Revisiting the Encoding of Satellite Image Time Series

Xin Cai[1] Yaxin Bi[1] Peter Nicholl[1] Roy Sterritt[1]

[1]Artificial Intelligence Research Centre (AIRC), Ulster University (UU), Belfast, UK

BMVC 2023

Background: Satellite Image Time Series (SITS)

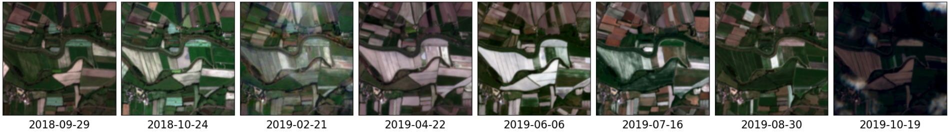
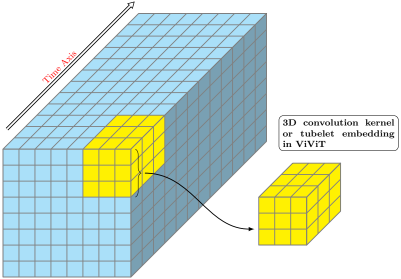
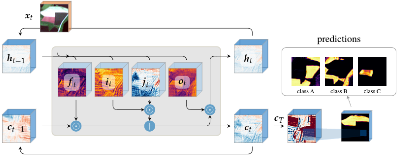


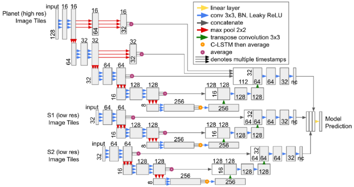
Figure: SITS from Sentinel-2



(a) Local Smoothness of Video Signals



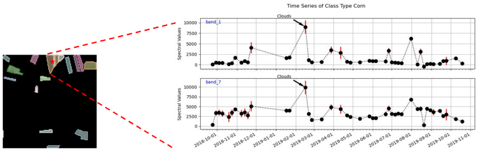
(b) CLSTM



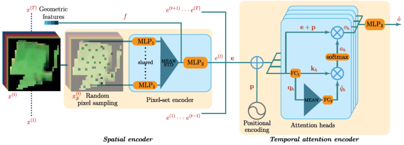
(c) 2D U-Net+CLSTM

- Arnab, Anurag, *et al.*"Vivit: A video vision transformer." CVPR. 2021.
- Rußwurm, Marc, *et al.*"Convolutional LSTMs for cloud-robust segmentation of remote sensing imagery." arXiv preprint (2018).
- M Rustowicz, Rose, *et al.*"Semantic segmentation of crop type in Africa: A novel dataset and analysis of deep learning methods." CVPR Workshops.2019.

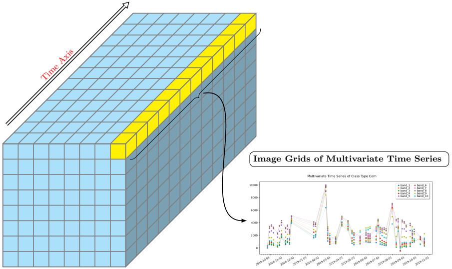
Background: SITS or TSSI?



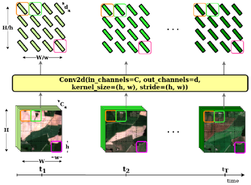
(a) Multivariate Time Series



(b) PSE+TAE



(c) TSSI = Image Grids of TS



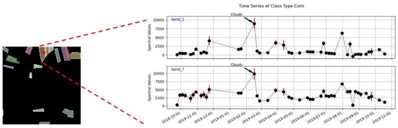
(d) TSViT

- Garnot, Vivien Sainte Fare, *et al.*"Satellite image time series classification with pixel-set encoders and temporal self-attention." CVPR. 2020.
- Tarasiou, Michail, *et al.*"ViTs for SITS: Vision Transformers for Satellite Image Time Series." CVPR. 2023.

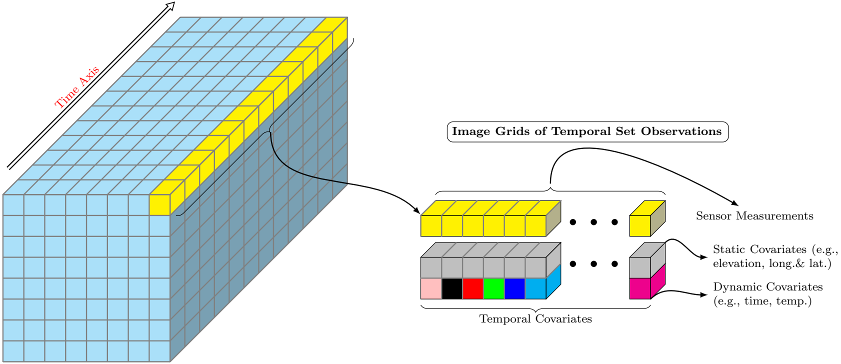
Motivation: SITS (TSSI) is a **NEW** data modality

- Two data storage format for SITS: pixel-set format ($T \times C \times N$) and image sequence format ($T \times C \times H \times W$) (*e.g.*, PSE+TAE only works with pixel-set format, and TSViT treats SITS as image sequences)
- Pixel-set format is a resource-efficient format for pre-training
- Characteristics of the temporal dimension of SITS: Irregularity & Asynchronization
- Do we really need to build bespoke neural architectures for SITS?

Method: SITS (TSSI) = Image Grids of Temporal Set Observations



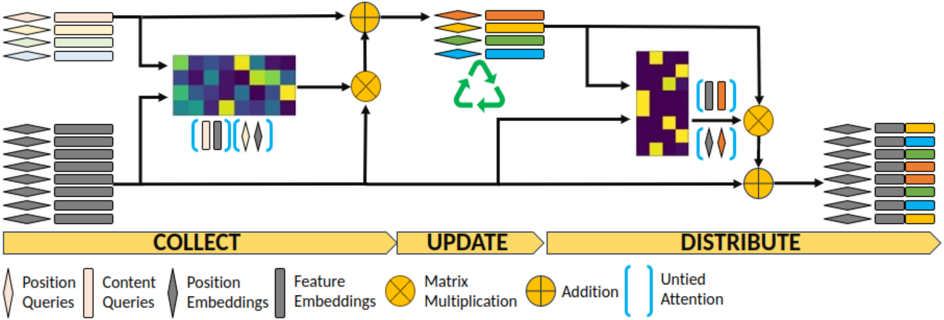
(a) TSSI



(b) Reformulation of SITS representation

Method: A Novel Learning Paradigm – Exchanger

- Self-attention is not suitable for modelling complex temporal relations in Time Series.
- Irregularity & Asynchronization in the temporal axis.



(a) The schematic illustration of the proposed collect–update–distribute process for generic representation learning of SITS.

- Zeng, Ailing, *et al.*"Are transformers effective for time series forecasting?." AAAI 2023.

Method: A Specific Instantiation

▷ **COLLECT**

$$\boldsymbol{C}^v = \mathrm{Concat}_h \left(\mathrm{Softmax} \left(\frac{1}{\sqrt{2d}} \boldsymbol{C}^v \boldsymbol{W}_h^Q \left(\boldsymbol{V} \boldsymbol{W}_h^K \right)^T + \frac{1}{\sqrt{2d}} \boldsymbol{C}^p \boldsymbol{U}_h^Q \left(\boldsymbol{P} \boldsymbol{U}_h^K \right)^T \right) \boldsymbol{V} \boldsymbol{W}_h^V \right) \quad (1)$$

▷ **UPDATE**

$$\begin{aligned} \boldsymbol{C}^v &= \boldsymbol{C}^v + \mathrm{MLP}_1 \left(\mathrm{LayerNorm} \left(\boldsymbol{C}^v \right)^T \right)^T \\ \boldsymbol{C}^v &= \boldsymbol{C}^v + \mathrm{MLP}_2 \left(\mathrm{LayerNorm} \left(\boldsymbol{C}^v \right) \right) \end{aligned} \quad (2)$$

▷ **DISTRIBUTE**

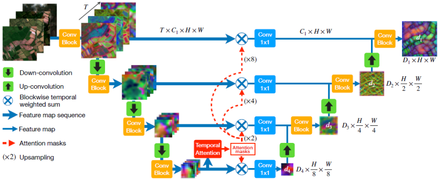
$$\begin{aligned} \boldsymbol{Z} &= \mathrm{Concat}_h \left(\mathrm{Softmax} \left(\frac{1}{\sqrt{2d}} \boldsymbol{V} \tilde{\boldsymbol{W}}_h^Q \left(\boldsymbol{C}^v \tilde{\boldsymbol{W}}_h^K \right)^T + \frac{1}{\sqrt{2d}} \boldsymbol{P} \tilde{\boldsymbol{U}}_h^Q \left(\boldsymbol{C}^p \tilde{\boldsymbol{U}}_h^K \right)^T \right) \boldsymbol{C}^v \tilde{\boldsymbol{W}}_h^V \right) \\ \boldsymbol{Z}' &= \mathrm{Concat} \left(\boldsymbol{Z}, \boldsymbol{V} \right) \tilde{\boldsymbol{W}}_{proj} \\ \boldsymbol{V}' &= \boldsymbol{Z}' + \mathrm{FFN} \left(\boldsymbol{Z}' \right) \end{aligned} \quad (3)$$

- Yang, Chenhongyi, et al. "GPViT: A High Resolution Non-Hierarchical Vision Transformer with Group Propagation." arXiv preprint (2022).

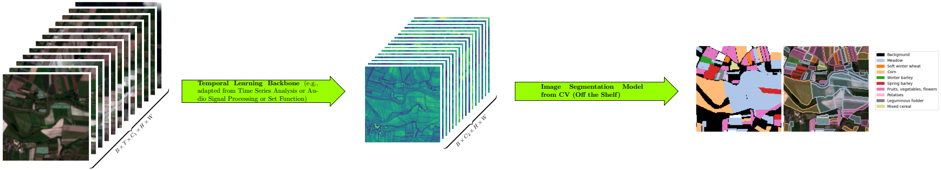
Method: Distinct Advantages

- subsumes PSE+TAE and TSViT as special cases
- works well both with the pixel-set ($T \times C \times N$) and image sequence ($T \times C \times H \times W$) format
- a resource-efficient pretrain (pixelset format) -finetune (image sequence format) paradigm for SITS
- linear computational complexity w.r.t. the length of input sequence
- streamlined dense prediction pipeline of SITS

Method: SITS is no longer an isolated island



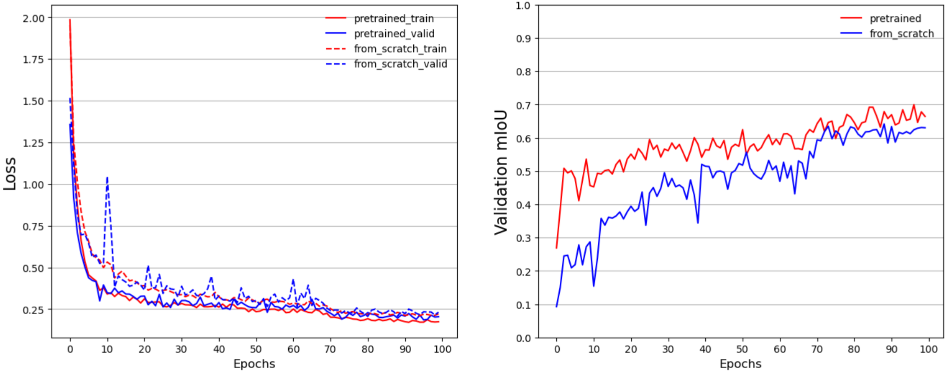
(a) Previous dense prediction pipeline of SITS.



(b) Streamlined dense prediction pipeline of SITS.

- Garnot, Vivien Sainte Fare, *et al.*"Panoptic segmentation of satellite image time series with convolutional temporal attention networks." CVPR. 2021.

Experimental Results: Convergent Analysis



(a) Convergence analysis for Exchanger+Unet with pre-trained backbones or training from scratch on PASTIS validation dataset (Fold-1). The left figure shows the training and validation losses. The right figure shows the evaluation metric mIoU on the validation dataset.

- Exchanger+Mask2Former cannot be trained completely from scratch.

Experimental Results: Semantic Segmentation

	mIoU (%)		#Params(M)	FLOPs
	PASTIS	MTLCC		
FPN + ConvLSTM	57.1	73.7	1.45	714 G
Unet + ConvLSTM	57.8	76.2	2.33	55 G
Unet-3D	58.4	75.2	1.55	92G
U-TAE	63.1	77.1	1.09	47 G
TSViT	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: 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 $T \times C \times H \times W = 30 \times 10 \times 128 \times 128$.

Experimental Results: Panoptic Segmentation

	SQ	RQ	PQ	#Params(M)	FLOPs	IT(s)
Unet+ConvLSTM+PaPs	80.2	43.9	35.6	2.50	55 G	660
U-TAE+PaPs	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: 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 $\approx$ 490 sequences on a single A100 GPU.

Experimental Results: Visualisations

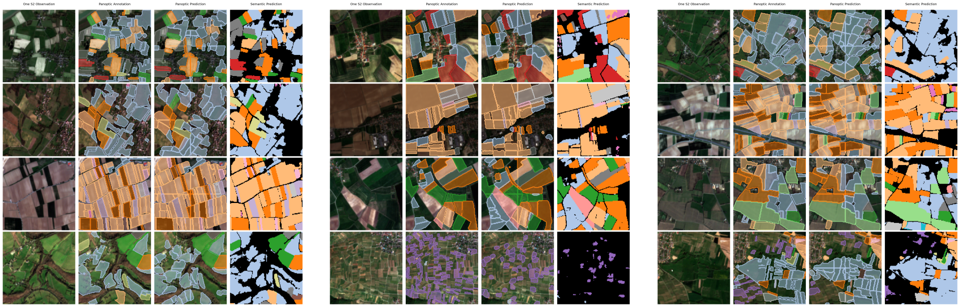


Figure: Qualitative results from predictions of Exchanger+Mask2Former. Please note the semantic & panoptic segmentation models are separately trained.

Conclusion

- reformulate SITS representation as image grids of temporal set observations
- explicitly decompose the representation learning procedure of SITS into three steps: collect–update–distribute
- the successful introduction of resource-efficient pretrain-finetune paradigm into SITS for the first time
- a streamlined dense prediction pipeline and marked performance gains over the previous SOTA models

Thanks for listening