Development of Deep Satellite Imagery Analytics for Vegetation and Urban Growth Monitoring

Initial Assessment Presentation

Xin Cai

Supervisors: Dr. Yaxin Bi, Dr. Peter Nicholl, Mr. Roy Sterritt

14th December 2021

Outline

Project Background

The Overall Research Aim & Research Objectives

Progress To Date & Research and Training Activities Planned

Project Background

- Remote sensing is entering a new era where modern satellites monitor the Earth surface in ever-shorter time intervals and ever-increasing spatial resolution.
- Copernicus program by the European Space Agency (ESA)
- Landsat program in the U.S.

The Overall Research Aim & Specific Research Objectives

- The overall aim of my PhD research is to develop deep learning algorithms that can exploit satellite imagery time series (SITS) with a particular focus on vegetation and urban growth monitoring.

  i. Exploiting the spatiotemporal structural information in raw SITS.
  ii. Employing unsupervised learning methods to distil transferable feature representations.
  iii. Exploring unsupervised domain adaptation (UDA) techniques to improve the generalization capability of deep learning models to unseen testing scenarios.

Spatiotemporal Structural Information

- Conclusion: The existing deep learning models developed for crop type classification are not designed to take advantage of the dense temporal density of SITS, especially in combination with the high spatial resolution.

Tables Credit: from the paper DENETHOR

Spatial Encoder	Temporal Encoder			
	TempCNN [27]	MSResNet [44]	LSTM [34]	Transformer [36]
ResNet18 [14]	52.22%	49.53%	44.64%	43.61%
SqueezeNet [17]	53.94%	49.78%	35.89%	42.58%
MobileNetv3 [16]	53.20%	54.33%	43.46%	48.06%
Pixel Average [34]	64.46%	58.83%	48.40%	52.56%
	Pixel-Set Encoding + Self-Attention			
PseTae [12]	**67.25%**			
PseTae [38]	64.95%			
	Ablation Scores			
PseTae (2018)	78.77%			
PseTae (Val)	88.02%			

Data Type	# Features	Accuracy	Macro F1-Score
Sentinel 1 (S1)	42	0.58	0.43
Sentinel 2 (S2)	91	0.59	0.42
Planet (PL)	35	0.37	0.12
S1 + S2	133	0.62	**0.46**
S1 + PL	77	0.60	0.42
S2 + PL	126	0.59	0.41
S1 + S2 + PL	168	**0.63**	**0.46**

• L. Kondmann, A. Toker, M. Rußwurm, A. Camero, D. Peressuti, G. Milcinski, P.-P. Mathieu, N. Longépé, T. Davis, G. Marchisio et al., "Denethor: The dynamic earthnet dataset for harmonized, inter-operable, analysis-ready, daily crop monitoring from space," NeurIPS Track on Datasets and Benchmarks, 2021.

Unsupervised Representation Learning

- Disentangled Representation Learning
  - Image-to-Image (I2I) Translation
- Deep Clustering Methods
  - Adapting traditional clustering methods for being differentiable
- Deep Generative Models
  - GANs, VAEs, Flow-based Models, ……
- Self-Supervised Learning (SSL)
  - Pretext tasks, Contrastive Learning, ……

Unsupervised Domain Adaptation (UDA)

Domain Adaptation refers to developing algorithms that can generalize well to the target domain by training models on a semantic related but distribution different source domain.

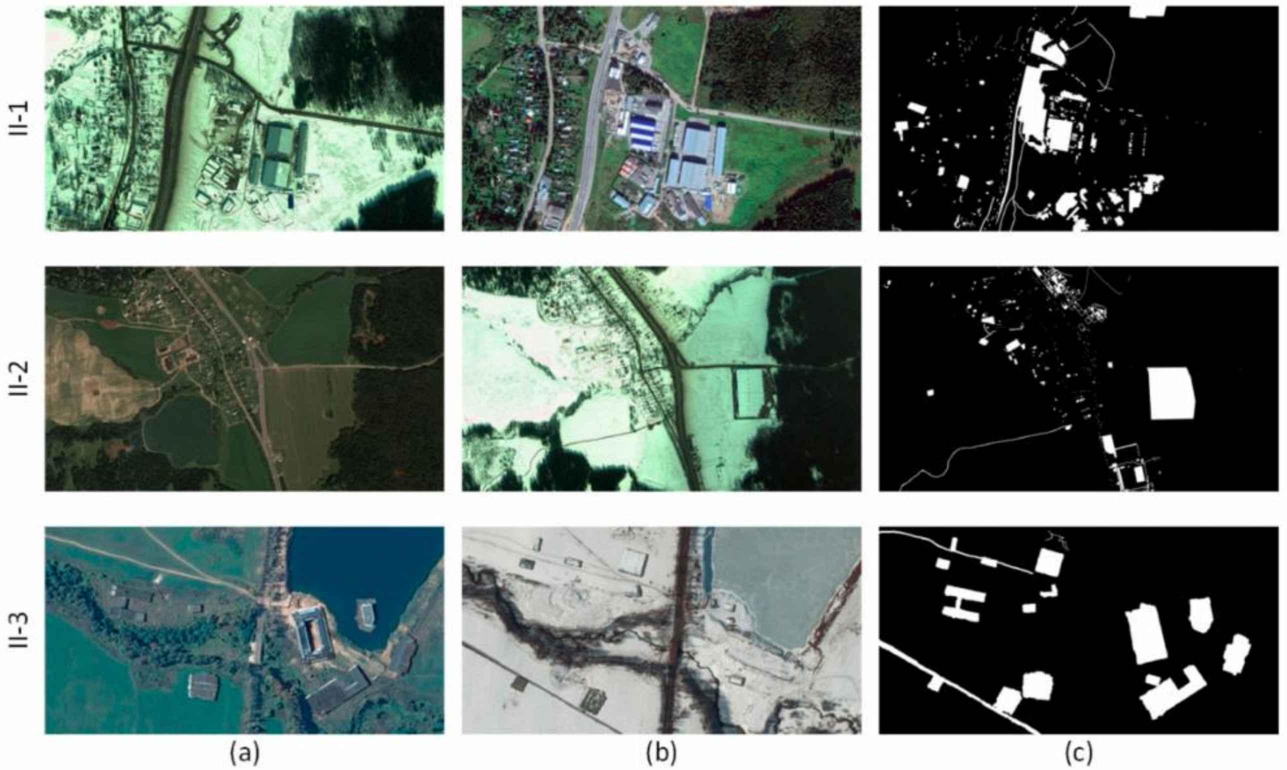
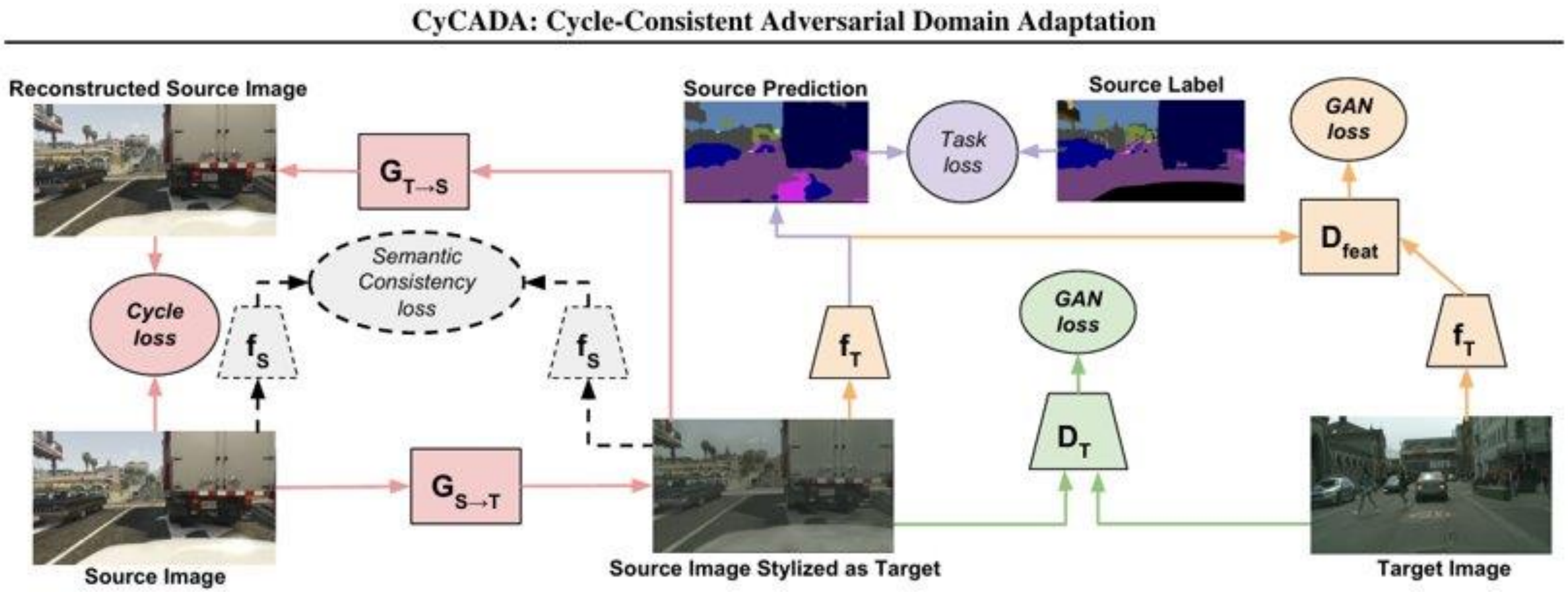


Figure 7. Overview of dataset II with major seasonal variation and corresponding reference change maps: (**a**) pre-event images, (**b**) post-event images, (**c**) references.

- Kou, Rong, et al. "Progressive Domain Adaptation for Change Detection Using Season-Varying Remote Sensing Images." *Remote Sensing* 12.22 (2020): 3815.

Adversarial Domain Adaptation



- Hoffman, Judy, et al. "Cycada: Cycle-consistent adversarial domain adaptation." *International conference on machine learning*. PMLR, 2018.

Recently-released Large-scale Benchmark Datasets in the Remote Sensing Community

- **DENETHOR** (*NeurIPS 2021, Vegetation Monitoring*)
- **Multi-Temporal Urban Development SpaceNet Dataset (MUDS)** (*CVPR 2021, Urban Growth Monitoring*)

• L. Kondmann, A. Toker, M. Rußwurm, A. Camero, D. Peressuti, G. Milcinski, P.-P. Mathieu, N. Longépé, T. Davis, G. Marchisio et al., "Denethor: The dynamic earthnet dataset for harmonized, inter-operable, analysis-ready, daily crop monitoring from space," NeurIPS Track on Datasets and Benchmarks, 2021.
• A. Van Etten, D. Hogan, J. M. Manso, J. Shermeyer, N. Weir, and R. Lewis, "The multi-temporal urban development spacenet dataset," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6398–6407.

Progress To Date

Literature Review

- Remote Sensing (e.g., Crop Type Mapping, Change Detection), Deep Generative Models, Deep Clustering Methods, Video Understanding, Time-Series Analysis, Transformers, … ...

$\mathbf{z} \sim p(\mathbf{z})$ to the desired data distribution $\mathbf{x} \sim p(\mathbf{x})$ based on the change of variable formula, provided that g is a bijective function:

$$p_{\mathbf{X}}(x) = p_{\mathbf{Z}}(\mathbf{z}) \left| det\left(\frac{\partial g(\mathbf{z})}{\partial \mathbf{z}^T}\right) \right|^{-1}$$

To facilitate efficient computations of Jacobian determinants and the inverse functions, the researchers proposed affine coupling layers based on the simple observation that the determinant of a triangular matrix can be efficiently computed as the product of its diagonal entries.

$$y_{1:d} = x_{1:d}$$

$$y_{d+1:D} = x_{d+1:D} \odot exp\left(s\left(x_{1:d}\right)\right) + t\left(x_{1:d}\right)$$

Then, they proposed two conditioning schemes: 1. spatial checkerboard masks; and 2. channel-wise masks. At last, they stack a series of these bijective functions having the above-mentioned functional form to compose a deep model.
It is still necessary to dive deep into code to gain a further understanding.

- **Glow: Generative Flow with Invertible 1×1 Convolutions**
PaperLink: Paper
CodeLink: glow
Reference Value: ★★★★★
NeurIPS
Classic Work in Flow-based Generative Models

Table 1: The three main components of our proposed flow, their reverses, and their log-determinants. Here, $\mathbf{x}$ signifies the input of the layer, and $\mathbf{y}$ signifies its output. Both $\mathbf{x}$ and $\mathbf{y}$ are tensors of shape $[h \times w \times c]$ with spatial dimensions (h, w) and channel dimension c. With (i, j) we denote spatial indices into tensors $\mathbf{x}$ and $\mathbf{y}$. The function NN() is a nonlinear mapping, such as a (shallow) convolutional neural network like in ResNets (He et al., 2016) and RealNVP (Dinh et al., 2016).

Description	Function	Reverse Function	Log-determinant
Actnorm. See Section 3.1	$\forall i,j: \mathbf{y}_{i,j} = \mathbf{s} \odot \mathbf{x}_{i,j} + \mathbf{b}$	$\forall i,j: \mathbf{x}_{i,j} = (\mathbf{y}_{i,j} - \mathbf{b})/\mathbf{s}$	$h \cdot w \cdot \mathtt{sum}(\log\|\mathbf{s}\|)$
Invertible 1×1 convolution. $\mathbf{W}: [c \times c]$. See Section 3.2	$\forall i,j: \mathbf{y}_{i,j} = \mathbf{W}\mathbf{x}_{i,j}$	$\forall i,j: \mathbf{x}_{i,j} = \mathbf{W}^{-1}\mathbf{y}_{i,j}$	$h \cdot w \cdot \log\|\det(\mathbf{W})\|$ or $h \cdot w \cdot \mathtt{sum}(\log\|\mathbf{s}\|)$ (see eq. (10))
Affine coupling layer. See Section 3.3 and (Dinh et al., 2014)	$\mathbf{x}_a, \mathbf{x}_b = \mathtt{split}(\mathbf{x})$ $(\log \mathbf{s}, \mathbf{t}) = \mathtt{NN}(\mathbf{x}_b)$ $\mathbf{s} = \exp(\log \mathbf{s})$ $\mathbf{y}_a = \mathbf{s} \odot \mathbf{x}_a + \mathbf{t}$ $\mathbf{y}_b = \mathbf{x}_b$ $\mathbf{y} = \mathtt{concat}(\mathbf{y}_a, \mathbf{y}_b)$	$\mathbf{y}_a, \mathbf{y}_b = \mathtt{split}(\mathbf{y})$ $(\log \mathbf{s}, \mathbf{t}) = \mathtt{NN}(\mathbf{y}_b)$ $\mathbf{s} = \exp(\log \mathbf{s})$ $\mathbf{x}_a = (\mathbf{y}_a - \mathbf{t})/\mathbf{s}$ $\mathbf{x}_b = \mathbf{y}_b$ $\mathbf{x} = \mathtt{concat}(\mathbf{x}_a, \mathbf{x}_b)$	$\mathtt{sum}(\log(\|\mathbf{s}\|))$

The work [18] extends RealNVP to a powerful generative model on par with the performance of GANs, which is capable of producing realistic-looking images but only relying on maximizing the log-likelihood. The paper is clearly presented, and the results are very impressive.
Given the fact that different generative models (GANs, VAEs, NFs)

- **Swin Transformer: Hierarchical Vision Transformer using Shifted Windows**
PaperLink: Paper
CodeLink: Swin-Transformer
Reference Value: ★★★★★
ICCV

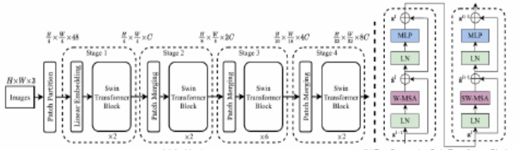


Figure 3. (a) The architecture of a Swin Transformer (Swin-T); (b) two successive Swin Transformer Blocks (notation presented with Eq. (3)). W-MSA and SW-MSA are multi-head self attention modules with regular and shifted windowing configurations, respectively.

The work [22] proposed to build a general-purpose backbone for computer vision tasks by adapting the prevalent Transformer architecture in NLP. As pointed out in the paper, there are challenges needing to be overcome for the adaptation due to significant differences between CV and NLP, e.g., visual entities can very substantially in scales, the much higher resolution of pixels in images compared to words in texts, especially for those dense prediction tasks. As a result, the authors proposed to use patch merging layers to build hierarchical representations, window partitioning scheme to ensure linear complexity of computations for self-attention, and shifted window strategy to enable connections among different windows. The experiments are extensive and can well support their claims and proposed modifications.

- **Vision Transformers for Dense Prediction**
PaperLink: Paper

- **Learning Disentangled Representations of Satellite Image Time Series**
PaperLink: Paper
CodeLink: N.A.
Reference Value: ★★★★☆
ECML PKDD

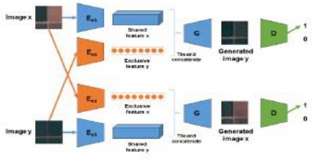


Figure 1. Model overview. The model goal is to learn both image transitions: $\mathbf{x} \rightarrow \mathbf{y}$ and $\mathbf{y} \rightarrow \mathbf{x}$. Both images are passed through the network $\mathbf{E}_{sh}$ in order to extract their shared representations. On the other hand, the network $\mathbf{E}_{ex}$ extracts the exclusive representations corresponding to images $\mathbf{x}$ and $\mathbf{y}$. The exclusive representation encoder output is constrained to follow a standard normal distribution. In order to generate the image $\mathbf{y}$, the decoder network $\mathbf{G}$ takes the shared feature of image $\mathbf{x}$ and the exclusive feature of image $\mathbf{y}$. A similar procedure is performed to generate the image $\mathbf{x}$. Finally, the discriminator $\mathbf{D}$ is used to evaluate the generated images.

Figure 4. Multimodal generation. The first row corresponds to a time series sampled from the test dataset. The second row corresponds to a time series where each image is generated by using the same shared feature and only modifying the exclusive feature.

The work [30] proposed to employ multi-modal Image-to-Image (I2I) translation techniques to extract disentangled representations from satellite imagery time series (SITS) in an unsupervised manner. The main idea is borrowed from BicycleGAN and cross-domain representation disentanglement. Specifically, the researchers postulated SITS can be decomposed into the shared component that captures common features existing in time series data and the exclusive components which contain specific information for each image. As a result, images acquired at different times but on the same geographical locations can be used as supervisory signals to guide the unsupervised learning. Based on this pre-trained model, they further performed a series of downstream tasks, such as image retrieval, image classification, and image segmentation, demonstrating the powerfulness of the proposed unsupervised

Progress To Date

Basic Usage of SLURM-based HPC for Distributed Training

- SLURM commands, Configuration of Virtual Environments, Horovod, Torch.Distributed

Toolboxes & Packages for Processing EO Data

- geopandas, eo-learn, sentinelhub, radiant-mlhub, rasterio, fiona

Progress To Date

AI4FoodSecurity Challenge

- Building a preprocessing pipeline dedicated to EO data (using specialized toolboxes)
- Testing the combination of ResNet18 + TempCNN
- Testing the state-of-the-art crop classification model: PseLTae
- Testing FocalBCE Loss for imbalanced classes
- Modifying a recently proposed Vision Transformer Model (CPVT) for processing time-series data

- Garnot, Vivien Sainte Fare, and Loic Landrieu. "Lightweight Temporal Self-attention for Classifying Satellite Images Time Series." *International Workshop on Advanced Analytics and Learning on Temporal Data*. Springer, Cham, 2020.
- Chu, Xiangxiang, et al. "Conditional positional encodings for vision transformers." *arXiv preprint arXiv:2102.10882* (2021).

AI4FoodSecurity Challenge Leaderboard

Leaderboard

Your best submission will appear on the leaderboard.

Entries: 8

TEAM	SCORE
TCSA-AI Total submissions: 79 Last submission: 5 days ago	4.436
EagleEyes Total submissions: 9 Last submission: 1 hour ago	4.615
MEOTEQ Total submissions: 65 Last submission: 7 hours ago	4.795
Adrián Cal Total submissions: 26 Last submission: 16 hours ago	5.092
Fer Total submissions: 19 Last submission: 12 hours ago	5.373
Microcosm Total submissions: 8 Last submission: 2 days ago	5.396
AIRC_Ulster Total submissions: 1 Last submission: 37 days ago	8.692
vecxoz Total submissions: 5 Last submission: 10 days ago	14.509

Leaderboard

Your best submission will appear on the leaderboard.

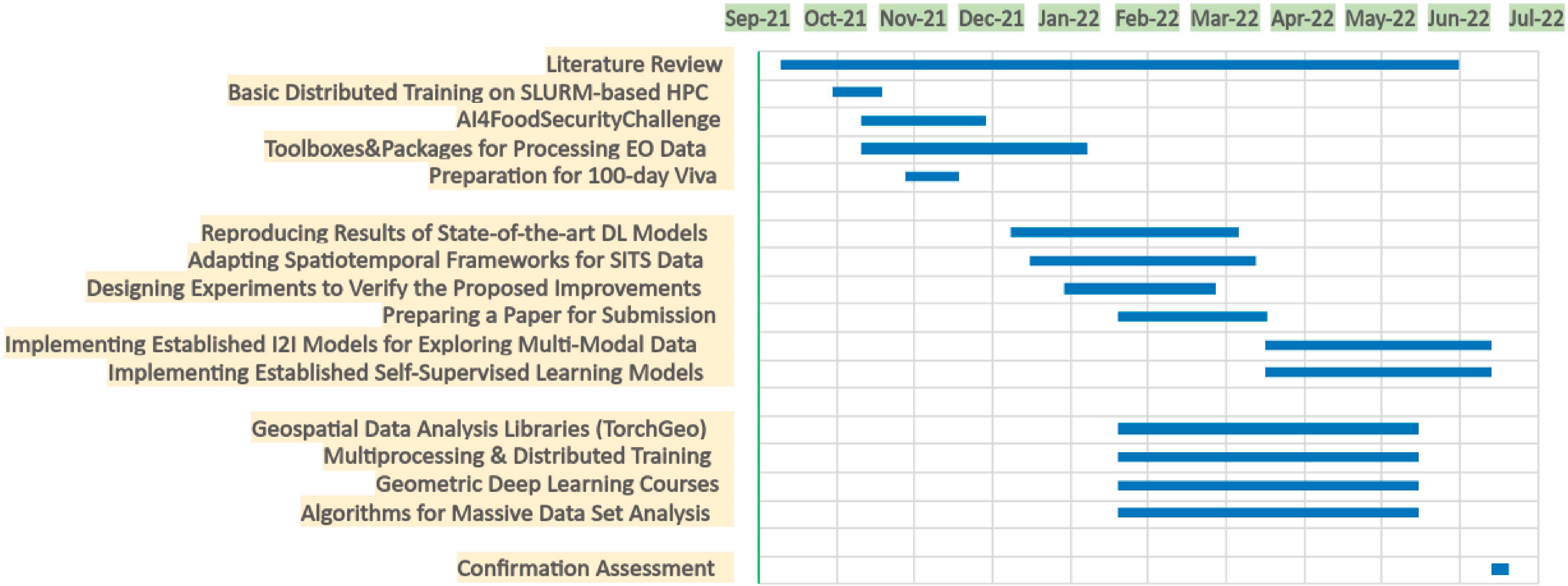
Entries: 8

TEAM	SCORE
EagleEyes Total submissions: 43 Last submission: 21 hours ago	3.828
MEOTEQ Total submissions: 83 Last submission: 18 hours ago	3.901
TCSA-AI Total submissions: 117 Last submission: 10 hours ago	3.908
Panopterra Total submissions: 37 Last submission: 15 hours ago	3.928
AIRC_Ulster Total submissions: 1 Last submission: 37 days ago	6.008
Ivan Total submissions: 12 Last submission: 1 hour ago	6.522
ma2okalab Total submissions: 9 Last submission: 8 days ago	7.015
vecxoz Total submissions: 4 Last submission: 10 days ago	13.584

Research Activities Planned

- **1st Research Focus – Spatiotemporal Learning (the primary focus in the following 10 months)**
- Investigate methods proposed for video understanding/video action recognition and other research fields which involve spatiotemporal learning to study how the spatiotemporal structural information can be exploited.
- Investigate methods proposed for multivariate time-series analysis to study how to deal with problems related to time-series data such as imputation and forecasting.
- Investigate methods proposed for 3D point cloud processing or Graph Convolutional Neural Networks (GCNs) to study how irregular data can be efficiently processed by deep learning models.
- Submit at least one paper to international conferences or journals before the confirmation assessment:
  - *Remote Sensing, IEEE Transactions on Geoscience and Remote Sensing,*
  - *IEEE International Geoscience and Remote Sensing Symposium (IGARSS),*
  - *IEEE Geoscience and Remote Sensing Magazine,*
  - *British Machine Vision Conference (BMVC)*

Summary of Progress To Date & Training Planned




thank
you