Towards Privacy-preserved Pre-training of Remote Sensing Foundation Models with Federated Mutual-guidance Learning

Supplementary Material

A. Overview

We provide the following materials to supplement our paper and divide them into two sections.

- We provide the theoretical analysis of our proposed FedSense in Sec. B.
- We provide the details of our pre-training datasets and downstream datasets in Sec. C

B. Theoretical Analysis

B.1. Assumptions

Assumption 1 (Smoothness) *The self-supervised loss $\mathcal{L}_m^{ssl}$ is L-smooth:*

$$\|\nabla\mathcal{L}_m^{ssl}(\theta_1) - \nabla\mathcal{L}_m^{ssl}(\theta_2)\| \leq L\|\theta_1 - \theta_2\|, \quad \forall\theta_1, \theta_2 \tag{19}$$

Assumption 2 (Bounded Gradient) *Local gradients are bounded:*

$$\mathbb{E}[\|\nabla\mathcal{L}_m^{total}(\theta_m)\|^2] \leq G^2, \quad \forall m \tag{20}$$

Assumption 3 (Parameter Discrepancy) *The discrepancy between local and global models satisfies:*

$$\|\theta_m - \Theta\| \leq \delta, \quad \forall m \in [M] \tag{21}$$

where δ quantifies the maximum client drift.

B.2. Key Lemmas

Lemma 1 (Optimal Perturbation Bound) *Under Assumption 2, the optimal perturbation $\widetilde{\epsilon}$ in SCG satisfies:*

$$\|\widetilde{\epsilon}\| \leq \lambda\sqrt{\beta^2\delta^2 + G^2} \tag{22}$$

Proof 1 *From the perturbation approximation:*

$$\widetilde{\epsilon} \approx \lambda\frac{\nabla\mathcal{L}_m^{disc}}{\|\nabla\mathcal{L}_m^{disc}\|}$$

$$\|\widetilde{\epsilon}\| \leq \lambda\sqrt{\frac{\|\nabla\mathcal{L}_m^{disc}\|^2}{\|\nabla\mathcal{L}_m^{disc}\|^2}} = \lambda$$

Using the parameter discrepancy term $\nabla\mathcal{L}_m^{disc} = \beta(\theta_m - \Theta)$ and Assumption 3:

$$\|\nabla\mathcal{L}_m^{disc}\| \leq \beta\delta$$

Combining with the gradient bound G via the triangle inequality completes the proof.

Lemma 2 (Quantization Error Decay) *Let e_m^t be the feedback error in CSG. With momentum factor $\alpha \in (0, 1)$, the error decays geometrically:*

$$\|e_m^t\| \leq \alpha^t\|e_m^0\| + \frac{1-\alpha}{1-\alpha^{t+1}}\sum_{k=0}^{t}\alpha^{t-k}\|\epsilon_q^k\| \tag{23}$$

where ϵ_q^k is the quantization error at round k.

Proof 2 *Unrolling the recursive error update:*

$$\begin{aligned} e_m^t &= \alpha e_m^{t-1} + (1-\alpha)\epsilon_q^t \\ &= \alpha^t e_m^0 + (1-\alpha)\sum_{k=1}^{t}\alpha^{t-k}\epsilon_q^k \end{aligned}$$

Taking norms and applying the triangle inequality:

$$\begin{aligned} \|e_m^t\| &\leq \alpha^t\|e_m^0\| + (1-\alpha)\sum_{k=1}^{t}\alpha^{t-k}\|\epsilon_q^k\| \\ &\leq \alpha^t\|e_m^0\| + \frac{1-\alpha}{1-\alpha^{t+1}}\sum_{k=0}^{t}\alpha^{t-k}\|\epsilon_q^k\| \end{aligned}$$

The geometric series bound completes the proof.

B.3. Main Convergence Result

Theorem 1 (Convergence Guarantee) *Under Assumptions 1-3, let learning rate $\gamma = \frac{1}{L\sqrt{T}}$. After T rounds, the averaged gradient satisfies:*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla\mathcal{L}^{total}(\Theta^t)\|^2 \leq \frac{2L(\mathcal{L}^0 - \mathcal{L}^*)}{\sqrt{T}} + \frac{C}{T}\sum_{t=1}^{T}(\delta^2 + \|e^t\|^2) \tag{24}$$

where C is a constant combining L, G, β, λ.

Proof 3 (Proof Sketch) *Using smoothness (Assump. 1):*

$$\mathcal{L}^{t+1} \leq \mathcal{L}^t + \langle\nabla\mathcal{L}^t, \Theta^{t+1} - \Theta^t\rangle + \frac{L}{2}\|\Theta^{t+1} - \Theta^t\|^2$$

Substituting the update rule $\Theta^{t+1} = \Theta^t - \gamma(\nabla\mathcal{L}^{total} + e^t)$:

$$\begin{aligned} \mathbb{E}[\mathcal{L}^{t+1}] \leq \mathbb{E}[\mathcal{L}^t] - \gamma\mathbb{E}\|\nabla\mathcal{L}^t\|^2 + \gamma\mathbb{E}\langle\nabla\mathcal{L}^t, e^t\rangle \\ + \frac{L\gamma^2}{2}\mathbb{E}\|\nabla\mathcal{L}^t + e^t\|^2 \end{aligned}$$

Hyperparameters analysis. We provide more analysis on some hyperparameters λ, ρ, and α in Fig. 4. They are insensitive to the performance of FedSense with our self-stabilized design.

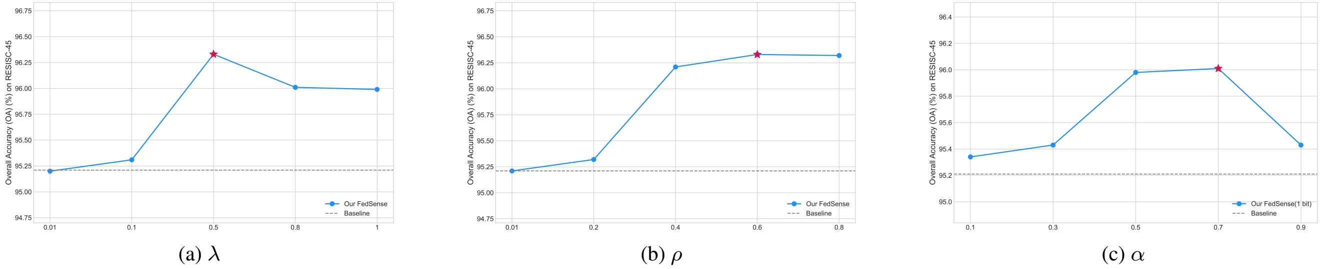

Figure 4. **Hyperparameters analysis on λ, ρ, and α.**

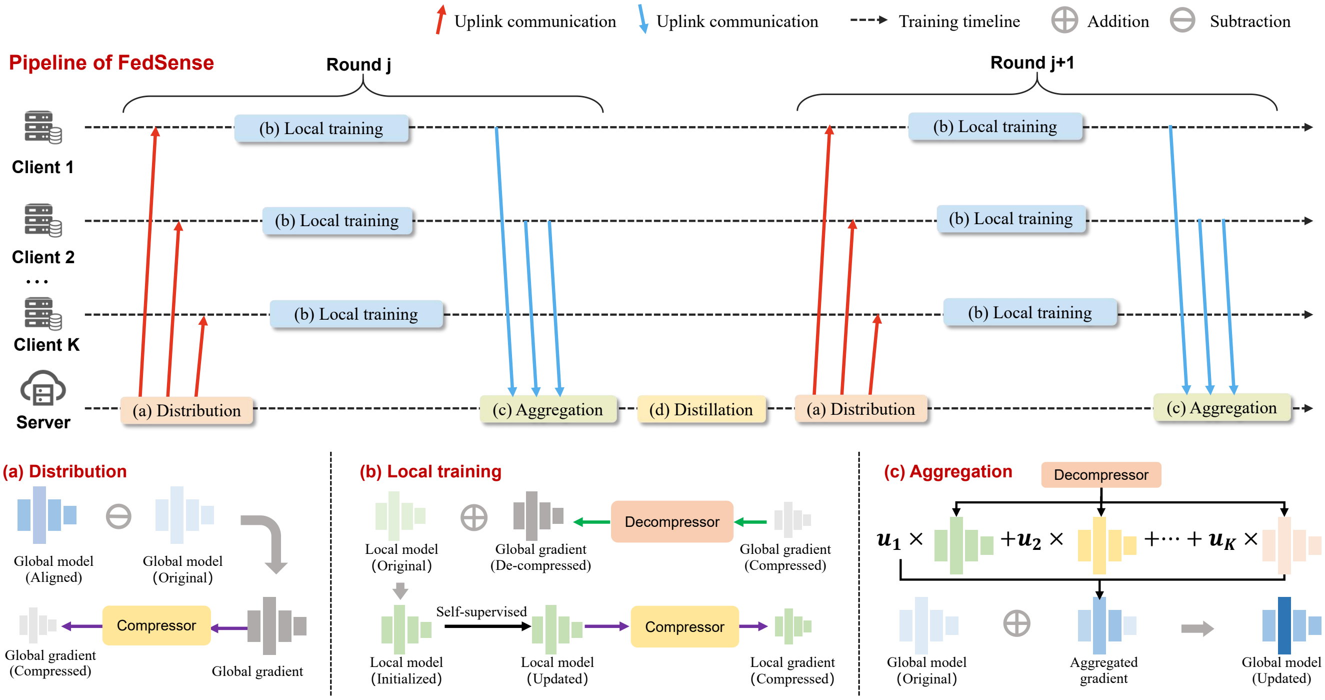

Figure 5. **Pipeline of privacy-preserved pre-training of RSFMs.**

C. Dataset details and implementation details

Scene Classification.

(1) *Aerial Image Dataset (AID) [55]*. This dataset is comprised of 10,000 images across 30 classes, all sourced from Google Earth and cropped to 600×600 pixels. It also includes diverse resolutions from 0.5 to 8 meters per pixel and geographic variations to enhance robustness.

(2) *NWPU-RESISC45 [8]*. This dataset comprises 31,500 RGB images at resolutions from 0.2m to 30m across 45 scene classes, each with 700 samples with a size of 256 times 256 pixels. It offers significant variability in translation, scale, viewpoint, illumination, and occlusion. It also has high within-class diversity and inter-class similarity.

Object Detection.

(1) *DIOR-R [9]*. This dataset consists of 23,463 remote sensing images, with 192,472 annotated object instances spanning 20 categories. The size of each image is 800×800 pixels, and spatial resolutions range from 0.5m to 30m. With emphasis on high inter-class similarity, intra-class diversity, and object size variability, it is designed to benchmark object detection methods in diverse conditions such as different imaging times, weathers, and resolutions.

(2) *DOTA-v1.0 [56]*. This dataset consists of 2,806 aerial images, measuring from 800×800 to 4000×4000 pixels, annotated with 188,282 instances across 15 categories that include airplanes, ships, vehicles, and bridges. The objects in this dataset are presented in diverse scales, orientations and aspect ratios.

Semantic Segmentation.

(1) *LoveDA [52]*. This dataset for domain-adaptive seman-

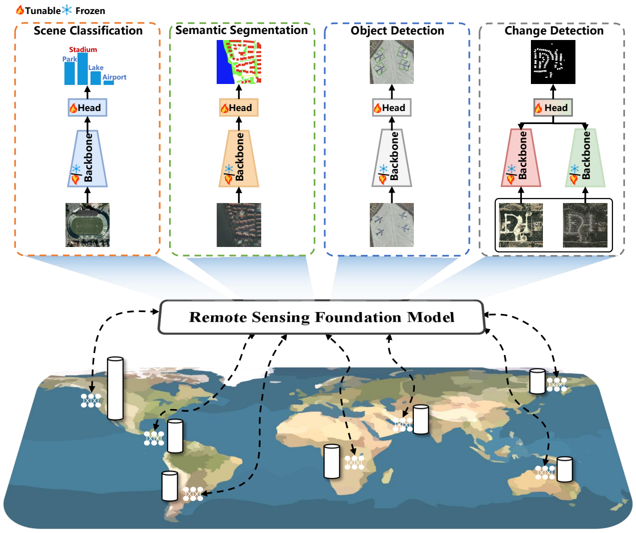

Figure 6. **Illustration on downstream usage of collaboratively pre-trained RSFMs to accommodate various Earth Observation tasks.**

tic segmentation features 5,987 images with a spatial resolution of 0.3 m, each sized at 1024×1024 pixels in RGB format. Covering 536.15 km^2, it spans urban and rural areas across Nanjing, Changzhou and Wuhan, and includes pixel-level annotations across seven land-cover categories. It addresses challenges of multi-scale objects, complex backgrounds, and inconsistent class distributions, supporting semantic segmentation and unsupervised domain adaptation.

(2) *Inria [38]*. This dataset comprises high-resolution RGB aerial imagery, with 180 training and 180 test tiles (each 1500×1500 pixels, 0.3 m resolution). It focuses on two classes: building and non-building, covering a total of 405 km^2 of urban areas across five cities in the U.S. and Austria. The dataset emphasizes generalization challenges, supporting semantic segmentation across diverse urban landscapes.

Change Detection.

(1) *LEVIR-CD+ [6]*. This dataset is a high-resolution urban building change detection dataset comprised of 985 RGB image pairs from Google Earth, each measuring 1024×1024 pixels with a spatial resolution of 0.5 meters per pixel. Spanning 20 regions in Texas, the dataset includes building and land use change masks, covering the years 2002 to 2020 with a 5-year observation interval. It features predominantly urban areas and near-nadir imagery, making it accessible for change detection studies.

(2) *SECOND [62]*. This dataset is a large-scale semantic change detection benchmark, comprising 4,662 image pairs, each with a size of 512×512 pixels. The images were collected from multiple platforms across multiple cities including Hangzhou, Chengdu, and Shanghai. It focuses on six land-cover classes: non-vegetated ground surface, tree, low vegetation, water, buildings, and playgrounds. SECOND also offers approximately 30 change types, including changes within the same land-cover class, with pixel-level annotations ensuring high diversity and label accuracy.

ID	Source	#samples	GSD (m)	Coverage
Server	WorldView-3/4	240,000	0.5-2.5	Global
Client 01	NOAA	22,292	0.25	USA
Client 02	GF-2	27,300	4.0	China
Client 03	WorldView-2	88,272	0.3-0.5	Global
Client 04	Mixed	125,474	0.3-25	Global
Client 05	QB-2/GE-1	180,562	0.3	Global
Client 06	JL-1/GF-7	40,816	0.8	China
Client 07	Mixed	90,000	0.3-25	Global
Client 08	QB-2/GE-1	180,000	0.3-25	Global
Client 09	NAIP	45,000	1.25	USA
Client 10	Mixed	50,800	0.3-0.75	Global
Total	Multi-source	1,000,000	0.25-25	Global

Table 7. **Details of the pre-training datasets.**