

# SkySense V2: A Unified Foundation Model for Multi-modal Remote Sensing

Yingying Zhang<sup>1</sup> Lixiang Ru<sup>1</sup> Kang Wu<sup>2</sup> Lei Yu<sup>1</sup> Lei Liang<sup>1</sup> Yansheng Li<sup>2</sup> Jingdong Chen<sup>1</sup>

<sup>1</sup>Ant Group <sup>2</sup>Wuhan University

qichu.zyy@antgroup.com

## Abstract

The multi-modal remote sensing foundation model (MM-RSFM) has significantly advanced various Earth observation tasks, such as urban planning, environmental monitoring, and natural disaster management. However, most existing approaches generally require the training of separate backbone networks for each data modality, leading to redundancy and inefficient parameter utilization. Moreover, prevalent pre-training methods typically apply self-supervised learning (SSL) techniques from natural images without adequately accommodating the characteristics of remote sensing (RS) images, such as the complicated semantic distribution within a single RS image. In this work, we present SkySense V2, a unified MM-RSFM that employs a single transformer backbone to handle multiple modalities. This backbone is pre-trained with a novel SSL strategy tailored to the distinct traits of RS data. In particular, SkySense V2 incorporates an innovative adaptive patch merging module and learnable modality prompt tokens to address challenges related to varying resolutions and limited feature diversity across modalities. In addition, we incorporate the mixture of experts (MoE) module to further enhance the performance of the foundation model. SkySense V2 demonstrates impressive generalization abilities through an extensive evaluation involving 16 datasets over 7 tasks, outperforming SkySense by an average of 1.8 points.

## 1. Introduction

Advancements in multi-modal remote sensing foundation models (MM-RSFM) have significantly enhanced the utilization of remote sensing (RS) data across numerous Earth observation (EO) applications[5, 12, 47, 83], such as environmental monitoring, agriculture management, natural disaster response, land cover (or use) mapping etc. Recently, Guo *et al.* introduced SkySense [24], the largest MM-RSFM to date. SkySense has demonstrated impressive generalization capabilities through extensive evaluations across many different tasks. A key factor in its success is the simultaneous training of geographically aligned multi-modal

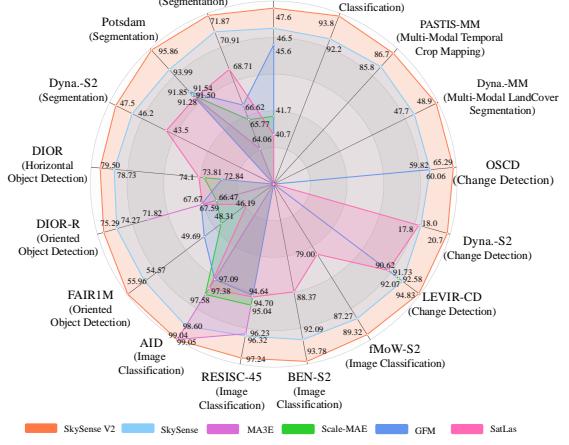


Figure 1. SkySense V2 has achieved superior performance on 16 datasets over 7 distinct tasks compared with its predecessor SkySense and supports a broad range of tasks.

and multi-temporal RS data. To accommodate the varying resolutions of different modalities, SkySense employs the Swin Transformer [41] for processing high-resolution (HR) optical images, while using the Vision Transformer (ViT) [18] for medium-resolution multi-spectral (MS) data and synthetic aperture radar (SAR) data. However, this separate backbone design results in redundancy and inefficient use of parameters. On the other hand, SkySense’s pre-training approach is primarily based on DINOv2 [53], renowned for its efficiency in self-supervised learning (SSL). Although it has demonstrated strong performance with RS images, there remains considerable room for improvement by adequately accommodating the involved semantic distribution within a single RS image.

In this paper, we introduce SkySense V2, an advanced MM-RSFM pre-trained through a novel SSL approach tailored to the distinctive traits of RS data. To be specific, SkySense V2 leverages a unified transformer backbone capable of processing data across diverse modalities and resolutions. It is pre-trained using a new SSL strategy that utilizes query-based attention to gather the similar semantic features distributed in different regions of RS images.

Designing a unified backbone for geographically aligned multi-modal data presents two primary challenges. The first is processing feature resolutions across data with varying ground sample distance (GSD) during simultaneous pre-training of multiple modalities. To address this challenge, we introduce an innovative Adaptive Patch Merging (APM) module, which is integrated after each stage of the unified backbone. This module determines whether to reduce feature resolution according to the specific feature resolution requirements of each modality. For instance, when dealing with HR optical images, the APM module activates resolution reduction at each stage. In contrast, for medium-resolution data, such as MS and SAR data, the feature resolution is preserved throughout the stages. This design enables the backbone to efficiently process data of different resolutions while adapting output feature resolutions according to actual requirements, which is crucial for the joint training and fusion of multi-modal features.

The second challenge is that fully sharing parameters across the different modalities can reduce the feature diversity. To address this, inspired by the work of [16] and [35], we introduce learnable modality prompt tokens for each modality. By interacting with these individual modal prompt tokens through an attention mechanism, the pre-trained model can capture the unique characteristics of each modality better. Compared to SkySense, our unified design significantly enhances the efficiency of parameter utilization. To be specific, SkySense uses three separate backbones, Swin-H for HR optical data, ViT-L for MS data, and ViT-L for SAR data, collectively totaling 1.26 billion parameters. In contrast, SkySense V2’s unified transformer backbone is capable of simultaneously processing all three modalities while utilizing only 665 million parameters. This efficient use of parameters not only optimizes the architecture but also allows for potential model scaling to further boost performance. Therefore, we integrate the mixture of experts module [33], a technique commonly used in various large language models [19, 59, 86] to further enhance the performance of the foundation model.

To effectively adapt SSL techniques developed for natural images to the pre-training of RSFM, it is essential to consider the distinct differences in data characteristics. Natural images typically feature a single, clear subject, such as a cat, dog, or person. In contrast, RS images comprise multiple subjects distributed across various regions of the image. For instance, as illustrated in Figure 2, one section of an RS image may show a **building**, while other areas might show a **forest**, **pond**, or **land**. Current SSL methods perform contrastive learning directly between different views using various augmented crops, which is effective for natural images due to their singular focus. However, when applied to RS images, this approach can result in semantic inaccuracies, as different views may capture different subjects. For

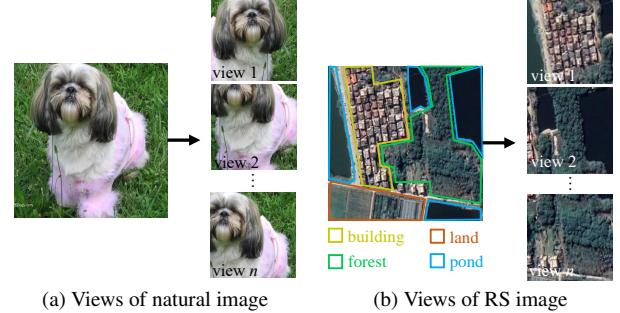


Figure 2. Comparison of different augmented views from natural and RS images in SSL. Natural image views often concentrate on a singular subject, such as a “dog”. RS image views encompass a variety of subjects, including building, forest, pond, and land.

example, in Figure 2b, view 1 of the RS image primarily contains **building**, whereas view 2 shows **forest** and **pond**.

To tackle this challenge, we propose a novel approach called Query-based Semantic Aggregation Contrastive Learning (QSACL). This method employs multiple learnable queries to perform cross-attention with features from different views, generating semantically aggregated features. We then apply contrastive learning to these aggregated features pairs derived from the same query. As demonstrated in Figure 9 from our ablation study, utilizing different queries can aggregate consistent semantic features across multiple views, which enhances the accuracy of contrastive learning.

We evaluated SkySense V2 on a diverse set of 16 datasets [9, 13–15, 17, 21, 36, 62, 65, 67, 76, 79], covering a range of task types, modalities, and spatial scales. As illustrated in Figure 1, SkySense V2 demonstrates substantial improvements in performance over its predecessor, SkySense, achieving state-of-the-art (SOTA) results across various modalities of EO tasks. The experimental results across all test scenarios highlight its competitive advantage over existing RSFMs in a wide array of EO interpretation tasks.

## 2. Related Work

### 2.1. Remote Sensing Foundation Model

RSFMs are characterized by their ability to leverage vast amounts of data through self-supervised learning (SSL) technologies, enabling them to learn robust feature representations without the need for extensive annotations. The predominant RSFMs utilize Contrastive Learning (CL) or Masked Image Modeling (MIM). Typical works in CL include RS-BYOL [34], GASSL [82], DINO-MC [75], SeCo [49], and CACo [48]. Other research efforts focus on enhancing MIM framework, such as RingMo [64], S2MAE [38], MA3E [39], SatMAE [15], and SatMAE++ [52]. Additionally, studies like CMID [51] and GFM [50] explore the intersection of CL and MIM through a self-distillation

approach. Recently, Guo *et al.* introduced a comprehensive MM-RSFM known as SkySense [24], which features a factorized multi-modal spatiotemporal encoder. This architecture facilitates independent spatial feature extraction and multi-modal temporal fusion. As the largest MM-RSFM to date, SkySense exhibits exceptional generalization capabilities across a wide range of RS datasets. However, it does face a challenge: the inefficient utilization of parameters due to the separate backbone designs for different modalities. In this work, we propose SkySense V2 to address this issue by employing a unified transformer backbone design.

## 2.2. Unified Framework for Multi-modal Learning

Multi-modal learning aims to train models that effectively process and relate information from various modalities. Substantial advancements have been made in this field, culminating in numerous studies focused on creating unified networks for the integration and processing of diverse modalities. The main work includes VLMO [60], Meta-Transformer [84], Uni-Perceiver [85], and UniTR [25], *etc.* There are several methods that utilize a unified backbone design for multi-modal RS data. For example, OFA-Net [81] pre-trains a single transformer backbone on a curated multi-modal dataset using MIM. Han *et al.* proposed a RSFM called msGFM[26], which utilizes a shared transformer encoder while employing different decoders for each modality. In contrast to our approach, both OFA-Net and msGFM simply share all of parameters of backbone and learn feature representations for multi-modal data separately, which results in a lack of geographical alignment and fusion between features from different modalities. AnySat [2] employs a joint embedding predictive architecture for a multi-modal model, training a single model on heterogeneous data in a SSL manner. However, the backbone design and SSL pre-training methods of our SkySense V2 differ significantly from AnySat.

## 3. SkySense V2

### 3.1. Model Architecture

#### 3.1.1. Unified Transformer Backbone

As illustrated in Figure 3, our unified transformer backbone is a hierarchical encoder structure with four stages. In the first two stages, we employ Swin Transformer V2 Blocks (SwinV2B) [42] to incorporate essential visual priors, such as locality and translation invariance. The window-based self-attention mechanisms in SwinV2B also help reduce computational complexity compared to global self-attention, particularly given the high spatial resolution of features in these initial stages. In the last two stages, we utilize vanilla Transformer Blocks (TB) [18] for two main reasons: First, the spatial resolution of features in these stages is relatively low, making the computational

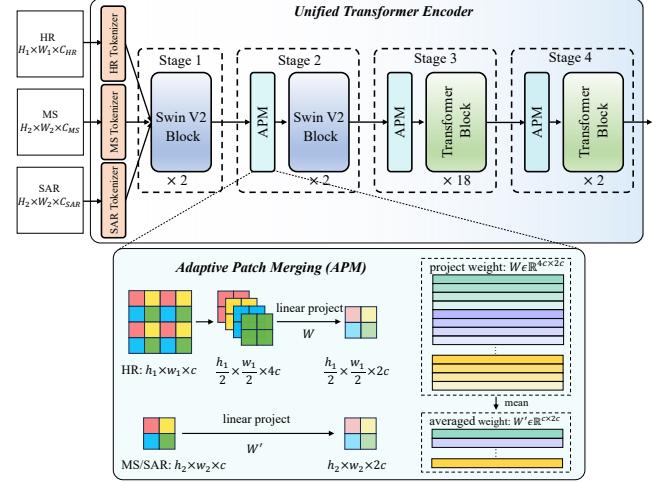


Figure 3. Overview of the unified transformer backbone model in SkySense V2: The entire backbone shares all parameters across different modalities, with the exception of distinct tokenizers.

costs of global self-attention more manageable. Second, our ablation study demonstrates that global self-attention can be complementary to window-based self-attention, enabling the model to achieve stronger representations.

Given a group of multi-modal inputs consisting of a high-resolution optical image  $x_{HR}$ , multi-spectral data  $x_{MS}$ , and synthetic aperture radar data  $x_{SAR}$  where each 'pixel' in the different modalities is naturally aligned by geo-location, we employ three distinct tokenizers to process these inputs into tokens. In each tokenizer, the input data is first divided into non-overlapping  $4 \times 4$  patches. Subsequently, a linear embedding layer is applied to this raw, patched data to project it into patch tokens. Subsequently, four stages of either SwinV2B or TB, with shared parameters across modalities, are applied to these patch tokens.

Due to the varying GSD of satellite sensors across different modalities, input data corresponding to the same region at specific geo-locations during joint aligned training exhibit differing spatial resolutions. To harmonize these spatial resolutions for tokens from various modalities, we propose Adaptive Patch Merging (APM) and utilize it to selectively reduce spatial resolution at each stage, except for Stage 1. Specifically, for the tokens of optical image data with high spatial resolution, APM concatenates the features from groups of  $2 \times 2$  neighboring patches and applies a linear layer to the  $4c$ -dimensional concatenated features. This process reduces the number of tokens by a factor of  $2 \times 2 = 4$  downsampling of spatial resolution, while the output dimension is set to  $2c$ . In contrast, for the tokens of MS and SAR data with lower spatial resolution, APM maintains the resolution by applying a linear projection with averaging the weights across the input dimensions. By integrating the APM module, our unified backbone can efficiently process

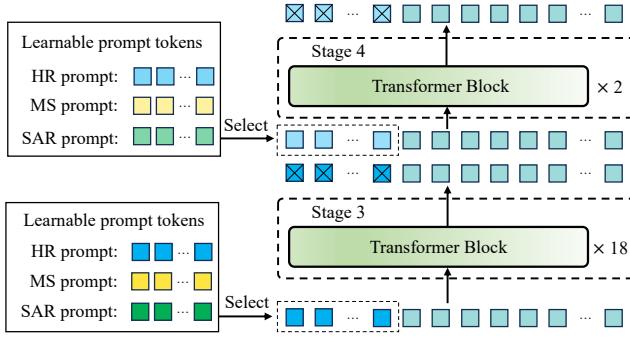


Figure 4. The illustration of modality-specific prompt tokens added to the unified backbone.

multi-modal data with varying resolutions while maintaining the spatial alignment of features across modalities

### 3.1.2. Modality-specific Prompt Tokens

Our unified transformer backbone utilizes fully shared parameters across different modalities. To enhance feature diversity, we introduce learnable modality-specific prompt tokens for each modality. By interacting with these individual modal prompt tokens through an attention mechanism, the pre-trained model can better capture the unique characteristics of each modality. As illustrated in Figure 4, we incorporate  $N$  learnable prompt tokens for each modality in the last two stages. We denote the input tokens of each stage as  $E_i^j \in \mathbb{R}^{h_j w_j \times c_j}$ ,  $i \in \{HR, MS, SAR\}$ ,  $j \in \{3, 4\}$ , where  $h_j$  and  $w_j$  are the height and width of the spatial resolution at stage  $j$ , while  $c_j$  refers to the dimension of the tokens. For each modality  $i$ , we insert the modality-specific learnable prompt tokens  $P_i^j \in \mathbb{R}^{N \times c_j}$  into the input of the first block in  $j$ -th stage  $\mathcal{F}_j$ . At the last block of each stage, these prompt tokens are discarded. The entire process can be formulated as:

$$\begin{aligned} [E_{drop}, E_i^4] &= \mathcal{F}_3([P_i^3, E_i^3]), \\ [E_{drop}, E_i^{out}] &= \mathcal{F}_4([P_i^4, E_i^4]). \end{aligned} \quad (1)$$

Here,  $E_{drop}$  denotes the discarded tokens, and  $E_i^{out}$  represents the final output tokens of backbone. Our design of modality-specific prompts enhances feature diversity while maintaining full parameter sharing by incorporating only a small number of modality-specific parameters.

### 3.1.3. Scaling Up Model Capacity

Our unified backbone design across various modalities significantly enhances parameter utilization efficiency. To further boost the model’s capabilities, we integrate a Mixture of Experts (MoE) [33] into the backbone network. We choose MoE instead of simply increasing the backbone’s width and depth because MoE enables models to be pre-trained with substantially less computation by leveraging

sparse feed-forward layers (*i.e.*, experts) for individual tokens. Following common approaches [43, 78], we insert MoE modules into the last  $L$  transformer blocks, replacing the original feed-forward network (FFN) layers. Each MoE module consists of  $M$  experts (denoted as  $\mathcal{E}_i(\cdot)$ ,  $i = 1, 2, \dots, M$ ), which share the same structure as the FFN and operate as independent networks. For the gating network in MoE, we employ a learnable linear layer followed by a Softmax function,  $\mathcal{G}(x) = \text{Softmax}(Wx)$ , where  $W$  is the gating parameter. Finally, the output of the MoE is computed as a linear combination of the outputs from the selected experts, weighted by the corresponding gate values. This process is formulated as follows:

$$MOE(x) = \sum_{i \in \mathcal{T}} \mathcal{G}_i(x) \cdot \mathcal{E}_i(x), \quad (2)$$

$\mathcal{T}$  represent the set of the top- $k$  indices. In our SkySense V2 backbone, we set  $L = 6$ ,  $M = 8$ , and  $k = 1$ .

## 3.2. Overall Pre-training Pipeline

Figure 5 provides an overview of our pre-training procedure. The pre-training framework of SkySense V2 primarily adopts the teacher-student architecture from SkySense [24], where the teacher network’s parameters are updated using an exponential moving average (EMA) [53] of the student network’s parameters. To train SkySense V2, we employ the Multi-Granularity Contrastive Learning (MGCL) loss  $\mathcal{L}_{MGCL}$  as proposed in SkySense. For each modality, MGCL utilizes contrastive loss [53] to align the representations of teacher and student networks at pixel, object and image-level granularity, respectively. Additionally, we incorporate the unsupervised Geo-Context Prototype Learning (GCPL), as validated in SkySense, to enhance the learning of complementary regional context clues, aiding downstream tasks. To further enhance dense interpretation capabilities, we introduce an auxiliary supervision loss  $\mathcal{L}_{ITA}$  to employ dense Image-Text Alignment (ITA) according to OpenStreetMap (OSM) labels<sup>1</sup>. Detailed implementations of MGCL, GCPL, and ITA are included in the Appendix A. In following paragraph, we will focus on proposed query-based semantic aggregation contrastive learning (QSACL).

**Query-based Semantic Aggregation Contrastive Learning.** QSACL utilizes learnable queries to aggregate similar semantics across different regions of images, enabling more accurate contrastive learning. Given features  $g_1, g_2$  from two global views and  $l_1, l_2, \dots, l_n$  from  $n$  local views (where  $g$  and  $l$  denotes the fused features  $F_{fus}$  from global and local view), we employ a transformer decoder layer to perform cross-attention between  $m$  learnable queries  $q_1, q_2, \dots, q_m$  and the feature sets  $g$  and  $l$ . This process generates sets of globally and locally aggregated features set

<sup>1</sup><https://www.openstreetmap.org/>

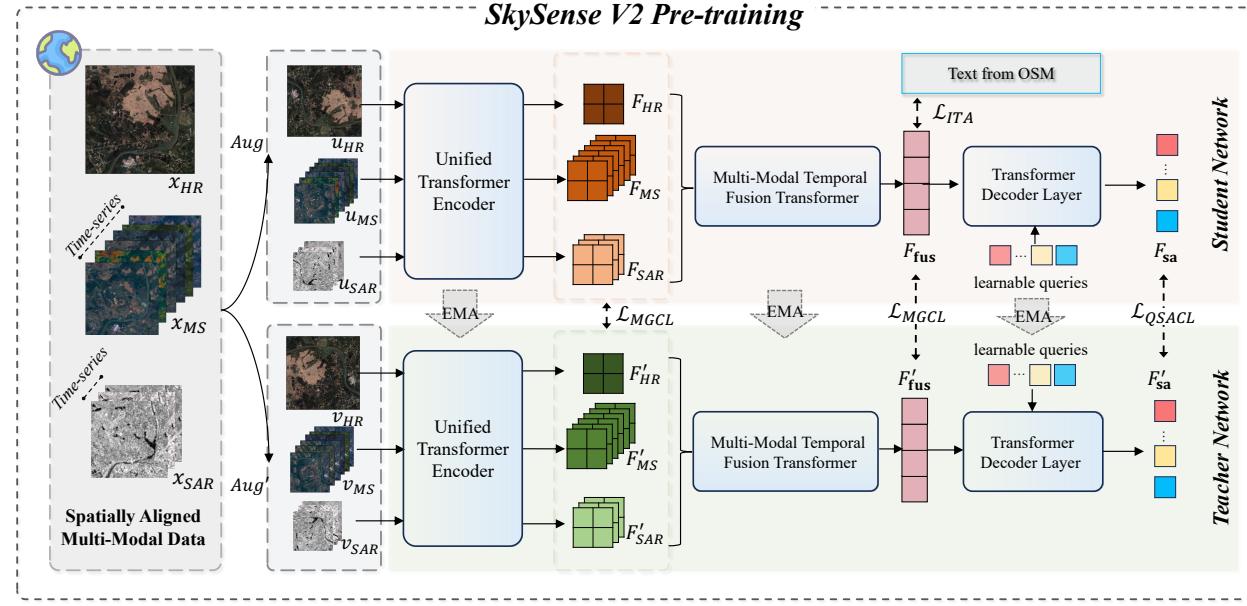


Figure 5. Overview of the SkySense V2 pre-training pipeline. The SkySense V2 employs data augmentations on multi-modal inputs, which are then fed into both the student and teacher networks. The pre-training process incorporates Multi-Granularity Contrastive Learning, Dense Image-Text Alignment, and Query-based Semantic Aggregation Contrastive Learning to effectively train the network. Once pre-training is complete, the parameters from the teacher branch are utilized for downstream applications.

$z_i^g$  and  $z_i^l$  for each  $i = 1, 2, \dots, m$ . As is shown in Figure 6, we use superscript  $'$  denotes features from the teacher network, the QSACL loss  $\mathcal{L}_{QSACL}$  is computed by averaging the loss  $\mathcal{L}_{CL}$  of aggregation features between local and global views for each query:

$$\mathcal{L}_{QSACL} = \frac{1}{2m} \sum_{i=1}^m (\mathcal{L}_{CL}(z_i^g, z_i^l) + \mathcal{L}_{CL}(z_i^l, z_i^{g'})), \quad (3)$$

$$\mathcal{L}_{CL}(x, x') = -\mathcal{H}(x) \log(\mathcal{H}'(x')). \quad (4)$$

Here,  $\mathcal{H}$  and  $\mathcal{H}'$  denotes the learning head and corresponding EMA part as defined in [7]. The overall training objective is the weighted sum of  $\mathcal{L}_{MGCL}$ ,  $\mathcal{L}_{ITA}$ , and  $\mathcal{L}_{QSACL}$ , i.e.,

$$\mathcal{L} = \lambda_1 \mathcal{L}_{MGCL} + \lambda_2 \mathcal{L}_{ITA} + \lambda_3 \mathcal{L}_{QSACL}. \quad (5)$$

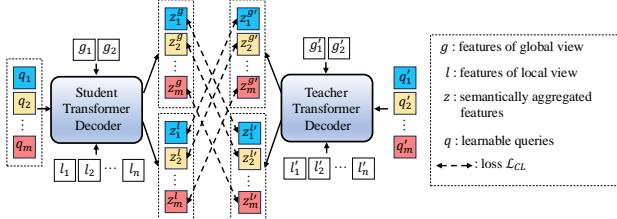


Figure 6. Computation pipelines for QSACL with multi-crop augmentation (2 global and  $n$  local views).

## 4. Experiments

Following SkySense [24], we conducted experiments on 16 datasets across various modalities and tasks to thoroughly evaluate SkySense V2. The applications of pre-trained SkySense V2 to diverse tasks are detailed in Appendix B. Each module is engineered for combined or individual use, providing the flexibility to be frozen or fine-tuned as needed.

### 4.1. Pre-training Implementation

**Training Datasets.** We use the same training dataset of SkySense [24] to train SkySense V2. This extensive dataset comprises approximately 21 million multi-modal RS imagery sets. Each set comprises three components: a high-resolution optical RGB image of spatial dimension  $2048 \times 2048$ , a sequence of Sentinel-2 (S2) multi-spectral (MS) images with 10 spectral bands and spatial dimensions of  $64 \times 64$  (average sequence length of 65 images), and a sequence of Sentinel-1 (S1) synthetic aperture radar (SAR) images with 2 bands, also having spatial dimensions of  $64 \times 64$  (average sequence length of 10 images). In each training iteration, we randomly sample 10 S2 images and 5 S1 images of different time series to accommodate the GPU memory constraints.

**Setting of Unified Backbone.** We initially set the token dimension to  $C = 352$ . In the first two stages, the win-

dow size for the Swin Transformer V2 Blocks (SwinV2B) is configured to 8. Each head has a query dimension of 32, and each MLP layer consists of two full-connected layers with dimensions expanding by a factor of 4 consistently across all blocks within the unified backbone. Prompt tokens are incorporated during the last two stages, with the number of prompt tokens being the same for each modality. We allocate 4 prompt tokens for stage-3 and stage-4.

**Pre-training Settings.** SkySense V2 is trained using a batch size of 1024, distributed across 128 H20 GPUs. The model undergoes a total of 600k iterations, utilizing the AdamW optimizer [46]. The learning rate is initially set to  $2 \times 10^{-4}$  and decays to  $1 \times 10^{-6}$  following a cosine annealing schedule [45]. Additional implementation details can be found in the Appendix C.

## 4.2. Single-Modal Tasks

We evaluate SkySense V2 on four common single-modal RS interpretation tasks: scene classification, object detection, semantic segmentation, and change detection. The experiments are conducted via fine-tuning a pre-trained SkySense V2 backbone along with a task-specific head. Further implementation details can be found in the Appendix D.

**Scene Classification.** In Table 1, we report the performance of scene classification on four widely used datasets: AID [79] and RESISC-45 [13], both featuring static RGB optical images; BEN-S2 [62], which includes static MS images; and fMoW-S2 [15], which contains temporal MS images. For each dataset, the performance is evaluated using different training ratios (TR) following the methodology [15, 49]. In our experiments, we employ a simple linear classifier as the classification head. For multi-label datasets, we use the mean average precision (mAP) as the evaluation metric, whereas for single-label datasets, we report the overall accuracy (OA). As illustrated in Table 1, SkySense V2 consistently outperforms previous RSFMs across various datasets and training configurations, particularly in low TR settings. This demonstrates that SkySense V2 exhibits a stronger representation capability compared to other RSFMs, including its predecessor, SkySense [24].

**Semantic Segmentation.** In Table 2a, we report the segmentation results from four representative semantic segmentation datasets: iSAID [76] and Potsdam [61] for high-resolution optical image segmentation, and Dyna.-Pla [67] and Dyna.-S2 [67] for multi-spectral and multi-temporal image segmentation. Following previous work, we use the mean F1-score (mF1) as the evaluation metric for the Potsdam dataset. While for the other datasets, we report the mean intersection over union (mIoU). All segmentation experiments employ the UperNet [80] as the decoder head.

Models	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 [3]	93.55/95.92	90.86/93.06	79.24/87.40	50.69/77.99	
SeCo [49]	93.47/95.99	89.64/92.91	82.62/87.81	51.65/77.40	
SatMAE [15]	95.02/96.94	91.72/94.10	86.18/89.50	63.84/-	
RingMo [64]	96.90/98.34	94.25/95.67	-	-	
RVSA [70]	97.03/98.50	93.93/95.69	-	-	
DINO-MC [75]	-	-	84.20/88.75	60.16/83.49	
TOV [66]	95.16/97.09	90.97/93.79	-	-	
SSL4EO [73]	91.06/94.74	87.60/91.27	87.10/91.80	51.70/76.77	
CMID [51]	96.11/97.79	94.05/95.53	-	-	
CACo [48]	90.88/95.05	88.28/91.94	81.30/87.00	50.72/76.31	
CROMA [20]	-	-	88.29/-	63.59/-	
SatLas [4]	94.96/97.38	92.16/94.70	82.80/88.37	57.95/79.00	
GFM [50]	95.47/97.09	92.73/94.64	86.30/-	-	
Scale-MAE [55]	96.44/97.58	92.63/95.04	-	-	
MA3E [39]	-/99.04	-/96.23	-	-	
SatMAE++ [52]	-	-/97.48	85.11/-	63.23/-	
SkySense [24]	97.68/98.60	94.85/96.32	88.67/92.09	64.38/87.27	
<b>SkySense V2</b>	<b>98.34/99.05</b>	<b>96.42/97.24</b>	<b>89.13/93.78</b>	<b>66.65/89.32</b>	

Table 1. Scene classification results. - means the task is not supported or the value is unavailable in the paper.

As illustrated in Table 2a, our SkySense V2 achieve higher performance than recent RSFMs across all four datasets. Specifically, compared to the previous SOTA method SkySense [24], SkySense V2 achieves 1.5% higher performance on average.

**Horizontal & Oriented Object Detection.** We utilize the DIOR dataset [36] to evaluate SkySense V2’s performance in horizontal object detection, while utilizing the DIOR-R [14] and FAIR1M [65] datasets for oriented object detection. These datasets only consist of RGB RS imagery. Consistent with previous research [24], we adopt Faster R-CNN [56] and Oriented R-CNN [37] as the basic detectors for horizontal and oriented object detection, respectively. As reported in Table 2b, our SkySense V2 outperforms previous RSFMs by a notable margin. Specifically, SkySense V2 surpass the previous SOTA, SkySense [24], by an average of 1.1% mAP.

**Change Detection.** For change detection, we conducted experiments on LEVIR-CD [9], OSCD [17] and Dyna.-S2 [67] datasets. Since LEVIR-CD and OSCD datasets only contain binary change detection results, we use the basic change detector in [10] and report the F1 score as the evaluation metric. While Dyna.-S2 focuses on semantic change detection, we use UperNet [80] as the segmentation network and calculate the semantic change detection score (SCS) on validation and test set following [24]. As presented in Table 2c, SkySense V2 consistently achieves higher performance than the previous SOTA method SkySense on all 3 datasets, and surpasses SkySense by 2.7% on average. Particularly, on the OSCD dataset, our SkySense V2 remark-

Models	Publication	Dyna.-Pla.	iSAID	Potsdam	Dyna.-S2
		mIoU	mIoU	mF1	mIoU
GASSL [3]	ICCV'21	34.0/40.8	65.95	91.27	28.1/41.0
SeCo [49]	ICCV'21	-	57.20	89.03	29.4/39.8
SatMAE [15]	NIPS'22	32.8/39.9	62.97	90.63	30.1/38.7
RingMo [64]	TGRS'22	-	67.20	91.27	-
RVSA [70]	TGRS'22	34.3/44.4	64.49	-	-
BFM [8]	Arxiv'23	-	-	92.12	-
TOV [66]	JSTARS'23	32.1/37.8	66.24	92.03	-
SSL4EO [73]	GRSM'23	35.3/42.1	64.01	91.54	31.8/42.7
CMID [51]	TGRS'23	36.4/43.5	66.21	91.86	-
CACo [48]	CVPR'23	35.4/42.7	64.32	91.35	30.2/42.5
SAMRS [71]	NIPS'23	-	66.26	91.43	-
SatLas [4]	ICCV'23	37.4/40.7	68.71	91.28	31.9/43.5
GFM [50]	ICCV'23	36.7/45.6	66.62	91.85	-
Scale-MAE [55]	ICCV'23	34.0/41.7	65.77	91.54	-
MA3E [39]	ECCV'24	-	64.06	91.50	-
SkySense [24]	CVPR'24	39.7/46.5	70.91	93.99	33.1/46.2
<b>SkySense V2</b>	-	<b>41.2/47.6</b>	<b>71.87</b>	<b>95.86</b>	<b>35.2/47.5</b>

(a) Semantic segmentation results.

Models	Horizontal	Oriented	mAP <sub>50</sub>	mAP	mAP
	DIOR	DIOR-R			
GASSL [3]	67.40	65.65	48.15		
SatMAE [15]	70.89	65.66	46.55		
RingMo [64]	75.90	-	46.21		
RVSA [70]	73.22	71.05	47.04		
BFM [8]	-	73.62	-		
TOV [66]	70.16	66.33	49.62		
SSL4EO [73]	64.82	61.23	49.37		
CMID [51]	75.11	66.37	50.58		
CACo [48]	66.91	64.10	47.83		
SatLas [4]	74.10	67.59	46.19		
GFM [50]	72.84	67.67	49.69		
Scale-MAE [55]	73.81	66.47	48.31		
MA3E [39]	-	71.82	-		
SkySense [24]	78.73	74.27	54.57		
<b>SkySense V2</b>	<b>79.50</b>	<b>75.29</b>	<b>55.96</b>		

(b) Object detection results.

Models	LEVIR-CD	OSCD	Dyna.-S2
	F1	F1	SCS
GASSL [3]	78.19	46.26	13.6/16.7
SeCo [49]	90.14	47.67	13.9/16.0
SatMAE [15]	87.65	52.76	14.8/16.2
RingMo [64]	91.86	-	-
RVSA [70]	90.86	-	-
SpectralGPT [28]	-	54.29	-
MATTER [1]	-	59.37	-
DINO-MC [75]	-	52.70	14.5/15.6
SSL4EO [73]	89.05	35.08	12.3/17.5
CMID [51]	91.72	-	-
CACo [48]	81.04	52.11	15.3/15.8
SatLas [4]	90.62	-	13.3/17.8
GFM [50]	91.73	59.82	-
Scale-MAE [55]	92.07	-	-
SkySense [24]	92.58	60.06	15.4/18.0
<b>SkySense V2</b>	<b>94.83</b>	<b>65.29</b>	<b>16.0/20.7</b>

(c) Change detection results.

Table 2. Results of semantic segmentation, object detection and change detection.

Task & Dataset	Data Source	SkySense[24]	SkySense V2
(a) Multi-modal Seg: Dyna.-MM	(i) Planet.	46.5	<b>47.6</b>
	(ii) S2	46.2	<b>47.5</b>
	(iii) Planet. + S2	47.3	<b>48.7</b>
	(iv) Planet. + S2 + S1	47.7	<b>48.9</b>
(b) Multi-modal Seg: PASTIS-MM	(i) S2	73.5	<b>75.0</b>
	(ii) S2-Ts	84.6	<b>85.5</b>
	(iii) S2-Ts + S1-Ts	84.8	<b>85.6</b>
	(iv) S2-Ts + GEP	85.8	<b>86.7</b>
(c) Multi-modal Cls: BEN-MM	(i) S1	86.2	<b>86.5</b>
	(ii) S2 + S1	92.2	<b>93.8</b>

Table 3. Fine-tuning results on multi-modal tasks.

ably outperform SkySense by 5.2% on F1-score.

### 4.3. Multi-modal Tasks

In this section, we evaluate SkySense V2 on two representative multi-modal tasks, *i.e.*, Multi-modal Segmentation and Multi-modal Scene Classification, to demonstrate its generalization capability across data from various modalities.

**Multi-modal Segmentation.** In Table 3 (a) and (b), we present the performance results of Dyna.-MM [67] and PASTIS-MM [21], respectively. Dyna.-MM contains high-resolution optical imagery from PlanetFusion (Planet.), multi-spectral imagery from Sentinel-2 (S2), and SAR imagery from Sentinel-1 (S1). For the experiments conducted on Dyna.-MM, we employ a basic UperNet [80] as the decoder head and utilize mIoU as the evaluation metric. As depicted in Table 3 (a), SkySense V2 consistently outperforms the SkySense across various modalities, demonstrating a stronger representation ability. Furthermore, integrating multiple modalities yields higher performance compared to using a single modality. This results confirms that the proposed SkySense V2 can effectively extract representations from diverse modalities using a unified backbone.

PASTIS-MM is a crop mapping dataset that contains high-resolution optical imagery from Google Earth Pro (GEP), multi-temporal multi-spectral imagery from Sentinel-2 (S2-Ts), and multi-spectral SAR imagery from Sentinel-1 (S1-Ts). We employ a straightforward FCN head to decode the semantic segmentation results and report the overall accuracy as the evaluation metric in Table 3 (b). The results demonstrate that SkySense V2 outperforms the previous SOTA method SkySense by an average of 1.0%. Additionally, it is worth noting that S2-Ts can remarkably improve the segmentation accuracy compared to S2, underscoring the importance of temporal information for crop mapping.

**Multi-modal Scene Classification.** We conduct multi-modal scene classification experiments on the BEN-MM dataset [63]. As is shown in Table 3 (c), the proposed SkySense V2 surpasses SkySense, demonstrating a stronger representation ability. Furthermore, by integrating S1 and S2 imagery, SkySense V2 achieves higher improvements, highlighting its enhanced ability to extract superior representations from different modalities compared to SkySense.

### 4.4. Ablation and Discussion

**Ablation Study of Components.** To understand the contributions of each component, we conduct a k-nearest neighbors (k-NN) evaluation on the AID and RESISC-45 datasets. Figure 7 illustrates the performance of different component combinations across various pre-training iterations. Notably, the unified backbone design significantly accelerates representation learning. This is primarily due to shared parameters across different modalities, which allow gradients to aggregate, thereby speeding up convergence. Moreover, the unified design allows the backbone to be trained with data from different modalities, thereby enhancing the model’s generalization capabilities (see Ap-

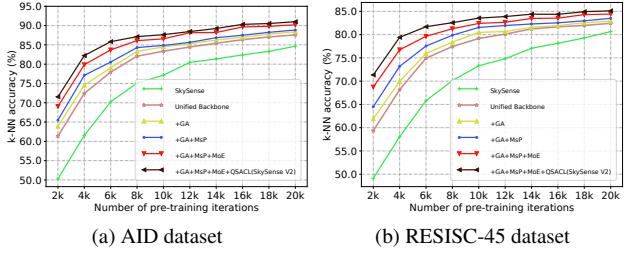


Figure 7. Ablation study using k-NN classification was conducted on the AID and RESISC-45 datasets. In this context, the abbreviations have the following meanings: GA refers to replacing window-based self-attention with global self-attention in the last two stages; MsP involves the addition of modality-specific prompt tokens; MoE indicates scaling up the model through a mixture of experts approach; and QSACL means adding additional query-based semantic aggregation contrastive learning.

pendix F.3 for more details). The results also indicate that global self-attention complements window-based self-attention, facilitating improved representations. Additionally, modality-specific prompt tokens and MoE scaling enhance representation learning by increasing feature diversity and model capacity, respectively. QSACL enhances model performance by improving the semantic precision of features in contrastive learning.

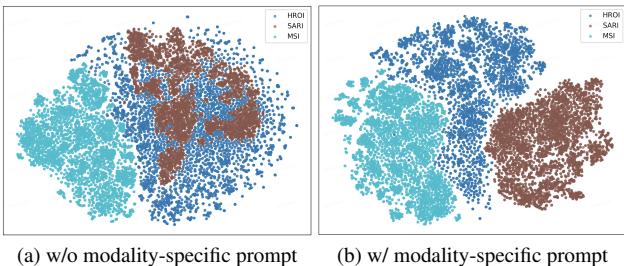


Figure 8. t-SNE visualization of the output feature representation from the final block of the unified transformer backbone. HROI, SARI, and MSI represent the features from high-resolution optical imagery, synthetic aperture radar imagery, and multi-spectral imagery, respectively.

**How Do Modality-Specific Prompt Tokens Facilitate Pre-training?** To investigate the impact of modality-specific prompt (MsP) tokens on the multi-modal pre-training process within a unified backbone, we employed t-SNE [69] for visualizing the features corresponding to different modalities. These features are extracted from the last block of the unified transformer backbone, and the comparative results are presented in Figure 8. Without the use of MsP, the features of high-resolution optical imagery (HROI) and synthetic aperture radar imagery (SARI) appear closely clustered, leading to overlapping feature distributions.

However, with the inclusion of MsP, the features of HROI and SARI become distinctly separable. These visualization outcomes clearly demonstrate that MsP effectively enhances feature diversity and imparts modality-specific characteristics.

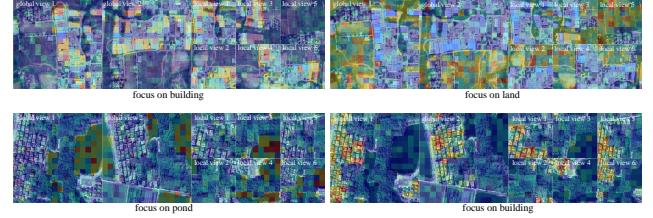


Figure 9. Visualization of cross-attention weights for features patch corresponding to different queries in QSACL. Queries can effectively aggregate features with specific semantics.

**What Types of Features Does Query-based Semantic Aggregation Contrastive Learning Capture?** To provide a clearer understanding of our proposed QSACL, we visualize the attention weights assigned to different features during their interaction with certain query, as illustrated in Figure 9. During pre-training, we use two global augmentation crops and six local augmentation crops. Here, we present two distinct queries, each focusing on different semantic features of image patches. For example, one query focuses on features with building semantics (left part of the first row in Figure 9), while another focuses on features with land semantics (right part of the first row in Figure 9). The attention weights of different queries aggregate features from each patch into distinct feature representations. Subsequently, we perform contrastive learning between aggregated features from the same query across teacher and student models. This approach significantly enhances semantic accuracy compared to traditional contrastive learning applied to entire image features.

## 5. Conclusion

In this paper, we introduce SkySense V2, a MM-RSFM that utilizes a unified backbone to accommodate various modalities. This unified approach enhances parameter utilization efficiency and improves the model’s generalization ability. SkySense V2 is pre-trained with an innovative QSACL strategy, specifically designed to leverage the unique characteristics of RS images. As a result, SkySense V2 significantly boosts performance while maintaining the multi-modal flexibility benefits of its predecessor, SkySense. Looking ahead, our future work will focus on integrating the language modality and incorporating a geographical knowledge graph to develop a more powerful and versatile MM-RSFM.

## A. Pre-training Module & Loss

### A.1. Multi-Granularity Contrastive Learning

We implement the multi-granularity contrastive learning proposed in SkySense[24] for self-supervised learning across multiple modalities and spatial granularities. Given the input set  $\{x_{HR}, x_{MS}, x_{SAR}\}$ , two separate collections of augmented views, denoted as  $\{u_i\}$  and  $\{v_i\}$ , are generated through random augmentations, where  $i \in \{HR, MS, SAR\}$ . These views  $u_i$  and  $v_i$  are then input into the student and teacher branches respectively. In the student branch, let  $\mathcal{T}_i$  represent the tokenizer for each modality and  $\mathcal{U}$  the unified transformer backbone of SkySense V2. The weights for the teacher branch are calculated as the exponential moving average (EMA) of the student branch weights:  $\mathcal{T}'_i = EMA(\mathcal{T}_i)$ ,  $\mathcal{U}' = EMA(\mathcal{U})$ s. This procedure yields spatial features as described in Equation 6:

$$F_i = \mathcal{U}(\mathcal{T}_i(u_i)), F'_i = \mathcal{U}'(\mathcal{T}'_i(v_i)) \quad i \in \{HR, MS, SAR\}. \quad (6)$$

By applying multi-modal temporal fusion and geo-context integration [24] to  $F_i$  and  $F'_i$ , we obtain the final features  $F_{\text{fus}}$  and  $F'_{\text{fus}}$ . We then initiate pixel-level, object-level, and image-level contrastive learning to progressively acquire coarse-to-fine spatial features for various tasks.

**Pixel-level Loss.** Each temporal slice of spatial feature  $F_i$  can be viewed as a pixel-level feature  $F_i^{\text{pix}} \in \mathbb{R}^{Ns \times d}$ . The pixel-level contrastive learning loss, denoted as  $\mathcal{L}_{\text{pix}}$  is calculated by averaging all  $\mathcal{L}_{CL}$  over both spatial ( $s$ ) and temporal ( $t$ ) dimensions, as described in Equation 7. Here,  $f_i^{\text{pix}} \in \mathbb{R}^d$  represents a feature vector from  $F_i^{\text{pix}}$  in specific location, and  $f_i^{\text{pix}'}$  is its correspondence at the same geo-location.  $\mathcal{L}_{CL}$  denotes the learning loss [7] between  $f_i^{\text{pix}}$  and  $f_i^{\text{pix}'}$ :

$$\mathcal{L}_{\text{pix}}(F_i, F'_i) = \frac{1}{N_S T_i} \sum_s \sum_t \mathcal{L}_{CL}(f_i^{\text{pix}}, f_i^{\text{pix}'}). \quad (7)$$

**Object-level Loss.** The object-level features  $F_i^{\text{obj}} \in \mathbb{R}^{N_C \times d}$  are generated from unsupervised clustering on pixel-level feature vectors  $f_i^{\text{pix}}$  in a single RSI, where  $N_C$  is the number of clusters. For clustering, we employ the Sinkhorn-Knopp algorithm [6], as used in [24]. Each cluster center, denoted as  $f_i^{\text{obj}} \in \mathbb{R}^d$  serves as a generalized representation for a collection of  $f_i^{\text{pix}}$ . This cluster center typically corresponds to a specific ground object or semantic concept. We calculate the object-level contrastive learning loss as follows:

$$\mathcal{L}_{\text{obj}}(F_i, F'_i) = \frac{1}{N_C T_i} \sum_s \sum_t \mathcal{L}_{CL}(f_i^{\text{obj}}, f_i^{\text{obj}'}). \quad (8)$$

**Image-level Loss.** The image-level feature  $F_i^{\text{img}} \in \mathbb{R}^d$  is simply an average pooling result from  $F_i^{\text{pix}}$ . The image-level contrastive learning loss is defined as follows:

$$\mathcal{L}_{\text{img}}(F_i, F'_i) = \frac{1}{T_i} \sum_t \mathcal{L}_{CL}(F_i^{\text{img}}, F_i^{\text{img}'}). \quad (9)$$

Finally, the fine-grained contrastive learning loss  $\mathcal{L}_{FGCL}$  is the sum of pixel-, object- and image-level contrastive learning losses, as described in Equation 10. Subsequently, we develop multi-modal loss  $\mathcal{L}_{MGCL}$  as shown in Equation 11. The multi-granularity concept is reflected in two main dimensions: spatial and modal. From a spatial perspective, contrastive learning is executed at the pixel, object, and image levels, enabling representation learning that comprehensively captures different spatial dimensions. From a modal perspective, we perform contrastive learning on both the features of individual modalities, denoted as  $F_i$ , the fused multi-modal features, represented as,  $F_{\text{fus}}$ :

$$\mathcal{L}_{FGCL}(F_i, F'_i) = \sum_{n \in \{\text{pix, obj, img}\}} \mathcal{L}_n(F_i, F'_i), \quad (10)$$

$$\begin{aligned} \mathcal{L}_{MGCL} = & \sum_{i \in \{HR, MS, SAR\}} \mathcal{L}_{FGCL}(F_i, F'_i) \\ & + \mathcal{L}_{FGCL}(F_{\text{fus}}, F'_{\text{fus}}). \end{aligned} \quad (11)$$

### A.2. Dense Image-Text Alignment

In addition to the  $\mathcal{L}_{MGCL}$  and  $\mathcal{L}_{QSACL}$  losses, we introduce an auxiliary supervision strategy using OpenStreetMap (OSM)<sup>2</sup> to enhance dense interpretation capabilities. OSM is an open-source, global-scale database that provides pixel-level land-cover and land-use categories. For multi-modal input imagery, we first collect the corresponding pixel-level OSM labels. Each pixel's class name is converted into a text representation using the CLIP [54] text encoder, and its visual representation is aligned with this text representation. Our experiments demonstrate that this dense image-text alignment encourages SkySense V2 to learn dense and semantic-aware representations.

Specifically, assuming the category set of OSM includes  $K$  classes, we first encode all class names to text representations  $F^{\text{text}} \in \mathbb{R}^{K \times D}$  with the CLIP text encoder, where  $D$  denotes the number of feature dimensions. Given a vision feature  $F \in \mathbb{R}^{N \times D}$  extracted by the SkySense V2 backbone, we maximize the similarity between each pixel's vision feature and its corresponding text feature while minimizing the similarity with non-matching text features. The dense image-text alignment loss  $\mathcal{L}_{ITA}$  is then formulated as

$$\mathcal{L}_{ITA} = -\frac{1}{n} \log \left( \sum_{i \in n} \frac{\exp(F_i * F_j^{\text{text}} / \tau)}{\sum_{k=1}^K \exp(F_i * F_k^{\text{text}} / \tau)} \right), \quad (12)$$

<sup>2</sup><https://www.openstreetmap.org/>

where  $j$  denotes the label index of the  $i$ -th vision feature, and  $\tau$  is a temperature parameter that controls the smoothness of the logits. By aligning the vision and text representations for every pixel as described in Eq. 12, SkySense V2 generates a more fine-grained interpretation of the input imagery.

### A.3. Unsupervised Geo-Context Prototype Learning

Different regions are characterized distinct geographic landscapes [29, 30] influenced by variations in culture, topography, and climate. SkySense [24] has demonstrated that this regional geo-context benefits the interpretation of remote sensing imagery [11, 23, 30, 40]. Following the approach of SkySense [24], we employ unsupervised geo-context prototype learning (GCPL) to group similar  $F_{\text{fus}}^{\text{mm}}$ . And these features are integrated as implicit geo-knowledge over a wide geo-spatial range to augment original feature during pre-training. Specifically, 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 based on  $F_{\text{fus}}^{\text{mm}}$ . 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  $F_{\text{fus}}^{\text{mm}}$  and  $\mathcal{P}_r$ :

$$\mathbf{M} = \frac{F_{\text{fus}}^{\text{mm}} \cdot \mathcal{P}_r^T}{\|F_{\text{fus}}^{\text{mm}}\| \|\mathcal{P}_r\|}. \quad (13)$$

The Sinkhorn-Knopp (SK) algorithm [6] on  $\mathbf{M}$  is utilized to find the optimal assignment matrix  $\mathbf{S} \in \mathbb{R}^{N_S \times N_p}$  between  $F_{\text{fus}}^{\text{mm}}$  and the prototypes. The SK algorithm incorporates a uniform distribution constraint to circumvent trivial solutions while striving to achieve the highest similarity possible. Subsequently, we utilize  $\mathbf{S}$  to generate an updated value for current sample's corresponding  $\mathcal{P}_r$ , denoted as  $\overline{\mathcal{P}}_r$ . This process is detailed as follows:

$$\overline{\mathcal{P}}_r = \mathbf{S}^T F_{\text{fus}}^{\text{mm}}. \quad (14)$$

Afterwards, we update  $\mathcal{P}_r$  through EMA [27] as in Equation 15, where  $m \in [0, 1]$  is a momentum coefficient.

$$\mathcal{P}_r \leftarrow m\mathcal{P}_r + (1 - m)\overline{\mathcal{P}}_r. \quad (15)$$

Each  $\mathcal{P}_r$  is updated during pre-training and serves as a fixed geo-context for downstream tasks. GCPL is applied exclusively to the student branch, extracting generalized region-aware representations from numerous RSI within a consistent region. This provides complementary information to enhance the features of individual RSI.

## B. Downstream Usage of SkySense V2

After pre-training, we utilize the parameters from the teacher branch for downstream tasks, as shown in Figure 10.

Each pre-trained module can be used independently or in combination with others, with the selected modules either frozen or fine-tuned. For single-modal static downstream tasks, we retain the unified transformer backbone and activate the specific tokenizer. Additionally, we add a task-specific head tailored to the particular task. In single-modal temporal downstream tasks, we incorporate the pre-trained fusion transformer to process time series feature data from a single modality. This fusion transformer integrates temporal information, enabling the model to capture dynamic patterns and trends over time, which are crucial for applications such as crop identification or change detection. For multi-modal downstream tasks, the fusion transformer is employed to integrate features from different modalities. This integration addresses both modality-specific and temporal aspects, allowing the model to leverage complementary information from various data sources. By fusing multi-modal data, SkySense V2 enhances its ability to perform complex tasks that require the synthesis of diverse information. This flexibility ensures that SkySense V2 can be effectively applied to a wide range of downstream applications, maintaining high performance while adapting to varying task demands.

## C. Pre-training Implementation Details

SkySense V2 is pre-trained using a batch size of 1024, distributed across 128 H20 GPUs. The model undergoes a total of 600k iterations, utilizing the AdamW optimizer [46] with  $\beta_1 = 0.9, \beta_2 = 0.999$ . The learning rate is initially set to  $2 \times 10^{-4}$  and decays to  $1 \times 10^{-6}$  following a cosine annealing schedule [45]. Similarly, the weight decay follows a cosine schedule, starting at 0.04 and increasing to 0.2. Additionally, to maintain stable training, the gradient is clipped at an  $L_2$  norm of 3.0 for all parameters. The momentum in EMA updating for teacher network is initialized as 0.996 and decay to 1.0 with cosine schedule. The loss weights for loss  $\mathcal{L}_{MGCL}, \mathcal{L}_{QSACL}, \mathcal{L}_{ITA}$  are set as:  $\lambda_1 = 1.0, \lambda_2 = 1.0, \lambda_3 = 1.0$ . The weight of MoE auxiliary loss is set to 0.01. The number of queries of QSACL is set to 16. The whole pre-training progress takes 44500 H20 GPU hours, and its computational complexity is 8109.52 GFLOPs.

For high-resolution optical imagery (HROI), we apply augmentations including Gaussian blur, solarization [22], random color jitter, random flips, and random rotations. In terms of multi-spectral imagery (MSI) and synthetic aperture radar imagery (SARI) time series, we randomly select a fixed-sized sequence (20 for MSI and 10 for SARI) from the original one and perform random disturbances on the RSI acquisition date. We follow the global and local multi-view cropping strategy in [6, 24], with 2 global views and 6 local views being used respectively.

Following SkySense[24], the multi-modal temporal fu-

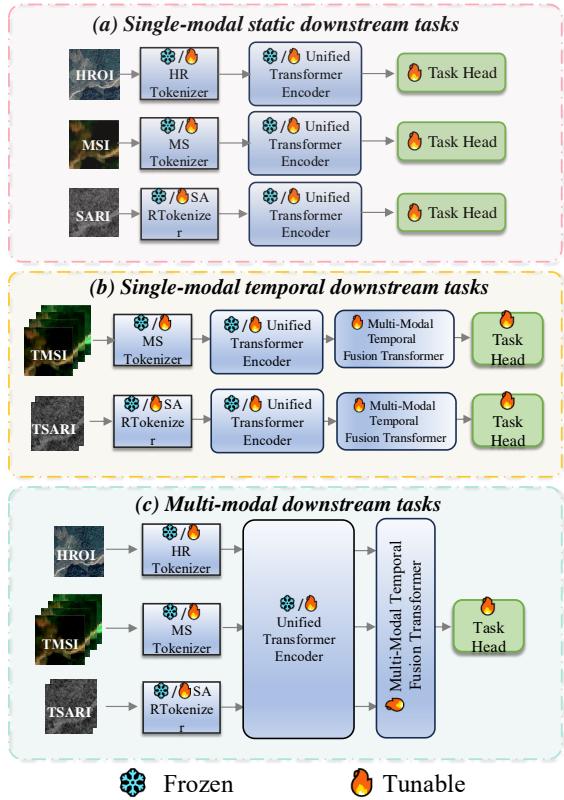


Figure 10. Overview of Downstream Usage of SkySense V2. Each pre-trained module can be utilized independently or in combination, with options to freeze or fine-tune the selected modules based on the specific downstream task requirements.

sion transformer module contains 24 basic transformer encoder layers. Additionally, a single basic transformer decoder layer is employed for query-based semantic aggregation contrastive learning. For GCPL, the globe is segmented into 4096 regions, each covering an area of roughly 4294 square kilometers and consisting of 100 prototypes.

## D. Downstream Tasks Training Implementation Details

### D.1. Semantic Segmentation

Semantic segmentation is widely used in remote sensing to automatically extract land use classes and ground instances. Considering factors such as spatial resolution, spectrum and number of categories, we select four popular datasets for the semantic segmentation task: DynamicEarthNet-PlanetFusion (Dyna.-Pla.) [67], iSAID [76], Potsdam [61], and DynamicEarthNet-Sentinel2 (Dyna.-S2). We employ the UperNet [80] as the unified segmentation head, implemented based on the MMSegmentation<sup>3</sup>, in line

<sup>3</sup><https://github.com/open-mmlab/mms Segmentation>

Dataset	Dyna.-Pla.	iSAID	Potsdam	Dyna.-S2
Activated modality	HR	HR	HR	MS
Optimizer	AdamW	AdamW	AdamW	AdamW
Input size	1024×1024	896×896	512×512	256×256
Input channel	RGBNIR	RGB	NIRRG	B02-08, B8A, B11-12
Base lr.	1e-4	1e-4	1e-4	1e-4
Lr. scheduler	poly	poly	poly	poly
Weight decay	0.01	0.01	0.01	0.01
Layer-wise lr decay	0.8	0.8	0.8	0.8
Max iters.	80k	80k	80k	80k
Warmup	linear	linear	linear	linear
Warmup iters.	1.5k	1.5k	1.5k	1.5k
Warmup ratio	1e-6	1e-6	1e-6	1e-6
Drop path rate	0.2	0.2	0.2	0.2
Augmentations				
RandomScaling		✓	✓	
RandomCrop	✓	✓	✓	✓
RandomFlip	✓	✓	✓	✓

Table 4. The finetuning setting in single-modal semantical segmentation tasks. The minimum and maximum values for random scaling are 0.5 and 2.0, respectively, and the probability of a random flip is 0.5.

with the approaches of [8, 64, 70]. Detailed fine-tuning settings are provided in Table 4.

### D.2. Horizontal & Oriented Objection Detection

Remote sensing images encompass a diverse array of objects, including buildings, vehicles, bridges and so on. These objects are densely distributed and vary widely in size, scale, and orientation, making their detection and identification a challenging task [77]. To evaluate the effectiveness of RSFMs in oriented object detection, we use the DIOR-R and FAIR1M datasets and implement the Oriented RCNN [37] as the detection algorithm, in line with prior studies [8, 24, 64, 70]. For assessing the horizontal object detection capabilities of SkySense V2, we utilize the DIOR dataset. Following the methodology of [24, 64], we employ the Faster RCNN [56] as the detector. Additional details are provided in Table 5.

### D.3. Change Detection

Change detection focuses on identifying pixel-level regional changes using bi-temporal or multi-temporal images. Building upon the work of Sun et al. [64], we incorporate the backbones of various RSFMs into the BIT framework [10] to evaluate their performance on the LEVIR-CD dataset. Following previous approaches [24, 48, 49], we utilize U-Net [58] as the segmentation head to assess the effectiveness of RSFMs in bi-temporal change detection tasks using the OSCD dataset with multi-spectral imagery. Addi-

Dataset	DIOR	DIOR-R	FAIR1M
Activated modality	HR	HR	HR
Optimizer	AdamW	AdamW	AdamW
Input size	800×800	800×800	512×512
Input channel	RGB	RGB	RGB
Base lr.	8e-5	8e-5	8e-5
Lr. scheduler	multistep	multistep	multistep
Layer-wise lr decay	0.85	0.85	0.85
Weight decay	0.05	0.05	0.05
Max epoch	12	12	8
Warmup	linear	linear	linear
Warmup iters.	1k	1k	0.5k
Warmup ratio	1e-3	1e-3	1e-3
Drop path rate	0.2	0.2	0.2
Augmentations			
RandomFlip	✓	✓	✓
RadnomRotate			✓
Head	Faster RCNN	Oriented RCNN	Oriented RCNN

Table 5. The finetuning setting in object detection tasks. The probability of a random flip is 0.5.

Dataset	LEVIR-CD	OSCD	Dyna.-S2
Activated modality	HR	MS	MS
Optimizer	AdamW	AdamW	AdamW
Input size	256×256	96×96	256×256
Input channel	RGB	B02-08, B8A,	B02-08, B8A,
Base lr.	6e-5	B11-12	B11-12
Lr. scheduler	LambdaLR	ExponentialLR	poly
Layer-wise lr decay	0.9	0.9	0.8
Weight decay	0.01	1e-4	0.05
Max iters./epoch	200 epochs	100 epochs	80k iters
Warmup	-	-	linear
Warmup iters.	-	-	1.5k
Warmup ratio	-	-	1e-6
Drop path rate	0.2	0.2	0.2
Augmentations			
RandomCrop	✓		✓
RandomFlip	✓	✓	✓
Head/Detector	BIT	U-Net	UperNet
Loss	CrossEntropy	BCE	CrossEntropy

Table 6. The finetuning setting in change detection tasks. The probability of a random flip is 0.5.

tionally, we use the DynamicEarthNet-Sentinel2 dataset to evaluate model performance on semantic change detection tasks, maintaining the same configuration as the segmentation task. Further settings are detailed in Section 6.

Dataset	AID	RESISC-45	BEN-S2	fMoW-S2
Activated modality	HR	HR	MS	MS
Optimizer	AdamW	AdamW	AdamW	AdamW
Input size	320×320	320×320	128×128	96×96
Input channel	RGB	RGB	B02-08, B8A,	B02-08, B8A,
Base lr.	6e-5	6e-5	5e-5	8e-4
Lr. scheduler	cosine	cosine	multistep	cosine
Weight decay	0.05	0.05	0.01	0.05
Layer-wise lr decay	0.9	0.9	0.9	0.9
Max epoch	200	200	100	30
Warmup	linear	linear	-	linear
Warmup epoch	5	5	-	5
Warmup ratio	0.01	0.01	-	0.2
Drop path rate	0.2	0.2	0.2	0.2
Augmentations				
RandomErasing	✓	✓		
RandomCrop	✓	✓		✓
Mixup				✓
RandomFlip	✓	✓	✓	✓

Table 7. The finetuning setting in single-modal semantical segmentation tasks. The minimum and maximum area ratio of random erasing are 0.03 and 0.333, respectively, and the probability of a random erasing is 0.3. The mixup ratio and probability are 0.8 and 1.0, respectively. The probability of a random flip is 0.5.

#### D.4. Scene Classification

We select two widely-used single-label scene classification datasets: AID and NWPU-RESISC45. Additionally, we utilize a multi-label multispectral scene classification dataset, BigEarthNet-Sentinel2, and a temporal multispectral scene classification dataset, fMoW-Sentinel2. The AID and NWPU-RESISC45 (RESISC-45) datasets consist of high-resolution optical images, while BigEarthNet-Sentinel2 (BEN-S2) and fMoW-Sentinel2 (fMoW-S2) are extensive multispectral image datasets. Our scene classification experiments are carried out using a standard linear classifier. Detailed implementation settings can be found in Table 7.

#### D.5. Multi-Modal Semantic Segmentation

By integrating multi-modal data from a variety of sensors, imaging techniques, resolutions, and spectral bands, we can extract a richer and more distinctive set of features. These features improve the ability to understand and interpret the shape, size, and relationships among ground objects. To evaluate the tasks of Time-insensitive Land Cover Mapping and Time-sensitive Crop Mapping, we use the DynamicEarthNet-MM (Dyna.-MM) dataset and the PASTIS-MM dataset, respectively.

Task	(i) Multi-Modal Segmentation: Time-insensitive LandCover Mapping	(ii) Multi-Modal Segmentation: Time-sensitive Crop Mapping	(iii) Multi-Modal Classification
Dataset	Dyna.-MM	PASTIS-MM	BEN-MM
Optimizer	AdamW	AdamW	AdamW
Input Size	planet: $1024 \times 1024$ sentinel2: $1024 \times 1024$ sentinel1: $1024 \times 1024$ planet: RGBNIR sentinel2: B02-08, B8A, B11-12 sentinel1: VV, VH	gep: $4096 \times 4096$ sentinel2: $128 \times 128$ sentinel1: $128 \times 128$ gep: RGB sentinel2: B02-08, B8A, B11-12 sentinel1: VV, VH	sentinel2: $128 \times 128$ sentinel1: $128 \times 128$ sentinel2: B02-08, B8A, B11-12 sentinel1: VV, VH
Input channel	6e-05 linear 0.01 8 6k iters linear 150 iters 1e-6 0.2 RandomFlip UperNet CrossEntropy	6e-05 linear 0.01 8 20k iters linear 1500 iters 1e-6 0.2 RandomFlip FCN CrossEntropy	5e-05 MultiStepLR 0.01 256 100 epoch - - - 0.2 RandomFlip Linear Classifier MultiLabel SoftMargin
Base learning rate			
Learning rate scheduler			
Weight decay			
Batch size			
Max iteration/epoch			
Warmup			
Warmup iteration/epoch			
Warmup ratio			
Drop path rate			
Augmentation			
Head/Detector			
Loss function			

Table 8. The finetuning setting in multi-modal downstream tasks.

**Dyna.-MM** contains spatially and temporally aligned multi-modal data, which include PlanetFusion imagery from the DynamicEarthNet-PlanetFusion dataset, Sentinel-2 multispectral imagery from the DynamicEarthNet-Sentinel2 dataset, and Sentinel-1 SAR imagery. For the SAR data, we utilize standard-calibrated Sentinel-1 GRD data with VV and VH polarizations, selecting it based on the geographical coordinates of the optical imagery. This approach is the same as SkySense [24] and ensures the validity of our multi-modal experiments. For segmentation tasks, UperNet is used as the segmentation head, and we report the mean Intersection over Union (mIoU) metric. Additional implementation details can be found in Table 8 (i).

**PASTIS-MM** [21, 24] is a dataset sourced from SkySense[24], which is designed for fine-grained, time-sensitive crop mapping. This dataset extends the PASTIS-R dataset [21] by incorporating spatially aligned high-resolution RGB images. PASTIS-MM aims to explore the combined impact of high-resolution optical imagery, medium-resolution temporal multispectral data, and temporal synthetic aperture radar (SAR) data in the context of time-sensitive crop mapping. The dataset was collected based on geo-coordinates and acquisition dates from the image tiles of the original PASTIS-R dataset, sourced from [24]. PASTIS-MM comprises 2433 Sentinel-2 image tiles, each with dimensions of  $128 \times 128$  pixels, 10 spectral bands, and a GSD of 10 meters. For each tile, the dataset includes

all available Sentinel-2 and Sentinel-1 acquisition data from September 2018 to November 2019, along with additional high-resolution visible imagery. For segmentation, we employ a naive Fully Convolutional Network (FCN) head [44] and report Overall Accuracy (OA) based on the official five-fold cross-validation of the dataset. Further implementation details can be found in Table 8 (ii).

## D.6. Multi-Modal Scene Classification

Following SkySense [24], we utilize the representative BigEarthNet-MM (BEN-MM) dataset to evaluate the performance of SkySense V2 in large-scale scene classification tasks, with a focus on integrating optical and SAR data. This dataset builds upon the BigEarthNet-Sentinel2 dataset by adding corresponding Sentinel-1 SAR data, thereby enabling the assessment of multi-label scene classification using both MS and SAR modalities. BEN-MM enriches each Sentinel-2 image patch from the BigEarthNet-Sentinel2 dataset with a preprocessed Sentinel-1 image patch taken around the same time. Each Sentinel-1 patch retains the annotation information from its corresponding Sentinel-2 patch and features a GSD of 10 meters. These patches provide dual-polarization information channels (VV and VH) and are collected in interferometric wide-swath mode. Consistent with prior studies [20, 24, 72, 73], we keep the same data splits as employed in the BigEarthNet-Sentinel2 dataset. Further implementation details can be found in Table 8 (iii).

## E. Comparison of Parameter Numbers with SkySense

Model Name	SkySense	SkySense V2 w/o MoE	SkySense V2
Tokenizer	0.21M <b>HR: 0.02</b> <b>MS: 0.16</b> <b>SAR: 0.03</b>	0.09M <b>HR: 0.02</b> <b>MS: 0.06</b> <b>SAR: 0.01</b>	0.09M <b>HR: 0.02</b> <b>MS: 0.06</b> <b>SAR: 0.01</b>
	1260.31M <b>HR: 655.17</b> <b>MS: 302.57</b> <b>SAR: 302.57</b>	661.40M	1994.10M
Modality prompt	-	9.94M	9.94M
Fusion module	398.20M	347.01M	347.01M
Others	404.13M	490.49M	490.49M
Total	2062.85M	1508.93M	2841.63M

Table 9. Comparison of the number of parameters in different modules between SkySense V2 and SkySense.

SkySense [24] employed three distinct backbones: Swin-H for high-resolution (HR) optical data, ViT-L for multi-spectral (MS) data, and ViT-L for synthetic aperture radar (SAR) data. In SkySense V2, the backbone parameters are shared across different modalities, maintaining a few separate parameters for modality-specific tokenizers and prompts. Detailed comparisons are presented in Table 9. By adopting this unified design, the total number of backbone parameters for the three modalities has been reduced from 1,260 million to 661 million. Additionally, we incorporated a mixture of experts (MoE) approach [33], which allowed us to scale up the number of parameters to 1,994 million (with 661 million activated). To sum up, our unified transformer backbone employs full parameter sharing across different modalities, presenting several key benefits: 1) As discussed in the ablation part in our paper, this parameter sharing aggregates gradients from all modalities, thereby accelerating the convergence process. 2) It significantly boosts parameter utilization efficiency, leaving enough room for increasing additional capacity by incorporating MoE modules, which further enhances representation learning. 3) Our unified model architecture and complete parameter sharing simplifies the alignment of features across different modalities.

## F. Experiments

### F.1. Influence of Image-text Alignment with OSM

OpenStreetMap is a global open-source data providing a wealth of semantic classes. We utilize the CLIP text encoder [54] to transform categories into text representations and then apply dense image-text alignment (ITA) to enhance pre-trained model’s capability for dense interpreta-

tion. To validate this approach, we conducted ablation experiments on segmentation datasets, specifically iSAID and Potsdam. Due to the resource-intensive nature of the whole pre-training, we ensured a fair comparison by limiting it to 20,000 iterations. The fine-tuning process was kept consistent with the approach outlined in Section D.1. The results, presented in Table 10, demonstrate that image-text alignment effectively improves the performance of dense tasks.

Dataset	iSAID	Potsdam
w/o ITA	67.45	88.77
w/ ITA	68.24	90.05

Table 10. Ablation results of image-text alignment in SkySense V2.

### F.2. Features of Different Resolutions Derived from Adaptive Patch Merging

Our Adaptive Patch Merging (APM) module, integrated after each stage of the unified backbone, can flexibly generate features with various resolutions based on specific requirements. To evaluate the impact of different subsampling activation conditions within APM, we conducted ablation experiments on the segmentation datasets iSAID and Potsdam. The fine-tuning process remained consistent with the methodology outlined in Section D.1, and all models utilized parameters from the same pre-trained model. As shown in Table 11, generating higher-resolution features through APM enhances the model’s performance. This improvement makes the model particularly advantageous for deployment in environments where sufficient computing resources are available.

Sub-sampling activation of APM			downscale	Dataset	
Stage 2	Stage 3	Stage 4		iSAID	Potsdam
✓	✓	✓	1/8	71.87	95.86
✓	✓		1/4	71.92	95.85
✓			1/2	72.55	96.76
			—	72.88	97.03

Table 11. Experiment results of different sub-sampling activation conditions within APM. All models were initialized with identical parameters, differing only in their subsampling activation strategies in APM.

### F.3. Performance on Sensor Data Outside of Training

To further validate the generalizability of the pre-trained model, we conducted experiments on three datasets collected from different sensors: Five-Billion-Pixels (FBP) [68] from the Gaofen-2 satellite, SPARCS [31] from the Landsat-8 satellite, and AIR-PoSAR-Seg (APS) [74] from

the Gaofen-3 satellite. All these datasets utilize sensors different from those used in the training data. FBP comprises over 5 billion labeled pixels across 150 high-resolution images, annotated into 24 categories covering artificially constructed, agricultural, and natural classes. SPARCS includes 80 images with a resolution of  $1000 \times 1000$  pixels, annotated into 7 categories. APS consists of a PolSAR image with a region of  $9082 \times 9805$  pixels and 2000 image patches, each sized  $512 \times 512$  pixels. The experimental results on these three datasets are presented in Table 12. SkySense V2 surpasses SkySense by an average of 1.8% in mIoU, indicating that SkySense V2 possesses stronger generalization capabilities than SkySense. We attribute this improvement to the unified design, which allows the backbone parameters to be trained with data from different modalities, thereby enhancing the model’s ability to generalize effectively.

Dataset	Sensor	SkySense	SkySense V2
Five-Billion-Pixels	Gaofen-2	65.31	66.82
SPARCS	Landsat-8	72.57	74.32
AIR-PoSAR-Seg	Gaofen-3(SAR)	53.21	55.32

Table 12. Results on datasets built from various sensors. The evaluation metric is mIoU.

#### F.4. Ablation of Modality-specific Prompt Tokens in Downstream Tasks

After pre-training the model, there are two options for handling Modality-specific Prompt Tokens (MSPT) during downstream fine-tuning: 1) retaining the MSPT or 2) removing the MSPT entirely. We assess the impact of MSPT in two different settings: 1) single-modal tasks, where only one modality is activated, and 2) multi-modal tasks, where at least two modalities are activated. For the single-modal setting, we conduct experiments using the RESISC-45 and BEN-S2 datasets. For the multi-modal setting, we utilize the BEN-MM dataset. As demonstrated in Table 13, our findings indicate that MSPT can significantly enhance performance in multi-modal tasks, primarily due to its ability to increase the diversity of features of different modalities.

Dataset	RESISC-45	BEN-S2	BEN-MM
	(TR=10%)	(TR=10%)	
Activated modality	HR	MS	MS,SAR
w/o MSPT	96.15	88.97	92.64
w/ MSPT	96.42	89.13	93.81

Table 13. Results of ablation study of modality-specific prompt token in downstream tasks. "TR" refers to training ratio, representing the proportion of training data relative to the entire dataset.

#### F.5. Ablation Studies about MoE in Pre-training

To quickly assess the impact of MOE-related configurations, we pre-trained the model with 20,000 iterations for each configuration. After pre-training, we evaluated the model on the AID and RESISC-45 datasets using the k-NN accuracy.

**Varying the number of experts.** We configured the unified backbone with varying numbers of experts to evaluate performance relative to parameter size. The results, shown in Table 14, indicate that as the number of experts increases, the representational capacity of the model improves. Although the configuration with 16 experts outperforms that with 8 experts, it requires an additional 1.6 billion parameters. This increase in parameters does not match the marginal gain in performance. Consequently, we set the number of experts to 8 in our SkySense V2 model.

#experts	#parameters	AID	RESISC-45
4	1232.61M	89.05	82.57
8	1994.10M	91.00	85.11
16	3517.08M	91.23	85.97

Table 14. Ablation results of varing number of experts in MoE. We report k-NN classification accuracy on AID and RESISC-45 datasets.

**Varying the number of MoE blocks.** Following prior methods utilizing Mixture of Experts (MoE) [43, 78], we integrate MoE modules into the last  $L$  transformer blocks, substituting the original feed-forward network (FFN) layers. Each MoE module comprises 8 experts, all of which maintain the FFN’s structural design but function as independent networks. We present ablation studies exploring various configurations with different numbers of MoE blocks ( $L = 2, 4, 6, 8$ ) within the backbone. As shown in Table 15, the results indicate that performance tends to plateau at 6 MoE blocks.

$L$	#parameters	AID	RESISC-45
2	1486.37M	89.63	83.45
4	1740.23M	90.14	84.37
6	1994.10M	91.00	85.11
8	2247.97M	91.11	85.43

Table 15. Ablation results of varing number of MoE blocks in backbone.

**Exploring different distributions of MoE blocks in backbone.** Previous studies [43, 78] typically incorporate Mixture of Experts (MoE) into the last few layers of a network. This approach is motivated by two main factors: 1) deeper

routing decisions are more closely related to image classes and contain richer semantic information [57], and 2) the last layers have the most significant impact on classification performance. However, the official implementation of Swin-MoE [32]<sup>4</sup> introduces an alternative strategy, distributing MoE blocks evenly across all layers in whole backbone. We tested both distribution strategies within the backbone of SkySense V2. Our findings indicate that while the performance difference between the two methods is minimal, incorporating MoE blocks into the last layers offers a slight advantage, as detailed in Table 16.

MoE block(layer) index Total: 24	AID	RESISC-45
3,7,11, 15,19,23	90.93	84.87
18,19,20, 21,22,23	91.00	85.11

Table 16. Comparison of different MoE distribution strategies within the backbone of SkySense V2.

## F.6. Ablation Studies about MoE in Downstream Tasks

To further investigate the MoE, we conducted ablation experiments during the downstream fine-tuning phase of a pre-trained MoE backbone. We selected the RESISC-45, BEN-S2, and BEN-MM datasets, which encompass three modalities: high-resolution (HR), multispectral (MS), and synthetic aperture radar (SAR). Firstly, we examined whether to fix the routing gate during downstream fine-tuning. As shown in Table 17, despite the routing gate being trained with a substantial amount of data during the pre-training stage, fine-tuning for a specific task proves to be necessary. Secondly, we experimented by randomly keeping one expert from the MoE block while removing the others, effectively reducing the MoE feed-forward network (FFN) to a plain FFN. The performance of this modified model is comparable to the fully pre-trained backbone without MoE, indicating that each expert has been sufficiently trained and possesses individual representational capabilities.

## F.7. Ablation Studies about the Number of Queries in Query-based Semantic Aggregation Contrastive Learning

In Query-based Semantic Aggregation Contrastive Learning (QSACL) ablation study, we explore the influence of different  $m$  learnable queries, which are used to aggregate features with different semantics across multiple augmented views of an image. We pre-trained the model with 20,000 iterations, experimenting with various numbers of queries.

<sup>4</sup><https://github.com/microsoft/Swin-Transformer>

Dataset Activated modality	RESISC-45 (TR=10%)	BEN-S2 (TR=10%)	BEN-MM
	HR	MS	MS,SAR
SkySense V2	96.42	89.13	93.81
SkySense V2 w/o MoE	95.61	88.76	92.95
SkySense V2 fixed routing gate	95.73	88.65	92.80
SkySense V2 random 1 expert	95.47	88.69	92.81

Table 17. Results of ablation study of MoE in downstream tasks. "TR" refers to training ratio, representing the proportion of training data relative to the entire dataset.

After pre-training, we evaluated the model's performance on the AID and RESISC-45 datasets using k-NN accuracy. The results, as shown in Table 18, indicate that with a small number of queries, such as 4 or 8, performance drops significantly. This decline occurs because the number of queries is insufficient to capture the diverse semantic categories within an image, resulting in inadequate pre-training. Conversely, when  $m = 8$ , the performance remains similar to when  $m = 16$ , suggesting that 16 queries are sufficient to capture the different semantics in an image, with additional queries offering no significant improvement in performance.

$m$	AID	RESISC-45
4	90.21	84.32
8	90.68	84.87
16	91.00	85.11
24	91.05	85.07

Table 18. Ablation results of varying number of MoE learnable queries in QSACL.

## F.8. Comparison with Random Initialization.

In this section, we use both SkySense pre-trained weights and randomly initialized weights to fine-tune the same backbone network across three datasets, each corresponding to a different task. These tasks include scene classification with the AID dataset [79], object detection with the DIOR dataset [36], and semantic segmentation with the iSAID dataset [76]. The experimental results, which are shown in Table 19, indicate a significant performance advantage for our pre-trained model compared to the model trained from scratch across all three datasets.

## 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*, pp. 1033–1042. 2020.

Model	AID	DIOR	iSAID
	OA(TR=20/50%)	mAP <sub>50</sub>	mIoU
Randm Init	66.82/90.78	56.36	48.34
SkySense V2	98.34/99.05	79.50	71.87

Table 19. Comparison of SkySense V2 with random initialization and SkySense V2 with pre-training.

- ference on Computer Vision and Pattern Recognition, pages 8203–8215, 2022. 7
- [2] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. AnySat: An Earth observation model for any resolutions, scales, and modalities. *arXiv preprint arXiv:2412.14123*, 2024. 3
- [3] 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. 6, 7
- [4] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Fernando, 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. 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
- [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. 9, 10
- [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. 5, 9
- [8] Keumgang Cha, Junghoon Seo, and Taekyung Lee. A billion-scale foundation model for remote sensing images. *arXiv preprint arXiv:2304.05215*, 2023. 7, 11
- [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] Hao Chen, Zipeng Qi, and Zhenwei Shi. Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021. 6, 11
- [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. 10

- [12] 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
- [13] 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, 6
- [14] 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. 6
- [15] 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, 6, 7
- [16] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024. 2
- [17] Rodrigo Caye Daudt, Bertrand 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
- [18] 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. 1, 3
- [19] William Fedus, Barret Zoph, and Noam M. Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *ArXiv*, abs/2101.03961, 2021. 2
- [20] 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. 6, 13
- [21] 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, 13
- [22] 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. 10
- [23] Shaohua Guo, Liang Liu, Zhenye Gan, Yabiao Wang, Wuhan 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. 10

- [24] Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, Huimei He, Jian Wang, Jingdong Chen, Ming Yang, Yongjun Zhang, and Yansheng 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 (CVPR)*, pages 27672–27683, 2024. 1, 3, 4, 5, 6, 7, 9, 10, 11, 13, 14
- [25] Wang Haiyang, Tang Hao, Shi Shaoshuai, Li Aoxue, Li Zhenguo, Schiele Bernt, and Liwei Wang. Unitr: A unified and efficient multi-modal transformer for bird’s-eye-view representation. In *ICCV*, 2023. 3
- [26] Boran Han, Shuai Zhang, Xingjian Shi, and Markus Reichstein. Bridging remote sensors with multisensor geospatial foundation models. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27852–27862, 2024. 3
- [27] 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. 10
- [28] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Pedram Ghamisi, Naoto Yokoya, Hao Li, Xiuping Jia, Antonio Plaza, Paolo Gamba, Jon Atli Benediktsson, and Jocelyn Chanussot. Spectralgpt: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (8):5227–5244, 2024. 7
- [29] 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. 10
- [30] 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 (gis). *International Journal of Applied Earth Observation and Geoinformation*, 109:102787, 2022. 10
- [31] M. Joseph Hughes and Robert H. Kennedy. High-quality cloud masking of landsat 8 imagery using convolutional neural networks. *Remote. Sens.*, 11:2591, 2019. 14
- [32] Changho Hwang, Wei Cui, Yifan Xiong, Ziyue Yang, Ze Liu, Han Hu, Zilong Wang, Rafael Salas, Jithin Jose, Prabhakar Ram, Joe Chau, Peng Cheng, Fan Yang, Mao Yang, and Yongqiang Xiong. Tutel: Adaptive mixture-of-experts at scale, 2022. 16
- [33] Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87, 1991. 2, 4, 14
- [34] Pallavi Jain, Bianca Schoen-Phelan, and Robert J. Ross. Self-supervised learning for invariant representations from multi-spectral and sar images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15: 7797–7808, 2022. 2
- [35] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision (ECCV)*, 2022. 2
- [36] Ke Li, Gang Wan, Gong Cheng, Lijiu 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, 16
- [37] Wentong Li, Yijie Chen, Kaixuan Hu, and Jianke Zhu. Oriented repoints for aerial object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1829–1838, 2022. 6, 11
- [38] Xuyang Li, Danfeng Hong, and Jocelyn Chanussot. S2mae: A spatial-spectral pretraining foundation model for spectral remote sensing data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24088–24097, 2024. 2
- [39] Zhihao Li, Biao Hou, Siteng Ma, Zitong Wu, Xianpeng Guo, Bo Ren, and Licheng Jiao. Masked angle-aware autoencoder for remote sensing images. In *European Conference on Computer Vision*, pages 260–278. Springer, 2024. 2, 6, 7
- [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. 10
- [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. 1
- [42] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12009–12019, 2022. 3
- [43] Zhili Liu, Kai Chen, Jianhua Han, Lanqing Hong, Hang Xu, Zhenguo Li, and James Tin-Yau Kwok. Task-customized masked autoencoder via mixture of cluster-conditional experts. *ArXiv*, abs/2402.05382, 2024. 4, 15
- [44] 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. 13
- [45] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv: Learning*, 2016. 6, 10
- [46] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CorR*, abs/1711.05101, 2017. 6, 10
- [47] 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
- [48] 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, 11

- [49] 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](#), [11](#)
- [50] 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](#), [6](#), [7](#)
- [51] 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](#), [6](#), [7](#)
- [52] Mubashir Noman, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan. Rethinking transformers pre-training for multi-spectral satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27811–27819, 2024. [2](#), [6](#)
- [53] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. [1](#), [4](#)
- [54] 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. [9](#), [14](#)
- [55] 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. [6](#), [7](#)
- [56] 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](#), [11](#)
- [57] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. In *Neural Information Processing Systems*, 2021. [16](#)
- [58] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. [11](#)
- [59] Sheng Shen, Le Hou, Yan-Quan Zhou, Nan Du, S. Longpre, Jason Wei, Hyung Won Chung, Barret Zoph, William Fedus, Xinyun Chen, Tu Vu, Yuexin Wu, Wuyang Chen, Albert Webson, Yunxuan Li, Vincent Zhao, Hongkun Yu, Kurt Keutzer, Trevor Darrell, and Denny Zhou. Mixture-of-experts meets instruction tuning: A winning combination for large language models. In *International Conference on Learning Representations*, 2023. [2](#)
- [60] Sheng Shen, Le Hou, Yanqi Zhou, Nan Du, Shayne Longpre, Jason Wei, Hyung Won Chung, Barret Zoph, William Fedus, Xinyun Chen, Tu Vu, Yuexin Wu, Wuyang Chen, Albert Webson, Yunxuan Li, Vincent Y Zhao, Hongkun Yu, Kurt Keutzer, Trevor Darrell, and Denny Zhou. Mixture-of-experts meets instruction tuning: A winning combination for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. [3](#)
- [61] Jamie Sherrah. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. *arXiv preprint arXiv:1606.02585*, 2016. [6](#), [11](#)
- [62] 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](#), [6](#)
- [63] 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](#)
- [64] 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](#), [11](#)
- [65] 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](#)
- [66] 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. [6](#), [7](#)
- [67] Aysim Toker, Lukas Kondmann, Mark Weber, Marvin Eisenberger, Andrés Camero, Jingliang Hu, Ariadna Pregel Hoderlein, Çağlar Şenaras, Timothy Davis, Daniel Creimers, 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](#), [11](#)
- [68] Xin-Yi Tong, Guisong Xia, and Xiaoxiang Zhu. Enabling country-scale land cover mapping with meter-resolution satellite imagery. *Isprs Journal of Photogrammetry and Remote Sensing*, 196:178 – 196, 2022. [14](#)
- [69] Laurens van der Maaten and Geoffrey E. Hinton. Visualizing

- data using t-sne. *Journal of Machine Learning Research*, 9: 2579–2605, 2008. 8
- [70] 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. 6, 7, 11
- [71] 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. 7
- [72] Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Chenying Liu, Zhitong Xiong, and Xiao Xiang Zhu. De-cur: decoupling common & unique representations for multimodal self-supervision. *arXiv preprint arXiv:2309.05300*, 2023. 13
- [73] 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. 6, 7, 13
- [74] Zhirui Wang, X.-M. Zeng, Zhiyuan Yan, Jian Kang, and Xian Sun. Air-polsar-seg: A large-scale data set for terrain segmentation in complex-scene polsar images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:3830–3841, 2022. 14
- [75] 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, 7
- [76] 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, 11, 16
- [77] Long Wen, Yu Cheng, Yi Fang, and Xinyu Li. A comprehensive survey of oriented object detection in remote sensing images. *Expert Systems with Applications*, page 119960, 2023. 11
- [78] Lemeng Wu, Mengchen Liu, Yinpeng Chen, Dongdong Chen, Xiyang Dai, and Lu Yuan. Residual mixture of experts. *ArXiv*, abs/2204.09636, 2022. 4, 15
- [79] 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, 6, 16
- [80] 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, 7, 11
- [81] Zhitong Xiong, Yi Wang, Fahong Zhang, and Xiao Xiang Zhu. One for all: Toward unified foundation models for earth vision. *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 2734–2738, 2024. 3
- [82] Longqi Yang, Liangliang Zhang, and Wenjing Yang. Graph adversarial self-supervised learning. In *Advances in Neural Information Processing Systems*, 2021. 2
- [83] 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
- [84] Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. Meta-transformer: A unified framework for multimodal learning. *arXiv preprint arXiv:2307.10802*, 2023. 3
- [85] Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Xiaogang Wang, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16783–16794, 2021. 3
- [86] Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam M. Shazeer, and William Fedus. Stmoe: Designing stable and transferable sparse expert models. 2022. 2