MaRS: A Multi-Modality Very-High-Resolution Remote Sensing Foundation Model with Cross-Granularity Meta-Modality Learning

Ruoyu Yang, Yinhe Liu*, Heng Yan, Yiheng Zhou, Yihan Fu, Han Luo, Yanfei Zhong†

Wuhan University, Wuhan, China

{yangruoyu, liuyinhe, yanheng, zhouyiheng, fuyihan, luo_han, zhongyanfei}@whu.edu.cn

Abstract

The multi-modality remote sensing foundation model (MM-RSFM) has made notable progress recently. However, most existing approaches remain limited to medium-resolution, single-modality, restricting their performance in fine-grained downstream applications such as disaster response and urban planning. In this work, MaRS is proposed, a multi-modality very-high-resolution (VHR) remote sensing foundation model designed for cross-modality granularity interpretation of complex scenes. To achieve this, a multi-modality VHR SAR-optical dataset, MaRS-16M, is constructed through large-scale collection and semi-automated processing, comprising over 16 million paired samples. Unlike previous work, MaRS tackles two fundamental challenges in VHR SAR-optical self-supervised learning (SSL) techniques. Cross-granularity contrastive learning (CGCL) is introduced to alleviate alignment inconsistencies caused by imaging differences, and meta-modality attention (MMA) is designed to unify heterogeneous physical characteristics across modalities. Compared to existing remote sensing foundation models (RSFMs) and general vision foundation models (VFMs), MaRS performs better as a pre-trained backbone across nine multi-modality VHR downstream tasks.

Code — https://rsidea.whu.edu.cn/mars.htm

Introduction

Foundation models have revolutionized natural language processing and computer vision by enabling strong generalization across tasks via large-scale pretraining (Radford et al. 2021; Devlin et al. 2019; Oquab et al. 2023). This paradigm is increasingly influencing remote sensing (RS), where remote sensing foundation models (RSFMs) promise to unify diverse earth observation tasks (Bastani et al. 2023; Cong et al. 2022). Existing MM-RSFMs such as Prithvi 2.0 (Jakubik et al. 2023; Szwarcman et al. 2024), FlexiMo (Li et al. 2025), and SkySense V2 (Zhang et al. 2025) are predominantly trained in medium-resolution data sets (e.g., Sentinel 1 / 2) or focus on single-modality inputs. Meanwhile, focusing on high-resolution foundation models remains limited to the optical modality, leaving the high-resolution SAR modality unexplored (Sun et al. 2023), which limits MM-RSFMs' effectiveness in fine-grained applications requiring high spatial resolution and robust multi-modality interpretation.

Remote sensing foundation models (RSFMs) capable of interpreting VHR-SAR imagery are urgently needed in time-sensitive and low-visibility scenarios, such as post-disaster response (Chen et al. 2025; Gupta et al. 2019; Zheng et al. 2024). During critical windows (the first 24 to 48 hours following wildfires, floods, or earthquakes), optical imagery is often obstructed by smoke, cloud cover, or night-time conditions. At the same time, medium-resolution Sentinel-1 SAR data lacks the spatial granularity needed for fine-grained interpretation. However, existing RSFMs, trained mostly on optical imagery and medium-resolution from Sentinel-1/2 and NAIP, cannot generalize to VHR-SAR modality, especially in cases where only VHR-SAR is available (Astruc et al. 2024; Li et al. 2024a). Therefore, in this paper, we are motivated to ***advance multi-modal high-resolution remote sensing foundation model*** for fine-grained interpretation in complex scenes.

To achieve this goal, we contribute from two key aspects: First, we construct the **MaRS-16M dataset** by collecting and semi-automatically processing large-scale VHR-optical and VHR-SAR data from commercial remote sensing satellite providers. Second, a novel **MaRS model** is trained and evaluated on nine multi-modal VHR downstream tasks, as shown in Figure 1.

In remote sensing, although several pioneering works have leveraged contrastive learning to build multi-modal foundation models using medium-resolution data (MM-RSFMs) (Fuller, Millard, and Green 2023; Astruc et al. 2025), extending this paradigm to very-high-resolution (VHR) remains highly challenging. Specifically, developing a multi-modal VHR remote sensing foundation model faces two key challenges: 1) Imaging discrepancy: VHR-SAR imagery is often affected by layover effects and speckle noise, resulting in severe local distortions that make pixel-level alignment with VHR-optical imagery challenging. 2) Modal representation gap: High-resolution optical and SAR images describe objects differently. Optical imagery captures texture, while SAR imagery reflects physical structure. This inherent difference leads to heterogeneous fea-

*Corresponding author

†Corresponding author

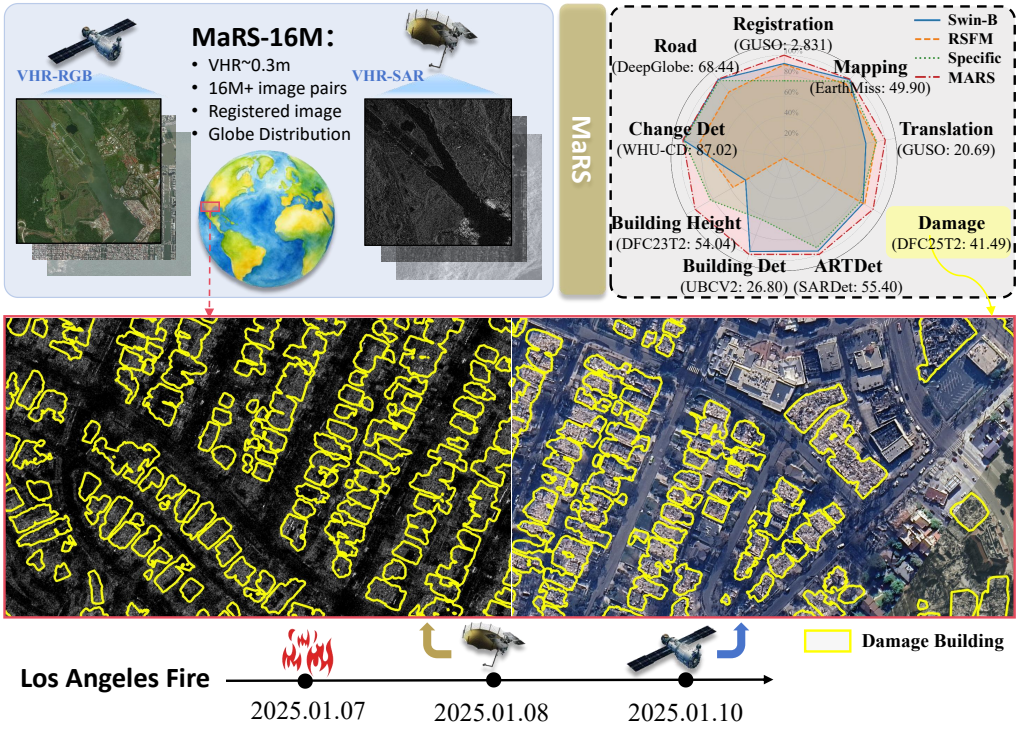


Figure 1: MaRS-16M is a globally distributed dataset of paired VHR optical and SAR imagery. The MaRS foundation model, pretrained on MaRS-16M, achieves state-of-the-art performance across nine VHR multi-modality tasks, demonstrating strong generalization to complex real-world scenarios.

tures that are difficult to align or integrate. These challenges motivate the design of MaRS, a cross-granularity and meta-modality foundation model for VHR multi-modal remote sensing. To address local discrepancies caused by imaging differences, we propose Cross-Granularity Contrastive Learning (CGCL), which performs cross-contrastive learning between patch-level and batch-level feature granularities. This design leverages global scene semantics to supervise the representation learning of local regions. Meanwhile, meta-modality attention (MMA) is introduced, incorporating an alternating intra-modality and cross-modality attention structure. This alternating design enables the Transformer to iteratively capture both modality-specific patterns and cross-modal dependencies, facilitating the emergence of a unified meta-modality representation. To validate the effectiveness of MaRS, we conducted extensive experiments on 9 VHR multi-modal tasks, comparing them against existing foundation models and specialized task-specific approaches. We summarize the main contributions as follows.

- We construct the MaRS-16M dataset, a large-scale paired VHR optical-SAR dataset comprising more than 16 million globally distributed 0.3m SAR and RGB image pairs. This dataset captures various land cover types, urban patterns, and disaster scenarios.
- We propose MaRS, a VHR multi-modal remote sensing foundation model that combines Cross-Granularity Contrastive Learning (CGCL) for local-global alignment and Meta-Modality Attention (MMA) for bridging SAR-optical heterogeneity through alternating attention design.
- We validate MaRS on nine VHR SAR-optical benchmarks, achieving strong performance across diverse tasks including cross-modal registration, cross-modal generation, land cover mapping under modality-missing, damage assessment, multi-modal object extraction, and change detection.

Related Work

Multi-modal Representation Learning

Multi-modal representation learning has attracted increasing attention due to its potential to integrate complementary information from different modalities(Caron et al. 2021; Radford et al. 2021). These models typically adopt transformer-based architectures and fall into two broad categories based on their training objectives. The first category includes masked image modeling (MIM) methods (He et al. 2022; Xie et al. 2022), which learn to reconstruct missing image patches in a self-supervised manner. These approaches capture rich semantic features by forcing the model to infer global context from partial information. The second category involves contrastive learning(Chen et al. 2020; Oquab et al. 2023), which aligns representations of different views or augmentations of the same image. More recently, hybrid methods have emerged that combine reconstruction and contrastive signals (e.g., iBOT (Zhou et al. 2021), data2vec

	Model	Dataset	Modality	Size	MM	MM-VHR
RSFMs	Prithvi	Prithvi	S2 (10m)	1TB	✗	✗
	RingMo	RingMo	GF2 (0.8m)	2M	✗	✗
	SatMAE++	fMoW + S2	S2 (10m)	883K	✗	✗
	Satlas	Satlas	NAIP (0.5∼2m)/S2 (10m)	13M	✗	✗
	ScaleMAE	fMoW-RGB	S2 (10m)	363.6K	✗	✗
	CrossMAE	fMoW-RGB	S2 (10m)	363.6K	✗	✗
MM-RSFMs	CROMA	SSL4EO	**S1**-S2 (10m)	250K	✗	✗
	DeCUR	SSL4EO	**S1**-S2 (10m)	250K	✓	✗
	USat	Satlas	NAIP (0.5-2m) - S2 (10m)	13M	✓	✗
	SkySense	SkySense	Optical (0.5m) - **S1**-S2 (10m)	21.5M	✓	✗
	DOFA	GEO-Bench	NAIP + Gaofen + **S1** / - S2 + EnMAP (1-30m)	-	✗	✗
	OmniSat	-	NAIP (0.5-2m) - **S1**-S2 (10m)	300K	✓	✗
	SeaMo	SSL4EO	**S1**-S2 (10m)	250K	✓	✗
	AnySat	-	NAIP (0.5 2m) - **S1**-S2 (10m)	-	✓	✗
	MaRS	**MaRS-16M**	**Optical (0.35m)/ Umbra & Capella (0.35m)**	**16.8M**	✓	✓

Table 1: Comparison with existing Remote Sensing Foundation Models (RSFMs). S1/S2 denote Sentinel-1 and Sentinel-2. MM: Multi-modality. MM-VHR: Multi-modality with very-high-resolution.

(Baevski et al. 2022)), aiming to integrate the strengths of both paradigms. At the same time, recent advances are dominated by vision transformers (ViT) (He et al. 2016; Woo et al. 2023; Vaswani et al. 2017; Dosovitskiy et al. 2021) and hierarchical variants (Liu et al. 2021, 2022), which support global receptive fields and multi-scale reasoning. Additionally, pretraining frameworks (Radford et al. 2021; Li et al. 2021) have pushed the boundary further by aligning visual and language modalities at scale, inspiring many cross-modal extensions. However, most of these models are trained on web-scale natural image datasets (Deng et al. 2009; Schuhmann et al. 2022)) and struggle to generalize to domain-specific modalities like SAR, which exhibit fundamentally different signal characteristics and structural priors.

Multi-modal Remote Sensing Foundation Model

Adapting foundation model techniques to remote sensing has recently gained attention, with most existing work focusing on either optical imagery or medium-resolution optical and SAR data (Jakubik et al. 2023; Guo et al. 2024). In the single-modal optical domain, models such as Prithvi (Szwarcman et al. 2024), SatMAE (Cong et al. 2022; Noman et al. 2024), and Scale-MAE (Reed et al. 2023) leverage large-scale optical imagery (e.g., Sentinel-2, NAIP) using masked autoencoding or contrastive pretraining. These models often use ViT backbones and perform strongly on tasks like land cover classification and segmentation. However, they are typically restricted to medium resolution imagery (10–30m) and lack adaptability to SAR-specific tasks. In the multi-modal domain, works like CROMA (Fuller, Millard, and Green 2023), SeaMo (Li et al. 2024a), AnySat (Astruc et al. 2024), and USat (Irvin et al. 2023) extend self-supervised learning to SAR-RGB pairs. For instance, CROMA and SeaMo use contrastive learning across Sentinel-1/2 image pairs (10m), while USat incorporates NAIP (0.5–2m) and Sentinel-2 in a shared embedding space. Most of these models rely on coarsely aligned medium-resolution multi-modal datasets (Wang et al. 2024), restricting their effectiveness in fine-grained scenarios.

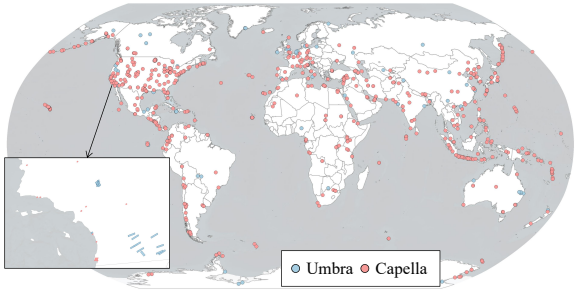


Figure 2: Dataset Distribution. MaRS-16M is composed of 2 open-source SAR companies, Umbra & Capella.

MaRS

MaRS-16M Dataset

To construct MaRS, we first collected 4,225 very-high-resolution (VHR) SAR images from Umbra and Capella Space, with acquisition dates up to December 30, 2024. After manually filtering out images with severe geometric distortion or poor quality, we retrieved the corresponding VHR optical images. To address the georeferencing misalignment between VHR optical and VHR-SAR imagery, we developed a semi-automated preprocessing pipeline, which includes a registration checker, resampling, and image registration.

Finally, the MaRS-16M dataset contains 16,785,168 paired VHR optical and SAR image patches at a spatial resolution of 0.35 meters, as shown in Figure 2. The SAR modality uses the X-band with HH or VV polarization. This dataset covers many land cover types, urban, and disaster scenarios, providing dense, precisely-aligned supervi-

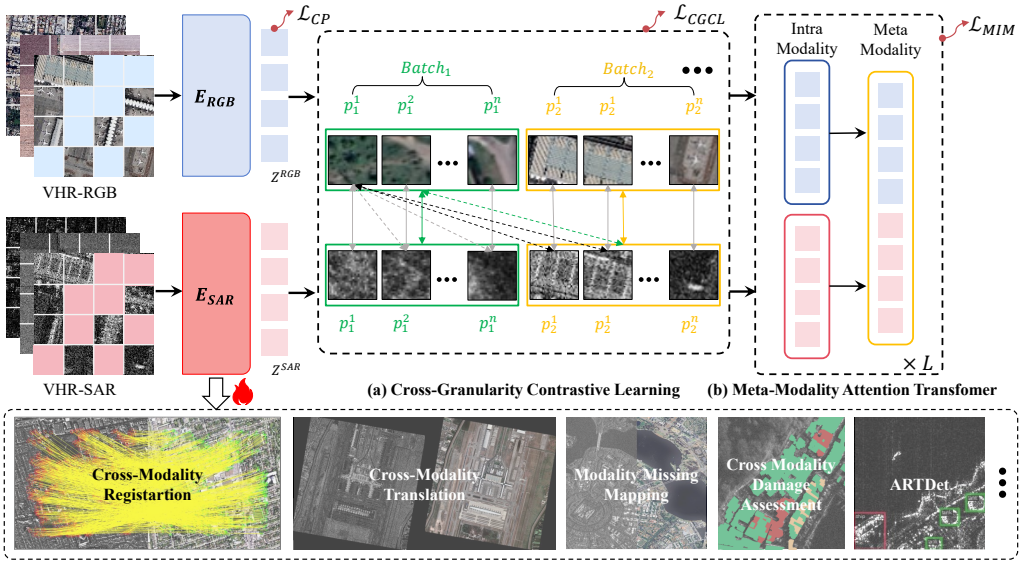


Figure 3: Overview of the MaRS pretraining pipeline. The MaRS framework first encodes VHR optical and SAR images through modality-specific encoders, followed by Cross-Granularity Contrastive Learning (CGCL) to enforce spatial consistency across patches and batches. A Meta-Modality Attention Transformer is then applied to enhance intra- and meta-modality feature representations. The model is jointly supervised by contrastive learning, masked image modeling, and continual pretraining. The encoders are utilized for downstream tasks.

sion essential for robust cross-modal learning under geometric distortion and signal noise conditions. Compared to existing foundation model training datasets, MaRS is distinguished by three key characteristics: very-high spatial resolution, precisely paired multi-modality (optical-SAR) data, and significantly larger scale, as illustrated in Table 1.

Overview Architecture

Let the input paired VHR optical and SAR image patches be denoted as $\chi^{\text{RGB}} \in \mathbb{R}^{H\times W\times 3}$ and $\chi^{\text{SAR}} \in \mathbb{R}^{H\times W\times 1}$, respectively. MaRS adopts a dual-encoder architecture to accommodate the distinct physical properties of optical and SAR imagery. Specifically, the two modality-specific encoders E_{RGB} and E_{SAR} are both implemented using Swin Transformer V2 (Liu et al. 2022), which extracts patch-wise representations $\mathbf{z}^{\text{RGB}} = E_{\text{RGB}}(\chi^{\text{RGB}})$ and $\mathbf{z}^{\text{SAR}} = E_{\text{SAR}}(\chi^{\text{SAR}})$. The encoded features are fed into the Meta Modality Attention Transformer (MMA), followed by light decoders D_{RGB} and D_{SAR} for dense prediction. Note that the $E_{\text{RGB}}, E_{\text{SAR}}, D_{\text{RGB}}, D_{\text{SAR}}$ are basic modules in masked image modeling (MIM) (Xie et al. 2022). While the MMA is newly proposed to use an alternating design to learn a unified meta-modality representation,as shown in Figure 3.

To pretrain MaRS effectively, we integrate three complementary self-supervised strategies. First, cross-granularity contrastive learning (CGCL) between paired SAR and optical representations using local and global feature granularity alignment. Second, we incorporate masked image modeling objectives within each modality branch, encouraging the encoders to model intra-modal spatial structure from partially masked inputs. Finally, inspired by GFM (Mendieta et al. 2023), we apply continued pretraining on Swin Transformer V2 backbones on the VHR-optical branch using our VHR multi-modal dataset, which has proven especially beneficial in enhancing optical representation quality and enabling downstream generalization.

Cross-Granularity Contrastive Learning

Contrastive learning has proven effective in aligning representations across different modalities by pulling positive pairs closer and pushing negative pairs apart. This paradigm has been widely applied in vision-language pretraining frameworks such as CLIP (Radford et al. 2021) and DINOv2 (Oquab et al. 2023), and recently extended to remote sensing foundation models including SkySense V2 (Zhang et al. 2025) and OmniSat (Astruc et al. 2025), where it facilitates SAR-optical representation alignment. However, two critical challenges emerge under the very-high-resolution (VHR) condition. First, due to the sparse distribution of land-cover objects in VHR imagery, the overlap between modalities is often limited, making patch-level alignment difficult. Second, severe local distortions introduced by SAR-specific effects, including speckle noise and geometric deformation, degrade the information of local features, compromising the effectiveness of contrastive learning.

To address these issues, we propose Cross-Granularity Contrastive Learning(CGCL), a hierarchical contrastive framework that jointly leverages patch-level and image-level features. The key idea is to use stable global semantics to align local patches,as shown in Figure 4.

Formally, let $\mathbf{Z}^{\text{RGB}} \in \mathbb{R}^{B\times N\times D}$ and $\mathbf{Z}^{\text{SAR}} \in \mathbb{R}^{B\times N\times D}$ denote the patch tokens extracted from VHR-RGB and VHR-SAR images, where B is the batch size, N is the number of patch tokens, and D is the feature dimension. We define

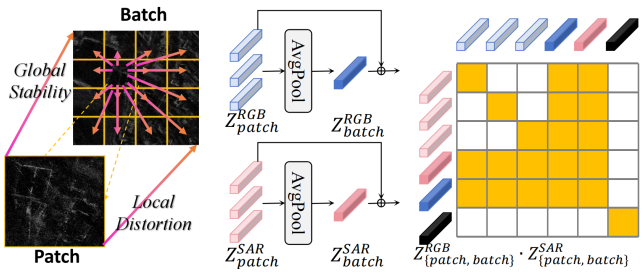


Figure 4: Illustration of Cross-Granularity Contrastive Learning (CGCL) enforces consistency between global-local patches across modalities to solve the local distortion problem.

three levels of contrastive objectives:

- **Patch-to-Patch contrast:** directly aligns local features across modalities,
- **Image-to-Image contrast:** aligns mean patch features to encourage global semantic consistency,
- **Patch-to-Global contrast:** performs hierarchical alignment by matching each patch to global context features from the other modality.

The total CGCL loss is defined as a weighted combination:

$$L_{\text{CGCL}} = \alpha \cdot L_{\text{patch}} + \beta \cdot L_{\text{global}} + \gamma \cdot L_{\text{cross}} \tag{1}$$

where α, β, and γ denote the weights of each granularity level. This cross-granularity design mitigates sensitivity to local distortions and provides robust multi-modal alignment for downstream high-resolution tasks.

Meta-Modality Attention

Inspired by the recent theoretical insights in meta-modality fusion (Zheng et al. 2021), we introduce a modality-aware attention mechanism tailored for dual-branch fusion in VHR remote sensing. We aim to dynamically integrate modality-specific and cross-modal cues while maintaining global coherence across spatially dense representations. To this end, we design a dedicated transformer module, Meta-Modality Attention (MMA), that interleaves two self-attention layers: modality-wise and global.

Given the tokenized feature sequences from E_{RGB} and E_{SAR}, denoted as $\mathbf{Z}^{\text{RGB}} \in \mathbb{R}^{B\times N\times D}$ and $\mathbf{Z}^{\text{SAR}} \in \mathbb{R}^{B\times N\times D}$, the MMA module concatenates them into a unified token stream $\mathbf{T} = [\mathbf{Z}^{\text{RGB}}; \mathbf{Z}^{\text{SAR}}] \in \mathbb{R}^{B\times 2N\times D}$. The overall module consists of L alternating layers, where each layer performs the following two-stage operation:

- **Intra-Modality Attention**: We first apply intra-modality self-attention to T_{RGB} and T_{SAR} independently. This step enables the model to capture modality-specific inductive patterns, such as structural geometry in SAR and texture-rich cues in optical imagery.
- **Meta-Modality Attention**: The full token sequence $\mathbf{T}$ is then refined via self-attention across both modalities. This layer promotes global interaction and facilitates information propagation between modalities.

These two operations alternate across L transformer blocks. Repeating the interleaved update, the MMA transformer preserves modality-specific semantic representations and bridges the modality gap through iterative alignment. The final outputs from MMA are modality-specific enhanced features $T^{\text{out}}_{\text{RGB}}$, $T^{\text{out}}_{\text{SAR}} \in \mathbb{R}^{B\times S\times C}$, which are reshaped back to the spatial domain by decoders. The entire process can be formulated as follows.

$$\boldsymbol{H}_{\text{intra}}(T^{l-1}) = F_{\text{MHA}}(T^{l-1}_{\text{RGB}}) \oplus F_{\text{MHA}}(T^{l-1}_{\text{SAR}}) \tag{2}$$

$$\boldsymbol{H}_{\text{meta}}(T^{l-1}) = F_{\text{MHA}}(T^{l-1}) \tag{3}$$

$$T^{l} = \begin{cases} \boldsymbol{H}_{\text{intra}}(T^{l-1}), & \text{if } l \bmod 2 = 1 \\ \boldsymbol{H}_{\text{meta}}(T^{l-1}), & \text{if } l \bmod 2 = 0 \end{cases} \tag{4}$$

$l \in \mathscr{L}$ represent the l-th **transformer** block. $F_{\text{MHA}}(\cdot)$ represents a standard Transformer block composed of a Layer Normalization, a Multi-Head Self-Attention module, an MLP feed-forward network, and residual connections.

Experiments

Pre-training Implementation

MaRS is pre-trained for 12 epochs with a batch size 16 per GPU on 8 A800 GPUs, totaling 128 samples per iteration. The training lasts for approximately 48 hours. Following recent masked image modeling paradigms, we set the masking ratio to 60% and apply random horizontal flipping as the primary data augmentation strategy. Each input image is resized to 512×512 before being fed into the model. The frozen teacher model is instantiated with the same architecture as the RGB encoder and provides pixel-wise guidance for accelerating convergence and stabilizing training. The experiments are conducted by fine-tuning a pretrained MaRS backbone and task-specific architecture.

Performance on Multi-modal VHR Tasks

We evaluate the proposed MaRS on seven challenging downstream tasks under complex, multi-modal, and very-high-resolution (VHR) scenarios. The used benchmarks cover a wide range of fine-grained urban and disaster perception tasks. Specifically, GUSO contains 20,000 image pairs of optical and SAR data from 441 global regions, with 0.3m resolution and standardized into 512×512 patches for cross-modality registration and translation evaluation. EarthMiss comprises 3,355 pairs of 0.6m optical and SAR images covering 13 cities across five continents, and is annotated with eight semantic categories to support modality-miss land cover mapping. DFC25T2 (Chen et al. 2025) is the first open-access, globally distributed, event-diverse dataset for building damage assessment. SARDdet-100K (Li et al. 2024b) contains over 120,000 SAR image chips with bounding box annotations for target detection. UBCv2 (Huang et al. 2023) includes multi-modal rooftop and building instance labels, and we only use the building detection annotations in this work. DFC23T2 (Huang et al. 2022) provides paired multi-modal imagery and height labels for building height estimation in urban areas.

Quantitative and qualitative comparisons are shown in Table 2 and Figure 5. Compared to VFM models trained

		Registration GUSO		Mapping EarthMiss	Translation GUSO		Damage DFC25T2	SAR ATRDet SARDet-100K		Building Detection UBCV2		Height DFC23T2
Model	Backbone	RMSE↓	NCM↑	mIoU↑	PSNR↑	SSIM↑	IoU↑	mAP↑	mAP75↑	mAP↑	mAP50↑	Delta↑
ResNet	R50	-	-	48.0	18.75	0.32	34.77	48.4	53.1	17.1	34.1	-
ViT	ViT-L	-	-	48.7	17.13	0.27	37.00	-	-	-	-	23.26
SwinV2	SwinV2-B	3.08	1788.5	49.6	16.71	0.22	36.82	53.5	58.7	26.0	44.7	23.38
DoFA	DoFA-B	-	-	46.8	-	-	37.59	-	-	-	-	30.99
SatMAE	ViT-L	3.52	1607.3	35.0	15.32	0.22	29.39	-	-	-	-	14.19
ScaleMAE	ViT-L	3.22	1688.5	35.1	15.72	0.22	-	-	-	-	-	7.11
GFM	Swin-B	-	-	39.6	16.02	0.22	28.24	-	-	-	-	21.60
CROMA	CROMA-B	-	-	39.2	-	-	-	-	-	-	-	18.30
CrossMAE	ViT-B	3.12	1786.5	39.4	15.63	0.22	29.49	-	-	-	-	12.57
Prithvi	PriV2-300M	-	-	-	15.27	0.21	-	-	-	-	-	22.28
Satlas	Swin-B	-	-	-	15.72	0.21	37.06	-	-	-	-	24.45
SSL4EO	R50	-	-	-	18.99	0.34	-	-	-	-	-	-
Task-specific		*SuperGlue*		*Semantic*	*HFGAN*		*SiamCRNN*	*GridRCNN*		*Cascade*		*DFC23-B*
		3.91	36.19	48.66	18.934	0.310	35.57	51.5	56.3	16.5	26.9	30.10
		XoFTR		-	-	-	-	*FCOS*		*Mask+CGT*		*Light*
		3.77	1019.23	-	-	-	-	48.5	51.4	16.3	26.3	38.30
		TopicFM		-	-	-	-	*Deform-DETR*		*Cascade+CGT*		*Light-G*
		4.27	21.07	-	-	-	-	51.3	54.0	17.1	27.1	44.60
MaRS	**SwinV2-B**	**2.83**	**1836.9**	**49.9**	**20.69**	**0.44**	**41.49**	**55.4**	**61.4**	**26.8**	**45.9**	**54.04**

Table 2: The multi-modality VHR downstream tasks results compared to existing foundation modal and task-specific methods.

on medium-resolution data, we observe that RSFM suffers from severe performance drops across tasks, likely due to a lack of fine-scale representation that limits generalization to complex VHR settings. Among VFM baselines, DoFA achieves relatively strong results, which suggests the importance of large model capacity and spectrum-adaptive design. SwinV2 consistently outperforms ViT-based backbones, particularly in dense prediction tasks such as target detection and height estimation, which confirms the necessity of hierarchical multi-scale features. MaRS achieves the best performance across six out of seven tasks, and even outperforms several task-specific methods. The qualitative results in Figure 5 highlight the visual advantages of MaRS, especially in cross-modal translation, where fine-grained details like farmland and building edges are better recovered.

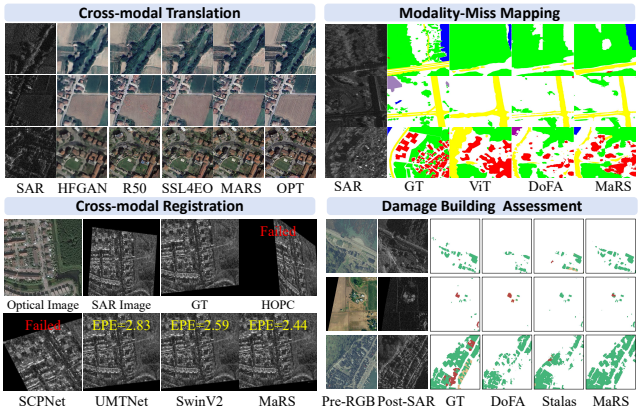


Figure 5: Visualization of the multi-modality VHR downstream tasks.

Performance on VHR-optical Tasks

In addition to multi-modal tasks, we further evaluate MaRS on high-resolution optical tasks to assess its performance under single-modal, fine-grained scenarios. WHU-CD (Ji, Wei, and Lu 2019) is a building binary change detection dataset with high-resolution (0.3 m) image pairs and pixel-wise annotations. DeepGlobe is a benchmark for road extraction from VHR optical images with diverse global coverage and fine road annotations.

As illustrated in Table 3, MaRS achieves a remarkable IoU of 86.66 on WHU-CD, significantly outperforming most vision foundation models. In comparison, ViT-based models (e.g., ViT-L, SatMAE) fall behind due to their limited ability to capture spatially localized details. SwinV2-based VFM models perform better, reflecting the advantage of hierarchical features in dense prediction tasks. MaRS surpasses task-specific baselines like STANet and DASNet, demonstrating its generalizability and fine-scale discrimination capacity even without task-specific tuning. On the DeepGlobe road extraction benchmark, MaRS attains 68.14 IoU, again outperforming other VFMs and approaching the recently task-specific performance.

Real-world Case Study on Los Angeles Fire

To validate the practical utility of MaRS in real-world disaster response, we applied our framework to assess wildfire damage during the Los Angeles Hill Fire incident on January 7, 2024. The imagery was collected by Umbra. As shown in Figure 6, MaRS successfully aligns the multi-modality images, reducing the original registration offset of

		Change Det WHU-CD (IoU↑)	Road Extraction DeepGlobe (IoU↑)
ResNet	R50	78.38	65.82
ViT	ViT-L	72.44	56.51
SwinV2	SwinV2-B	86.66	68.14
DoFA	DoFA-B	66.51	57.27
SatMAE	ViT-L	65.60	28.01
ScaleMAE	ViT-L	70.43	25.52
GFM	Swin-B	-	41.55
CROMA	CROMA-B	-	57.03
CrossMAE	ViT-B	67.83	27.24
Prithvi	PriV2-300M	72.80	46.47
Satlas	Swin-B	-	32.73
SSL4EO	R50	49.70	29.13
Task-specific		*STANet* 82.99	*LinkNet* 66.89
		DASNet 84.42	*CoANet* 66.75
		USSFCNet 85.96	*Connectivity* 67.21
MaRS	**SwinV2-B**	**87.02**	**68.44**

Table 3: The VHR-optical downstream tasks results compared to existing foundation and task-specific methods.

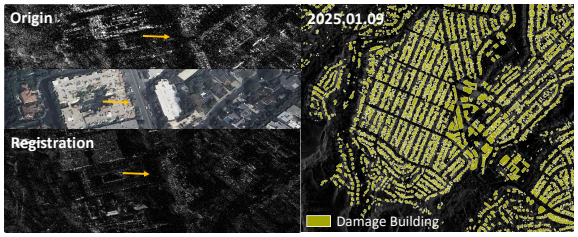
	Registration RMSE↓	Mapping mIoU↑	SAR ATRDet PSNR↑	Change Det IoU↑
BaseLine	2.835	42.0	53.5	67.1
+CGCL	2.795	49.1	53.8	76.0
+MMA	2.854	48.4	55.1	86.2
MaRS	**2.831**	**49.9**	**55.4**	**87.0**

Table 4: Ablation Study. The Baseline model adopts masked image modeling (MIM) and standard dual-encoder contrastive learning, without specialized alignment or fusion strategies.

over 100 pixels to within 10 pixels, thereby enabling precise pixel-level comparison. The damage assessment visualization on the right highlights temporal structural changes, revealing fine-grained damage patterns over 24 hours. This global stability highlights the framework's robustness to domain shifts, sensor types, and regional variations, making MaRS a practical and scalable solution for real-world multimodal remote sensing applications.

Ablation Study

To understand the contributions of each component in our framework, we conduct a systematic ablation study across four representative VHR multi-modal tasks, as shown in Table 4. The CGCL enhances the model's capability for fine-grained semantic interpretation by enforcing consistent representations across spatial hierarchies. Meanwhile, the MMA facilitates comprehensive cross-modal fusion by explicitly modeling intra- and inter-modal dependencies. When integrated, the complete MaRS model delivers consistently superior performance across all tasks, validating the effectiveness and complementarity of its components.



(a) Cross-Registration results (b) Damage assessment results

Figure 6: The cross-modality registration and modality miss building damage assessment of MaRS pretraining weights on the Los Angeles Fire.

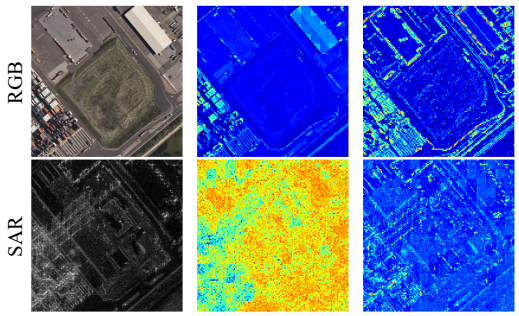


Figure 7: Feature heatmap of the output feature representation from the final block of backbone.

Modality-Aware Feature Analysis

To further validate the advantages of MaRS in fine-grained representation and modality alignment, we conduct a feature heatmap visualization experiment. Figure 7 shows that MaRS produces sharper activations along object boundaries and structural textures, indicating enhanced detail modeling capabilities. Moreover, compared to SwinV2, which relies on unimodal features, MaRS exhibits more consistent activation regions across RGB and SAR inputs, demonstrating superior modality-invariant representation and fusion effectiveness.

Conclusion

This paper introduces MaRS, a multi-modality remote sensing foundation model pre-trained on MaRS-16M, a large-scale, globally distributed dataset of paired VHR optical and SAR imagery. MaRS employs a unified dual-branch backbone and two core strategies, Cross-Granularity Contrastive Learning and the Meta-Modality Attention Transformer, to jointly enhance fine-grained representation and modality fusion. This design enables MaRS to generalize effectively across diverse VHR tasks and achieve state-of-the-art performance on nine challenging benchmarks, ranging from registration and translation to change detection and building analysis. Beyond its strong benchmark results, MaRS demonstrates real-world utility in complex disaster scenarios such as damage assessment.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 42325105.

References

Astruc, G.; Gonthier, N.; Mallet, C.; and Landrieu, L. 2024. AnySat: an Earth observation model for any resolutions, scales, and modalities.

Astruc, G.; Gonthier, N.; Mallet, C.; and Landrieu, L. 2025. OmniSat: self-supervised modality fusion for Earth observation. In *Computer Vision – ECCV 2024*, volume 15086, 409–427. Cham: Springer Nature Switzerland.

Baevski, A.; Hsu, W.-N.; Xu, Q.; Babu, A.; Gu, J.; and Auli, M. 2022. data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language.

Bastani, F.; Wolters, P.; Gupta, R.; Ferdinando, J.; and Kembhavi, A. 2023. SatlasPretrain: a large-scale dataset for remote sensing image understanding. 16772–16782.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Chen, H.; Song, J.; Dietrich, O.; Broni-Bediako, C.; Xuan, W.; Wang, J.; Shao, X.; Wei, Y.; Xia, J.; Lan, C.; et al. 2025. BRIGHT: A globally distributed multimodal building damage assessment dataset with very-high-resolution for all-weather disaster response. *Earth System Science Data Discussions*, 2025: 1–51.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PmLR.

Cong, Y.; Khanna, S.; Meng, C.; Liu, P.; Rozi, E.; He, Y.; Burke, M.; Lobell, D.; and Ermon, S. 2022. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35: 197–211.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv:2010,11929 [cs]*.

Fuller, A.; Millard, K.; and Green, J. 2023. CROMA: remote sensing representations with contrastive radar-optical masked autoencoders. *Adv. Neural Inf. Process. Syst.*, 36: 5506–5538.

Guo, X.; Lao, J.; Dang, B.; Zhang, Y.; Yu, L.; Ru, L.; Zhong, L.; Huang, Z.; Wu, K.; Hu, D.; et al. 2024. 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*, 27672–27683.

Gupta, R.; Hosfelt, R.; Sajeev, S.; Patel, N.; Goodman, B.; Doshi, J.; Heim, E.; Choset, H.; and Gaston, M. 2019. xbd: A dataset for assessing building damage from satellite imagery. *arXiv preprint arXiv:1911.09296*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollar, P.; and Girshick, R. 2022. Masked Autoencoders Are Scalable Vision Learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15979–15988. New Orleans, LA, USA: IEEE.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, X.; Chen, K.; Tang, D.; Liu, C.; Ren, L.; Sun, Z.; Hänsch, R.; Schmitt, M.; Sun, X.; Huang, H.; et al. 2023. Urban building classification (UBC) V2—A benchmark for global building detection and fine-grained classification from satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–16.

Huang, X.; Ren, L.; Liu, C.; Wang, Y.; Yu, H.; Schmitt, M.; Hänsch, R.; Sun, X.; Huang, H.; and Mayer, H. 2022. Urban building classification (ubc)-a dataset for individual building detection and classification from satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1413–1421.

Irvin, J.; Tao, L.; Zhou, J.; Ma, Y.; Nashold, L.; Liu, B.; and Ng, A. Y. 2023. USat: A unified self-supervised encoder for multi-sensor satellite imagery. *arXiv preprint arXiv:2312.02199*.

Jakubik, J.; Roy, S.; Phillips, C.; Fraccaro, P.; Godwin, D.; Zadrozny, B.; Szwarcman, D.; Gomes, C.; Nyirjesy, G.; Edwards, B.; et al. 2023. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660*.

Ji, S.; Wei, S.; and Lu, M. 2019. Fully Convolutional Networks for Multisource Building Extraction From an Open Aerial and Satellite Imagery Data Set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1): 574–586.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705.

Li, X.; Hong, D.; Li, C.; and Chanussot, J. 2024a. Seamo: A multi-seasonal and multimodal remote sensing foundation model. *CoRR*.

Li, X.; Li, C.; Ghamisi, P.; and Hong, D. 2025. Fleximo: A flexible remote sensing foundation model. *arXiv preprint arXiv:2503.23844*.

Li, Y.; Li, X.; Li, W.; Hou, Q.; Liu, L.; Cheng, M.-M.; and Yang, J. 2024b. Sardet-100k: Towards open-source

benchmark and toolkit for large-scale sar object detection. *Advances in Neural Information Processing Systems*, 37: 128430–128461.

Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. 2022. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12009–12019.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9992–10002. New York: IEEE.

Mendieta, M.; Han, B.; Shi, X.; Zhu, Y.; and Chen, C. 2023. Towards Geospatial Foundation Models via Continual Pretraining. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 16760–16770. Paris, France: IEEE.

Noman, M.; Naseer, M.; Cholakkal, H.; Anwer, R. M.; Khan, S.; and Khan, F. S. 2024. Rethinking transformers pre-training for multi-spectral satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27811–27819.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763.

Reed, C. J.; Gupta, R.; Li, S.; Brockman, S.; Funk, C.; Clipp, B.; Keutzer, K.; Candido, S.; Uyttendaele, M.; and Darrell, T. 2023. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4088–4099.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35: 25278–25294.

Sun, X.; Wang, P.; Lu, W.; Zhu, Z.; Lu, X.; He, Q.; Li, J.; Rong, X.; Yang, Z.; Chang, H.; He, Q.; Yang, G.; Wang, R.; Lu, J.; and Fu, K. 2023. RingMo: A Remote Sensing Foundation Model With Masked Image Modeling. *IEEE Trans. Geosci. Remote Sensing*, 61: 1–22.

Szwarcman, D.; Roy, S.; Fraccaro, P.; Gíslason, . E.; Blumenstiel, B.; Ghosal, R.; de Oliveira, P. H.; Almeida, J. L. d. S.; Sedona, R.; Kang, Y.; et al. 2024. Prithvi-eo-2.0: A versatile multi-temporal foundation model for earth observation applications. *arXiv preprint arXiv:2412.02732*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Y.; Albrecht, C. M.; Braham, N. A. A.; Liu, C.; Xiong, Z.; and Zhu, X. X. 2024. Decoupling common and unique representations for multimodal self-supervised learning. In *European Conference on Computer Vision*, 286–303. Springer.

Woo, S.; Debnath, S.; Hu, R.; Chen, X.; Liu, Z.; Kweon, I. S.; and Xie, S. 2023. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16133–16142.

Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; and Hu, H. 2022. SimMIM: a simple framework for masked image modeling. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9643–9653. New Orleans, LA, USA: IEEE.

Zhang, Y.; Ru, L.; Wu, K.; Yu, L.; Liang, L.; Li, Y.; and Chen, J. 2025. SkySense V2: A Unified Foundation Model for Multi-modal Remote Sensing. *arXiv preprint arXiv:2507.13812*.

Zheng, Z.; Ma, A.; Zhang, L.; and Zhong, Y. 2021. Deep multisensor learning for missing-modality all-weather mapping. *ISPRS Journal of Photogrammetry and Remote Sensing*, 174: 254–264.

Zheng, Z.; Zhong, Y.; Zhang, L.; Burke, M.; Lobell, D. B.; and Ermon, S. 2024. Towards transferable building damage assessment via unsupervised single-temporal change adaptation. *Remote Sensing of Environment*, 315: 114416.

Zhou, J.; Wei, C.; Wang, H.; Shen, W.; Xie, C.; Yuille, A.; and Kong, T. 2021. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*.