

Not every day is a sunny day: Synthetic cloud injection for deep land cover segmentation robustness evaluation across data sources

Sara Mobsite¹, Renaud Hostache¹, Laure Berti-Équille¹, Emmanuel Roux¹,
and Joris Guérin¹

¹ ESPACE-DEV, French National Research Institute for Sustainable Development
(IRD)
`{firstname.lastname}@ird.fr`

Abstract. Supervised deep learning for land cover semantic segmentation (LCS) relies on labeled satellite data. However, most existing Sentinel-2 datasets are cloud-free, which limits their usefulness in tropical regions where clouds are common. To properly evaluate the extent of this problem, we developed a cloud injection algorithm that simulates realistic cloud cover, allowing us to test how Sentinel-1 radar data can fill in the gaps caused by cloud-obstructed optical imagery. We also tackle the issue of losing spatial and/or spectral details during encoder downsampling in deep networks. To mitigate this loss, we propose a lightweight method that injects Normalized Difference Indices (NDIs) into the final decoding layers, enabling the model to retain key spatial features with minimal additional computation. Injecting NDIs enhanced land cover segmentation performance on the DFC2020 dataset, yielding improvements of 1.99% for U-Net and 2.78% for DeepLabV3 on cloud-free imagery. Under cloud-covered conditions, incorporating Sentinel-1 data led to significant performance gains across all models compared to using optical data alone, highlighting the effectiveness of radar-optical fusion in challenging atmospheric scenarios. The code for our cloud location mask generation algorithm is publicly available at: [code](#).

Keywords: land cover · segmentation · Cloud Injection · Normalized Difference Indices.

1 Introduction

High-resolution land cover segmentation (LCS) increasingly relies on deep neural networks, which facilitate precise pixel-wise classification of satellite imagery by learning from labeled datasets. LCS can be utilized in environmental studies to track deforestation [5], identify forest disturbances [20], and map the coverage of sea ice [23], among other applications. The type of satellite data selected for a given application is guided by the specific objectives and requirements of each use case. Some studies leverage Synthetic Aperture Radar (SAR) data, such as Sentinel-1 [3, 13, 16], while others utilize optical imagery from satellites such as

Sentinel-2 [1, 2]. SAR and optical data are increasingly combined in research due to their complementary properties [7, 21]: optical imagery provides rich textural and spatial details, while SAR data is almost unaffected by clouds and adverse weather conditions.

In deep learning-based LCS, SAR data provides complementary information that can help compensate for gaps in optical imagery caused by atmospheric interference, such as clouds. However, existing datasets that combine Sentinel-1 and Sentinel-2 data, such as DFC2020 [12], typically exclude cloud-covered samples and do not include a dedicated cloud class. This limitation is significant, as cloud presence is common in real-world satellite observations, especially in tropical and subtropical regions [25]. The lack of optical images containing cloud cover in training samples can hinder generalization and reduce segmentation accuracy in real-world applications. Moreover, this limits the model’s ability to learn how to effectively utilize SAR features to compensate for missing optical information during the training process.

In this work, we propose to compare the performance of state-of-the-art segmentation networks while assessing the influence of the input satellite data type (SAR, optical, or a combination of both) and the effect of clouds. To remain in a controlled environment, we propose a more realistic evaluation approach for comparing LCS networks by injecting synthetic clouds into originally cloud-free optical data during both the training and testing processes. This approach helped leverage labeled optical data and make it more suitable for real-world applications. To do so, we introduce an automated method that generates cloud masks based on real cloud patterns and insert them into Sentinel-2 imagery. Our goal is therefore to evaluate the impact of cloud cover on segmentation performance and to assess the ability of SAR features to compensate for cloud-induced gaps.

One of the challenges with autoencoder-based networks for LCS is the downsampling applied during the encoding stage. This downsampling often results in the loss of spatial details and features necessary for accurately identifying and distinguishing land cover classes at high resolution. Therefore, in this work, we explore a second approach aimed at enhancing the spatial and textural details in optical data under minimal cloud coverage. To mitigate the loss of fine features caused by downsampling in the encoder stage, we employ a decoder-level feature injection strategy using normalized difference indices (NDIs). Specifically, NDIs for water (NDWI), buildings (NDBI), and vegetation (NDVI) are computed and concatenated with decoder features at the final segmentation layer. This approach reintroduces important spatial and textural cues associated with key land cover types, thereby improving the model’s ability to distinguish between land cover classes.

2 Related Work

LCS heavily relies on spectral features that provide reflectance values across different wavelengths, as well as spatial features that capture patterns and textures

based on pixel arrangements. Sentinel-2 offers rich detail but is hindered by cloud and atmospheric interference, while Sentinel-1 is weather-resilient. Consequently, many studies explore the fusion of data from both sources using early, middle, or late-level fusion. In early-level fusion, data is combined at the input stage before encoding. In [10], the authors focused on forest identification in the Bajo Cauca Subregion of Colombia by applying early-stage fusion. They proposed a Random Forest algorithm that uses dual-polarization C-band SAR data, combined with vegetation indices extracted from multispectral optical data. In [21], a modified U-TAE model based on channel attention was proposed, using different combinations of satellite data at the input level, including spectral indices and elevation data, to improve LCS performance.

In [7], feature-level fusion was applied by independently encoding SAR and optical data using Vision Transformer encoders, followed by a second Vision Transformer encoder that performed the fusion of the extracted features. In [15], attention-based fusion at the channel level was used to learn from both SAR and optical data. The attention-based multi-modal network combined features from both modalities by learning channel-wise relationships between the outputs of the multi-scale feature extractors. In [14], a multi-stream U-Net architecture was proposed to detect forest burned areas, incorporating residual and stream-to-stream connections for feature fusion. The fusion process was based on multi-level integration of features extracted from two independent encoding branches. These feature maps were then passed into a unified decoder, where the combined outputs from each encoder stream were reintroduced through residual connections to enhance spatial feature recovery after downsampling. Rather than relying solely on feature-level fusion, the authors in [9] implemented fusion at the final stage. In their approach, each processing stream handled optical and SAR data separately, and the final pixel-wise classification was derived by combining the feature maps generated by each autoencoder. Consequently, instead of using a U-Net architecture, the study proposed a W-Net network.

Regardless of the fusion technique used, most available multi-modal datasets for supervised LCS only include cloud-free images. This does not reflect real-world conditions, especially for Sentinel-2 imagery, which is highly susceptible to cloud cover and other atmospheric disturbances. This omission presents a key gap in the literature, restricting insight into how effectively SAR features can compensate for cloud-related gaps in optical data.

3 Method

3.1 SAR and Optical Data Fusion Architecture

Our work aims to identify the most effective and reliable data sources for high-resolution LCS. We evaluate three input configurations: optical data alone, SAR data alone, and a combination of both at the input level. This setup enables a comprehensive assessment of segmentation performance for each data modality. For this analysis, we use two state-of-the-art neural network architectures for semantic segmentation: U-Net [18], and DeepLabV3 [4].

3.2 Improving Segmentation with NDIs

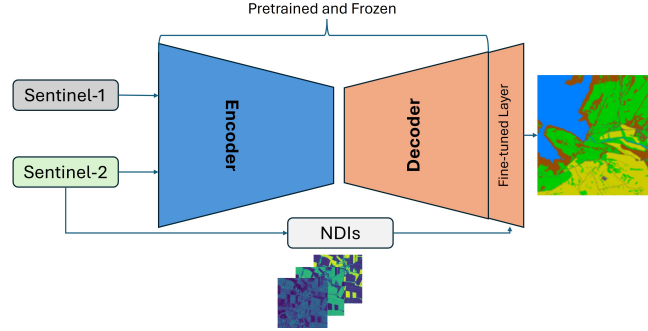


Fig. 1. NDIs injection in the decoder of a pretrained autoencoder

Sentinel-1 provides dual-polarization based data, including VV (vertical transmit and vertical receive) and VH (vertical transmit and horizontal receive) modes. Sentinel-2 data, on the other hand, provides extensive spectral coverage through its 13 bands spanning the visible, near-infrared (NIR), and shortwave-infrared (SWIR) ranges.

In segmentation networks, the encoder stage often leads to the loss of important spatial details due to repeated downsampling. To mitigate this issue, our approach incorporates normalized difference indices (NDIs), including vegetation, water, and built-up area indices. These NDIs are injected into the later stages of the decoder to enhance spatial feature recovery across the studied networks. We specifically selected these indices to be reintroduced in the final layers to reinforce information related to the primary land-cover target classes (water, built-up areas, and vegetation).

As illustrated in Fig. 1, the networks are initially trained on the DFC2020 dataset using either optical data alone or a fusion of optical and SAR data, without including the NDIs. In the second stage, transfer learning is applied: all previously trained layers are frozen, and only the final 1×1 convolutional layer is retrained. At this stage, the output feature maps from the decoder (before the final layer) are concatenated with the NDIs. This combined feature representation is then passed through a 3×3 convolutional layer followed by a final 1×1 convolution to perform LCS.

In our study, we inject three widely used NDIs in the decoding final stage: the Normalized Difference Vegetation Index (NDVI) [17], the Normalized Difference Water Index (NDWI) [8], and the Normalized Difference Built-up Index (NDBI) [24]. These indices were selected because they are specifically designed to highlight key land cover types, such as vegetation, water bodies, and built-up areas, which are essential for accurate LCS. In [22], these indices were used to train a self-supervised network based on masked autoencoders, encouraging the

model to learn important features from optical data. Instead of using raw optical data, these indices provide a concise representation of key land cover types using only three channels instead of all 13 spectral inputs. This makes the approach more efficient and helps emphasize relevant spatial features for the main land cover classes.

3.3 Synthetic Cloud Generation and Injection into Optical Data

Existing datasets based on Sentinel-2 are mostly cloud-free, which does not accurately represent real-world conditions. To address this, we developed an algorithm that synthetically injects cloud patterns into clean Sentinel-2 images. Rather than manually generating cloud-covered samples by downloading cloud data for each image from open-access sources, we create unique, cloud-like patterns by generating binary cloud location masks and embedding real-world cloud radiometric data into them. These binary cloud location masks are generated by randomly shaping ellipses based on a normal Gaussian distribution and bilinear interpolation, which mimics the irregular appearance of natural clouds (Algorithm 1). In our approach, the deformation, density, and cloud coverage percentages are uniquely defined for each mask. This helps prevent repetitive patterns in Sentinel-2 data, which can lead to overfitting during the learning process.

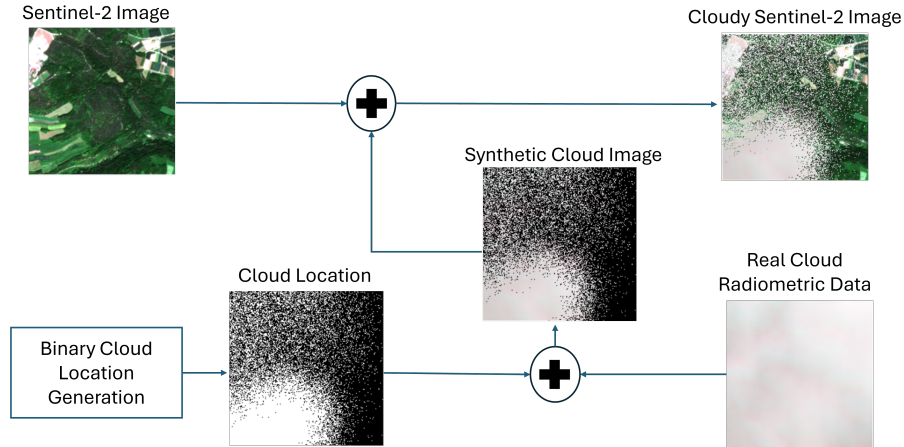


Fig. 2. Example of synthetic cloud generation based on a binary cloud location mask and cloud optical data

The cloudy areas of the mask are filled with real thick cloud radiometric values, while clear sky regions preserve the original Sentinel-2 imagery targeted for modification. Sentinel-1 was not modified, as it is not affected by cloud cover. This method allows for scalable and rather realistic injection of clouds into cloud-

free datasets, supporting the training of more robust models capable of handling diverse atmospheric conditions.

Algorithm 1 Cloud Location Mask Generation

Require: Number of masks N , mask size $S \times S$, max coverage fraction $C_{\max}$, ellipse axes bounds $E_{\min}, E_{\max}$, deformation range $D_{\max}$

- 1: **for** $i = 1$ to N **do**
- 2: Initialize empty binary mask $M \in \mathbb{R}^{S \times S}$
- 3: Calculate the number of cloud pixels $T \leftarrow S^2 \cdot \text{random } C \in [0, C_{\max}]$
- 4: Set filled area $A \leftarrow 0$
- 5: **while** $A < T$ **do**
- 6: Create blank image $B \in \mathbb{R}^{S \times S}$
- 7: Draw filled ellipse on B with:
 - Random center within bounds
 - random Axes $\in [E_{\min}, E_{\max}]$
 - random Angle $\in [0^\circ, 180^\circ]$
- 8: Generate random flow field $F \in \mathbb{R}^{S \times S \times 2}$:
 - Each component $\sim \mathcal{N}(0, 1)$
 - Multiply by random $D \in [5, D_{\max}]$
- 9: Deform B using F via bilinear remapping (reflect border) $\rightarrow B_{\text{def}}$
- 10: Threshold B_{def} to binary $\rightarrow B_{\text{bin}}$
- 11: Merge: $M \leftarrow M \vee B_{\text{bin}}$
- 12: Update $A \leftarrow \text{countNonZero}(M)$
- 13: **end while**
- 14: Save M
- 15: **end for**

Fig. 2 presents an example of a binary cloud location mask generated randomly based on Algorithm 1, alongside the resulting cloud mask after injecting radiometric values of real cloud data derived from Sentinel-2 **100% cloud coverage** observation. The cloud data is extracted from a tile captured on **May 19, 2025, at 14:27:19 UTC**, covering a $100 \text{ km} \times 100 \text{ km}$ region in northeastern Brazil, centered near **3° south latitude** and **44° west longitude**, within the state of Maranhão. In particular, this Sentinel-2 image is used in our work solely as a source of real cloud data. The region is divided into **1,500 non-overlapping tiles of 256×256 pixels**, and each tile is used to generate a unique cloud mask. To simulate realistic atmospheric interference, the generated cloud masks are individually applied to each optical image in the DFC2020 dataset, ensuring unique cloud coverage and patterns per sample.

4 Experiments

4.1 Dataset

The DFC2020 dataset, introduced for the IEEE GRSS Data Fusion Contest 2020 as an extension of the SEN12MS dataset [12, 19], contains Sentinel-1 and Sentinel-2 image pairs along with pixel-level land cover annotations at a resolution of 10 meters. The dataset covers 7 land cover classes. Following prior studies [7, 22], we used the original test split (5,128 samples) for model training (with an 85%/15% train/validation split), and the original validation set (986 samples) for final evaluation. In addition to the clean dataset, we evaluate the robustness of different approaches using a modified version with simulated cloud cover. To reflect realistic atmospheric conditions, clouds were synthetically injected into both the training and evaluation samples. Each image was overlaid with a unique, generalized cloud mask to ensure diverse and realistic cloud patterns. The deformation applied to the binary cloud location masks varied between 5 and 20 pixels, the random axes ranged from 10 to 100, and the simulated cloud coverage ranged from 0% to 70% per sample.

4.2 Implementation Details

Each of the proposed networks was based on a ResNet-50 [11] backbone as an encoder, with weights initialized from a pretrained model on the ReBEN dataset [6] for land cover classification. The models were implemented using the PyTorch library and trained on an NVIDIA RTX 4000 GPU. The input size for the Sentinel images was set to 256×256 pixels. For all experiments, cross-entropy loss was used as the objective function. Models were trained for up to 200 epochs, with early stopping applied using a patience of 20 epochs. The Adam optimizer was used with an initial learning rate of 0.001, and the batch size was set to 32. For the NDIs injection phase, the models were trained using the same configuration.

5 Results and Discussion

5.1 The Impact of NDIs Decoder-Level Injection on Model Performance

Table 1 shows the results of each network with and without NDIs injection. The networks were first trained without this operation. Then, pretrained weights were transferred, and NDIs were injected into the final decoder layer, while all earlier layers were frozen to train only the NDI-related part. This approach avoids full retraining and focuses on enhancing the decoder. Adding NDIs improved the U-Net’s mean IoU by 0.59% with optical input alone and by 2.00 with SAR+Optical input. For DeepLabV3, the gains were 5.18 and 2.78, respectively, showing a stronger benefit, especially with optical-only input. This larger impact on DeepLabV3 is likely due to its architecture, which relies on

an encoder-decoder structure with upsampling that lacks skip connections from early layers of the encoder. In contrast, U-Net incorporates skip connections from early encoding layers, which already preserve spatial details, making the addition of NDIs a more complementary enhancement rather than a critical compensatory input.

Table 1. Mean IoU of Each Model Using Different Input Types on the Cloud-Free DFC2020 Test Set, Including Performance of Decoder-Level NDIs Injection.

Model	SAR	Optical	Optical+NDIs	SAR+Optical	SAR+Optical+NDIs
U-Net	38.89	47.80	48.39 (+0.59%)	51.62	53.61 (+1.99%)
DeepLabV3	36.88	39.24	44.42 (+5.18%)	43.22	46.00 (+2.78%)

5.2 The Impact of Cloud Injection on Model Performance

To assess the impact of using SAR data alone or in combination with optical data under cloudy conditions, we evaluated the performance of each model on cloud-injected data. The training was conducted under two scenarios: using cloud-free data and using data with injected clouds. The results show that training with cloud-injected data (TCI) significantly improves performance when using optical inputs alone. For instance, DeepLabV3’s mean IoU rises from 15.82 to 43.48, and U-Net’s from 15.44 to 28.36, highlighting that without exposure to cloudy conditions during training, models struggle to handle occlusions in the test set. When using SAR data alone, the performance remains stable because SAR is cloud-penetrating and unaffected by optical obstructions. However, combining SAR with optical data only improves performance when the model is trained with cloudy data. For example, DeepLabV3 without TCI sees a drop to 14.16 with SAR+Optical (worse than using SAR alone), while training with TCI boosts it to 45.71, showing that SAR is only useful for filling gaps if the model learns to fuse SAR and optical information under cloud-obscured conditions. This suggests that SAR data does not inherently compensate for clouds unless the model is explicitly trained to associate SAR with missing optical context.

Table 2. Mean IoU of Each Optical-Based Model Trained With Cloud Injection (TCI) and Without it. Evaluated on the Cloud-Contaminated DFC2020 Test Set

Model	Optical	Optical TCI	SAR	SAR+Optical	SAR+Optical TCI
U-Net	15.44	28.36	38.89	17.44	47.72
DeepLabV3	15.82	43.48	36.88	14.16	45.71

Table 3. Mean IoU of Each Optical-Based Model Trained With Cloud Injection (TCI) and Without it. Evaluated on the Cloud-free DFC2020 Test Set

Model	Optical	Optical TCI	SAR	SAR+Optical	SAR+Optical TCI
U-Net	47.8	35.42	38.89	51.62	51.26
DeepLabV3	39.24	45.64	36.88	43.22	45.66

Conclusion

This study demonstrates the importance of training with cloud-injected data to enable models to effectively handle cloud occlusion in remote sensing imagery. SAR data alone offers robust performance under cloudy conditions, but its full potential is only realized when combined with optical data in models specifically trained to fuse both modalities under such conditions. Even optical-only models benefit from cloud-injected training by learning to infer missing data from nearby clear regions. Decoder-level NDI injection further enhances spatial detail recovery, particularly in DeepLabV3, by facilitating the reconstruction of features lost during downsampling. Future work could explore cloud synthesis using generative deep learning networks and assess performance based on cloud coverage percentage in SAR-optical fusion tasks.

Acknowledgement

This work has received funding from the European Union’s Horizon Europe Research and Innovation program under Grant Agreement N° 101137398

References

1. Giuseppina Andresini, Annalisa Appice, and Donato Malerba. Leveraging sentinel-2 time series for bark beetle-induced forest dieback inventory. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, pages 875–882, 2024.
2. Preetpal Kaur Buttar and Manoj Kumar Sachan. Generating land cover maps in semi-arid regions based on a 3d semantic segmentation architecture using multi-temporal sentinel-2 satellite images: a case study of ludhiana district in punjab, india. *Journal of the Indian Society of Remote Sensing*, 52(2):383–398, 2024.
3. Thibault Catry, Zhichao Li, Emmanuel Roux, Vincent Herbreteau, Helen Gurgel, Morgan Mangeas, Frédérique Seyler, and Nadine Dessay. Wetlands and malaria in the amazon: Guidelines for the use of synthetic aperture radar remote-sensing. *International Journal of Environmental Research and Public Health*, 15(3):468, 2018.
4. Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Re-thinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.
5. Yeonju Choi, Dongoo Lee, and SungTae Moon. Deforestation segmentation approach based on time of event occurrence using multi-temporal satellite data. *IEEE Sensors Letters*, 2024.

6. Kai Norman Clasen, Leonard Hackel, Tom Burgert, Gencer Sumbul, Begum Demir, and Volker Markl. reben: Refined bigearthnet dataset for remote sensing image analysis. *arXiv preprint arXiv:2407.03653*, 2024.
7. 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.
8. Bo-Cai Gao. Ndwī—a normalized difference water index for remote sensing of vegetation liquid water from space. *Remote sensing of environment*, 58(3):257–266, 1996.
9. Massimiliano Gargiulo, Domenico AG Dell’Aglío, Antonio Iodice, Daniele Riccio, and Giuseppe Ruello. Integration of sentinel-1 and sentinel-2 data for land cover mapping using w-net. *Sensors*, 20(10):2969, 2020.
10. Ana Guisao-Betancur, Luis Gómez Déniz, and Alejandro Marulanda-Tobón. Forest/nonforest segmentation using sentinel-1 and-2 data fusion in the bajo cauca subregion in colombia. *Remote Sensing*, 16(1):5, 2023.
11. 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.
12. 2020 ieee grss data fusion contest. <http://www.grss-ieee.org/community/technical-committees/data-fusion>, 2020. Accessed: 2025-05-16.
13. Ali Jamali, Swalpa Kumar Roy, Leila Hashemi Beni, Biswajeet Pradhan, Jonathan Li, and Pedram Ghamisi. Residual wave vision u-net for flood mapping using dual polarization sentinel-1 sar imagery. *International Journal of Applied Earth Observation and Geoinformation*, 127:103662, 2024.
14. Ehsan Khankeshizadeh, Sahand Tahermanesh, Amin Mohsenifar, Armin Moghimi, and Ali Mohammadzadeh. Fba-dpattresu-net: Forest burned area detection using a novel end-to-end dual-path attention residual-based u-net from post-fire sentinel-1 and sentinel-2 images. *Ecological Indicators*, 167:112589, 2024.
15. Wanli Ma, Oktay Karakuş, and Paul L Rosin. Amm-fusenet: Attention-based multi-modal image fusion network for land cover mapping. *Remote Sensing*, 14(18):4458, 2022.
16. Fernando Pech-May, Raúl Aquino-Santos, Omar Álvarez-Cárdenas, Jorge Lozoya Arandia, and German Rios-Toledo. Segmentation and visualization of flooded areas through sentinel-1 images and u-net. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.
17. Nathalie Pettorelli. *The normalized difference vegetation index*. Oxford University Press, USA, 2013.
18. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.
19. Michael Schmitt, Lloyd Haydn Hughes, Chunping Qiu, and Xiao Xiang Zhu. Sen12ms—a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. *arXiv preprint arXiv:1906.07789*, 2019.
20. Yuping Tian, Feng Zhao, Ran Meng, Rui Sun, Yuan Zhang, Yanyan Shen, Bin Wang, Jie Liu, and Mingze Li. A vision foundation model-based method for large-scale forest disturbance mapping using time series sentinel-1 sar data. *Remote Sensing of Environment*, 325:114775, 2025.

21. Anastasios Tzepkenlis, Konstantinos Marthoglou, and Nikos Grammalidis. Efficient deep semantic segmentation for land cover classification using sentinel imagery. *Remote Sensing*, 15(8):2027, 2023.
22. Yi Wang, Hugo Hernández Hernández, Conrad M Albrecht, and Xiao Xiang Zhu. Feature guided masked autoencoder for self-supervised learning in remote sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.
23. Zuomin Wang, Ying Li, Jiazhu Wang, and Bingxin Liu. Lh-unet: A lighted histoformer-encoded u-net for sea ice recognition with high-resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 2025.
24. Yong Zha, Jay Gao, and Shaoxiang Ni. Use of normalized difference built-up index in automatically mapping urban areas from tm imagery. *International journal of remote sensing*, 24(3):583–594, 2003.
25. Shuyan Zhang, Yong Ma, Fu Chen, Erping Shang, Wutao Yao, Yubao Qiu, and Jianbo Liu. Global land high-resolution cloud climatology based on an improved mod09 cloud mask. *Remote Sensing*, 13(19):3997, 2021.