

FLASH-SAR: Fast Learning Self-supervised Hierarchical Architecture for SAR

Sai Shruti Prakhya and Uttam Kumar

Spatial Computing Laboratory

Department of Data Science and Artificial Intelligence

International Institute of Information Technology Bangalore

26/C, Electronics City, Hosur Road, Bengaluru, Karnataka 560100

{saishruti.praekhya018, uttam}@iiitb.ac.in

Abstract

While deep supervised methods have advanced the synthetic aperture radar (SAR) data analysis, the potential of self-supervised methods remains relatively underexplored. Prevailing self-supervised approaches largely focus on masked image modeling (MIM), which despite its effectiveness, demand extensive training time and massive datasets to achieve convergence. In this paper, we introduce FLASH-SAR (Fast Learning Self-supervised Hierarchical Architecture for SAR), a computationally efficient, self-supervised SAR model, designed to excel at both image level classification and dense-prediction tasks such as semantic segmentation. FLASH-SAR integrates the dense contrastive learning paradigm with the hierarchical Swin transformer backbone. This combination enables the model to learn local discriminative features that capture the neighborhood geometry, while preserving global semantic discriminability. We evaluate our model on a comprehensive set of tasks, including multi-label classification and land use and land cover (LULC) segmentation across 2 global benchmark dataset and 7 region specific datasets using the mIoU and mAP metrics. Our proposed model achieved state-of-the-art performance that are competitive with heavier multimodal baselines, notably surpassing them by an average of 1.07 % mAP on BigEarthNet-S1, while attaining up to 50% faster fine-tuning speeds than existing baselines. These results demonstrate that dense contrastive pre-training can be a promising alternative to MIM-based self-supervised learning approaches for SAR imagery.

1. Introduction

The proliferation of Earth Observation (EO) satellites have led to a massive influx of data, offering an unprecedented opportunity to monitor planetary changes, from urbanization [1, 12, 42, 49–52], greenhouse gas emissions [5, 14, 21,

65] to disaster management [4, 26, 53]. While the volume of data is large and spans multiple modalities, including optical [11, 13] and synthetic aperture radar (SAR) [46], the human capacity to annotate these data is significantly limited. SAR imagery in particular is extremely difficult to annotate, requiring domain experts to interpret complex back-scattering mechanisms that are counter intuitive to human visual perception. This discrepancy in the abundance of raw data and the scarcity of labeled samples is a huge bottleneck for supervised learning approaches in remote sensing. Without sufficient diversity in annotated training datasets, deep learning models struggle to generalize across varying geographies [18, 68]. Furthermore, in data-constrained regimes such as SAR automatic target recognition (ATR), limited labeled data often causes the model to overfit to irrelevant background details rather than discriminative object features [7, 54, 56, 66, 67].

To bridge this gap, foundation models pre-trained through principles of self supervised learning (SSL) have emerged as a powerful paradigm. By constructing pretext tasks such as masked image modeling (MIM) or contrastive learning, these foundation models learn generalizable representations from unlabeled data which can then be fine-tuned for various downstream tasks with limited supervision. While generic computer vision foundation models have shown remarkable success on standard RGB images [2], they often struggle with the domain specific nuances of remote sensing including the coherent speckle noise and geometric distortions such as foreshortening, layover and shadows that are inherent to the side-looking acquisition geometry of SAR. Consequently, the remote-sensing community has initiated the development of domain specific foundation models.

Although the field of foundation models has seen rapid progress, existing methods encounter specific limitations. First, the foundation model landscape remains heavily skewed toward optical imagery. Since these models are optimized for passive reflectance properties, they do not nat-

urally extend to the active back-scattering mechanisms of SAR. Second, current methods typically rely on MIM or global contrastive learning, and while these paradigms excel at representation learning, they also present inherent trade-offs. For example, MIM approaches prioritize pixel-level reconstruction, creating a texture bias that often comes at the expense of the global shape information essential for semantic discrimination. As a result, they typically require significant fine-tuning to adapt effectively to downstream tasks [38]. On the other hand, global contrastive learning struggles with dense prediction tasks such as land use and land cover (LULC) segmentation. By optimizing for image-level invariance, these methods tend to induce attention collapse in deeper layers, reducing the spatial diversity of learned representations [37]. Finally, the current trend in foundation models prioritizes massive scale, in terms of dataset size as well as parameter count, often disregarding the environmental and operational costs of such compute-intensive architectures. This hinders practical deployment in constrained settings, such as on-board satellite processing or real-time disaster response [19, 43, 47].

This work addresses these challenges by proposing a computationally efficient, self-supervised, SAR specific model tailored for dense prediction tasks. We adopt a dense contrastive learning framework that forces the model to learn both global as well as local discriminative features while preserving the shape, geometry and texture information that are critical for SAR interpretation. We demonstrate this by merging the objectives of global and dense local supervision during pre-training, and achieve competitive performance with existing state-of-the-art methods; requiring significantly less training data and compute resources. The main contribution of this work is summarised as follows:

- Introduction of a data-efficient self-supervised SAR model through dense contrastive pretraining, where multi-temporal positive pairing and local negative sampling are employed, together with a hierarchical Swin transformer backbone. It utilizes shifted window attention and hence scales linearly in time complexity with respect to image size.
- The approach matches or surpasses state-of-the-art frameworks while training on less than half the data typically required.
- The model enables faster adaptation, with fine-tuning speeds accelerated by 1.5–2× compared to existing baselines.

2. Related works

The application of SSL in remote sensing has predominantly focused on optical imagery. Early approaches adapted global contrastive frameworks such as MoCo [24] and SimCLR [8] to the geospatial domain. For instance, incorporating geolocation as a pretext task and defining posi-

tive pairs through spatio-temporal alignment [3]. Similarly, Seasonal Contrast (SeCo) [32] explicitly models temporal dependencies by learning multiple embedding subspaces, some invariant and others sensitive to seasonal change, allowing the model to generalize effectively to both classification and change detection. In the SAR domain, pre-training strategies have largely restricted to either specific tasks such as ATR [39, 61] or multi-modal frameworks that perform radar-optical contrastive learning [29, 58]. More recently, this paradigm was advanced by introducing a masked Siamese framework for SAR feature extraction [35]. This approach showed strong generalization by pre-training on a diverse multi-resolution dataset consisting of X, C and L bands with quad-polarization and relies on airborne acquisitions and high resolution inputs. Most of these approaches have primarily focussed on consistency at the global or patch level, which limits their effectiveness for dense prediction tasks. Further, as these models are optimized to distinguish global image identities, they exhibit a strong shape bias (low-frequency focus) that comes at the cost of high-frequency textural details (e.g. in built-up areas), which are critical for interpreting SAR backscatter mechanisms. Consequently, the learned representations discard fine-grained spatial information, rendering these representations suboptimal for dense prediction tasks such as LULC segmentation [44, 45] or object detection, where local discriminability is of importance.

To address the limitations of contrastive learning, recent works have shifted toward masked image modeling (MIM), which reconstructs masked portions of an input signal. MIM is typically implemented via encoder-decoder architectures, for example, masked autoencoder (MAE) [25]. However, majority of these foundation models remain restricted to the optical domain, for instance, SatMAE [9] adapted the MAE framework to multi-spectral optical data. Addressing the issue of scale ambiguity, scale-aware MAE utilized a unique masking strategy [40] to determine the scale of its vision transformer (ViT) positional encoding. The decoder reconstructed images through a bandpass filter to retain scale-specific information, leading to robust, multiscale representations. Few attempts to bridge the modality gap include CROMA [17] that sought to combine the best of both worlds by integrating contrastive alignment with masked reconstruction. While this multimodal approach leverages vast archives like SSL4EO [59] to learn rich representations, it incurs a significant computational overhead (uses 3 ViT heads as encoders), and lacks empirical validation on SAR semantic segmentation. MERLIN [10] takes a complementary direction by introducing a compute-efficient U-Net trained with auxiliary self-supervised tasks such as despeckling, segmentation, and regression, on 1 m TerraSAR-X imagery, where despeckling explicitly encourages learning fine-grained pixel-level structures rather than

coarse patch-level representations.

SUMMIT [16] represents one of the few foundation models tailored exclusively for SAR. Recognizing the vulnerability of standard MIM to speckle noise, SUMMIT augments the masked reconstruction objective with auxiliary tasks, specifically denoising, corner detection and edge reconstruction to guide the model toward physical structure rather than overfitting speckle noise patterns. While this helps mitigate texture bias, they compound the computational and representational trade-offs inherent to the MIM paradigm. Computationally, optimizing multiple reconstruction heads alongside heavy ViT backbones increase the training time, requiring 2-3x more epochs for convergence [25]. Despite the emphasis on learning structural elements, the fundamental objective remains signal reconstruction. This forces the later layers of the encoder to function as an implicit decoder, dedicating capacity to restoring pixel-level details rather than abstracting semantic concepts. Consequently, even with auxiliary denoising, the features are not semantically aligned for any task without extensive fine-tuning [38, 64], which ultimately limits the model’s efficiency as a general-purpose feature extractor. Notably, while the framework demonstrated strong performance on instance segmentation, it remained largely unevaluated on dense LULC semantic segmentation benchmarks.

To overcome the trade-offs between global invariance and masked reconstruction, we adapt dense contrastive learning [28, 55, 57, 63] paradigm. Unlike standard discriminative approaches restricted to image level optimization, the proposed framework extends similarity constraints at the pixel level. This ensures that the model preserves fine-grained structural details required for SAR data analysis, while maintaining a solely encoder based architecture. Our work focuses on maximizing the utility of standard dual-pol C band Sentinel-1 imagery (10 m spatial resolution), ensuring wider applicability to EO applications.

3. Methodology

In this section, we describe the intuition behind model architecture with global and dense contrastive learning.

3.1. Model architecture

A Swin-Tiny [30] architecture is employed for pre-training, operating at an input resolution of 224×224 . Leveraging hierarchical shifted window attention, the four encoder stages output feature embeddings with channel dimensions of 96, 192, 384 and 768, respectively. The final stage output is reshaped into a 2D feature map to enable processing by the convolutional dense head, while the globally pooled features are processed by the MLP (Multilayer Perceptron) head. For fine-tuning, we utilize a UPerNet-style decoder [62], which integrates a Pyramid Pooling Module (PPM) with Feature Pyramid Network (FPN) fusion. This architec-

Algorithm 1: Dense Contrastive Pre-training for SAR (FLASH-SAR)

Input: Unlabeled SAR dataset \mathcal{D} , Batch size $k = 16$, Temperature $\tau = 0.1$, Balance weight $\lambda = 0.5$

Output: Pre-trained Swin Encoder g_θ

- 1 Initialize Swin-Tiny encoder g_θ , MLP head h_{mlp} , and Dense head h_{dense} with random weights
- 2 **for** $epoch \leftarrow 1$ **to** 100 **do**
- 3 **for** each mini-batch $\{A, P, N_1, \dots, N_{k-2}\}$ sampled from \mathcal{D} **do**
- 4 // 1. Forward Pass
- 4 Generate backbone feature maps:
- 5 $f_A, f_P, \{f_{N_i}\} \leftarrow g_\theta(A), g_\theta(P), \{g_\theta(N_i)\}$
- 5 // 2. Global Contrastive
- 6 Generate global embeddings:
- 7 $e_A, e_P, \{e_{N_i}\} \leftarrow h_{mlp}(\text{GlobalPool}(f_A, f_P, \{f_{N_i}\}))$
- 8 Compute \mathcal{L}_{gc} using $\{e_A, e_P, \{e_{N_i}\}\}$ via Eq. (1)
- 9 // 3. Dense Contrastive
- 9 Generate dense embeddings:
- 10 $d_A, d_P \leftarrow h_{dense}(f_A), h_{dense}(f_P)$
- 10 // Correspondence Matching
- 11 Downsample f_A, f_P to match spatial resolution of d
- 12 Determine pixel-wise correspondence indices c between f_A and f_P via Eq. (3)
- 13 Compute \mathcal{L}_{dc} using d_A, d_P and indices c via Eq. (4)
- 14 // 4. Optimization
- 14 Compute total loss:
- 15 $\mathcal{L}_{total} \leftarrow \lambda \mathcal{L}_{gc} + (1 - \lambda) \mathcal{L}_{dc}$
- 15 Backpropagate $\nabla \mathcal{L}_{total}$ and update parameters $\theta, h_{mlp}, h_{dense}$
- 16 **end**
- 17 **end**

ture fuses features from all four encoder stages via skip connections and upsampling. Specifically, the PPM is applied to the final encoder map to capture global context; subsequently, multi-scale features are concatenated and upsampled to generate dense predictions at the original resolution.

3.2. Pre-training objectives

Global contrastive learning is based on the simple idea of pulling similar data points (positive pairs) as close together as possible, while pushing dissimilar ones (negative pairs) far apart [8, 22, 24]. The definition of ‘positive pairs’ is crucial in this approach as the model is forced to learn invariant representations for the samples that are deemed similar.

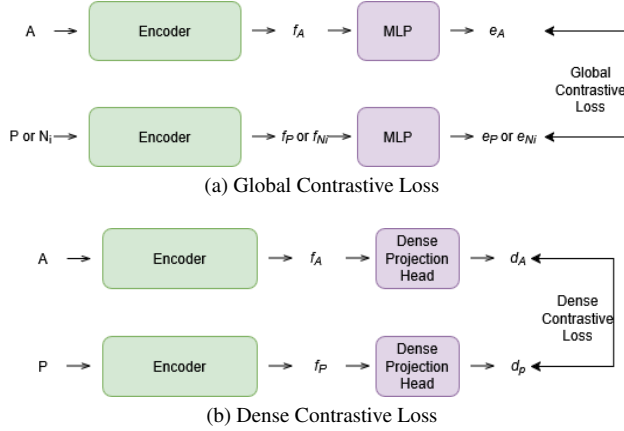


Figure 1. Overview of dense contrastive pre-training. (a) Illustrates the global contrastive loss computation using the multi-layer perception (MLP) which acts as the global projection head. (b) Illustrates the dense contrastive loss computation using the dense projection head.

In this framework, positive pairs are defined through geographic correspondence, that is, two SAR images of the same geo-location acquired at different time stamps. The considered dataset in this work consists of imagery captured between March 2020 and December 2021, indicating that the temporal variability is mainly seasonal. This forces the model to learn seasonally invariant representations of the data. Negative samples are defined as ‘everything else’ i.e. SAR images covering different locations are all defined as negative with respect to each-other. The training approach for global contrastive learning is described below.

Assume a mini-batch of size k containing an anchor A , its temporal positive P , and a set of negatives N_1, \dots, N_{k-2} sourced from different locations. The encoder parametrized by θ , transforms these inputs into intermediate features $f_A, f_P, f_{N_1}, \dots, f_{N_{k-2}}$, which are subsequently mapped by a MLP head to lower dimensional embeddings $e_A, e_P, e_{N_1}, \dots, e_{N_{k-2}}$. We formulate the global contrastive loss function denoted by L_{gc} , based on the InfoNCE objective [36], on these projected embeddings to maximize the similarity between the positive pairs (A, P) while minimizing the similarity with the negatives. Here, both positive and negative pairs are defined with respect to the anchor image. This objective back-propagates gradients through the entire architecture, thus updating both the encoder and the projection head.

$$L_{gc} = -\log \left(\frac{\exp(e_A \cdot e_P / \tau)}{\exp(e_A \cdot e_P / \tau) + \sum_{i=1}^{k-2} \exp(e_A \cdot e_{N_i} / \tau)} \right) \quad (1)$$

Probabilistically, the InfoNCE objective in Eq. (1) functions as a form of implicit negative sampling [20, 33, 34],

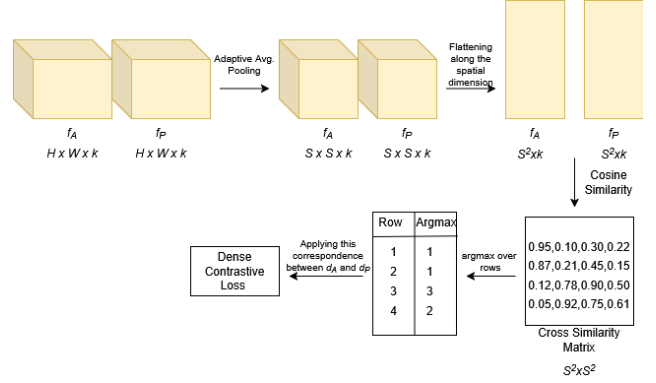


Figure 2. Dense correspondence matching. Illustration of the pixel-wise alignment strategy used to compute L_{dc} .

maximizing the log-likelihood of correctly identifying the true positive samples among a set of negatives. The negatives in the denominator serve as contrasting evidence that regularizes the embedding space. Otherwise, the model minimizes the loss trivially by mapping all inputs to a constant vector. This mechanism prevents such representation collapse by enforcing a probabilistic competition, which ensures that the model learns discriminative representations by pulling semantically related features together while actively repelling dissimilar ones to structure the embedding space. The temperature hyper-parameter τ controls the degree of the softmax.

In order to bridge the gap between the global level feature invariance and local discriminability, we adopt the dense contrastive learning framework [57]. This extends the contrastive objective to local features, ensuring that the model captures the fine-grained spatial structure for dense prediction.

As illustrated in Fig. 1b, we supplement the standard contrastive pipeline with a parallel dense head. Input images are processed through the encoder to return spatial maps. While the globally pooled features are projected via an MLP head for the global objective, the unpooled feature maps are simultaneously passed through a dense projection head, consisting of a series of 1×1 convolution blocks to obtain the dense feature embeddings d . The network is optimized via a linear combination of the global contrastive loss, L_{gc} in Eq. (1) and the dense contrastive loss, L_{dc} . Therefore,

$$L_{total} = \lambda \cdot L_{gc} + (1 - \lambda) \cdot L_{dc} \quad (2)$$

The process for defining positive pairs for dense features is illustrated in Fig. 2. Correspondence is derived between the anchor A and its temporal positive P . Let the encoder features f have dimensions $\mathbb{R}^{H \times W \times K}$ and the dense features obtained from the dense projection embeddings d have

dimensions $\mathbb{R}^{S \times S \times E}$ (assuming a square spatial grid of S).

We determine the dense feature correspondence of d_A and d_P using the encoder feature maps f_A and f_P . First, f_A and f_P are down-sampled through adaptive average pooling to match the $S \times S$ spatial dimensions of the dense embeddings d_A and d_P , then flattened into matrices of shape $S^2 \times k$. A pairwise cosine similarity is computed between these two matrices i.e. for each of the S^2 feature vectors in f_A we identify the vector in f_P that maximizes cosine similarity; this recognizes the dense feature correspondence used to align d_A and d_P , and is formulated as:

$$c_i = \arg \max_j \text{sim}(f_{A_i}, f_{P_j}). \quad (3)$$

Analogous to the global objective in Eq. (1), we formulate the dense contrastive loss L_{dc} by applying the InfoNCE principle to the dense features:

$$L_{dc} = \frac{1}{S^2} \sum_{s=1}^{S^2} -\log \left(\frac{\exp(d_A^s \cdot d_{P+}^s / \tau)}{\exp(d_A^s \cdot d_{P+}^s / \tau) + \sum_{d_{P-}} \exp(d_A^s \cdot d_{P-} / \tau)} \right). \quad (4)$$

Here, d_A^s denotes the feature vector at spatial index s in the anchor map d_A , and d_{P+}^s denotes its corresponding positive match in d_P . The set of negatives, d_{P-} , comprises all other feature vectors within the positive view d_P that are not the matched d_{P+}^s . Notably, this definition diverges from the original DenseCL formulation [57], which typically incorporates pooled vectors from other images as negatives. By treating spatially distinct regions within the paired view as negatives, we explicitly force the model to discriminate between different semantic parts of the same scene, thereby enhancing local feature distinctiveness. This dense correspondence step leads to a negligible computational overhead in total training time ($< 1\%$) when compared to the MoCo-v2 baseline [24].

It is important to jointly optimize both L_{dc} and L_{gc} . Setting λ in Eq. (2) to 0, hinders convergence as good global features are essential in order to learn good dense features. Therefore the optimum value of λ was considered as 0.5 based on several experiments [57]. The complete pre-training procedure, integrating both global and dense objectives is summarized in Algorithm 1.

4. Experimental setup

4.1. Pre-training

Data: FLASH-SAR was pre-trained on the SSL4EO-S12 dataset [59], utilising only the C-band, dual-pol (VV and VH), Sentinel-1 SAR modality. From the original corpus of over one million images spanning over 10000 global locations, we sampled 2 non-overlapping patches per location (derived from metadata) to minimize spatial redundancy

Table 1. Hyperparameter configurations. All FLASH-SAR experiments used Swin-Tiny, AdamW and 224×224 input. When fine-tuning on CROMA, we used its native 120×120 input. Segmentation experiments employed an 80-20 train-test split. All other hyper-parameter configurations remained the same across models.

Param	Pre-train	Single-Cls	Multi-Cls	Seg
Dataset	SSL4EO	EuroSAT	BigEarth	Urban
Batch	16	32	32	8
LR	$1e^{-4}$	$1e^{-3}$	$1e^{-3}$	$1e^{-4}$
Schedule	Const.	Cosine (5)	Cosine (5)	Patience
Epochs	100	100	100	100
Loss	DenseCL	CrossEntropy (CE)	Binary CE	CE+Dice
Metric	-	Acc	mAP	mIoU

and bias. This resulted in a balanced training set of approximately 400,000 patches.

Implementation details: The framework was implemented using PyTorch, with training conducted on a single NVIDIA Quadro RTX 8000 GPU (48 GB VRAM). Optimization was performed using AdamW [31], chosen for its stability in transformer training. Hyperparameter settings for pre-training and fine-tuning are summarized in Tab. 1.

Table 2. Classification results on EuroSat-SAR (Top 1 Accuracy %) and BigEarthNet S1 (mAP %) datasets. Best results are highlighted in bold.

Dataset	Model	Full FT	LP	LoRA	Part-FT	Scratch
EuroSat	FLASH-SAR	82.08	73.47	81.32	81.43	39.88
	CROMA	84.64	71.82	83.11	82.41	-
	ResNet50	82.72	64.77	84.54	78.38	-
BigEarth	FLASH-SAR	69.98	69.99	69.99	63.31	54.28
	CROMA	66.03	67.35	68.92	66.69	-
	ResNet50	64.07	53.29	65.67	62.36	-

4.2. Fine-tuning

Data: The model was fine-tuned on the EuroSAT-SAR dataset [60] (10 classes) for single-label classification and the BigEarthNet-S1 dataset [48] (19 classes) for multi-label tasks. To test the few-shot generalization capability of the model, we conducted the experiments using a randomly sampled 1% (for BigEarthNetS1 - 2349 samples) and 10% (for EuroSAT-SAR - 2700 samples) subset of the training data, while validating on the entire validation split (BigEarthNetS1 - 122347 samples, EuroSAT-SAR - 24300), only after the final fine-tuning epoch was completed. Region-specific adaptation was evaluated using locally curated LULC segmentation datasets spanning seven major Indian cities: Delhi, Mumbai, Hyderabad, Kolkata, Ahmedabad, Pune and Bangalore. For each city, we obtained Sentinel-1 SAR imagery along with the corresponding land cover labels derived from the Dynamic World product [6] using Google Earth Engine for the month of March

Table 3. LULC segmentation performance. We report mIoU scores across seven major Indian cities under five adaptation protocols: full fine-tuning (Full-FT), frozen encoder (Frozen-Enc), LoRA [27], random initialization (Scratch), and partial fine-tuning (Part-FT). Computational efficiency is reported in total training minutes and seconds per epoch. Best results per city are highlighted in bold.

City	Model	Full-FT	Frozen-Enc	LoRA	Scratch	Part-FT	Tot. Time (Min)	Time/Ep (Sec)
Delhi	FLASH-SAR	0.5730	0.5543	0.5593	0.4999	0.5612	12.93	20.94 ± 3.18
	CROMA	0.5576	0.5522	0.5641	–	0.5552	23.34	48.50 ± 0.96
	ResNet50	0.5363	0.5121	0.5011	–	0.5319	14.73	19.52 ± 0.67
Mumbai	FLASH-SAR	0.5741	0.5665	0.5688	0.5628	0.5672	6.78	4.33 ± 0.85
	CROMA	0.5599	0.5680	0.5752	–	0.5666	9.42	8.52 ± 4.13
	ResNet50	0.5516	0.5714	0.5346	–	0.5556	7.07	2.61 ± 0.30
Hyderabad	FLASH-SAR	0.5704	0.5645	0.5612	0.4518	0.5368	9.60	16.68 ± 0.81
	CROMA	0.5220	0.5183	0.5307	–	0.5147	20.02	39.17 ± 1.13
	ResNet50	0.4835	0.4807	0.4803	–	0.4848	5.69	12.35 ± 3.62
Kolkata	FLASH-SAR	0.5701	0.5636	0.5694	0.4605	0.5674	14.54	17.56 ± 0.65
	CROMA	0.5584	0.4976	0.5270	–	0.5580	22.85	40.88 ± 0.69
	ResNet50	0.4780	0.4633	0.4468	–	0.4745	5.69	9.21 ± 0.52
Ahmedabad	FLASH-SAR	0.4968	0.5007	0.4931	0.4073	0.4942	5.11	5.07 ± 0.38
	CROMA	0.5133	0.4972	0.5168	–	0.5125	8.93	11.99 ± 0.31
	ResNet50	0.4817	0.4670	0.4489	–	0.4819	6.09	2.34 ± 0.50
Pune	FLASH-SAR	0.5314	0.5395	0.5382	0.4261	0.5409	7.04	5.53 ± 0.42
	CROMA	0.4674	0.4692	0.4685	–	0.4716	8.97	13.42 ± 0.36
	ResNet50	0.4646	0.4509	0.4372	–	0.4614	7.10	3.53 ± 0.64
Bangalore	FLASH-SAR	0.4998	0.4975	0.4974	0.4489	0.4978	10.66	17.50 ± 0.67
	CROMA	0.5030	0.5118	0.5230	–	0.5198	18.16	38.36 ± 3.51
	ResNet50	0.4629	0.4463	0.4408	–	0.4618	7.74	9.15 ± 0.47

2025. The SAR imageries were preprocessed using the SNAP software, where speckle noise was reduced through filtering, and the images were co-registered with their corresponding ground truth maps to achieve pixel-level alignment. The dataset was categorized into 5 semantic classes: water, built-up, agriculture, trees and open land.

Implementation details: In this work, all the tasks were benchmarked against CROMA [17] and ResNet50 [23] (pre-trained on ImageNet [15]) architectures using 4 distinct validation strategies: full fine-tuning, linear probing (encoder frozen), LoRA based fine-tuning (injecting trainable low-rank adapters), and partial fine-tuning (freezing early encoder layers). Comparisons with SUMMIT [16] are omitted, as pre-trained weights are not publicly available.

For both single and multi-label classification, a linear classification head was attached to the global feature representations of each backbone. For semantic segmentation, the decoder architecture was tailored to maximize the performance of each backbone. While FLASH-SAR employs a UPerNet, we attached a UNet decoder to ResNet50 and a simple deconvolutional decoder to CROMA, as empirical analysis indicated that the latter yields superior results compared to UPerNet-style heads for CROMA’s global representations. To ensure a strictly fair internal comparison, we reproduced all baselines in our environment using identical hyperparameters and data splits (Tab. 1). While this comparison may yield different results compared to original publications [17, 23], it demonstrates the efficiency of

the FLASH-SAR architecture in resource constrained environments.

Evaluation metrics: Top-1 Accuracy [16] is reported for single-label classification on EuroSAT-SAR. For multi-label classification on BigEarthNet-S1, mean average precision (mAP) [17] was utilised, calculated as the mean area under the precision-recall curve across all the categories to account for class imbalance and multi-class co-occurrences. For semantic segmentation, performance was assessed using mean intersection over union (mIoU) [16], calculated as the average overlap between predicted and ground truth masks across all the classes.

5. Results and discussion

As evident from Tab. 2, FLASH-SAR achieved superior performance on the challenging BigEarthNet data. In the standard full fine-tuning and linear probing regimes, it outperformed the multimodal baseline by 3.95% and 2.64% respectively. With LoRA adaptation, FLASH-SAR maintained a lead of 1.07%, demonstrating that dense pre-training yields highly adaptive features even under parameter efficient constraints. While CROMA showed a significant advantage in partial fine-tuning, FLASH-SAR consistently dominated in the high-performance experiments. Notably, FLASH-SAR surpassed the ResNet50 baseline by over 4.3% (LoRA) and 16.7% (linear probing), confirming that generic transfer learning struggled to capture the semantic complexity of multi-label SAR scenes when com-

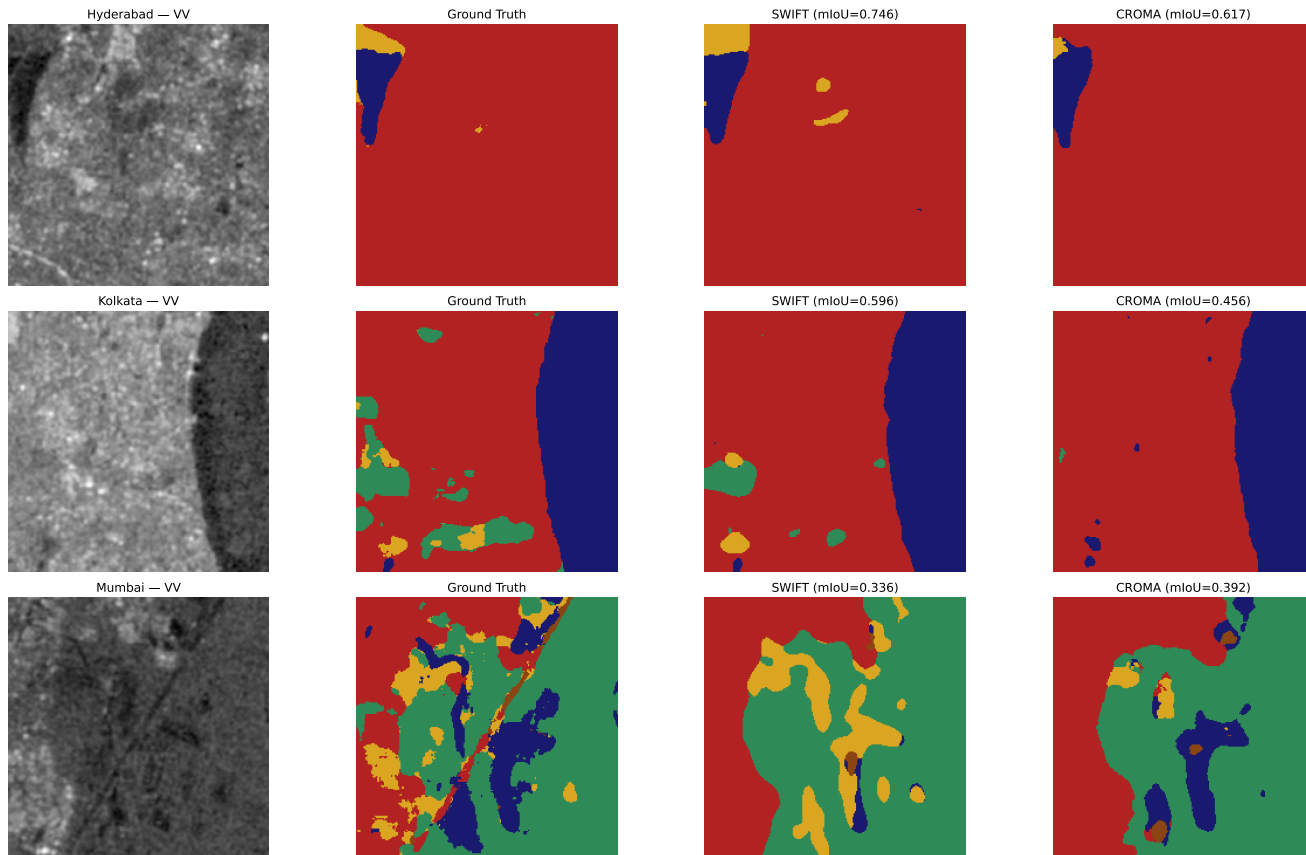


Figure 3. Qualitative Segmentation. LULC comparison (water: blue, trees: green, built-up: red, open land: brown and agriculture: yellow). FLASH-SAR resolves fine-grained agriculture (Top) and vegetation (Middle) missed by baselines. Bottom: Shared failure case due to low-contrast water features and multiple overlapping land-use categories.

pared to domain-specific pre-training.

However, on the EuroSAT-SAR single-label benchmark, CROMA and ResNet50 generally outperformed FLASH-SAR in full fine-tuning and LoRA experiments, with ResNet50 achieving a peak accuracy of 84.54% using LoRA. FLASH-SAR retained a significant lead in linear probing (73.47% vs. (64.77% for ResNet50) and (71.82% for CROMA)), indicating that its intrinsic, pre-trained representations were more robust and discriminative than features obtained through multi-modal or ImageNet pre-training. This suggested that while strategies permitting encoder adaptation (LoRA/full fine-tuning) allowed high-capacity baselines to bridge the domain gap for simple classification, FLASH-SAR’s dense objective provided superior initial feature alignment, which is a critical advantage for data-efficient adaptation.

In semantic segmentation tasks, FLASH-SAR achieved consistent performance across various fine-tuning protocols. It matched or often exceeded the multi-modal baseline and outperformed the standard ResNet50 + UNet [41] baseline across all evaluated regions (Tab. 3). In terms of ef-

iciency, by forgoing the computationally expensive MIM, FLASH-SAR was up to 50% more efficient in total fine-tuning time than CROMA. Performance gains were especially pronounced in New Delhi, Hyderabad, Kolkata and Pune, where FLASH-SAR achieved an average improvement of 3.5% over CROMA during the full fine-tuning experiments. The gains proved substantial in cities like Hyderabad and Pune where FLASH-SAR surpassed the multi-modal baseline by 4.7 and 6.3% respectively. Performance degradation while freezing the encoder was minimal (often $< 1\%$), providing strong empirical evidence that the pre-trained features are intrinsically discriminative even without extensive fine-tuning.

Fig. 3 visualizes the semantic segmentation performance across diverse urban landscapes (Hyderabad, Kolkata and Mumbai). FLASH-SAR demonstrated superior sensitivity to fine-grained semantic features, particularly in complex mixed-use areas. As seen in the Hyderabad (top row) and Kolkata (middle row) samples, our model successfully delineated minority classes such as agriculture (shown in yellow colour) and trees (in green) which CROMA

missed. The baseline exhibited a tendency towards over-regularization, often collapsing small, textured regions into the dominant built-up class (red). FLASH-SAR preserved these minority classes, which resulted in higher mIoU scores. However, the Mumbai sample (bottom row) illustrates a shared failure mode where the low contrast between water features and the surrounding terrain led to significant ambiguity. Both models struggled to identify the classes, resulting in noisy and fragmented predictions.

Overall, our experimental analysis provided three primary insights regarding the design and efficacy of SAR foundation models. First, a substantial performance drop was observed when training from scratch, and it highlighted the critical role of FLASH-SAR’s pre-training in stabilizing convergence, particularly in data-constrained regimes. Second, FLASH-SAR maintained high performance across the full spectrum of applications, ranging from image-level classification to dense pixel-level predictions. This task-agnostic robustness suggested that the dense contrastive objective effectively encoded fundamental scattering properties, establishing FLASH-SAR as a highly adaptable model for diverse scenarios. Finally, the results demonstrated that efficient pre-training strategies on moderate-scale datasets could rival or exceed the performance of larger models. These gains were achieved despite using less than 50% of the data volume compared to the baselines, effectively validating the potential of pure-SAR foundation models to learn rich, generalizable representations. Despite these encouraging results, the current study is limited by its reliance on static single-sensor, dual-polarized Sentinel-1 data, a restricted set of baselines, and the absence of evaluation on object detection benchmarks. Addressing these limitations, through extensions to multi-sensor, multi-polarization settings and broader downstream tasks, represents a natural and important direction for future work.

Ablation: The impact of pre-training data size on downstream performance was analysed by conducting experiments with 1.6%, 3.2%, 40% and 100% of the available training samples. Fig. 4 illustrates that dataset scaling yielded more pronounced performance gains for multi-label classification compared to semantic segmentation. The model exhibited rapid early learning in segmentation, where doubling the pre-training data from 1.6% to 3.2% led to an average gain of 1.05% mIoU, whereas scaling from 40% to 100% resulted in a marginal gain of only 0.5% mIoU. This diminishing return at higher data volumes highlights the early convergence and high data efficiency of the FLASH-SAR framework, suggesting that the dense contrastive objective learned robust structural features even in very low-data regimes. We attribute this efficiency to the framework’s strong inductive bias, since the deterministic properties of SAR surfaces (such as roughness and dielectric constant)

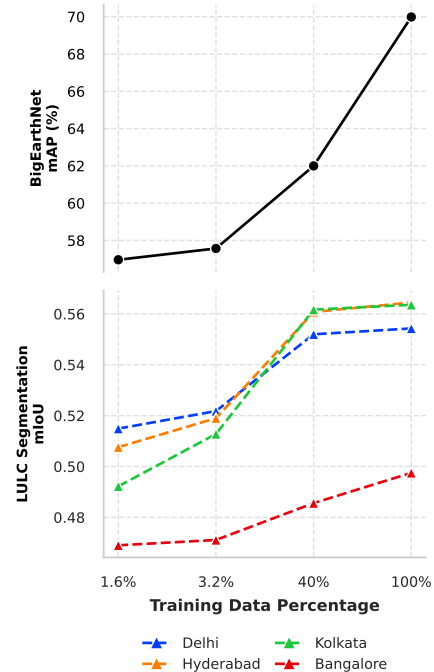


Figure 4. Data scaling. Top: Classification improves log-linearly with data. Bottom: Segmentation saturates early, peaking at 40% data.

were effectively captured with minimal pre-training data by enforcing local feature consistency. The continued linear growth in classification performance, in contrast, indicated that global semantic modeling remains a key area for algorithmic refinement.

6. Conclusion

This work proposed FLASH-SAR, a computationally efficient SAR model trained via a novel adaptation of dense contrastive learning. By defining dense negative pairs within globally aligned views, the framework prioritizes the pixel-level granularity that is critical for studying regional changes and spatiotemporal dynamics in remote sensing. Future work will focus on architectural evolutions to enhance high-level scene abstraction for single-label classification tasks, without compromising the structural efficiency established here in order to evolve FLASH-SAR into a comprehensive foundation model for SAR data.

Acknowledgments

We are grateful to International Institute of Information Technology Bangalore, India for the infrastructure support and acknowledge Mphasis Cognitive Computing Centre of Excellence for the financial assistance.

References

- [1] Kabila Abass, Selase Kofi Adanu, and Seth Agyemang. Peri-urbanisation and loss of arable land in kumasi metropolis in three decades: Evidence from remote sensing image analysis. *Land use policy*, 72:470–479, 2018. 1
- [2] Deeksha Aggarwal, Sai Shruti Prakhya, and Uttam Kumar. Agriculture crop monitoring for yield estimation with zero-shot fruit detection: A deep learning approach. In *Remote Sensing of Land Cover and Land Use Changes in South and Southeast Asia, Volume 1*, pages 115–132. CRC Press. 1
- [3] Kumar Ayush, Burak Uz Kent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning, 2022. 2
- [4] Olalekan Mumin Bello and Yusuf Adedoyin Aina. Satellite remote sensing as a tool in disaster management and sustainable development: towards a synergistic approach. *Procedia-Social and Behavioral Sciences*, 120:365–373, 2014. 1
- [5] Hartmut Boesch, Yi Liu, Johanna Tamminen, Dongxu Yang, Paul I Palmer, Hannakaisa Lindqvist, Zhaonan Cai, Ke Che, Antonio Di Noia, Liang Feng, et al. Monitoring greenhouse gases from space. *Remote Sensing*, 13(14):2700, 2021. 1
- [6] Christopher F Brown, Steven P Brumby, Brookie Guzder-Williams, Tanya Birch, Samantha Brooks Hyde, Joseph Mazzariello, Wanda Czerwinski, Valerie J Pasquarella, Robert Haertel, Simon Ilyushchenko, et al. Dynamic world, near real-time global 10 m land use land cover mapping. *Scientific data*, 9(1):251, 2022. 5
- [7] Changjie Cao, Zongjie Cao, and Zongyong Cui. Ldgan: A synthetic aperture radar image generation method for automatic target recognition. *IEEE Transactions on Geoscience and Remote Sensing*, 58(5):3495–3508, 2019. 1
- [8] 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. 2, 3
- [9] 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. 2
- [10] Emanuele Dalsasso, Clément Rambour, Loïc Denis, and Florence Tupin. Learning a versatile representation of sar data for regression and segmentation by leveraging self-supervised despeckling with merlin. In *EUSAR 2024; 15th European Conference on Synthetic Aperture Radar*, pages 1265–1270. VDE, 2024. 2
- [11] Anindita Dasgupta and Uttam Kumar. Urban heat island and its impact on impervious surfaces during two seasons: A case study of bangalore. In *2021 IEEE International India Geoscience and Remote Sensing Symposium (InGARSS)*, pages 250–253, 2021. 1
- [12] Anindita Dasgupta and Uttam Kumar. Bangalore coming closer to it’s satellite town : Overview of urban sprawl in the city. In *2024 IEEE Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS)*, pages 549–552, 2024. 1
- [13] Anindita Dasgupta and Uttam Kumar. Interactive influence of urban heat island and urban pollution island in two major cities of india (bangalore and delhi). In *2024 International Conference on Machine Intelligence for GeoAnalytics and Remote Sensing (MIGARS)*, pages 1–3, 2024. 1
- [14] Anindita Dasgupta and Uttam Kumar. Atmospheric pollution and land surface temperature intensity in covid-19 pandemic: A case of major indian cities. In *Geospatial Science for Urban Ecosystems: Insights from India and Beyond*, pages 513–536. Springer, 2026. 1
- [15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6
- [16] Yuntao Du, Yushi Chen, Lingbo Huang, Yahu Yang, Pedram Ghamisi, and Qian Du. Summit: A sar foundation model with multiple auxiliary tasks enhanced intrinsic characteristics. *International Journal of Applied Earth Observation and Geoinformation*, 141:104624, 2025. 3, 6
- [17] Anthony Fuller, Koreen Millard, and James Green. Croma: Remote sensing representations with contrastive radar-optical masked autoencoders. *Advances in Neural Information Processing Systems*, 36:5506–5538, 2023. 2, 6
- [18] Jie Geng, Wen Jiang, and Xinyang Deng. Multi-scale deep feature learning network with bilateral filtering for sar image classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 167:201–213, 2020. 1
- [19] Gianluca Giuffrida, Luca Fanucci, Gabriele Meoni, Matej Batič, Léonie Buckley, Aubrey Dunne, Chris Van Dijk, Marco Esposito, John Hefele, Nathan Verduyssen, et al. The ϕ -sat-1 mission: The first on-board deep neural network demonstrator for satellite earth observation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021. 2
- [20] Yoav Goldberg and Omer Levy. word2vec explained: deriving mikolov et al.’s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014. 4
- [21] Meng Guo, Xiufeng Wang, Jing Li, Hongmei Wang, and Hiroshi Tani. Examining the relationships between land cover and greenhouse gas concentrations using remote-sensing data in east asia. *International journal of remote sensing*, 34(12):4281–4303, 2013. 1
- [22] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR’06)*, pages 1735–1742. IEEE, 2006. 3
- [23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6
- [24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2, 3, 5
- [25] 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. 2, 3
- [26] Muhammad Al-Amin Hoque, Stuart Phinn, Chris Roelfsema, and Iraphne Childs. Tropical cyclone disaster management using remote sensing and spatial analysis: A review. *International journal of disaster risk reduction*, 22:345–354, 2017. 1
- [27] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 6
- [28] Xiaoni Li, Yu Zhou, Yifei Zhang, Aoting Zhang, Wei Wang, Ning Jiang, Haiying Wu, and Weiping Wang. Dense semantic contrast for self-supervised visual representation learning. In *Proceedings of the 29th ACM international conference on multimedia*, pages 1368–1376, 2021. 3
- [29] Chenfang Liu, Hao Sun, Yanjie Xu, and Gangyao Kuang. Multi-source remote sensing pretraining based on contrastive self-supervised learning. *Remote Sensing*, 14(18):4632, 2022. 2
- [30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3
- [31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5
- [32] Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9414–9423, 2021. 2
- [33] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 4
- [34] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013. 4
- [35] Max Muzeau, Joana Frontera-Pons, Chengfang Ren, and Jean-Philippe Ovarlez. Safe: a sar feature extractor based on self-supervised learning and masked siamese vits. *arXiv preprint arXiv:2407.00851*, 2024. 2
- [36] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4
- [37] Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022. 2
- [38] Namuk Park, Wonjae Kim, Byeongho Heo, Taekyung Kim, and Sangdoo Yun. What do self-supervised vision transformers learn? *arXiv preprint arXiv:2305.00729*, 2023. 2, 3
- [39] Hao Pei, Mingjie Su, Gang Xu, Mengdao Xing, and Wei Hong. Self-supervised feature representation for sar image target classification using contrastive learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:9246–9258, 2023. 2
- [40] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4088–4099, 2023. 2
- [41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 7
- [42] Nur Aulia Rosni, Noorzailawati Mohd Noor, and Alias Abdullah. Managing urbanisation and urban sprawl in malaysia by using remote sensing and gis applications. *Planning Malaysia*, (4), 2016. 1
- [43] Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. Green ai. *Communications of the ACM*, 63(12):54–63, 2020. 2
- [44] Himanshi Srivastava and Uttam Kumar. Tesa-net: A novel triplet attention-enhanced skip atrous network for urban land use and land cover segmentation using sar. pages 204–209, 2025. 2
- [45] Himanshi Srivastava, Uttam Kumar, and Nihal Pattanshetty. 71 land use and land cover change (lulcc) detection in bangalore, india using multi-temporal synthetic aperture radar through a novel spatial attention enhanced-dual path-siamese neural network. In *Remote Sensing of Land Cover and Land Use Changes in South and Southeast Asia, Volume 1: Mapping and Monitoring*, pages 71–94. Routledge, 2
- [46] Himanshi Srivastava, Uttam Kumar, and Lalith Kumar Reddy. Sar-only few-shot learning for urban land use classification. In *IGARSS 2025 - 2025 IEEE International Geoscience and Remote Sensing Symposium*, pages 1970–1974, 2025. 1
- [47] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3645–3650, 2019. 2
- [48] 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. 5
- [49] Rahisha Thottolil and Uttam Kumar. Automatic building footprint extraction using random forest algorithm from high resolution google earth images: A feature-based approach. In *2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pages 1–6. IEEE, 2022. 1
- [50] Rahisha Thottolil, Uttam Kumar, and Tanujit Chakraborty. Prediction of transportation index for urban patterns in small and medium-sized indian cities using hybrid ridgegan model. *Scientific Reports*, 13(1):21863, 2023.
- [51] Rahisha Thottolil, Uttam Kumar, and Aswathi Mundayatt. Predicting urban expansion using a patch-generating land use simulation (plus) model: A case study of bangalore city,

- india. In *2023 IEEE India Geoscience and Remote Sensing Symposium (InGARSS)*, pages 1–4. IEEE, 2023.
- [52] Manjunath Bhimappa Ujjinakoppa, Uttam Kumar, Rahisha Thottolil, and Anindita Dasgupta. Multimodal and multitemporal spatial data analysis in google earth engine cloud computing platform to detect human settlements without electricity: A case study of bangalore city. In *2021 IEEE International India Geoscience and Remote Sensing Symposium (InGARSS)*, pages 238–241, 2021. 1
- [53] CJ Van Westen. Remote sensing for natural disaster management. *International archives of photogrammetry and remote sensing*, 33(B7/4; PART 7):1609–1617, 2000. 1
- [54] Chenwei Wang, Siyi Luo, Jifang Pei, Yulin Huang, Yin Zhang, and Jianyu Yang. Crucial feature capture and discrimination for limited training data sar atr. *ISPRS Journal of Photogrammetry and Remote Sensing*, 204:291–305, 2023. 1
- [55] Feng Wang, Huiyu Wang, Chen Wei, Alan Yuille, and Wei Shen. Cp 2: Copy-paste contrastive pretraining for semantic segmentation. In *European conference on computer vision*, pages 499–515. Springer, 2022. 3
- [56] Xianyuan Wang, Zongjie Cao, and Yiming Pi. Semisupervised classification with adaptive anchor graph for polsar images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021. 1
- [57] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3024–3033, 2021. 3, 4, 5
- [58] Yi Wang, Conrad M Albrecht, and Xiao Xiang Zhu. Self-supervised vision transformers for joint sar-optical representation learning. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 139–142. IEEE, 2022. 2
- [59] Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M Albrecht, and Xiao Xiang Zhu. Ssl4eos12: A large-scale multimodal, multitemporal dataset for self-supervised learning in earth observation [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 11(3):98–106, 2023. 2, 5
- [60] Yi Wang, Hugo Hernández Hernández, Conrad M Albrecht, and Xiao Xiang Zhu. Feature guided masked autoencoder for self-supervised learning in remote sensing. *arXiv preprint arXiv:2310.18653*, 2023. 5
- [61] Zaidao Wen, Zhunga Liu, Shuai Zhang, and Quan Pan. Rotation awareness based self-supervised learning for sar target recognition with limited training samples. *IEEE Transactions on Image Processing*, 30:7266–7279, 2021. 2
- [62] 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. 3
- [63] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16684–16693, 2021. 3
- [64] Zhenda Xie, Zigang Geng, Jingcheng Hu, Zheng Zhang, Han Hu, and Yue Cao. Revealing the dark secrets of masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14475–14485, 2023. 3
- [65] LIU Yi, WANG Jing, CHE Ke, CAI Zhaonan, YANG Dongxu, and WU Lin. Satellite remote sensing of greenhouse gases: Progress and trends. *National Remote Sensing Bulletin*, 25(1):53–64, 2021. 1
- [66] Zhenyu Yue, Fei Gao, Qingxu Xiong, Jun Wang, Teng Huang, Erfu Yang, and Huiyu Zhou. A novel semi-supervised convolutional neural network method for synthetic aperture radar image recognition. *Cognitive Computation*, 13(4):795–806, 2021. 1
- [67] Tianwen Zhang, Xiaoling Zhang, Chang Liu, Jun Shi, Shunjun Wei, Israr Ahmad, Xu Zhan, Yue Zhou, Dece Pan, Jianwei Li, et al. Balance learning for ship detection from synthetic aperture radar remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 182:190–207, 2021. 1
- [68] Wenbin Zhu, Hong Gu, and Xiaochun Zhu. Synthetic aperture radar image classification based on constrictive learning with limited data. *IET Radar, Sonar & Navigation*, 16(9): 1530–1537, 2022. 1