RoMA: Scaling up Mamba-based Foundation Models for Remote Sensing

Fengxiang Wang[1], Yulin Wang[2], Mingshuo Chen [3], Haiyan Zhao[2], Yangang Sun[2], Shuo Wang[2], Hongzhen Wang[2] , Di Wang[4,5]*, Long Lan[1], Wenjing Yang [1]*, Jing Zhang[4]*

[1] College of Computer Science and Technology, National University of Defense Technology, China
[2] Tsinghua University, China [3] Beijing University of Posts and Telecommunications, China
[4] School of Computer Science, Wuhan University, China [5] Zhongguancun Academy, China

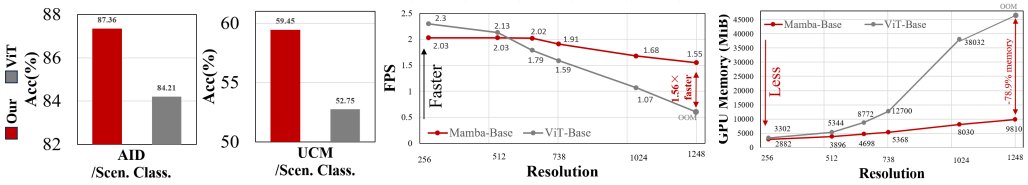


Figure 1: Comparison of ViT [1] (pretrained with MAE [2]) and our Mamba model (pretrained with RoMA) in scene classification, change detection, and semantic segmentation. Mamba outperforms ViT while being more computationally and memory efficient for high-resolution images. Notably, Mamba-B achieves 1.56× faster inference and reduces GPU memory usage by 78.9% on 1248×1248 resolution images (6084 tokens per image) on a single NVIDIA 4090 GPU (batch size = 2).

Abstract

Recent advances in self-supervised learning for Vision Transformers (ViTs) have fueled breakthroughs in remote sensing (RS) foundation models. However, the quadratic complexity of self-attention poses a significant barrier to scalability, particularly for large models and high-resolution images. While the linear-complexity Mamba architecture offers a promising alternative, existing RS applications of Mamba remain limited to supervised tasks on small, domain-specific datasets. To address these challenges, we propose RoMA, a framework that enables scalable self-supervised pretraining of Mamba-based RS foundation models using large-scale, diverse, unlabeled data. RoMA enhances scalability for high-resolution images through a tailored auto-regressive learning strategy, incorporating two key innovations: 1) a rotation-aware pretraining mechanism combining adaptive cropping with angular embeddings to handle sparsely distributed objects with arbitrary orientations, and 2) multi-scale token prediction objectives that address the extreme variations in object scales inherent to RS imagery. Systematic empirical studies validate that Mamba adheres to RS data and parameter scaling laws, with performance scaling reliably as model and data size increase. Furthermore, experiments across scene classification, changing detection, and semantic segmentation tasks demonstrate that RoMA-pretrained Mamba models consistently outperform ViT-based counterparts in both accuracy and computational efficiency. The source code and pretrained models were released at RoMA.

1 Introduction

Over the past decade, advancements in remote sensing (RS) technology and more efficient data acquisition have significantly enhanced applications in ecosystem monitoring [3], and natural disaster management [4]. These applications rely on crucial model capabilities such as scene classification [5],

*Corresponding authors

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

object detection [6], change detection [7], and semantic segmentation [8]. However, training solely on limited task-specific data restricts the scale and generalizability of current RS deep learning models.

Recent breakthroughs in self-supervised learning (SSL) [2, 9] have led to the development of RS foundation models (RSFMs) [10, 11, 12, 13, 14, 15, 16] that offer robust feature representations and excel across various remote sensing tasks. However, many of these tasks involve high-resolution imagery—such as the 4,000×4,000 pixel images in the DOTA dataset for object detection. Most RSFMs rely on Vision Transformer (ViT)-based attention architectures, whose quadratic complexity limits their practicality on high-resolution data. To overcome this challenge, researchers are exploring pretraining RSFMs on architectures with linear complexity, with Mamba [17] emerging as a promising alternative.

The Mamba architecture is well-regarded in remote sensing for its efficient inference with high-resolution images in downstream tasks [18, 19]. However, current Mamba-based studies are limited to small-scale training datasets, restricting their exposure to diverse remote sensing data. This contrasts with trends in ViT-based RSFMs, which use self-supervised pretraining to harness extensive unlabeled data. Therefore, exploring self-supervised learning for Mamba to harness large-scale remote sensing data—and thereby compete with ViT—presents a promising, yet underexplored, direction.

Autoregressive pretraining [20, 21] offers a principled solution to Mamba's sequence continuity challenges by representing images as 1-D sequences and employing next-token prediction. Its causal token dependencies naturally align with Mamba's unidirectional linear-time scanning, preserving spatial coherence without the disruptions introduced by masking. While this approach has been successfully applied to natural images [20, 21, 22], RS images present unique challenges that remain largely unaddressed. We highlight three key challenges: (1) RS images often contain sparsely distributed foreground objects amid complex backgrounds. (2) Unlike objects in natural images, which typically maintain fixed orientations due to gravity, overhead RS images feature objects at varying orientations. (3) The wide range of object sizes in RS images complicates the extraction of effective representations. These challenges naturally lead to the question of whether Mamba-based RSFMs can scale efficiently with both increasing model size and larger data volumes—mirroring the performance improvements observed in self-supervised pretrained ViT architectures [23, 24].

To address these challenges, we propose Rotation-aware Multi-scale Autoregressive learning (RoMA), a framework that enables scalable self-supervised pretraining of Mamba-based RSFMs using large-scale, diverse, unlabeled data. Specifically, RoMA enhances scalability for high-resolution images through a tailored autoregressive learning strategy, incorporating two key innovations: (1) a rotation-aware pretraining mechanism combining adaptive cropping with angular embeddings to handle sparsely distributed objects with arbitrary orientations. By identifying key regions for rotation augmentation, it enhances rotation-invariant representation learning. Additionally, it embeds angle information during rotated cropping and requires the model to predict angular changes during autoregressive pretraining, further reinforcing rotation-invariant visual representations; and (2) multi-scale token prediction objectives that address the extreme variations in object scales inherent to RS imagery. By aggregating predicted token information across multiple spatial scales, this strategy helps Mamba capture more complete and structurally robust object representations during autoregressive pretraining.

Building on RoMA, we investigate its potential for pretraining Mamba-based RSFMs. Through systematic empirical studies, we confirm that Mamba aligns with RS data and parameter scaling laws, exhibiting reliable performance improvements as model and data size increase. Additionally, experiments across scene classification, changing detection, and semantic segmentation tasks show that RoMA-pretrained Mamba models consistently surpass its ViT-based counterparts in both accuracy and computational efficiency.

The contributions of this study are as follows:

(1) We introduce RoMA, the first self-supervised autoregressive pretraining framework for Mamba architectures in remote sensing, enabling efficient scaling to high-resolution RS imagery. RoMA validates that Mamba-based RSFMs follow scaling laws, achieving consistent performance gains with larger models and datasets.

(2) We propose a dynamic rotation-aware mechanism that integrates adaptive region cropping and angle-aware embeddings. By guiding the model to predict angular variations dur-

ing autoregressive learning, it effectively addresses rotational diversity and sparse target distributions, enhancing rotation-invariant feature learning.

(3) We design a multi-scale prediction objective that addresses the extreme variations in object scales, enabling the model to learn robust object representations for downstream tasks.

2 Related Work

Remote Sensing Foundation Models. While vast amounts of RS data exist, much of it remains unlabeled and thus inaccessible for supervised learning [25]. Recently, self-supervised learning frameworks have been employed to learn representations for tasks such as scene classification, object detection, and semantic segmentation, with methods falling into generation-based [2] and contrastive learning-based [26] categories. Notably, GASSL [27] and CACo [28] utilize spatio-temporal information, while SeCo [29] focuses on multiple Earth locations at different timestamps. Beyond representation learning, rotation-aware detection has also been widely studied in RS. ReDet [30] introduces a rotation-equivariant backbone, CSL [31] addresses angle boundary issues by turning regression into fine-grained classification, and S2A-Net [32] enhances detection accuracy by aligning features with rotated anchors. Most recent work in RS has primarily focused on Masked Image Modeling (MIM), categorized by general image knowledge [33], large parameter scales [34], spatio-temporal information [11], and multi-sensor data [35, 36, 37, 38], with multi-scale methods [13, 14, 15] improving performance. A recent study further explores plain ViT as a remote sensing foundation model by introducing a Rotated Varied-Size Attention (RVSA) mechanism to better handle arbitrarily oriented objects [39]. MA3E [40] incorporates angle factors into MIM training. In parallel, multimodal foundation models have emerged to bridge heterogeneous RS modalities. CROMA [41] combines contrastive radar–optical pretraining with masked reconstruction to learn rich multimodal RS representations, while AnySat [42] adopts a JEPA-based joint-embedding framework with scale-adaptive encoders to unify various resolutions, scales, and modalities. Despite these advances, most methods focus on ViT-based RSFMs and MIM pretraining, while the Mamba-based autoregressive models remain unexplored.

Vision Mamba in Remote Sensing. Recently, the Mamba architecture has excelled in NLP and has been adapted to the vision domain to address visual problems. Vision Mamba (Vim) [17] uses Vim blocks, consisting entirely of Mamba layers, with forward and backward scanning to model bidirectional representations. Vmamba [43] incorporates Visual State Space (VSS) blocks, combining Mamba and 2D convolution layers, supported by a Swin Transformer [44]-like pyramid architecture. The vision Mamba architecture has also expanded into remote sensing, producing various Mamba-based projects, categorized into four types: classification, detection, segmentation, and others. For classification, SSMamba [45] and SpectralMamba [18] handle hyperspectral data, while RSMamba [46] focuses on visible light. Detection methods like ChangeMamba [19] and RSCaMa [47] focus on change detection. Segmentation methods include Samba [48] and RS-Mamba [49], which use Mamba alone. Despite the growing research in RS using Mamba, current work is limited to supervised training on small-scale datasets, not fully exploiting the vast RS data.

Self-Supervised Learning in Vision. Inspired by the success of self-supervised learning in NLP, visual self-supervised learning methods are thriving in three main categories: contrastive learning [26, 50], autoregressive learning [20, 51, 52], and Masked Image Modeling (MIM) [9, 2]. Current research on MIM focuses on regression targets and masking strategies. Various targets include discrete tokens [53], HOG features [54], deep features [55], and frequencies [56], have already been explored. However, MIM methods often face training issues while pretraining the Mamba architeture [22]. Recently, autoregressive pretraining in the visual domain has been explored. Most works, like VAR [21], have explored the application of autoregression in image generation. We are more focused on the work of autoregression in self-supervised pretraining. iGPT [20] first highlighted the potential of autoregressive pretraining as a general self-supervised visual representation strategy. SAIM [51], RandSAC [57] and AIM [52] explored further. These works mainly focus on pretraining the ViT series and have not explored pretraining the Mamba series. ARM [22] firstly explored the compatibility of autoregressive pretraining with Mamba on the ImageNet [58] dataset, but it has not considered the specific issues of the RS field, like the rotation-invariant representations and various sizes information in RS images Notably, while ARM has explored models of different sizes on natural images, it has not evaluated Mamba's autoregressive pretraining performance across varying data

scales. In the RS field, we are the first to establish the relationships between Mamba's pretraining performance with data volume, and model size .

3 Method

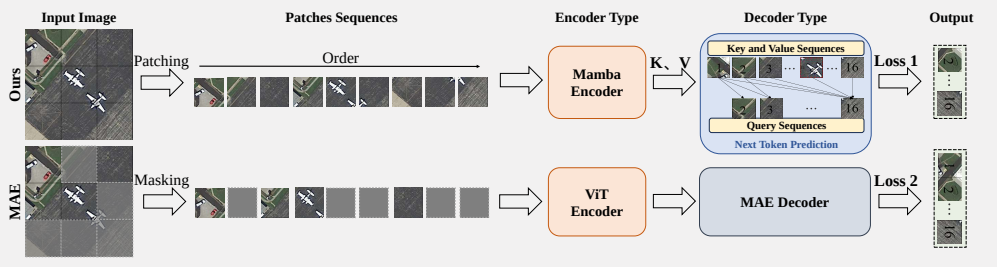


Figure 2: **Comparison between our autoregressive pretraining strategy and the standard MAE method.** (1) RoMA encodes all patches using a Mamba encoder, whereas MAE encodes only a randomly sampled subset. (2) RoMA predicts the next token in a sequence to capture continuity, while MAE only reconstructs masked patches.

3.1 Preliminaries

Autoregressive Model. Considering a sequence of discrete tokens $x = (x_1, x_2, \ldots, x_N)$, where $x_n \in [S]$ is an integer from a vocabulary of size S. The next-token autoregressive model posits that the probability of observing the current token x_n depends only on its preceding tokens $(x_1, x_2, \ldots, x_{n-1})$. This unidirectional token dependency assumption enables the factorization of the sequence x's likelihood as follows:

$$p(x_1, x_2, \ldots, x_N) = \prod_{n=1}^{N} p(x_n \mid x_1, x_2, \ldots, x_{n-1}). \tag{1}$$

Training an autoregressive model p_ϕ involves optimizing $p_\phi(x_n \mid x_1, x_2, \ldots, x_{n-1})$ over a dataset. This process, known as the next-token prediction, allows the trained p_ϕ to generate new sequences.

MAE-based Pretraining of Mamba. Previous work [10, 16, 11, 15, 13] primarily used MAE-based methods for pretraining Remote-Sensing Fundamental Models (RSFMs), where ViT is served as their visual backbones in often.

3.2 RoMA: Rotation-aware Multi-scale Autoregressive learning

We propose the RoMA autoregressive pretraining framework for the Mamba architecture in RS field. Specially, as shown in the Figure 3, we extend the iGPT [20] series with a KV cache-based prediction method. The Mamba-Encoder processes the entire image to compute the Key and $Value$ for all tokens. Then we calculate the learnable query vector from the Key and $Value$, and compute the loss between the $Query$ and the target ground truth. Building on the autoregressive pretraining structure for natural images, RoMA introduces two key contributions: an adaptive rotation encoding strategy and a multi-scale prediction strategy.

3.2.1 Auto-regressive Pre-training of Mamba on RS Imagery

Disadvantages of MAE-based Pre-training of Mamba. First, we compare properties of general visual pre-training tasks and RS tasks, together with reflection on *why MAE-based pre-training are suboptimal choice towards RS imagery data:*

- **Explosion of visual tokens on high-resolution RS data.** As informed by Section 1, high-solution RS imagery exhibits numerous visual tokens. In contrast, the quadratic complexity of ViT-based MAT pre-training protocols is computational infeasible towards increasing visual tokens in high-resolution RS tasks (see detailed comparisons on speed, GPU usage and accuracy in Figure. 1).

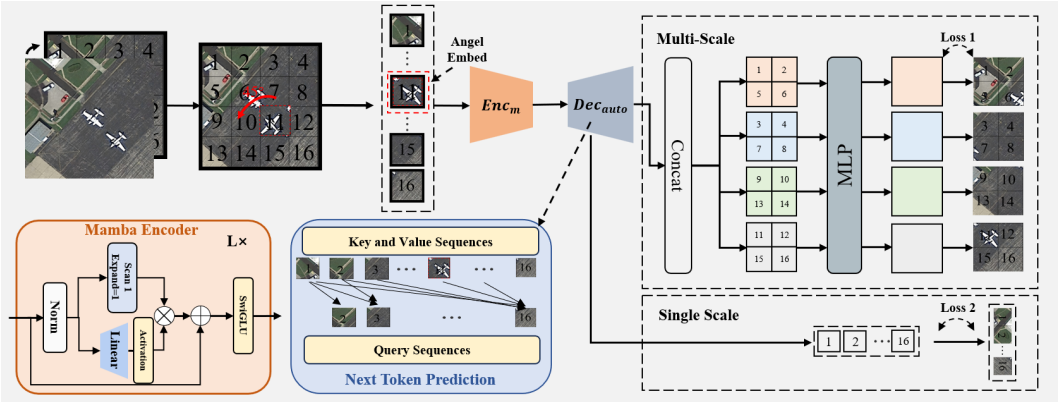


Figure 3: **Overview of the RoMA Pretraining Pipeline.** The input image is first divided into patches, and high-value patches are selected for random rotation using the Adaptive Rotation Encoding Strategy. These patches are then tokenized and processed by the Mamba encoder. The encoded features undergo autoregressive next-token prediction, followed by a multi-scale strategy that computes loss at different scales for gradient updates.

- **Disruption of MAE on Visual Tokens in Mamba's Architecture.** To be specific, MAE learns semantic representations by a first-masking-then-reconstructing pipeline. However, we observe that the mask operation in MAE disrupts between adjacent tokens, i.e., breaking sequential orders among tokens. As illustrated in Figure 2, the mask operation conflicts with the linear scanning operation embodied in Mamba, which aggregates temporally related tokens but not randomly related tokens.

Advantage of Autoregressive Pre-training of Mamba. Autoregressive pre-training aligns naturally with the sequential nature of Mamba's architecture, which processes input tokens in a temporally ordered manner. Specifically, autoregressive modeling constructs image patches sequentially and predicts the next token based on previous context, mirroring Mamba's token-by-token scanning mechanism. This architectural alignment facilitates more coherent temporal dependencies and better token transition modeling, allowing Mamba to learn more structured and semantically meaningful representations. Therefore, the autoregressive training paradigm not only complements Mamba's design but also enhances its ability to model spatial continuity and visual context in remote sensing imagery.

3.2.2 Adaptive Rotation Encoding Strategy

Rotation-Invariant Pre-training is critical for RS Data. As shown in Figure 4, RS images contain redundant airport runway pixels, and varying airplane angles lead to different postures and shapes. RS images often contain redundant airport runway pixels, while airplanes appear at various orientations, leading to different postures and shapes. Such directional diversity has been extensively addressed in supervised change detection through rotation-equivariant or rotation-invariant designs [30, 31, 32]. However,Autoregressive pretraining for natural images does not consider the uneven, sparse information distribution and rotational invariance in RS images. For instance, as shown in Figure 4, RS images contain redundant airport runway pixels, and varying airplane angles lead to different postures and shapes. These unique characteristics prompt us to rethink how the encoder can learn high knowledge density features with rotational invariance. In RoMA, we outline an adaptive rotation encoding strategy to enhance autoregressive pretraining for remote sensing. RoMA omits explicit angle prediction. Instead, angle embeddings introduce directional priors that help the model learn rotation-invariant representations during pre-training, without supervision from angle labels.

1. Split the input image.
2. Associate each patch x^p with a score.
3. Selecting the patch (16×16) with the highest score.
4. Compute all 96×96 candidate boxes containing the patch and select the one with the highest

value.
5. Compare the 96×96 patch to the image-wide patch mean. If it exceeds the mean, select it; otherwise, proceed to 64×64 patches until one surpasses the mean.

We then detail each step outlined above for rotation-invariant encoding strategy: (1) **Step 1 of ARES:** We split the input image. $x \in \mathbb{R}^{H\times W\times C}$ into $N = (H \times W)/p^2$ non-overlapping patches $x^p \in \mathbb{R}^{N\times(p^2C)}$, where p is the patch size, (H, W) is the size of the input image, and C is the number of channels; (2) **Step 2 of ARES:** We associate each patch x^p with a score, computed via a efficient feature descriptor F, e.g., LBP [59], and then select the token with the highest values; (3) **Step 3 of ARES:** Let $token_{top}$ denote the selected token. Then, centered on the patch represented by $token_{top}$, we expand its size to obtain a larger, more suitable region. (4) **Step 4 of ARES:** After extracting a square token region with an edge length of $L = \{96, 64, 32\}$, we compute its average feature value and compare it with the average feature value of each patch in the original image; (5) **Step 5 of ARES:** If $token_L$ has a higher average feature value, it is identified as a high-value region and proceeds to the rotation step. Otherwise, a smaller L is selected, and step 4 is reapplied. This process repeats up to three times until a region with a higher average feature value than the original image is reached.

Migrating Information Loss. The selected patch is cropped to generate diverse rotated remote sensing scenes, while potential information loss on the edge pixels might occur. To mitigate this, we follow MA3E [40] and apply center cropping to retain the inscribed square (marked in yellow) within the largest inscribed circle. This region, oriented in any direction, replaces the original scene and introduces explicit angular variations. In addition to positional embeddings, we also incorporates learnable angle embeddings shared across patches within the rotated crop, i.e., served as implicit cues, aiding the model perceive angular changes while distinguishing them from the background.

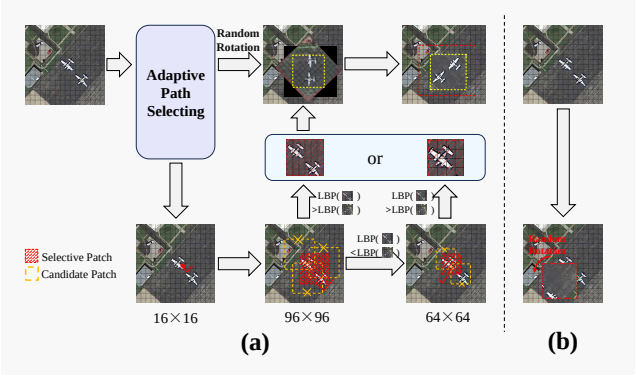


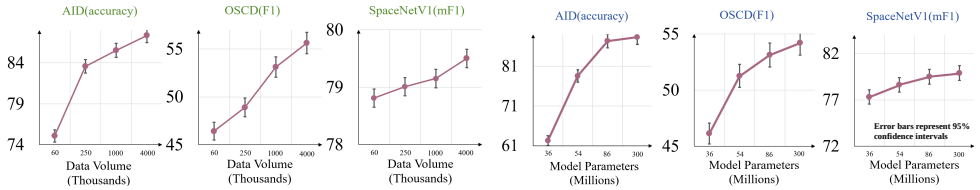
Figure 4: **Illustration of the Adaptive Rotation Encoding Strategy.** (a) Pipeline of the Adaptive Rotation Encoding Strategy. LBP refers to Local Binary Pattern. (b) Random patch selection for rotation without adaptive selection. The random approach in (b) disrupts object information in the RS image.

Finally, the Adaptive Rotation Encoding Strategy processes the image before feeding it into the Mamba-based encoder for representation learning. RoMA follows the standard Mamba architecture [22] without modifications, focusing purely on pretraining Mamba for RS field. While architectural improvements could enhance performance and efficiency, RoMA prioritizes pretraining strategies, leaving further Mamba optimizations to future research.

3.2.3 Multi-scale Prediction Strategy

Images are continuous 2D signals. To apply autoregressive pretraining via next-token prediction, two steps are required: (1) Convert images into discrete tokens. (2) Define them as a one-dimensional sequence for unidirectional modeling. Methods like iGPT [20] and VAR [21] tackle these challenges by slicing images into segments and arranging them into a feature sequence in a specified one-dimensional order.

However, as discussed in Section 1, directly applying this method to RS images fails to consider key factors. Unlike natural images, which focus on visual semantic understanding, RS images focus on surface measurement information [60]. Arbitrarily disrupting spatial relationships in RS images leads to fundamentally different interpretations. For example, token $x^{(i,j)}$ and its neighbors $x^{(i\pm1,j)}$, $x^{(i,j\pm1)}$ are closely related due to planar surface measurements. As seen in Figure 2, the token $x^{(i,j)}$



(a) Data Volume Scaling (b) Model Size Scaling

Figure 5: **Scaling with Data Volume and Model Size.** Each experiment was conducted three times, and the average was reported as the final result. (a) We showcase the Mamba-Base model's performance on three downstream tasks after RoMA pretraining with different data scales. (b) We compare the performance of various Mamba model sizes on three downstream tasks, all pretrained with 4 million data using RoMA. Details on pretraining and downstream task configurations are provided in Section 4.

representing the part of plane is closely related to vertical neighboring tokens and some distant vertical tokens. The unidirectional flattening of autoregressive methods compromises surface information representation in RS images. Therefore, we introduce a multi-scale prediction strategy to mitigate the effects of unidirectional flattening.

The Mamba-based encoder generates key and $value$ feature representations for each token. During decoding, we apply cross-attention with token-level causal masking to sequentially predict tokens, ensuring each token relies only on previously observed ones. The decoder performs autoregressive prediction at the token level. After the decoder generates the reconstructed outputs, the autoregressive module aligns them with the corresponding regions of the original image for supervision. The reconstruction quality is optimized using the mean squared error (MSE) loss between the predicted and original pixel values.

Building on MSE loss function, we concatenate token representations from each decoder block at a higher scale following a predefined raster order. A fully connected multi-layer perceptron (MLP) then reconstructs the pixel values of the next block. Notably, $x_k \in \mathcal{R}^{16\times16}$, while at a higher scale, spatial aggregation results in $y_n \in \mathcal{R}^{s\times s}$. s is the pixel size value of larger scale. The formula is as follows:

$$\ell(\theta) = \frac{1}{K-1}\sum_{k=2}^{K} \|\hat{x}_k(\theta; x_{<k}) - x_k\|_2^2 + \frac{\lambda}{N-1}\sum_{n=2}^{N} \|\hat{y}_n(\theta; y_{<n}) - y_n\|_2^2 \tag{2}$$

where θ represents the network parameters, N is the number of cluster blocks in an image, y_n is the ground truth pixel value of the n-th cluster block, and $\hat{y}_n(\theta; y_{<n})$ denotes the reconstructed value based on θ and preceding tokens ($y_{<n}$), K is the number of all tokens in an image, x_k denotes the ground truth pixel value of the k-th token, and $\hat{x}_k(\theta; x_{<k})$ is the reconstructed value based on the network parameters (θ) and preceding tokens ($x_{<k}$) in the sequence. The parameter λ regulates the contribution of the MSE loss from higher-scale cluster blocks to the overall loss.

3.3 Scaling Mamba-based RSFMs

To investigate the scaling potential of the self-supervised pretrained Mamba architecture for developing powerful RSFMs. With RoMA, we analyze the relationships between Mamba's performance with model parameters, and data scale. The scalability of ViT architectures pretrained with MAE has been well established, demonstrating performance gains with increasing data volume and model size [23, 24]. However, no prior work has systematically examined whether the Mamba architecture follows a similar trend in RS field. For the first time, we explore Mamba's scaling behavior in RS domain using the RoMA pretraining method.

Table 1: **The configuration of different architecture variants.**

Model	Block	Width	Depth	Param.(M)
ViT-T	Attention+MLP	192	12	5.7
Mamba-T	Mamba+MLP	192	12	5.3
ViT-S	Attention+MLP	384	12	22
Mamba-S	Mamba+MLP	384	12	21
ViT-B	Attention+MLP	768	12	86
Mamba-B	Mamba+MLP	768	12	85
ViT-L	Attention+MLP	1024	24	307
Mamba-L	Mamba+MLP	1024	24	297

Table 2: **Results for scene classification, change detection, and semantic segmentation.** "TR" represents the ratio of training data. ⋆ indicates results from MA3E [40] and MTP [61]. † denotes our reproduction with their official code.

Methods	Publication	Backbone	Params	Scene Classification		Change Detection	Semantic Segmentation
				AID [62]	UCM [63]	OSCD [64]	SpaceNetv1 [65]
				OA(TR=50%)	OA(TR=50%)	F1	mF1
Natural Image pretraining							
MoCo v3 ⋆ [50]	ICCV'21	ViT-B	86M	78.72	38.34	-	-
DINO ⋆ [26]	ICCV'21	ViT-B	86M	78.51	40.04	-	-
MAE ⋆ [2]	CVPR'22	ViT-B	86M	84.21	52.75	-	-
SimMIM ⋆ [9]	CVPR'22	ViT-B	86M	83.19	51.48	-	-
LoMaR ⋆ [66]	Arxiv'22	ViT-B	86M	82.26	51.89	-	-
MixMAE ⋆ [67]	CVPR'23	Swin-B/W14	88M	81.53	50.63	-	-
ARM †[22]	ICLR'25	Mamba-B	85M	81.14	50.41	47.28	77.89
RS Image pretraining							
SeCo ⋆ [29]	ICCV'21	ResNet-50	25.6M	78.26	47.45	47.67	77.09
CACo ⋆ [28]	CVPR'23	ResNet-50	25.6M	77.81	40.53	52.11	77.94
SatMAE ⋆ [11]	NIPS'22	ViT-L	307M	55.10	34.28	52.76	78.07
ScaleMAE ⋆ [13]	ICCV'23	ViT-L	307M	48.46	28.19	-	-
GFM ⋆ [33]	ICCV'23	Swin-B	88M-	80.58	49.73	-	-
RVSA ⋆ [10]	TGRS'23	ViT-B+RVSA	86M	84.06	50.86	50.28	**79.56**
SatMAE++ † [15]	CVPR'24	ViT-L	307M	85.98	55.72	53.10	79.21
MA3E ⋆ [40]	ECCV'24	ViT-B	86M	85.86	55.69	-	-
RoMA	-	Mamba-B	85M	**87.36**	**59.45**	**55.63**	79.50

Scaling with Data Volume: Mamba shows a clear performance boost on downstream tasks as the pretraining data volume grows. We pretrain the Mamba-Base model with RoMA across various data scales and evaluate its performance in the downstream tasks. As illustrated in Figure 5a, larger datasets lead to significant improvements. Mamba-based RSFMs exhibit no significant performance bottlenecks across a broad pretraining data scale from 62.5K to 4M, achieving data learning capabilities on par with ViT-based RSFMs. We look forward to future advancements of data volume in remote sensing, where larger datasets can further enhance Mamba-based RSFMs through our RoMA pretraining framework.

Scaling with Model Size: Mamba's performance also improves with increasing model size. We conduct extensive pretraining on four model variants—Tiny, Small, Base, and Large—following the configurations in our code. As shown in Figure 5b, larger models consistently achieve superior results on downstream tasks. Although Mamba-Large surpasses Mamba-Base in AID dataset, its performance gain remains limited, likely due to insufficient pretraining. With only 300 epochs on 4 million samples, the training may not be adequate for a 297M-parameter model. Due to experimental constraints, we did not extend pretraining to 800 epochs as in MAE. The OSCD and SpaceNet experiments are ongoing, with updates to follow. However, these results do not alter our key findings: Mamba-based RSFMs pretrained with RoMA demonstrate performance gains as model parameters scale. While this growth remains inconclusive in more large-scale experiments, we anticipate future research will further explore Mamba's scaling potential.

4 Experiments

We pretrain Mamba extensively using RoMA and assess its effectiveness across diverse downstream tasks. Finally, we conduct thorough ablation studies on RoMA's design choices.

Pretraining Setup. Our pretraining experiment setup largely follows ARM [22]. We train both the Mamba-B on the OpticalRS-4M [16]. We adjust the input image to a size of 196×196, with a patch size of 16, using the AdamW optimizer and a cosine learning rate scheduler. The initial learning rate is set to 1.5e-4, and batch size is set to 256, with a epoch of 400.

Downstream Tasks. We further evaluated RoMA across three key downstream tasks: scene classification, changing detection, and semantic segmentation. In addition to benchmarking against ViT-based RSFMs, we compared RoMA with other pretraining methods for natural images. These encompass methods leveraging contrastive learning and generative learning and autogressive pretraining approaches, ARM [22]. The Mamba-B architectures strictly adhere to the simplest Mamba design from ARM, without any modifications, allowing us to exhaustively test the advantages of the RoMA pretraining framework.

Table 3: **Ablation study on the design choices of RoMA with Mamba-B backbone.** We report the top-1 accuracy (%). The default settings of RoMA are highlighted in grey.

(a) **Main ablation.** Adaptive Rotation Encoding Strategy (ARE) and Multi-scale Prediction Strategy (MSP) significantly improve RoMA.

ARE	MSP	AID	
		OA (TR=20%)	OA (TR=50%)
		69.59	76.80
✓		71.70	78.00
✓	✓	72.69	79.16

(b) **Feature Descriptor** in Adaptive Rotation Encoding Strategy. Local Binary Pattern (LBP) measurement outperforms the Wavelet Transform and Histogram of Oriented Gradients (HOG).

Feature Descriptor	AID	
	OA (TR=20%)	OA (TR=50%)
Wavelet	71.42	77.00
HOG	71.94	78.32
LBP	71.70	78.00

(c) **Selecting Patch Size** in Adaptive Rotation Encoding Strategy. Three layers is the most effective choice.

Patch Size	AID	
	OA (TR=20%)	OA (TR=50%)
96	70.48	76.82
96-32	71.23	77.12
96-64-32	71.70	78.00

(d) **Threshold** for patch selection in the Adaptive Rotation Encoding Strategy is based on the image's overall average computed ($Avg.$) from the Feature Descriptor.

Threshold	AID	
	OA (TR=20%)	OA (TR=50%)
$1.5\times Avg.$	68.33	72.39
$0.5\times Avg.$	69.71	74.18
$1.0\times Avg.$	71.70	78.00

(e) **Coefficient λ.** The variation of the coefficient λ in Multi-scale Prediction Strategy. λ balances autoregressive reconstruction and sparsity regularization.

Coefficient λ	AID	
	OA (TR=20%)	OA (TR=50%)
0.01	71.81	78.23
1.0	70.92	77.49
0.1	72.69	79.16

(f) **Various Scales** choices in Multi-scale Prediction Strategy. Experiments were conducted using the standard 16×16 patch as a baseline, with additional combinations at multiples of 2–6.

Multi-scale Prediction Strategy	AID	
	OA (TR=20%)	OA (TR=50%)
2×+4×+6×	63.17	71.34
2×+4×	68.06	74.74
4×+6×	67.81	74.32
4×	72.46	78.72
6×	72.69	79.16

Scene Classification. By freezing the model's parameters and fine-tuning only the final fully connected (FC) layer, linear probing effectively demonstrates its feature extraction ability. Since full-parameter fine-tuning already achieves over 99% performance on classification datasets, like AID [62], we prefer linear probing rather than fine tuning. We use two scene classification datasets: AID [62] and UCM [63], with training details, including the train-test split ratio, following [10, 13]. Evaluation is based on overall accuracy (OA). The results in Table 2 show RoMA's competitive performance compared to other pretraining methods.

Change Detection. We used the OSCD [64] dataset consisting of RGB images for change detection. Following previous works [61], we kept the experimental setups consistent, using UNet [68] as the decoder. On the OSCD dataset, our method outperforms ARM [22] and other methods that overlook rotational invariance and information sparsity and varying object size issues in RS.

Semantic Segmentation. We further evaluate the pretrained model on semantic segmentation tasks, using common remote sensing datasets: SpaceNetv1 [65]. Our implementation follows [61], using UperNet [69] as the segmentation framework. However, in pixel-level tasks, Mamba-based RSFMs show less pronounced advantages compared to other downstream tasks. We attribute this to RoSA's autoregressive pretraining, which prioritizes multi-scale patch-based targets over pixel-level prediction.

Ablation Study. Due to resource constraints, we conducted the ablation experiments using the MillionAID [70] dataset and trained for 400 epochs. Table 3a presents the performance of RoMA's two main contributions, with ARM as the baseline. The experiment shows a significant gain in feature extraction capability. Tables 3b and 3c adding the Adaptive Rotation Encoding Strategy on the baseline. Table 3b comparing the effects of different Feature Descriptors. We believe the Feature Descriptor method can be further optimized without affecting the Adaptive Rotation Encoding Strategy's effectiveness. Table 3c evaluates patch sizes for rotation, showing that various sizes improve performance. Tables 3d, 3e, and 3f examine the Multi-scale Prediction Strategy. The parameter and threshold selections are shown in Tables 3d and 3e, while Table 3f presents the performance of aggregating spatial features from multiple scales. Our results indicate that adding excessive multi-scale information doesn't guarantee improved performance; instead, using only large-scale aggregated information along with the original 16×16 patch data yields better results.

Table 4: **Peak GPU memory usage (MB) across different input resolutions.**

Resolution	768	1024	1248	1520	2048	3072	4096
RoMA-Base	2693	4504	6526	9434	16934	37485	66357
ViT-Base	4229	11726	24531	52106	OOM	OOM	OOM

Table 5: **Inference speed (samples/sec) across different input resolutions.**

Resolution	768	1024	1248	1520	2048	3072	4096
RoMA-Base	24.98	15.91	11.43	7.86	4.37	2.00	1.15
ViT-Base	22.11	9.94	4.99	2.57	OOM	OOM	OOM

5 Further Analyses

Scalability to Ultra-High-Resolution Inputs. To further examine RoMA's scalability, we evaluated it on inputs ranging from 768×768 to 4096×4096. Both RoMA-Base and ViT-Base were tested for GPU memory usage and inference speed on a single NVIDIA A100 (batch size = 1). As shown in Table 4 and Table 5, RoMA scales stably up to 4096×4096, while ViT-Base fails beyond 2048×2048. These results further verify RoMA's computational efficiency and suitability for ultra-high-resolution remote sensing imagery.

Ability to Learn Small Targets. To further analyze RoMA's ability to capture local information, we evaluate its performance on small-object categories in the iSAID dataset [71]. We compare UperNet [69] with different backbones, following the RingMo [72] fine-tuning protocol. As shown in Table 6.Our method achieves the highest overall mIoU and also shows notable improvements on small-object classes, especially *Small Vehicle* (average width 15 pixels), where RoMA surpasses all other backbones. These results suggest that the proposed **adaptive region cropping strategy** effectively increases the visibility of small foreground objects during pretraining, allowing the model to learn more discriminative representations.

Table 6: **Fine-tuning performance on iSAID [71] (following RingMo [72] settings).**

Method	Backbone	mIoU	Ship (33^2 px)	Small Vehicle (15^2 px)	Swimming Pool (34^2 px)	Plane (53^2 px)
UperNet	IMP-ResNet-50	61.9	65.9	48.8	44.5	83.8
UperNet	SeCo-ResNet-50	57.2	63.9	44.8	9.3	83.3
UperNet	RSP-ResNet-50	61.6	64.2	47.5	43.8	82.8
UperNet	ViT-B+RVSA	63.8	68.9	51.9	46.7	85.6
UperNet	ViTAE-B+RVSA	63.5	69.6	51.9	47.5	85.4
UperNet	RingMo	67.2	73.5	51.2	48.9	85.7
UperNet	**RoMA-B (Ours)**	**67.4**	**73.8**	**53.7**	**51.8**	**86.0**

Effectiveness of Top-k Region Selection. To further validate the effectiveness of the top-k token selection mechanism, we conducted an additional experiment on the SpaceNet V1 [65] building segmentation dataset. This experiment aims to examine whether the proposed adaptive region cropping can consistently capture target regions compared with random cropping under the same resolution settings. As shown in Table 7, our method achieves a foreground capture accuracy of 75.09%, significantly higher than the 38.56% of random cropping. These results demonstrate that the top-k strategy effectively preserves informative regions, leading to more reliable semantic representations for autoregressive modeling.

Table 7: **Foreground object capture rate on SpaceNet V1 [65].**

Cropping Strategy	Accuracy (%)
Random Cropping	38.56
Top-k Adaptive Region Selection (Ours)	**75.09**

Feasibility of Token-space Reconstruction Loss. To evaluate the flexibility of RoMA, we compared pixel-space and token-space reconstruction losses. Following BEiT [53] and CAE [73], the token-space loss was computed using a frozen CLIP [74] teacher, where the student predicted its latent features through a lightweight MIM head. As shown in Table 8, token-space loss consistently outperforms pixel-space loss on UCM and AID, indicating its stronger semantic representation capability and the adaptability of the RoMA framework to diverse pretraining objectives.

Table 8: **Comparison between pixel-space and token-space reconstruction loss.**

Pretraining Loss Type	UCM-55 (%)	AID-55 (%)
RoMA-Base (Pixel Space)	52.39	80.34
RoMA-Base (Token Space)	**56.67**	**81.72**

6 Conclusion

We introduce RoMA, the first self-supervised autoregressive framework to scale Mamba-based foundation models for RS. By leveraging large-scale, diverse, unlabeled data, RoMA enables scalable self-supervised pretraining of Mamba-based RS models. Extensive experiments across scene classification, changing detection, and semantic segmentation show that RoMA-pretrained Mamba models consistently outperform ViT-based counterparts. Additionally, these models achieve significant efficiency improvements, reducing GPU memory consumption by 78.9% and accelerating inference speed by 1.56× at 1,248×1,248 resolution, while maintaining linear computational scaling. Our findings also provide new insights into Mamba's scaling laws, demonstrating consistent performance gains with increasing data volume and model size, highlighting its potential for large-scale Earth observation tasks. **Limitations.** The current RoMA framework is evaluated mainly on optical imagery, and future work will extend it to multi-source remote sensing data (e.g., SAR and hyperspectral) .

Acknowledgements: This work was partially supported by the National Natural Science Foundation of China (No. 62372459, No.62376282 and No. 624B2109).

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[2] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 15979–15988. IEEE, 2022.

[3] Nathalie Pettorelli, Henrike Schulte to Bühne, Ayesha Tulloch, Grégoire Dubois, Cate Macinnis-Ng, Ana M Queirós, David A Keith, Martin Wegmann, Franziska Schrodt, Marion Stellmes, et al. Satellite remote sensing of ecosystem functions: Opportunities, challenges and way forward. *Remote Sensing in Ecology and Conservation*, 4(2):71–93, 2018.

[4] Olalekan Mumin Bello and Yusuf Adedoyin Aina. Satellite remote sensing as a tool in disaster management and sustainable development: Towards a synergistic approach. *Procedia-Social and Behavioral Sciences*, 120:365–373, 2014.

[5] Liang Huang, Fengxiang Wang, Yalun Zhang, and Qingxia Xu. Fine-grained ship classification by combining cnn and swin transformer. *Remote Sensing*, 14(13):3087, 2022.

[6] Wentong Li, Yijie Chen, Kaixuan Hu, and Jianke Zhu. Oriented reproints for aerial object detection. In *CVPR*, pages 1829–1838, June 2022.

[7] Curtis E Woodcock, Thomas R Loveland, Martin Herold, and Marvin E Bauer. Transitioning from change detection to monitoring with remote sensing: A paradigm shift. *Remote Sensing of Environment*, 238:111558, 2020.

[8] 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.

[9] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, pages 9653–9663, 2022.

[10] 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, 2023.

[11] 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.

[12] Fanglong Yao, Wanxuan Lu, Heming Yang, Liangyu Xu, Chenglong Liu, Leiyi Hu, Hongfeng Yu, Nayu Liu, Chubo Deng, Deke Tang, Changshuo Chen, Jiaqi Yu, Xian Sun, and Kun Fu. Ringmo-sense: Remote sensing foundation model for spatiotemporal prediction via spatiotemporal evolution disentangling. *IEEE Trans. Geosci. Remote. Sens.*, 61:1–21, 2023.

[13] 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 *ICCV*, pages 4065–4076. IEEE, 2023.

[14] Maofeng Tang, Andrei Cozma, Konstantinos Georgiou, and Hairong Qi. Cross-scale mae: A tale of multiscale exploitation in remote sensing. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[15] Mubashir Noman, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwer, Salman H. Khan, and Fahad Shahbaz Khan. Rethinking transformers pre-training for multi-spectral satellite imagery. In *CVPR*, pages 27811–27819. IEEE, 2024.

[16] Fengxiang Wang, Hongzhen Wang, Di Wang, Zonghao Guo, Zhenyu Zhong, Long Lan, Jing Zhang, Zhiyuan Liu, and Maosong Sun. Scaling efficient masked autoencoder learning on large remote sensing dataset. *arXiv preprint arXiv:2406.11933*, 2024.

[17] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *ICML*, 2024.

[18] Jing Yao, Danfeng Hong, Chenyu Li, and Jocelyn Chanussot. Spectralmamba: Efficient mamba for hyperspectral image classification. *arXiv preprint arXiv:2404.08489*, 2024.

[19] Hongruixuan Chen, Jian Song, Chengxi Han, Junshi Xia, and Naoto Yokoya. Changemamba: Remote sensing change detection with spatiotemporal state space model. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–20, 2024.

[20] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 1691–1703. PMLR, 2020.

[21] Keyu Tian, Yi Jiang, Zehuan Yuan, BINGYUE PENG, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 84839–84865. Curran Associates, Inc., 2024.

[22] Sucheng Ren, Xianhang Li, Haoqin Tu, Feng Wang, Fangxun Shu, Lei Zhang, Jieru Mei, Linjie Yang, Peng Wang, Heng Wang, Alan Yuille, and Cihang Xie. Autoregressive pretraining with mamba in vision. In *ICLR*, 2025.

[23] Mannat Singh, Quentin Duval, Kalyan Vasudev Alwala, Haoqi Fan, Vaibhav Aggarwal, Aaron Adcock, Armand Joulin, Piotr Dollar, Christoph Feichtenhofer, Ross Girshick, Rohit Girdhar, and Ishan Misra. The effectiveness of mae pre-pretraining for billion-scale pretraining. In *ICCV*, pages 5484–5494, October 2023.

[24] Cheng-Ze Lu, Xiaojie Jin, Qibin Hou, Jun Hao Liew, Ming-Ming Cheng, and Jiashi Feng. Delving deeper into data scaling in masked image modeling. *arXiv preprint arXiv:2305.15248*, 2023.

[25] Di Wang, Jing Zhang, Bo Du, Gui-Song Xia, and Dacheng Tao. An empirical study of remote sensing pretraining. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–20, 2023.

[26] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9630–9640, 2021.

[27] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *ICCV*, pages 10161–10170, 2021.

[28] Utkarsh Mall, Bharath Hariharan, and Kavita Bala. Change-aware sampling and contrastive learning for satellite images. In *CVPR*, pages 5261–5270, 2023.

[29] Oscar Mañas, Alexandre Lacoste, Xavier Giró-i-Nieto, David Vázquez, and Pau Rodríguez. Seasonal contrast: Unsupervised pre-Training from uncurated remote sensing data. In *ICCV*, pages 9394–9403. IEEE, 2021.

[30] Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. Redet: A rotation-equivariant detector for aerial object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2786–2795, 2021.

[31] Xue Yang and Junchi Yan. Arbitrary-oriented object detection with circular smooth label. In *European conference on computer vision*, pages 677–694. Springer, 2020.

[32] Jiaming Han, Jian Ding, Jie Li, and Gui-Song Xia. Align deep features for oriented object detection. *IEEE transactions on geoscience and remote sensing*, 60:1–11, 2021.

[33] Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen. Towards geospatial foundation models via continual pretraining. In *ICCV*, pages 16806–16816, 2023.

[34] Keumgang Cha, Junghoon Seo, and Taekyung Lee. A billion-scale foundation model for remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pages 1–17, 2024.

[35] 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 *CVPR*, pages 27672–27683, June 2024.

[36] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, Antonio Plaza, Paolo Gamba, Jón Atli Benediktsson, and Jocelyn Chanussot. Spectralgpt: Spectral remote sensing foundation model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(8):5227–5244, 2024.

[37] Xuyang Li, Danfeng Hong, and Jocelyn Chanussot. S2mae: A spatial-spectral pretraining foundation model for spectral remote sensing data. In *CVPR*, pages 24088–24097, 2024.

[38] Vishal Nedungadi, Ankit Kariryaa, Stefan Oehmcke, Serge Belongie, Christian Igel, and Nico Lang. Mmearth: Exploring multi-modal pretext tasks for geospatial representation learning. In *ECCV*, pages 164–182. Springer, 2024.

[39] 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.

[40] Zhihao Li, Biao Hou, Siteng Ma, Zitong Wu, Xianpeng Guo, Bo Ren, and Licheng Jiao. Masked angle-aware autoencoder for remote sensing images. In *ECCV*, pages 260–278. Springer, 2025.

[41] Anthony Fuller, Koreen Millard, and James Green. Croma: Remote sensing representations with contrastive radar-optical masked autoencoders. *Advances in Neural Information Processing Systems*, 36:5506–5538, 2023.

[42] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. Anysat: One earth observation model for many resolutions, scales, and modalities. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19530–19540, 2025.

[43] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 103031–103063. Curran Associates, Inc., 2024.

[44] 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 *ICCV*, pages 9992–10002. IEEE, 2021.

[45] Lingbo Huang, Yushi Chen, and Xin He. Spectral-spatial mamba for hyperspectral image classification. *Remote Sensing*, 16(13), 2024.

[46] Keyan Chen, Bowen Chen, Chenyang Liu, Wenyuan Li, Zhengxia Zou, and Zhenwei Shi. Rsmamba: Remote sensing image classification with state space model. *IEEE Geoscience and Remote Sensing Letters*, pages 1–5, 2024.

[47] Chenyang Liu, Keyan Chen, Bowen Chen, Haotian Zhang, Zhengxia Zou, and Zhenwei Shi. Rscama: Remote sensing image change captioning with state space model. *IEEE Geosci. Remote. Sens. Lett.*, 21:1–5, 2024.

[48] Qinfeng Zhu, Yuanzhi Cai, Yuan Fang, Yihan Yang, Cheng Chen, Lei Fan, and Anh Nguyen. Samba: Semantic segmentation of remotely sensed images with state space model. *Heliyon*, 10(19):e38495, 2024.

[49] Sijie Zhao, Hao Chen, Xueliang Zhang, Pengfeng Xiao, Lei Bai, and Wanli Ouyang. Rs-mamba for large remote sensing image dense prediction. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024.

[50] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9726–9735. Computer Vision Foundation / IEEE, 2020.

[51] Yu Qi, Fan Yang, Yousong Zhu, Yufei Liu, Liwei Wu, Rui Zhao, and Wei Li. Exploring stochastic autoregressive image modeling for visual representation. In Brian Williams, Yiling Chen, and Jennifer Neville, editors, *AAAI*, pages 2074–2081. AAAI Press, 2023.

[52] Alaaeldin El-Nouby, Michal Klein, Shuangfei Zhai, Miguel Angel Bautista, Vaishaal Shankar, Alexander Toshev, Joshua M Susskind, and Armand Joulin. Scalable pre-training of large autoregressive image models. In *ICML*, ICML'24. JMLR.org, 2024.

[53] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *ICLR*, 2021.

[54] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, pages 14668–14678, 2022.

[55] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image bert pre-training with online tokenizer. In *ICLR*, 2021.

[56] Hao Liu, Xinghua Jiang, Xin Li, Antai Guo, Yiqing Hu, Deqiang Jiang, and Bo Ren. The devil is in the frequency: Geminated gestalt autoencoder for self-supervised visual pre-training. In *AAAI*, volume 37, pages 1649–1656, 2023.

[57] Tianyu Hua, Yonglong Tian, Sucheng Ren, Michalis Raptis, Hang Zhao, and Leonid Sigal. Self-supervision through random segments with aautoregressive coding (randsac). In *ICLR*, 2022.

[58] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009.

[59] Marko Heikkilä, Matti Pietikäinen, and Cordelia Schmid. Description of interest regions with local binary patterns. *Pattern recognition*, 42(3):425–436, 2009.

[60] Mi Zhang, Bingnan Yang, Xiangyun Hu, Jianya Gong, and Zuxun Zhang. Foundation model for generalist remote sensing intelligence: Potentials and prospects. *Science Bulletin*, 69(23):3652–3656, 2024.

[61] Di Wang, Jing Zhang, Minqiang Xu, Lin Liu, Dongsheng Wang, Erzhong Gao, Chengxi Han, Haonan Guo, Bo Du, Dacheng Tao, and Liangpei Zhang. Mtp: Advancing remote sensing foundation model via multi-task pretraining. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pages 1–24, 2024.

[62] 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.

[63] Maryam Mehmood, Ahsan Shahzad, Bushra Zafar, Amsa Shabbir, and Nouman Ali. Remote sensing image classification: A comprehensive review and applications. *Mathematical problems in engineering*, 2022(1):5880959, 2022.

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

[65] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018.

[66] Jun Chen, Ming Hu, Boyang Li, and Mohamed Elhoseiny. Efficient self-supervised vision pretraining with local masked reconstruction. *arXiv preprint arXiv:2206.00790*, 2022.

[67] Jihao Liu, Xin Huang, Jinliang Zheng, Yu Liu, and Hongsheng Li. Mixmae: Mixed and masked autoencoder for efficient pretraining of hierarchical vision transformers. In *CVPR*, pages 6252–6261, 2023.

[68] 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.

[69] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, pages 418–434, 2018.

[70] Yang Long, Gui-Song Xia, Shengyang Li, Wen Yang, Michael Ying Yang, Xiao Xiang Zhu, Liangpei Zhang, and Deren Li. On creating benchmark dataset for aerial image interpretation: Reviews, guidances and million-aid. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:4205–4230, 2021.

[71] 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.

[72] Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang, Hao Chang, Qinglin He, Guang Yang, Ruiping Wang, Jiwen Lu, and Kun Fu. RingMo: A remote sensing foundation model with masked image modeling. *IEEE Trans. Geosci. Remote. Sens.*, 61:1–22, 2023.

[73] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *International Journal of Computer Vision*, 132(1):208–223, 2024.

[74] 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 *ICML*, 2021.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our abstract and introduction accurately reflect our contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations and future directions at the end of the paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide analysis for each theoretical result presented in the paper, including formulas where necessary and visualizations when appropriate.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All experimental results are reproducible. Detailed information is provided in the appendix. We also plan to open-source our dataset and code to further contribute to the community.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide detailed descriptions of the dataset and experimental setup in the paper. Due to time constraints and submission size limits, the data and code cannot be included at this stage, but we commit to open-sourcing all datasets and code as soon as possible.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All results are reproducible, with details in the appendix. We will also open-source our dataset and code to support the community.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We conducted extensive experiments in strict accordance with prior work in the field and commit to open-sourcing our code to ensure reproducibility.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Similar to previous work, we used 16–24 A100 GPUs as computing resources. Variations in computational resources do not affect our experimental results. Detailed information is provided in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We strictly adhered to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the limitations of our work and directions for future research in the final section of the paper.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We did not collect any potentially sensitive data from the internet. All provided data comply with established protocols in the remote sensing imagery community, ensuring there are no security concerns.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All data, models, and API utilized in this paper are publicly available and appropriately cited. The license and terms of use are properly respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide a new dataset, which will be released in the future for use by the research community.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We only used LLMs to assist with writing, complying with the LLM policy. For benchmark evaluation, MLLMs were employed as assessment tools, following a widely accepted evaluation protocol in this field.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.