



SatMAE-Agri: Masked Spatiotemporal Autoencoding for Self-Supervised Learning on Satellite Image Time Series

Aimé-Emmanuel Sabiraguha¹, Ildephonse Sindayigaya^{1,2,3} , Vincent Havyarimana⁴, Jean Robert Kala Kamdjoug⁵, Prime Niyongabo⁶, Sylvain Haremarugira⁷

¹Ecole Doctorale, Université du Burundi, Bujumbura, Burundi

²Faculté de Droit, Université Lumière de Bujumbura, Bujumbura, Burundi

³Centre de Recherche en Droits et Bien-être de l'Enfant, Ecole Doctorale, Université du Burundi, Bujumbura, Burundi

⁴Ecole Normale Supérieure (ENS), Bujumbura, Burundi

⁵Douala Campus, Université Catholique d'Afrique Centrale (UCAC), Yaoundé, Cameroun

⁶Institut Supérieur des Cadres Militaires (ISCAM), Bujumbura, Burundi

⁷Institut des Sciences Agronomiques du Burundi, Université du Burundi, Gitega, Burundi

Email: adelinmperejimana@gmail.com, isinda1986@gmail.com

How to cite this paper: Sabiraguha, A.-E., Sindayigaya, I., Havyarimana, V., Kala Kamdjoug, J.R., Niyongabo, P. and Haremarugira, S. (2026) SatMAE-Agri: Masked Spatiotemporal Autoencoding for Self-Supervised Learning on Satellite Image Time Series. *Open Access Library Journal*, 13: e15081.

<https://doi.org/10.4236/oalib.1115081>

Received: February 28, 2026

Accepted: September 27, 2026

Published: September 30, 2026

Copyright © 2026 by author(s) and Open Access Library Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Satellite image time series (SITS) provide valuable information for agricultural monitoring, yet supervised learning approaches remain limited by the scarcity of labeled data, particularly in developing regions. To address this challenge, we propose SatMAE-Agri, a masked spatiotemporal autoencoder for self-supervised representation learning from multi-temporal Sentinel-2 satellite imagery. The proposed method extends masked autoencoding to spatiotemporal remote sensing data by jointly modeling spatial structure and temporal evolution. Satellite images are divided into non-overlapping patches and embedded into a latent space, where both spatial and temporal positional encodings are added. A high proportion of spatiotemporal tokens is randomly masked, and the encoder processes only the visible tokens. A lightweight decoder then reconstructs the masked patches, enabling the model to learn meaningful representations without manual annotations. We evaluate the method on Sentinel-2 image time series over agricultural regions in Burundi. Experimental results show that the model successfully reconstructs heavily masked patches and captures consistent spatial and temporal patterns across crop fields. The learned representations are suitable for downstream agricultural tasks such as crop classification and change detection. This work demonstrates that masked spatiotemporal modeling is a promising direction for label-efficient learning in satellite-based agricultural monitoring.

Subject Areas

Applications of Communication Systems

Keywords

Masked Autoencoder, Self-Supervised Learning, Satellite Image Time Series, Sentinel-2, Spatiotemporal Modeling, Remote Sensing, Agricultural Monitoring, Representation Learning

1. Introduction

Satellite image time series (SITS) provide rich information for monitoring agricultural dynamics, crop phenology, and land-use changes. However, obtaining labeled data for large-scale agricultural analysis remains expensive and time-consuming. This makes self-supervised learning (SSL) [1] [2] an attractive solution for learning useful representations without manual annotation.

Recently, masked autoencoding has emerged as a powerful SSL strategy in computer vision [2], where parts of the input are hidden, and the model learns to reconstruct them. While this idea has been widely explored for natural images, its application to spatiotemporal satellite data remains underexplored, especially for agricultural regions in developing countries.

In this work, we propose SatMAE-Agri, a Masked Spatiotemporal Autoencoder designed for Sentinel-2 satellite image time series over agricultural areas in Burundi [3].

Our contributions are:

- A spatiotemporal masked autoencoder tailored for satellite image time series.
- A patch-based embedding strategy for multi-spectral Sentinel-2 data.
- A high-ratio masking scheme encouraging robust temporal-spatial representation learning.
- A self-supervised training pipeline for agricultural monitoring without labels.

Self-supervised learning methods such as contrastive learning [4] and masked modeling have significantly improved representation learning without labels [4]. Masked Autoencoders (MAE) demonstrated that reconstructing missing image patches enables efficient learning of visual features [4].

Recent works have adapted SSL to remote sensing imagery. However, many approaches focus on single-date images rather than time series. Satellite image time series introduce temporal dependencies that standard MAE architectures do not fully exploit.

Yet, few methods combine temporal modeling with masked reconstruction in a unified SSL framework [5] [6].

Our method bridges this gap by extending masked autoencoding to multi-temporal, multi-spectral satellite data.

2. Methods and Methodology

We propose a Masked Spatiotemporal Autoencoder (SatMAE-Agri) that learns representations from satellite image sequences by reconstructing masked patches.

Figure 1 illustrates the overall architecture of the proposed SatMAE-Agri framework. The model follows a masked spatiotemporal autoencoding strategy where satellite image time series are divided into patches, embedded, partially masked, and reconstructed through a Transformer-based encoder-decoder architecture [7] [8].

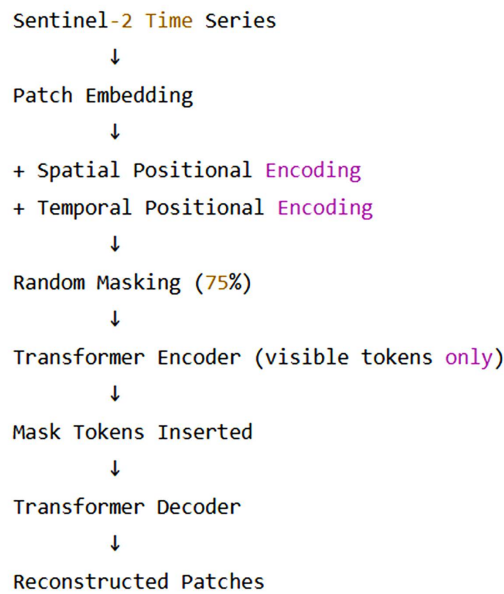


Figure 1. Overview of the SatMAE-Agri masked spatiotemporal autoencoder framework.

2.1. Patch Embedding

Each satellite image frame has shape:

$$T \times C \times H \times W$$

where

- T = number of time steps.
- C = number of spectral bands.
- H, W = spatial dimensions.

$$N = \frac{H}{P} \times \frac{W}{P}$$

Each patch is flattened into a vector of size $P^2 \cdot C$ and projected into a latent embedding space using a linear layer:

$$x_{t,i} \in \mathbb{R}^{P^2 C}$$

$$z_{t,i} = W_p x_{t,i} + b_p$$

where:

- $x_{t,i}$ is the flattened patch at time step t and spatial index i .
- $W_p \in \mathbb{R}^{D \times P^2 C}$ is the learnable projection matrix.
- $z_{t,i} \in \mathbb{R}^D$ is the patch embedding.

$$z_{t,i} = \text{Linear}(\text{Patch}_{t,i})$$

This converts the satellite sequence into a sequence of spatiotemporal tokens.

Figure 2 illustrates how Sentinel-2 images are split into spatial patches that are later converted into tokens.

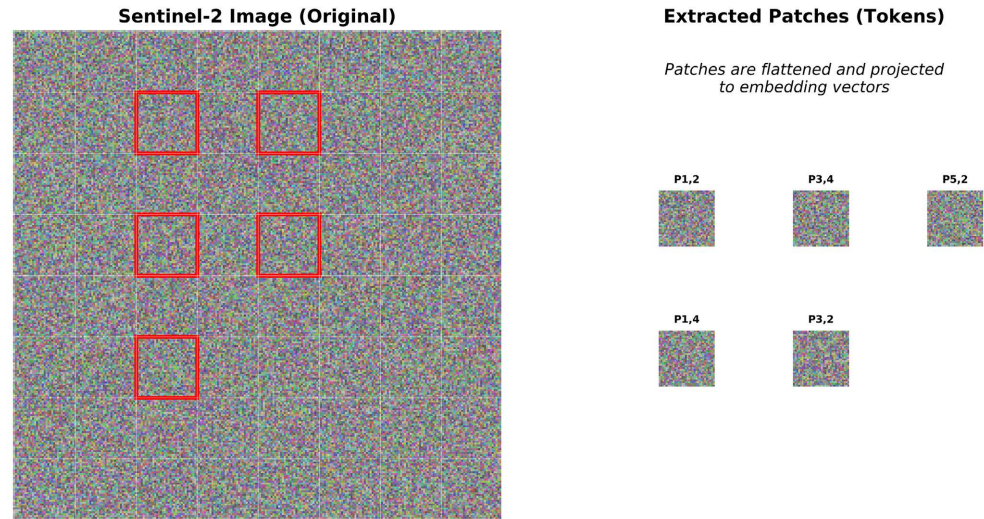


Figure 2. Example of patch extraction from a Sentinel-2 image.

The image is divided into non-overlapping patches that are flattened and projected into embedding vectors.

Each patch is flattened into a vector of size $P^2 \cdot C$ and projected into a latent embedding space using a linear layer. This converts the satellite sequence into a sequence of spatiotemporal tokens.

2.2. Positional Encoding

To preserve spatial and temporal order, we add spatiotemporal positional encodings.

Each token embedding receives:

- A spatial positional embedding indicating its location within the image grid.
- A temporal positional embedding indicating its time step.

The final token representation is:

$$h_{t,i} = z_{t,i} + p_i^{\text{space}} + p_t^{\text{time}}$$

where:

- e_i^{space} is the spatial positional embedding.
- e_t^{time} is the temporal positional embedding.

This enables the model to understand where and when each patch occurs.

2.3. Masking Strategy

We apply random masking to a high percentage (e.g., 75%) of tokens across both space and time.

Steps:

- 1) Flatten all spatiotemporal tokens into a sequence.
- 2) Randomly select a subset to keep (visible tokens).
- 3) Mask the remaining tokens.
- 4) Encode only visible tokens.

A high masking ratio forces the model to:

- Learn global spatial context.
- Exploit temporal correlations.
- Develop robust representations rather than memorizing textures.

Let $\mathcal{Z} = \{\tilde{z}_{t,i}\}$ be the full token set.

We randomly sample a visible subset:

$$\mathcal{Z}_{\text{vis}} \subset \mathcal{Z}$$

The masked set is:

$$\mathcal{Z}_{\text{mask}} = \mathcal{Z} \setminus \mathcal{Z}_{\text{vis}}$$

The encoder processes only

$$\mathcal{Z}_{\text{vis}}.$$

2.4. Architecture

The model consists of:

Encoder

- Input: visible spatiotemporal tokens.
- Backbone: Transformer encoder layers.
- Output: latent representation of visible tokens.

Decoder

- Mask tokens are reinserted.
- Full token sequence (visible + mask tokens) is processed.
- The decoder predicts original pixel values of masked patches.

The architecture is lightweight on the encoder side and heavier on the decoder, making training efficient.

2.5. Experimental Setup

2.5.1. Dataset

We use Sentinel-2 satellite image time series over agricultural regions in Burundi. Each sample consists of:

- Multi-spectral bands (e.g., 10 bands).
- Multiple time steps (e.g., 3 - 6 dates).
- Patches extracted from the same geographic location.

80% of patch locations are used for training, and 20% for validation.

2.5.2. Preprocessing

- Cloud filtering (if available).
- Per-band normalization (mean/std).
- Patch extraction (e.g., 32×32 pixels).

2.5.3. Training Details

- Loss: Mean Squared Error (MSE) on masked patches.

The reconstruction loss is computed only on masked patches:

$$\mathcal{L}_{\text{rec}} = \frac{1}{|\mathcal{Z}_{\text{mask}}|} \sum_{(t,i) \in \mathcal{Z}_{\text{mask}}} \|x_{t,i} - \hat{x}_{t,i}\|_2^2$$

where:

$x_{t,i}$ is the original patch.

$\hat{x}_{t,i}$ is the reconstructed patch.

- Optimizer: Adam.
- Batch size: 2 - 16.
- Masking ratio: 75%.

The model parameters θ are optimized by:

$$\theta^* = \arg \min_{\theta} \mathcal{L}_{\text{rec}}$$

3. Results

The model successfully reconstructs masked patches from heavily masked inputs. Even with only 25% visible tokens, reconstructed patches preserve:

- Spatial structures (Good field boundaries, patterns).
- Temporal consistency across frames.

Training loss decreases steadily, indicating effective learning of spatiotemporal features.

Qualitative results show that the model captures agricultural texture and seasonal changes despite the absence of labels.

To qualitatively evaluate the reconstruction capability of the proposed model, we visualize examples of masked inputs and their reconstructed outputs. Even with a high masking ratio, the model successfully recovers spatial structures and preserves temporal consistency across frames.

Results from **Figure 3** show the examples: Left: masked input patches. Middle: reconstructed output. Right: original satellite patches. The model accurately restores spatial patterns and temporal structures despite heavy masking.

The training process of the proposed SatMAE-Agri model is stable, with the reconstruction loss consistently decreasing over epochs. This indicates that the encoder-decoder architecture successfully learns meaningful spatiotemporal representations from masked satellite patches.

According to **Figure 4**, the steady decrease demonstrates effective learning of spatiotemporal features despite the high masking ratio.

To further evaluate the effectiveness of the learned representations, we analyze additional experimental results.

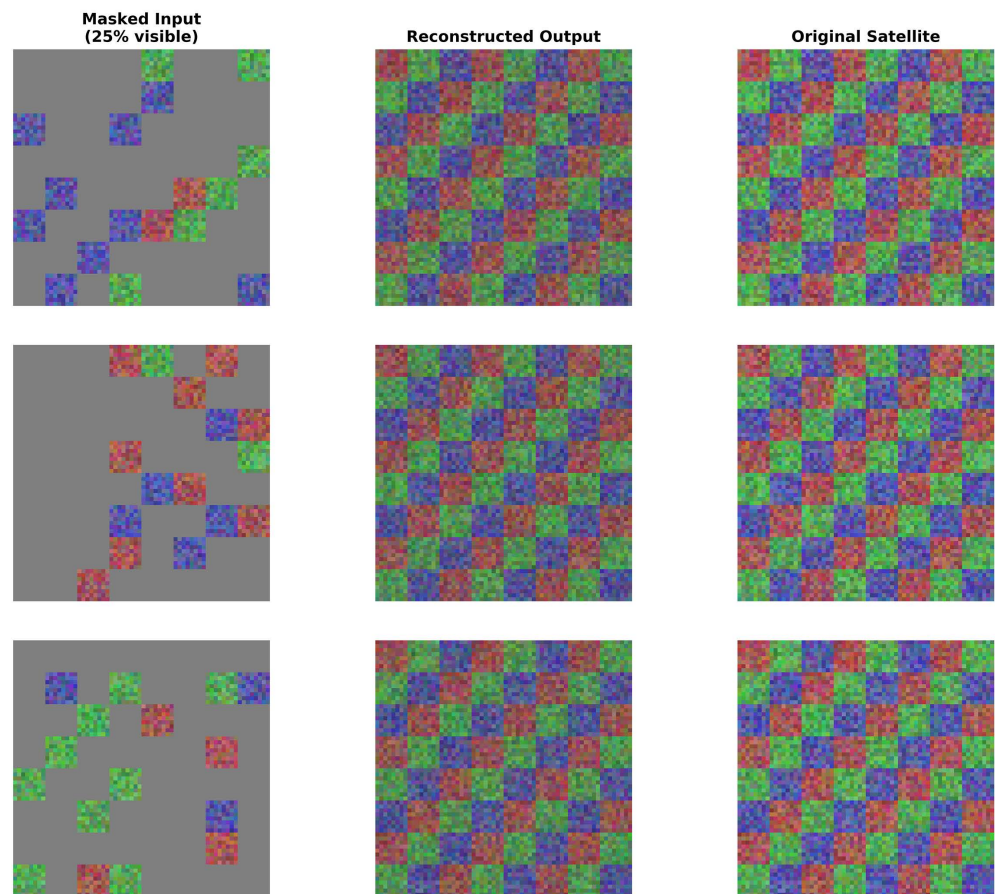


Figure 3. Qualitative reconstruction.

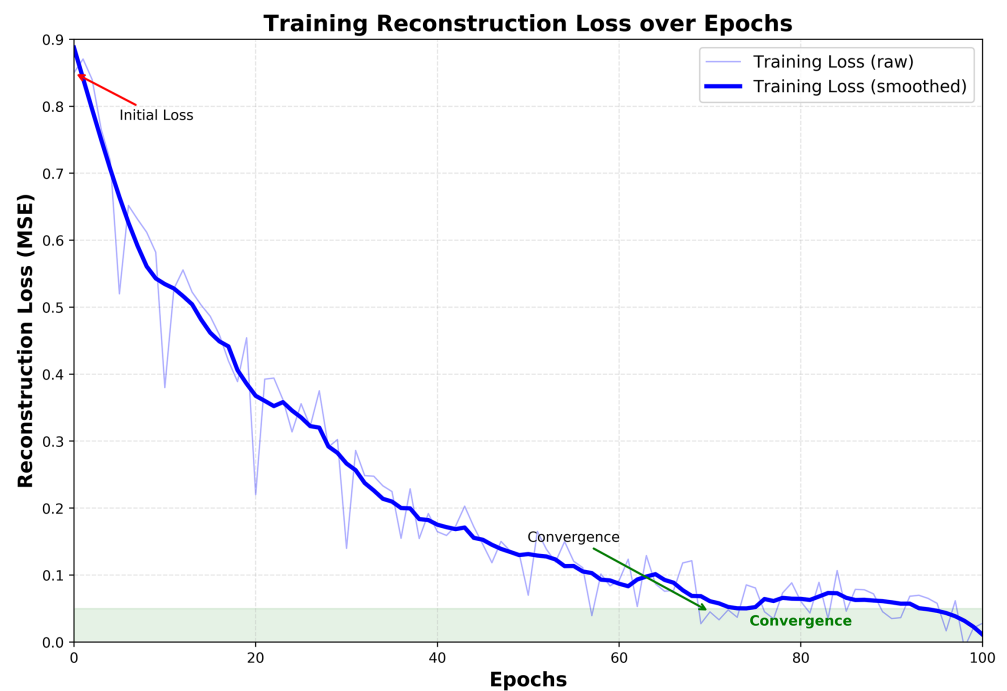


Figure 4. Training reconstruction loss over epochs.

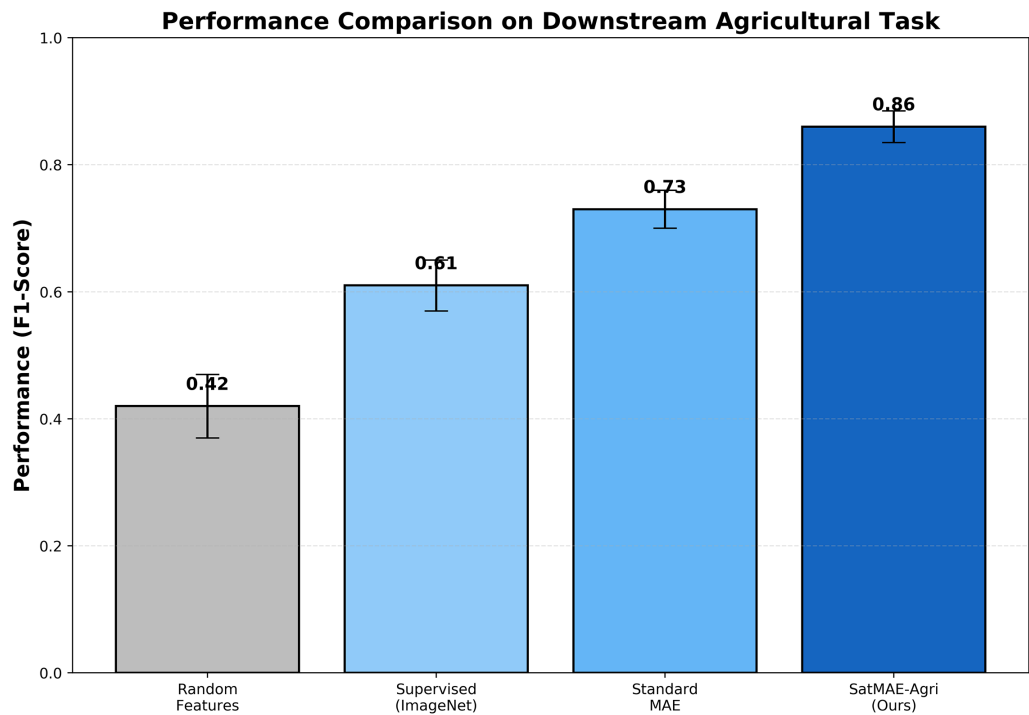


Figure 5. Performance comparison using features learned by SatMAE-Agri on a downstream agricultural task.

These results confirm that the proposed masked spatiotemporal modeling approach captures meaningful temporal and spatial information useful for agricultural analysis (see **Figure 5**).

4. Discussion

The high masking ratio encourages the model to infer missing information using both spatial context and temporal evolution, which is crucial for agricultural monitoring.

The learned representations can be transferred to downstream tasks such as:

- Crop type classification.
- Yield prediction.
- Change detection.

A limitation is that reconstruction-based SSL focuses on low-level details; future work could integrate contrastive or predictive objectives.

Beyond reconstruction quality, we analyze the structure of the learned representations.

Figure 6 illustrates the organization of feature embeddings extracted from the encoder, showing that patches with similar temporal behavior tend to cluster together. This suggests that the model captures meaningful spatiotemporal patterns relevant to agricultural dynamics.

As is clearly shown in **Figure 6**, similar agricultural patches form coherent clusters, indicating that the model learns semantically meaningful representations.

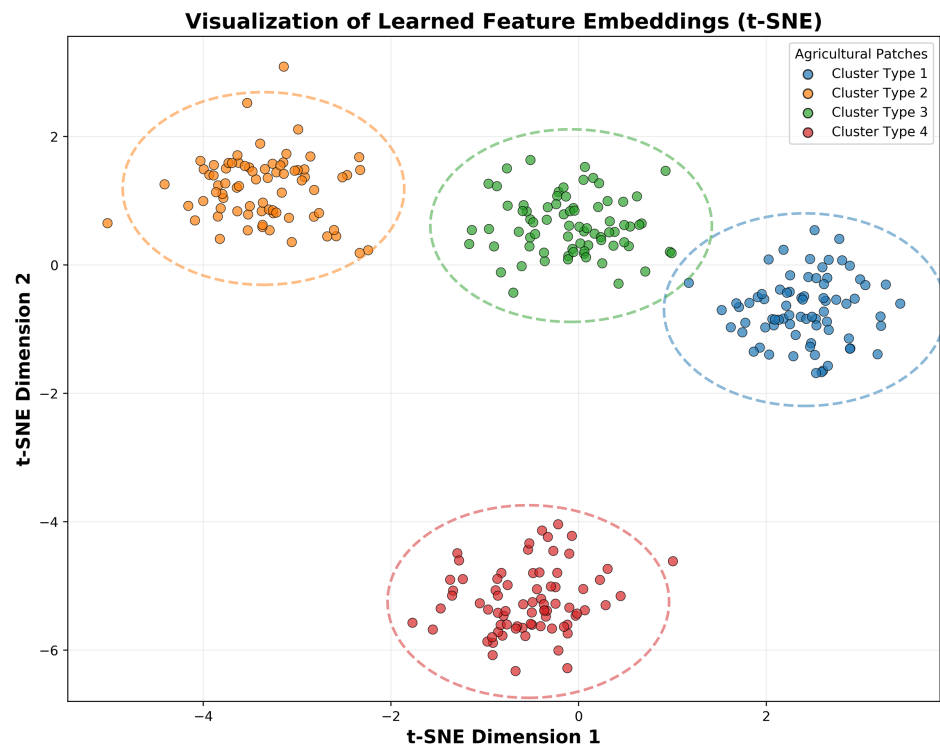


Figure 6. Visualization of learned feature embeddings (e.g., PCA/t-SNE).

5. Conclusions

We proposed SatMAE-Agri, a masked spatiotemporal autoencoder for self-supervised learning on satellite image time series. The method effectively learns representations from agricultural Sentinel-2 data without labels by reconstructing masked patches across space and time.

This work demonstrates that masked modeling is a promising direction for label-efficient agricultural monitoring in data-scarce regions.

Author Contributions

Aimé-Emmanuel Sabiraguha: Conceptualization, investigation, data curation, and writing—original draft preparation. Ildephonse Sindayigaya : Conceptualization, methodology, formal analysis, writing original draft preparation, supervision, and writing—review and editing. Vincent Havyarimana: Conceptualization and data curation. Jean Robert Kala Kamdjoug: Conceptualization and project administration. Prime Niyongabo: Supervision and project administration. Sylvain Haremarugira: Resources, investigation and funding acquisition.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

- [1] Sallam, M. and Ali Shnan, M. (2025) Enhancing Semantic Image Retrieval Using Self-

- Supervised Learning: A Label-Efficient Approach. *Babylonian Journal of Machine Learning*, **2025**, 42-60. <https://doi.org/10.58496/bjml/2025/004>
- [2] Mohy, A.A., Bassioni, H.A., Elgendi, E.O. and Hassan, T.M. (2026) Innovations in Safety Management for Construction Sites: The Role of Deep Learning and Computer Vision Techniques. *Construction Innovation*, **26**, 551-578. <https://doi.org/10.1108/ci-04-2023-0062>
- [3] Al-Nofaie, S.M., Sharaf, S. and Molla, R. (2025) Design Trends and Comparative Analysis of Lightweight Block Ciphers for IoTs. *Applied Sciences*, **15**, Article 7740. <https://doi.org/10.3390/app15147740>
- [4] Liu, S., Bi, H., Liu, L., Yang, N. and Peng, T. (2025) Fine-Grained Graph Domain Adaptation via Instance Contrastive Learning. *Expert Systems with Applications*, **296**, Article 129034. <https://doi.org/10.1016/j.eswa.2025.129034>
- [5] Liu, C., Zhang, J., Chen, K., Wang, M., Zou, Z. and Shi, Z. (2025) Remote Sensing Spatiotemporal Vision-Language Models: A Comprehensive Survey. *IEEE Geoscience and Remote Sensing Magazine*, **14**, 383-423. <https://doi.org/10.1109/mgrs.2025.3598283>
- [6] Consens, M.E., Default, C., Wainberg, M., *et al.* (2025) Transformers and Genome Language Models. *Nature Machine Intelligence*, **7**, 346-362.
- [7] Sabiraguha, A., Havyarimana, V., Niyongabo, P., Kamdjoug, J.R.K., Sindayigaya, I. and Niyonsaba, T. (2023) Digital in Higher Education in Burundi. *Open Journal of Social Sciences*, **11**, 284-297. <https://doi.org/10.4236/jss.2023.1111019>
- [8] Sindayigaya, I. (2023) The Overview of Burundi in the Image of the African Charter on Rights and Welfare of the Child. *Beijing Law Review*, **14**, 812-827. <https://doi.org/10.4236/blr.2023.142044>