SkySense: A Multi-Modal Remote Sensing Foundation Model Towards Universal Interpretation for Earth Observation Imagery

Xin Guo[1]*, Jiangwei Lao[1]*, Bo Dang[§2]*,
Yingying Zhang[1], Lei Yu[1], Lixiang Ru[1], Liheng Zhong[1], Ziyuan Huang[1], Kang Wu[§2], Dingxiang Hu[3,1],
Huimei He[3,1], Jian Wang[1], Jingdong Chen[1], Ming Yang[1†], Yongjun Zhang[2], Yansheng Li[2†]
[1]Ant Group [2]Wuhan University [3]MYBank
{bangzhu.gx, wenshuo.ljw}@antgroup.com, bodang@whu.edu.cn

Abstract

Prior studies on Remote Sensing Foundation Model (RSFM) reveal immense potential towards a generic model for Earth Observation. Nevertheless, these works primarily focus on a single modality without temporal and geo-context modeling, hampering their capabilities for diverse tasks. In this study, we present SkySense, a generic billion-scale model, pre-trained on a curated multi-modal Remote Sensing Imagery (RSI) dataset with 21.5 million temporal sequences. SkySense incorporates a factorized multi-modal spatiotemporal encoder taking temporal sequences of optical and Synthetic Aperture Radar (SAR) data as input. This encoder is pre-trained by our proposed Multi-Granularity Contrastive Learning to learn representations across different modal and spatial granularities. To further enhance the RSI representations by the geo-context clue, we introduce Geo-Context Prototype Learning to learn region-aware prototypes upon RSI's multi-modal spatiotemporal features. To our best knowledge, SkySense is the largest Multi-Modal RSFM to date, whose modules can be flexibly combined or used individually to accommodate various tasks. It demonstrates remarkable generalization capabilities on a thorough evaluation encompassing 16 datasets over 7 tasks, from single- to multi-modal, static to temporal, and classification to localization. SkySense surpasses 18 recent RSFMs in all test scenarios. Specifically, it outperforms the latest models such as GFM, SatLas and Scale-MAE by a large margin, i.e., 2.76%, 3.67% and 3.61% on average respectively. We will release the pre-trained weights to facilitate future research and Earth Observation applications.

1. Introduction

Remote Sensing Imagery (RSI) interpretation is crucial in understanding our common home, the Earth [16, 60], via quite diverse tasks [5, 13, 43, 72], *e.g.* crop monitoring, natural disaster management, *etc*. Every task may require significant dedicated efforts and resources to build a task-specific model. Recently, Foundation Model emerges as a pre-trained generic model that excels in a wide range of downstream tasks [70, 76]. Hence, there is a soaring interest in exploring a comprehensive Remote Sensing Foundation Model (RSFM) for many Earth Observation (EO) tasks.

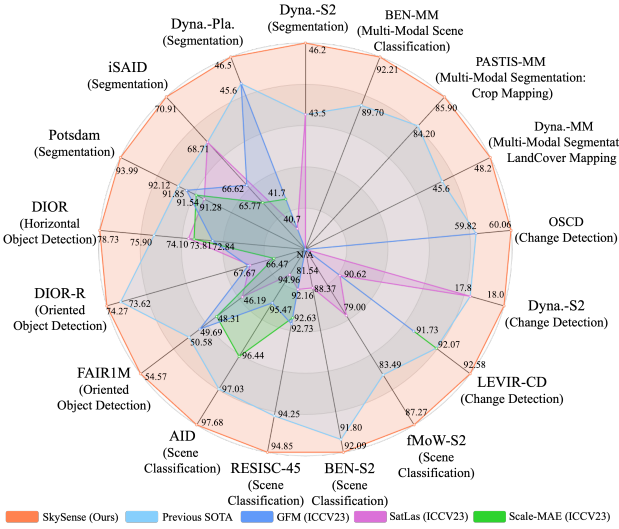

Figure 1. SkySense has achieved superior performance on 16 datasets over 7 distinct tasks compared with 18 state-of-the-art RSFMs and supports a board range of EO imagery interpretations.

The key question naturally arises: *What is essential for a RSFM?* First of all, an ideal RSFM should possess the ability to perceive multi-modal temporal RSI. EO heavily relies on multi-modal time series of remote sensing data, including temporal optical and Synthetic Aperture Radar (SAR) data. Individual modality offers unique advantages and complements to each other. For example, optical images provide rich spectral bands and texture details but are susceptible to weather [77]. In contrast, SAR sensors cap-

*Equally contributing first authors. †Corresponding authors. §Work done during the internship of the author at Ant Group.

Model	Different EO Interpretation Input Types			
	Single-Modal O(RGB)	Single-Modal O(Ms)	Multi-Modal Static O & SAR	Multi-Modal Temporal O & SAR
SkySense	✔	✔	✔	✔
SatLas[4]	✔	✔		
GFM[47]	✔			
Scale-MAE[50]	✔			

Table 1. SkySense supports various input types. O(RGB): Optical RGB images; O(Ms): Optical multispectral images.

ture clear imagery in all weather conditions [31, 38]. Moreover, the time series of such data provide the crucial temporal clue to various tasks [5, 22, 73] like change prediction. Second, a RSFM should be easy to tailor when being deployed for EO tasks using different modalities (*i.e.*, single- and multi-modal) at different spatial (*i.e.*, pixel-, object-, and image-level) granularities. Last but not the least, remote sensing data is inherently contingent on their space-time coordinates, which provide rich regional and seasonal geo-context that benefits RSI interpretation a lot, as indicated in [11, 24, 32, 39, 40]. Therefore, a RSFM shall bear the vital capability of effective geo-context learning and utilization.

Previous works on RSFM [1, 2, 4, 8, 17, 33, 45–48, 50, 55, 57, 62, 63, 65, 66] have demonstrated their preliminary success on several specific datasets. However, these RSFMs, while proficient in certain areas, are limited in their applications to EO tasks, due to factors such as single-modal pre-training and the neglect of geo-context.

In this paper, we propose SkySense, a billion-scale Multi-Modal Remote Sensing Foundation Model (MM-RSFM). SkySense incorporates 2.06 billion parameters and is pre-trained on a large-scale multi-modal dataset which comprises 21.5 million RSI temporal sequences extracted from high-spatial-resolution optical images (HSROIs), medium-resolution temporal multispectral imagery (TMsI) and temporal SAR imagery (TSARI). To handle the multi-modal temporal RSI sequences, SkySense employs a factorized multi-modal spatiotemporal encoder to perform spatial feature extraction and multi-modal temporal fusion independently, since RSI sequence are spatially-aligned in nature. It leads to a modular design allowing flexible use of its modules, *i.e.*, the spatial encoder can be either used alone or in combination of the fusion module to support tasks from static single-modal to temporal multi-modal. This design delivers strong modeling of RSI sequences while using substantially less parameters compared to common 3D structures [44, 75]. The factorized encoder is pre-trained by Multi-Granularity Contrastive Learning to construct features from different modal and spatial granularities. Furthermore, we propose Geo-Context Prototype Learning to generate regional prototypes from RSI features given geo-locations. This approach enhances multi-modal spatiotemporal representation learning by leveraging the regional context clue hidden in numerous unlabeled RSI.

SkySense has achieved the state-of-the-art (SOTA) performance across a variety of modalities and EO tasks, as shown in Fig. 1. We evaluate SkySense on a diverse set of 16 datasets [9, 14, 15, 17, 18, 22, 36, 53, 56, 59, 67, 68], where the selection covers different task types, modalities and spatial scales. The results demonstrate that SkySense outperforms 18 advanced RSFMs [1, 2, 4, 8, 17, 45–48, 50, 55, 57, 62, 63, 65, 66] in all test scenarios, validating its competitive edge for a broad range of EO interpretation tasks. Tab. 1 compares our work with latest representative studies w.r.t. various input types of EO interpretation.

In summary, our technical contributions are:

- We propose SkySense, the largest MM-RSFM to date with a modular design, which is capable of handling diverse tasks, from single- to multi-modal, static to temporal, and classification to localization.
- The design of SkySense involves three novel technical components: a) A factorized multi-modal spatiotemporal encoder to effectively process multi-modal temporal RSI; b) Multi-Granularity Contrastive Learning that learns features at various levels of granularities to facilitate different tasks; c) Geo-Context Prototype Learning to extract region-aware geo-context clue to enable implicit geo-knowledge integration.
- We extensively compare SkySense with 18 recently published RSFMs. Our model has achieved the SOTA performance, supparssing the latest models like GFM, SatLas and Scale-MAE by over 2.5% on average. We hope the release of pre-trained weights will contribute to the Remote Sensing community and facilitate future research.

2. Related Work

2.1. Remote Sensing Foundation Model

Recent Remote Sensing Foundation Models draw their primary inspiration from the research on Vision Foundation Model [3, 7, 10, 19, 23, 25–27, 41, 49, 61]. Remote sensing data inherently integrates space-time coordinates and has diverse spatial scales. The maintstream RSFMs extend the foundation model techniques to space-time RS data, such as Contrastive Learning. For instance, GASSL [2] utilized geo-location prediction as an additional pre-text task in the MoCo-v2 framework [12]. Multiple views with different sizes were utilized by DINO-MC [66] for self-supervised learning within the DINO framework [7]. SeCo [46] and CACo [45] both proposed Contrastive Learning to perceive short-term and long-term changes by using the spatiotemporal structure of temporal RSI sequences. Besides, there are works either improving the MIM-based framework [50, 55, 62] or exploring the model scale-up [8]. For example, RingMo [55] modified MAE to adapt to the dense

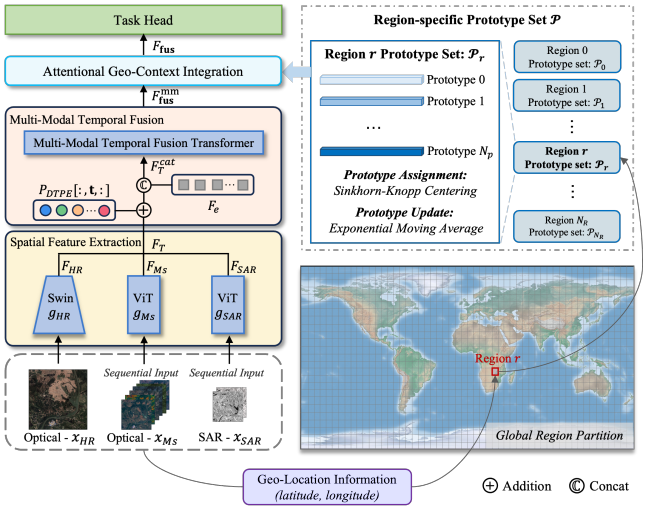

Figure 2. The overview of our SkySense model architecture.

objects in RSI. SatMAE [17] employed TMsI to enhance the performance on temporal sequences. Scale-MAE [50] built a framework with scale-aware encoder. Recent efforts such as CMID [48] and GFM [47] have commenced to explore amalgamation of CL and MIM strategies. Concurrently, CROMA [21] and DeCUR [64] investigated multi-modal pre-training for single- and multi-modal tasks using static imagery. In this study, we propose a comprehensive MM-RSFM, SkySense, to fill the gap in existing RSFMs, *i.e.*, single modality of RingMo, CACo, *etc.*, static input of Scale-MAE, CROMA, *etc.*, and the neglect of geo-context of SatLas, RVSA, *etc.*

3. SkySense

In this section, we introduce the pre-training dataset and the design choices for individual module respectively.

3.1. Pre-training Dataset

We curate an extensive multi-modal remote sensing dataset with temporal sequences, containing RSI from various sources: HSROIs from WorldView-3, 4, *etc.* (RGB band), TMsI from Sentinel-2 (B2-8, B8A, B11-12 band) and TSARI from Sentinel-1 (VV, VH Polarization). All data is geo-spatially aligned. Strictly speaking, HSROIs and TMsI shall be categorized to the optical modality, while TSARI falls to the SAR modality. However, due to HSROIs and TMsI's significant difference in spectral band and ground sample distance, we regard HSROIs and TMsI as two distinct modalities for simplicity in this paper. The dataset comprises 21.5 million training samples, each consisting of a static HSROI with rich texture details, a TMsI containing temporal and multispectral data, a TSARI providing backscatter polarization under cloud coverage, and the metadata like geo-location and acquisition date for geo-context modeling. This dataset covers a great variety of scenarios across resolution, spectrum, and imaging mechanism. More details of the data are included in the supplementary materials. We construct the input for SkySense as $\{x_{HR}, x_{Ms}, x_{SAR}\}$, where x_{HR} represents a static HSROI; x_{Ms} is a Sentinel-2 TMsI after filtering cloudy images, where we randomly select 20 images to form the sequence; and x_{SAR} stands for a standard-calibrated TSARI, from which we randomly select 10 images for training.

3.2. Model Architecture

Factorized Multi-Modal Spatiotemporal Encoder. The overall architecture of our method is illustrated in Fig. 2. In a multi-modal input $\{x_{HR}, x_{Ms}, x_{SAR}\}$, the pixels within each RSI naturally align with the others given the same geo-location. Upon this, we propose a factorized encoder that initially extracts spatial features from each RSI independently and then fuses them to capture a multi-modal spatiotemporal representation. The design separates spatial feature extraction from the feature fusion, enabling the integration of the clues from modality, time and geo-context.

Spatial Feature Extraction. To handle the spatially aligned sequence input $\{x_{HR}, x_{Ms}, x_{SAR}\}$, we utilize the spatial encoder g_{HR}, g_{Ms} and g_{SAR} for each individual RSI from HSROI, TMsI and TSARI respectively. As shown in Eq. (1), the obtained feature $F_i \in \mathbb{R}^{h\times w\times T_i\times d}, i \in \{HR, Ms, SAR\}$ are of the same size in spatial dimension, where h and w are the height and width of F_i, T_{HR}, T_{Ms}, T_{SAR} represent the sequence lengths of HSROI, TMsI, and TSARI respectively, and d is the feature dimension. The initial multi-modal temporal feature representation $F_T \in \mathbb{R}^{N_S\times N_T\times d}$ is generated by concatenating all F_i along the time dimension, where $N_S = h \times w$ represents the feature size in the spatial dimension, and $N_T = \sum_{i\in\{HR,Ms,SAR\}} T_i$ represents the total sequence length across all modalities,

$$\begin{aligned} F_i &= g_i(x_i), i \in \{HR, Ms, SAR\}, \\ F_T &= \text{Concat}\left[F_{HR}, F_{Ms}, F_{SAR}\right]. \end{aligned} \tag{1}$$

Multi-modal Temporal Fusion. Next, we incorporate the date-specific temporal positional encoding $P_{DTPE}[:, \mathbf{t}, :] \in \mathbb{R}^{1\times N_T\times d}$ to F_T through broadcasting, creating F_T^{date} for date-aware modeling. F_T^{date} is then concatenated with an extra token $F_\mathbf{e} \in \mathbb{R}^{N_S\times 1\times d}$ [20] (see Eq. (2)),

$$\begin{aligned} F_T^{date} &= F_T + P_{DTPE}[:, \mathbf{t}, :], \\ F_T^{cat} &= \text{Concat}\left[F_\mathbf{e}, F_T^{date}\right] \in \mathbb{R}^{N_S\times(1+N_T)\times d}, \end{aligned} \tag{2}$$

where $\mathbf{t} \in \mathbb{R}^{N_T}$ is a vector containing the acquisition dates of all RSI in the current batch. $P_{DTPE} \in \mathbb{R}^{1\times 365\times d}$ is a learnable parameter representing different dates of a year, which is essential for tasks affected by seasons (*e.g.*, crop recognition). F_T^{cat} is then fed into the Multi-modal

Temporal Fusion Transformer, composed of multiple Naive Transformer encoder layers. This module employs self-attention to integrate multi-modal temporal data, generating the multi-modal spatiotemporal feature $F_{\mathbf{fus}}^{\mathrm{mm}} \in \mathbb{R}^{N_S \times 1 \times d}$.

Attentional Geo-Context Integration. Each RSI's geographical location may reveal rich region-specific geo-context. It is valuable for RSI interpretation as indicated by [11, 24, 32, 40]. To utilize this contextual clue to enhance $F_{\mathbf{fus}}^{\mathrm{mm}}$, we employ a region-specific prototype set $\mathcal{P} \in \mathbb{R}^{N_R \times N_p \times d}$ (shown on the right side of Fig. 2), where N_R is the number of regions, N_p represents the number of prototypes for each region and d denotes the feature dimension. The learning procedure of $\mathcal{P}$ will be elaborated in Sec. 3.3. Specifically, a regional prototype subset $\mathcal{P}_r \in \mathbb{R}^{N_p \times d}$ is chosen from $\mathcal{P}$ based on the geo-location embedded with $F_{\mathbf{fus}}^{\mathrm{mm}}$. $F_{\mathbf{fus}}^{\mathrm{mm}}$ is then attended to the prototypes of $\mathcal{P}_r$ through the attention mechanism, as shown in Eq. (3). The weights, computed from $\mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)$, facilitate a soft selection of prototypes in accordance with their similarity to $F_{\mathbf{fus}}^{\mathrm{mm}}$. The final representation $F_{\mathbf{fus}} \in \mathbb{R}^{N_S \times 2d}$ is generated by concatenation of $F_{\mathbf{fus}}^{\mathrm{mm}}$ and weighted sum of prototypes from $\mathcal{P}_r$ along the feature dimension. The prototypes represent a set of discriminative features linked to certain semantics like water body, cropland, *etc*. By finding the similar ones to $F_{\mathbf{fus}}^{\mathrm{mm}}$, we provide the standard representations of certain semantics to complement $F_{\mathbf{fus}}^{\mathrm{mm}}$,

$$F_{\mathbf{fus}} = \mathrm{Concat}\left[F_{\mathbf{fus}}^{\mathrm{mm}}, \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V\right], \quad Q = F_{\mathbf{fus}}^{\mathrm{mm}}, K = V = \mathcal{P}_r. \tag{3}$$

3.3. Pre-training

An overview of our pre-training procedure is illustrated in Fig. 3. We build the pre-training framework on a common teacher-student structure [7], since it conducts self-supervised learning using only positive pairs, which is easily accessible given spatially aligned RSI, avoiding complicated design of negative pairs. Teacher's parameter set θ' is updated through exponential moving average (EMA) [26] from student's parameter set θ.

Multi-Granularity Contrastive Learning. We propose Multi-Granularity Contrastive Learning for self-supervised learning on different modal and spatial granularities for diverse tasks. Given the input $\{x_{HR}, x_{Ms}, x_{SAR}\}$, two sets of random augmentations are employed, generating two groups of views $\{u_i\}$ and $\{v_i\}$, where $i \in \{HR, Ms, SAR\}$. u_i and v_i are subsequently fed into the spatial encoders from the student and teacher branches respectively. g_i is the student's spatial encoder and g_i' is the teacher's. The features are generated as in Eq. (4),

$$F_i = g_i(u_i), F_i' = g_i'(v_i) \quad i \in \{HR, Ms, SAR\}. \tag{4}$$

After applying the multi-modal temporal fusion and geo-context integration on F_i and F_i', the final feature $F_{\mathbf{fus}}$ and $F_{\mathbf{fus}}'$ are obtained. Initially, we establish pixel-, object- and image-level contrastive learning to progressively learn coarse-to-fine spatial features for various tasks.

Each temporal slice of F_i can be viewed as a pixel-level feature $F_i^{\mathbf{pix}} \in \mathbb{R}^{N_S \times d}$. Pixel-level contrastive learning loss $\mathcal{L}_{\mathbf{pix}}$ is obtained by averaging all $\mathcal{L}_{CL}$ across the spatial (s) and temporal (t) dimensions, as shown in Eq. (5). $f_i^{\mathbf{pix}} \in \mathbb{R}^d$ represents a feature vector from $F_i^{\mathbf{pix}}$ and $f_i^{\mathbf{pix}\prime}$ is its correspondence at the same geo-location. $\mathcal{L}_{CL}$ denotes the learning loss [7] between $f_i^{\mathbf{pix}}$ and $f_i^{\mathbf{pix}\prime}$, and

$$\mathcal{L}_{\mathbf{pix}}(F_i, F_i') = \frac{1}{N_S T_i} \sum_s \sum_t \mathcal{L}_{CL}(f_i^{\mathbf{pix}}, f_i^{\mathbf{pix}\prime}). \tag{5}$$

$F_i^{\mathbf{obj}} \in \mathbb{R}^{N_C \times d}$ denotes object-level feature generated from unsupervised clustering on pixel-level feature vectors $f_i^{\mathbf{pix}}$ in a single RSI, where N_C is the number of clusters. The clustering employs the same Sinkhorn-Knopp algorithm [6] we apply for Geo-Context Prototype Learning, as shown later. $f_i^{\mathbf{obj}} \in \mathbb{R}^d$ is the vector representing the cluster centers in $F_i^{\mathbf{obj}}$, which can be viewed as a general representation for a set of collected $f_i^{\mathbf{pix}}$. It usually corresponds to a certain ground object or semantics. The object-level contrastive learning loss is computed as Eq. (6),

$$\mathcal{L}_{\mathbf{obj}}(F_i, F_i') = \frac{1}{N_C T_i} \sum_s \sum_t \mathcal{L}_{CL}(f_i^{\mathbf{obj}}, f_i^{\mathbf{obj}\prime}). \tag{6}$$

$F_i^{\mathbf{img}} \in \mathbb{R}^d$ corresponds to the image-level feature, which is an average pooling result from $F_i^{\mathbf{pix}}$. Image-level contrastive learning loss is illustrated by Eq. (7),

$$\mathcal{L}_{\mathbf{img}}(F_i, F_i') = \frac{1}{T_i} \sum_t \mathcal{L}_{CL}(F_i^{\mathbf{img}}, F_i^{\mathbf{img}\prime}). \tag{7}$$

The fine-grained contrastive learning loss $\mathcal{L}_{FGCL}$ is the sum of pixel-, object- and image-level contrastive learning losses as Eq. (8). Finally we form the Multi-Granularity Contrastive Learning loss $\mathcal{L}_{MGCL}$ in Eq. (9). The concept of multi-granularity is reflected in two aspects: space and modality. In terms of space, contrastive learning is performed at the pixel-, object-, and image-level, facilitating representation learning that encapsulates diverse spatial dimensions. Regarding modality, we conduct contrastive learning on the feature of each single modality, *i.e.*, F_i, and the multi-modal feature after fusion, *i.e.*, $F_{\mathbf{fus}}$,

$$\mathcal{L}_{FGCL}(F_i, F_i') = \sum_{n \in \{\mathbf{pix}, \mathbf{obj}, \mathbf{img}\}} \mathcal{L}_n(F_i, F_i'), \tag{8}$$

$$\mathcal{L}_{MGCL} = \sum_{i \in \{HR, Ms, SAR\}} \mathcal{L}_{FGCL}(F_i, F_i') + \mathcal{L}_{FGCL}(F_{\mathbf{fus}}, F_{\mathbf{fus}}'). \tag{9}$$

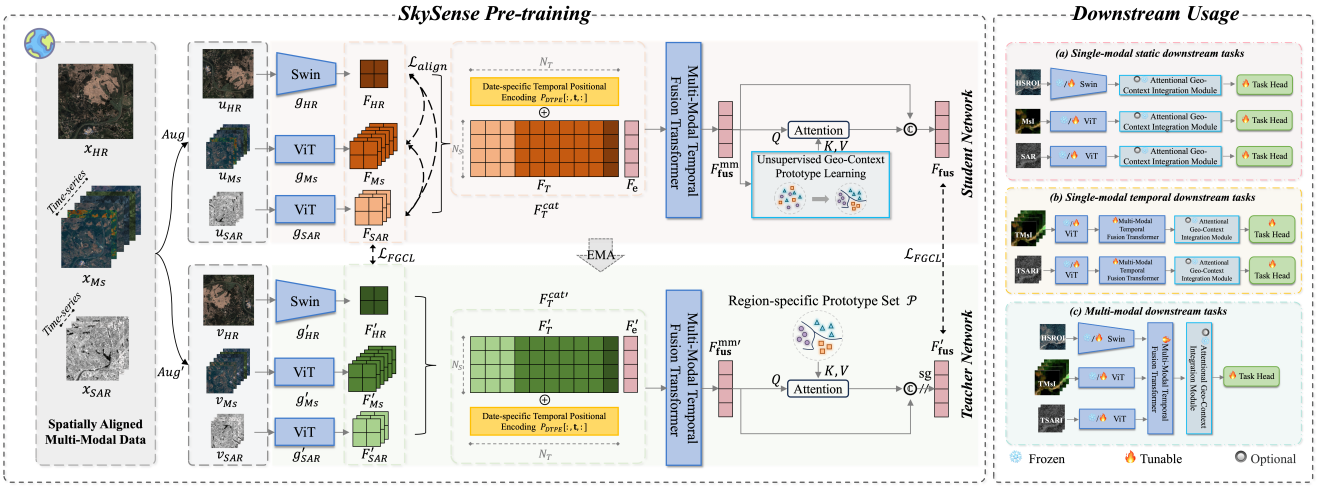

Figure 3. Overview of SkySense pre-training and downstream usage. SkySense employs data augmentations on the input and then feeds the augmented data into the student and teacher networks respectively. Multi-Granularity Contrastive Learning and Cross-Modal Alignment are proposed to pre-train the overall network. The region-specific prototype set $\mathcal{P}$ is learned on the student branch and it is frozen for downstream usage. Enhancing feature with $\mathcal{P}$ is optional. After pre-training, we adopt the parameters of the teacher branch for downstream tasks. Each pre-trained module can be used alone or combined with the others, with the chosen ones either frozen or fine-tuned.

Cross-Modal Alignment. The heterogeneity of multi-modal data poses a challenge for effective multi-modal feature fusion. We address this issue by adopting multi-modal contrastive loss $\mathcal{L}_{MMCL}$ [35] to form the alignment loss $\mathcal{L}_{align}$, as shown in Eq. (10),

$$\mathcal{L}_{align} = \sum_{i \neq j} \mathcal{L}_{MMCL}\left(F_i, F_j\right), \quad i, j \in \{HR, Ms, SAR\}. \tag{10}$$

$\mathcal{L}_{MMCL}$ maximizes the similarity of cross-modal features from the same geo-location, while minimizing it otherwise. Cross-modal alignment is performed on the student branch.

Unsupervised Geo-Context Prototype Learning. Different regions characterize distinct geographic landscapes and seasonal dynamics [30, 32] due to disparities in topography and climate. Prior arts have shown that the enlarged context can benefit the RSI interpretations [11, 24, 32, 40]. In this work, $F^{\text{mm}}_{\textbf{fus}}$ captures rich spatiotemporal clues for a small area. By clustering on numerous $F^{\text{mm}}_{\textbf{fus}}$, higher-level regional semantics are obtained as implicit geo-knowledge for a vast geo-spatial scope (see Fig. 5). Thus, we propose Geo-Context Prototype Learning to unsupervisedly extract regional geo-context from $F^{\text{mm}}_{\textbf{fus}}$ during pre-training.

We divide the globe into N_R regions and initialize a region-specific prototype set $\mathcal{P} \in \mathbb{R}^{N_R \times N_p \times d}$. Each prototype is learned from $F^{\text{mm}}_{\textbf{fus}}$. We leverage the geo-location of the RSI to retrieve the regional subset $\mathcal{P}_r \in \mathbb{R}^{N_p \times d}$ from $\mathcal{P}$. Then, we calculate the cosine similarity matrix $\mathbf{M} \in \mathbb{R}^{N_S \times N_p}$ between $F^{\text{mm}}_{\textbf{fus}}$ and $\mathcal{P}_r$ as in Eq. (11),

$$\mathbf{M} = \frac{F^{\text{mm}}_{\textbf{fus}} \cdot \mathcal{P}_r^{\mathrm{T}}}{\|F^{\text{mm}}_{\textbf{fus}}\| \|\mathcal{P}_r\|}, \tag{11}$$

We utilize the Sinkhorn-Knopp algorithm [6] on $\mathbf{M}$ to find the optimal assignment matrix $\mathbf{S} \in \mathbb{R}^{N_S \times N_p}$ between $F^{\text{mm}}_{\textbf{fus}}$ and the prototypes. This algorithm introduces the uniform distribution constraint to avoid trivial solution while achieving maximal similarity possible. We then use $\mathbf{S}$ to generate an update value for current sample's corresponding $\mathcal{P}_r$, denoted as $\overline{\mathcal{P}_r}$, as shown in Eq. (12),

$$\overline{\mathcal{P}_r} = \mathbf{S}^{\mathrm{T}} F^{\text{mm}}_{\textbf{fus}}. \tag{12}$$

Afterwards, we update $\mathcal{P}_r$ through EMA [26] as in Eq. (13), where $m \in [0, 1)$ is a momentum coefficient,

$$\mathcal{P}_r \leftarrow m\mathcal{P}_r + (1 - m)\overline{\mathcal{P}_r}. \tag{13}$$

Each $\mathcal{P}_r$ is updated during pre-training and used as the fixed geo-context for downstream tasks. Geo-Context Prototype Learning is only conducted on the student branch. It extracts generalized region-aware representations from numerous RSI within a consistent region, offering a complementary clue to enhance the feature of a single RSI.

As Geo-Context Prototype Learning is incorporated without an explicit loss term, our pre-training objective is shown in Eq. (14), where α and β are trade-off weights,

$$\mathcal{L} = \alpha\mathcal{L}_{MGCL} + \beta\mathcal{L}_{align}. \tag{14}$$

4. Experiments

Fig. 1 demonstrates SkySense's superior performance in all test scenarios. We conduct experiments on 16 datasets, covering different modalities and tasks, to ensure a comprehensive assessment. The right side of Fig. 3 shows how to apply SkySense to different tasks. Each pre-trained module is

Model	Publication	Dyna.-Pla.	iSAID	Potsdam	Dyna.-S2
		mIoU	mIoU	mF1	mIoU
GASSL [2]	ICCV'21	34.0/40.8	65.95	91.27	28.1/41.0
SeCo [46]	ICCV'21	-	57.20	89.03	29.4/39.8
SatMAE [17]	NIPS'22	32.8/39.9	62.97	90.63	30.1/38.7
RingMo† [55]	TGRS'22	-	67.20	91.27	-
RVSA [62]	TGRS'22	34.3/44.4	64.49	-	-
BFM† [8]	Arxiv'23	-	-	92.12	-
TOV [57]	JSTARS'23	32.1/37.8	66.24	92.03	-
SSL4EO [65]	GRSM'23	35.3/42.1	64.01	91.54	31.8/42.7
CMID [48]	TGRS'23	36.4/43.5	66.21	91.86	-
CACo [45]	CVPR'23	35.4/42.7	64.32	91.35	30.2/42.5
SAMRS† [63]	NIPS'23	-	66.26	91.43	-
SatLas [4]	ICCV'23	37.4/40.7	68.71	91.28	31.9/43.5
GFM [47]	ICCV'23	36.7/45.6	66.62	91.85	-
Scale-MAE [50]	ICCV'23	34.0/41.7	65.77	91.54	-
SkySense	-	**39.7/46.5**	**70.91**	**93.99**	**33.1/46.2**

(a) Semantic segmentation results.

Model	Horizontal	Oriented	
	DIOR	DIOR-R	FAIR1M
	mAP_{50}	mAP	mAP
GASSL [2]	67.40	65.65	48.15
SatMAE [17]	70.89	65.66	46.55
RingMo† [55]	75.90	-	46.21
RVSA [62]	73.22	71.05	47.04
BFM† [8]	-	73.62	-
TOV [57]	70.16	66.33	49.62
SSL4EO [65]	64.82	61.23	49.37
CMID [48]	75.11	66.37	50.58
CACo [45]	66.91	64.10	47.83
SatLas [4]	74.10	67.59	46.19
GFM [47]	72.84	67.67	49.69
Scale-MAE [50]	73.81	66.47	48.31
SkySense	**78.73**	**74.27**	**54.57**

(b) Object detection results.

Model	LEVIR-CD	OSCD	Dyna.-S2
	F1	F1	SCS
GASSL [2]	78.19	46.26	13.6/16.7
SeCo [46]	90.14	47.67	13.9/16.0
SatMAE [17]	87.65	52.76	14.8/16.2
RingMo† [55]	91.86	-	-
RVSA [62]	90.86	-	-
SpectralGPT† [28]	-	54.29	-
MATTER† [1]	-	59.37	-
DINO-MC [66]	-	52.70	14.5/15.6
SSL4EO [65]	89.05	35.08	12.3/17.5
CMID [48]	91.72	-	-
CACo [45]	81.04	52.11	15.3/15.8
SatLas [4]	90.62	-	13.3/17.8
GFM [47]	91.73	59.82	-
Scale-MAE [50]	92.07	-	-
SkySense	**92.58**	**60.06**	**15.4/18.0**

(c) Change detection results.

Table 2. Results of semantic segmentation, object detection and change detection. † means the code and weights are not released until November 11th, 2023, thus we report the metrics from the paper. - means the task is not supported or the value is unavailable in the paper.

Model	Single-label		Multi-label	Temporal
	AID (TR=20%/50%)	RESISC-45 (TR=10%/20%)	BEN-S2 (TR=10%/100%)	fMoW-S2 (TR=100%)
	OA	OA	mAP	Top-1/5 Acc
GASSL [2]	93.55/95.92	90.86/93.06	79.24/87.40	50.69/77.99
SeCo [46]	93.47/95.99	89.64/92.91	82.62/87.81	51.65/77.40
SatMAE [17]	95.02/96.94	91.72/94.10	86.18/89.50	63.84/-
RingMo† [55]	96.90/98.34	94.25/95.67	-	-
RVSA [62]	97.03/98.50	93.93/95.69	-	-
DINO-MC [66]	-	-	84.20/88.75	60.16/83.49
TOV [57]	95.16/97.09	90.97/93.79	-	-
SSL4EO [65]	91.06/94.74	87.60/91.27	87.10/91.80	51.70/76.77
CMID [48]	96.11/97.79	94.05/95.53	-	-
CACo [45]	90.88/95.05	88.28/91.94	81.30/87.00	50.72/76.31
CROMA† [21]	-	-	88.29/-	63.59/-
SatLas [4]	94.96/97.38	92.16/94.70	82.80/88.37	57.95/79.00
GFM [47]	95.47/97.09	92.73/94.64	86.30/-	-
Scale-MAE [50]	96.44/97.58	92.63/95.04	-	-
SkySense	**97.68/98.60**	**94.85/96.32**	**88.67/92.09**	**64.38/87.27**

Table 3. Scene classification results.

designed to allow for combined or individual use, with the flexibility to be either frozen or fine-tuned as needed. More details are included in the supplementary materials.

4.1. Pre-training Implementation

The model is pre-trained with a batch size of 240 samples, distributed over 80 A100-80GB GPUs. For HSROIs, we apply data augmentations including multi-crop [6], Gaussian blur, solarization [23], *etc*. As for TMsI and TSARI, we randomly select a fixed-sized sequence from the original one and perform random disturbances on the RSI acquisition date. We employ the huge version of the Swin Transformer (Swin-H) [41] as the spatial encoder of HSROIs, for its design efficiency in minimizing computational costs for high-resolution imagery [74]. RSI from TMsI or TSARI is processed with corresponding ViT-L [20]. For Geo-Context Prototype Learning, we divide the globe into 4096 regions, each containing 100 prototypes.

4.2. Performance on Single-Modal Tasks

We evaluate SkySense on 4 representative single-modal tasks. All experiments are conducted using consistent fine-tuning settings for fairness. The supplementary materials include implementation details, visualization results and additional experiments on frozen backbone tuning.

Semantic Segmentation. We adopt Dyna.-Pla. [59], iSAID [67], Potsdam [52] and Dyna.-S2 [59] for the segmentation experiment. They are chosen considering factors such as spatial resolution, spectrum and category type. UperNet [69] serves as the segmentation head. For Dyna.-Pla. and Dyna.-S2 datasets, we report mIoU results on official validation and test sets. For iSAID and Potsdam, we follow the settings of [55]. As depicted in Tab. 2a, SkySense has achieved the SOTA performance on all four segmentation datasets. On average, it surpasses the previous SOTA by an impressive improvement of 1.86%.

Horizontal & Oriented Object Detection. We employ the widely recognized DIOR [36] dataset for Horizontal Object Detection and its enhanced version DIOR-R [15], along with FAIR1M [56], for Oriented Object Detection. All datasets consist of optical RGB images. Faster RCNN [51] and Oriented RCNN [37] are used for the experiment, following the setup of [55, 62]. SkySense excels on all three datasets (Tab. 2b). Notably, we surpass the second best CMID by 3.99% mAP and have achieved the best performance on the FAIR1M v2.0[1] leaderboard. More importantly, our results are accomplished without using any sophisticated Oriented Detection designs [29, 37, 71].

Change Detection. We assess SkySense's Change Detection performance on LEVIR-CD [9], OSCD [18], and Dyna.-S2 [59] datasets. For LEVIR-CD and OSCD, we follow the frameworks of [55, 66] and report the F1 metric. For Dyna.-S2, we utilize the UperNet head since the re-

[1]https://www.gaofen-challenge.com/benchmark *(2023.11.17)*

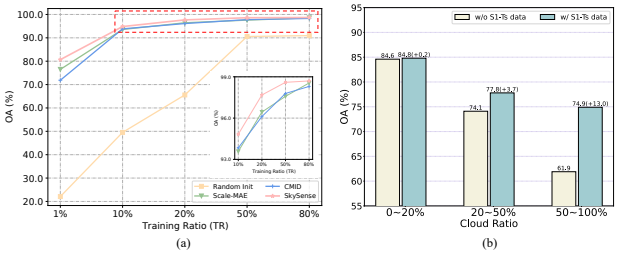

Figure 4. (a) Experiment on fine-tuning using different percentages of training data on the AID dataset. (b) The impact of S1-Ts data under varying cloud coverage conditions.

Task & Dataset	Data Source & Geo-Context	Previous SOTA	SkySense
(a) Multi-Modal Seg: Dyna.-MM	(i) Planet.	45.6 [47]	46.5 ↑ **0.9**
	(ii) Planet. with GCP	-	47.0
	(iii) S2	43.5 [4]	46.2 ↑ **2.7**
	(iv) Planet. + S2	-	47.3
	(v) Planet. + S2 + S1	-	47.7
	(vi) Planet. + S2 + S1 with GCP	-	48.2
(b) Multi-Modal Seg: PASTIS-MM	(i) S2-static	-	73.5
	(ii) S2-Ts	83.4 [58]	84.6 ↑ **1.2**
	(iii) S2-Ts + S1-Ts	84.2 [22]	84.8 ↑ **0.6**
	(iv) S2-Ts + GEP	-	85.8
	(v) S2-Ts + S1-Ts + GEP	-	85.9
(c) Multi-Modal Cls: BEN-MM	(i) S1	83.70 [64]	86.25 ↑ **2.55**
	(ii) S2 + S1	89.70 [64]	92.21 ↑ **2.51**

Table 4. Fine-tuning results on multi-modal tasks.

ported semantic change segmentation (SCS) score is calculated from the segmentation results [59]. Both results from the validation and test sets of Dyna.-S2 are presented. The remarkable generalization ability of SkySense is evident in the consistent improvements shown in Tab. 2c. Unlike CACo [45], which shows proficiency mainly on the Dyna.-S2 validation set, our model excels across all datasets.

Scene Classification. We utilize four scene classification datasets: AID [68] and RESISC-45 [14] with static RGB images, BEN-S2 [53] with static multispectral images, and fMoW-S2 [17] with temporal multispectral images. The training ratio (TR) follows [17, 46, 55]. We use a linear classifier head for experiment. For AID and RESISC-45, we report Overall Accuracy (OA), for BEN-S2 we report mAP, and for fMoW-S2 we report both Top-1 and Top-5 Accuracy. SkySense overally outperforms competitive baselines and achieves the best results on all datasets (Tab. 3). Additionally, with limited labeled data on the AID dataset, SkySense consistently outperforms CMID, Scale-MAE, and random initialization, with a 4.17% higher OA than the second best Scale-MAE using only 1% training data (Fig. 4a). These results highlight the robustness and generalization ability of SkySense's pre-trained features.

4.3. Performance on Multi-Modal Tasks

Multi-Modal Segmentation: Time-insensitive Land Cover Mapping. We employ the Dyna.-MM dataset [59] for fine-tuning and report mIoU on the official test set. The dataset comprises HSROIs from PlanetFusion (Planet.), multispectral imagery from Sentinel-2 (S2), and SAR imagery from Sentinel-1 (S1). We use a simple UperNet head. As shown in Tab. 4a, SkySense achieves the best results in single-modal scenarios (i) and (iii), clearly outperforming the previous SOTA by roughly 1% mIoU. Moreover, combining all three modalities as (v) further improves the mIoU by 1.2% compared to (i). Notably, without bells and whistles, SkySense ranks No.1 on the challenging DynamicEarthNet leaderboard[2].

Multi-Modal Segmentation: Time-sensitive Crop Mapping. We evaluate SkySense's fine-tuning result on the PASTIS-MM dataset, an enhanced version of PASTIS-R [22]. PASTIS-MM includes HSROIs from Google Earth Pro (GEP), TMsI from Sentinel-2 (S2-Ts), and TSARI from Sentinel-1 (S1-Ts). We use a naive FCN head [42] and report the OA from the official five-fold validation on PASTIS-MM dataset. In Tab. 4b, comparing S2-Ts and static multispectral data (S2-static), we observe a significant 11.1% OA increase, highlighting the importance of incorporating temporal clue for crop mapping.

Furthermore, both (ii) and (iii) exceed the performance of the previous SOTA, affirming the superior capabilities of SkySense. When more modalities are added as (ii), (iv), and (v), the OA increases accordingly. However, integrating Sentinel-1 data yields no substantial improvement, presumably because of the cloud-free imagery from PASTIS-MM dataset. To further investigate, we compare OA using S2-Ts data at different cloud ratios with and without Sentinel-1. Fig. 4b illustrates that the performance difference between utilizing and foregoing Sentinel-1 data becomes more pronounced with an increasing cloud ratio. Specifically, when the cloud ratio exceeds 50%, the result of using Sentinel-1 outperforms its counterpart by 13%. This highlights the importance of SAR data in situations with cloud coverage and rainfall.

Multi-Modal Scene Classification. We utilize the BEN-MM [54] dataset for evaluating the multi-modal scene classification task. This dataset includes both Sentinel-1 (S1) and Sentinel-2 (S2) imagery. The evaluation protocol from DeCUR [64] is followed, and we report the mAP metric of fine-tuning with 100% training data. In Tab. 4c, both (i) and (ii) significantly outperform the previous SOTA by more than 2.5% mAP. Furthermore, the inclusion of Sentinel-2 imagery greatly enhances performance compared to using Sentinel-1 imagery alone.

All these results show a notable gain for the tasks using multi-modal data, affirming the necessity of SkySense's multi-modal pre-training from one perspective.

[2]https://codalab.lisn.upsaclay.fr/competitions/2882#results*(2023.11.17)*

Pre-training	iSAID	fMoW-S2
	mIoU	Top-5 Acc
Simple Ver.	68.98	85.69
SkySense	70.91 ↑ **1.93**	87.27 ↑ **1.58**

(a)

Pre-training	Dyna.-MM
	mIoU
Baseline	42.2
+ MGCL	44.4 ↑ 2.2
+ MM	47.0 ↑ 2.6
+ CMA	47.7 ↑ 0.7
+ GCPL	48.2 ↑ 0.5

(b)

Table 5. (a) Discussion on multi-modal pre-training effectiveness. (b) Ablation study on the pre-training design.

5. Discussions & Ablation Studies

Multi-modal Pre-training Effectiveness. In addition to confirming the effectiveness of using multi-modal data in downstream tasks, we investigate the impact of multi-modal pre-training on single-modal tasks, compared with pre-training on fewer modalities. We conduct experiments on iSAID for static HSROI segmentation and fMoW-S2 for temporal multispectral classification. Two versions of the pre-trained model are tested: a simple version pre-trained only with optical imagery (HSROIs, TMsI), and SkySense, which includes HSROIs, TMsI, and TSARI for pre-training. The rest of the settings remain consistent. The results in Tab. 5a show that SkySense consistently outperforms the simple version, suggesting that the introduction of SAR data benefits representation learning of other modalities. This may attribute to the implicit clue brought by SAR data through Cross-Modal Alignment. It provides another perspective on the necessity of SkySense's multi-modal pre-training.

What does Geo-Context Prototype (GCP) Learn? We utilize the Dyna.-MM dataset for experiment as it contains diverse geo-locations worldwide. For the segmentation task in Tab. 4a, adding GCP in downstream tasks leads to a further gain of 0.5% mIoU compared to the strong multi-modal baseline `(v)`. Moreover, a comparison between `(i)` and `(ii)` shows a 0.5% mIoU improvement using GCP for the single-modal task. It demonstrates GCP's consistent performance gain in single- and multi-modal scenarios.

In Fig. 5, we visualize the learned prototypes on the Map by calculating the pre-trained feature of each pixel and assigning the most similar prototype to it. A comparison with the ESRI LandCover Map [34] reveals GCP's promising results in segmenting different areas. Moreover, GCP exhibit fine-grained advantage, as shown in the middle of Fig. 5. The prototypes learned from unsupervised clustering segment cropland within the town, which is overlooked by the LandCover Map. Notably, the visualization shares the same spatial resolution with ESRI LandCover Map.

Design of Pre-training. Tab. 5b presents the ablation study to assess our pre-training design, namely Multi-Granularity Contrastive Learning (MGCL), Multi-Modal (MM) integration, Cross-Modal Alignment (CMA) and Geo-Context Prototype Learning (GCPL). We utilize the Dyna.-MM dataset for the experiment and report mIoU metric on the official test set.

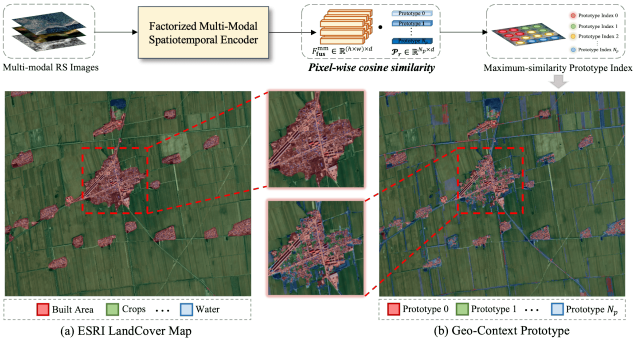

Figure 5. Comparison between (a) ESRI LandCover Map and (b) Geo-Context Prototype. The visualization process of Geo-Context Prototype is illustrated in the upper part of this figure.

Initially, we utilize a single-modal version, using g_{HR} spatial encoder and HSROIs for pre-training. The training settings are kept the same as described in Sec. 4.1. The results show that MGCL leads to a notable improvement compared to a simple baseline [7]. Then we integrate further multi-modal data (*i.e.*, TMsI and TSARI) into the pre-training and downstream evaluation. This effectively improves the performance on the test set to 47.0% mIoU, validating the necessity of multi-modal pre-training.

CMA is another necessary design for SkySense's pre-training, which explicitly pulls features from different modalities together, encouraging cross-modal interactions. The results show that the incorporation of modal alignment leads to 0.7% mIoU improvement. Finally, we introduce GCPL, which learns complementary regional context clue to facilitate downstream tasks and further pushes the very strong performance to 48.2% mIoU.

6. Conclusion & Future Work

In this paper, we present SkySense, a large-scale MM-RSFM for interpretation of EO imagery. SkySense allows using its modules flexibly to accommodate different scenarios and consistently outperforms other models on a variety of tasks, showcasing its exceptional generalization ability and strong performance. We hope SkySense will inspire further research on MM-RSFM and its release may contribute to sustainable innovations thriving in the Remote Sensing community. As part of our future work, we plan to incorporate the language modality, thereby extending SkySense's applications to more EO tasks.

7. Acknowledgment

This work was in part supported by the National Natural Science Foundation of China under Grants 42030102 and 42371321.

References

[1] Peri Akiva, Matthew Purri, and Matthew Leotta. Self-supervised material and texture representation learning for remote sensing tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8203–8215, 2022. 2, 6

[2] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10181–10190, 2021. 2, 6

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022. 2

[4] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. Satlaspretrain: A large-scale dataset for remote sensing image understanding. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16772–16782, 2023. 2, 6, 7

[5] Yinxia Cao, Xin Huang, and Qihao Weng. A multi-scale weakly supervised learning method with adaptive online noise correction for high-resolution change detection of built-up areas. *Remote Sensing of Environment*, 297:113779, 2023. 1, 2

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 4, 5, 6

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2, 4, 8

[8] Keumgang Cha, Junghoon Seo, and Taekyung Lee. A billion-scale foundation model for remote sensing images. *arXiv preprint arXiv:2304.05215*, 2023. 2, 6

[9] Hao Chen and Zhenwei Shi. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing*, 12(10):1662, 2020. 2, 6

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[11] Wuyang Chen, Ziyu Jiang, Zhangyang Wang, Kexin Cui, and Xiaoning 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*, pages 8924–8933, 2019. 2, 4, 5

[12] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Mocov2: Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[13] Gong Cheng and Junwei Han. A survey on object detection in optical remote sensing images. *ISPRS journal of photogrammetry and remote sensing*, 117:11–28, 2016. 1

[14] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 2, 7

[15] Gong Cheng, Jiabao Wang, Ke Li, Xingxing Xie, Chunbo Lang, Yanqing Yao, and Junwei Han. Anchor-free oriented proposal generator for object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022. 2, 6

[16] Mingmin Chi, Antonio Plaza, Jon Atli Benediktsson, Zhongyi Sun, Jinsheng Shen, and Yangyong Zhu. Big data for remote sensing: Challenges and opportunities. *Proceedings of the IEEE*, 104(11):2207–2219, 2016. 1

[17] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022. 2, 3, 6, 7

[18] Rodrigo Caye Daudt, Bertr Le Saux, Alexandre Boulch, and Yann Gousseau. Urban change detection for multispectral earth observation using convolutional neural networks. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 2115–2118. Ieee, 2018. 2, 6

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 6

[21] Anthony Fuller, Koreen Millard, and James R. Green. Croma: Remote sensing representations with contrastive radar-optical masked autoencoders. *Advances in Neural Information Processing Systems*, 2023. 3, 6

[22] Vivien Sainte Fare Garnot, Loic Landrieu, and Nesrine Chehata. Multi-modal temporal attention models for crop mapping from satellite time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 187:294–305, 2022. 2, 7

[23] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 2, 6

[24] Shaohua Guo, Liang Liu, Zhenye Gan, Yabiao Wang, Wuhao Zhang, Chengjie Wang, Guannan Jiang, Wei Zhang, Ran Yi, Lizhuang Ma, et al. 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*, pages 4361–4370, 2022. 2, 4, 5

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[26] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 4, 5

[27] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2

[28] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Xiuping Jia, Antonio Plaza, Gamba Paolo, Jon Atli Benediktsson, and Jocelyn Chanussot. Spectralgpt: Spectral foundation model. *arXiv preprint arXiv:2311.07113*, 2023. 6

[29] Liping Hou, Ke Lu, Jian Xue, and Yuqiu Li. Shape-adaptive selection and measurement for oriented object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 923–932, 2022. 6

[30] Jingliang Hu, Lichao Mou, and Xiao Xiang Zhu. Unsupervised domain adaptation using a teacher-student network for cross-city classification of sentinel-2 images. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43:1569–1574, 2020. 5

[31] Lanqing Huang, Bin Liu, Boying Li, Weiwei Guo, Wenhao Yu, Zenghui Zhang, and Wenxian Yu. Opensarship: A dataset dedicated to sentinel-1 ship interpretation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(1):195–208, 2017. 2

[32] Xin Huang, Yihong Song, Jie Yang, Wenrui Wang, Huiqun Ren, Mengjie Dong, Yujin Feng, Haidan Yin, and Jiayi Li. Toward accurate mapping of 30-m time-series global impervious surface area (gisa). *International Journal of Applied Earth Observation and Geoinformation*, 109:102787, 2022. 2, 4, 5

[33] Johannes Jakubik, Sujit Roy, C. E. Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Nyirjesy Gabby, Blair Edwards, Daiki Kimura, Naomi Simumba, Linsong Chu, S. Karthik Mukkavilli, Devyani Lambhate, Kamal Das, Ranjini Bangalore, Dario Oliveira, Michal Muszynski, Kumar Ankur, Muthukumaran Ramasubramanian, Iksha Gurung, Sam Khallaghi, Hanxi Li, Michael Cecil, Maryam Ahmadi, Fatemeh Kordi, Hamed Alemohammad, Manil Maskey, Raghu Ganti, Kommy Weldemariam, and Rahul Ramachandran. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660*, 2023. 2

[34] Krishna Karra, Caitlin Kontgis, Zoe Statman-Weil, Joseph C Mazzariello, Mark Mathis, and Steven P Brumby. Global land use/land cover with sentinel 2 and deep learning. In *2021 IEEE international geoscience and remote sensing symposium IGARSS*, pages 4704–4707. IEEE, 2021. 8

[35] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 5

[36] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307, 2020. 2, 6

[37] Wentong Li, Yijie Chen, Kaixuan Hu, and Jianke Zhu. Oriented reppoints for aerial object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1829–1838, 2022. 6

[38] Xue Li, Guo Zhang, Hao Cui, Shasha Hou, Shunyao Wang, Xin Li, Yujia Chen, Zhijiang Li, and Li Zhang. Mcanet: A joint semantic segmentation framework of optical and sar images for land use classification. *International Journal of Applied Earth Observation and Geoinformation*, 106: 102638, 2022. 2

[39] Yansheng Li, Wei Chen, Xin Huang, Gao Zhi, Siwei Li, Tao He, and Yongjun Zhang. Mfvnet: a deep adaptive fusion network with multiple field-of-views for remote sensing image semantic segmentation. *Science China Information Sciences*, 2023. 2

[40] Yinhe Liu, Sunan Shi, Junjue Wang, and Yanfei Zhong. Seeing beyond the patch: Scale-adaptive semantic segmentation of high-resolution remote sensing imagery based on reinforcement learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16868–16878, 2023. 2, 4, 5

[41] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2, 6

[42] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 7

[43] ZhiYong Lv, HaiTao Huang, Xinghua Li, MingHua Zhao, Jon Atli Benediktsson, WeiWei Sun, and Nicola Falco. Land cover change detection with heterogeneous remote sensing images: Review, progress, and perspective. *Proceedings of the IEEE*, 2022. 1

[44] Rose M Rustowicz, Robin Cheong, Lijing Wang, Stefano Ermon, Marshall Burke, and David Lobell. Semantic segmentation of crop type in africa: A novel dataset and analysis of deep learning methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 75–82, 2019. 2

[45] Utkarsh Mall, Bharath Hariharan, and Kavita Bala. Change-aware sampling and contrastive learning for satellite images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5261–5270, 2023. 2, 6, 7

[46] Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9414–9423, 2021. 2, 6, 7

[47] Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, Chen Chen, and Mu Li. Towards geospatial foundation models via continual pretraining. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16806–16816, 2023. 2, 3, 6, 7

[48] Dilxat Muhtar, Xueliang Zhang, Pengfeng Xiao, Zhenshi Li, and Feng Gu. Cmid: A unified self-supervised learning framework for remote sensing image understanding. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 2, 3, 6

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[50] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4088–4099, 2023. 2, 3, 6

[51] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 6

[52] Jamie Sherrah. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. *arXiv preprint arXiv:1606.02585*, 2016. 6

[53] Gencer Sumbul, Jian Kang, Tristan Kreuziger, Filipe Marcelino, Hugo Costa, Pedro Benevides, Mario Caetano, and Begüm Demir. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IEEE International Geoscience and Remote Sensing Symposium*, pages 5901–5904, 2019. 2, 7

[54] Gencer Sumbul, Arne de Wall, Tristan Kreuziger, Filipe Marcelino, Hugo Costa, Pedro Benevides, Mario Caetano, Begüm Demir, and Volkerl Mark. BigEarthNet-MM: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval. *IEEE Geoscience and Remote Sensing Magazine*, 9(3):174–180, 2021. 7

[55] Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang, Hao Chang, et al. Ringmo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 2022. 2, 6, 7

[56] Xian Sun, Peijin Wang, Zhiyuan Yan, Feng Xu, Ruiping Wang, Wenhui Diao, Jin Chen, Jihao Li, Yingchao Feng, Tao Xu, et al. Fair1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 184:116–130, 2022. 2, 6

[57] Chao Tao, Ji Qi, Guo Zhang, Qing Zhu, Weipeng Lu, and Haifeng Li. Tov: The original vision model for optical remote sensing image understanding via self-supervised learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023. 2, 6

[58] Michail Tarasiou, Erik Chavez, and Stefanos Zafeiriou. Vits for sits: Vision transformers for satellite image time series. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10418–10428, 2023. 7

[59] Aysim Toker, Lukas Kondmann, Mark Weber, Marvin Eisenberger, Andrés Camero, Jingliang Hu, Ariadna Pregel Hoderlein, Çağlar Şenaras, Timothy Davis, Daniel Cremers, et al. Dynamicearthnet: Daily multi-spectral satellite dataset for semantic change segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21158–21167, 2022. 2, 6, 7

[60] Devis Tuia, Konrad Schindler, Begüm Demir, Gustau Camps-Valls, Xiao Xiang Zhu, Mrinalini Kochupillai, Sašo Džeroski, Jan N van Rijn, Holger H Hoos, Fabio Del Frate, et al. Artificial intelligence to advance earth observation: a perspective. *arXiv preprint arXiv:2305.08413*, 2023. 1

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[62] Di Wang, Qiming Zhang, Yufei Xu, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. Advancing plain vision transformer toward remote sensing foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2022. 2, 6

[63] Di Wang, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. Scaling-up remote sensing segmentation dataset with segment anything model. *Advances in Neural Information Processing Systems*, 2023. 2, 6

[64] Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Chenying Liu, Zhitong Xiong, and Xiao Xiang Zhu. Decur: decoupling common & unique representations for multimodal self-supervision. *arXiv preprint arXiv:2309.05300*, 2023. 3, 7

[65] Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M Albrecht, and Xiao Xiang Zhu. Ssl4eo-s12: A large-scale multi-modal, multi-temporal dataset for self-supervised learning in earth observation. *IEEE Geoscience and Remote Sensing Magazine*, 11(3):98–106, 2023. 2, 6

[66] Xinye Wanyan, Sachith Seneviratne, Shuchang Shen, and Michael Kirley. Dino-mc: Self-supervised contrastive learning for remote sensing imagery with multi-sized local crops. *arXiv preprint arXiv:2303.06670*, 2023. 2, 6

[67] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–37, 2019. 2, 6

[68] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017. 2, 7

[69] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 6

[70] Binbin Yang, Xinchi Deng, Han Shi, Changlin Li, Gengwei Zhang, Hang Xu, Shen Zhao, Liang Lin, and Xiaodan Liang. Continual object detection via prototypical task correlation guided gating mechanism. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9255–9264, 2022. 1

[71] Yi Yu and Feipeng Da. Phase-shifting coder: Predicting accurate orientation in oriented object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13354–13363, 2023. 6

[72] Xiaohui Yuan, Jianfang Shi, and Lichuan Gu. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Systems with Applications*, 169: 114417, 2021. 1

[73] Hankui K Zhang, David P Roy, and Dong Luo. Demonstration of large area land cover classification with a one dimensional convolutional neural network applied to single pixel temporal metric percentiles. *Remote Sensing of Environment*, 295:113653, 2023. 2

[74] Pengchuan Zhang, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, and Jianfeng Gao. Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2998–3008, 2021. 6

[75] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11656–11665, 2021. 2

[76] Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*, 2023. 1

[77] Stefano Zorzi, Shabab Bazrafkan, Stefan Habenschuss, and Friedrich Fraundorfer. Polyworld: Polygonal building extraction with graph neural networks in satellite images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1848–1857, 2022. 1