Revisiting the Encoding of Satellite Image Time Series

Xin Cai
cai-x@ulster.ac.uk
Yaxin Bi
y.bi@ulster.ac.uk
Peter Nicholl
p.nicholl@ulster.ac.uk
Roy Sterritt
r.sterritt@ulster.ac.uk

School of Computing
Ulster University
Belfast, UK

Abstract

Satellite Image Time Series (SITS) representation learning is complex due to high spatiotemporal resolutions, irregular acquisition times, and intricate spatiotemporal interactions. These challenges result in specialized neural network architectures tailored for SITS analysis. The field has witnessed promising results achieved by pioneering researchers, but transferring the latest advances or established paradigms from Computer Vision (CV) to SITS is still highly challenging due to the existing suboptimal representation learning framework. In this paper, we develop a novel perspective of SITS processing as a direct set prediction problem, inspired by the recent trend in adopting query-based transformer decoders to streamline the object detection or image segmentation pipeline. We further propose to decompose the representation learning process of SITS into three explicit steps: collect–update–distribute, which is computationally efficient and suits for irregularly-sampled and asynchronous temporal satellite observations. Facilitated by the unique reformulation, our proposed temporal learning backbone of SITS, initially pre-trained on the resource efficient pixel-set format and then fine-tuned on the downstream dense prediction tasks, has attained new state-of-the-art (SOTA) results on the PASTIS benchmark dataset. Specifically, the clear separation between temporal and spatial components in the semantic/panoptic segmentation pipeline of SITS makes us leverage the latest advances in CV, such as the universal image segmentation architecture, resulting in a noticeable 2.5 points increase in mIoU and 8.8 points increase in PQ, respectively, compared to the best scores reported so far.

1 Introduction

Recent years have witnessed a surge of interest in automating the monitoring of the Earth surface based on satellites with high revisit frequency, such as European Space Agency (ESA) Sentinel satellites. In particular, automated large-scale crop type mapping benefits most from leveraging complex temporal dynamics contained in SITS, which can promote the fair allocation of agricultural subsidies and the regulation of the best crop practices being observed by farmers. However, applying deep learning models to extract representative features from SITS is non-trivial, e.g., some of which with a naïve concatenation of spatial and temporal encoders even struggle to surpass the performance of a random forest classifier [15], forcing researchers to devote great efforts to develop bespoke neural architectures.

The pioneering work PSE+TAE[8]/PSE+L-TAE[6] has introduced a promising learning paradigm for SITS, where statistics of spectral values are first summarized across the spatial extent of crop parcels by Multi-Layer Perceptrons (MLPs) that operate independently on unordered sets of pixels. These summarized spatial features are then fed into a temporal encoder with self-attention to uncover underlying temporal patterns, following a spatio-then-temporal factorization order. With the empirical evidence provided by the recent work TSViT [27], however, it argues that the temporal-then-spatial factorization order is a more intuitive design choice for SITS analysis as spatial contexts in medium-resolution satellite imagery provide non-informative information in contrast to high resolution optical imagery, especially for vegetation monitoring or crop type mapping. This line of research has demonstrated one important aspect when designing deep learning models for SITS: decoupling the learning framework into spatially and temporally separated components. However, the lack of flexibility to operate on different input formats, i.e., the pixel-set or image sequence format, imposes restrictions on PSE+TAE or TSViT. Consequently, the classical pretrain-finetune paradigm in CV, i.e., pre-training a classification model on large-scale datasets (e.g., ImageNet [2]) with fully-/self-supervised learning [4, 9] and fine-tuning on downstream tasks such as object detection [22] or semantic segmentation [17], has not been successfully adopted in SITS analysis yet.

Meanwhile, as pointed out by previous work [6, 8], another great challenge for effectively learning representations for SITS is to capture the complex temporal dynamics in crop phenology, i.e., the precise timings of plant events are crucial for distinguishing various crop types [20]. However, recent work for SITS analysis [6, 7, 8, 20] advocates adopting self-attention [28] as a core compute unit without questioning its validity for temporal modelling, especially considering its permutation-invariant nature. Based on the latest findings in time series forecasting [32, 33], the capability of self-attention operations for modelling complex temporal relations is exaggerated due to a lack of rich semantics in numerical time series data. Modules with strong built-in priors or inductive biases on temporal ordering such as the classical exponential smoothing [32] or frequency analysis methods [34] have proven to be superior over the vanilla self-attention mechanism for temporal pattern extraction. But irregularity in the temporal axis which is prevalent in satellite image sequences, e.g., optical acquisitions obstructed by clouds, complicates the problem even further, which usually calls for imputation or interpolation as a preprocessing step [15] or developing an end-to-end learning framework which should reconcile potentially conflicted optimization objectives [26] between interpolation and classification. Except for the validity of self-attention for temporal modelling that has been questioned recently, the quadratic space and time complexity w.r.t. the processed sequence length introduces extra computational concerns for model designs and limits its applicability to dense prediction tasks in SITS [7, 27].

These two observations motivated us to reconsider the existing encoding schemes for SITS: *Do we really need to develop bespoke neural architectures for SITS? Is it possible to adapt established CV paradigms to SITS through a simple yet generic representation learning framework?* Specifically, we propose to frame SITS as sets of observations, inspired by the formulation proposed by [10] for classifying irregularly-sampled and asynchronous time series, where each element is represented by its spectral signatures augmented with static or dynamic covariates such as calendar time or thermal time [20]. Facilitated by this unique perspective, we propose a simple yet effective representation learning framework, dubbed as Exchanger, for SITS processing by decomposing the encoding process into three steps: collect–update–distribute, which excludes the use of self-attention to circumvent its limitations. By simply concatenating the proposed Exchanger with a commonly-used segmentation model from CV, we have showcased for the first time that pre-training a classification model on pixel-set format datasets and fine-tuning it on downstream dense prediction tasks with image sequences as input can lead to the new SOTA performance on PASTIS [7] compared to highly-specialized network architectures. Furthermore, we can directly introduce the latest universal image segmentation architecture Mask2Former [1] into semantic/panoptic segmentation of SITS without any modifications by simply letting it consume output feature maps encoded by Exchanger, outperforming the previous SOTA models by a significant margin. To sum up, the contributions of this work include:

- redefining SITS representation as sets of instances, eliminating restrictions on model design to accommodate different input data formats of SITS. This allows us to utilize the resource efficient pixel-set format for pre-training, followed by fine-tuning on downstream dense prediction tasks, which we argue is a more desirable way to introduce the pretrain-finetune paradigm from CV to SITS.
- explicitly decomposing the representation learning process of SITS into three steps: collect–update–distribute, leading to a conceptually clear and computationally efficient learning framework, dubbed as Exchanger, for generic feature extraction of SITS.
- in contrast to the existing work where temporal and spatial components are intricately interwoven with each other in the dense prediction pipeline, we argue that a clear separation of temporal and spatial encoders can greatly reduce the complexity in model design and facilitate leveraging the latest advances in CV, mitigating the gap between CV and SITS.
- having conducted extensive experiments to verify the effectiveness of our proposed model, which outperforms the previous SOTA models by a significant margin across semantic and panoptic segmentation tasks on PASTIS benchmark dataset.

2 Related Work

Encoding of SITS The high frequency revisit time of satellites enables the exploitation of rich temporal dynamics captured for crop type mapping or vegetation monitoring. Traditional machine learning methods [29] rely on handcrafted features where the encoding has not been properly tackled despite the heavy domain expertise required. Recently, differential neural architectures have dominated the field. Specifically, Convolutional Neural Networks (CNNs) [21] and Recurrent Neural Networks (RNNs) [25] have been adopted as a de facto choice to encode spatial and temporal features, respectively. Furthermore, the convolutional-recurrent hybrid models [24] have been proposed to process SITS by viewing it as spatiotemporal signals. Despite the promising results attained, these early attempts have overlooked the significant differences between natural images/videos and SITS. The pioneering work PSE+TAE [8] has proposed to use MLPs to summarize spatial statistics given the lack of rich spatial semantics in medium-resolution Sentinel-2 images and self-attention to encode temporal patterns, followed by PSE+L-TAE [6] where a light-weight transformer decoder has been used to extract temporal features. Pixel-Set Encoder (PSE) is particularly effective for dealing with the irregularity in parcel geometry by simplifying parcel representation from $T \times C \times H \times W$ to $T \times C \times N$, where T is the length of temporal sequence, C is the number channels, H/W denotes the height/width, and N denotes the number of pixels, and consequently requires significantly less storage memory [8] compared to the patch format. But, when it comes to downstream dense prediction tasks, TAE needs to be integrated into spatial encoders in a complicated manner as shown in the previous SOTA model U-TAE [7], which prevents the replication of the successful pretrain-finetune paradigm. TSViT [27] is the first attempt to bridge the gap between SITS analysis and CV by incorporating a unique inductive bias into ViT [3], which is the temporal-then-spatial factorization based on the observation that spatial contexts provide marginal information for crop type recognition. However, the patch tokenization scheme in ViT is naturally built for images, therefore making TSViT incapable to directly consume unordered pixel-set format, which is a more efficient format for SITS classification and pre-training. Furthermore, the intense computation required by self-attention is exacerbated because the spatial dimension is maintained throughout the whole temporal learning process, which causes TSViT problematic for dense prediction tasks.

3 Proposed Method

In this section, we first reformulate the representation of SITS as sets of observations in contrast to the conventional spatiotemporal signals. Then, we simplify the current encoding process of SITS by eliminating the need to specially account for the spatial dimension and further decompose the temporal feature learning procedure into three explicit steps: collect–update–distribute. The specific network instantiation is deferred to the supplementary material.

Definition. *We describe satellite image sequences captured at a particular geo-referenced location with a certain spatial extent as a set $\mathfrak{S}_i$ of instances/sets $\mathfrak{S}_i = \left\{\boldsymbol{S}^1, \ldots, \boldsymbol{S}^n\right\}$, where each instance/set $\boldsymbol{S}^j$ is comprised of a set of temporal acquisitions $\boldsymbol{S}^j = \left\{\boldsymbol{s}^j_{t_1}, \ldots, \boldsymbol{s}^j_{t_m}\right\}$. And we assume each observation $\boldsymbol{s}^j_{t_k}$ is represented by $\left[\boldsymbol{v}^j_{t_k}, \boldsymbol{p}^j_{t_k}, \odot\right]$, where $\boldsymbol{v}^j_{t_k}$ is feature embedding of sensor measurements, $\boldsymbol{p}^j_{t_k}$ is temporal positional embedding for a particular acquisition time, and $\odot$ serves as a placeholder for other static or dynamic covariates such as geometric boundaries or modality information, opening up the possibility of arriving at a universal representation for SITS. $[\cdot]$ denotes an arbitrary operator to mix the features included in it such as summation or concatenation. Note that the superscript and subscript of $\boldsymbol{s}^j_{t_k}$ denote a spatial and temporal identifier, respectively, and we omit the index i for differentiating parcels to avoid notational clutter.*

In contrast to the commonly-adopted representation of satellite observations as spatiotemporal signals $\mathcal{X}_i \in \mathbb{R}^{T \times C \times H \times W}$, we relax the constraints on spatial dimensions imposed by regular grids, for the spatial structure prior is not indispensable for SITS processing [1] and further restricts the flexibility when it comes to model design. We argue that more emphasis

[1] Note that we restrict the assumption to crop type mapping or vegetation monitoring from SITS. As demonstrated in [12], spatial proximity can be exploited for contrastive representation learning of satellite imagery. Besides, specific land cover recognition, e.g., building footprints, relies most on monotemporal but high resolution imagery [5].

should be placed on the temporal dimension and the aggregation of spatial information can be flexibly dealt with according to output requirements of various tasks. With such a more universal reformulation, the classification problem of SITS is intimately linked to Multiple Instance Learning (MIL) [11] where a single class label is assigned to a bag of instances with no ordering or strong dependencies among each other, i.e., treating each temporal sequence of observations sampled from different sub-locations within a parcel field as independent instances with uneven contributing weights to the final bag-level classification results. Concerning the dense prediction problem, the regular grid arrangement is only retained for matching the required output format rather than being used for mining high-level spatial semantics. And we have observed in experiments that simply appending well-established semantic segmentation models such as U-Net [23] after first summarizing temporal information of SITS leads to superior performance to highly-specialized segmentation networks for SITS such as U-TAE [7], which reveals that rich semantics emerge after temporal processing of SITS and resonates with the temporal-then-spatial factorization order advocated in TSViT [27].

3.1 Temporal Context Clusters

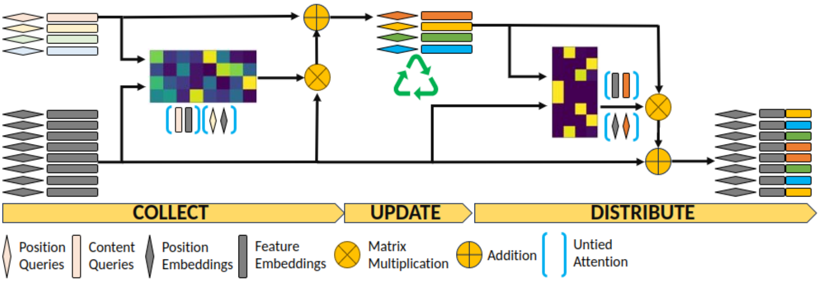

Figure 1: The schematic illustration of the proposed collect–update–distribute procedure for generic representation learning of SITS.

Thanks to our reformulated SITS representation, spatial modeling is not included in the SITS representation learning pipeline due to weak spatial dependencies. As for dense prediction tasks, mining high-level semantics can be accomplished by appending a semantic segmentation model after temporal feature extraction of SITS, which greatly simplifies the existing dense prediction model design for SITS where temporal encoding components are intricately interwoven with spatial encoding components. Motivated by the success of substituting self-attention with other temporal modelling blocks in time series analysis [32, 34], we propose to use a set of learnable queries as an external memory module to exchange temporal information with the input, given that the extra complexity caused by the irregularity in SITS acquisition times, and therefore dub our model "Exchanger".

Formally, we distil the representation learning process of SITS into three steps: collect–update–distribute, as illustrated in Fig.1, with the aid of a set of temporal context clusters, which is further split into two components: content and position queries: $\boldsymbol{C}^v \in \mathbb{R}^{N\times d}, \boldsymbol{C}^p \in \mathbb{R}^{N\times d}$ to avoid blemishing each other, where N is the number of clusters.

- ▷ **COLLECT** Given the input feature embeddings $\boldsymbol{V} \in \mathbb{R}^{T\times d}$ and temporal positional embeddings $\boldsymbol{P} \in \mathbb{R}^{T\times d}$, temporal clusters $\boldsymbol{C}^v$ first collect information from feature embeddings $[\boldsymbol{v}_1,\ldots,\boldsymbol{v}_T]$ by calculating pair-wise similarities followed by a selective function $\mathcal{S}$ to filter out the least significant ones, which is formulated as follows:

$$\begin{aligned}\boldsymbol{A}_1 &= \text{cal\_simlarity}\left([\boldsymbol{C}^v,\boldsymbol{V}],[\boldsymbol{C}^p,\boldsymbol{P}]\right)\\ \boldsymbol{W} &= \mathcal{S}\left(\boldsymbol{A}_1\right)\\ \boldsymbol{C}^v &= \boldsymbol{C}^v+\boldsymbol{W}\boldsymbol{V}\end{aligned} \tag{1}$$

where $\boldsymbol{A}_1 \in \mathbb{R}^{N\times T}$ is the affinity matrix and is further refined by the selective function $\mathcal{S}$ to obtain $\boldsymbol{W}$ to be multiplied by $\boldsymbol{V}$, achieving the collection process.

▷ **UPDATE** Then temporal clusters are updated by solely relying on $\boldsymbol{C}^v,\boldsymbol{C}^p$ to allow for global information exchange among different temporal segments, which is formulated as follows:

$$\boldsymbol{C}^v = \text{Update}\left(\boldsymbol{C}^v,\boldsymbol{C}^p\right) \tag{2}$$

▷ **DISTRIBUTE** After updating the clusters, the more robust and representative features of temporal context clusters are distributed back by assigning each temporal element $\boldsymbol{v}_i$ to $\boldsymbol{C}^v_j$ in a hard or soft manner, which is formulated as follows:

$$\begin{aligned}\boldsymbol{A}_2 &= \text{cal\_simlarity}\left([\boldsymbol{V},\boldsymbol{C}^v],[\boldsymbol{P},\boldsymbol{C}^p]\right)\\ \boldsymbol{I} &= \text{assign}\left(\boldsymbol{A}_2\right)\\ \boldsymbol{V} &= \boldsymbol{V}+\boldsymbol{I}\boldsymbol{C}^v\end{aligned} \tag{3}$$

where $\boldsymbol{A}_2 \in \mathbb{R}^{T\times N}$ is the affinity matrix and each row of $\boldsymbol{I} \in \mathbb{R}^{T\times N}$ contains a hard index or soft probability vector to indicate the temporal context cluster to which each temporal element $\boldsymbol{v}_i$ is assigned.

The proposed temporal representation learning paradigm collect–update–distribute is particularly effective for dealing with the irregularity and asynchronization in time series data as it imposes no prior assumption such as processing temporal observations in a sequential manner. The features of each temporal element can be updated by interacting with temporal context clusters and information flow among different temporal segments is realized through communication between context clusters, which is a more computationally efficient way for information exchange. Compared to the computation complexity of self-attention $\mathcal{O}\left(T^2d\right)$, it only requires $\mathcal{O}\left(NTd\right)$ where $N \ll T$ and therefore scales much better w.r.t. the number of temporal tokens. More importantly, the proposed representation learning framework for SITS can be seen as a generalization of current self-attention based models such as L-TAE [6] or TSViT [27]. To be concrete, L-TAE [6] is a lightweight transformer decoder where a set of learnable queries is used for extracting key features from outputs of the spatial encoder, which corresponds to the collect step we proposed. The lack of update and distribute steps renders L-TAE less effective for encoding as there is no mechanism implemented for feature updating. The temporal encoder of TSViT [27] prepends a set of class tokens to input temporal elements and relies on self-attention for feature learning, which can be seen as a special case of our proposed framework where collect–update–distribute steps are implicitly realized through self-attention. The added external tokens and input temporal elements communicate with each other synchronously, which is more computationally intensive and conceptually vague than our proposed decomposition scheme.

4 Experiments

In this section, we perform extensive ablation studies to verify the effectiveness of our proposed representation learning framework for SITS and make comparisons with previous SOTA models on semantic and panoptic segmentation tasks. Please note the implementation details are deferred to the supplementary material. The code has been made publicly available at https://github.com/TotalVariation/Exchanger4SITS.

4.1 Datasets

We choose PASTIS (Panoptic Agricultural Satellite TIme Series) [2] benchmark dataset [7] to evaluate the performance of our proposed model and make comparisons with previous SOTA models, which consists of 2433 sequences of multi-spectral images of shape $10 \times 128 \times 128$ and each sequence contains temporal acquisitions taken between September 2018 and November 2019 with varying sequence lengths between 38 and 61, for a total of over 2 billion pixels. Furthermore, PASTIS covers four different regions of France with diverse climates and crop distributions, spanning over 4000 km^2 and including 18 crop types plus a background class. In addition to the spatiotemporal format $T \times C \times H \times W$ with high-quality semantic and panoptic annotations, over 120,000 bounding boxes and pixel-precise masks, it is accompanied with a pixel-set format $T \times C \times N$ dataset [8] for parcel-based crop type classification. We mainly use the 5-Fold splits officially provided by PASTIS for extensive ablation studies and model performance evaluation and additionally report semantic segmentation results on another dataset MTLCC [25]. The MTLCC dataset covers a large area of interest (AOI) of 102 km × 42 km north of Munich, Germany, with 17 distinct crop classes and temporal observations of two different lengths of 46 and 52 gathered in two growing seasons in 2016 and 2017 [3].

4.2 Ablation Studies

	Precision%	Recall%	F1 Score%	#Params(M)	FLOPs
w/o Pos. Queries (4)	80.0+0.8	77.0+1.0	78.3+0.9	0.50	117 G
w/ Pos. Queries (4)	83.5+0.6	80.9+0.7	82.0+0.5	0.52	125 G
Untied Cont. & Pos. Attention (4)	83.6+0.6	81.1+0.7	82.2+0.5	0.52	125 G
Untied Cont. & Pos. Attention (8)	83.9+0.5	81.7+1.0	82.6+0.7	0.52	138 G
Untied Cont. & Pos. Attention (16)	83.4+0.4	81.3+0.9	82.2+0.6	0.52	164 G
2-Stages (8)	84.3+0.4	82.3+0.4	83.1+0.3	0.94	283 G
Temp. Self-Attn (8)	83.8+0.6	81.9+1.0	82.6+0.6	0.55	277 G
Temp. & Spatio. Self-Attn (8)	**84.5+0.6**	**82.7+1.0**	**83.4+0.8**	0.95	332 G

Table 1: Ablation studies of core design choices in Exchanger on PASTIS validation dataset with 5-Fold cross-validation. The figure in parenthesis denotes the number of content/position queries used.

We first study the impact of several key design choices in Exchanger on PASTIS validation dataset compared to a strong baseline model where self-attention is employed to process

[2] https://github.com/VSainteuf/pastis-benchmark

[3] Please note that the individual samples in MTLCC have limited spatial resolutions of 24 × 24.

temporal and spatial features as done in TSViT [27]. As seen in Tab.1, not incorporating position queries results in the worst performance with around an absolute 4% drop compared to all other models, indicating date-specific temporal embeddings are key to capture crop phenological profiles. Instead of mixing the content and position information in attention calculation, adopting untied cont. & pos. attention as proposed in TUPE[13] slightly improves F1-Score by 0.2%, which is set to the default choice for all the subsequent experiments, unless stated otherwise. Then we evaluate the performance of Exchanger w.r.t. the number of content & position tokens by increasing it from 4 to 8 to 16. As shown in Tab. 1, Exchanger has achieved the best scores across precision, recall and F1 metrics with 8 tokens. In contrast to the only 1 class token prepended to the input sequence in NLP, we hypothesize that requiring slightly more tokens for crop type recognition is due to the significant intra-class variation and multi-mode nature which we will show the latent embeddings in supplementary materials. Contradicting with fixing the number of tokens to that of classes needed to be identified in TSViT [27], we found that continually increasing the number of content/position queries did not bring the expected performance boost but with a noticeable increase in computational cost. When comparing untied cont. & pos. attention (8) with its self-attention counterpart (Temp. Self-Attn (8)), it shows that Exchanger can achieve nearly identical results with a similar number of parameters but with a drastic drop in computational cost (almost 50% saving in GFLOPs). Last, with stacking of two identical Exchanger blocks (2-Stages (8)), it reached a F1-Score of 83.1, which is on par with that obtained by Temp. & Spatio. Self-Attn (8) which is a modified TSViT[27] whilst being computationally-light (around 15% saving in GFLOPs). Additionally, the latter (Temp. & Spatio. Self-Attn (8)) can be seen as adding an attentive MIL pooling component [11] after the temporal self-attention block to identify key spatial instances. However, we have demonstrated solely increasing the depth of Exchanger can bring a similar performance boost, enjoying the advantage that it can be reused in downstream tasks rather than being discarded in TSViT[27] for dense prediction.

4.3 Comparison with SOTA

4.3.1 Semantic Segmentation

	mIoU (%)		#Params(M)	FLOPs
	PASTIS	MTLCC		
FPN + ConvLSTM[19]	57.1	73.7	1.45	714 G
Unet + ConvLSTM[18]	57.8	76.2	2.33	55 G
Unet-3D[18]	58.4	75.2	1.55	92G
U-TAE[7]	63.1	77.1	1.09	47 G
TSViT[27]	65.4	84.8	2.16	558 G
Exchanger+Unet	66.8(+1.2)	**90.7**	8.08	300 G
Exchanger+Mask2Former	**67.9(+1.2)**	90.5	24.59	329 G

Table 2: Comparison with SOTA models on PASTIS and MTLCC test dataset. The figure in parenthesis denotes the standard deviation across the official 5-Fold splits in PASTIS. FLOPs are calculated based on a single SITS sample with $T \times C \times H \times W = 30 \times 10 \times 128 \times 128$.

As shown in Tab. 2, coupling the Exchanger which serves as a pure temporal encoder with a plain Unet [23] which exclusively focuses on spatial semantic mining has easily led to 66.8% and 90.7% mIoU on PASTIS and MTLCC, surpassing the previous state-of-the-art results attained by TSViT[27] by 1.4 and 5.9 points respectively while only using 53% FLOPs. The dissociation between temporal and spatial components further allows us to explore the potential of adopting the recently proposed powerful universal image segmentation framework Mask2Former[1] with PVT2[30] as backbone and FPN[16] as the pixel decoder, resulting in a significant improvement of around an absolute 2.5% compared to the best results reported in the literature and a boost of about 1.1% compared to Exchanger+Unet but only with less than 10% increase in the computational cost. It is notable that all previous semantic segmentation models for SITS except for TSViT[27] feature a complicated composition of spatial and temporal components, hindering them from leveraging the latest advances in CV. Although TSViT[27] is the first fully-attentional neural architecture for SITS processing, it faces extra obstacles when deployed in the pretrain-finetune paradigm because of the patch tokenization layer which prevents it from being directly operated on the pixel-set format, and the self-attention operation can incur prohibitive computational cost for dense prediction tasks. Another marked fact is that the temporal-then-spatial processing order, which has been demonstrated is a more desirable inductive bias[27] for SITS analysis, would cause the temporal encoder to consume a drastic proportion of the requested computation, e.g., the Exchanger accounts for nearly 96% of the total computational cost in Exchanger+Unet. And it should be pointed out that our proposed model only has a linear computational complexity $\mathcal{O}(NTd)$ w.r.t. the input sequence length.

4.3.2 Panoptic Segmentation

	SQ	RQ	PQ	#Params(M)	FLOPs	IT(s)
Unet+ConvLSTM+PaPs [7]	80.2	43.9	35.6	2.50	55 G	660
U-TAE+PaPs [7]	81.5	53.2	43.8	1.26	47 G	207
Exchanger+Unet+PaPs	80.3(+0.1)	58.9(+0.6)	47.8(+0.4)	9.99	301 G	252
Exchanger+Mask2Former	**84.6(+0.9)**	**61.6(+1.6)**	**52.6(+1.8)**	24.63	332 G	**154**

Table 3: Comparison with SOTA models on PASTIS test dataset. The figure in parenthesis denotes the standard deviation across the official 5-Fold splits in PASTIS. FLOPs are calculated based on a single SITS sample with $T \times C \times H \times W = 30 \times 10 \times 128 \times 128$. Inference Time (IT) is calculated on Fold-1 with ≈ 490 sequences on a single A100 GPU.

To further demonstrate the effectiveness of our proposed representation learning framework, we tested its performance on the panoptic segmentation task [14] on PASTIS, which unifies semantic and instance segmentation into a joint task and therefore delivers a holistic scene understanding vision system. Despite the pioneering effort made in [7] where a single-stage instance segmentation network CenterMask[31] has been adapted to a panoptic segmentation module named Parcels-as-Points (PaPs), the task still remains extremely difficult as the majority of existing panoptic segmentation networks proposed for natural images or videos is not particularly effective for directly processing SITS. We argue that a strong temporal encoder is key to extracting high-level semantics from SITS, converting the low signal-to-noise ratio 4-D satellite data $T \times C \times H \times W$ to rich semantic 3-D feature maps $C \times H \times W$, which can be fed into off-the-shelf panoptic segmentation models. We report the class-averaged Segmentation Quality (SQ), Recognition Quality (RQ), and Panoptic Quality [4] (PQ) in Tab.3. It can be seen that Exchanger, equipped with Unet [23] as the spatial encoder and the PaPs module[7] for panoptic prediction, has increased RQ and PQ by a significant margin of 5.7% and 4.0%, respectively, compared to U-TAE+PaPs. Furthermore, it is prominent to see that Exchanger combined with Mask2Former[1] consistently outperforms Exchanger+Unet+PaPs by 4.3, 2.7 and 4.8 points in SQ, RQ, and PQ, respectively, setting a new state-of-the-art. Besides, it is noticeable that the required inference time on A100 GPU for Exchanger+Mask2Former is much lower because of the streamlined pipeline and high parallelizability.

5 Conclusion

To conclude, in this paper, we first present a unique reformulation of SITS representation as sets of instances, which relaxes the constraints caused by traditional spatiotemporal grids and further enables designing models that can flexibly process both pixel-set and image sequence format of SITS. Then, we propose to explicitly decompose the representation learning procedure of SITS into three steps: collect–update–distribute, resulting in a conceptually clear and computationally efficient feature learning framework called Exchanger. Facilitated by the previous two innovations, we have demonstrated for the first time the successful transfer of pretrain-finetune paradigm from CV to SITS, leading to a streamlined semantic & panoptic segmentation pipeline and marked performance gains over the previous SOTA models.

6 Acknowledgements

The work was supported by Department for the Economy (DfE) international studentship at Ulster University (UU). All the experiments presented in the paper were performed at the High Performance Computing (HPC) Centre at UU. We appreciate the constructive and insightful comments of the reviewers.

References

[1] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[4] Note that we follow the evaluation protocol in [7] where the calculation of PQ only involves thing classes.

[4] Yutong Feng, Jianwen Jiang, Mingqian Tang, Rong Jin, and Yue Gao. Rethinking supervised pre-training for better downstream transferring. *arXiv preprint arXiv:2110.06014*, 2021.

[5] Anatol Garioud, Stéphane Peillet, Eva Bookjans, Sébastien Giordano, and Boris Wattrelos. Flair: French land cover from aerospace imagery., 2022.

[6] Vivien Sainte Fare Garnot and Loic Landrieu. Lightweight temporal self-attention for classifying satellite images time series. In *Advanced Analytics and Learning on Temporal Data: 5th ECML PKDD Workshop, AALTD 2020, Ghent, Belgium, September 18, 2020, Revised Selected Papers 6*, pages 171–181. Springer, 2020.

[7] Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4872–4881, 2021.

[8] Vivien Sainte Fare Garnot, Loic Landrieu, Sebastien Giordano, and Nesrine Chehata. Satellite image time series classification with pixel-set encoders and temporal self-attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12325–12334, 2020.

[9] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[10] Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten Borgwardt. Set functions for time series. In *International Conference on Machine Learning*, pages 4353–4363. PMLR, 2020.

[11] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

[12] Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon. Tile2vec: Unsupervised representation learning for spatially distributed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3967–3974, 2019.

[13] Guolin Ke, Di He, and Tie-Yan Liu. Rethinking positional encoding in language pre-training. *arXiv preprint arXiv:2006.15595*, 2020.

[14] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019.

[15] Lukas Kondmann, Aysim Toker, Marc Rußwurm, Andres Camero Unzueta, Devis Peressuti, Grega Milcinski, Nicolas Longépé, Pierre-Philippe Mathieu, Timothy Davis, Giovanni Marchisio, et al. Denethor: The dynamicearthnet dataset for harmonized, inter-operable, analysis-ready, daily crop monitoring from space. In *35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, pages 1–13, 2021.

[16] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[17] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[18] Rose M Rustowicz, Robin Cheong, Lijing Wang, Stefano Ermon, Marshall Burke, and David Lobell. Semantic segmentation of crop type in africa: A novel dataset and analysis of deep learning methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 75–82, 2019.

[19] Jorge Andres Chamorro Martinez, Laura Elena Cué La Rosa, Raul Queiroz Feitosa, Ieda Del'Arco Sanches, and Patrick Nigri Happ. Fully convolutional recurrent networks for multidate crop recognition from multitemporal image sequences. *ISPRS Journal of Photogrammetry and Remote Sensing*, 171:188–201, 2021.

[20] Joachim Nyborg, Charlotte Pelletier, and Ira Assent. Generalized classification of satellite image time series with thermal positional encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1392–1402, 2022.

[21] Charlotte Pelletier, Geoffrey I Webb, and François Petitjean. Temporal convolutional neural network for the classification of satellite image time series. *Remote Sensing*, 11 (5):523, 2019.

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[24] Marc Rußwurm and Marco Körner. Convolutional lstms for cloud-robust segmentation of remote sensing imagery. *arXiv preprint arXiv:1811.02471*, 2018.

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

[26] Satya Narayan Shukla and Benjamin M Marlin. Interpolation-prediction networks for irregularly sampled time series. *arXiv preprint arXiv:1909.07782*, 2019.

[27] Michail Tarasiou, Erik Chavez, and Stefanos Zafeiriou. Vits for sits: Vision transformers for satellite image time series. *arXiv preprint arXiv:2301.04944*, 2023.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[29] Francesco Vuolo, Martin Neuwirth, Markus Immitzer, Clement Atzberger, and Wai-Tim Ng. How much does multi-temporal sentinel-2 data improve crop type classification? *International journal of applied earth observation and geoinformation*, 72: 122–130, 2018.

[30] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvtv2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):1–10, 2022.

[31] Yuqing Wang, Zhaoliang Xu, Hao Shen, Baoshan Cheng, and Lirong Yang. Centermask: single shot instance segmentation with point representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9313–9321, 2020.

[32] Gerald Woo, Chenghao Liu, Doyen Sahoo, Akshat Kumar, and Steven Hoi. Etsformer: Exponential smoothing transformers for time-series forecasting. *arXiv preprint arXiv:2202.01381*, 2022.

[33] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? *arXiv preprint arXiv:2205.13504*, 2022.

[34] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning*, pages 27268–27286. PMLR, 2022.