

SHD-Mamba: A Spectral–Temporal Hybrid Mamba with Dual-Suppression Decoding for Satellite Image Time Series

Qizhao Long*, Xinming Chen[†], Yunzhuo Dai^{†*}

**College of Computer Science and Technology, Zhejiang University, Hangzhou, China
22421142@zju.edu.cn*

*[†]Territorial Consolidation Center in Zhejiang Province, China
{zjtdzl, 408191358}@{126.com, qq.com}*

Abstract—Satellite Image Time Series (SITS) crop segmentation is crucial for precision agriculture and food security. However, existing methods still face three key limitations when handling irregularly sampled multispectral sequences: (i) spectral processing that lacks physical priors, (ii) difficulty in jointly modeling discrete phenological events and continuous growth dynamics, and (iii) limited receptive fields that hinder large-scale parcel segmentation.

To address the above three limitations, we propose SHD-Mamba, a unified framework integrating spectral-aware input, hybrid temporal modeling, and global spatial decoding for SITS segmentation. Specifically, our framework features: (1) a Spectral Band Grouping Module (SGM) that organizes bands by physical characteristics to inject spectral priors; (2) a Mamba-enhanced Temporal Attention Encoder (M-TAE) that unifies attention-based phenology recognition with Mamba-based continuous dynamics modeling at linear complexity; (3) a DSMamba Block decoder with a parameter-efficient Dual-Suppression Attention Module (DSAM) for efficient global spatial modeling and channel–spatial joint modulation to suppress redundant activations.

SHD-Mamba achieves 64.11% and 76.55% mIoU on PASTIS and MTLCC, surpassing the state-of-the-art by 1.54 and 0.53 points. Ablations reveal that global spatial modeling, hybrid temporal encoding, and physical priors are all critical. This work demonstrates the synergistic potential of state space models and attention mechanisms for SITS segmentation.

Index Terms—Satellite Image Time Series, Crop segmentation, State Space Models, Remote Sensing, Deep Learning

I. INTRODUCTION

Satellite Image Time Series (SITS) crop segmentation is a critical component in precision agriculture, food security assessment, and agricultural policy formulation. Unlike single-date remote sensing semantic segmentation, SITS temporal semantic segmentation requires models to predict crop categories for each pixel across long-term, irregularly sampled multi-temporal multispectral sequences: given observation sequences of the same geographic region throughout an entire growing season, models must simultaneously comprehend the temporal dynamics of phenological evolution (sowing—growth—maturation) and the spatial patterns of par-

cel structures. This temporal-spectral-spatial triadic coupling poses unique modeling challenges for SITS crop segmentation.

Recent methods such as U-TAE [1] and TSViT [2] have achieved notable progress by introducing temporal attention and transformer-style modeling, building upon earlier works [3]. However, existing approaches still face systematic bottlenecks in spectral processing, temporal modeling, and spatial decoding. These limitations restrict further performance improvements, particularly when generalizing across regions with different crop calendars and parcel distributions.

From a spectral perspective, most SITS methods directly concatenate multispectral bands as input channels, treating them uniformly without considering physical characteristics. Different spectral bands encode distinct physical properties (e.g., vegetation vigor, moisture, and atmospheric conditions), and indiscriminate concatenation weakens structured spectral priors, leading to confusion among spectrally similar classes and unstable cross-region generalization. Moreover, the contribution of each band group can vary across growth stages, making a single monolithic encoder less robust. While spectral grouping has been explored in hyperspectral analysis and pre-training [19]–[21], it has not been systematically integrated into end-to-end SITS segmentation frameworks.

On the temporal side, early recurrent approaches (e.g., ConvLSTM [16], [17]) suffer from gradient instability and limited long-range modeling capacity. Attention-based encoders (e.g., LTAE [4] and U-TAE [1]) effectively capture global phenological cues by focusing on key timestamps, but they tend to treat the time series as discrete events and scale poorly with sequence length. Such discrete aggregation is effective for identifying critical dates, yet may be insufficient to represent smooth growth trajectories under irregular sampling. In contrast, selective state-space models (e.g., Mamba [6]), built on structured state space foundations [5], provide efficient long-range dynamics modeling with linear complexity, yet purely SSM-based SITS designs [13], [24], [25] may lose global temporal summarization. While Mamba has shown promise across various vision tasks [7], [9]–[12] and remote sensing applications [14], [15], this motivates a hybrid temporal encoder that combines attention-based phenology recognition with SSM-

*Corresponding author: Yunzhuo Dai (408191358@qq.com).

based continuous dynamics modeling, aiming to retain both global temporal selectivity and long-horizon continuity.

For spatial decoding, traditional U-Net decoders mainly rely on local convolutions for progressive upsampling, where the receptive field grows slowly and struggles to model large-scale irregular parcels and long-range dependencies in agricultural scenes. In crop segmentation, errors often concentrate at parcel boundaries and transition zones, which require both global context for intra-parcel consistency and fine-grained boundary discrimination. Recent 2D state-space models (e.g., VMamba [8]) enable linear-complexity global modeling via selective scanning, providing efficient global receptive fields for decoding. However, such global modeling still requires effective gating. For parcel segmentation, the key is not merely enhancing salient features, but suppressing confusing responses at boundaries to avoid error amplification during upsampling. Yet channel importance is highly position-dependent: edge-related channels matter at boundaries but are irrelevant in homogeneous interiors. Existing gating designs (e.g., CAM+SAM [23]) model channel and spatial importance separately, failing to capture this coupling and introducing extra parameters that increase overfitting on small remote sensing datasets. This motivates a lightweight joint modulation mechanism for robust decoding.

To overcome these limitations, we propose SHD-Mamba, a unified SITS segmentation framework that jointly optimizes spectral processing, temporal modeling, and spatial decoding in an end-to-end manner. **First**, to address the lack of structured spectral priors caused by indiscriminate band concatenation, we introduce a Spectral Grouping Module (SGM) that organizes multispectral bands into physically meaningful groups and performs group-aware feature extraction, thereby improving class separability and cross-region robustness. **Second**, to reconcile global phenology summarization with continuous long-horizon dynamics under irregular sampling, we design a hybrid temporal encoder M-TAE, which couples attention-based timestamp selection with Mamba-based state-space evolution, enabling the model to capture both key phenological stages and smooth growth trajectories efficiently. **Third**, to mitigate boundary-dominated errors and prevent confusion from being amplified during progressive upsampling, we propose a global decoder built upon DSMamba blocks, where a lightweight Dual-Suppression Attention Module (DSAM) performs joint channel-spatial suppression to adaptively dampen redundant activations in transition regions while preserving intra-parcel consistency. By integrating these components, SHD-Mamba provides a collaborative spectral-temporal-spatial modeling pipeline, where domain priors, mechanism complementarity, and global context modeling reinforce each other for robust crop segmentation. The contributions of our research are summarized as follows:

- (1) We introduce spectral grouping priors into end-to-end SITS segmentation frameworks, improving discriminability and generalization;
- (2) We propose a hybrid temporal encoder that collabora-

tively fuses attention and SSM for complementary temporal aggregation;

- (3) We design a linear-complexity global decoder with parameter-efficient dual-suppression attention for channel-spatial joint modulation;

- (4) We conduct extensive experiments on PASTIS and MTLCC benchmarks, achieving state-of-the-art performance and providing comprehensive ablation studies that validate the effectiveness of each component.

II. METHOD

A. Overall Architecture

We propose SHD-Mamba, a hybrid CNN-Mamba architecture for satellite image time series crop segmentation. As shown in Fig. 1, our model comprises three key innovations: (1) the Spectral Band Grouping Module (SGM), which structures multispectral inputs according to physical band characteristics; (2) the M-TAE encoder (Mamba-enhanced Temporal Attention Encoder), which combines LTAE and Mamba for hybrid temporal modeling; and (3) a DSMamba Block-based decoder equipped with the parameter-efficient DSAM (Dual-Suppression Attention Module) for global spatial decoding.

Given input SITS $\mathbf{X} \in \mathbb{R}^{B \times T \times C \times H \times W}$, where B is the batch size, T is the number of time steps, C is the number of spectral bands (10 for PASTIS, 13 for MTLCC), and H and W are the spatial resolution, the model first processes $\mathbf{X}$ through the spectral band grouping module to obtain grouped features. These features are then fed into a CNN encoder consisting of four ResBlock stages for spatial feature extraction, with each stage halving the spatial resolution through strided convolution while increasing channel dimensions. At the bottleneck layer, M-TAE performs hybrid temporal aggregation, compressing the temporal dimension T into a single feature map. Finally, the DSMamba Block-based decoder progressively restores spatial resolution through four upsampling stages while fusing encoder features via temporal skip connections, ultimately outputting a segmentation map $\mathbf{Y} \in \mathbb{R}^{B \times K \times H \times W}$ with resolution $H \times W$, where K is the number of crop classes.

Unlike U-TAE, which employs simple channel concatenation and CNN-based decoding, our architecture explicitly models spectral structure, combines attention and state space models for temporal encoding, and adopts SSM-based global spatial decoding. This unified design achieves a balance between parameter efficiency and long-range dependency modeling across the spectral, temporal, and spatial dimensions.

B. Spectral Band Grouping Module

Existing SITS methods typically concatenate all spectral bands directly as input channels, processing them uniformly without considering their unique physical characteristics. However, Sentinel-2 bands have well-defined functional groupings: RGB and NIR bands are sensitive to vegetation vigor, red-edge bands capture chlorophyll content variations, shortwave infrared bands indicate soil moisture, while atmospheric bands measure aerosols and water vapor. Direct

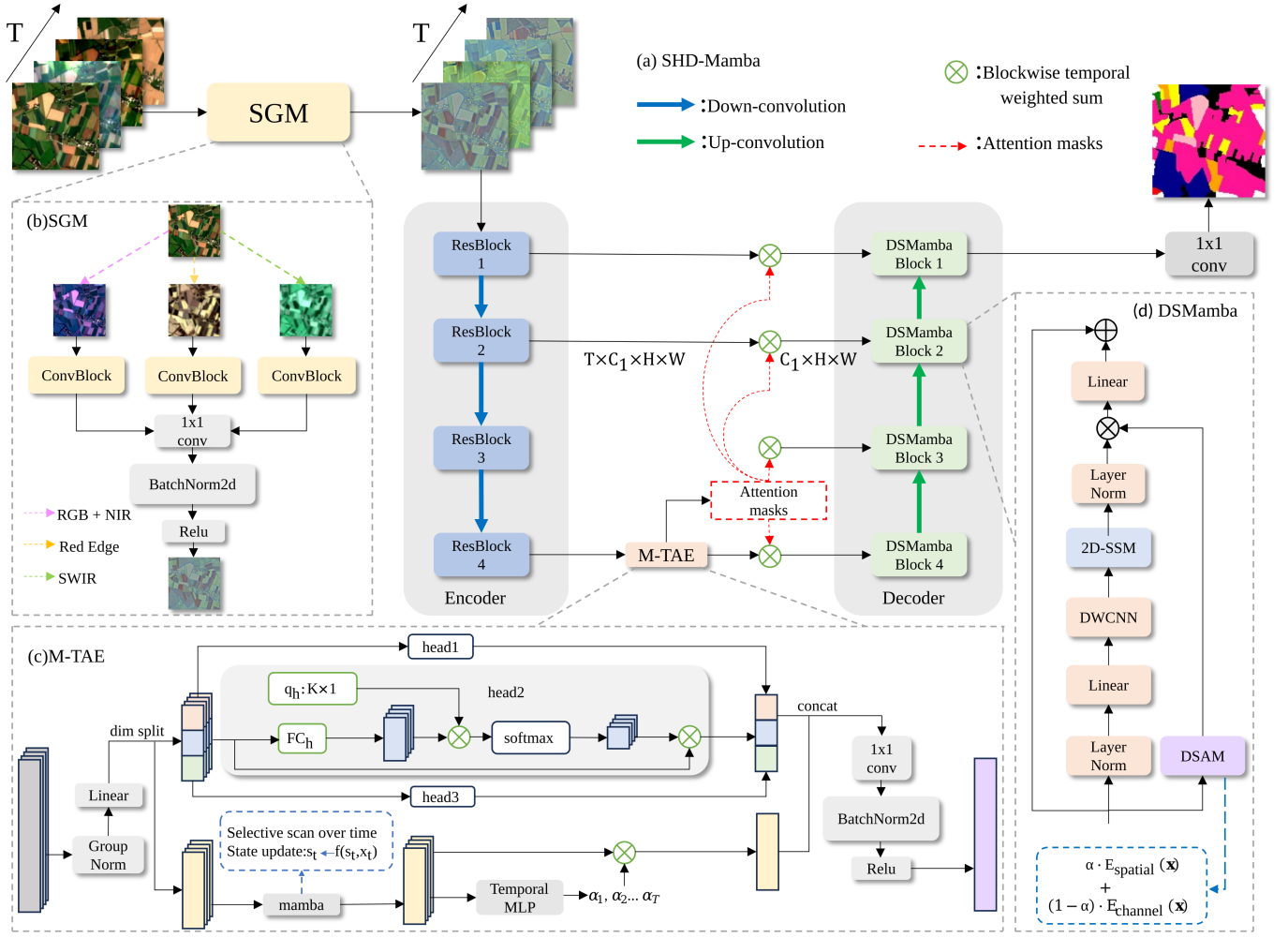


Fig. 1. Overall architecture of SHD-Mamba. (a) The complete framework integrating three key innovations. (b) Spectral Band Grouping Module (SGM) that organizes multispectral bands into physically meaningful groups. (c) M-TAE encoder combining LTAE’s global phenological recognition with Mamba’s continuous dynamics modeling for hybrid temporal aggregation. (d) DSMamba Block decoder with Dual-Suppression Attention Module (DSAM) for parameter-efficient global spatial decoding.

concatenation of these heterogeneous channels forces convolutional kernels to learn cross-band relationships that may not align with natural spectral correlations.

Inspired by spectral grouping explorations in remote sensing pre-training [19], [21], we propose an end-to-end Spectral Band Grouping Module (SGM) that explicitly exploits spectral structure. We divide the C input channels into G groups according to physical characteristics. Each band group $\mathcal{G}_i$ is processed by an independent convolutional block $f_i(\cdot)$ consisting of two 3×3 convolutions with batch normalization and ReLU activation.

The outputs from all groups are then concatenated and fused through a 1×1 convolution to produce the final grouped representation:

$$\mathbf{Z} = \text{Conv}_{1 \times 1}(\text{Concat}[f_1(\mathbf{X}_1), \dots, f_G(\mathbf{X}_G)]) \quad (1)$$

where $\mathbf{X}_i = \mathbf{X}_{\mathcal{G}_i}$ for group i , and $\mathbf{Z} \in \mathbb{R}^{B \times T \times C_1 \times H \times W}$ is the output feature map.

where $\mathbf{Z} \in \mathbb{R}^{B \times T \times C_1 \times H \times W}$ is the output feature map. This design allows the network to first extract group-specific spectral features before cross-group fusion, better aligning with the physical interpretation of multispectral data compared to direct concatenation.

C. Mamba-Enhanced Temporal Attention Encoder

Crop segmentation in SITS requires handling temporal features at two levels: identifying global phenological patterns across the entire growing season, and capturing continuous dynamic changes during crop growth. Existing temporal modeling methods struggle to address both aspects simultaneously: attention-based methods like LTAE [4] excel at identifying discrete phenological events through global comparison, but their computational complexity and attention-based aggregation mechanism limit their ability to model fine-grained temporal continuity.

To this end, we introduce Mamba to SITS temporal modeling for the first time, proposing the Mamba-enhanced Tempo-

ral Attention Encoder (M-TAE). Mamba’s [6] selective state space mechanism is naturally suited for modeling continuous temporal processes: through recursive state updates, it can capture subtle changes and gradual evolution between adjacent time steps at $O(T)$ linear complexity.

Mamba Temporal Modeling: Given the feature sequence at all time steps from encoder layer l , $\{\mathbf{e}_t^l\}_{t=1}^T$, where $\mathbf{e}_t^l \in \mathbb{R}^{D/2 \times H_l \times W_l}$, the Mamba branch processes temporal information through selective state space equations:

$$\mathbf{h}_t = \mathbf{A}(\mathbf{e}_t^l)\mathbf{h}_{t-1} + \mathbf{B}(\mathbf{e}_t^l)\mathbf{e}_t^l \quad (2)$$

$$\mathbf{y}_t = \mathbf{C}(\mathbf{e}_t^l)\mathbf{h}_t \quad (3)$$

where $\mathbf{h}_t \in \mathbb{R}^N$ is the hidden state (state dimension $N = 16$), and $\mathbf{A}(\cdot)$, $\mathbf{B}(\cdot)$, and $\mathbf{C}(\cdot)$ are input-dependent state transition matrices. After processing the entire sequence, we take the output of the final time step as the aggregated feature from the Mamba branch: $\mathbf{o}_{\text{Mamba}} = \mathbf{y}_T$.

Hybrid Architecture Design: We propose a hybrid architecture that combines Mamba with LTAE [4] in parallel: LTAE computes global attention weights and performs weighted aggregation of temporal features $\mathbf{o}_{\text{LTAE}}$, thereby capturing long-range phenological patterns. The outputs of both branches are fused through concatenation:

$$\mathbf{o}_{\text{M-TAE}} = \text{Concat}[\mathbf{o}_{\text{LTAE}}, \mathbf{o}_{\text{Mamba}}] \quad (4)$$

By evenly splitting the feature space between two branches, M-TAE achieves a balance between global phenological pattern recognition and local continuous dynamics modeling: Mamba captures continuous processes while LTAE identifies critical growth stages. The complementarity of both enables comprehensive understanding of crop temporal evolution.

D. Dual-Suppression Mamba Block

Traditional CNN decoders are limited by local receptive fields and require stacking deep networks to capture long-range spatial dependencies. To address this problem, we propose the Dual-Suppression Mamba Block (DSMamba Block), which implements global spatial modeling based on the selective scanning mechanism [8] and achieves precise feature importance modulation through the Dual-Suppression Attention Module (DSAM).

1) *Dual-Suppression Attention Module (DSAM):* Standard Mamba-based decoders adopt a gating mechanism to modulate outputs. Existing implementations (such as CM-UNet [23]) generate gating weights through sequential Channel Attention Modules (CAM) and Spatial Attention Modules (SAM). However, sequential processing assumes channel importance and spatial importance can be modeled independently, failing to capture their coupling relationship in deep features. Additionally, these attention modules introduce large parameter overhead, posing overfitting risks on small-scale datasets. While parameter-efficient attention mechanisms like SimAM [22] have been proposed for general vision tasks, they employ spatial suppression only and do not explicitly model the

channel-spatial coupling relationship critical for deep feature modulation in dense prediction tasks.

From Suppression Theory to Dual Energy Functions: Drawing inspiration from both the spatial suppression theory [28] in neuroscience and recent advances in parameter-efficient attention [22], we propose a unified dual suppression mechanism. We define a dual energy function for each neuron $x_{c,i,j}$ in the feature map:

$$e(x_{c,i,j}) = \alpha \cdot \frac{(x_{c,i,j} - \mu_c)^2}{4(\sigma_c^2 + \lambda)} + (1 - \alpha) \cdot \frac{(x_{c,i,j} - \mu_{i,j})^2}{4(\sigma_{i,j}^2 + \lambda)} \quad (5)$$

where the first term is spatial energy (measuring pixel distinctiveness within the channel), the second term is channel energy (measuring channel distinctiveness at that location), and α is the only learnable parameter for balancing the two energies. The final attention weight is: $A(x) = \sigma(e(x) + 0.5)$.

Compared to CAM+SAM, DSAM achieves channel-spatial joint modeling through a unified energy function, capturing their coupling relationship while introducing only 1 parameter per decoder layer, reducing parameters by 99%. This design significantly reduces overfitting risk on small-scale remote sensing datasets while maintaining expressive capability.

2) *2D Selective Scanning (2DSS):* DSMamba Block achieves global modeling through 2D Selective Scanning (2DSS), which flattens 2D feature maps into 1D sequences along four different directions for processing. Given feature map $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$, 2DSS scans in four directions. Each direction’s scan applies the selective state space model to process that sequence. The scan outputs from four directions are then merged:

$$\mathbf{F}_{\text{out}} = \text{Merge}[\mathbf{S}_1(\mathbf{F}), \mathbf{S}_2(\mathbf{F}), \mathbf{S}_3(\mathbf{F}), \mathbf{S}_4(\mathbf{F})] \quad (6)$$

where $\mathbf{S}_i(\cdot)$ denotes selective scanning in the i -th direction. This multi-directional scanning strategy ensures each pixel can integrate information from all other positions, achieving global receptive fields without quadratic computational complexity.

III. EXPERIMENTS

A. Experimental Setup

Datasets. We evaluate our method on two widely-used SITS crop classification benchmarks:

PASTIS [1] contains 2,433 agricultural parcels from France covering the 2018-2019 growing season. Each parcel includes 38-61 time steps with 10 Sentinel-2 spectral bands at 10m spatial resolution. Images are 128×128 pixels. The dataset contains 19 crop classes. We follow the official 5-fold split protocol: folds 1-3 for training, fold 4 for validation, and fold 5 for testing.

MTLCC [27] covers a 102×42 km agricultural landscape in Munich, Germany, with 6,248 annotated samples from the 2017 growing season. It includes 30 time steps with 13 Sentinel-2 bands. Images are 48×48 pixels at 10m spatial resolution, containing 18 land cover classes. We adopt the official train/validation/test split with a 4:1:1 ratio.

Evaluation Metrics. We adopt two standard metrics: (1) Overall Accuracy (OA); (2) Mean Intersection over Union

TABLE I
PERFORMANCE COMPARISON ON PASTIS AND MTLCC DATASETS

Method	PASTIS		MTLCC		Params (M)	Time (s)	
	mIoU (%)	OA (%)	mIoU (%)	OA (%)		PASTIS	MTLCC
U-ConvLSTM [16]	56.87	81.02	71.82	89.60	1.508	62.32	8.82
3D U-Net [18]	57.53	81.42	72.98	89.44	1.555	65.21	8.93
U-TAE [1]	60.30	83.16	75.11	92.75	1.088	36.81	8.52
TSViT [2]	58.17	81.99	73.87	91.70	1.723	69.13	9.76
UnetMamba [26]	62.24	82.17	75.53	91.98	0.931	32.91	8.13
SITSMamba [13]	62.57	82.46	76.02	92.05	1.275	42.81	8.69
SHD-Mamba (Ours)	64.11	83.45	76.55	93.11	1.128	39.28	8.60

TABLE II
PROGRESSIVE ABLATION STUDY RESULTS

Components	PASTIS		MTLCC	
	mIoU	OA	mIoU	OA
LTAE + CNN	60.30	82.96	75.11	92.75
LTAE + DSMamba	62.16	83.11	75.49	92.91
M-TAE + DSMamba	63.29	83.41	75.80	92.95
M-TAE + DSMamba + SGM	64.11	83.45	76.55	93.11
Δ Total (Full vs Baseline)	+3.81	+0.49	+1.44	+0.36

TABLE III
GATING MECHANISM COMPARISON

Gating	Params	PASTIS		MTLCC	
		mIoU	OA	mIoU	OA
SiLU	0	63.21	83.36	75.77	92.89
+ CAM	294	63.79	83.41	76.19	92.97
+ SAM	768	63.62	83.39	75.93	92.93
CAM+SAM	1,062	63.87	83.44	76.22	93.04
DSAM	3	64.11	83.45	76.55	93.11

(mIoU), which serves as our primary metric as it provides fairer evaluation on imbalanced datasets.

Implementation Details. Our model is implemented in PyTorch and trained on NVIDIA A800 GPUs. The encoder consists of 4 downsampling stages with channel dimensions [64, 64, 64, 128], and the decoder uses [32, 32, 64, 128]. M-TAE inherits LTAE configuration with 16 attention heads, $d_{\text{model}} = 256$, $d_k = 4$. DSMamba Block adopts 4-directional SS2D scanning with Mamba state dimension of 16. We use Adam optimizer with learning rate 1e-3, weight decay 5e-4, batch size 4, trained for 100 epochs.

Baseline Methods. We compare against: (1) U-ConvLSTM [16]; (2) 3D U-Net [18]; (3) U-TAE [1]; (4) TSViT [2]; (5) UnetMamba [26] (equipped with LTAE for temporal modeling); (6) SITSMamba [13].

B. Quantitative Evaluation

Main Results. Table I shows comprehensive comparison on both datasets. SHD-Mamba achieves state-of-the-art performance: on PASTIS, obtaining 64.11% mIoU and 83.45% OA, surpassing U-TAE by 3.81 points and SITSMamba by 1.54 points. On MTLCC, we obtain 76.55% mIoU and 93.11% OA, exceeding U-TAE by 1.44 points and SITSMamba by 0.53 points.

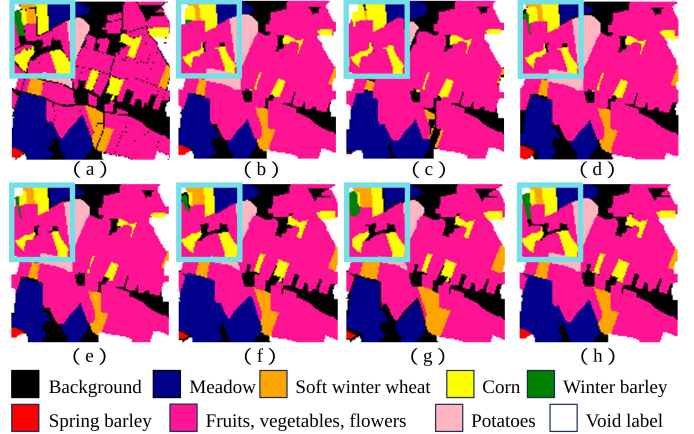


Fig. 2. Qualitative comparison on PASTIS dataset. (a) Ground Truth, (b) U-ConvLSTM, (c) 3D U-Net, (d) U-TAE, (e) TSViT, (f) UnetMamba, (g) SITSMamba, (h) SHD-Mamba (Ours).

Efficiency Analysis. SHD-Mamba demonstrates competitive efficiency with 1.128M parameters, outperforming TSViT (1.723M) while achieving higher accuracy than SITSMamba (1.275M). Inference time is 39.28 seconds on PASTIS and 8.60 seconds on MTLCC, comparable to baseline methods.

Ablation Studies. Tables II and III show comprehensive ablation results. Table II presents progressive ablation on both datasets, starting from U-TAE baseline and sequentially adding each component. On PASTIS: DSMamba Block brings +1.86 point improvement (62.16%), M-TAE adds +1.13 points (63.29%), and SGM contributes +0.82 points (64.11%). On MTLCC: DSMamba contributes +0.38 points, M-TAE adds +0.31 points, and SGM improves +0.75 points.

Gating Mechanism Comparison. Table III validates DSAM effectiveness. Despite having only 3 parameters vs. 1,062 for CAM+SAM, DSAM achieves best performance: 64.11%/76.55% mIoU, outperforming CAM+SAM by +0.24/+0.33 points. This validates that channel-spatial coupling is better captured through DSAM’s unified dual energy function.

C. Qualitative Analysis

Figures 2 and 3 present qualitative comparisons on representative regions. SHD-Mamba produces more coherent segmentation results that better align with parcel structures in complex

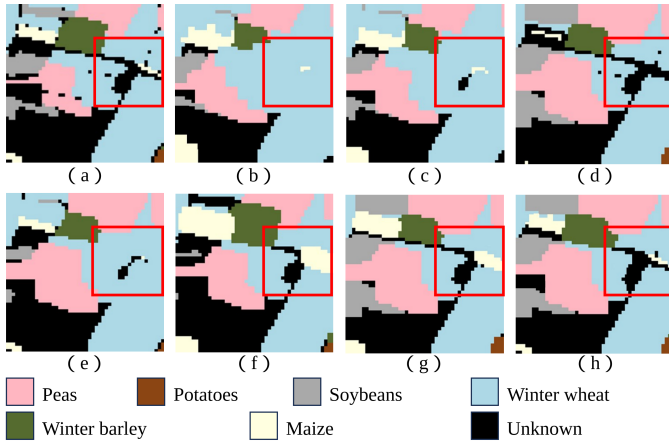


Fig. 3. Qualitative comparison on MTLCC dataset. (a) Ground Truth, (b) U-ConvLSTM, (c) 3D U-Net, (d) U-TAE, (e) TSViT, (f) UnetMamba, (g) SITSMamba, (h) SHD-Mamba (Ours).

agricultural scenes. Benefiting from DSMamba Block’s global spatial modeling, the model yields clearer and more continuous boundaries between adjacent crop types and transition areas. Multi-directional SS2D scanning further improves structural consistency for elongated and irregular parcels, resulting in outputs that better reflect true field geometry. In addition, the combination of SGM’s spectral-aware modeling and M-TAE’s hybrid temporal modeling enhances discrimination for small parcels and spectrally similar crop types, reducing confusion in ambiguous regions.

IV. CONCLUSION

This paper addresses three fundamental limitations in existing SITS crop segmentation methods—insufficient spectral priors, incomplete temporal modeling, and locally constrained decoding—by proposing SHD-Mamba, which integrates SGM for spectral structure priors, M-TAE combining LTAE and Mamba for hybrid temporal modeling, and DSMamba Block with DSAM for parameter-efficient global spatial decoding. Achieving 64.11% and 76.55% mIoU on PASTIS and MTLCC, our method significantly outperforms existing state-of-the-art. Ablation studies confirm that global spatial modeling (+1.86 points), hybrid temporal modeling (+1.13 points), and spectral priors (+0.82 points) are all critical, demonstrating the synergistic potential of state space models and attention mechanisms for SITS analysis.

REFERENCES

- [1] V. Sainte Fare Garnot and L. Landrieu, “Panoptic segmentation of satellite image time series with convolutional temporal attention networks,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 4872–4881.
- [2] M. Tarasiou, E. Chavez, and S. Zafeiriou, “ViTs for SITS: Vision transformers for satellite image time series,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 10418–10428.
- [3] V. Sainte Fare Garnot, L. Landrieu, S. Giordano, and N. Chehata, “Satellite image time series classification with pixel-set encoders and temporal self-attention,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12325–12334.
- [4] V. Sainte Fare Garnot and L. Landrieu, “Lightweight temporal self-attention for classifying satellite images time series,” in *Proc. Int. Workshop Advanced Analytics Learn. Temporal Data*, Springer, 2020, pp. 171–181.
- [5] A. Gu, K. Goel, and C. Ré, “Efficiently modeling long sequences with structured state spaces,” in *Proc. Int. Conf. Learn. Represent.*, 2022.
- [6] A. Gu and T. Dao, “Mamba: Linear-time sequence modeling with selective state spaces,” in *First Conf. Language Model.*, 2024.
- [7] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, “Vision mamba: Efficient visual representation learning with bidirectional state space model,” in *Proc. 41st Int. Conf. Mach. Learn.*, 2024, pp. 62429–62442.
- [8] Y. Liu, Y. Tian, Y. Zhao *et al.*, “VMamba: Visual state space model,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 37, 2024, pp. 103031–103063.
- [9] J. Ma, F. Li, and B. Wang, “U-Mamba: Enhancing long-range dependency for biomedical image segmentation,” *arXiv preprint arXiv:2401.04722*, 2024.
- [10] J. Ruan, J. Li, and S. Xiang, “VM-UNet: Vision Mamba UNet for medical image segmentation,” *ACM Trans. Multimedia Comput. Commun. Appl.*, 2024.
- [11] K. Li, X. Li, Y. Wang *et al.*, “VideoMamba: State space model for efficient video understanding,” in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2024, pp. 237–255.
- [12] G. Chen, Y. Huang, J. Xu *et al.*, “Video mamba suite: State space model as a versatile alternative for video understanding,” *Int. J. Comput. Vision*, vol. 134, no. 1, p. 20, 2026.
- [13] X. Qin, X. Su, and L. Zhang, “SITSMamba for crop classification based on satellite image time series,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2025, pp. 7096–7099.
- [14] H. Chen, J. Song, C. Han, J. Xia, and N. Yokoya, “ChangeMamba: Remote sensing change detection with spatiotemporal state space model,” *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–20, 2024.
- [15] Z. Dewis, Z. Xu, Y. Zhu *et al.*, “Spatial-temporal-spectral Mamba with sparse deformable token sequence for enhanced MODIS time series classification,” *arXiv preprint arXiv:2508.02839*, 2025.
- [16] M. Rußwurm and M. Körner, “Convolutional LSTMs for cloud-robust segmentation of remote sensing imagery,” in *NeurIPS Workshops*, 2018.
- [17] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, “Convolutional LSTM network: A machine learning approach for precipitation nowcasting,” in *Proc. NeurIPS*, 2015.
- [18] R. M. Rustowicz *et al.*, “Semantic segmentation of crop type in Africa: A novel dataset and analysis of deep learning methods,” in *Proc. CVPR Workshops*, 2019, pp. 75–82.
- [19] Y. Cong *et al.*, “SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 197–211.
- [20] J. Yao, D. Hong, C. Li, and J. Chanussot, “SpectralMamba: Efficient mamba for hyperspectral image classification,” *arXiv preprint arXiv:2404.08489*, 2024.
- [21] D. Hong *et al.*, “SpectralGPT: Spectral remote sensing foundation model,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5227–5244, 2024.
- [22] L. Yang, R.-Y. Zhang, L. Li, and X. Xie, “SimAM: A simple, parameter-free attention module for convolutional neural networks,” in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 11863–11874.
- [23] M. Liu, J. Dan, Z. Lu *et al.*, “CM-UNet: Hybrid CNN-Mamba UNet for remote sensing image semantic segmentation,” *arXiv preprint arXiv:2405.10530*, 2024.
- [24] S. Zhao, H. Chen, X. Zhang, P. Xiao, L. Bai, and W. Ouyang, “RS-Mamba for large remote sensing image dense prediction,” *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024.
- [25] X. Ma, X. Zhang, and M.-O. Pun, “RS³Mamba: Visual state space model for remote sensing image semantic segmentation,” *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.
- [26] E. Zhu, Z. Chen, D. Wang *et al.*, “Unetmamba: An efficient unet-like mamba for semantic segmentation of high-resolution remote sensing images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 22, 2025.
- [27] M. Rußwurm and M. Körner, “Multi-temporal land cover classification with sequential recurrent encoders,” *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 4, p. 129, 2018.
- [28] B. S. Webb, N. T. Dhruv, S. G. Solomon, C. Tailby, and P. Lennie, “Early and late mechanisms of surround suppression in striate cortex of macaque,” *J. Neurosci.*, vol. 25, no. 50, pp. 11666–11675, 2005.