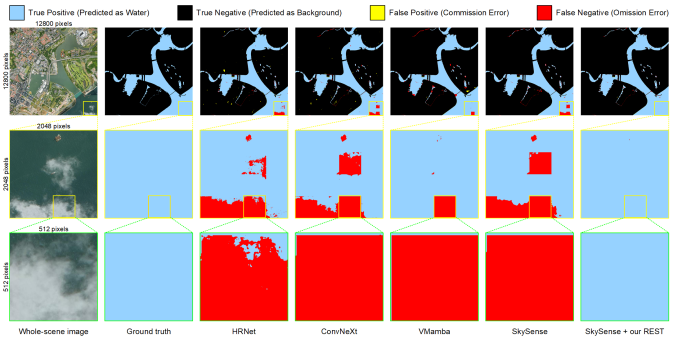


Fig. 3. **Visualization of segmentation results on GLH-Water dataset.** REST accurately extracts the complete water body although it is obscured by clouds, while the competing methods present omission errors in different locations. All models use the UPerNet as the decoder. Best viewed in color.

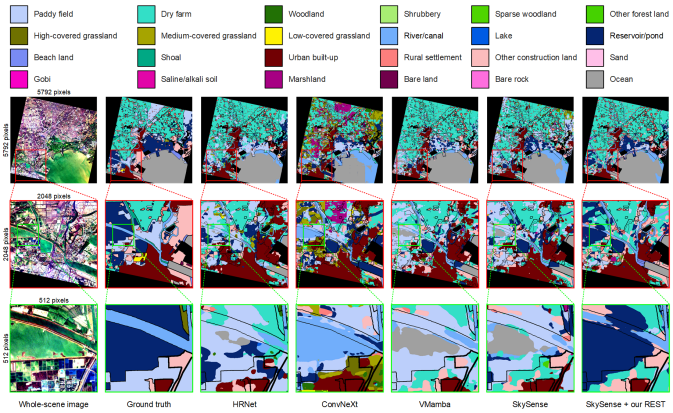


Fig. 4. **Visualization of segmentation results on WHU-OHS dataset.** Even on the challenging hyperspectral imagery datasets, REST demonstrates better segmentation results than the competing methods. All models use the UPerNet as the decoder. Best viewed in color.

segmentation [53], [56], and hyperspectral image segmentation [57], [58]. For fusion-based methods [12], [20], [21], [22], they aim to improve performance by combining local tile information with global context. Typically, these methods involve downsampling the WRI, leading to information degradation and the introduction of noise or artifacts. Furthermore, fusion-based methods require extensive data reading and writing operations, resulting in a longer training time (Table 3). To verify the effectiveness of REST, we compare its performance with these comparison methods across different semantic segmentation tasks. As shown in Table 1, REST shows improved performance across all tasks. In the following, we elaborate on the superiority of REST in various semantic segmentation tasks.

4.2 Results and Analysis

4.2.1 *Overall performance*

To analyze the performance of REST on single-class semantic segmentation task with optical satellite imagery, we conduct experiments on the GLH-Water [49] dataset. This dataset features large-size images and significant intra-class variations. It requires a significant amount of contextual information to accurately segment the water bodies within the WRI, which is a very good condition for validating REST. As displayed in Table 1, although cropping-based methods are limited to performing predictions on cropped image tiles, Swin-Large-SkySense, a recently proposed RSFM, achieves

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2025.3609767

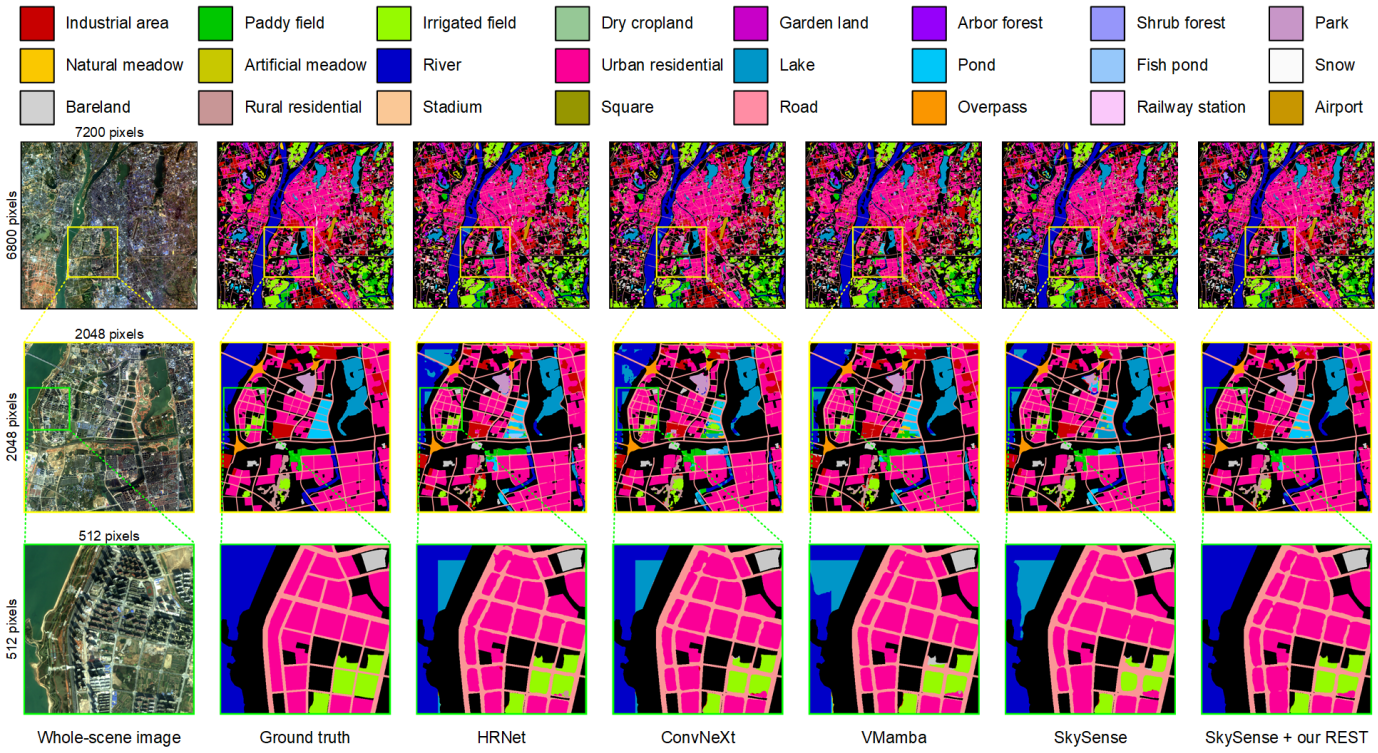


Fig. 5. **Segmentation results on Five-Billion-Pixels dataset.** REST successfully distinguishes between fine-grained categories (e.g., river, lake, pond) due to the enriched spatial context, while the results of other competing methods mostly show confusion among these categories. All models use the UPerNet as the decoder. Best viewed in color.

a remarkable IoU score of 86.64%, significantly outperforming all existing methods. The integration of REST further enhances Swin-Large-SkySense's capability by enabling HSW, resulting in a substantial increase of IoU score to 89.15%. Furthermore, REST demonstrates notable versatility by significantly boosting the performance of VMamba-Base and ConvNeXt-V2-Large. These improvements primarily originate from REST's holistic segmentation capability. Notably, fusion-based methods fail to converge to satisfactory performance despite extended training periods, due to their inherent limitation in parallel processing. By effectively perceiving and leveraging rich contextual information in the WRI, REST achieves precise segmentation of geographical elements while maintaining their spatial integrity, as demonstrated in Fig. 3 and Supplementary Figure S6. Besides, REST also demonstrates a good trade-off between accuracy and efficiency owing to its high degree of parallelism. As displayed in Fig. 10b and e, Fig. 11 and Table 4, REST delivers stable performance gains compared to SOTA baselines (Fig. 10g), incurring only a marginal computational cost (Table 3). In general, contrary to the intuitive assumption that WRI is superfluous for single-class semantic segmentation, our study demonstrates the fundamental necessity of HSW, as empirically confirmed through the experimental results.

To verify the effectiveness of REST on multi-class semantic segmentation of multispectral satellite imagery and optical unmanned aerial vehicle (UAV) imagery, we conduct experiments on the Five-Billion-Pixels [50] and the UAVid [51] datasets. In addition to their large-size attributes, these two datasets possess distinctive features. The Five-Billion-Pixels dataset comprises 24 fine-grained classes, presenting challenging inter-class similarities that cannot be resolved through analysis of local image tiles alone. It is essential for models to comprehensively utilize global contextual information and capture the relationships among geographical elements, which have been empirically demonstrated as distinctive advantages of REST. As illustrated in Table 2 and Fig. 5, the incorporation of REST substantially enhances the IoU score in most categories, especially for fine-grained categories with high inter-class similarity. Additionally, we verify the efficacy and versatility of REST on several other encoders and decoders, consistently demonstrating its capacity, as illustrated in Fig. 10d and Supplementary Figure S2. When considering the UAV imagery, its oblique viewing geometry poses distinct challenges for semantic segmentation, primarily manifested as substantial scale heterogeneity for the same object category depending on its distance from the sensor. As presented in Supplementary Table S16, there is consistent improvement across almost all categories by integrating REST. VMamba-Base achieves optimal performance on the UAVid dataset with REST, even outperforming some methods specifically tailored for UAV images [53], [56]. Visual comparisons in Fig. 6 demonstrate that REST substantially enhances the spatial integrity of segmented objects relative to competing methods. To sum up, experimental results demonstrate that REST can achieve the best performance on both satellite and UAV imagery, showing its versatility.

To explore the most arduous task of multi-class semantic

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2025.3609767

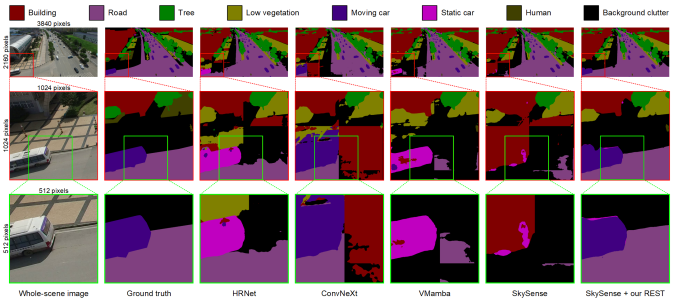


Fig. 6. **Visualization of segmentation results on UAVid dataset.** Compared with other methods, REST precisely identifies the vehicle in the image as a moving car instead of a static car, completely extracts the road, and significantly reduces misclassification problems. All models use the UPerNet as the decoder. Best viewed in color.

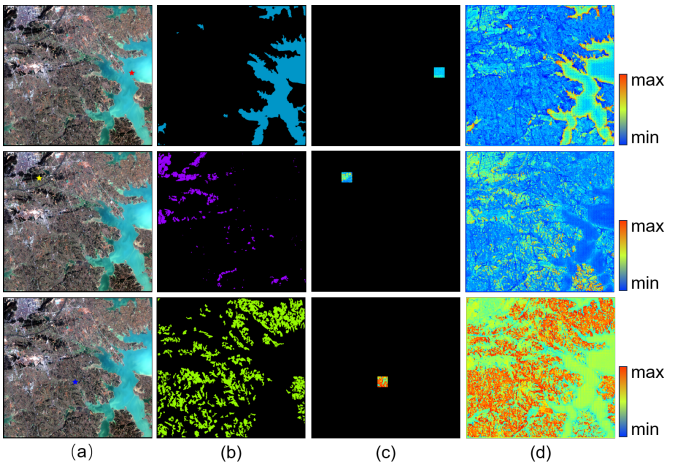


Fig. 7. **Visualization of feature maps on Five-Billion-Pixels dataset.** (a) Query pixels (star). (b) Ground truth maps of query category. (c) Attention maps of baseline. (d) Attention maps of baseline + our REST. Enhanced with REST, the model can exploit the features of the entire spatial region in the WRI. Swin-Large-SkySense with UPerNet is chosen as the baseline. Best viewed in color.

segmentation of hyperspectral satellite imagery, we conduct experiments on the WHU-OHS dataset. This dataset is characterized by two properties: large image size and large number of spectral channels (each image has a size around 6,272×6,272 pixels and 24 spectral channels). Simultaneously, it exhibits a severe class imbalance issue (Supplementary Figure S8) and demonstrates significant intra-class variability due to the diverse visual appearances present across WRIs. These factors render it difficult for models to learn robust feature representations, preventing both the comparison methods and our REST from attaining good performance. Notwithstanding these challenges, results point out that the utilization of REST still engenders improvement in overall accuracy. Supplementary Table S15 illustrates the category-level results on the WHU-OHS dataset. Certain categories appear in only a scant few scenes within the validation and testing sets, thereby heightening the difficulty of model learning and occasionally leading to zero accuracy for those specific categories. Even in such challenging circumstances, REST achieves noticeable improvement, demonstrating its effectiveness. To sum up, REST stands out as a superior solution across various tasks for HSW, encompassing diverse spectral bands, different sensor types, and various category granularities.

4.2.2 *Global perception and robust representation ability*

From an explainability perspective, we conduct a focused analysis of REST's global perception and representation ability, which are intrinsically coupled characteristics that interact synergistically. Global perception refers to the ability of REST to focus both on local details and the entire scene, including the structures, arrangements, and interactions of different geographical elements. This involves aggregating information from different regions of WRI to form a comprehensive understanding, enabling the identification of global patterns and dependencies that might not be apparent when looking at individual components in isolation. As for the representation ability, it denotes the capacity to represent and convey meaningful information about the WRI in the form of features, involving learning complex non-linear relationships in the image, highlighting important characteristics, and suppressing noise or redundant information. In the following, we will discuss these two capabilities separately, as well as the effects of their coupled interaction.

To investigate the global perception ability, we conduct experimental analysis across multiple tasks. By facilitating communication and information exchange among multiple GPUs, REST enables cross-GPU global attention calculation. REST enables models to perceive global information and mine long-distance spatial context. As presented in Fig. 3 and Supplementary Figure S6, the global perception of REST ensures precise segmentation in single-class semantic segmentation. Utilizing the long-distance context, REST can preserve the complete structure of geographical elements, offering a more reliable segmentation result with fewer errors in both false positives and false negatives. Besides, the results on multi-class semantic segmentation further demonstrate REST's global perception capability. As depicted in Fig. 5, distinguishing between river and lake within the fine-grained categories of water bodies poses significant challenges for existing methods, primarily due to limited receptive fields. These methods often misclassify the water bodies within a given tile, leading to serious block effects in the final result. In contrast, the results of REST show a substantial alleviation in misclassification. Leveraging global context, REST correctly segments rivers, lakes, and ponds, highlighting its effectiveness. One can see that the global perception capabilities of REST endow the model with a stronger fine-grained discrimination ability. Fig. 4 provides similar visualized results on the hyperspectral WHU-OHS dataset. Compared to other methods, REST maximally distinguishes the fine-grained category of reservoir/pond. Moreover, visualization results also confirm the effectiveness of REST on the UAVid dataset, where the oblique view of UAV imagery brings additional challenges to segmentation. As illustrated in Fig. 6, comparison methods exhibit various issues such as block effects, commission errors, and omission errors. In contrast, REST effectively distinguishes cars and accurately extracts roads. Additional visualization results for the Five-Billion-Pixels

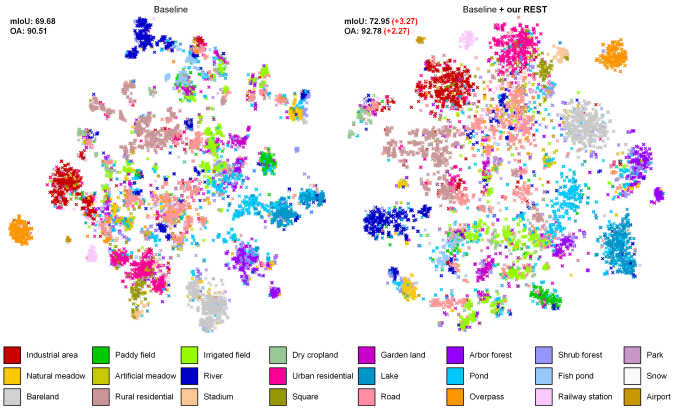


Fig. 8. **Visualization of t-SNE results on Five-Billion-Pixels dataset.** Swin-Large-SkySense with UPerNet is chosen as the baseline. When combined with REST, the features exhibit more distinct classification boundaries, demonstrating the robust feature representation ability brought by REST. Best viewed in color.

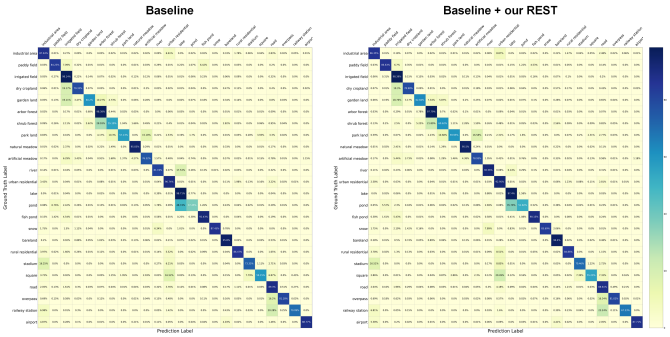


Fig. 9. **The confusion matrices of results on the Five-Billion-Pixels dataset.** (a) The confusion matrix of the baseline. (b) The confusion matrix of baseline integrating our REST. Swin-Large-SkySense with UPerNet is chosen as the baseline. After the introduction of REST, the accuracy performance across various categories improves, and the confusion between fine-grained categories decreases. Best viewed in color.

dataset, WHU-OHS dataset, and UAVid dataset are provided in Supplementary Figure S7, S8, and S9, respectively.

To analyze the representational ability of REST, we perform experiments on the Five-Billion-Pixels dataset and visualize the feature distribution. As exemplified in Fig. 7, the first column shows the original input WRI, where the red, yellow, and blue stars represent the query pixels. The second column displays regions of the same category as the query pixel in the ground truth maps. The third column illustrates the perceivable context information using the baseline method, while the last column shows the perception area with REST, capturing the whole-scene context. Besides, the response values of feature maps demonstrate that areas with high values align closely with the ground truth maps. This alignment reflects that REST adaptively perceives meaningful information for the query pixel. Additionally, Fig. 8 reflects the features extracted on the Five-Billion-Pixels dataset. By comparing the two figures on the left and right, one can see that the feature points in the right figure are more compactly clustered, and the separation between different categories is more distinct. This indicates that REST helps to improve the model's ability to distinguish features from different categories, thereby enhancing the segmentation performance. Overall, Fig. 7 and Fig. 8 illustrate the robust representational capabilities of REST.

The combination of global perception and robust representation ability in REST helps to achieve discrimination between fine-grained categories, improving performance while reducing confusion. To support this, we visualize the confusion matrices in Fig. 9. The baseline is on the left and the baseline with REST is on the right. In the matrix, the color transition from blue to red indicates an increase in values from small to large. As one can observe from the figure, after incorporating REST, the prediction accuracies significantly improve for the vast majority of categories. Specifically, compared with the baseline, REST achieves a remarkable increase of 17% to 20% in true positives for categories such as garden land, artificial meadow, river, and dry cropland. For categories like industrial area, park land, pond, fish pond, stadium, overpass, and railway station, there is a relatively substantial increase between 6% and 10%. The adoption of REST is of great assistance in discriminating fine-grained categories that require rich contextual information. When compared with the baseline, REST shows a slight decrease in categories such as bareland, rural residential, and airport. This decrease might be attributed to parameter optimization deviation and class imbalance. Yet this degree of accuracy reduction is almost negligible when considering the overall accuracy improvement. The Five-Billion-Pixels dataset demonstrates typical challenges in fine-grained segmentation. Certain inter-class similarities among categories lead to confusion. For example, such confusion arises among dry cropland, irrigated field, and garden land; among river, lake, pond, and fish pond; among square, road, overpass, and railway station; as well as among industrial area, urban residential, rural residential, stadium, square, and road. Following the utilization of REST, the degree of confusion among these fine-grained categories is markedly diminished, indicating a notable enhancement in the model's fine-grained discrimination ability and overall performance.

4.2.3 *Near-linear throughput-scalability with expansion of GPUs*

Theoretically, REST achieves near-linear throughput-scalability in handling WRI by distributing the memory load across multiple GPUs. Here, we define throughput-scalability as the ability of REST to scale up the throughput and handle WRI with a larger image size as the number of GPUs expands. During the HSW, REST divides the WRI into non-overlapping spatial regions based on the number of available GPUs, with each region processed on a separate GPU, ensuring balanced memory utilization. Thus, REST scales up the handleable image size near-linearly along with expansion of GPUs, as shown in Fig. 11b. To validate this theoretical analysis, we detail the experimental validation of REST's near-linear throughput-scalability. In our implementation, when using 16 NVIDIA A100 GPUs with 40 GB memory, REST can support HSW with an image size of 12,800×12,800 during testing and handle images with a size of 4,096×4,096 during training (with the batch size set to 8).

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2025.3609767

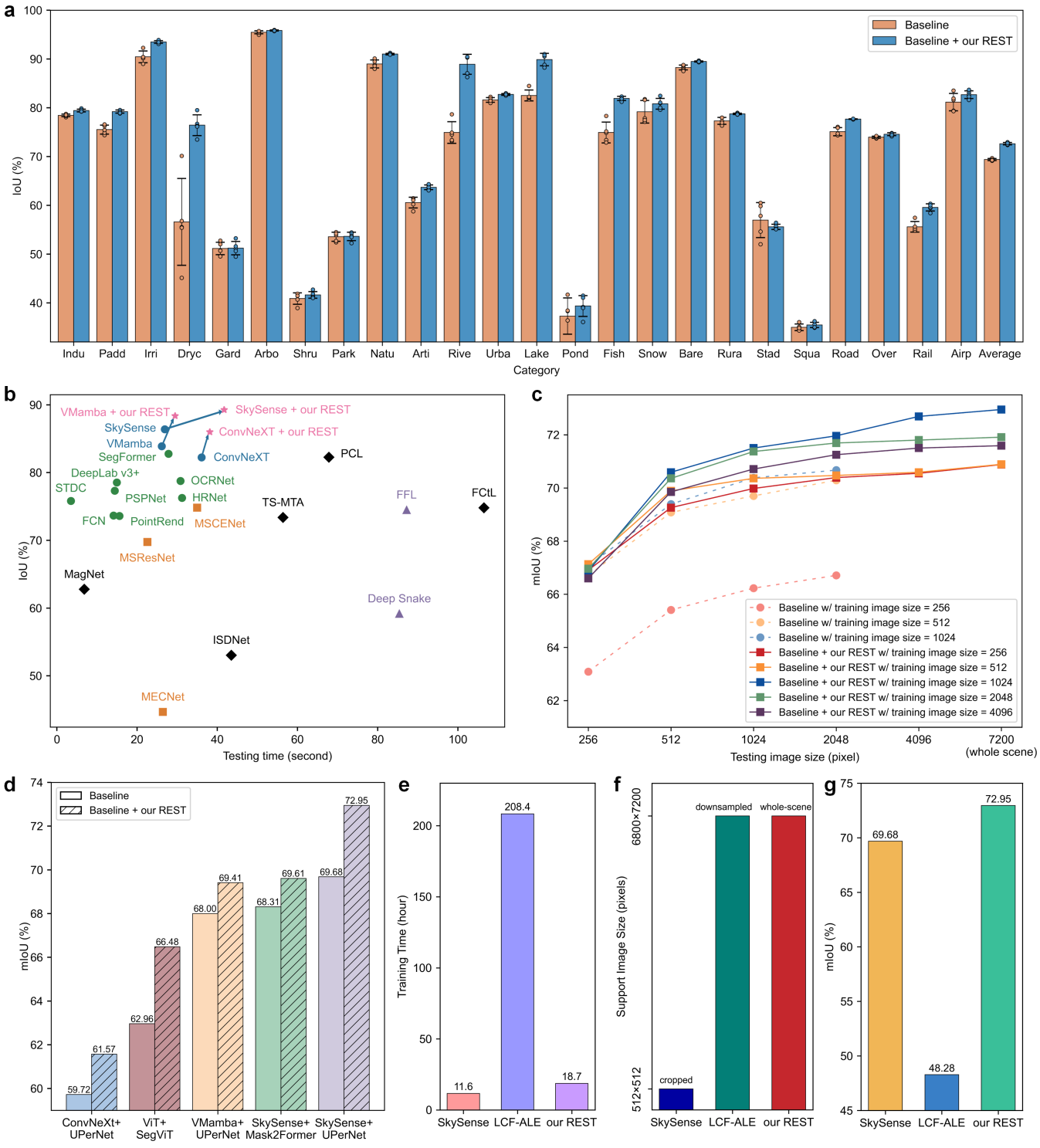


Fig. 10. **Comprehensive performance of REST.** (a) Per-class result comparison on the Five-Billion-Pixels dataset is shown, and Swin-Large-SkySense with UPerNet is chosen as the baseline. The bar chart shows the average performance over five repeated experiments, and the dots indicate the individual performance of each experiment. The error bars indicate the interval spanning one standard deviation about the mean ($mean \pm std$), derived from five independent experimental repetitions. (b) Comparison of performance (IoU) and efficiency (testing time) of different methods on the GLH-Water dataset. (c) Performance of REST with different baselines (i.e., encoders and decoders) on the Five-Billion-Pixels dataset. (d) Performance of REST on the Five-Billion-Pixels dataset with different image sizes using Swin-Large-SkySense with UPerNet as the baseline. Here we adopt the notation size = Y as a concise representation for size = $Y \times Y$ to simplify size descriptions. (e-g) Comparison of different methods in terms of training time, supported image size, and mIoU. SkySense and LCF-ALE represent the top-performing cropping-based and fusion-based methods, respectively. As shown, SkySense struggles with large-size WRI and exhibits limited accuracy. LCF-ALE can only leverage degraded (i.e., downsampled) information, resulting in suboptimal performance in both efficiency and accuracy. Specifically, our REST effectively utilizes whole-scene context, achieving superior accuracy while maintaining acceptable training time. Best viewed in color.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2025.3609767

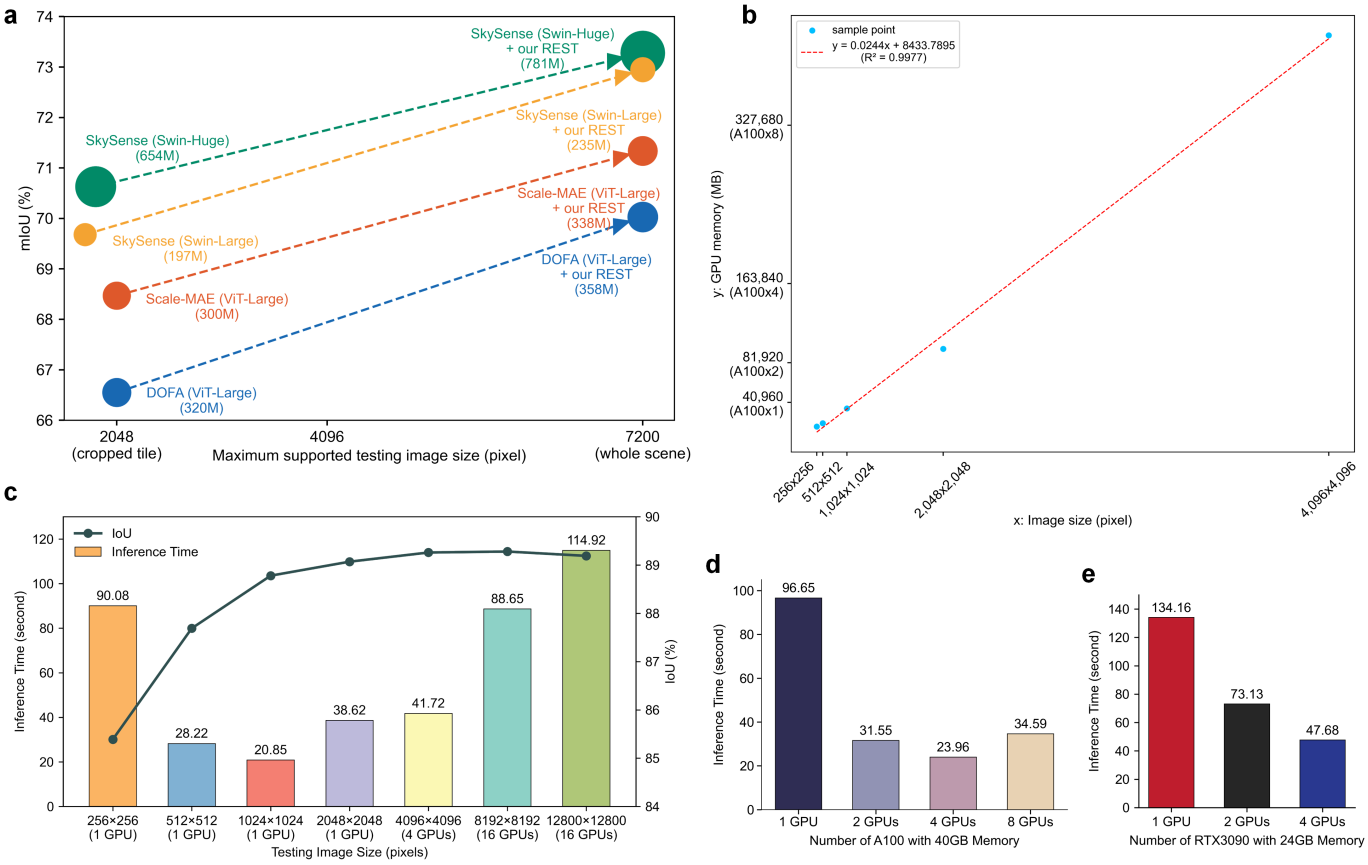


Fig. 11. **Efficiency analysis of REST.** (a) REST further improves the strong capabilities of various RSFMs. (b) The GPU memory consumption of REST with different image sizes during training. (c) IoU score improvement and inference time variation across different testing image sizes, with diminishing returns and communication bottlenecks. (d) Inference time reduction with 1–4 NVIDIA A100 GPUs and increase with 8 GPUs due to inter-node communication at 2,048 testing image size. (e) Inference time trends with NVIDIA RTX3090 GPUs at 2,048 testing image size.

As expected, one can observe that the training stage requires more GPUs than the testing stage when handling images of the same size due to the requirement to store intermediate features and gradients. It is noted that, whether in training or testing, REST can achieve HSW when a sufficient number of GPUs are available.

To analyze the throughput-scalability, we conduct a series of experiments on the Five-Billion-Pixels dataset, with Swin-Large-SkySense with UPerNet as the baseline. It should be noted that due to GPU memory limitations, the baseline methods are limited to supporting a maximum training image size of 1,024×1,024 pixels. In contrast, REST features throughput-scalability, enabling progressive expansion of supported training image sizes with an increasing number of GPUs. However, the trend of segmentation performance varying with image size is not entirely consistent with our expectations. The investigation, as shown in Fig. 10c, elucidates three critical phenomena through empirical analysis. First, REST demonstrates consistent performance enhancement across all evaluated training image sizes, exemplified by its 2.28% mIoU improvement over the baseline at the training image size of 1,024×1,024, validating its capacity to enhance feature representation. Second, our experimental analysis reveals distinct stage-dependent size-accuracy relationships. On the one hand, testing-stage evaluations exhibit a 3.60% mIoU gain for the baseline at a fixed training image size of 1,024×1,024 when testing image sizes increase, pointing out the necessity of holistic segmentation during testing. On the other hand, training-stage experiments demonstrate a 3.96% mIoU improvement as training image sizes expand from 256×256 to 1,024×1,024, confirming that enhanced feature learning capacity at larger training image sizes. Third, REST's scalability boundaries emerge at extreme training image sizes, which is a limitation stemming from the dual mechanisms of expanded spatial context perception and introduced information redundancy in the training stage. Although maintaining an advantage over the baseline performance, further expansion to 4,096×4,096 induces a 1.36% mIoU degradation, This phenomenon exhibits stage-specific manifestations: large-size training inputs propagate noise prior to feature stabilization. However, the trained models strategically leverage expanded contexts through attention mechanisms in the testing stage, evidenced by the mIoU gain when testing image sizes increase under all training conditions. Consequently, the empirically validated optimal configuration combines a training image size of 1,024×1,024 with a testing image size of 6,800×7,200.

To analyze from the perspective of the model parameters, we conduct comprehensive experiments across all four

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2025.3609767

TABLE 3
Comparison of performance and efficiency on the Five-Billion-Pixels dataset. Training time represents the duration required for the model to train until convergence, while testing time denotes the time needed to complete the inference for a WRI.

Type	Encoder	Decoder	$mIoU(\%)$	$OA(\%)$	Training time (hour)	Testing time (second)
Cropping-based methods	ResNet-101 [46]	PSPNet [32]	64.48	89.02	19.6	6.8
	ResNet-101 [46]	DeepLab-v3+ [33]	64.01	89.00	17.2	7.1
	HRNet-48 [34]	FCN [47]	64.84	88.95	10.2	12.5
	HRNet-48 [34]	OCRNet [29]	63.91	89.33	15.0	9.1
	ResNet-101 [46]	PointRend [30]	60.93	88.16	7.0	5.8
	STDC-1446 [31]	STDC [31]	50.94	85.70	8.0	5.1
	MIT-B5 [37]	SegFormer [37]	65.73	89.30	14.1	8.6
	ConvNeXt-v2-Large [43]	UPerNet [42]	59.72	88.13	16.3	12.6
	VMamba-Base [44]	UPerNet [42]	68.00	90.52	12.8	9.5
	Swin-Large-SkySense [13]	UPerNet [42]	69.68	90.51	11.6	12.4
Fusion-based methods	ResNet-50 [46]	GLNet [12]	44.73	-	336.2	207.1
	VGG-16 [48]	LCF-ALE [20]	48.28	-	208.4	91.3
	ResNet-50 [46]	MagNet [21]	44.20	-	39.5	119.7
	ResNet-18 [46]	ISDNet [22]	21.98	-	10.3	15.8
Our REST (holistic segmentation methods)	ConvNeXt-v2-Large [43]	UPerNet [42]	61.57	88.50	33.4	13.5
	VMamba-Base [44]	UPerNet [42]	69.41	91.96	16.6	12.5
	Swin-Large-SkySense [13]	UPerNet [42]	72.95	92.78	18.7	14.0

WRI datasets. Since the WRI is typically replete with a variety of details, in order to precisely represent such intricate information, the model should be equipped with a large number of parameters to capture features across different scales and positions. As the input image size increases, the complexity of the information encapsulated within the WRI increases significantly. Theoretically, the model demands more parameters to process this information effectively. For instance, consider a scenario where the model can learn effectively with small-size images using N parameters. As the input size expands, the model might need $4N$, $16N$, or an even greater quantity of parameters to extract the information and achieve an accuracy equivalent to or surpassing that of the scenario with a small input size. However, if the model parameter remains static at N, the model may face a parameter shortage when confronted with large-size images. This can impede the model from fully leveraging all the information in the input image, making it difficult to capture complex features and relationships, and even resulting in a decline in performance. We surmise that an optimal input image size for specific tasks and particular categories likely exists. Supplementary Figure S3 presents more results to support this.

4.2.4 Efficiency Analysis

The proposed REST framework significantly enhances the efficiency of RSFMs by enabling holistic processing of large-size imagery while maintaining manageable computational demands. As shown in Fig. 11a, unlike traditional RSFMs constrained to processing cropped 2,048×2,048 tiles, REST allows these models to handle entire remote sensing scenes with improved performance, despite introducing only minimal parameter increases. Training efficiency analysis in Fig. 11b reveals REST's near-linear GPU memory consumption growth with increasing image sizes on the Five-Billion-Pixels dataset, as shown by the blue data points and their red fitted curve. This predictable memory pattern directly supports the observed sublinear throughput scalability when expanding GPU resources.

TABLE 4
Comparison of baseline methods and our proposed REST on the GLH-Water dataset. UperNet is utilized as the decoder for all methods.

Method	Params	FLOPs	IoU (%)	F1 (%)
ConvNeXt	233	1570	82.24	90.25
ConvNeXt + our REST	271 (+38)	1686 (+116)	86.46 (+4.22)	92.74 (+2.49)
VMamba	121	1161	83.12	90.78
VMamba + our REST	138 (+17)	1212 (+51)	89.08 (+5.96)	94.09 (+3.31)
SkySense	234	1771	86.64	92.84
SkySense + our REST	272 (+38)	1888 (+117)	89.15 (+2.51)	94.26 (+1.42)

When evaluating inference efficiency (Fig. 11c), three clear operational patterns emerge. First, for smaller inputs ranging from 256×256 to 1,024×1,024, parallel processing accelerates inference with underutilized GPUs. Second, at moderate sizes between 1,024×1,024 and 2,048×2,048, GPU utilization approaches saturation, forcing the framework to offload computations to CPU and host memory, which increases latency. Third, with ultra-large images exceeding 4,096×4,096, multi-GPU deployments maintain per-GPU loads equivalent to single-GPU processing of 2,048×2,048 tiles. However, this scaling introduces critical

communication bottlenecks, particularly from 8,192×8,192 to 12,800×12,800, where inter-node coordination increases inference time by approximately 25% compared to ideal expectations. These limitations persist in fixed-size experiments with 2,048×2,048 inputs (Fig. 11d): scaling from 1 to 4 GPUs reduces per-GPU computational load, but the accompanying communication overhead diminishes potential speed improvements. The pattern further manifests in 8-GPU configurations where cross-node communication replicates the latency issues observed in variable-size testing.

Notably, the framework's behavior remains consistent across hardware environments, as confirmed by single-node RTX3090 experiments (Fig. 11e) that exhibit similar diminishing returns in multi-GPU scaling. These systematic observations collectively validate REST's ability to process large-size imagery while exposing a fundamental constraint: current communication architectures, especially inter-node links, become the dominant bottleneck when distributing ultra-large remote sensing workloads. The results suggest that advanced networking solutions such as InfiniBand, though unavailable in the current test environment, could potentially mitigate these limitations.

Table 4 provides a comprehensive comparison of the baseline segmentation models and their REST-enhanced counterparts on the GLH-Water dataset, using UperNet as the decoder for all methods. We report both the total number of parameters and the floating-point operations (FLOPs), alongside the resulting segmentation accuracy in terms of IoU and F1 score. Integrating REST consistently boosts accuracy by up to 5.96% IoU and 3.31% F1, while only incurring a modest increase in computational cost. Notably, SPIM's own FLOPs and parameter count remain invariant to the number of GPUs employed. To further enhance efficiency and versatility, we introduce a selective KV compression mechanism to substantially reduce communication and computational overhead and develop lightweight variants to ensure scalability in resource-constrained scenarios. See Supplementary Section 3 for details.

4.2.5 Ablation Study and Parameter Analysis

We conduct a comprehensive ablation study to evaluate the individual contributions and sensitivity of key components in REST. The analysis focuses on the effects of spatial partitioning, inter-region communication, and attention configuration. Through controlled comparisons, we quantify the trade-offs between model accuracy and computational cost, and identify design choices that most effectively enhance segmentation performance.

In the ablation study (Table 5), the spatial partitioning baseline without attention or communication achieves 69.08% mIoU in 73 minutes. Adding only attention raises mIoU slightly to 69.18% but increases training to 114 minutes. Enabling both attention and communication in full SPIM pushes mIoU to 69.51% at 213 minutes, demonstrating a clear precision benefit for the added module.

The communication sensitivity results (Table 6) show that full four-stage exchanges yield the highest mIoU (71.50%) at 24.51 seconds per forward pass. Disabling stage 0 reduces accuracy by just 0.07% while saving around 7.6 seconds, and disabling later stages incurs less than 0.22% drops with modest speedups. Removing early-stage communication entirely drops mIoU to 70.99% but cuts about 9 seconds, highlighting a tunable speed–accuracy trade-off.

TABLE 5
Ablation study of REST on the Five-Billion-Pixels dataset. The table reports the resulting mIoU and training time.

Method	Attn	Comm	mIoU (%)	Time (min)
Spatial partition baseline			69.08	**73**
+ attention parameters	✓		69.18	114
our REST (full SPIM)	✓	✓	**69.51**	213

TABLE 6
Communication frequency sensitivity analysis on the Five-Billion-Pixels dataset. The table reports the resulting mIoU and inference time impact of disabling communication modules at each stage (✓: *enabled*, '–': *disabled*).

Stage 0	Stage 1	Stage 2	Stage 3	mIoU (%)	Time (s)
✓	✓	✓	✓	**71.50**	24.51
–	✓	✓	✓	71.43	16.90
✓	–	✓	✓	71.46	24.02
✓	✓	–	✓	71.28	24.26
✓	✓	✓	–	71.46	24.26
–	–	✓	✓	71.39	16.45
–	–	✓	–	71.36	15.86
–	–	–	✓	71.15	15.55
–	–	–	–	70.99	15.39

TABLE 7
Parameter sensitivity study of attention heads on the Five-Billion-Pixels dataset. The table reports the resulting mIoU. Rows and columns correspond to the number of heads used during training and testing.

Train/Test	4	8	12	16
4	70.93	70.83	70.79	70.76
8	71.27	71.21	71.19	71.18
12	**71.50**	71.47	71.44	71.38
16	70.91	70.92	70.86	70.84

Table 7 shows that increasing training heads from four to twelve raises mIoU, reaching 71.50% when trained with twelve heads and tested with four, while using sixteen heads offers no further gain. Under any fixed training setting, raising the number of heads at test time leads to lower accuracy. For example, training with twelve heads and testing with eight, twelve or sixteen heads yields mIoU values of 71.47%, 71.44% and 71.38% respectively.

In addition to the core component ablations, we further validate the versatility and robustness of our framework in the Supplementary Section 3. Specifically, we integrate a scale-adaptive attention module to enhance performance on small-scale objects, demonstrate the synergy of our method with various class-balancing techniques for handling rare categories, and confirm its resilience against input and label noise through dedicated noise interference experiments. We also incorporate an edge alignment loss to improve stitching consistency at multi-GPU boundaries with negligible computational impact. These comprehensive analyses affirm that REST is a robust and adaptable framework for tackling diverse real-world challenges.

5 Discussion

REST enhances the emerging remote sensing foundation models. RSFMs [13], [14], [15], [16], [17], [18] have emerged as a transformative solution to address the challenges of remote sensing interpretation. By significantly scaling up model parameters, RSFMs demonstrate superior feature representation and generalization capabilities compared to traditional deep learning-based methods. While their substantially larger parameter sizes endow RSFMs with considerable potential for handling HSW, this potential remains underutilized due to their substantial GPU memory requirements, which currently restrict them to processing cropped image tiles at reduced size. Notably, REST effectively unlocks the full potential of existing RSFMs by scaling them up to support a larger image size, as evidenced in Fig. 11a. The achievement of HSW through REST requires between 4 to 16 NVIDIA A100 GPUs with 40 GB memory, depending on the specific RSFM's parameter size. Experimental results confirm that REST can be seamlessly integrated with RSFMs, and their synergistic combination significantly enhances HSW performance. Additional technical analysis and implementation details are provided in Supplementary Section 2.

REST offers advantages of versatility and minimal overhead. The proposed REST framework demonstrates remarkable versatility in supporting diverse data types and model architectures, as evidenced by comprehensive results in Table 1, Fig. 10d, and Supplementary Figure S2. It significantly enhances both overall and fine-grained category accuracy across various encoder-decoder combinations, including Swin, VMamba, ConvNeXt, and ViT encoders paired with UPerNet, Mask2Former, and SegViT decoders, underscoring its wide applicability and effectiveness. At its core, REST leverages SPIM, an architecture-agnostic mechanism that seamlessly integrates with mainstream models in a plug-and-play manner, ensuring adaptability to emerging high-performance models for HSW. Unlike existing fusion-based methods, which are constrained by their inability to perform HSW and suffer from excessive computational redundancy, REST introduces minimal overhead, requiring only 5% to 10% additional computational resources and time to enable HSW while delivering consistent accuracy improvements, as quantitatively demonstrated in Fig. 10b, Table 3, and Table 4.

REST shows promising potential for applications in the medical field. Beyond RSIs, the segmentation of large-size medical images remains a formidable challenge that has garnered extensive research interest [23], [24], [25], [39], [40]. While most existing methods target whole-slide image classification, their direct translation to pixel-level segmentation is often inadequate. To assess REST's generality, we applied it to three medical benchmarks, including the ISIC dermoscopic dataset [59], the CRAG histopathology dataset [60], and the Synapse multi-organ CT dataset. REST consistently outperforms classic convolutional and transformer-based baselines in each domain, achieving state-of-the-art accuracy while robustly handling diverse scales and complex anatomical patterns. These results underscore REST's adaptability across vastly different medical imaging tasks. See Supplementary Section 4 for details and further analysis.

6 Conclusion

The REST architecture marks a significant step forward in HSW. By employing the proposed SPIM, REST addresses GPU memory constraints, enabling precise and efficient whole-scene segmentation. Its plug-and-play compatibility with various encoders and decoders, including foundation models, ensures versatility across diverse segmentation tasks. REST consistently outperforms traditional cropping and fusion-based methods, establishing itself as a robust solution for HSW. However, REST's reliance on multiple GPUs may restrict its use in resource-limited settings, and SPIM's all-to-all communication could introduce overhead in real-time scenarios. Future research could focus on optimizing SPIM for reduced communication costs, like exploring hybrid parallelization. Additionally, integrating lightweight models could enable deployment on edge devices, enhancing practical applicability. In general, REST lays a solid foundation for end-to-end semantic segmentation for WRIs.

Acknowledgment

Numerical calculations in this paper have been done on the GPUs of the Supercomputing Center of Wuhan University.

References

[1] P. M. Vitousek, H. A. Mooney, J. Lubchenco, and J. M. Melillo, "Human domination of earth's ecosystems," *Science*, vol. 277, pp. 494–499, 1997.

[2] V. Zalles, M. C. Hansen, P. V. Potapov, D. Parker, S. V. Stehman, A. H. Pickens, L. L. Parente, L. G. Ferreira, X.-P. Song, A. Hernandez-Serna, and I. Kommareddy, "Rapid expansion of human impact on natural land in south america since 1985," *Science Advances*, vol. 7, p. eabg1620, 2021.

[3] W. Zhang, G. Villarini, G. A. Vecchi, and J. A. Smith, "Urbanization exacerbated the rainfall and flooding caused by hurricane harvey in houston," *Nature*, vol. 563, pp. 384–388, 2018.

[4] T. P. Roland, O. T. Bartlett, D. J. Charman, K. Anderson, D. A. Hodgson, M. J. Amesbury, I. Maclean, P. T. Fretwell, and A. Fleming, "Sustained greening of the antarctic peninsula observed from satellites," *Nature Geoscience*, 2024.

[5] K. Wu, Y. Zhang, L. Ru, B. Dang, J. Lao, L. Yu, J. Luo, Z. Zhu, Y. Sun, J. Zhang *et al.*, "A semantic-enhanced multi-modal remote sensing foundation model for earth observation," *Nature Machine Intelligence*, pp. 1–15, 2025.

[6] Y. Li, L. Wang, T. Wang, X. Yang, J. Luo, Q. Wang, Y. Deng, W. Wang, X. Sun, H. Li *et al.*, "Star: A first-ever dataset and a large-scale benchmark for scene graph generation in large-size satellite imagery," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 47, no. 3, pp. 1832–1849, 2025.

[7] N. Casagli, E. Intrieri, V. Tofani, G. Gigli, and F. Raspini, "Landslide detection, monitoring and prediction with remote-sensing techniques," *Nature Reviews Earth & Environment*, vol. 4, pp. 51–64, 2023.

[8] Y. Li, Y. Wu, G. Cheng, C. Tao, B. Dang, Y. Wang, J. Zhang, C. Zhang, Y. Liu, X. Tang *et al.*, "Meet: A million-scale dataset for fine-grained geospatial scene classification with zoom-free remote sensing imagery," *IEEE/CAA Journal of Automatica Sinica*, vol. 12, no. 5, pp. 1004–1023, 2025.

[9] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat, "Deep learning and process understanding for data-driven earth system science," *Nature*, vol. 566, p. 195 – 204, 2019.

[10] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, p. 8 – 36, 2017.

[11] Y. Li, W. Chen, Y. Zhang, C. Tao, R. Xiao, and Y. Tan, "Accurate cloud detection in high-resolution remote sensing imagery by weakly supervised deep learning," *Remote Sensing of Environment*, vol. 250, p. 112045, 2020.

[12] W. Chen, Z. Jiang, Z. Wang, K. Cui, and X. Qian, "Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8924–8933.

[13] X. Guo, J. Lao, B. Dang, Y. Zhang, L. Yu, L. Ru, L. Zhong, Z. Huang, K. Wu, D. Hu, H. He, J. Wang, J. Chen, M. Yang, Y. Zhang, and Y. Li, "Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 662–27 673.

[14] F. Bastani, P. Wolters, R. Gupta, J. Ferdinando, and A. Kembhavi, "Satlaspretrain: A large-scale dataset for remote sensing image understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 726–16 736.

[15] C. J. Reed, R. Gupta, S. Li, S. Brockman, C. Funk, B. Clipp, K. Keutzer, S. Candido, M. Uyttendaele, and T. Darrell, "Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4065–4076.

[16] X. Sun, P. Wang, W. Lu, Z. Zhu, X. Lu, Q. He, J. Li, X. Rong, Z. Yang, H. Chang, Q. He, G. Yang, R. Wang, J. Lu, and K. Fu, "Ringmo: A remote sensing foundation model with masked image modeling," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–22, 2023.

[17] M. Mendieta, B. Han, X. Shi, Y. Zhu, and C. Chen, "Towards geospatial foundation models via continual pretraining," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2023, pp. 16 806–16 816.

[18] M. Noman, M. Naseer, H. Cholakkal, R. M. Anwar, S. Khan, and F. S. Khan, "Rethinking transformers pre-training for multi-spectral satellite imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 811–27 819.

[19] Y. Li, W. Chen, X. Huang, Z. Gao, S. Li, T. He, and Y. Zhang, "Mfvnet: a deep adaptive fusion network with multiple field-of-views for remote sensing image semantic segmentation," *Science China Information Sciences*, vol. 66, p. 140305, 2023.

[20] W. Liu, Q. Li, X. Lin, W. Yang, S. He, and Y. Yu, "Ultra-high resolution image segmentation via locality-aware context fusion and alternating local enhancement," *International Journal of Computer Vision*, vol. 132, pp. 5030–5047, 2024.

[21] C. Huynh, A. T. Tran, K. Luu, and M. Hoai, "Progressive semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 755–16 764.

[22] S. Guo, L. Liu, Z. Gan, Y. Wang, W. Zhang, C. Wang, G. Jiang, W. Zhang, R. Yi, L. Ma, and K. Xu, "Isdnet: Integrating shallow and deep networks for efficient ultra-high resolution segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4361–4370.

[23] H. Pinckaers, B. van Ginneken, and G. Litjens, "Streaming convolutional neural networks for end-to-end learning with multi-megapixel images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, pp. 1581–1590, 2022.

[24] H. Xu, N. Usuyama, J. Bagga, S. Zhang, R. Rao, T. Naumann, C. Wong, Z. Gero, J. González, Y. Gu, Y. Xu, M. Wei, W. Wang, S. Ma, F. Wei, J. Yang, C. Li, J. Gao, J. Rosemon, T. Bower, S. Lee, R. Weerasinghe, B. J. Wright, A. Robicsek, B. Piening, C. Bifulco, S. Wang, and H. Poon, "A whole-slide foundation model for digital pathology from real-world data," *Nature*, vol. 630, pp. 181–188, 2024.

[25] C.-L. Chen, C.-C. Chen, W.-H. Yu, S.-H. Chen, Y.-C. Chang, T.-I. Hsu, M. Hsiao, C.-Y. Yeh, and C.-Y. Chen, "An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning," *Nature Communications*, vol. 12, p. 1193, 2021.

[26] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. Le, and R. Salakhutdinov, "Transformer-XL: Attentive language models beyond a fixed-length context," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2978–2988.

[27] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[28] S. A. Jacobs, M. Tanaka, C. Zhang, M. Zhang, R. Y. Aminadabi, S. L. Song, S. Rajbhandari, and Y. He, "System optimizations for enabling training of extreme long sequence transformer models," in *Proceedings of the 43rd ACM Symposium on Principles of Distributed Computing*, 2024, p. 121–130.

[29] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 173–190.

[30] A. Kirillov, Y. Wu, K. He, and R. Girshick, "Pointrend: Image segmentation as rendering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9799–9808.

[31] M. Fan, S. Lai, J. Huang, X. Wei, Z. Chai, J. Luo, and X. Wei, "Rethinking bisenet for real-time semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9716–9725.

[32] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2881–2890.

[33] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 801–818.

[34] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 3349–3364, 2020.

[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of the International Conference on Learning Representations*, 2021.

[36] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2021, pp. 10 012–10 022.

[37] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 12 077–12 090.

[38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[39] S.-C. Huang, C.-C. Chen, J. Lan, T.-Y. Hsieh, H.-C. Chuang, M.-Y. Chien, T.-S. Ou, K.-H. Chen, R.-C. Wu, Y.-J. Liu, C.-T. Cheng, Y.-J. Huang, L.-W. Tao, A.-F. Hwu, I.-C. Lin, S.-H. Hung, C.-Y. Yeh, and T.-C. Chen, "Deep neural network trained on gigapixel images improves lymph node metastasis detection in clinical settings," *Nature Communications*, vol. 13, p. 3347, 2022.

[40] O. S. M. El Nahhas, M. van Treeck, G. Wölflein, M. Unger, M. Ligero, T. Lenz, S. J. Wagner, K. J. Hewitt, F. Khader, S. Foersch, D. Truhn, and J. N. Kather, "From whole-slide image to biomarker prediction: end-to-end weakly supervised deep learning in computational pathology," *Nature Protocols*, vol. 20, pp. 293–316, 2025.

[41] Q. Xu and Y. You, "An efficient 2d method for training super-large deep learning models," in *Proceedings of the IEEE International Parallel and Distributed Processing Symposium*, 2023, pp. 222–232.

[42] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 418–434.

[43] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "Convnext v2: Co-designing and scaling convnets with masked autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 133–16 142.

[44] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, and Y. Liu, "Vmamba: Visual state space model," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 103 031–103 063.

[45] S. Li, Y. Zhao, R. Varma, O. Salpekar, P. Noordhuis, T. Li, A. Paszke, J. Smith, B. Vaughan, P. Damania *et al.*, "Pytorch distributed: Experiences on accelerating data parallel training," *Proceedings of the VLDB Endowment*, vol. 13, no. 12.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[47] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[48] K. Simonyan and A. Zisserman, "Very deep convolutional net-

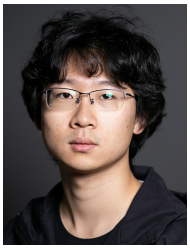
works for large-scale image recognition," in *Proceedings of the International Conference on Learning Representations*, 2015.
[49] Y. Li, B. Dang, W. Li, and Y. Zhang, "Glh-water: A large-scale dataset for global surface water detection in large-size very-high-resolution satellite imagery," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 22 213–22 221.
[50] X.-Y. Tong, G.-S. Xia, and X. X. Zhu, "Enabling country-scale land cover mapping with meter-resolution satellite imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 196, pp. 178–196, 2023.
[51] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz, and M. Y. Yang, "Uavid: A semantic segmentation dataset for uav imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 165, pp. 108–119, 2020.
[52] J. Li, X. Huang, and L. Tu, "Whu-ohs: A benchmark dataset for large-scale hersepctral image classification," *International Journal of Applied Earth Observation and Geoinformation*, vol. 113, p. 103022, 2022.
[53] W. Diao, H. Yu, K. Kang, T. Ling, D. Liu, Y. Feng, H. Bi, L. Ren, X. Li, Y. Mao, and X. Sun, "Ringmo-aerial: An aerial remote sensing foundation model with a affine transformation contrastive learning," *arXiv preprint arXiv:2409.13366*, 2024.
[54] B. Dang and Y. Li, "Msresnet: Multiscale residual network via self-supervised learning for water-body detection in remote sensing imagery," *Remote Sensing*, vol. 13, 2021.
[55] J. Kang, H. Guan, D. Peng, and Z. Chen, "Multi-scale context extractor network for water-body extraction from high-resolution optical remotely sensed images," *International Journal of Applied Earth Observation and Geoinformation*, vol. 103, p. 102499, 2021.
[56] S. Yi, X. Liu, J. Li, and L. Chen, "Uavformer: A composite transformer network for urban scene segmentation of uav images," *Pattern Recognition*, vol. 133, p. 109019, 2023.
[57] X. Huang, M. Dong, J. Li, and X. Guo, "A 3-d-swin transformer-based hierarchical contrastive learning method for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.
[58] L. Tu, J. Li, X. Huang, J. Gong, X. Xie, and L. Wang, "S2hm2: A spectral–spatial hierarchical masked modeling framework for self-supervised feature learning and classification of large-scale hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–19, 2024.
[59] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, p. 180161, 2018.
[60] S. Graham, H. Chen, J. Gamper, Q. Dou, P.-A. Heng, D. Snead, Y. W. Tsang, and N. Rajpoot, "Mild-net: Minimal information loss dilated network for gland instance segmentation in colon histology images," *Medical image analysis*, vol. 52, pp. 199–211, 2019.

Wei Chen received the BS degree in remote sensing science and technology and the MS degree in surveying and mapping engineering from Wuhan University, Wuhan, China in 2019. He is currently pursuing his PhD degree at the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China. He has published several papers in Remote Sensing of Environment, Science China Information Sciences, etc. His research interests include remote sensing semantic segmentation and large-size remote sensing image interpretation.

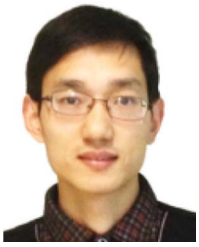
Lorenzo Bruzzone (Fellow, IEEE) received the Laurea (MS) degree in electronic engineering in 1993 and the PhD degree in telecommunications in 1998 from the University of Genoa, Italy. He is currently a Full Professor of Telecommunications at the University of Trento. He is the Founder and the Director of the Remote Sensing Laboratory with the Department of Information Engineering and Computer Science. He has authored over 400 journal articles, 390 conference papers, and 25 book chapters, accumulating more than 56 500 citations with an h-index of 109. His research interests include remote sensing, radar and SAR, signal processing, machine learning, and pattern recognition. He is editor/co-editor of 18 books/conference proceedings and two scientific books. He has been the founder of the IEEE Geoscience and Remote Sensing Magazine for which he has been Editor-in-Chief between 2013-2017. He currently serves as Associate Editor for IEEE Transactions on Geoscience and Remote Sensing.

Bo Dang received the BS degree in remote sensing science and technology from Wuhan University, Wuhan, China in 2022. He is currently working toward the PhD degree with the School of Remote Sensing and Information Engineering, Wuhan University. He has published several papers in CVPR, AAAI, ISPRS Journal of Photogrammetry and Remote Sensing, etc. His research interests include remote sensing semantic segmentation and remote sensing foundation model.

Yuan Gao received the BS degree in biomedical engineering and the MS degree in pattern recognition and intelligent systems from the Huazhong University of Science and Technology, Wuhan, China, in 2009 and 2012, respectively, and the PhD degree in electronic engineering from the City University of Hong Kong, Hong Kong, in 2016. He was a Visiting Student with the Department of Statistics, University of California at Los Angeles, Los Angeles, in 2015. He is currently an Associate Professor of the Computational and Photogrammetric Vision Team at Wuhan University. His research interests include multi-modality and Multi-Task AI, with its efficiency in training and inference, and its applications mainly to Computer Vision.



Youming Deng received the BS degree in remote sensing science and technology from Wuhan University, Wuhan, China in 2022. He is currently working toward the PhD degree with the Department of Computer Science, Cornell University, New York, United States. He has published a first-authored paper in ECCV. His research interests include scene graph generation, computer vision, and 3D vision.



Jin-Gang Yu received the BS degree from Xi'an Jiaotong University, Xi'an, China, in 2005, and the MS and PhD degrees from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2007 and 2014, respectively. He was a Postdoctoral Research Associate with the Department of Computer Science and Technology, University of Nebraska at Lincoln, Lincoln, NE, USA, from 2014 to 2016. He spent three years as a Research and Development Engineer with ZTE Corporation, Shenzhen, China, and with Nortel Networks Corporation, Guangzhou, China, before starting the PhD Program at HUST. He joined the South China University of Technology, Guangzhou, China, in 2016, where he is currently an Associate Professor. His research interests include computer vision, pattern recognition, and machine learning.

Liangqi Yuan (Student Member, IEEE) received the BE degree from Beijing Information Science and Technology University, Beijing, China, in 2020, and the MS degree from Oakland University, Rochester, MI, USA, in 2022. He is currently pursuing the PhD degree with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA. His research interests are in the areas of sensors, the Internet of Things, signal processing, and machine learning.

Yansheng Li (Senior Member, IEEE) received the BS degree in information and computing science from Shandong University, Weihai, China, in 2010, and the PhD degree in pattern recognition and intelligent system from the Huazhong University of Science and Technology, Wuhan, China, in 2015. He is currently a full professor and vice dean with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China. He has authored more than 100 peer-reviewed papers, such as Nature Machine Intelligence, IEEE Transactions on Pattern Analysis and Machine Intelligence, International Journal of Computer Vision. His research interests include knowledge graph, deep learning and their applications in remote sensing Big Data mining. He is an associate editor of IEEE Transactions on Geoscience and Remote Sensing.