EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

FreqEdgeViT: A Scalable and Efficient Reliability-Aware Transformer for Large-Scale Agricultural Information Systems

Li Qu[1], Le Sun[1,*], Yimin Yu[2], Hemant Ghayvat[3]

[1]Department of Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET), Nanjing University of Information Science and Technology, Nanjing 210044, China
[2]School of Information, Yunnan University of Finance and Economics, Kunming 650221, China
[3]Linnaeus University, Växjö 35195, Sweden

Abstract

With the exponential growth of data in modern agriculture, satellite image time series (SITS) has become an important data source for scalable information systems that analyze global crop distribution. However, processing these massive, high-dimensional data streams poses significant challenges; existing semantic segmentation models suffer from prohibitive computational overhead and lack scalability. Furthermore, they are vulnerable to high-frequency non-phenological perturbations and mixed-pixel boundary ambiguity, which degrades reliability in agricultural Internet of Things (IoT) applications. In this work, we propose FreqEdgeViT, an efficient, reliability-aware, and boundary-guided Vision Transformer. Designed for scalable SITS processing, FreqEdgeViT integrates a factorized spatiotemporal architecture with two novel mechanisms. First, in the temporal domain, we introduce a Phenology-Aware Frequency Filter (PAFF) combined with Reliability-Aware Token Merging (Ra-ToMe). This combination utilizes spectral analysis to filter environmental noise and dynamically prunes temporal redundancy based on signal reliability, significantly reducing data throughput requirements. Second, in the spatial domain, we propose a Boundary-Guided Spatial Encoder (BGSE) that enforces explicit geometric constraints to resolve edge blurring in mixed pixels. Experimental results on two public SITS datasets demonstrate that FreqEdgeViT achieves state-of-the-art accuracy with significantly reduced computational costs. The proposed architecture offers a scalable solution for processing large-scale agricultural data, providing precise support for crop policy formulation and enhancing the value of agricultural information systems.

Received on 08 February 2026; accepted on 24 March 2026; published on 31 March 2026

Keywords: Vision Transformer, Satellite image time series, Crop type semantic segmentation, Scalable Data Processing

doi:10.4108/eetsis.11866

1. Introduction

The rapid adoption of the Internet of Things and remote sensing technologies has led to a surging growth in agricultural data volume for the sector's scalable information systems[1–3]. Modern scalable agricultural systems increasingly require processing massive data streams to support scientific decision-making[4–7]. Within this data-intensive ecosystem, accurate crop type distribution serves as a critical information layer, bridging macro-scale governmental policymaking [8] with field-level production optimization [9]. However, efficiently processing such large-scale data poses a significant challenge for scalable information systems [10–12].

Traditional methods for acquiring crop information, such as field surveys, are labor-intensive and clearly unscalable for vast geographical regions. While ground-based IoT sensors provide local precision, they suffer from limited coverage and high deployment costs in remote areas with weak infrastructure [13–15]. Consequently, Satellite Image Time Series (SITS) has

*Corresponding author. Email: lesun1@nuist.edu.cn

emerged as a dominant data source for scalable agricultural monitoring. SITS offers distinct advantages for information systems: it provides high-frequency multispectral data capable of capturing dynamic growth patterns over thousands of square kilometers [16, 17].

Driven by the availability of this massive data, deep learning-based crop type segmentation has become a focal point in research [18, 19]. Similar scalable deep learning architectures have also demonstrated robustness in other data-intensive fields, providing valuable reference for agricultural monitoring[20–23]. For instance, advanced neural networks have shown exceptional feature extraction capabilities in medical diagnostics and social behavior analysis [24–27]. Existing approaches generally fall into two architectures: CNN-based methods [28, 29] and Transformer-based methods [30, 31]. Despite their accuracy, deploying these models in scalable systems faces critical hurdles. High-accuracy Transformers, such as TSViT [30], incur considerable computational costs and memory footprints. This complexity creates a bottleneck for large-scale monitoring applications, where processing massive streams of high-dimensional SITS data efficiently is paramount. Furthermore, SITS data inherently contains temporal redundancy; consecutive observations often convey repetitive information[32]. Most existing architectures fail to effectively prune this redundancy, leading to inefficient resource utilization that hinders system scalability [2].

To improve computational efficiency, the general computer vision community has explored lightweight strategies such as Token Merging (ToMe) [33] to reduce data dimensionality. However, directly transplanting these general-purpose strategies to SITS is non-trivial; it exposes the system to two domain-specific challenges that can severely compromise the reliability of the extracted agricultural information:

1. Sensitivity to High-Frequency Noise: SITS data is frequently contaminated by non-phenological noise, such as atmospheric scattering and cloud shadows [34]. These artifacts introduce high-frequency fluctuations. Standard compression methods based on simple similarity metrics may inadvertently discard critical phenological information or preserve noise, compromising data integrity.

2. Mixed-Pixel Boundary Ambiguity: Field boundaries in remote sensing are naturally plagued by mixed pixels due to limited sensor spatial resolution. The patch-based processing inherent in standard Transformers further aggravates this ambiguity by dropping local fine-grained spatial details, making precise boundary delineation highly challenging [35].

We introduce FreqEdgeViT to solve these issues. Unlike standard models that trade accuracy for speed, FreqEdgeViT integrates physics-informed and geometric-constrained mechanisms to ensure both high efficiency and precision. We first address temporal redundancy and noise by introducing a Phenology-Aware Frequency Filter (PAFF). Based on the signal processing principle that phenological trends manifest as low-frequency components, PAFF adapts frequency-domain filtering [36] to purify features. Building on this, we propose a Reliability-Aware Token Merging (Ra-ToMe) strategy. This mechanism merges tokens based on signal reliability rather than simple similarity, effectively compressing data volume while preserving information quality. Spatially, to resolve boundary ambiguity, we design a Boundary-Guided Spatial Encoder (BGSE), which incorporates an explicit edge branch to enforce geometric constraints.

Our contributions to scalable agricultural information processing are as follows:

1. We propose a physics-informed Phenology-Aware Frequency Filter and a Reliability-Aware Token Merging mechanism. By distinguishing low-frequency phenological trends from high-frequency perturbations, we achieve aggressive token reduction, which directly leads to a significant reduction in FLOPs while maintaining robust feature representation.

2. We introduce a Boundary-Guided Spatial Encoder (BGSE) with explicit geometric constraints. By integrating a dual-stream interaction mechanism, BGSE effectively mitigates the mixed-pixel problem, effectively improving the precision of information extraction compared to standard Transformers.

3. We conduct extensive experiments on two public SITS datasets. FreqEdgeViT achieves state-of-the-art performance with low computational cost, demonstrating a promising balance between efficiency and accuracy suitable for large-scale smart agriculture applications.

The remainder of this paper is organized as follows: Section 2 reviews Related Work; Section 3 details the proposed architecture; Section 4 presents experimental results; Section 5 discusses implications; and Section 6 concludes the paper.

2. Related Work

2.1. Crop type recognition

Crop type recognition aims to assign each pixel in a Satellite Image Time Series (SITS) to one of K crop classes based on geographic location. The key

to successfully achieving this goal lies in modeling the temporal patterns of crop growth [37]. Recent deep neural networks trained on raw optical data have outperformed traditional classifiers [38]. In the task of semantic segmentation for temporal images, Rustowicz et al. [28] investigated the application of CNN architectures and found that the model combining a UNET2D feature extractor with a subsequent CLSTM achieves performance comparable to that of the fully convolutional UNET3D model. U-TAE [29] first applies a UNET2D encoder to all images in parallel, then compresses the temporal feature dimension via a temporal attention mechanism. These features, which contain only spatial information, are ultimately processed by a UNET2D decoder to obtain dense prediction results.

2.2. Self-attention in vision

Inspired by the Vision Transformer (ViT) [39], numerous studies have developed attention-based architectures for visual tasks. In satellite imaging, Kaselimi et al. [40] found that ViT-based classifiers outperform CNN architectures, particularly under class imbalance. Tarasiou et al. [30] introduced temporal positional encodings aligned with acquisition times to enhance SITS modeling. Nguyen et al. [41] proposed improvements to the ViT pipeline, variable tokenization and token aggregation, to handle heterogeneous climate and weather data sources.

2.3. Scalable Information Systems in Agriculture

The increasing volume of data in modern agriculture necessitates computational architectures capable of efficient processing. Knapen et al. [42] investigated the scalability of crop growth models, specifically benchmarking the WISS-WOFOST simulation model within a big data framework. Their work demonstrated that leveraging distributed computing technologies significantly reduces the runtime of complex simulations for large-scale yield forecasting. In the context of system integration, Sarkar et al. [43] proposed the Cyber-Agricultural Systems (CAS) framework, which aligns scalable cyberinfrastructure with advanced artificial intelligence models. They emphasized that scalable computing backbones are essential for supporting complex deep learning algorithms in crop breeding and production tasks. Furthermore, to address the fragmentation of agricultural data, López-Morales et al. [44] developed a unified data model within an interoperable platform. This approach ensures that predictive models can generate scalable predictions for irrigation management by standardizing heterogeneous data streams throughout the farm lifecycle.

2.4. Efficient Transformers

Several works have attempted to create more efficient transformers in both NLP and Vision. Choromanski et al. [45] and Bolya et al. [46] focus on faster attention. Meng et al. [47] prune attention heads or feature dimensions to cut costs. Additionally, SwinV2 [48] and CSWin [49] have focused on incorporating domain-specific modules into the model. Recent advances also explore frequency domain learning [36] to improve feature robustness. Furthermore, explicit boundary modeling has gained attention in remote sensing to address the mixed-pixel problem. Previous works often employ heavy multi-task learning frameworks to jointly predict semantic masks and field contours [50]. While effective, these approaches introduce substantial computational overhead. To evaluate boundary quality rigorously, metrics like Boundary IoU [35] have been proposed. Leveraging these insights, our parameter-free Boundary-Guided constraints offer a highly scalable alternative to resolve mixed-pixel ambiguity without inflating the model size. In this paper, we adopt a factorized temporo-spatial encoder that separately processes temporal and spatial information. Unlike previous approaches, we optimize the temporal encoder with Reliability-Aware Token Merging and frequency filtering, and enhance the spatial encoder with Boundary-Guided constraints.

3. Method

In this section, we detail the architecture of FreqEdgeViT. We adopt a factorized spatiotemporal design, consisting of a Physics-Informed Temporal Encoder (Section 3.2) and a Boundary-Guided Spatial Encoder (Section 3.3).

3.1. Overview and Patch Embedding

Following the TSViT framework [30], the input SITS data $\mathbf{X} \in \mathbb{R}^{T\times H\times W\times C}$ is divided into non-overlapping patches. Let D denote the number of patches along the spatial height and width. The flattened patches are projected into a d-dimensional space, resulting in input tokens $\mathbf{Z} \in \mathbb{R}^{T\times N\times d}$, where $N = D \times D$ is the total number of spatial tokens. The architecture processes these tokens through separate temporal and spatial encoders to optimize efficiency. Figure 1 illustrates the overall architecture of the model.

3.2. Physics-Informed Temporal Encoder

The temporal encoder processes the time series of each spatial patch independently. To address the redundancy and noise issues in SITS, we introduce the Phenology-Aware Frequency Filter (PAFF) and Reliability-Aware Token Merging (Ra-ToMe).

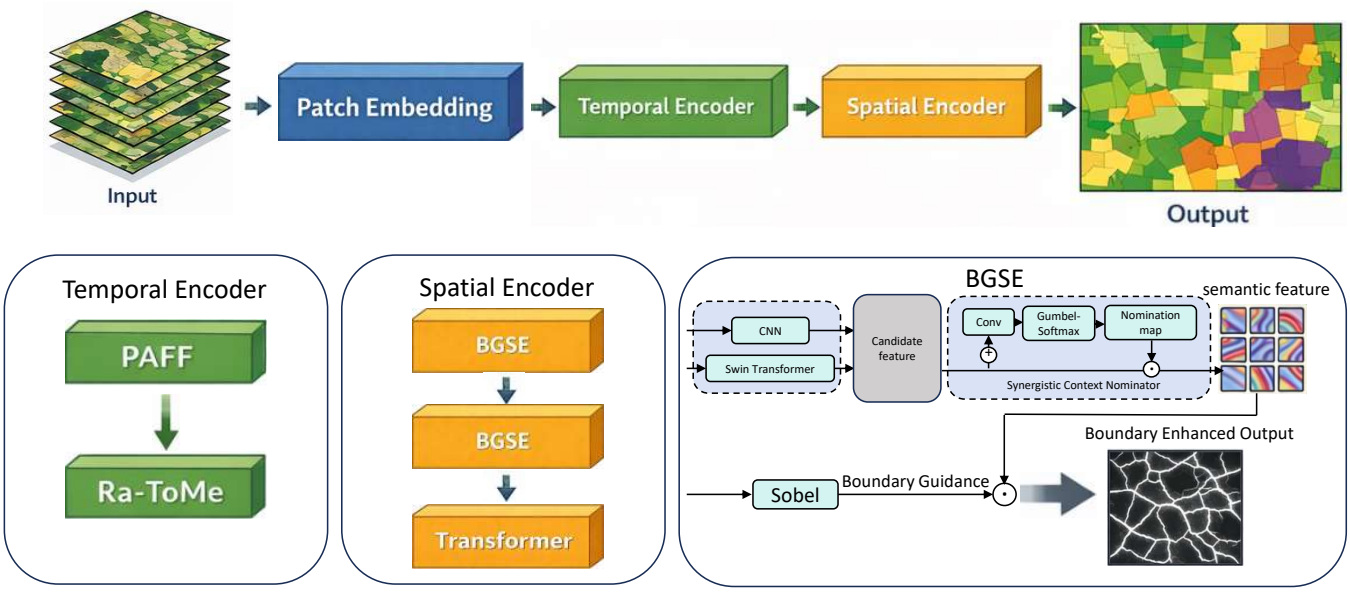

Figure 1. Overall architecture. The overview of the proposed framework architecture is as follows: The framework includes a temporal encoder and a spatial encoder, processing temporal and spatial features respectively. In the temporal encoder, the input first suppresses high-frequency noise via PAFF, then extracts features through a Transformer integrated with Ra-ToMe. The spatial encoder consists of two stacked BGSE blocks and one global Transformer layer. Each BGSE block has two branches: the Semantic Stream and the Boundary Stream, focusing on extracting local contextual features and boundary features respectively

Phenology-Aware Frequency Filter (PAFF). Standard temporal encoders typically rely on temporal convolutions or local attention, which possess restricted receptive fields and struggle to distinguish long-term biological trends from sudden environmental anomalies. From a biophysical perspective, crop phenological stages evolve gradually, manifesting as continuous, low-frequency signals. Conversely, non-phenological environmental artifacts (e.g., transient clouds) appear as abrupt, high-frequency spikes. Based on this signal processing principle, we introduce the PAFF module to explicitly encode this physical prior. By transforming the sequence into the frequency domain, PAFF instantly obtains a global temporal receptive field, allowing it to explicitly separate and purify features more effectively than local temporal convolutions. Let $\mathbf{Z}_T \in \mathbb{R}^{N\times T\times d}$ be the reshaped temporal tokens. We apply the Fast Fourier Transform (FFT) along the temporal dimension:

$$\mathcal{F}(\mathbf{Z}_T) = \mathrm{FFT}(\mathbf{Z}_T) \tag{1}$$

A learnable complex-valued filter $\mathbf{W}_{freq}$ is applied via element-wise multiplication to suppress high-frequency artifacts. The refined features $\mathbf{Z}_{clean}$ are recovered via Inverse FFT:

$$\mathbf{Z}_{clean} = \mathrm{Real}(\mathrm{IFFT}(\mathcal{F}(\mathbf{Z}_T) \odot \mathbf{W}_{freq})) + \mathbf{Z}_T \tag{2}$$

This process ensures that the subsequent processing is driven by true phenological trends.

Reliability-Aware Token Merging (Ra-ToMe). We integrate Token Merging [33] to reduce computational cost. It is crucial to note that this merging operates strictly along the temporal dimension (T) for each independent spatial patch. Therefore, the spatial resolution remains unaltered, ensuring that small field parcels are not inadvertently merged into the spatial background. However, instead of merging based solely on similarity, we propose a reliability-aware strategy. The process is shown in Figure 2. First, a reliability score S is computed by evaluating the deviation between the raw input and the purified signal using a Multi-Layer Perceptron (MLP):

$$S = \sigma(\mathrm{MLP}(|\mathbf{Z}_T - \mathbf{Z}_{clean}|)) \in \mathbb{R}^{N\times T\times 1} \tag{3}$$

where $|\cdot|$ denotes the absolute difference, and σ is the Sigmoid function. The MLP maps multi-channel noise patterns to a scalar reliability score. During the bipartite matching process, tokens are weighted by S. When merging a source token a and a destination token b, the new token is computed as:

$$\mathbf{z}_{new} = \frac{S_a \cdot \mathbf{z}_a + S_b \cdot \mathbf{z}_b}{S_a + S_b} \tag{4}$$

This ensures that the merged tokens retain high-fidelity information, reducing the sequence length from T to T' while preserving phenological integrity. This dynamic reduction in data dimensionality directly contributes to system scalability. By aggressively pruning redundant

tokens early, Ra-ToMe curtails the memory footprint and the computational complexity of subsequent layers. In a large-scale agricultural information system, this reduction translates to higher data throughput, enabling the processing of broader geographical extents using the same hardware infrastructure. After processing via several Transformers integrated with Ra-ToMe, features are extracted while the number of tokens is drastically reduced.

3.3. Boundary-Guided Spatial Encoder

The output of the temporal encoder is reshaped into spatial feature maps $\mathbf{Z}_S \in \mathbb{R}^{(B\cdot K)\times D\times D\times d}$, where K is the number of classes. To resolve boundary ambiguity, we propose the Boundary-Guided Spatial Encoder (BGSE). Structurally, the encoder consists of two stacked BGSE blocks to progressively refine features, followed by a final Global Transformer layer.

Semantic Stream: LCA with SCN. To capture both rigid textures and flexible contexts, we use the Local Context Aggregation (LCA) module equipped with a Synergistic Context Nominator (SCN)[51]. The input flows into two parallel branches: a CNN Branch using a bottleneck convolution[52] to extract local texture features $\mathbf{F}_{cnn}$, and a Swin Branch using Window-based Self-Attention to capture local context $\mathbf{F}_{swin}$ [53].

Instead of simple summation, the SCN dynamically determines the optimal feature combination for each pixel. We concatenate the features from both branches and project them via a lightweight 1×1 convolutional layer to generate the selection logits $\mathbf{L} \in \mathbb{R}^{D\times D\times 2}$. To enable differentiable hard selection during training, we use the Gumbel-Softmax strategy to generate the discrete selection mask $\mathbf{I} \in \{0, 1\}^{D\times D\times 2}$. Importantly, this selection mechanism only introduces a lightweight 1×1 convolution, adding negligible computational overhead and parameters while providing the significant architectural benefit of adaptive feature routing.

The final semantic feature $\mathbf{F}_{sem}$ is obtained by selecting the features from the corresponding branch based on $\mathbf{I}$:

$$\mathbf{F}_{sem} = \mathbf{I}_{:,:,0} \odot \mathbf{F}_{cnn} + \mathbf{I}_{:,:,1} \odot \mathbf{F}_{swin} \tag{5}$$

This mechanism allows the model to explicitly nominate the most appropriate feature extractor for different spatial regions. For instance, in homogeneous field interiors, the SCN tends to select the Swin branch to leverage flexible global contexts, thereby preventing local noise overfitting. Conversely, at the mixed-pixel field borders, the SCN favors the CNN branch to preserve sharp, high-frequency texture cues. By dynamically optimizing the feature representation pixel-by-pixel, the SCN inherently resists the over-smoothing tendency of standard Transformers, laying a high-quality semantic foundation for the subsequent boundary extraction.

Boundary Stream and Interaction. To explicitly model geometric boundaries, we introduce a lightweight Sobel branch. Compared to complex learnable edge-detection heads, employing fixed Sobel kernels G_x, G_y provides a robust inductive bias suited for scalable applications. Learnable convolutions might degenerate into generic texture extractors, whereas a fixed gradient operator strictly defines boundaries as high-frequency semantic transitions, forcing the upstream backbone to learn sharper representations. We perform Depth-wise Convolution (DWConv) using these fixed kernels. DWConv computes gradients independently for each feature channel, ensuring that distinct semantic changes are captured without cancellation. The gradient magnitude map $\mathbf{E} \in \mathbb{R}^{D\times D\times 1}$ is then aggregated via a point-wise convolution:

$$\mathbf{E} = \sigma(\mathrm{Conv}_{1\times 1}(\sqrt{\mathrm{DWConv}(\mathbf{Z}_S, G_x)^2 + \mathrm{DWConv}(\mathbf{Z}_S, G_y)^2})) \tag{6}$$

Finally, the boundary information refines the semantic features via residual interaction:

$$\mathbf{F}_{out} = \mathbf{F}_{sem} \odot (1 + \mathbf{E}) \tag{7}$$

This interaction acts as a spatial attention mechanism, forcing the network to pay heightened attention to boundary pixels where classification errors are most likely to occur. After passing through the stacked BGSE blocks, the refined features are fed into a single standard Global Transformer layer. Since the preceding LCA modules have already thoroughly aggregated local contexts, a single global layer is sufficient to capture field-level long-range dependencies across the entire image, ensuring global coherence without introducing the prohibitive quadratic computational overhead of multiple global layers before the final prediction. Finally, the output of the spatial encoder is processed by an MLP layer to restore its dimensions, and then mapped back to the class probability maps for each pixel location of the original image.

3.4. Loss Function

We employ a composite loss function with Deep Supervision. The total loss is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{MaskedCE}(\mathbf{Y}_{pred}, \mathbf{Y}_{gt}) + \lambda \cdot \mathcal{L}_{BCE}(\mathbf{E}_{last}, \mathbf{E}_{gt}) \tag{8}$$

Here, $\mathcal{L}_{MaskedCE}$ denotes the Cross-Entropy loss calculated only on valid pixels. $\mathcal{L}_{BCE}$ is the auxiliary Binary Cross-Entropy loss. While traditional deep supervision applies auxiliary losses to multiple intermediate layers, we calculate the auxiliary loss exclusively using the edge map $\mathbf{E}_{last}$ output from the final BGSE block.

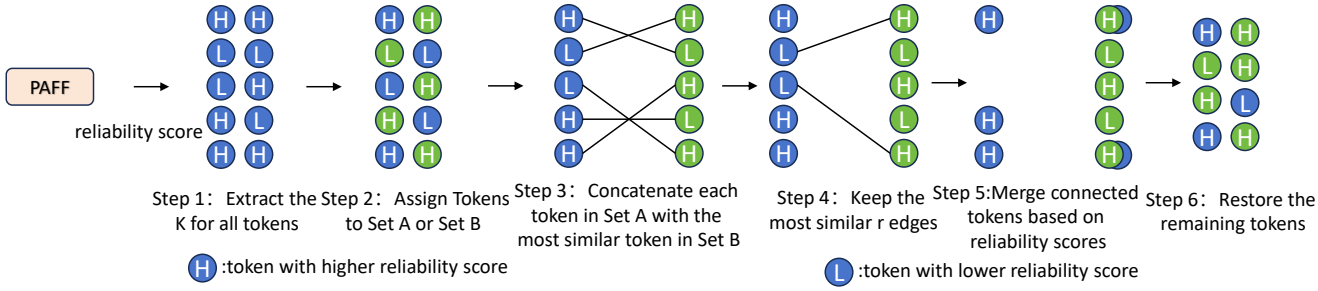

Figure 2. The Process of Ra-ToMe

This design choice provides sufficient gradient guidance for the spatial encoder to learn geometric boundary representations while keeping the training process computationally efficient and stable. The ground truth edge map $\mathbf{E}_{gt} \in \mathbb{R}^{D\times D}$ is derived from the segmentation labels using an edge detection operator and downsampled to match the feature map resolution using Max Pooling, ensuring that fine-grained boundary signals are preserved at lower resolutions. In our experiments, we empirically set the balancing hyperparameter $\lambda = 0.15$ to prioritize the primary semantic classification task while providing sufficient geometric boundary constraints without overshadowing the main objective.

4. Experiments

We employ the FreqEdgeViT model to tackle the semantic segmentation task with input satellite image time series $\mathbf{X} \in \mathbb{R}^{T\times H\times W\times C}$. The semantic segmentation task aims to predict the class probability of each pixel within the spatial range of the input data. We demonstrate our research findings on two publicly available semantic segmentation datasets.

4.1. Experiment Settings

Datasets To evaluate the performance of our proposed semantic segmentation model, we utilize two publicly available datasets. The PASTIS dataset [29] contains images from four distinct regions in France, covering a total area of over 4,000 square kilometers and encompassing 18 crop types. This dataset comprises 2,400 SITS samples, each with a size of 128 × 128 pixels, including 33 to 61 acquisitions and 10 image bands. Because the original PASTIS sample size (128 × 128) is too large for efficient spatiotemporal sequence modeling, we spatially crop each sample into 24 × 24 non-overlapping patches while retaining all acquisition times, resulting in a total of approximately 60,000 training samples. The T31TFM-1618 dataset [54] covers a densely cultivated Sentinel-2 tile in France spanning the period 2016–2018, containing 20 categories in total. It includes 140,000 samples, each of size 48 × 48, with 14 to 33 image acquisitions and 13 bands. For both datasets, we strictly utilize the officially provided data splits to ensure a fair comparison and prevent spatial data leakage.

Implementation Details We train on all datasets using the provided training sets and report results on the test set. When training the FreqEdgeViT model, we adopt the AdamW optimizer [55] with a weight decay of 0.05, combined with a learning rate scheduling strategy. A warmup period is employed in the first ten epochs, where the learning rate linearly rises from zero to a peak of 10^{-3}. Subsequently, via cosine decay [56], it drops to 5×10^{-6} by the end of training. We conduct experiments on an NVIDIA GeForce RTX 3090 with a batch size of 32, which is selected to maximize GPU memory utilization while maintaining training stability. Furthermore, our architecture employs a parameter-free Sobel operator, which helps mitigate the risk of overfitting; this operator provides a strong structural inductive bias without introducing learnable parameters. All models use Masked Cross-Entropy loss, and the influence of the background class is masked in both training loss and evaluation metrics. We report the pixel-level averaged overall accuracy (OA) and the class-averaged mean Intersection over Union (mIoU). Regarding the model's computational speed, we report the model's params. (in millions, M) and FLOPs (in billions, G).

Evaluation metrics To evaluate the performance of FreqEdgeViT, we adopt four evaluation metrics: Overall Accuracy (OA), mean Intersection over Union (mIoU), Floating-point Operations (FLOPs), and Parameters (params). Additionally, we report Boundary IoU [35]. In agricultural SITS, crop fields typically occupy large, homogeneous areas. Consequently, standard mIoU scores are heavily dominated by the easily classified interior pixels, which can mask severe misclassifications at the field edges. However, the precise delineation of these mixed-pixel borders is critical for accurate acreage estimation. To rigorously evaluate this, Boundary IoU computes the Intersection over Union exclusively within a fixed narrow band

Table 1. Comparison with state-of-the-art methods on PASTIS and T31TFM-1618 datasets. We report the number of parameters (M), Floating Point Operations (FLOPs), Overall Accuracy (OA), and mean Intersection over Union (mIoU).

Model	#Params (M)	FLOPs (G)	PASTIS		T31TFM-1618	
			OA	mIoU	OA	mIoU
BiCGRU[57]	4.5	30.5	80.5	56.2	88.6	57.7
FPN-CLSTM[58]	1.2	102.3	81.9	59.5	88.4	57.8
UNET3D[28]	6.2	72.3	82.3	60.4	88.4	57.6
UNET3Df[54]	7.2	143.8	82.1	60.2	88.6	57.7
UNET2D-CLSTM[28]	2.3	194.8	82.7	60.7	89.0	58.8
U-TAE[29]	1.1	12.7	82.9	62.4	88.9	58.5
TSViT[30]	1.7	19.9	83.4	65.1	90.3	63.1
FreqEdgeViT (Ours)	**1.2**	**10.9**	**84.8**	**66.8**	**91.3**	**64.6**

Table 2. Ablation study of component contributions. Boundary IoU is reported on the PASTIS dataset to explicitly evaluate the segmentation quality at field boundaries.

Model	Modules						FLOPs (G)	B-IoU	PASTIS		T31TFM	
	ToMe	LCA	PAFF	Ra-ToMe	BGSE	Aux			OA	mIoU	OA	mIoU
Baseline	×	×	×	×	×	×	19.9	59.5	83.4	65.1	90.3	63.1
+ Token Merging	✓	×	×	×	×	×	11.6	57.8	82.4	63.8	88.9	62.1
+ LCA	✓	✓	×	×	×	×	10.5	58.2	82.6	64.0	89.4	62.3
+ PAFF	✓	✓	✓	×	×	×	10.6	58.8	83.1	64.7	90.1	62.9
+ Ra-ToMe	✓	✓	✓	✓	×	×	10.7	59.6	83.8	65.4	90.9	63.4
+ BGSE	✓	✓	✓	✓	✓	×	10.9	63.5	84.5	66.4	91.1	64.3
Ours (Full)	✓	✓	✓	✓	✓	✓	**10.9**	**64.9**	**84.8**	**66.8**	**91.3**	**64.6**

from the object contours, providing a highly sensitive metric for spatial fidelity. The formal definitions and calculation formulas of OA and mIoU are detailed as follows.

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{9}$$

$$\mathrm{IoU}_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k + \mathrm{FN}_k} \tag{10}$$

$$\mathrm{mIoU} = \frac{1}{K} \sum_{k=0}^{K-1} \mathrm{IoU}_k \tag{11}$$

In Equations (9), (10), and (11), TP, TN, FP, and FN denote the number of correctly predicted positives, correctly predicted negatives, incorrectly predicted positives, and incorrectly predicted negatives, respectively. In Equations (10) and (11), the subscript k represents the k-th class, and K denotes the total number of classes.

4.2. Comparison with State-of-the-Art

In Table 1, we compare FreqEdgeViT with state-of-the-art models, which can be categorized into recurrent architectures (BiCGRU, FPN-CLSTM), CNN-RNN hybrids (UNET2D-CLSTM), pure CNNs (UNET3D, UNET3Df), and Transformer-based models (U-TAE, TSViT). Specifically, U-TAE combines a spatial U-Net with a lightweight temporal attention encoder, while TSViT employs a heavy factorized architecture applying self-attention sequentially in both temporal and spatial dimensions. While heavy architectures like UNET2D-CLSTM incur prohibitive computational overhead, lightweight alternatives such as U-TAE, despite their efficiency, often exhibit a trade-off in segmentation accuracy. The visualization of prediction results on the PASTIS dataset is illustrated in Figure 3. As shown in the results, FreqEdgeViT outperforms previous methods in terms of efficiency and accuracy. Compared to the baseline TSViT, our model achieves a reduction in FLOPs of approximately 45% while improving mIoU by 1.7% on the PASTIS dataset. Even when compared to the lightweight U-TAE, FreqEdgeViT demonstrates lower computational complexity and significantly higher accuracy (+4.4% mIoU). This confirms that our physics-informed and geometric-constrained design enables extremely efficient learning without sacrificing capacity.

4.3. Ablation Studies

To investigate the contribution of each component, we conduct a comprehensive ablation study on the PASTIS dataset. The results are summarized in Table 2.

Efficiency vs. Accuracy Trade-off. The introduction of Token Merging and LCA initially reduces FLOPs significantly (from 19.9G to 10.5G). However, this

Table 3. Robustness analysis against temporal noise. We simulate high-frequency perturbations by randomly masking 20% of the time steps. "Retention Rate" indicates the percentage of performance maintained relative to the clean baseline.

Model	Clean mIoU	Noisy mIoU	Drop	Retention Rate
TSViT	65.1	60.5	-4.6	92.9%
+ Token Merging	63.8	58.2	-5.6	91.2%
FreqEdgeViT (Ours)	**66.8**	**64.5**	**-2.3**	**96.6%**

Table 4. Parameter sensitivity analysis of the token reduction count r on the PASTIS dataset. We compare computational cost (FLOPs), relative speedup, and segmentation performance.

Reduction Count (r)	FLOPs (G)	Relative Speedup	OA	mIoU
2	15.8	1.00×	84.8	66.9
4	14.1	1.13×	84.8	66.9
6	12.4	1.27×	84.9	67.1
8	10.9	1.46×	84.8	66.8

aggressive reduction leads to a performance drop (mIoU 65.1% → 64.0%), confirming that naive pruning may discard valuable spatiotemporal information. By integrating PAFF and upgrading ToMe to Ra-ToMe, the model recovers from the accuracy drop, surpassing the baseline with negligible computational overhead (+0.2G). This verifies that filtering high-frequency noise allows the model to retain higher-quality features even with fewer tokens.

Geometric Boost. Adding the BGSE module triggers a sharp increase in Boundary IoU (59.6% → 63.5%). Finally, with Deep Supervision (Aux Loss), the model achieves optimal performance (66.8% mIoU, 64.9% Boundary IoU), proving that explicit geometric constraints effectively resolve the mixed-pixel ambiguity that standard Transformers struggle with.

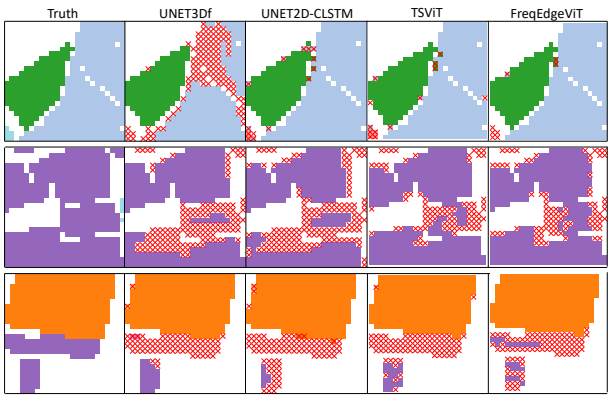

Figure 3. Visualization of Prediction Results on the PASTIS Dataset. "x" Indicates Incorrect Predictions

4.4. Robustness and Parameter Analysis

Robustness to Temporal Noise. To validate the effectiveness of our PAFF module, we evaluate model performance under simulated high-frequency noise (20% random masking of time steps). As shown in Table 3, the standard lightweight model (+Token Merging) suffers a severe performance drop. In contrast, FreqEdgeViT exhibits remarkable stability, with a much smaller degradation. This confirms that our frequency-aware design effectively filters out non-phenological perturbations.

Parameter Sensitivity. We further analyze the impact of the token reduction count r. As shown in Table 4, increasing r significantly boosts inference speed. Although $r = 6$ yields the peak accuracy, we select $r = 8$ as our default configuration. It offers a substantial speedup compared to lower reduction counts with only a negligible accuracy trade-off (-0.3%).

5. Discussion

Our research aims to address the inherent scalability bottlenecks in agricultural information systems: the prohibitive computational costs of processing high-dimensional vision models, the vulnerability to environmental noise in distributed satellite data, and the loss of spatial fidelity i n e fficiency-oriented architectures. To this end, we proposed FreqEdgeViT, a reliability-aware and boundary-guided Vision Transformer designed for scalable SITS semantic segmentation. The experimental results on the PASTIS and T31TFM-1618 datasets validate the efficiency an d precision of our design. The following sections discuss the underlying mechanisms, implications for scalable information applications, limitations, and future directions.

5.1. Mechanism of Efficiency and Reliability

FreqEdgeViT achieves a balance between computational efficiency and segmentation accuracy by optimizing its architecture to address the specific characteristics of SITS data streams. Unlike general image data, agricultural SITS is characterized by temporal redundancy, high-frequency environmental noise, and mixed-pixel boundaries.

FreqEdgeViT solves these challenges by combining physical principles and geometric constraints:

- Scalable Processing via Spectral Analysis: Standard data reduction methods often compromise information integrity due to noise sensitivity. By integrating the Phenology-Aware Frequency Filter (PAFF) and Ra-ToMe, our model utilizes the spectral property that crop growth manifests as low-frequency trends. This allows the system to aggressively compress the temporal dimension (reducing FLOPs by ~45%) while simultaneously filtering out high-frequency artifacts caused by atmospheric scattering.

- Information Precision via Geometric Constraints: A persistent challenge in remote sensing information extraction is the ambiguity of "mixed pixels" at field borders. The proposed Boundary-Guided Spatial Encoder (BGSE) explicitly models these geometric transitions. By reinforcing the attention mechanism with explicit edge cues, the model ensures that the efficiency gains do not result in the loss of critical spatial details, bridging the gap between coarse patch-based processing and the pixel-level precision required by downstream applications.

5.2. Implications for Scalable Information Applications

Beyond model performance, FreqEdgeViT offers significant value for agricultural information systems, extending technical segmentation results to practical data-driven decision support.

For cultivated land resource management, the precise boundary delineation capability of FreqEdgeViT enhances the reliability of area estimation algorithms. Accurate extraction of field borders enables precise statistics of planting areas across large-scale regions. This provides a robust data foundation for government systems to verify agricultural subsidies and monitor crop rotation compliance, ensuring the integrity of ecological sustainability policies. Crucially, the suppression of environmental noise via the PAFF module ensures the temporal consistency of the extracted features. In practical operations, misclassifications caused by unmitigated noise directly skew regional acreage statistics. By filtering out these high-frequency perturbations, our model delivers highly reliable crop distribution maps, which serve as a trustworthy foundation for downstream tasks such as regional yield estimation and market supply forecasting.

In terms of precision irrigation systems, the crop type maps generated by FreqEdgeViT serve as a reliable static layer for fusion with dynamic environmental data (e.g., soil moisture sensors) in IoT networks. By accurately distinguishing crop types with varying water demands, the system can optimize irrigation schedules. This integration reduces resource waste and enhances the overall resource utilization efficiency of smart agriculture platforms.

5.3. Scalability in Distributed IoT Ecosystems

The lightweight and robust nature of FreqEdgeViT establishes a foundation for its deployment in distributed agricultural IoT ecosystems. Unlike heavy models that depend on centralized high-performance computing, FreqEdgeViT's low computational footprint facilitates deployment on edge servers or regional data nodes closer to the data source. The model contains only 1.2 million parameters, making it exceptionally well-suited for the stringent storage and memory constraints faced by edge devices deployed in the field. Furthermore, its inherent robustness to temporal noise (verified by our masking experiments) ensures system reliability even when input data quality fluctuates due to weather conditions. This capability supports a "satellite + ground" collaborative monitoring paradigm, enabling real-time, comprehensive crop information processing for intelligent management.

5.4. Limitations

Despite these advantages, this study is subject to certain limitations. First, regarding data availability and climatic patterns, FreqEdgeViT currently relies primarily on optical SITS. While the PAFF module effectively filters high-frequency noise, optical sensors face an insurmountable physical limitation in penetrating thick, persistent cloud cover. If critical phenological transition periods coincide with a prolonged rainy season, the optical time series suffers from massive continuous data gaps, constraining the system's information retrieval capability. Furthermore, the effectiveness of the PAFF module inherently assumes distinct phenological frequency signatures among different crops. In regions where multiple crop types share nearly identical planting and harvesting cycles, or in the case of evergreen agriculture, the low-frequency spectral differences may become marginal, challenging the filter's discriminative power. Second, regarding generalization, while the model performs effectively on the tested datasets, agricultural patterns vary significantly across climatic

zones. The transferability of the learned frequency filters and boundary priors to vastly different agricultural landscapes requires further validation to ensure global scalability. Third, regarding geometric constraints, the BGSE module relies on fixed Sobel operators. While this ensures computational efficiency and provides a strong inductive bias against overfitting, its effectiveness heavily depends on the upstream layers' ability to initially disentangle class features. If the semantic differences between adjacent crop patches are entirely lost in the early representation stages, the fixed gradient operator cannot dynamically recover the boundary, unlike heavier, learnable edge-prediction decoders.

5.5. Future Prospects

Future research will focus on two main directions to enhance system capabilities. First, we aim to explore multi-modal data fusion by integrating Synthetic Aperture Radar (SAR) data. Since SAR signals penetrate clouds, fusing them with optical SITS can compensate for data gaps, enhancing the system's all-weather adaptability. Second, we plan to investigate semi-supervised learning strategies to reduce reliance on large-scale annotated labels, facilitating rapid adaptation to new regions and further improving the scalability of the information system.

6. Conclusion

FreqEdgeViT addresses the critical bottlenecks of processing massive, noisy satellite data streams under resource constraints. We move beyond the standard speed-accuracy trade-off by integrating frequency-domain filtering and explicit boundary guidance. Specifically, the PAFF and Ra-ToMe mechanisms suppress non-phenological noise to enable safe data compression, while the BGSE resolves spatial ambiguity at field borders. Experimental results on two datasets show that our method achieves state-of-the-art performance while significantly reducing computational cost. These results provide a viable technical foundation for scalable, data-driven agricultural monitoring systems. Moving forward, we plan to incorporate multi-modal data to further enhance system adaptability in complex environments.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (NSFC) through "Research on Distributed Technology Governance Model for Cross-Border Data Flow Based on Blockchain" under Grant 62462065 and through "Security and Governance Framework for China-ASEAN Cross-Border Data Circulation" under Grant 72441009; in part by the Major Science and Technology Special Project of Yunnan province under Grant 202402AD080005 through "Research on Key Technologies and Innovative Applications for Smart Ports"

References

[1] Rustemi, A. and Dalipi, F. (2025) Synergizing iot, ai, and blockchain for smart agriculture: Challenges, opportunities, and future directions. *Internet of Things* : 101778.

[2] Zhang, S., Han, Q., Wang, H., Liu, J. and Li, B. (2025) Federated learning with dual dynamic quantization optimization in smart agriculture. *Internet of Things* : 101798.

[3] Zhang, Q., Ahmed, K., Sharda, N. and Wang, H. (2023) Australian animal species selection and image data collection. In *2023 27th International Conference Information Visualisation (IV)* (IEEE): 55–63. doi:10.1109/IV60694.2023.00018.

[4] Sánchez, J., Rodríguez, J. and Espitia, H. (2020) Review of artificial intelligence applied in decision-making processes in agricultural public policy. *Processes* **8**(11): 1374.

[5] Chandraprabha, M. and Dhanraj, R. (2023) Ensemble deep learning algorithm for forecasting of rice crop yield based on soil nutrition levels. *EAI Endorsed Transactions on Scalable Information Systems* **10**(4).

[6] Hao, X., Li, X., Wang, H., Zheng, Z., Jiang, Y. and Zhang, Y. (2025) Itcohd-mrec: An independent topological preference-aware and cooperative hypergraph diffusion-based multimodal recommender model. *ACM Trans. Inf. Syst.* **44**(1). doi:10.1145/3767337.

[7] Pham, T.T.L., Taherdoost, H. and Le, T.V. (2024) Scalable information systems for agribusiness: Developing farmers' digital capabilities for e-commerce platform adoption. *EAI Endorsed Transactions on Scalable Information Systems* **12**(1).

[8] Xie, W., Zhu, A., Ali, T., Zhang, Z., Chen, X., Wu, F., Huang, J. *et al.* (2023) Crop switching can enhance environmental sustainability and farmer incomes in china. *Nature* **616**(7956): 300–305.

[9] Saleem, H., Ahmad, S., Shafeeque, U.B. and Khan, N.A. (2025) Fuzzy topsis method for sustainable supplier assortment in green supply chain management. *EAI Endorsed Transactions on Scalable Information Systems* **12**(3).

[10] Ge, Y.F., Wang, H., Bertino, E., Cao, J., Zhang, Y. and Zheng, Z. (2025) Distributed bandit-based cooperative coevolution for large-scale multi-objective data publishing. *IEEE Transactions on Services Computing* doi:10.1109/TSC.2024.3517403. Early Access.

[11] You, M., Ge, Y.F., Yin, J., Wang, K., Zheng, Z., Zhang, Y. and Wang, H. (2025) Akief: Adaptive knowledge inheritance evolutionary framework for dynamic privacy-preserving data publishing. *ACM Trans. Web* doi:10.1145/3779413.

[12] Ge, Y.F., Wang, H., Bertino, E., Zhan, Z.H., Cao, J., Zhang, Y. and Zhang, J. (2024) Evolutionary dynamic

database partitioning optimization for privacy and utility. *IEEE Transactions on Dependable and Secure Computing* **21**(4): 2296–2311. doi:10.1109/TDSC.2023.3302284.

[13] Hoque, M., Islam, M. and Ahmed, I.e.a. (2024) Enhancing precision agriculture efficiency through edge computing-enabled wireless sensor networks: A data aggregation perspective. *Engineering Proceedings* **82**(1): 90.

[14] Chen, G., Wang, M., Han, S., Yin, J., Wang, H. and Cao, J. (2025) Deep reinforcement learning-based cloud-edge offloading for wbans. *IEEE Transactions on Consumer Electronics* **71**(2): 4053–4064. doi:10.1109/TCE.2024.3504545.

[15] Ge, Y.F., Wang, H., Bertino, E., Cao, J. and Zhang, Y. (2025) Multiobjective privacy-preserving task assignment in spatial crowdsourcing. *IEEE Transactions on Cybernetics* **55**(8): 3584–3597. doi:10.1109/TCYB.2025.3573292.

[16] Alexopoulos, A., Koutras, K. and Ali, S.e.a. (2023) Complementary use of ground-based proximal sensing and airborne/spaceborne remote sensing techniques in precision agriculture: A systematic review. *Agronomy* **13**(7): 1942.

[17] Chen, X.e.a. (2025) Mask-guided frequency feature fusion for visible–infrared remote sensing object detection. *IEEE Trans. Geosci. Remote Sens.* **63**: 1–15.

[18] Ghosh, H., Rahat, I. and Shaik, K.e.a. (2023) Potato leaf disease recognition and prediction using convolutional neural networks. *EAI Endorsed Transactions on Scalable Information Systems* **10**(6).

[19] Huang, C.Q., Huang, Q.H., Huang, X., Wang, H., Li, M., Lin, K.J. and Chang, Y. (2024) Xkt: Toward explainable knowledge tracing model with cognitive learning theories for questions of multiple knowledge concepts. *IEEE Transactions on Knowledge and Data Engineering* **36**(11): 7308–7325. doi:10.1109/TKDE.2024.3418098.

[20] Chatterjee, S., Satpathy, S. and Nibedita, A. (2023) Digital investigation of network traffic using machine learning. *EAI Endorsed Transactions on Scalable Information Systems* **11**(1).

[21] Gayathri, B. (2024) Opin-itp: Optimized physics informed network with trimmed score regression based insider threats prediction in cloud computing. *EAI Endorsed Transactions on Scalable Information Systems* **12**(1).

[22] Shi, F., Meng, Y., Zhao, Z., Yin, J., Cao, J. and Wang, H. (2026) Liteghost-yolo: scale-aware lightweight traffic sign recognition in complex environments. *Journal of Real-Time Image Processing* **23**(1): 19. doi:10.1007/s11554-025-01584-6.

[23] Huang, E., Zhao, Z., Yin, J., Cao, J. and Wang, H. (2025) Transformer-enhanced adaptive graph convolutional network for traffic flow prediction. *ACM Transactions on Intelligent Systems and Technology* doi:10.1145/3702144, URL https://doi.org/10.1145/3702144. Just Accepted.

[24] Alvi, A.M., Khan, M.J., Manami, N.T., Miazi, Z.A., Wang, K., Siuly, S. and Wang, H. (2024) Xcr-net: A computer aided framework to detect covid-19. *IEEE Transactions on Consumer Electronics* **70**(4): 7551–7561. doi:10.1109/TCE.2024.3446793.

[25] Tawhid, M.N.A., Siuly, S., Wang, K. and Wang, H. (2024) Genet: A generic neural network for detecting various neurological disorders from eeg. *IEEE Transactions on Cognitive and Developmental Systems* **16**(5): 1829–1842. doi:10.1109/TCDS.2024.3386364.

[26] Vardhan, K.B., Nidhish, M., Kiran C., S., Dudekula, N.S., Varanasi, S.C. and Bhavadharini, R. (2024) Eye disease detection using deep learning models with transfer learning techniques. *EAI Endorsed Transactions on Scalable Information Systems* **12**(1).

[27] Singh, R., Subramani, S., Du, J., Zhang, Y., Wang, H., Miao, Y. and Ahmed, K. (2023) Antisocial behavior identification from twitter feeds using traditional machine learning algorithms and deep learning. *EAI Endorsed Transactions on Scalable Information Systems* **10**(4).

[28] Rustowicz, R., Cheong, R., Wang, L., Ermon, S., Burke, M. and Lobell, D. (2019) Semantic segmentation of crop type in africa: A novel dataset and analysis of deep learning methods. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops*: 75–82.

[29] Garnot, V. and Landrieu, L. (2021) Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*: 4872–4881.

[30] Tarasiou, M., Chavez, E. and Zafeiriou, S. (2023) Vits for sits: Vision transformers for satellite image time series. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*: 10418–10428.

[31] Chen, L., Liu, W., Wang, H., Jeon, S.W., Jiang, Y. and Zheng, Z. (2025) Consistency-guided adaptive alternating training for semi-supervised salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **35**(7): 7033–7046.

[32] Li, R., Zhang, D., Wang, Y., Jiang, Y., Zheng, Z., Jeon, S.W. and Wang, H. (2025) Open-vocabulary multi-object tracking with domain generalized and temporally adaptive features. *IEEE Transactions on Multimedia* **27**: 3009–3022. doi:10.1109/TMM.2025.3557619.

[33] Bolya, D., Fu, C. and Dai, X.e.a. (2022), Token merging: Your vit but faster, arXiv preprint arXiv:2210.09461.

[34] Park, N. and Kim, S. (2022) How do vision transformers work? In *Proc. Int. Conf. Learning Representations (ICLR)*.

[35] Cheng, B., Girshick, R., Dollár, P., Berg, A. and Kirillov, A. (2021) Boundary iou: Improving object-centric image segmentation evaluation. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*: 15334–15342.

[36] Rao, Y., Zhao, W., Zhu, Z., Lu, J. and Zhou, J. (2021) Global filter networks for image classification. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*: 980–993.

[37] Garnot, V., Landrieu, L., Giordano, S. and Chehata, N. (2019) Time-space trade-off in deep learning models for crop classification on satellite multi-spectral image time series. In *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*: 6247–6250.

[38] Tarasiou, M. and Zafeiriou, S. (2022), Embedding earth: Self-supervised contrastive pre-training for dense land cover classification, arXiv preprint arXiv:2203.06041.

[39] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X. and Unterthiner, T.e.a. (2021)

An image is worth 16×16 words: Transformers for image recognition at scale. In *Proc. Int. Conf. Learning Representations (ICLR)*.

[40] Kaselimi, M., Voulodimos, A., Daskalopoulos, I., Doulamis, N. and Doulamis, A. (2022) A vision transformer model for convolution-free multilabel classification of satellite imagery in deforestation monitoring. *IEEE Trans. Neural Netw. Learn. Syst.* **34**(7): 3299–3307.

[41] Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. and Grover, A. (2023), Climax: A foundation model for weather and climate, arXiv preprint arXiv:2301.10343.

[42] Knapen, R., de Wit, A., Buyukkaya, E., Petrou, P., Paudel, D., Janssen, S. and Athanasiadis, I. (2025) Efficient and scalable crop growth simulations using standard big data and distributed computing technologies. *Computers and Electronics in Agriculture* **236**: 110392. doi:10.1016/j.compag.2025.110392.

[43] Sarkar, S., Ganapathysubramanian, B., Singh, A. *et al.* (2024) Cyber-agricultural systems for crop breeding and sustainable production. *Trends in Plant Science* **29**(2): 130–149.

[44] López-Morales, J.A., Martínez, J.A. and Skarmeta, A.F. (2020) Digital transformation of agriculture through the use of an interoperable platform. *Sensors* **20**(4): 1153. doi:10.3390/s20041153.

[45] Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A. and Sarlos, T.e.a. (2020), Rethinking attention with performers, arXiv preprint arXiv:2009.14794 [cs.LG].

[46] Bolya, D., Fu, C., Dai, X., Zhang, P. and Hoffman, J. (2022), Hydra attention: Efficient attention with many heads, arXiv preprint arXiv:2209.07484 [cs.CV].

[47] Meng, L., Li, H., Chen, B., Lan, S., Wu, Z., Jiang, Y. and Lim, S. (2022) Adavit: Adaptive vision transformers for efficient image recognition. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*.

[48] Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z. and Wei, Y.e.a. (2022) Swin transformer v2: Scaling up capacity and resolution. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*.

[49] Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N. and Yuan, L.e.a. (2022) Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*.

[50] Diakogiannis, F.I., Waldner, F., Caccetta, P. and Wu, C. (2020) Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing* **162**: 94–114. doi:10.1016/j.isprsjprs.2020.01.013.

[51] Liu, H., Jiang, X. and Li, X.e.a. (2022) Nommer: Nominate synergistic context in vision transformer for visual recognition. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*: 12073–12082.

[52] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep residual learning for image recognition. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*: 770–778.

[53] Liu, Z., Lin, Y., Cao, Y., Hu, H. and Wei, Y.e.a. (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*: 10012–10022.

[54] Tarasiou, M., Guler, R. and Zafeiriou, S. (2022) Context-self contrastive pre-training for crop type semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **60**: 1–11.

[55] Loshchilov, I. and Hutter, F. (2019) Decoupled weight decay regularization. In *Proc. Int. Conf. Learning Representations (ICLR)*.

[56] Loshchilov, I. and Hutter, F. (2017) Sgdr: Stochastic gradient descent with warm restarts. In *Proc. Int. Conf. Learning Representations (ICLR)*.

[57] Russwurm, M. and Körner, M. (2018) Multi-temporal land cover classification with sequential recurrent encoders. *ISPRS International Journal of Geo-Information* **7**(4): 129.

[58] Chamorro Martinez, J., Cue La Rosa, L., Queiroz Feitosa, R., Del'Arco Sanches, I. and Happ, P. (2021) Fully convolutional recurrent networks for multidate crop recognition from multitemporal image sequences. *ISPRS Journal of Photogrammetry and Remote Sensing* **171**: 188–201.