

# REJEPa: A Novel Joint-Embedding Predictive Architecture for Efficient Remote Sensing Image Retrieval

Shabnam Choudhury Yash Salunkhe\* Sarthak Mehrotra\* Biplab Banerjee  
Indian Institute of Technology Bombay

{choudhury.shabnam6, yashsalunkhe619}@gmail.com,  
{sarthak2002.mehrotra, getbiplab}@gmail.com

## Abstract

The rapid expansion of remote sensing image archives demands the development of strong and efficient techniques for content-based image retrieval (RS-CBIR). This paper presents REJEPa (Retrieval with Joint-Embedding Predictive Architecture), an innovative self-supervised framework designed for unimodal RS-CBIR. REJEPa utilises spatially distributed context token encoding to forecast abstract representations of target tokens, effectively capturing high-level semantic features and eliminating unnecessary pixel-level details. In contrast to generative methods that focus on pixel reconstruction or contrastive techniques that depend on negative pairs, REJEPa functions within feature space, achieving a reduction in computational complexity of 40–60% when compared to pixel-reconstruction baselines like Masked Autoencoders (MAE). To guarantee strong and varied representations, REJEPa incorporates Variance-Invariance-Covariance Regularisation (VICReg), which prevents encoder collapse by promoting feature diversity and reducing redundancy. The method demonstrates an estimated enhancement in retrieval accuracy of 5.1% on BEN-14K (S1), 7.4% on BEN-14K (S2), 6.0% on FMoW-RGB, and 10.1% on FMoW-Sentinel compared to prominent SSL techniques, including CSMAE-SESD, Mask-VLM, SatMAE, ScaleMAE, and SatMAE++, on extensive RS benchmarks BEN-14K (multispectral and SAR data), FMoW-RGB, and FMoW-Sentinel. Through effective generalization across sensor modalities, REJEPa establishes itself as a sensor-agnostic benchmark for efficient, scalable, and precise RS-CBIR, addressing challenges like varying resolutions, high object density, and complex backgrounds with computational efficiency.

## 1. Introduction

Content-based image retrieval (CBIR) plays a crucial role in remote sensing (RS), facilitating efficient retrieval of semantically relevant images from large and continuously ex-

\*Equal contribution

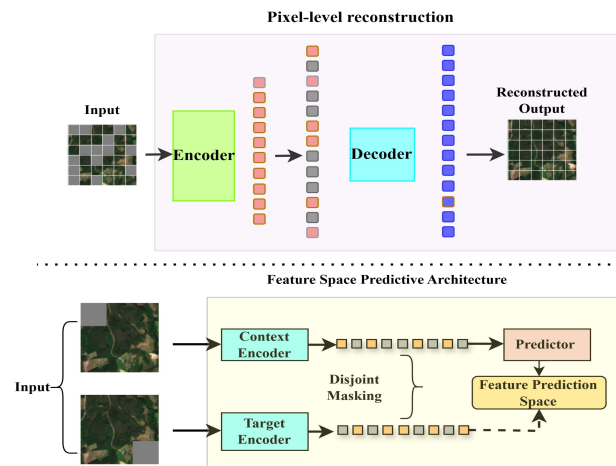


Figure 1. Conceptual illustration of REJEPa for RS-CBIR. The upper section represents traditional pixel-level reconstruction. The lower section highlights REJEPa feature-space prediction, The predicted features are then used for retrieval via k-NN

panding archives. The success of CBIR hinges on learning robust image representations that capture high-level semantics. Deep learning (DL)-based methods have become the cornerstone of CBIR, with self-supervised learning (SSL) emerging as a promising approach due to its ability to learn meaningful representations without the need for costly manual annotations [32, 40]. Among SSL methods, contrastive learning [11, 33] has shown exceptional promise by capturing semantic features from unannotated datasets. However, its reliance on negative pairs introduces challenges in RS, where different images often belong to the same class, potentially limiting its effectiveness.

Generative SSL methods, such as masked image modeling (MIM) [18, 38], reconstruct missing pixels to learn feature representations. In RS, MIM-based models like masked autoencoders (MAEs) [13, 18] leverage vision transformers (ViTs) [14] for feature extraction. However, these models emphasize pixel reconstruction, limiting their ability to capture high-level semantics [1, 39], especially in

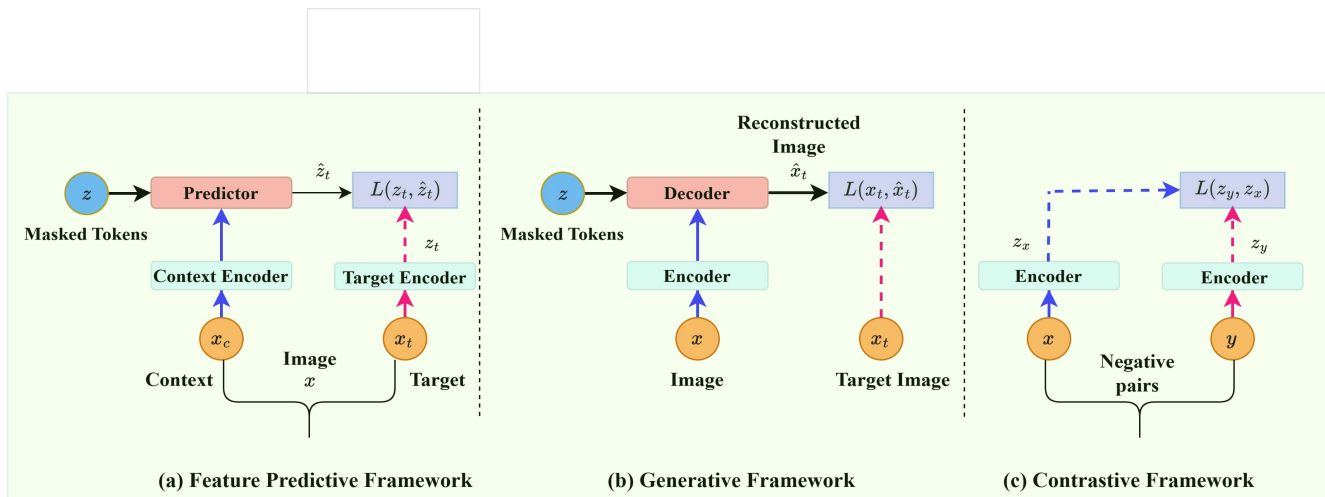


Figure 2. Illustration of self-supervised learning frameworks [23] based on their learning objectives. (a) Feature Predictive Framework (e.g., REJEPa) learns by predicting the feature representation  $\hat{z}_t$  of a target image  $x_t$  from the context representation  $z_c$ , using a predictor conditioned on a transformation variable  $z$ . (b) Generative Framework reconstructs the pixel-level output  $\hat{x}_t$  from a masked input  $x_t$ , employing a decoder for full-image reconstruction. (c) Contrastive Framework learns to align embeddings  $z_x$  and  $z_y$  of semantically similar images while pushing apart representations of negative samples, relying on augmentation-based training.

RS, where diverse resolutions, dense objects, and environmental noise pose additional challenges. To overcome these limitations, we propose REJEPa (Retrieval with Joint Embedding Predictive Architecture), a novel RS-CBIR framework that replaces pixel reconstruction with feature-space prediction. Built on joint-embedding predictive architectures [2], REJEPa predicts high-level feature representations of target regions using spatially distributed context tokens, capturing semantic structures while discarding pixel-level redundancies. This makes it well-suited for RS tasks involving multi-resolution data, complex object distributions, and noisy imaging conditions. REJEPa has broad applicability across various RS tasks, including land cover and land use mapping, disaster response and damage assessment, precision agriculture, surveillance, and climate change monitoring. For instance, in disaster response, REJEPa can quickly retrieve relevant pre- and post-disaster images, aiding in damage assessment and recovery efforts. Similarly, precision agriculture can retrieve images with similar crop health or soil conditions, supporting crop monitoring and yield prediction. Its ability to generalize across diverse sensor modalities (RGB, multispectral, SAR) makes it well-suited for cross-sensor data fusion, addressing complex challenges in remote sensing.

To further enhance REJEPa’s performance and address the risk of encoder collapse, where embeddings become constant or non-informative, we incorporate Variance-Invariance-Covariance Regularization (VICReg) [7]. VICReg employs two regularization terms: (1) maintaining the variance of each embedding dimension above a threshold, ensuring diversity in the learned features, and (2) decorrelating variables to minimize redundancy in the

embeddings. These enhancements are crucial for capturing the complex and diverse characteristics of RS imagery. By combining joint-embedding predictive design with VICReg, REJEPa produces semantically rich and robust representations, achieving faster convergence and superior performance in CBIR tasks. In this work, we propose the following key contributions:

- We introduce REJEPa, the first joint-embedding predictive framework for RS-CBIR, which learns high-level semantic representations without pixel-level reconstruction or reliance on negative pairs, reducing computational complexity by 40–60%.
- We incorporate Variance-Invariance-Covariance Regularization (VICReg) to prevent encoder collapse, ensuring robust, diverse, and non-redundant representations tailored for RS-CBIR.
- We demonstrate that REJEPa generalizes effectively across different sensor modalities (RGB, multispectral, and SAR), establishing itself as a sensor-agnostic benchmark for efficient and scalable RS-CBIR.

## 2. Related Work

### 2.1. Self-Supervised Learning for Remote Sensing

Self-supervised learning (SSL) has revolutionized remote sensing (RS) by alleviating the dependency on large-scale annotated datasets [37]. Existing SSL methods in RS predominantly fall into contrastive and generative paradigms, as illustrated in Figure 2. The research landscape of SSL-based foundation models for RS closely mirrors this dichotomy [4, 8, 35].

Contrastive methods, depicted in Figure 2(c) often referred to as Siamese structures, augmentation-based

techniques, or joint-embedding predictive architectures (JEPAs), have demonstrated significant promise in RS applications. For instance, studies such as [19, 41] utilized spatial neighbors as augmented data to create positive pairs, while [24] incorporated random rotations (90°, 180°, and 270°) to enhance data variability. Other efforts, such as [22], exploited geographical vegetation as augmentation cues. Additionally, some approaches predict missing modalities from available data to improve multi-modal representation learning [3, 16]. Despite these successes, contrastive methods face challenges in RS due to their reliance on negative pairs, which may inadvertently include semantically similar samples, thereby limiting their effectiveness.

Generative SSL methods, shown in Figure 2(b) particularly masked image modeling (MIM), have gained widespread recognition for their ability to learn semantic representations by reconstructing masked regions of an image. MIM-based models, such as masked autoencoders (MAEs) [18], have been extensively adapted for RS tasks. These adaptations incorporate domain-specific properties, such as scale invariance [30], temporal information [13], and temporal invariance [26]. Advanced generative models, including RingMo [34], billion-scale MAE [10], and VI-TAE [36], have pushed the boundaries of MIM by scaling the framework to larger datasets and architectures. Generative models in RS leverage unique properties of Earth observation data through spectral [13], temporal [15], and spatiotemporal [27] masking strategies.

## 2.2. Feature Predictive Architectures

Figure 2(a) illustrates predictive architectures as a promising paradigm in self-supervised learning (SSL), shifting from pixel-level reconstruction to high-level semantic feature prediction. Unlike contrastive or generative methods, which rely heavily on negative pairs or pixel-level reconstruction, feature predictive approaches aim to capture meaningful semantic relationships by leveraging contextual information from the data itself. At the core of this paradigm lies the joint-embedding predictive architecture (JEPA) [21], which employs a Siamese encoder-predictor design to infer missing information in feature space.

JEPAs have been successfully implemented across various modalities, including audio [5], image [2, 29, 42], and text [6]. These approaches have demonstrated that predicting in representation space leads to versatile representations that perform well in downstream tasks such as linear probing and low-shot adaptation [1, 2, 29], while offering significant efficiency gains during pretraining compared to pixel-level reconstruction [2, 5]. Additionally, feature-predictive methods exhibit competitive performance in end-to-end fine-tuning for image, audio, and text domains [5, 6]. Feature-predictive models can also incorporate contrastive objectives for improved stability and representation quality [5].

By bypassing the need for complex data augmentations or decoder networks, JEPAs are particularly well-suited for multi-modal applications, such as Earth observation, where data diversity and scale present unique challenges. For example, SAR-JEPA [23] introduced the application of JEPA concepts, focusing specifically on SAR data. JEPA architectures handle diverse sensor modalities (e.g., RGB, multi-spectral, SAR) without requiring modality-specific designs, adapt to varying resolutions and scales, and are resilient to noise and complex backgrounds. By avoiding pixel-level reconstruction, JEPA ensures computational efficiency and scalability, making it ideal for tasks like RS-CBIR, where robust and versatile representations are critical.

We propose REJEPA, a novel JEPA-based framework designed specifically for RS-CBIR, which predicts in feature space. REJEPA learns robust, high-level semantic representations while addressing the unique challenges of RS imagery, including varying sensor types, resolutions, and spatial complexities.

## 3. Method : REJEPA

### 3.1. Problem Definition

Let  $\mathcal{X} = \{x_i\}_{i=1}^N$  be a remote sensing (RS) image archive of  $N$  images. The goal of remote sensing content-based image retrieval (RS-CBIR) is to retrieve images from  $\mathcal{X}$  that are semantically similar to a query image  $x_q \in \mathcal{X}$ . To achieve this, we propose REJEPA, a joint-embedding predictive architecture (JEPA) based framework [2, 21], as shown in Figure 1. REJEPA learns representations by predicting the feature representation of a target image  $x_t$  from a context image  $x_c$ , conditioned on a transformation variable  $z$ :

$$\hat{z}_t = P_\phi(z_c, z_t), \quad (1)$$

where

$$z_c = E_\theta(x_c), \quad z_t = E_{\theta'}(x_t), \quad (2)$$

and  $z$  encodes the spatial and spectral relationships relationship between  $x_c$  and  $x_t$ . For retrieval, the representation of a query image  $x_q$  is computed as:

$$z_q = E_\theta(x_q), \quad (3)$$

and the nearest neighbors  $z_q$  are retrieved in feature space.

### 3.2. REJEPA Architecture and Predictive Learning

We introduce REJEPA, illustrated in Figure 3, a novel framework for remote sensing content-based image retrieval (RS-CBIR) built on the principles of the feature predictive architectures. The core idea of REJEPA is to learn semantically meaningful representations by predicting the feature representation of a target image from a context image, conditioned on a transformation variable  $z$ . This approach shifts the focus from pixel-level reconstruction (as in generative models) or contrastive learning (which relies

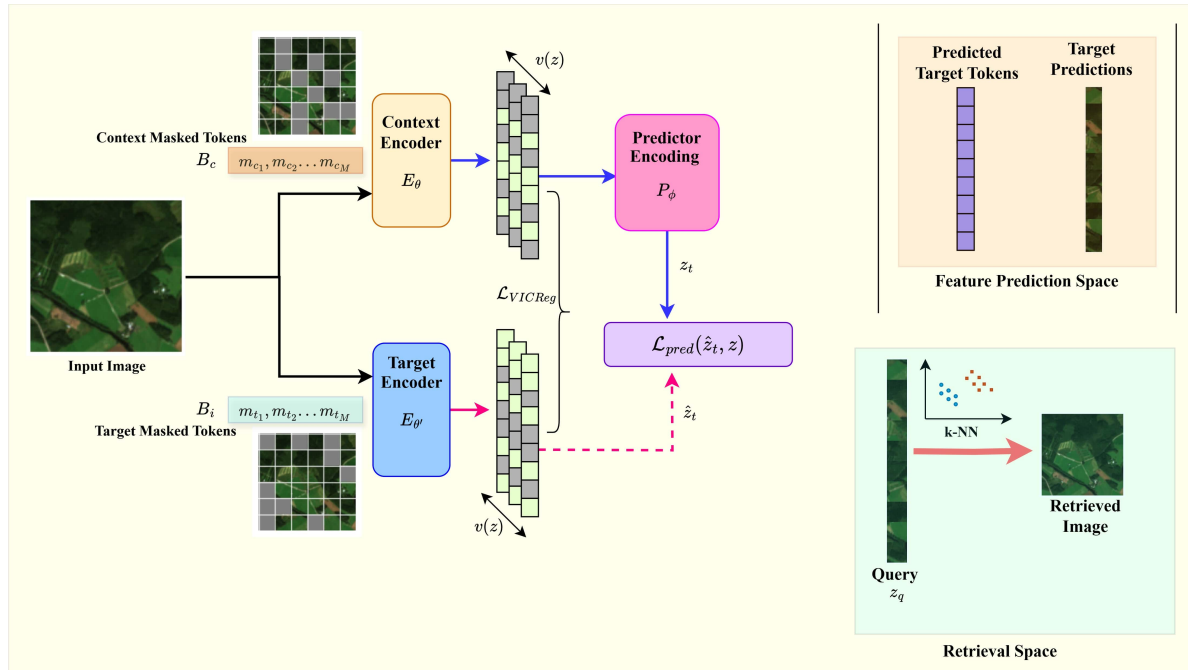


Figure 3. Architecture of REJEPa, employing a disjoint masking strategy, where context masked tokens ( $B_c$ ) and **target masked tokens** ( $B_i$ ) are independently sampled to ensure non-overlapping feature learning. The context encoder ( $E_\theta$ ) processes unmasked regions to extract spatial representations, while the target encoder ( $E_{\theta'}$ ) encodes the masked target patches, generating reference embeddings. The predictor ( $P_\phi$ ) estimates the missing target representations ( $\hat{z}_t$ ) from the context encoding, conditioned on positional tokens. VICReg regularization enforces feature variance ( $v(z)$ ), ensuring diverse and stable representations. The retrieval pipeline employs **k-NN search** on the learned feature space, retrieving semantically similar images from the archive.

on negative pairs) to feature-space prediction, enabling the model to capture high-level semantic information while discarding irrelevant pixel-level details. The overall architecture consists of three key components:

- **Context and Target Encoders** ( $E_\theta, E_{\theta'}$ ): Two encoders process the context and target images separately to generate meaningful feature representations. The *context encoder* extracts spatially distributed features while disregarding masked regions, whereas the *target encoder* reconstructs the missing regions by encoding target patches into an abstract feature space.
- **Predictor Network with Learned Transformations** ( $P_\phi$ ): Learns to predict the representation of masked target embeddings from the context embeddings while encoding transformation cues. By leveraging the context representation and an implicit transformation variable, the predictor estimates the feature distribution of the unseen target, allowing REJEPa to model high-level semantics beyond pixel-wise reconstruction.
- **Feature Regularization via VICReg**: Ensures robust and diverse representations by enforcing variance preservation, decorrelation, and invariance constraints, mitigating encoder collapse while improving feature discrimination in RS-CBIR.

### 3.3. Context and Target Encoders

**Context Encoder** ( $E_\theta$ ): A disjoint random masking strategy is employed, where the context mask ( $B_c$ ) and target mask ( $B_i$ ) are sampled independently, ensuring that no overlapping regions exist between the context and target tokens. This prevents information leakage and enforces a stronger predictive learning objective. The context tokens, masked by  $B_c$ , retain only spatially distributed visible regions and are then processed through  $E_\theta$  to generate their feature representation:

$$z_c = E_\theta(x_c) = \{z_{c1}, z_{c2}, \dots, z_{cM}\} \quad (4)$$

where  $z_c$  captures non-overlapping contextual features necessary for downstream prediction.

**Target Encoder** ( $E_{\theta'}$ ): The target encoder  $E_{\theta'}$  processes the masked regions (target patches) independently from the context encoder to generate feature representations that the model aims to predict. A disjoint target mask  $B_i$  is applied, ensuring that target tokens do not overlap with the context tokens, enforcing a stricter predictive setup. The target-encoded representation is then obtained as follows:

$$z_t = E_{\theta'}(x_t) = \{z_{t1}, z_{t2}, \dots, z_{tN}\} \quad (5)$$

### 3.4. Predictor Network

The prediction module ( $P_\phi$ ) in REJEPa is responsible for estimating the feature representations of masked target regions using only the available contextual features. Given the output of the context encoder  $z_c$ ,  $P_\phi$  learns to infer the representations of  $M$  masked target tokens, ensuring that the model captures high-level semantic relationships rather than low-level pixel dependencies. For a given target tokens representation  $z_t^{(i)}$ , associated with a target mask  $B_i$ , the predictor network  $P_\phi(\cdot, \cdot)$  takes as input:

- The context representation  $z_c$ , extracted from unmasked image regions.
- A set of mask tokens  $\{m_j\}_{j \in B_i}$  corresponding to the masked target patches.

The predictor then estimates the target encoder token representation as:

$$\hat{z}_t^{(i)} = P_\phi(z_c, \{m_j\}_{j \in B_i}) \quad (6)$$

where  $\hat{z}_t^{(i)}$  represents the predicted embeddings for the  $i$ -th target token.

**Mask Token Parameterization:** The prediction process relies on a set of learnable mask tokens. The mask tokens are parameterized by a shared learnable vector, ensuring that the model generalizes well across different masking patterns. Each mask token is enriched with a positional embedding, allowing the predictor to retain spatial information and ensure sensor-invariant feature learning. Since we predict  $M$  target tokens, we apply the predictor network  $M$  times, each time conditioning on the corresponding mask tokens for the target locations. This results in a set of predictions:

$$\{\hat{z}_t^{(1)}, \hat{z}_t^{(2)}, \dots, \hat{z}_t^{(M)}\} \quad (7)$$

which are compared with the actual target representations  $\{z_t^{(1)}, z_t^{(2)}, \dots, z_t^{(M)}\}$  during training.

### 3.5. Feature Regularization via VICReg

A major challenge in self-supervised learning is representation collapse, where the encoder outputs degenerate feature representations, mapping all inputs to a constant or non-informative vector. This issue is particularly critical in joint-embedding architectures, such as REJEPa, where contrastive negative pairs are absent. To mitigate this, we integrate Variance-Invariance-Covariance Regularization (VICReg) [7], which enforces stability and diversity in learned feature representations,  $z \in \mathbb{R}^d$ , where  $d$  represents the dimensionality of the feature embedding space. Given a batch of  $n$  feature representations, VICReg is formulated by applying three key constraints.

- **Variance Regularization:** Ensures feature diversity by preventing collapse into a single point, maintaining a min-

imum variance threshold across feature dimensions:

$$v(z) = \frac{1}{d} \sum_{j=1}^d \max(0, \gamma - \sqrt{\text{Var}(z_j) + \epsilon}), \quad (8)$$

where  $\gamma$  controls the variance threshold, and  $\epsilon$  stabilizes training.

- **Covariance Regularization:** Reduces redundancy by decorrelating feature dimensions, ensuring each captures distinct semantic information:

$$c(z) = \frac{1}{d} \sum_{i \neq j} \left[ \frac{1}{n-1} \sum_{i=1}^n (z_i - \bar{z})(z_i - \bar{z})^T \right]_{i,j}^2, \quad (9)$$

where  $\bar{z}$  represents the mean embedding.

- **Invariance Regularization:** Encourages consistency between embeddings from different transformations of the same image, ensuring robustness:

$$\mathcal{L}_{\text{inv}} = \frac{1}{n} \sum_{i=1}^n \|z_i - z'_i\|^2, \quad (10)$$

where  $z_i$  and  $z'_i$  are embeddings from different augmented views of the same image.

**VICReg Integration in REJEPa:** VICReg regularizes both context and target representations, preventing trivial solutions while enhancing retrieval distinctiveness through feature decorrelation. By enforcing statistical constraints, it ensures diverse, independent embeddings, enabling generalization across sensor modalities (RGB, SAR, multispectral). This regularization stabilizes training, prevents degenerate representations, and improves convergence and retrieval efficiency in large-scale RS-CBIR.

### 3.6. Training Objective

The training objective of REJEPa is designed to ensure both accurate predictive learning and stable feature representations by combining a prediction loss ( $\mathcal{L}_{\text{pred}}$ ) that aligns context and target representations with a regularization loss ( $\mathcal{L}_{\text{VICReg}}$ ) that prevents feature collapse and redundancy. The prediction loss minimizes the discrepancy between the predicted target representation  $\hat{z}_t^{(i)}$  and the actual target representation  $z_t^{(i)}$ , using an  $L_2$  loss formulation:

$$\mathcal{L}_{\text{pred}} = \frac{1}{M} \sum_{i=1}^M \|\hat{z}_t^{(i)} - z_t^{(i)}\|_2^2, \quad (11)$$

where  $M$  is the number of masked target tokens. To maintain feature diversity and avoid degenerate solutions, REJEPa incorporates a feature regularization loss that consists of three key terms: variance regularization  $v(z)$ , covariance regularization  $c(z)$ , and an invariance term  $\mathcal{L}_{\text{inv}}$ . These regularization terms are combined as follows:

Table 1. F1-score (%) comparison of self-supervised models for RS-CBIR on BEN-14K, FMoW-RGB, and FMoW-Sentinel datasets. REJEPa achieves the best retrieval performance with reduced computational complexity

Model Name	#Params (M)	BEN-14K		FMoW-RGB	FMoW-Sentinel
		S1→S1	S2→S2		
MAE [18]	224.87	60.81	72.04	58.73	61.77
MAE-RVSA [36]	227.75	55.40	71.47	55.26	60.28
SatMAE [13]	329.40	70.86	<b>78.71</b>	61.85	56.63
SatMAE++ [28]	329.14	67.29	<b>76.48</b>	60.09	57.75
SS-CMIR [33]	259.07	68.07	70.54	66.71	63.46
Scale-MAE [30]	284.35	62.73	NA	64.26	69.56
Mask-VLM [20]	225.82	68.10	71.02	61.52	65.23
CSMAE-SESD (Disjoint) [17]	210.64	70.62	39.01	68.42	57.13
<b>REJEPa</b>	<b>197.09</b>	<b>76.38</b>	75.42	<b>73.53</b>	<b>75.87</b>

$$\mathcal{L}_{\text{VICReg}} = \lambda_v v(z) + \lambda_c c(z) + \lambda_i \mathcal{L}_{\text{inv}}, \quad (12)$$

where  $\lambda_v, \lambda_c, \lambda_i$  are hyperparameters controlling the relative influence of each constraint. The final training objective of REJEPa is a weighted sum of the prediction and regularization losses:

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{VICReg}}, \quad (13)$$

where  $\alpha$  determines the balance between predictive accuracy and feature regularization. Once trained, the model generates latent feature representations for all images in the archive. Given a query image  $x_q$ , its corresponding representation is obtained via the context encoder,  $E_\theta(x_q)$ . To retrieve the most relevant images, a  $k$ -nearest neighbors ( $k$ -NN) search is performed in the learned feature space.

## 4. Experiments

### 4.1. Datasets

To evaluate the performance of REJEPa, we conduct experiments on three large-scale publicly available remote sensing datasets spanning multiple sensor modalities, including SAR, multispectral, and RGB satellite imagery.

**BigEarthNet-14K (BEN-14K)** This dataset comprises 14,832 Sentinel-1 SAR and Sentinel-2 multispectral image pairs, annotated with 19 land cover categories. The SAR images include VV and VH polarization bands, while multispectral bands at 10m and 20m resolution are standardized via bicubic interpolation.

**fMoW-RGB:** The fMoW-RGB dataset [12] consists of high-resolution satellite images categorized into 62 different classes, designed primarily for classification tasks. It contains approximately 363,000 training images and 53,000 test images.

**fMoW-Sentinel:** Derived from fMoW-RGB, fMoW-Sentinel [13] extends the dataset by incorporating Sentinel-2 multispectral imagery. It maintains the same 62-class tax-

onomy while offering a significantly larger collection of images, comprising 712,874 training samples, 84,939 validation samples, and 84,966 test images.

### 4.2. Implementation and Evaluation Details

REJEPa employs ViT-B/16 [14] architectures for the context encoder, target encoder, and predictor. The context and target encoders follow the standard ViT design, with the context encoder processing unmasked regions and the target encoder focusing on masked regions to generate target representations. The predictor is a lightweight ViT with 384-dimensional embedding and a depth of 12 layers, matching the number of attention heads in the context encoder. During evaluation, the target encoder’s output is average-pooled to produce a global image representation for retrieval tasks. We optimize REJEPa using AdamW [25] with an initial learning rate of  $10^{-4}$ , linearly increased to  $10^{-3}$  over the first 15 epochs and decayed to  $10^{-6}$  using a cosine schedule. The weight decay is gradually ramped up from 0.04 to 0.4 throughout training. The target encoder weights are updated via an exponential moving average (EMA) [1, 9], starting with a momentum of 0.996 and linearly increasing to 1 throughout training. REJEPa employs a random disjoint masking strategy with an optimal masking ratio of 0.25, distinguishing itself from conventional predictive learning approaches [2]. Context and target masks are independently sampled for each image in the mini-batch size of 128, ensuring diverse masking patterns. Additionally, the coefficients of VICReg [ $\lambda_v, \lambda_c, \lambda_i$ ], employed to prevent degenerate solutions, are set to [25, 25, 1] as followed by [7].

To evaluate the performance of REJEPa in content-based image retrieval (CBIR) tasks, we adopt the F1 score as the primary metric. The F1 score, defined as the harmonic mean of precision and recall, provides a balanced measure of retrieval accuracy by considering both the relevance of retrieved images (precision) and the completeness of the retrieval process (recall). The retrieval performance is evaluated based on the top-10 retrieved images for

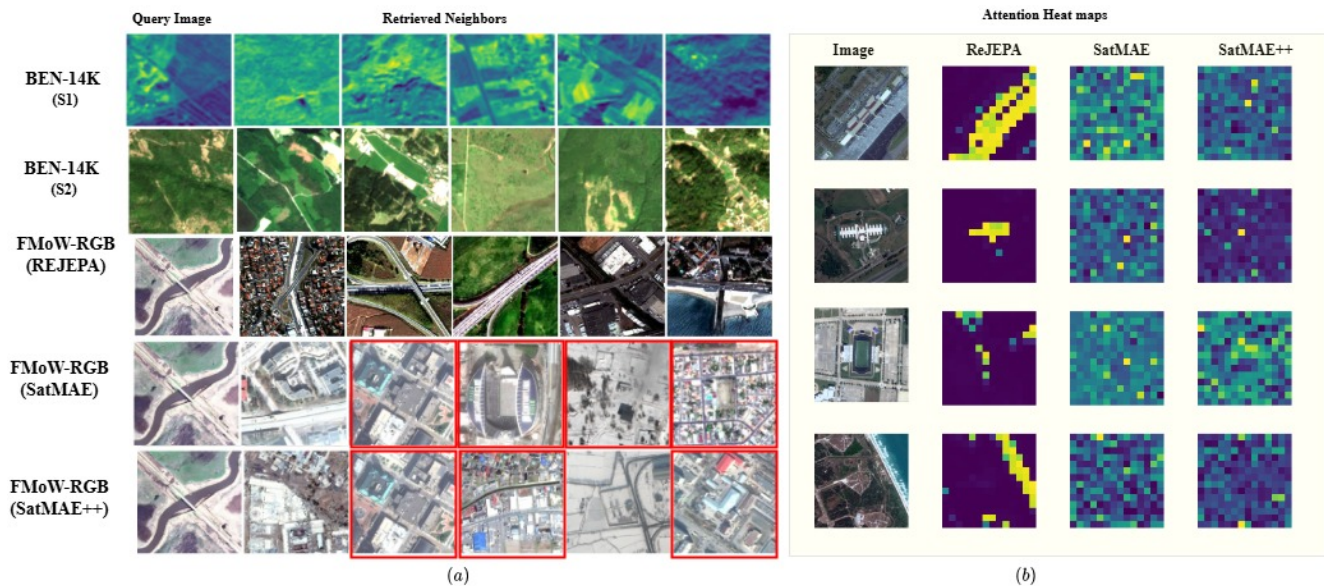


Figure 4. Qualitative retrieval and attention visualization results of REJEPa. (a) Retrieval results on BEN-14K and FMoW datasets. On FMoW, REJEPa retrieves more semantically coherent images than SatMAE and SatMAE++, particularly in complex scenes with high intra-class variance. (b) Context encoder attention heatmaps comparing REJEPa with SatMAE and SatMAE++. REJEPa attends to key structural regions while SatMAE and SatMAE++ exhibit more dispersed and less discriminative attention patterns, highlighting the advantage of feature-space prediction for remote sensing retrieval. Red colored boxes show wrong retrieval.

each query, where both the query and archive images are encoded into the feature space using the encoder. The encoded archive images serve as a retrieval bank, from which the most relevant images are retrieved based on the query, similar to CSMAE [17].

### 4.3. Retrieval Performance

We assess the effectiveness of REJEPa by benchmarking it against state-of-the-art self-supervised learning (SSL) models for content-based image retrieval (RS-CBIR). The competing methods include MAE [18], MAE-RVSA [36], SatMAE [13], SatMAE++ [28], SS-CMIR [33], Scale-MAE [30], MaskVLM [20], and CSMAE-SESD [17]. To ensure a fair comparison, all models were trained for the same number of epochs. The retrieval performance is evaluated across three diverse datasets—BEN-14K [31], FMoW-RGB [12], and FMoW-Sentinel [13], covering a range of SAR, multispectral, and high-resolution optical imagery. Table 1 shows that REJEPa surpasses existing SSL-based retrieval models (of same backbone ViT-B/16) while eliminating pixel reconstruction overhead, achieving superior performance with significantly lower computation and better scalability for large-scale RS data. Compared to parameter-heavy MAE-based models such as SatMAE, SatMAE++, Scale-MAE, and MAE, REJEPa reduces the model size by 10–40%.

On BEN-14K, REJEPa achieves an overall 4.5% improvement in F1-score, consistently excelling in  $S1 \rightarrow S1$  and  $S2 \rightarrow S2$  retrieval, surpassing MAE and SS-CMIR by

4–6% while maintaining efficiency across SAR and multi-spectral modalities. Notably, REJEPa performs 22% better for S1 and 2% better for S2 than all MAE-based setups, except for SatMAE and SatMAE++ in the S2 setup, where these models benefit from extensive scale-specific optimizations. Additionally, Scale-MAE [30] cannot be utilized with multi-spectral data ( $S2 \rightarrow S2$ ) due to its reliance on Ground Sample Distance Positional Encoding (GSDPE), which constrains its application to RGB channels only.

For a fair comparison, we take CSMAE-SESD (Disjoint) as a baseline model since it follows a similar non-overlapping masking strategy, aligning with our experimental setup. In FMoW-RGB, where scene complexity and intra-class variance pose challenges, REJEPa improves retrieval by 8.7% over SS-CMIR and 17.2% over SatMAE, demonstrating its effectiveness in learning robust semantic representations. Similarly, on FMoW-Sentinel, it generalizes effectively despite spectral distortions and resolution mismatches, achieving a 13.2% improvement over SS-CMIR and 10.1% over MaskVLM, highlighting its adaptability to multispectral satellite data.

### 4.4. Visualization Analysis

To illustrate the effectiveness of REJEPa in RS-CBIR, we present retrieval results on BEN-14K and FMoW datasets, along with comparative retrieval performance and attention visualizations. Figure 4(a) shows top-5 retrievals, demonstrating our model’s ability to retrieve semantically relevant

images across varying resolutions and sensor modalities. Unlike generative methods that emphasize pixel similarity, our feature-predictive learning prioritizes semantic consistency, ensuring structurally coherent retrievals. For FMoW datasets, we conduct a direct comparison against SatMAE and SatMAE++, on a particular query like *roadbridge*. As shown in Figure 4(a), while these models achieve strong performance, they tend to focus on texture-based similarities rather than structural and semantic relationships. On the contrary, REJEPa, by leveraging feature-space prediction, retrieves semantically meaningful scenes rather than relying on surface-level pixel cues, achieving higher retrieval accuracy even in complex urban and natural environments. Figure 4(b) visualizes attention heatmaps from the context encoder. Unlike pixel-reconstruction methods, which disperse attention across the entire image, REJEPa selectively focuses on salient geographic structures, such as roads, vegetation patches, and water bodies, while suppressing background clutter. This spatial selectivity enhances retrieval accuracy, particularly in multi-sensor settings where spectral distortions and resolution mismatches pose challenges.

#### 4.5. Ablation Studies

We conducted a series of ablation studies on BEN-14K to analyze the impact of predictor depth, masking strategy, masking ratio, and VICReg regularization on RS-CBIR performance.

**Effect of Predictor Depth:** Table 2 shows that increasing the predictor depth enhances retrieval performance, with 12 layers yielding a 26.5% and 27.8% improvement over a shallow 6-layer predictor for S1→S1 and S2→S2, respectively. This confirms that deeper predictors enable more expressive feature mapping to capture high-level semantic relationships between context and target representations.

**Impact of Masking Strategy:** As seen in Table 3, the random masking strategy outperforms the conventional multi-block masking [2] by 6.9% (S1→S1) and 7.9% (S2→S2), validating the effectiveness of our disjoint random masking approach in learning non-trivial contextual dependencies. Unlike multi-block masking, where structured regions of an image are masked together, random masking increases feature diversity and reduces spurious correlations between context and target regions for RS images.

**Impact of Masking Ratio:** Table 4 suggests that an optimal masking ratio of 0.25 achieves the best performance, improving retrieval by 11.5% (S1→S1) and 13.5% (S2→S2) compared to a higher 0.85 masking ratio.

**Effect of VICReg Regularisation:** As shown in Table 5, incorporating VICReg significantly enhances representation learning, improving retrieval by 18.4% (S1→S1) and 20.3% (S2→S2). Without VICReg, the model suffers from representation collapse, where embeddings become redundant, reducing retrieval effectiveness. VICReg mitigates

Table 2. Ablation Study on Predictor Depth for RS-CBIR Performance

Depth	S1→S1	S2→S2
6	61.62	59.01
8	64.96	62.61
10	67.61	66.82
12	<b>76.38</b>	<b>75.42</b>

Table 3. Ablation Study on Masking Strategy for RS-CBIR Performance

Masking Strategy	S1→S1	S2→S2
Multi-block	71.45	69.84
Random	<b>76.38</b>	<b>75.42</b>

Table 4. Ablation Study on Masking Ratio for RS-CBIR Performance

Masking Ratio	S1→S1	S2→S2
0.25	<b>76.38</b>	<b>75.42</b>
0.40	71.56	70.54
0.85	68.47	66.46

Table 5. Ablation Study on VICReg Regularization for RS-CBIR Performance

VICReg Applied	S1→S1	S2→S2
✓	<b>76.38</b>	<b>75.42</b>
✗	64.51	62.65

this by enforcing feature diversity (variance), ensuring distinct feature dimensions (decorrelation), and maintaining consistency between augmented views (invariance). This leads to more discriminative and compact feature embeddings, directly improving retrieval precision.

## 5. Takeaways

This work introduces REJEPa, a novel joint-embedding predictive architecture tailored for efficient and scalable RS-CBIR. By shifting from pixel-level reconstruction to feature-space prediction, the proposed framework addresses critical challenges in remote sensing image retrieval, including multi-sensor generalization, varying spatial resolutions, and complex scene structures.

## 6. Acknowledgement

B. Banerjee acknowledges the support from Anusandhan National Research Foundation (ANRF) - Grant no: **CRG/2023/004389**

## References

- [1] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *European Conference on Computer Vision*, pages 456–473. Springer, 2022. 1, 3, 6
- [2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023. 2, 3, 6, 8
- [3] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. Omnisat: Self-supervised modality fusion for earth observation. In *European Conference on Computer Vision*, pages 409–427. Springer, 2025. 3
- [4] 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. 2
- [5] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR, 2022. 3
- [6] Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli. Efficient self-supervised learning with contextualized target representations for vision, speech and language. In *International Conference on Machine Learning*, pages 1416–1429. PMLR, 2023. 3
- [7] Adrien Bardes, Jean Ponce, and Yann LeCun. Vircreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021. 2, 5, 6
- [8] Jules Bourcier, Gohar Dashyan, Karteek Alahari, and Jocelyn Chanussot. Learning representations of satellite images from metadata supervision. In *European Conference on Computer Vision*, pages 54–71. Springer, 2025. 2
- [9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 6
- [10] Keumgang Cha, Junghoon Seo, and Taekyung Lee. A billion-scale foundation model for remote sensing images. *arXiv preprint arXiv:2304.05215*, 2023. 3
- [11] 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. PMLR, 2020. 1
- [12] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018. 6, 7
- [13] 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. 1, 3, 6, 7
- [14] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 6
- [15] Iris Dumeur, Silvia Valero, and Jordi Inglada. Self-supervised spatio-temporal representation learning of satellite image time series. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024. 3
- [16] Anthony Fuller, Koreen Millard, and James Green. Croma: Remote sensing representations with contrastive radar-optical masked autoencoders. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [17] Jakob Hackstein, Gencer Sumbul, Kai Norman Clasen, and Begüm Demir. Exploring masked autoencoders for sensor-agnostic image retrieval in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 6, 7
- [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, 3, 6, 7
- [19] Heechul Jung, Yoonju Oh, Seongho Jeong, Chaehyeon Lee, and Taekyun Jeon. Contrastive self-supervised learning with smoothed representation for remote sensing. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021. 3
- [20] Gukyeong Kwon, Zhaowei Cai, Avinash Ravichandran, Erhan Bas, Rahul Bhotika, and Stefano Soatto. Masked vision and language modeling for multi-modal representation learning. *arXiv preprint arXiv:2208.02131*, 2022. 6, 7
- [21] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022. 3
- [22] Wenyuan Li, Keyan Chen, Hao Chen, and Zhenwei Shi. Geographical knowledge-driven representation learning for remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2021. 3
- [23] Weijie Li, Wei Yang, Tianpeng Liu, Yuenan Hou, Yuxuan Li, Zhen Liu, Yongxiang Liu, and Li Liu. Predicting gradient is better: Exploring self-supervised learning for sar atr with a joint-embedding predictive architecture. *ISPRS Journal of Photogrammetry and Remote Sensing*, 218:326–338, 2024. 2, 3
- [24] Zhihao Li, Biao Hou, Siteng Ma, Zitong Wu, Xianpeng Guo, Bo Ren, and Licheng Jiao. Masked angle-aware autoencoder for remote sensing images. In *European Conference on Computer Vision*, pages 260–278. Springer, 2025. 3
- [25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6
- [26] Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Un-supervised pre-training from uncurated remote sensing data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9414–9423, 2021. 3
- [27] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and

- Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023. [3](#)
- [28] Mubashir Noman, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan. Rethinking transformers pre-training for multispectral satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27811–27819, 2024. [6](#), [7](#)
- [29] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. [3](#)
- [30] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uytendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4088–4099, 2023. [3](#), [6](#), [7](#)
- [31] Gencer Sumbul, Arne De Wall, Tristan Kreuziger, Filipe Marcelino, Hugo Costa, Pedro Benevides, Mario Caetano, Begüm Demir, and Volker Markl. Bigearthnet-mm: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 9(3):174–180, 2021. [7](#)
- [32] Gencer Sumbul, Jian Kang, and Begüm Demir. Deep learning for image search and retrieval in large remote sensing archives. *Deep learning for the earth sciences: a comprehensive approach to remote sensing, climate science, and geosciences*, pages 150–160, 2021. [1](#)
- [33] Gencer Sumbul, Markus Müller, and Begüm Demir. A novel self-supervised cross-modal image retrieval method in remote sensing. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2426–2430. IEEE, 2022. [1](#), [6](#), [7](#)
- [34] Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang, Hao Chang, et al. Ringmo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–22, 2022. [3](#)
- [35] Wei-Hsin Tseng, Hoàng-Ân Lê, Alexandre Boulch, Sébastien Lefèvre, and Dirk Tiede. Croco: Cross-modal contrastive learning for localization of earth observation data. *arXiv preprint arXiv:2204.07052*, 2022. [2](#)
- [36] Di Wang, Qiming Zhang, Yufei Xu, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. Advancing plain vision transformer toward remote sensing foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2022. [3](#), [6](#), [7](#)
- [37] Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Lichao Mou, and Xiao Xiang Zhu. Self-supervised learning in remote sensing: A review. *IEEE Geoscience and Remote Sensing Magazine*, 10(4):213–247, 2022. [2](#)
- [38] 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](#)
- [39] Kun Yi, Yixiao Ge, Xiaotong Li, Shusheng Yang, Dian Li, Jianping Wu, Ying Shan, and Xiaohu Qie. Masked image modeling with denoising contrast. *arXiv preprint arXiv:2205.09616*, 2022. [1](#)
- [40] Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems*, 42(4):1–60, 2024. [1](#)
- [41] Zhicheng Zhao, Ze Luo, Jian Li, Can Chen, and Yingchao Piao. When self-supervised learning meets scene classification: Remote sensing scene classification based on a multitask learning framework. *Remote Sensing*, 12(20):3276, 2020. [3](#)
- [42] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. [3](#)