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

Yingying Zhang[1] Lixiang Ru[1] Kang Wu[2] Lei Yu[1] Lei Liang[1] Yansheng Li[2] Jingdong Chen[1]
[1]Ant Group [2] Wuhan University
qichu.zyy@antgroup.com

A. Pre-training Module & Loss

A.1. Multi-Granularity Contrastive Learning

We implement the multi-granularity contrastive learning proposed in SkySense[10] 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 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 1:

$$F_i = \mathcal{U}(\mathcal{T}_i(u_i)), F_i' = \mathcal{U}'(\mathcal{T}_i'(v_i)) \ \ i \in \{HR, MS, SAR\}. \tag{1}$$

By applying multi-modal temporal fusion and geo-context integration [10] to F_i and F_i', we obtain the final features $F_{\mathbf{fus}}$ and $F_{\mathbf{fus}}'$. 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 $F_i^{\mathbf{pix}} \in \mathbb{R}^{N_S \times d}$. The pixel-level contrastive learning loss, denoted as $\mathcal{L}_{\mathbf{pix}}$ is calculated by averaging all $\mathcal{L}_{CL}$ over both spatial (s) and temporal (t) dimensions, as described in Equation 2. Here, $f_i^{\mathbf{pix}} \in \mathbb{R}^d$ represents a feature vector from $F_i^{\mathbf{pix}}$ in specific location, and $f_i^{\mathbf{pix'}}$ is its correspondence at the same geo-location. $\mathcal{L}_{CL}$ denotes the learning loss [2] between $f_i^{\mathbf{pix}}$ and $f_i^{\mathbf{pix'}}$:

$$\mathcal{L}_{\mathbf{pix}}(F_i, F_i') = \frac{1}{N_S T_i} \sum_s \sum_t \mathcal{L}_{CL}(f_i^{\mathbf{pix}}, f_i^{\mathbf{pix'}}). \tag{2}$$

Object-level Loss. The object-level features $F_i^{\mathbf{obj}} \in \mathbb{R}^{N_C \times d}$ are generated from unsupervised clustering on pixel-level feature vectors $f_i^{\mathbf{pix}}$ in a single RSI, where N_C is the number of clusters. For clustering, we employ the Sinkhorn-Knopp algorithm [1], as used in [10]. Each cluster center, denoted as $f_i^{\mathbf{obj}} \in \mathbb{R}^d$ serves as a generalized representation for a collection of $f_i^{\mathbf{pix}}$. This cluster center typically corresponds to a specific ground object or semantic concept. We calculate the object-level contrastive learning loss as follows:

$$\mathcal{L}_{\mathbf{obj}}(F_i, F_i') = \frac{1}{N_C T_i} \sum_s \sum_t \mathcal{L}_{CL}(f_i^{\mathbf{obj}}, f_i^{\mathbf{obj'}}). \tag{3}$$

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

$$\mathcal{L}_{\mathbf{img}}(F_i, F_i') = \frac{1}{T_i} \sum_t \mathcal{L}_{CL}(F_i^{\mathbf{img}}, F_i^{\mathbf{img'}}). \tag{4}$$

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 5. Subsequently, we develop multi-modal loss $\mathcal{L}_{MGCL}$ as shown in Equation 6. 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, $F_{\mathbf{fus}}$:

$$\mathcal{L}_{FGCL}(F_i, F_i') = \sum_{n \in \{\mathbf{pix}, \mathbf{obj}, \mathbf{img}\}} \mathcal{L}_n(F_i, F_i'), \tag{5}$$

$$\begin{aligned} \mathcal{L}_{MGCL} = & \sum_{i \in \{HR, MS, SAR\}} \mathcal{L}_{FGCL}(F_i, F_i') \\ & + \mathcal{L}_{FGCL}(F_{\mathbf{fus}}, F_{\mathbf{fus}}'). \end{aligned} \tag{6}$$

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)[1] 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 [26] 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 $F^{text} \in \mathbb{R}^{K \times D}$ 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

$$\mathcal{L}_{ITA} = -\frac{1}{n} \log\left(\sum_{i \in n} \frac{\exp(F_i * F_j^{text}/\tau)}{\sum_{k=1}^{K} \exp(F_i * F_k^{text}/\tau)}\right), \quad (7)$$

where j denotes the label index of the i-th vision feature, and τ 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. 7, 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 [12, 13] influenced by variations in culture, topography, and climate. SkySense [10] has demonstrated that this regional geo-context benefits the interpretation of remote sensing imagery [5, 9, 13, 19]. Following the approach of SkySense [10], we employ unsupervised geo-context prototype learning (GCPL) to group similar $F_{\mathbf{fus}}^{\text{mm}}$. 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 $F_{\mathbf{fus}}^{\text{mm}}$. 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 $F_{\mathbf{fus}}^{\text{mm}}$ and $\mathcal{P}_r$:

$$\mathbf{M} = \frac{F_{\mathbf{fus}}^{\text{mm}} \cdot \mathcal{P}_r^{\text{T}}}{\|F_{\mathbf{fus}}^{\text{mm}}\| \|\mathcal{P}_r\|}. \quad (8)$$

The Sinkhorn-Knopp (SK) algorithm [1] on $\mathbf{M}$ is utilized to find the optimal assignment matrix $\mathbf{S} \in \mathbb{R}^{N_S \times N_p}$ between $F_{\mathbf{fus}}^{\text{mm}}$ 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:

$$\overline{\mathcal{P}_r} = \mathbf{S}^{\text{T}} F_{\mathbf{fus}}^{\text{mm}}. \quad (9)$$

Afterwards, we update $\mathcal{P}_r$ through EMA [11] as in Equation 10, where $m \in [0, 1)$ is a momentum coefficient.

$$\mathcal{P}_r \leftarrow m\mathcal{P}_r + (1 - m)\overline{\mathcal{P}_r}. \quad (10)$$

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 1. 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.

[1] https://www.openstreetmap.org/

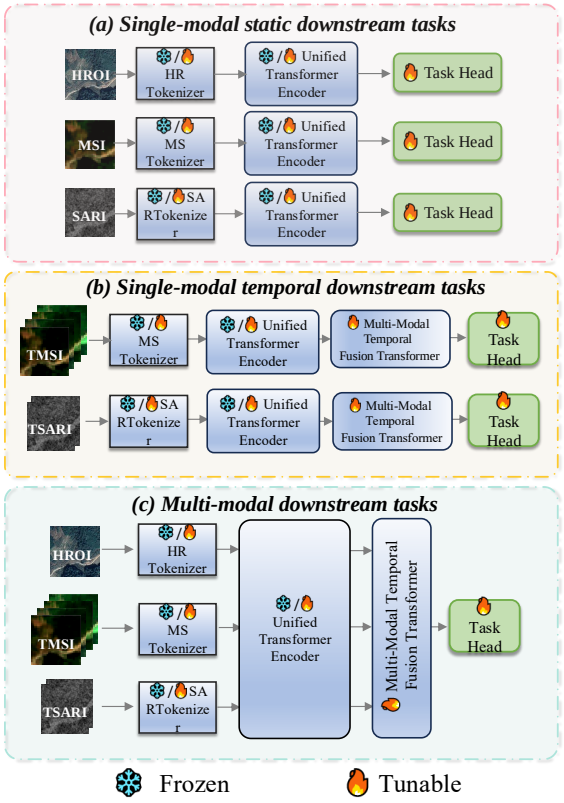

Figure 1. Overview of Downstream Usage of SkySense V2. Each pre-trained module can be utilized independently or in combination, with options to freeze or fine-tune the selected modules based on the specific downstream task requirements.

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 [23] with $\beta_1 = 0.9, \beta_2 = 0.999$. The learning rate is initially set to 2×10^{-4} and decays to 1×10^{-6} following a cosine annealing schedule [22]. 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 [8], 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 [1, 10], with 2 global views and 6 local views being used respectively.

Following SkySense[10], the multi-modal temporal fusion 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.) [32], iSAID [38], Potsdam [30], and DynamicEarthNet-Sentinel2 (Dyna.-S2). We employ the UperNet [42] as the unified segmentation head,implemented based on the MMSegmentation[2], in line with the approaches of [3, 31, 34]. Detailed fine-tuning settings are provided in Table 1.

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 [39]. To evaluate the effectiveness of RSFMs in oriented object detection, we use the DIOR-R and FAIR1M datasets and implement the Oriented RCNN [18] as the detection algorithm, in line with prior studies [3, 10, 31, 34]. For assessing the horizontal object detection capabilities of SkySense V2, we utilize the DIOR dataset. Following the methodology of [10, 31], we employ the Faster RCNN [27] as the detector. Additional details are provided in Table 2.

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. [31], we incorporate the backbones of various RSFMs into the BIT framework [4] to evaluate their performance on the LEVIR-CD

[2] https://github.com/open-mmlab/mmsegmentation

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 1. The finetuning setting in single-modal semantical segmetation 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.

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

dataset. Following previous approaches [10, 24, 25], we utilize U-Net [29] as the segmentation head to assess the effectiveness of RSFMs in bi-temporal change detection tasks using the OSCD dataset with multi-spectral imagery. Additionally, we use the DynamicEarthNet-Sentinel2 dataset to

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, B11-12	B02-08, B8A, B11-12
Base lr.	6e-5	6e-4	1e-4
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 3. The finetuning setting in change detection tasks. The probability of a random flip is 0.5.

evaluate model performance on semantic change detection tasks, maintaining the same configuration as the segmentation task. Further settings are detailed in Section 3.

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 4.

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.

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, B11-12	B02-08, B8A, B11-12
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 4. The finetuning setting in single-modal semantical segmetation 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.

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 [10] 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 5 (i).

PASTIS-MM [7, 10] is a dataset sourced from SkySense[10], which is designed for fine-grained, time-sensitive crop mapping. This dataset extends the PASTIS-R dataset [7] 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 [10]. PASTIS-MM comprises 2433 Sentinel-2 image tiles, each with dimensions of 128×128 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 [21] and report Overall Accuracy (OA) based on the official five-fold cross-validation of the dataset. Further implementation details can be found in Table 5 (ii).

D.6. Multi-Modal Scene Classification

Following SkySense [10], 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 [6, 10, 35, 36], we keep the same data splits as employed in the BigEarthNet-Sentinel2 dataset. Further implementation details can be found in Table 5 (iii).

E. Comparison of Parameter Numbers with SkySense

Model Name	SkySense	SkySense V2 w/o MoE	SkySense V2
Tokenizer	0.21M HR: 0.02 MS: 0.16 SAR: 0.03	0.09M HR: 0.02 MS: 0.06 SAR: 0.01	0.09M HR: 0.02 MS: 0.06 SAR: 0.01
Backbone	1260.31M HR: 655.17 MS: 302.57 SAR: 302.57	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 6. Comparison of the number of parameters in different modules between SkySense V2 and SkySense.

SkySense [10] employed three distinct backbones: Swin-H for high-resolution (HR) optical data, ViT-L for multi-

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: 1024×1024 sentinel2: 1024×1024 sentinel1: 1024×1024	gep: 4096×4096 sentinel2: 128×128 sentinel1: 128×128	sentinel2: 128×128 sentinel1: 128×128
Input channel	planet: RGBNIR sentinel2: B02-08, B8A, B11-12 sentinel1: VV, VH	gep: RGB sentinel2: B02-08, B8A, B11-12 sentinel1: VV, VH	sentinel2: B02-08, B8A, B11-12 sentinel1: VV, VH
Base learning rate	6e-05	6e-05	5e-05
Learning rate scheduler	linear	linear	MultiStepLR
Weight decay	0.01	0.01	0.01
Batch size	8	8	256
Max iteration/epoch	6k iters	20k iters	100 epoch
Warmup	linear	linear	-
Warmup iteration/epoch	150 iters	1500 iters	-
Warmup ratio	1e-6	1e-6	-
Drop path rate	0.2	0.2	0.2
Augmentation	RandomFlip	RandomFlip	RandomFlip
Head/Detector	UperNet	FCN	Linear Classifier
Loss function	CrossEntropy	CrossEntropy	MultiLabel SoftMargin

Table 5. The finetuning setting in multi-modal downstream tasks.

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 6. 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 [16], 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 [26] to transform categories into text representations and then apply dense image-text alignment (ITA) to enhance pre-trained model's capability for dense interpretation. 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 7, 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 7. 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 8, 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 8. 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) [33] from the Gaofen-2 satellite, SPARCS [14] from the Landsat-8 satellite, and AIR-PolSAR-Seg (APS) [37] 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 1000×1000 pixels, annotated into 7 categories. APS consists of a PolSAR image with a region of 9082×9805 pixels and 2000 image patches, each sized 512×512 pixels. The experimental results on these three datasets are presented in Table 9. 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-PolSAR-Seg	Gaofen-3(SAR)	53.21	55.32

Table 9. 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 10, 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 (TR=10%)	BEN-S2 (TR=10%)	BEN-MM
Activated modality	HR	MS	MS,SAR
w/o MSPT	96.15	88.97	92.64
w/ MSPT	96.42	89.13	93.81

Table 10. 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 11, 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 11. 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) [20, 40], 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 12, 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 12. Ablation results of varing number of MoE blocks in backbone.

Exploring different distributions of MoE blocks in backbone. Previous studies [20, 40] 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 [28], and 2) the last layers have the most significant impact on classification performance. However, the official implementation of Swin-MoE [15][3] 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 13.

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 13. 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 14, 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.

Dataset	RESISC-45 (TR=10%)	BEN-S2 (TR=10%)	BEN-MM
Activated modality	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 14. 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.

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. 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 15, 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.

[3] https://github.com/microsoft/Swin-Transformer

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 15. Ablation results of varing 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 [41], object detection with the DIOR dataset [17], and semantic segmentation with the iSAID dataset [38]. The experimental results, which are shown in Table 16, indicate a significant performance advantage for our pre-trained model compared to the model trained from scratch across all three datasets.

Model	AID	DIOR	iSAID
	OA(TR=20/50%)	mAP_{50}	mIoU
Randm Init	66.82/90.78	56.36	48.34
SkySense V2	98.34/99.05	79.50	71.87

Table 16. Comparison of SkySense V2 with random initialization and SkySense V2 with pre-training.

References

[1] 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. 1, 2, 3

[2] 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. 1

[3] Keumgang Cha, Junghoon Seo, and Taekyung Lee. A billion-scale foundation model for remote sensing images. *arXiv preprint arXiv:2304.05215*, 2023. 3

[4] 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. 3

[5] Wuyang Chen, Ziyu Jiang, Zhangyang Wang, Kexin Cui, and Xiaoning Qian. Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8924–8933, 2019. 2

[6] 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. 5

[7] 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. 5

[8] 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. 3

[9] Shaohua Guo, Liang Liu, Zhenye Gan, Yabiao Wang, Wuhao Zhang, Chengjie Wang, Guannan Jiang, Wei Zhang, Ran Yi, Lizhuang Ma, et al. Isdnet: Integrating shallow and deep networks for efficient ultra-high resolution segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4361–4370, 2022. 2

[10] 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 multimodal 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, 2, 3, 4, 5

[11] 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. 2

[12] 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. 2

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

[14] M. Joseph Hughes and Robert H. Kennedy. High-quality cloud masking of landsat 8 imagery using convolutional neural networks. *Remote. Sens.*, 11:2591, 2019. 7

[15] Changho Hwang, Wei Cui, Yifan Xiong, Ziyue Yang, Ze Liu, Han Hu, Zilong Wang, Rafael Salas, Jithin Jose, Prabhat Ram, Joe Chau, Peng Cheng, Fan Yang, Mao Yang, and Yongqiang Xiong. Tutel: Adaptive mixture-of-experts at scale, 2022. 8

[16] Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87, 1991. 6

[17] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A

survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307, 2020. 9

[18] Wentong Li, Yijie Chen, Kaixuan Hu, and Jianke Zhu. Oriented reppoints for aerial object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1829–1838, 2022. 3

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

[20] 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. 8

[21] 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. 5

[22] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv: Learning*, 2016. 3

[23] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017. 3

[24] 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. 4

[25] 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. 4

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 6

[27] 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. 3

[28] 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. 8

[29] 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. 4

[30] Jamie Sherrah. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. *arXiv preprint arXiv:1606.02585*, 2016. 3

[31] 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. 3

[32] Aysim Toker, Lukas Kondmann, Mark Weber, Marvin Eisenberger, Andrés Camero, Jingliang Hu, Ariadna Pregel Hoderlein, Çağlar Şenaras, Timothy Davis, Daniel Cremers, et al. Dynamicearthnet: Daily multi-spectral satellite dataset for semantic change segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21158–21167, 2022. 3

[33] 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. 7

[34] 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. 3

[35] Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Chenying Liu, Zhitong Xiong, and Xiao Xiang Zhu. Decur: decoupling common & unique representations for multimodal self-supervision. *arXiv preprint arXiv:2309.05300*, 2023. 5

[36] 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. 5

[37] 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. 7

[38] 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. 3, 9

[39] 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. 3

[40] Lemeng Wu, Mengchen Liu, Yinpeng Chen, Dongdong Chen, Xiyang Dai, and Lu Yuan. Residual mixture of experts. *ArXiv*, abs/2204.09636, 2022. 8

[41] 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. 9

[42] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understand-

ing. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 3