Can Generative Geospatial Diffusion Models Excel as Discriminative Geospatial Foundation Models?

Yuru Jia[1,2] Valerio Marsocci[3] Ziyang Gong[4] Xue Yang[5]
Maarten Vergauwen[1] Andrea Nascetti[2]

[1]KU Leuven [2]KTH [3]ESA Φ-lab [4]Shanghai AI Lab [5]SJTU

Abstract

Self-supervised learning (SSL) has revolutionized representation learning in Remote Sensing (RS), advancing Geospatial Foundation Models (GFMs) to leverage vast unlabeled satellite imagery for diverse downstream tasks. Currently, GFMs primarily focus on discriminative objectives, such as contrastive learning or masked image modeling, owing to their proven success in learning transferable representations. However, generative diffusion models—which demonstrate the potential to capture multi-grained semantics essential for RS tasks during image generation—remain underexplored for discriminative applications. This prompts the question: can generative diffusion models also excel and serve as GFMs with sufficient discriminative power? In this work, we answer this question with ***SatDiFuser****, a framework that transforms a diffusion-based generative geospatial foundation model into a powerful pretraining tool for discriminative RS. By systematically analyzing multi-stage, noise-dependent diffusion features, we develop three fusion strategies to effectively leverage these diverse representations. Extensive experiments on remote sensing benchmarks show that SatDiFuser outperforms state-of-the-art GFMs, achieving gains of up to +5.7% mIoU in semantic segmentation and +7.9% F1-score in classification, demonstrating the capacity of diffusion-based generative foundation models to rival or exceed discriminative GFMs. Code will be released.*

1. Introduction

Self-supervised learning (SSL) has emerged as a pivotal paradigm in computer vision, enabling models to learn robust representations without relying on labeled data. This capability is especially valuable for remote sensing (RS), where vast amounts of unlabeled satellite imagery can be leveraged for downstream tasks like land-cover classification and change detection [32, 53]. Modern SSL frameworks, such as contrastive learning [7, 41], self-distillation [6, 14], and masked image modeling (MIM) [18, 62], have driven substantial progress in developing Geospatial Foundation Models (GFMs) [35, 57], significantly advancing RS image analysis. Despite this progress, prevalent SSL paradigms exhibit inherent limitations under the RS context. Contrastive learning relies on constructing informative positive and negative pairs, a challenging task in complex RS imagery, and its global instance-level supervision tends to overlook spatially fine-grained details. Similarly, MIM's patch-level reconstruction objective may produce overly coarse representations, especially in many RS scenes dominated by homogeneous regions, where masking becomes trivial and limits learning effectiveness. These drawbacks restrict the adaptability of current SSL approaches to more complex and multi-scale geospatial data, as recently revealed by [35]. Motivated by these limitations, we explore an alternative yet underexplored SSL framework in geospatial domains: generative diffusion models. Diffusion models inherently perform self-supervision through a denoising process that models the underlying data distribution, offering a promising pathway toward capturing richer semantic representations from satellite imagery.

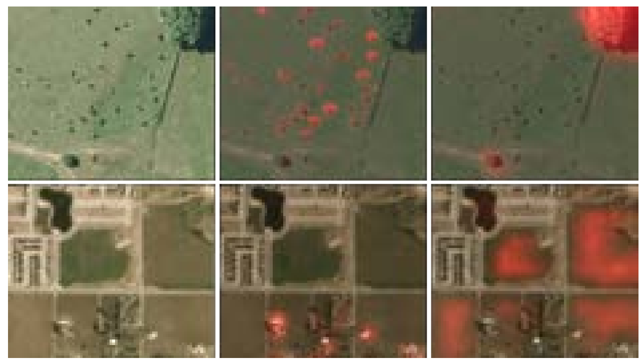


Figure 1. **Self-attention maps** from an **off-the-shelf** geospatial generative diffusion model [23] on satellite images. Semantically similar objects strongly attend to each other at different scales, highlighting the potential of generative diffusion models for discriminative tasks in remote sensing.

Diffusion models [19, 43, 46] have demonstrated extraordinary image generation capabilities by corrupting data with noise in a forward process and learning a reverse process to recover original data. Recent efforts [23, 28, 49, 66] have explored diffusion-based generative foundation models to produce high-fidelity RS scenes. While diffusion models have primarily been adopted for image synthesis, emerging works prove that these generative approaches can learn meaningful semantic representations [9, 11, 31, 64]. We further hypothesize that diffusion models provide distinct advantages for representation learning in RS imagery. During iterative denoising, these models simultaneously consider both global semantic structures and fine-grained local details, —necessary for coherent image synthesis. This aligns particularly well with RS data, where images typically contain objects and regions spanning various scales. As illustrated by Fig. 1, pretrained diffusion models naturally form meaningful self-attention patterns at multiple scales, from sparse cattle pixels to broader objects like trees and agricultural fields. Additionally, diffusion models explicitly model noisy data distributions, potentially providing robustness against sensor noise and atmospheric interference commonly present in RS data, enabling the extraction of more reliable representations [29].

These advantages indicate that generative diffusion models have untapped potential for discriminative tasks. However, one of the challenges preventing the widespread adoption of diffusion models for such tasks, both in RS and CV, is the lack of a unified and effective framework to leverage these fruitful features. Existing approaches adapt diffusion-derived representations differently depending on the task, leading to inconsistent performance and limiting broader applicability. In this work, we bridge this gap by repurposing diffusion-based generative models for self-supervised representation learning in RS. We introduce **SatDiFuser**, a flexible framework designed to efficiently harness multi-stage diffusion features, unlocking their full discriminative power for various RS tasks.

Specifically, we systematically analyze how noise-level–dependent features across diffusion stages affect performance on diverse RS tasks. To mitigate task-dependent feature selection, SatDiFuser hierarchically explore three feature fusion strategies: (i) a global weighted fusion for a broad aggregation, (ii) a localized weighted approach for input-dependent, fine-grained selection, and (iii) a mixture-of-experts design jointly modeling inter-timestep and inter-module relationships. When benchmarked against top GFMs pretrained via alternative objectives, SatDiFuser demonstrates superior accuracy on classification and semantic segmentation tasks, confirming the efficacy of generative diffusion models as a powerful SSL framework for RS. While we employ DiffusionSat [23]—a latent diffusion model (LDM) pretrained at scale on satellite imagery—as our backbone, SatDiFuser can be extended to other diffusion architectures, laying the foundation for a broader integration of diffusion-based generative modeling into geospatial analysis.

In summary, our **contributions** are threefold: **First**, to the best of our knowledge, we are the first to comprehensively adapt a large-scale diffusion-based generative model for self-supervised representation learning in RS, forming a diffusion-driven GFM. **Second**, we propose three efficient multi-stage feature fusion strategies, offering global weighted fusion, localized weighted fusion, and a mixture-of-experts fusion to maximize the discriminative power of diffusion-based features. **Third**, by benchmarking SatDiFuser against leading GFMs across various RS tasks, we show that diffusion-driven GFMs offer notable advantages, paving the way for broader synergies between diffusion-based generative modeling and discriminative geospatial analysis.

2. Related Work

Diffusion Models for Representation Learning. A number of recent works [9, 11, 22, 31, 38, 50, 60, 64, 67] in computer vision have started investigating the discriminative representations inherent in pretrained diffusion models. Some methods mine self-attention tensors for unsupervised segmentation, either by merging attention maps (DiffSeg [50]) or by constructing affinity graphs (DiffCut [9]). Others incorporate cross-attention signals (SLiMe [22]), fine-tuning text embeddings to segment objects at varied granularity. Meanwhile, DatasetDM [60] extracts multi-scale features from a Stable Diffusion UNet to train a dataset-generation model capable of producing densely annotated images. Diffusion HyperFeatures [31] further enhances feature aggregation by incorporating multi-timestep feature maps, creating a feature descriptor for semantic keypoint correspondence tasks. Additionally, REPA [65] demonstrates the improving synergies between representation learning and generative models by utilizing external high-quality representations. Inspired by these efforts, we exploit diffusion models that are trained on large-scale global satellite imagery, adapting their representational capacity to a wide range of RS tasks.

SSL for Remote Sensing. Supervised pretrained RS models (e.g., [5]) require extensive labeled data, which can be costly to obtain at scale. To circumvent this limitation, SSL has greatly advanced deep learning in RS by leveraging abundant unlabeled satellite imagery. Early efforts, such as SSL4EO-L [47] and SSL4EO-S12 [58], introduced globally distributed Landsat-8 and Sentinel-1/2 data, which have been used to train state-of-the-art SSL models like MAE [18] and DINO [6]. To address the unique characteristics of RS data, numerous studies have integrated RS-specific features, such as spatiotemporal embeddings and multi-spectral information, into SSL frameworks. These include masked image modeling (MIM)-based approaches (e.g., SatMAE [8], Scale-MAE [42], DOFA [63]), contrastive frameworks (e.g., GASSL [3], CROMA [12], SkySense [15]), and self-distillation methods (such as [34, 52]). Additionally, various multi-modal methods [2, 16, 20] extend these techniques by incorporating diverse RS modalities. Other learning strategies, including continual pretraining [37] and multi-task pretraining [13, 55], have also been explored to better adapt to satellite data. Despite this variety of RS-focused SSL methods, diffusion models remain largely unexplored as an SSL pretraining strategy. This work seeks to explore this promising direction.

Diffusion Models in RS. Diffusion models have gained increased traction in RS, being applied to image generation, enhancement, and interpretation [29]. Recent efforts [23, 48, 66, 69] have focused on developing diffusion-based generative foundation models for high-fidelity satellite image synthesis. For instance, DiffusionSat [23] generates data conditioned on semantic text and metadata, while MetaEarth [66] enables arbitrary-sized image generation using a resolution-guided approach. Beyond synthesis, numerous diffusion-based methods address image enhancement tasks, including denoising [17, 39], cloud removal [56, 71], and super-resolution [10, 54], showcasing their versatility in RS. Another line of research focuses on discriminative applications [26], though these often rely on labeled data and are limited to specific tasks, such as semantic segmentation [1, 24, 27, 40, 70] or change detection [21, 51, 59, 68]. For example, SegDiff [1] diffuse ground-truth masks, while others use class predictions or labeled guidance [24, 27, 40]. Although a few studies have explored diffusion as a label-free pretraining framework, they remain narrowly focused on a single application scenario, such as hyperspectral images segmentation [70], or change detection [4]. In contrast, our work provides a comprehensive investigation of the discriminative capabilities of diffusion-based generative models across multiple RS tasks. By moving beyond task-specific solutions and limited testing, our method advances the broader potential of diffusion-driven GFMs pretrained on global-scale data.

3. Methods

Our approach builds on DiffusionSat [23] - a satellite-adapted LDM based on Stable Diffusion v2-1 [43][1]. We first conduct an overview of the key internal components of the diffusion model and demonstrate the extraction of multi-scale multi-timesteps features in Sec. 3.1. We then propose three fusion strategies to systematically aggregate these features, i.e., via global weighted fusion (Sec. 3.2), via localized weighted fusion(Sec. 3.3), and via a mixture-of-experts mechanism (Sec. 3.4). An overview of our method is illustrated in Fig. 2.

3.1. Feature Extraction from Diffusion Process

In an LDM [43], an input image $\mathbf{x}$ is first mapped into a latent representation $\mathbf{z} \in \mathbb{R}^{H_0 \times W_0 \times C_0}$ via an autoencoder. To extract features from the diffusion process, we start with the clean latents $\mathbf{z}$ and employ DDIM inversion [46] to trace a reverse noise path, obtaining noisy latents. We then run the denoising diffusion model on these noisy latents to capture multi-scale multi-timestep feature maps. This inversion approach yields faithful latent representations, helping preserve fine-grained details in deterministic tasks. Complete equations are provided in the supplementary material.

Backbone Architecture. The denoising backbone follows a U-Net-like architecture that generates features at $S = 4$ scales with resolutions $\left\{\frac{H_0}{2^{s-1}} \times \frac{W_0}{2^{s-1}}\right\}_{s=1}^{S}$. Each scale contains multiple *residual blocks* capturing local spatial information, and *transformer blocks* including a self-attention (SA) and a cross-attention (CA) mechanism. The SA block captures contextual dependencies within the latent itself, while the CA block encodes interactions between the latent and additional conditioning signals (e.g., text prompts). Across the diffusion process, each noise level is conditioned on a *timestep* $t \in \{1, \dots, T\}$. At each t, the U-Net refines the noisy latent toward a cleaner state. This procedure naturally produces a variety of spatiotemporal features.

For simplicity, we denote the *SA* outputs at scale s and timestep t by $\mathbf{A}_{t,s} \in \mathbb{R}^{h_s \times w_s \times d_s^a}$, the *CA* outputs by $\mathbf{C}_{t,s} \in \mathbb{R}^{h_s \times w_s \times d_s^c}$, and the *ResNet* residual outputs by $\mathbf{R}_{t,s} \in \mathbb{R}^{h_s \times w_s \times d_s^r}$, where $h_s = H_0/2^{s-1}, w_s = W_0/2^{s-1}$, and d_s^a, d_s^c, d_s^r are channel dimensions that may vary across blocks. Note that for attention blocks, we recover spatial dimensions for outputs to maintain consistency with the ResNet outputs.

These multi-scale, multi-timestep features form the building blocks for subsequent recognition tasks, as they embed both coarse- and fine-grained cues from different stages of the diffusion process. A straightforward approach to utilize these features is to attach a task-specific decoder on top of any desired subset. However, effectively navigating which blocks and timesteps to pick can be cumbersome, and a simple concatenation often yields marginal improvements (see Sec. 4.4). To address this, we propose three feature fusion strategies in the following sections that combine diverse features effectively to optimize downstream task performance.

[1] Currently, DiffusionSat is the only openly available large-scale generative geospatial foundation model with pretrained weights and accessible training/inference code.

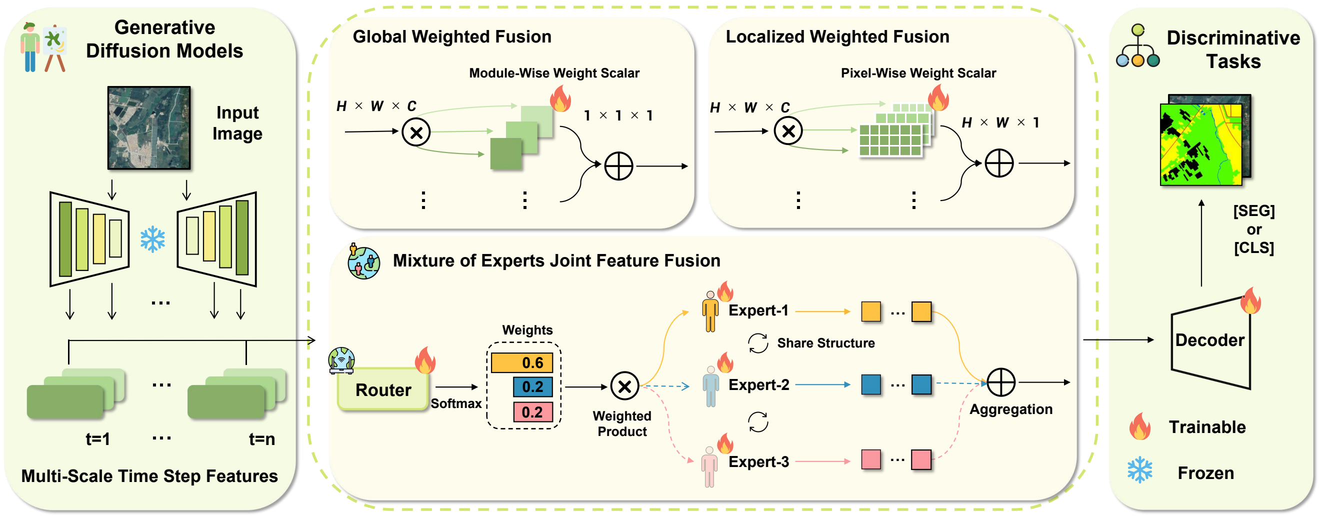


Figure 2. **Method Overview.** Our **SatDiFuser** framework leverages diffusion-based generative foundation models as self-supervised feature extractors for downstream discriminative remote sensing tasks. **(Left)**: The pre-trained geospatial diffusion model captures diverse representations at different scales and timesteps for satellite imagery. **(Right)**: Three feature fusion strategies are explored to effectively leverage these features: (1) Global Weighted Fusion applies learnable module-wise scalars for broad aggregation. (2) Localized Weighted Fusion learns pixel-wise weights for spatially varying importance. (3) Mixture of Experts (MoE) Joint Fusion uses specialized experts to model complex feature interactions. The fused features are fed into a task-specific decoder for different RS tasks.

3.2. Global Weighted Feature Fusion

Inspired by Diffusion Hyperfeatures [31], we adopt a learnable global-weight aggregation scheme across timesteps and feature blocks. However, unlike [31], which resizes all features to a uniform resolution, we maintain the original multi-scale resolutions, resulting in a feature pyramid $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_S\}$, allowing the features to capture the multi-scale nature inherent in RS images. At each scale s, the final aggregated feature is a weighted sum:

$$\mathbf{X}_s = \sum_{t=1}^{T_{sel}} \sum_{l=1}^{L_{sel}} w_{l,t} \cdot \Phi_s^l(\mathbf{F}_{t,s}^l), \tag{1}$$

where $\mathbf{F}_{t,s}^l$ denotes the l-th feature block at scale s and timestep t, and $w_{l,t}$ is a learnable weight scalar for each block-timestep pair. **A**, **C**, **R** are examples of possible feature types, and Φ is a projection network that aligns channels. T_{sel} and L_{sel} indicate the total number of selected timesteps and feature blocks.

By learning global importance weights, this method offers a simple and efficient way to integrate multi-scale, multi-timestep diffusion features, capturing a broad representation while minimizing additional computational costs.

3.3. Localized Weighted Feature Fusion

Unlike global weighting, which applies a uniform mixing factor to entire feature maps, we then investigate *pixel-level* weighting that dynamically emphasizes different features at each spatial location.

Specifically, for a given scale s, we first compute a *reference feature* by averaging the extracted feature maps $\{\mathbf{F}_{t,s}^l\}$. This reference is fed into a lightweight gating function (e.g., a small convolutional network) to generate pixel-wise weights $\{\mathbf{W}_{t,s}^l\} \in \mathbb{R}^{h_s \times w_s}$. These weights are then normalized and applied to the corresponding features:

$$\mathbf{X}_s(u,v) = \sum_{t=1}^{T_{sel}} \sum_{l=1}^{L_{sel}} \mathbf{W}_{t,s}^l(u,v) \cdot \Phi_s^l(\mathbf{F}_{t,s}^l)(u,v), \tag{2}$$

where (u,v) denotes a spatial location. Repeating the same gating process at each scale produces a pyramid of pixel-wise fused features.

By allowing a more nuanced feature aggregation, this spatially adaptive scheme can preserve local details more effectively. Its fine-grained emphasis is particularly suited to objects with intricate outlines or heterogeneous textures in RS images, offering richer spatial detail than a single global weighting factor. However, this sensitivity can also respond strongly to local variations such as illumination differences (see Sec. 4.5).

3.4. MoE Joint Feature Fusion

The previous fusion methods explicitly decouple each feature map, encouraging the model to learn patterns for "which modules within which timestep to emphasize". In contrast, to model the more complex interactions between timesteps and feature blocks, we introduce a joint modeling method using a mixture-of-experts mechanism.

Mixture of Experts (MoE) is a sparsely activated architecture that partitions the model's parameters into expert sub-networks, coordinated by a routing function that selects which experts to activate [45]. This divide-and-conquer approach allows the model to tackle complex tasks by assigning specialized experts to different data aspects. This capability is especially advantageous in remote sensing, where images display diverse patterns, ranging from fine-grained textures to large-scale contextual variations.

Building on this idea, we adapt the MoE paradigm to fuse diffusion features at each scale by jointly modeling different module outputs and multiple timesteps. Specifically, for each selected timestep $t \in \{1, \ldots, T_{sel}\}$ at scale s, we first concatenate the module-specific features $\{\mathbf{F}^{l}_{t,s}\}$ into a single vector $\mathbf{X}_{t,s}$ along the channel dimension:

$$\mathbf{X}_{t,s} = \mathrm{Concat}\big(\mathbf{F}^{1}_{t,s}, \ldots, \mathbf{F}^{L_{sel}}_{t,s}\big) \in \mathbb{R}^{B \times C_s \times H_s \times W_s}, \quad (3)$$

where C_s is the total channel dimension after concatenation. A shared MoE layer $f_{\mathrm{MoE}}(\cdot)$ then processes $\mathbf{X}_{t,s}$ via E expert sub-networks $\{f_1, \ldots, f_E\}$ and a gating function γ. Formally,

$$\mathbf{Y}_{t,s} \;=\; f_{\mathrm{MoE}}\big(\mathbf{X}_{t,s}\big) \;=\; \sum_{e=1}^{E} \gamma_e\big(\mathbf{X}_{t,s}\big)\; f_e\big(\mathbf{X}_{t,s}\big). \quad (4)$$

Each expert f_e focuses on certain patterns in the concatenated features, while $\gamma_e(\mathbf{X}_{t,s})$ indicates how strongly to activate the experts. Optionally, top-k routing [45] can reduce computational overhead by zeroing out less relevant experts. After processing each timestep, we sum the resulting outputs across all selected t to obtain $\mathbf{X}_s$.

Compared to scalar or pixel-wise weighting, this joint formulation explicitly captures the synergy among different network modules and timesteps. By leveraging specialized sub-networks to capture diverse diffusion features, it offers a robust and flexible representation that can adapt to varied patterns in RS data.

4. Experiments

4.1. Evaluation Protocol

To assess the discriminative power of generative diffusion features and validate the effectiveness of feature fusion strategies of SatDiFuser, we perform evaluations on a diverse set of classification and semantic segmentation tasks. Following standard evaluation protocols in recent GFMs [35, 63], we freeze the pretrained generative backbone, and only train SatDiFuser components with task-specific decoders: a linear head for classification tasks and a UPerNet decoder [61] for segmentation tasks.

We uniformly employ this setting to benchmark against state-of-the-art large-scale pretrained RS models, covering diverse pretraining paradigms, including MIM [37, 42, 58, 63], contrastive learning [12, 33], self-distillation [58], and supervised pretraining [5]. Detailed feature extraction settings for these models are provided in Sec. 4.2. We also include a fully supervised ConvNeXT [30] for classification and a UNet [44] for segmentation for reference. To handle discrepancies between the spectral bands available in the datasets and those required during GFMs pretraining, we follow the standard practice in [35] by matching available bands and zero-filling any missing ones for all the models. Specifically, we only match RGB bands for DiffusionSat across all tasks.

Downstream Tasks. We adopt GEO-Bench [25], which contains semantic segmentation and classification datasets, covering diverse application domains (e.g., agriculture, urban, forest, etc.) and geographic regions. Dataset-specific details, including dataset sizes and spectral properties, are summarized in the supplementary material.

Pretrained DiffusionSat. The original DiffusionSat model [23] supports text and metadata conditioning during image synthesis. In our experiments, we omit metadata and class-specific conditioning to avoid potential information leakage. For text prompts, we use a generic phrase, *"A satellite image"*, to keep the conditioning consistent across all tasks.

4.2. Implementation Details

For DiffusionSat, we select ResNet and self-attention outputs at timesteps $\{1, 100, 200\}$ from the decoder blocks of its UNet (details justified in Sec. 4.4). For comparison GFMs, we extract features from evenly spaced layers based on the specific GFM architecture, following common protocols [35, 42]. For instance, in a 12-layer ViT-based MAE model, we select features from layers indexed at (3, 5, 7, 11). All models are optimized using AdamW with an initial learning rate of 0.01, scheduled with cosine decay after a 5-epoch warm-up. Images are cropped or resized to match the pretraining resolution required by each GFM; specifically for DiffusionSat, which offers models trained on 512px and 256px resolutions, we resize images to these dimensions using bilinear interpolation based on their original resolution. Each spectral band is normalized individually using the minimum and maximum values calculated across the entire dataset. We use a training batch size of 32. Additional dataset-specific implementation details, such as loss functions and training epochs, are provided in the supplementary material.

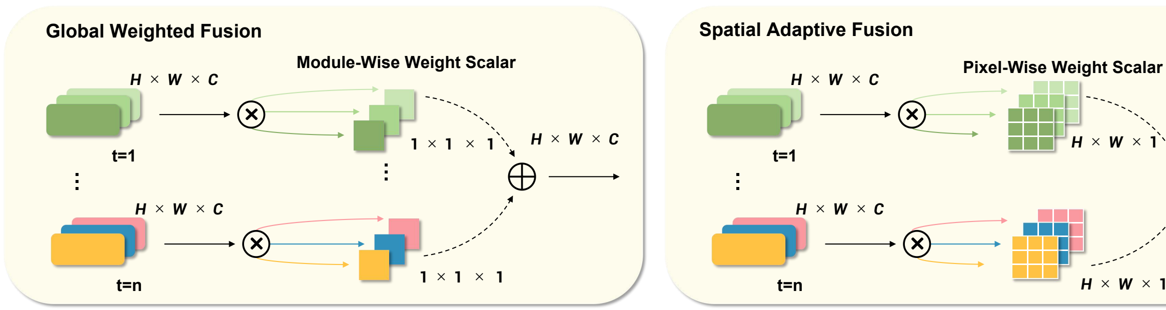
4.3. Main Results

In Tab. 1, we evaluate the semantic segmentation performance of SatDiFuser using three distinct fusion strategies, comparing it against other pretrained RS models. SatDiFuser achieves the highest mIoU on five of the six tasks,

Method	pv-s	nz-c	neon	cashew	sa-c	ches
Fully Supervised	94.7	85.1	64.2	79.9	34.4	70.4
Satlas [5]	92.3	83.1	52.0	49.1	31.6	52.2
SSL4EO-MAE [58]	89.2	78.7	53.1	57.8	28.6	52.0
ScaleMAE [42]	94.2	**84.1**	55.9	47.8	20.1	61.1
SSL4EO-DINO [58]						
GFM [36]						
RemoteCLIP [33]						
CROMA [12]						
DOFA [63]						
SatDiFuser(Ours)						
Global fusion						
Localized fusion						
MoE fusion						

Table 1. **Semantic seg**
decoder, reported as
highlighted as **first**,
other pretrained remo
row presents the perf
and the top-performin
s, nz-c, neon, cashew,
m-nz-cattle, m-NeonT
and m-chesapeake-lan

Method	ben	bk	es	fn	pv	s2s
Fully Supervised	69.4	98.9	97.7	56.8	98.0	58.1
Satlas [5]	54.3	88.7	92.1	40.8	95.2	55.3
SSL4EO-MAE [58]	46.0	92.0	82.4	40.5	90.9	47.3
ScaleMAE [42]	39.8	90.3	78.9	45.6	96.9	21.8



with particularly la
plantation, and m-c
5.9% improvement
tively). Notably, e
cashew-plan and n
superior performanc
its robustness. The
native knowledge embedded in large-scale pretrained generative diffusion models and validate the effectiveness of our approach in transferring that knowledge to downstream dense prediction tasks.

Turning to classification, Tab. 2 shows that SatDiFuser consistently outperforms competing RS models across the benchmark. The gains are especially pronounced on challenging datasets that are not saturated, with improvements of up to +7.9% on m-bigearthnet, +5.9% on m-forestnet, and +3.9% on m-so2sat. Especially, SatDiFuser even surpasses fully supervised settings on m-forestnet and m-so2sat datasets. These results underscore the versatility of diffusion-based features in RS tasks and confirm the potential of SatDiFuser as a robust foundation for both dense prediction and classification scenarios.

All three of our proposed fusion approaches present strong capabilities in leveraging pretrained diffusion features. A more detailed comparison and interpretation analysis can be found in Sec. 4.5.

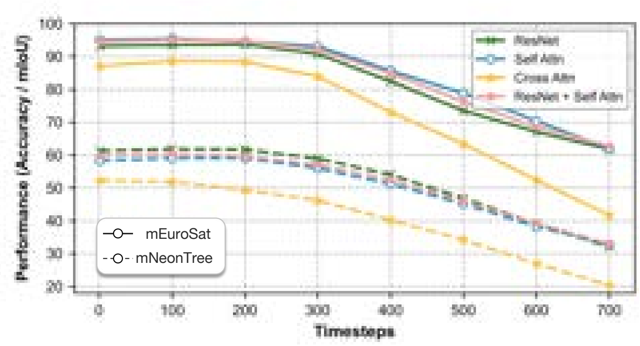


Figure 3. Performance of individual module blocks across sampling timesteps on mEuroSat and mNeonTree datasets.

4.4. Ablation Studies

We ablate key design choices to validate our approach. We first investigate the impact of individual diffusion features extracted at different diffusion stages. We then compare our fusion strategies against "raw" features, and also evaluate different pretrained backbones and data-scarcity scenarios.

Effects of Timesteps. Using one classification and one segmentation dataset as examples (Fig. 3), we observe that performance peaks when sampling within the first 20% of the diffusion timesteps. Traditionally, DDIM [46] sampling can take up to 1000 steps. At later timesteps, heavily noised latents lose too many fine-grained details, while very early steps (e.g., one-step denoising) provide limited learning signals.

Multi-T	Multi-L	Method	Classification		Segmentation	
			s2s	es	cashew	pv-s
✘	✘	ts=1, SA	53.6	94.3	55.3	92.5
✘	✘	ts=1, R	50.8	92.1	56.4	92.1
✘	✘	ts=100, SA	52.1	<u>94.7</u>	54.1	<u>93.2</u>
✘	✘	ts=100, R	50.5	92.4	57.9	92.6
✔	✔	simple concat	<u>55.4</u>	94.5	<u>59.1</u>	92.9
✔	✔	Global fusion	**59.3**	**97.7**	**66.5**	**95.1**
✔	✔	Localized fusion	**58.9**	**96.8**	**64.8**	**95.0**
✔	✔	MoE fusion	**58.8**	**97.3**	**66.1**	**95.3**

Table 3. Comparion of using raw features and different feature fusion strategies. The reported numbers are top-1 accuracy or mIoU.

	$E = 4$		$E = 8$		$E = 12$	
top-k	1	2	1	2	1	2
mIoU	60.5	69.8	60.2	**71.6**	59.6	68.3

Table 4. Ablation studies on number of experts E and top-k in MoE fusion on m-chesapeak-landcover dataset.

Effects of Module Blocks. As depicted in Fig. 3, ResNet and self-attention outputs contribute most significantly to performance, while cross-attention blocks provide minimal benefits. This aligns with our focus on vision foundation models, as cross-attention primarily encodes task-irrelevant textual information. Based on these findings, we primarily utilize ResNet and self-attention features from the initial 20% of timesteps.

Raw Features vs. Feature Fusion. A straightforward way to leverage diffusion-based representations is to feed raw features (without learnable fusion networks) directly into a task-specific decoder. As shown in Tab. 3, this naive approach already matches or surpasses the performance of other pretrained RS models, reflecting the discriminative capacity of generative diffusion. However, the optimal combination of timesteps and module blocks varies across datasets—one dataset might favor features from an early timestep's self-attention, whereas another benefits from a later timestep's ResNet activations. Simple concatenation of features occasionally improves results but suffers from high dimensionality and inconsistent gains. By contrast, our fusion strategies consistently outperform both raw features and simple concatenation, demonstrating that principled aggregation better exploits the diverse and complementary representations learned during diffusion and is crucial for maximizing downstream performance.

Number of Experts in MoE Fusion. We quantify the impact on the number of experts E and routing parameter top-k in MoE fusion method. Tab. 4 shows that fewer experts lead to insufficient learning, while more experts may introduce redundancy. Using top-k=2 consistently outperforms a single expert activation, suggesting the benefit of complementary expert representations. We select $E = 8$ and top-k=2 as our configuration to balance computational efficiency and performance.

Backbone	Binary-class			Multi-class		
	pv-s	nz-c	neon	cashew	sa-c	ches
SD v2-1	94.5	82.5	60.2	63.7	29.3	65.8
DiffusionSat	**95.1**	**83.5**	**61.8**	**66.5**	**32.6**	**69.5**

Table 5. Performance comparison of diffusion backbones using global weighted fusion strategy on segmentation tasks.

Method	pv-s		nz-c		cashew	
	100%	10%	100%	10%	100%	10%
Satlas [5]	92.3	88.6	83.1	77.8	49.1	25.7
SSL4EO-MAE [58]	89.2	84.8	78.7	68.1	57.8	26.8
ScaleMAE [42]	94.2	91.3	**84.1**	78.4	47.8	27.8
SSL4EO-DINO [58]	89.0	85.7	<u>78.9</u>	66.2	61.3	31.2
GFM [36]	93.1	90.3	82.4	74.1	53.5	22.8
RemoteCLIP [33]	93.2	90.0	80.7	73.7	51.7	27.1
CROMA [12]	92.5	88.7	83.4	75.4	<u>62.2</u>	<u>34.3</u>
DOFA [63]	<u>94.8</u>	<u>92.3</u>	82.8	<u>79.4</u>	53.9	29.5
SatDiFuser(Ours)						
Global fusion	95.1	93.5	83.5	80.0	**66.5**	**39.6**
Localized fusion	95.0	93.3	83.2	79.5	64.8	38.1
MoE fusion	**95.3**	**93.9**	83.7	**80.3**	66.1	38.5

Table 6. Semantic segmentation performance when using 100% and 10% of labeled data on the m-pv4ger-seg, m-nz-cattle and m-cashew-plantation datasets.

Comparison of Diffusion Backbones. In Tab. 5, we compare two pretrained diffusion backbones for semantic segmentation tasks using the global weighted fusion strategy. *SD v2-1* [43] is trained on a large-scale web-scraped dataset, while *DiffusionSat* [23] further finetunes it using large-scale satellite imagery. Despite not being domain-specific, SD v2-1 still achieves competitive results, likely due to its massive and diverse pretraining set. However, the specialized DiffusionSat model consistently outperforms SD v2-1, particularly on more complex multi-class tasks, highlighting the benefits of domain-focused finetuning.

Data Scarcity Scenario. We further evaluate our method under data scarcity by reducing the labeled training data to 10% while preserving the original data distribution. Results in Tab. 6 show that SatDiFuser maintains robust performance in limited-data scenarios, demonstrating its generalization capability even with fewer labeled samples.

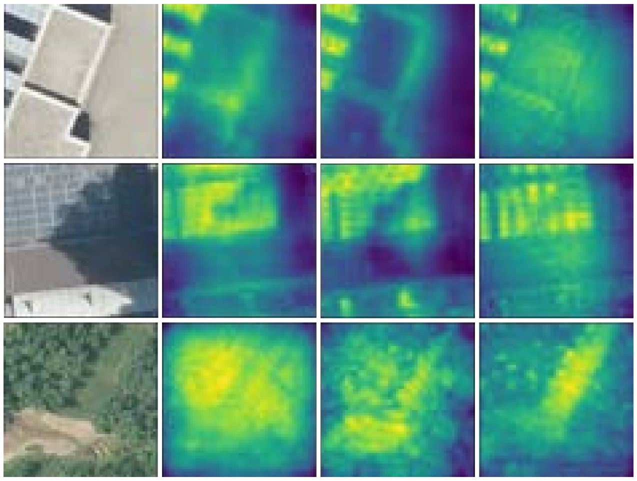
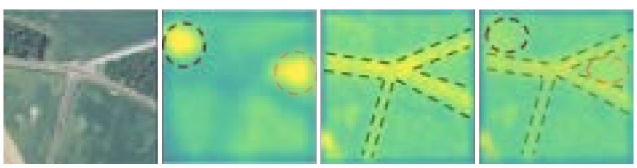


Image (a) G-W (b) L-W (c) MoE

Figure 4. Visualization of **fused feature maps** obtained by the three fusion strategies, demonstrating their distinct emphasis. G-W and L-W denote global weighted and localized weighted fusion.



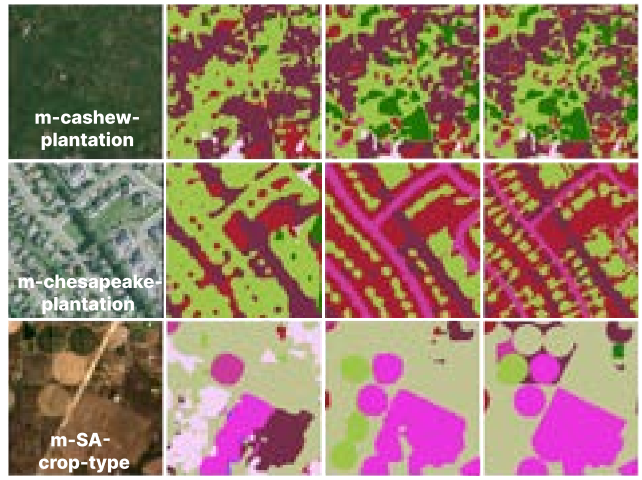
(a) Image (b) Expert-1 (c) Expert-2 (d) Fused

Figure 5. Visualization of individual **expert outputs** in the MoE fusion strategy, showing each expert specializing in distinct spatial patterns and textures.

4.5. Qualitative Analysis

Feature Visualization. Fig. 4 visualizes the fused feature maps learned by SatDiFuser's three fusion methods, highlighting their distinct characteristics. Localized weighted fusion produces detailed maps that can emphasize fine-grained object boundaries (e.g., buildings in the first row (b)), but it is also sensitive to local variations, such as shadows commonly observed in RS images (second row (b), split appearance of buildings). In contrast, Global Weighted fusion yields more stable representations across the shadowed regions, yet it may present limitations in preserving certain fine-grained details (as shown in the third row(a)). The MoE joint fusion balances these trade-offs by dynamically activating specialized expert sub-networks, preserving both global contexts and local details while mitigating sensitivity to perturbations (see the second row(c)). As further illustrated in Fig. 5, each expert learns distinct spatial or textural patterns, highlighting their abilities to adapt to the complexity of remote sensing imagery.

Prediction Results. Fig. 6 presents examples of semantic segmentation results obtained from our top-performing model and compared to the best-performing alternative GFMs on m-cashew-plantation, m-chesapeak-landcover and m-SA-crop-type. Our model demonstrates strong segmentation capability in distinguishing between different land cover types, showcasing its effectiveness in transferring robust representation for various RS downstream tasks.



(a) Image (b) Others (c) Ours (d) Label

Figure 6. **Example segmentation predictions** comparing SatDiFuser (ours) and the best-performing pretrained RS baselines per dataset. For m-cashew-plantation dataset, we colorize the clasees as follows: well-managed plantation, poorly-managed plantation, non-plantation, residential, background, and uncertain. For m-chesapeake-landcover dataset, the color maps are: tree-canopy-forest, low-vegetation-field, barren land, impervious-other, impervious-roads. For m-SA dataset, the classes are represented as: lucerne/medics, planted pastures, fallow, wine grapes, weeds, canola and rooibos.

5. Conclusion

we introduced SatDiFuser, a novel framework that systematically adapts diffusion-based generative models for self-supervised representation learning in remote sensing. By leveraging multi-stage diffusion features and novel fusion strategies, SatDiFuser achieves state-of-the-art performance on various discriminative RS tasks, demonstrating diffusion models' viability as scalable SSL alternatives.

Despite the great advantages, some limitations remain. First, our study focuses on RGB imagery due to limited pretraining on multi-band data of the utilized backbone. Extending SatDiFuser to multi-modal RS data could unlock further capabilities. Additionally, while we primarily leverage unconditional diffusion models, investigating conditioned image-to-image diffusion models—given their dense guidance mechanisms—could further enhance representations. Addressing these aspects in future work will contribute to a more comprehensive integration of diffusion models into GFMs framework, further bridging the gap between generative and discriminative paradigms in RS.

References

[1] Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021. 3

[2] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. Omnisat: Self-supervised modality fusion for earth observation, 2024. 3

[3] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10181–10190, 2021. 3

[4] Wele Gedara Chaminda Bandara, Nithin Gopalakrishnan Nair, and Vishal M Patel. Ddpm-cd: Denoising diffusion probabilistic models as feature extractors for change detection. *arXiv preprint arXiv:2206.11892*, 2022. 3

[5] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. Satlaspretrain: A large-scale dataset for remote sensing image understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16772–16782, 2023. 2, 5, 6, 7

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021. 1, 2

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607, 2020. 1

[8] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *arXiv preprint arXiv:2207.08051*, 2022. 3

[9] Paul Couairon, Mustafa Shukor, Jean-Emmanuel HAUGEARD, Matthieu Cord, and Nicolas THOME. Diffcut: Catalyzing zero-shot semantic segmentation with diffusion features and recursive normalized cut. In *Advances in Neural Information Processing Systems*, 2024. 2

[10] Runmin Dong, Shuai Yuan, Bin Luo, Mengxuan Chen, Jinxiao Zhang, Lixian Zhang, Weijia Li, Juepeng Zheng, and Haohuan Fu. Building bridges across spatial and temporal resolutions: Reference-based super-resolution via change priors and conditional diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27684–27694, 2024. 3

[11] Michael Fuest, Pingchuan Ma, Ming Gui, Johannes Schusterbauer, Vincent Tao Hu, and Bjorn Ommer. Diffusion models and representation learning: A survey. *arXiv preprint arXiv:2407.00783*, 2024. 2

[12] Anthony Fuller, Koreen Millard, and James R. Green. Croma: Remote sensing representations with contrastive radar-optical masked autoencoders, 2023. 3, 5, 6, 7

[13] Ziyang Gong, Zhixiang Wei, Di Wang, Xianzheng Ma, Hongruixuan Chen, Yuru Jia, Yupeng Deng, Zhenming Ji, Xiangwei Zhu, Naoto Yokoya, et al. Crossearth: Geospatial vision foundation model for domain generalizable remote sensing semantic segmentation. *arXiv preprint arXiv:2410.22629*, 2024. 3

[14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, pages 21271–21284, 2020. 1

[15] Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, Huimei He, Jian Wang, Jingdong Chen, Ming Yang, Yongjun Zhang, and Yansheng Li. Skysense: A multimodal remote sensing foundation model towards universal interpretation for earth observation imagery, 2024. 3

[16] Boran Han, Shuai Zhang, Xingjian Shi, and Markus Reichstein. Bridging remote sensors with multisensor geospatial foundation models, 2024. 3

[17] Jiang He, Yajie Li, Qiangqiang Yuan, et al. Tdiffde: A truncated diffusion model for remote sensing hyperspectral image denoising. *arXiv preprint arXiv:2311.13622*, 2023. 3

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 1, 2

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[20] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, Antonio Plaza, Paolo Gamba, Jon Atli Benediktsson, and Jocelyn Chanussot. Spectralgpt: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3

[21] Jia Jia, Geunho Lee, Zhibo Wang, Lyu Zhi, and Yuchu He. Siamese meets diffusion network: Smdnet for enhanced change detection in high-resolution rs imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024. 3

[22] Aliasghar Khani, Saeid Asgari Taghanaki, Aditya Sanghi, Ali Mahdavi Amiri, and Ghassan Hamarneh. Slime: Segment like me. *arXiv preprint arXiv:2309.03179*, 2023. 2

[23] Samar Khanna, Patrick Liu, Linqi Zhou, Chenlin Meng, Robin Rombach, Marshall Burke, David B. Lobell, and Stefano Ermon. Diffusionsat: A generative foundation model for satellite imagery. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2, 3, 5, 7

[24] Benedikt Kolbeinsson and Krystian Mikolajczyk. Multiclass segmentation from aerial views using recursive noise diffusion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8439–8449, 2024. 3

[25] Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, et al. Geo-

bench: Toward foundation models for earth monitoring. *Advances in Neural Information Processing Systems*, 2024. 5

[26] Georges Le Bellier and Nicolas Audebert. Detecting out-of-distribution earth observation images with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 481–491, 2024. 3

[27] Daixun Li, Weiying Xie, Jiaqing Zhang, and Yunsong Li. Mdfl: Multi-domain diffusion-driven feature learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8653–8660, 2024. 3

[28] Chenyang Liu, Keyan Chen, Rui Zhao, Zhengxia Zou, and Zhenwei Shi. Text2earth: Unlocking text-driven remote sensing image generation with a global-scale dataset and a foundation model. *arXiv preprint arXiv:2501.00895*, 2025. 2

[29] Yidan Liu, Jun Yue, Shaobo Xia, Pedram Ghamisi, Weiying Xie, and Leyuan Fang. Diffusion models meet remote sensing: Principles, methods, and perspectives. *arXiv preprint arXiv:2404.08926*, 2024. 2, 3

[30] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5

[31] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In *Advances in Neural Information Processing Systems*, 2023. 2, 4

[32] Gengchen Mai, Chris Cundy, Kristy Choi, Yingjie Hu, Ni Lao, and Stefano Ermon. Towards a foundation model for geospatial artificial intelligence (vision paper). *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, 2022. 1

[33] Utkarsh Mall, Cheng Perng Phoo, Meilin Kelsey Liu, Carl Vondrick, Bharath Hariharan, and Kavita Bala. Remote sensing vision-language foundation models without annotations via ground remote alignment. In *The Twelfth International Conference on Learning Representations*, 2024. 5, 6, 7

[34] Valerio Marsocci and Nicolas Audebert. Cross-sensor self-supervised training and alignment for remote sensing, 2024. 3

[35] Valerio Marsocci, Yuru Jia, Georges Le Bellier, David Kerekes, Liang Zeng, Sebastian Hafner, Sebastian Gerard, Eric Brune, Ritu Yadav, Ali Shibli, Heng Fang, Yifang Ban, Maarten Vergauwen, Nicolas Audebert, and Andrea Nascetti. Pangaea: A global and inclusive benchmark for geospatial foundation models, 2024. 1, 5

[36] Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen. Towards geospatial foundation models via continual pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 6, 7

[37] Matias Mendieta, Boran Han, Xingjian Shi, Yi Zhu, Chen Chen, and Mu Li. Gfm: Building geospatial foundation models via continual pretraining. *arXiv preprint arXiv:2302.04476*, 2023. 3, 5

[38] Soumik Mukhopadhyay, Matthew Gwilliam, Yosuke Yamaguchi, Vatsal Agarwal, Namitha Padmanabhan, Archana Swaminathan, Tianyi Zhou, Jun Ohya, and Abhinav Shrivastava. Do text-free diffusion models learn discriminative visual representations? In *European Conference on Computer Vision*, 2024. 2

[39] Li Pang, Xiangyu Rui, Long Cui, Hongzhong Wang, Deyu Meng, and Xiangyong Cao. Hir-diff: Unsupervised hyperspectral image restoration via improved diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3005–3014, 2024. 3

[40] Jiahui Qu, Yuanbo Yang, Wenqian Dong, and Yufei Yang. Lds2ae: Local diffusion shared-specific autoencoder for multimodal remote sensing image classification with arbitrary missing modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14731–14739, 2024. 3

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763, 2021. 1

[42] 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, 2023. 3, 5, 6, 7

[43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 7

[44] 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. 5

[45] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017. 5

[46] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 2, 3, 6

[47] Adam Stewart, Nils Lehmann, Isaac Corley, Yi Wang, Yi-Chia Chang, Nassim Ait Ait Ali Braham, Shradha Sehgal, Caleb Robinson, and Arindam Banerjee. Ssl4eo-l: Datasets and foundation models for landsat imagery. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[48] Datao Tang, Xiangyong Cao, Xingsong Hou, Zhongyuan Jiang, Junmin Liu, and Deyu Meng. Crs-diff: Controllable remote sensing image generation with diffusion model. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 3

[49] Maofeng Tang, Andrei Cozma, Konstantinos Georgiou, and Hairong Qi. Cross-scale mae: A tale of multiscale exploitation in remote sensing. *Advances in Neural Information Processing Systems*, 2024. 2

[50] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse attend and segment: Unsupervised zero-shot segmentation using stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3554–3563, 2024. 2

[51] Jiayuan Tian, Jie Lei, Jiaqing Zhang, Weiying Xie, and Yunsong Li. Swimdiff: Scene-wide matching contrastive learning with diffusion constraint for remote sensing image. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 3

[52] Jamie Tolan, Hung-I Yang, Benjamin Nosarzewski, Guillaume Couairon, Huy V Vo, John Brandt, Justine Spore, Sayantan Majumdar, Daniel Haziza, Janaki Vamaraju, et al. Very high resolution canopy height maps from rgb imagery using self-supervised vision transformer and convolutional decoder trained on aerial lidar. *Remote Sensing of Environment*, 2024. 3

[53] Devis Tuia, Konrad Schindler, Begüm Demir, Xiao Xiang Zhu, Mrinalini Kochupillai, Sašo Džeroski, Jan N. van Rijn, Holger H. Hoos, Fabio Del Frate, Mihai Datcu, Volker Markl, Bertrand Le Saux, Rochelle Schneider, and Gustau Camps-Valls. Artificial intelligence to advance earth observation: A review of models, recent trends, and pathways forward. *IEEE Geoscience and Remote Sensing Magazine*, pages 2–25, 2024. 1

[54] Ce Wang and Wanjie Sun. Semantic guided large scale factor remote sensing image super-resolution with generative diffusion prior. *ISPRS Journal of Photogrammetry and Remote Sensing*, 220:125–138, 2025. 3

[55] Di Wang, Jing Zhang, Minqiang Xu, Lin Liu, Dongsheng Wang, Erzhong Gao, Chengxi Han, Haonan Guo, Bo Du, Dacheng Tao, et al. Mtp: Advancing remote sensing foundation model via multi-task pretraining. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024. 3

[56] Meilin Wang, Yexing Song, Pengxu Wei, Xiaoyu Xian, Yukai Shi, and Liang Lin. Idf-cr: Iterative diffusion process for divide-and-conquer cloud removal in remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 3

[57] Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Lichao Mou, and Xiao Xiang Zhu. Self-supervised learning in remote sensing: A review. *arXiv preprint arXiv:2206.13188*, 2022. 1

[58] Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M. Albrecht, and Xiao Xiang Zhu. Ssl4eo-s12: A large-scale multimodal, multitemporal dataset for self-supervised learning in earth observation [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 2023. 2, 5, 6, 7

[59] Yihan Wen, Xianping Ma, Xiaokang Zhang, and Man-On Pun. Gcd-ddpm: A generative change detection model based on difference-feature guided ddpm. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 3

[60] Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. Datasetdm: Synthesizing data with perception annotations using diffusion models. *Advances in Neural Information Processing Systems*, 2023. 2

[61] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 5

[62] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9653–9663, 2022. 1

[63] Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural plasticity-inspired foundation model for observing the Earth crossing modalities. *arXiv preprint arXiv:2403.15356*, 2024. 3, 5, 6, 7

[64] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[65] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think, 2024. 2

[66] Zhiping Yu, Chenyang Liu, Liqin Liu, Zhenwei Shi, and Zhengxia Zou. Metaearth: A generative foundation model for global-scale remote sensing image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):1764–1781, 2025. 2, 3

[67] Manyuan Zhang, Guanglu Song, Xiaoyu Shi, Yu Liu, and Hongsheng Li. Three things we need to know about transferring stable diffusion to visual dense prediction tasks. In *European Conference on Computer Vision*, pages 128–145. Springer, 2025. 2

[68] Xiangrong Zhang, Shunli Tian, Guanchun Wang, Huiyu Zhou, and Licheng Jiao. Diffucd: Unsupervised hyperspectral image change detection with semantic correlation diffusion model. *arXiv preprint arXiv:2305.12410*, 2023. 3

[69] Zhuo Zheng, Stefano Ermon, Dongjun Kim, Liangpei Zhang, and Yanfei Zhong. Changen2: Multi-temporal remote sensing generative change foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3

[70] Jingyi Zhou, Jiamu Sheng, Peng Ye, Jiayuan Fan, Tong He, Bin Wang, and Tao Chen. Exploring multi-timestep multi-stage diffusion features for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 3

[71] Xuechao Zou, Kai Li, Junliang Xing, Yu Zhang, Shiying Wang, Lei Jin, and Pin Tao. Diffcr: A fast conditional diffusion framework for cloud removal from optical satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. 3