Efficient Representation Learning of Satellite Image Time Series and their Fusion for Spatiotemporal Applications

Poonam Goyal, Arshveer Kaur, Arvind Ram, Navneet Goyal

ADAPT Lab, Birla Institute of Technology and Science, Pilani
[poonam,p20170432,f20201210,goel]@pilani.bits-pilani.ac.in

Abstract

Satellite data bolstered by their increasing accessibility is leading to many endeavors of automated monitoring of the earth's surface for various applications. Such applications demand high spatial resolution images at a temporal resolution of a few days which entails the challenge of processing a huge volume of image time series data. To overcome this computing bottleneck, we present PatchNet, a bespoke adaptation of beam search and attention mechanism. PatchNet is an automated patch selection neural network that requires only a partial spatial traversal of an image time series and yet achieves impressive results. Satellite systems face a trade-off between spatial and temporal resolutions due to budget/technical constraints e.g., Landsat-8/9 or Sentinel-2 have high spatial resolution whereas, MODIS has high temporal resolution. To deal with the limitation of coarse temporal resolution, we propose FuSITSNet, a twofold feature-based generic fusion model with multimodal learning in a contrastive setting. It produces a learned representation after fusion of two satellite image time series leveraging finer spatial resolution of Landsat and finer temporal resolution of MODIS. The patch alignment module of FuSITSNet aligns the PatchNet processed patches of Landsat-8 with the corresponding MODIS regions to incorporate its finer resolution temporal features. The untraversed patches are handled by the cross-modality attention which highlights additional hot spot features from the two modalities. We conduct extensive experiments on more than 2000 counties of US for crop yield, snow cover, and solar energy prediction and show that even one-fourth spatial processing of image time series produces state-of-the-art results. FuSITSNet outperforms the predictions of single modality and data obtained using existing generative fusion models and allows for monitoring of dynamic phenomena using freely accessible images, thereby unlocking new opportunities.

Introduction

Satellite technology is extensively used to monitor the Earth's surface for different applications. Popular satellite systems that make their data publicly available include Landsat-8/9 (NASA 2016), Sentinel-2 (European Space Agency Signature 2017), and MODIS (NASA 2015). The last decade has witnessed a significant improvement in sensor technology leading to availability of higher spatial and temporal resolution satellite images. However, due to budgetary and technological constraints, it is not possible to capture satellite images with required high spatial and temporal resolutions using a single satellite system. This necessitates the development of efficient fusion algorithms that combine satellite image time series (SITS) from two satellites.

Many applications predicting crop yield, forest cover, forest fire (Gupta et al. 2023), etc. require SITS at high resolution along both spatial and temporal dimensions. Publicly available data from satellites like LANDSAT 8/9, SENTINEL-2, MODIS, etc. have high resolution only along one dimension, e.g., LANDSAT 8 has a spatial resolution of 30m and a 16-day revisit cycle whereas, MODIS has a spatial resolution of 250-500m and a revisit time of 8 days.

For high spatial resolution SITS, the amount of data we need to process increases manifolds, leading to a computing bottleneck. The amount of 7 years' data processed for 2000 counties considered in the paper is approximately 2.1 TB for MODIS and 10.0 TB for Landsat 8, respectively after applying the bits compression technique (Hubara et al. 2016).

The huge amount of data processing required seriously impedes the democratization of the use of satellite images for various applications. In this paper, we are addressing two major problems- 1) impractical computational requirements for processing high spatial resolution SITS & 2) dealing with coarse temporal resolution of high spatial resolution SITS.

For 1), we propose PatchNet which learns prominent patterns in a SITS by doing a spatial patch-based partial traversal, e.g., $(1/p)th$ spatial processing of SITS using the idea of beam search and attention mechanism for learnable patch selection. The learnable patch selection mechanism eliminates the need for full spatial processing of SITS, thereby reducing the amount of processing by a factor of p with some additional overheads and still achieves SOTA results for end tasks. Existing methods deal with the processing challenges by transforming the images into histograms (Sun et al.; You et al. 2017; Sun et al. 2020; Kaur et al. 2022). A few researchers have also tried to transform images into single-value numeric vegetation indices (Sakamoto 2020; Skakun et al. 2021; Ji et al. 2022; Choudhary et al. 2019). Both these approaches suffer from information loss.

For 2), we propose FuSITSNet, a twofold feature-based fusion model which can be used to fuse any two SITS. We applied it for Landsat-8 and MODIS SITS. FuSITSNet im-

proves the temporal features of Landsat SITS by aligning its PatchNet processed patches with the MODIS SITS. It takes care of the untraversed area of the time series by cross-modality attention which assimilates complementary features from two modalities. Its twofold feature fusion eliminates the need for image generation at mid-timestamps to increase the temporal resolution by a factor of 2. Our approach gives the learned representation directly from Landsat & MODIS SITS and thus does not increase the data volume for effectively increasing the temporal granularity. Code is available at (Poonam Goyal 2022)

Generative models such as Spatial and Temporal Adaptive Reflectance Fusion Model (STARFM) (Gao et al. 2006), Robust Spatiotemporal Fusion Network (RSFN) (Tan et al. 2022), and GAN ((Bouabid et al. 2020) have been used to generate Landsat-8 like images at higher temporal granularity. These models are constrained by the requirement of availability of Landsat & MODIS images captured on the same day and may also propagate existing noise. Moreover, the generation process is slow. If we use the generative models to enhance the temporal resolution of SITS, it increases the amount of data twofold and thus is computationally prohibitive. We used these models to generate Landsat-like images at improved temporal granularity of 8 days to create a baseline for comparisons.

FuSITSNet is set up in a contrastive learning framework (Fan, Zhang, and Gao 2020) which gives representation of SITS from both modalities. It can be applied to any two modalities having varied spatial, temporal, or spectral resolutions. The key contributions of the paper are as follows:

- To the best of our knowledge, this is the first attempt to efficiently process time series of high spatial resolution satellite images. We propose PatchNet which only needs to partially process the image time series using the concept of patches. The patch selection mechanism recommends most informative patches and achieves SOTA results for the end tasks considered.
- We also propose FuSITSNet, a twofold feature-based fusion model for fusing two image time series having different resolutions. We complementarily use a patch alignment module and cross-modality attention to learn high spatial resolution features of Landsat-8 and high temporal features of MODIS.
- We conduct extensive experiments to validate our models, PatchNet and FuSITSNet, for three applications – Crop yield prediction (CYP), snow cover prediction (SCP), and Solar energy prediction (SEP). The results of PatchNet are compared with those of existing models which use histogram time series and a significant improvement is observed. The direct feature-based learning from two SITS using FuSITSNet outperforms the enhanced SITS obtained from existing generative fusion models and all the baselines on single modality.

Related Work

Satellite data: Satellite systems like PlanetScope (planet 2019), CartoSat-1 (ISRO 2019), MODIS (NASA 2015), Landsat-8/9 (NASA 2016), Sentinel-2 (European Space Agency Signature 2017), and others are orbiting around the earth and collecting data at varying spatial, temporal and spectral resolutions. AVHRR has a coarse spatial resolution of 1km while PlanetScope, CartoSat-1 have a high resolution of 2-3m but their data is not freely available. Popular satellite systems are MODIS, Landsat-8/9, and Sentinel-2 due to their publicly available data which can be used in different real-world applications like disaster management, urban planning, agriculture, climate studies, etc. MODIS launched in 1999 provides data at a spatial resolution of 250-500m with a revisit time of daily or 8 days depending on the product. Landsat-8/9, launched in 2013/2021, has a spatial resolution of 30m with a revisit time of 16 days, and Sentinel-2 launched in 2015 has a spatial resolution of 20m with a revisit time of 10/5 days.

Spatiotemporal Applications: We consider applications viz. CYP, SCP, and SEP. These applications abide by the permutation invariant property where value of a pixel contributes to the end task irrespective of its position in the image (You et al. 2017). Accurate CYP is crucial for ensuring food security around the globe. Researchers have tried to predict crop yield with climate data (Fan et al. 2022; Verma et al. 2016; de Wit, Duveiller, and Defourny 2012; Guruprasad, Saurav, and Randhawa 2019) using traditional machine learning models. These models lack in capturing complex relationships between meteorological attributes and yield. A few researchers applied deep learning models and incorporated genotype (Khaki and Wang 2019; Måløy et al. 2021) and/or soil (Sun et al. 2020; Kaur et al. 2023) information. Recent studies attempt to include physics-guided patterns (He et al. 2023), and topological features (Jiang et al. 2022) along with climate data. Research shifted from meteorological data to the use of satellite image data after getting easy access to it. However, it is difficult to process image data due to its high volume. Therefore, vegetation indices are directly computed from MODIS product MOD13Q1 for a location. (Sun et al., 2020) and (You et al. 2017) converted MODIS images into histograms and used histogram time series to predict crop yield. Authors (Kaur et al. 2022) presented a deep learning model for MODIS, Landsat, and Sentinel histogram time series and fusion model (Kaur, Goyal, and Goyal 2023) to predict crop yield and highlighted the importance of high spatial and high temporal resolution of data required for the application. However, researchers faced data scarcity for training models using high-resolution satellites.

The other two applications have gained interest only recently, and very little work is available in the literature. (Xiao et al. 2022) applied a support vector machine on atmospheric-oceanic dynamics data SCP. However, satellite data has great importance for inaccessible and hazardous regions where gathering data physically is not possible (Xiao et al. 2022). SEP is done to find a suitable location for the installation of solar plants and reduce the dependence on fossil fuels for economic development. (Jebli et al. 2021) predicted solar energy using random forest on meteorological data. The existing studies for the listed applications do not use satellite data. However, they need analysis of SITS with high spatial & high temporal resolutions but data from a single satellite system has a trade-off between them.

Spatiotemporal Fusion: To handle the trade-off, spatiotemporal fusion is a possible solution. STARFM (Gao et al. 2006) creates synthetic Landsat-like image at timestamp $t+1$ by fusing MODIS and Landsat images at time t. It is a linear model which calculates the reflectance value of a pixel by a weighted sum of the neighboring pixels. It is a pixel-based method that needs at least one pair of images captured on the same day. Developed variants of the method also suffer from similar challenges. Another study uses a linear regression on pixels to generate images (Ping, Meng, and Su 2018). The pixel-based methods blindly use noisy pixels in the fusion process, thus propagating the noise in the neighboring pixels of the predicted image (Tan et al. 2022).

Given the limitations of pixel-based methods, learning-based approaches are gaining interest due to their ability to capture complex relationships in data without relying on predefined assumptions. (Wang et al. 2017) and (Wei et al. 2017) used downscaling & upscaling to generate an image having a spatial resolution of Landsat-8 with the help of a MODIS image. A few attempts have been made to use advanced Generative adversarial networks (GAN) for image generation. (Bouabid et al. 2020) generate a Landsat image at time t using MODIS and Landsat images at t and $t-1$, respectively. Similarly, (Tan et al. 2022) used GAN to handle noise while generating a Landsat-like image using a MODIS image at timestamp t and Landsat images at $t-1$ and $t+1$. Due to the complex image generation process, generative models have been applied to small-scale datasets having only a few locations (Bouabid et al. 2020). The applications under consideration require time series at a high temporal resolution and interpolating images between two consecutive images is inefficient and computationally expensive. Moreover, this approach increases data volume twofold. Also, the generated images can increase the already existing noise in the original images. Our proposed method overcomes these problems.

Problem Formulation

We have considered three spatiotemporal forecasting problems viz. CYP, SCP, and SEP. The goal is to predict $\hat{y}_{c,z} \epsilon$ {crop yield, percentage of area under the snow, and solar energy produced} for a county c at prediction time granularity z which is a year, a month and a fortnight for CYP, SCP, and SEP, respectively. Let input data set of TS be $X_z = \{[x_1^1, x_1^2, ..., x_1^t], [x_2^1, x_2^2, ..., x_2^t], ...[x_{z-1}^1, x_{z-1}^2, ..., x_{z-1}^t]\}$, where t represents the number of timestamps depending on the application and the satellite and x_1^t is its image at timestamp t. For example, for soybean crop t=15 and t=30 for Landsat-8 and MODIS, respectively.

Proposed Framework

We propose two models: PatchNet and FuSITSNet. PatchNet extracts prominent features from high spatial resolution SITS by partial spatial traversal and covering the hotspot areas using patch selection mechanism. FuSITSNet is a twofold fusion model that presents a way to fuse two SITS at the feature level. It uses two encoders - 1) image time series encoder (TSE) and 2) PatchNet (shown in blackbox in Figure 1 for image time series of high spatial resolution satellites. The representations of the two time series are then passed to a twofold Fusion module (shown as a black box in Figure 1). It learns complementary high spatial and temporal features to give a joint representation of the two TS. The regression module is applied to joint features for final prediction. The overview of the FuSITSNet is shown in Figure 1 & all the modules are described in subsequent subsections.

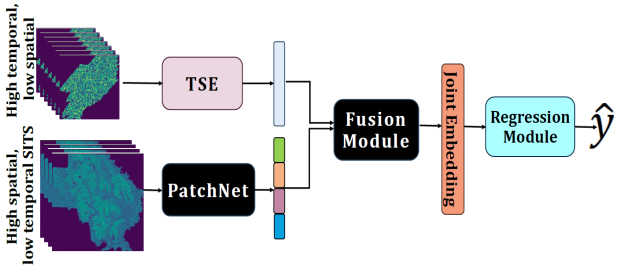

Figure 1: FuSITSNet

PatchNet

PatchNet is designed to encode high spatial resolution SITS which is otherwise impractical to process. It works on image times series iteratively for multiple patch time series (patchTS) and uses the idea of a beam search for optimizing the patch selection process. A patch is selected in the spatial dimension and patchTS consists of entire time series for the patch. The architecture of the PatchNet is given in Figure 2. We divide the image time series into a spatial virtual grid, resulting in multiple patchTS, one for each cell. We now onwards refer to patchTS as a patch. The patches are processed using TSE and their representations are passed to the Patch selection module (PSM). PSM uses attention score to identify top 'k' patches that are then forwarded to the Neighbor Selector (NS). NS determines the unprocessed neighboring patches of top 'k' patches and also creates a list of patches to be processed in the next iteration. The process continues till a fraction ($1/p$) of the SITS is processed. The enhanced patch representations obtained from PSM are passed to the embedding generation module which outputs the embedding of the entire SITS learned by the network in multiple iterations. The pseudo-code is given in Algorithm 1.

Time Series Encoder (TSE) gives a linear representation of the input patch. It consists of two submodules 3DCNN network and a Spatial Attention Mask (SAM) followed by a linear layer.

3DCNN Module (Gavahi, Abbaszadeh, and Moradkhani 2021) consists of three convolution layers having 10, 15, and 20 filters with zero padding. Each convolution layer is followed by a 3D-max pool layer. 3DCNN leverages both spatial and temporal features simultaneously and learns more informative representations of the volume.

Spatial attention mask (SAM) We followed (Mohla et al. 2020) and modified it for our problem. It has 6 2D convolution layers, each followed by a batch normalization layer to reduce the internal covariate shift and model overfitting.

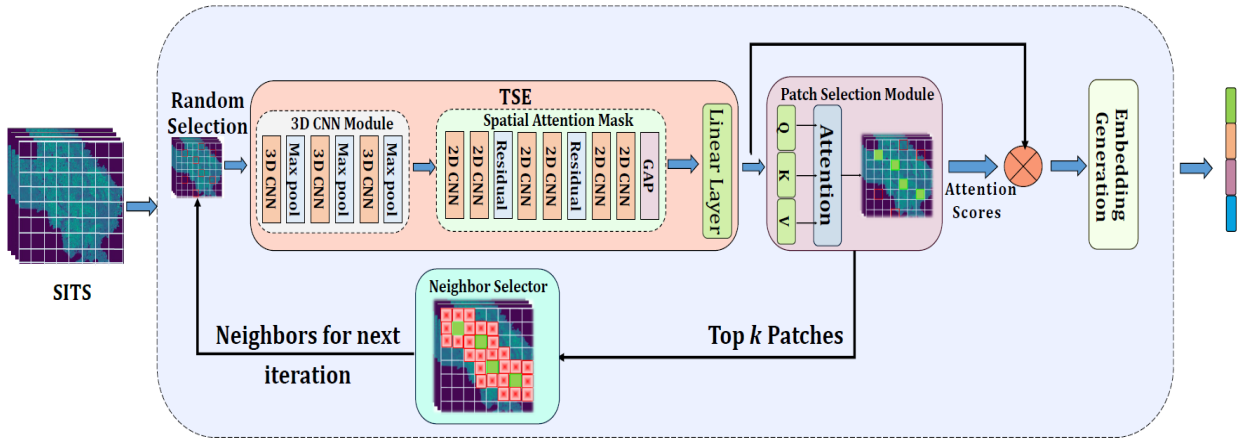

Figure 2: PatchNet

The skip connections are used after every two convolution layers to improve the information flow within the network and mitigate the problem of vanishing gradient. Global average pooling is done by two pooling operations 'average pooling' and 'max pooling' applied along the channel axis and are concatenated to create an efficient feature descriptor. A convolution layer is applied over the feature descriptor to get the highlighted regions.

Patch Selection Module (PSM) We utilize self-attention mechanism (given by eq. 1 and 2) to focus on the most important *k* patches from *n* input patches. PSM learns the enhanced representations of all the patches across iterations using the following process and gives the score of each patch based on its contribution to the end task. The input to PSM is $R = \{r_1, r_2, ...r_n\}$, where *n* is total number of patches and r_i is the linear representation of each patch after being processed by TSE. The mathematical representation is:
Query (*Q*), Key (*K*), and Value (*V*) for self attention are:

$$Q = R \times w_q, \quad K = R \times w_k, \quad V = R \times w_v \tag{1}$$

where w_q, w_k, and w_v are weight matrices for *Q*, *K*, and *V*, respectively.

$$A = softmax(QK^T) \tag{2}$$

where $A = \{a_1, a_2, ..., a_n\}$ is an attention score matrix for all the *n* patches, and each a_s is of size *b* equal to size of patch embedding.
To get the collective score i.e. contribution of the patch towards the end task is calculated as:

$$S = \sum_{i=0}^{b} a_s^i \quad s = 1 \; to \; n \tag{3}$$

$$l, \bar{l} = top\_k(S) \quad where \; l + \bar{l} = n \tag{4}$$

where top_k is the function that returns a list, *l*, the indices of top *k* patches and a list, $\bar{l}$, remaining patches to be used in the next iteration of the selection process.

Algorithm 1: PatchNet

Input: SITS
Output: Embedding of SITS
initialize: $m = 0$ *and*
$|P| = total\,patches\;in\;SITS$

while $(m) \neq |P|/p$ **do**
  select *n* random patches
  $R = TSE(patchTS)\;\forall\, n$ patches // apply TSE to get linear representation of *n* patches
  $l, \bar{l}, \tilde{R} = PSM(R)$ // list of top *'k'* patches and enhanced patch representations
  $n' = NS(l)$ // *NS* gives neighbors of each patch in *l*
  Select $n - n'$ random patches
  $m = m + n$
  $E_L = EG(\tilde{R})$
end while
return E_L

PSM also helps in enhancing the patch representations *R* as:

$$\tilde{R} = S \times V \tag{5}$$

Neighbor Selector (NS) finds the untraversed neighboring patches of all *k* patches. For a patch p_{ij}, set of neighbors is $\{p_{ef} - p_{ij}\}, e = i-1, i, i+1$ and $f = j-1, j, j+1$. Selecting the neighboring patches ensures that focus is maintained near the hotspots and this leverages the geospatial information to boost the prediction. We select a few untraversed random patches for the next iteration to make the number of patches *n*.

Embedding Generation (EG) is a two-level process and consists of two linear layers. The first layer is used to get the representation of each patch, p_{ij} in an iteration. The embeddings of all selected patches across iterations are then concatenated and passed to the second linear which gives a representation of the entire SITS.

FuSITSNet

FuSITSNet (Figure 1) consists of two encoders TSE and PatchNet and a fusion module. We use FuSITSNet for fusing two SITS from Landsat-8 and MODIS. We processed Landsat SITS using PatchNet. MODIS has a coarser spatial resolution and can be processed as a whole, thus we used TSE for processing its time series. However, we can replace TSE with PatchNet to generate embeddings if the second SITS also has a high spatial resolution.

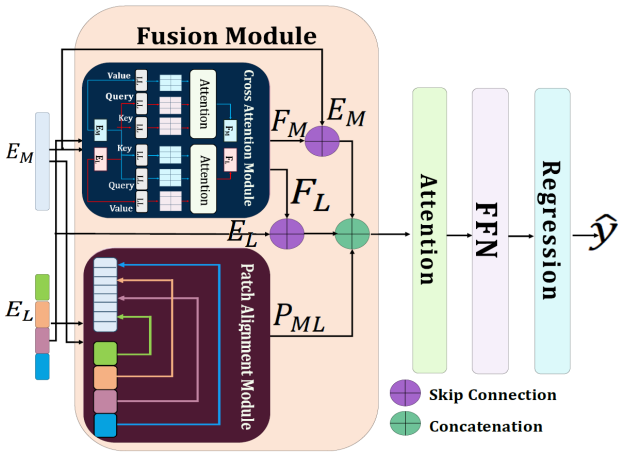

Figure 3: Fusion Module

Fusion Module: Given in Figure 3 is a twofold module that takes the embeddings E_M and E_L from the two encoders for MODIS and Landsat-8, respectively. It learns the features from the two modalities using two sub-modules, patch alignment module, and cross-modality attention.

Patch Alignment Module (PAM): We use PAM inspired by the text video correlation aware module (Chen et al. 2022) to align the patches of fine spatial resolution modality (Landsat-8) with the corresponding regions in MODIS time series and learn fine temporal patterns for the aligned patches. This also suppresses the noise present in two SITS and mitigates its effect on the end task. In the alignment process we calculate the similarity, Sim_L of Landsat-8 patches with MODIS as given below:

$$Sim_L = E_L(E_M)^T \tag{6}$$

Softmax is applied over Sim_L and we use an average patch-wise aggregator over the MODIS embeddings as:

$$P^i_{ML} = softmax(Sim_{L_i})E_M, \ \ (1 \leqslant i < n) \tag{7}$$

$$P_{ML} = [P^1_{ML}; P^2_{ML}; \cdots ; P^n_{ML}] \tag{8}$$

n is number of patches traversed, P^i_{ML} is similarity-aware aggregated MODIS representation of i^{th} patch of Landsat.

Cross-Modal attention (CMA) learns the inter-modality relationships from the two embeddings E_M and E_L by applying bi-directional cross-modality attention by taking queries from both modalities to leverage their profound features. The scalar dot product attention between the hotspot spatial features of Landsat and highlighted temporal features of MODIS gives the joint high quality features in both aspects. This helps the model to capture the complementary aspects of the two modalities and thus utilizes the information from one modality to compensate for the low quality of the other modality. Also, the module covers the untraversed Landsat-8 regions with the help of MODIS time series. The output of the module is represented by F_M and F_L.

We concatenate P_{ML}, F_M, and F_L and apply multi-head self-attention to highlight the combined hot spot features. It is followed by a feed-forward network comprising three linear layers with the Gaussian Error Linear Unit (GELU) activation function and followed by layer normalization. Lastly, a regression layer is applied to get the prediction.

Dataset Details

We considered the top producers of corn and soybean from the United States for CYP. The crop yield labels are collected from Quick Stats (USDA 2010) compiled by the United States Department of Agriculture (USDA). For SCP, we have considered the counties which experience average snowfall of more than 250 inches per year. The percentage of the area covered under snow is obtained from the MODIS product MOD10A1 (NASA 2000). For SEP, we considered 5 states. Details are given in Appendix A.

Data Preparation

Satellite images are by default in float values and require more bits for storage. We applied **bits precision** for compression where we replaced float values with unsigned integers (uint) (Hubara et al. 2016) which reduced the storage requirements by four. The difference between RMSE for CYP is less than 1% when we applied our model for 100 counties on float SITS and uint SITS. We used uint SITS with five surface reflectance bands common in both satellites (MODIS & Landsat-8) for all the experiments conducted. For data preparation details see Appendix B.

Learning Objectives

We use marginal contrastive loss for contrastive learning and the mean square error (MSE) for the prediction task.

Margin Contrastive Loss: In our twofold fusion model, we innovatively applied margin contrastive loss (Shah et al. 2022). Utilizing contrastive loss in regression problems is challenging since there are no explicit class categories to directly determine positive and negative pairs for training. The number of "classes" is roughly equivalent to the size of the dataset, rendering traditional contrastive loss implementation difficult. To overcome this challenge, we used batch-wise margin contrastive loss (Kaur, Goyal, and Goyal 2023). We selected the margin as 0.5 which minimized RMSE after experimenting with values [0.1,0.5,1.0]. Contrastive learning gives us two losses $loss_{pos}$ and $loss_{neg}$ for positive and negative pairs, respectively.

Mean squared error (MSE): We used mean squared error for the regression task. It gives the mean squared error between the actual target $y_{c,z}$ and the predicted output $\hat{y}_{c,z}$ for all the considered location-year pairs.

Model	Corn	Soybean
CNN(You et al. 2017)	24.617	8.346
CNN+GP(You et al. 2017)	23.881	8.343
CNN+LSTM*(Sun et al. 2020)	23.632	8.370
CYN**(Kaur et al. 2022)	21.819	7.370
PatchNet	**21.469**	**7.290**
PatchNet+M	**20.631**	**6.963**

*uses additional soil data, **uses both meteorological & soil

Table 1: Comparison: PatchNet vs histogram models

Model	CYP		SCP	SEP
	Corn	Soy		
TSE (MODIS)	23.335	7.545	17.167	8.863
PatchNet(Landsat-8)	21.469	7.290	12.813	8.543
PatchNet(STARFM)	20.289	6.308	—	7.227
PatchNet(RSFN)	22.839	6.432	12.329	8.012
PatchNet(GAN)	18.102	6.296	11.951	7.043
FuSITSNet	**16.1925**	**5.0389**	**9.2308**	**2.0447**

Table 2: FuSITSNet vs single modality baselines

The total loss (*L)* for the model is:

$$L = loss_{pos} + loss_{neg} + MSE$$

Baselines for Comparison

We considered three types of models for comparison: (i) **histogram models** which work on histogram TS of satellite data. (ii) **our baselines** for single modality SITS. (iii) our baselines on high temporal SITS generated using **generative fusion models**

Histogram models: We considered four existing CYP models CNN (Sun et al.), CNN+GP (Sun et al.), CNN+LSTM (Sun et al. 2020), and CYN (Kaur et al. 2022) working on histogram TS to compare with the proposed PatchNet. CNN and CNN+GP use only surface reflectance data and did not exploit the temporal dependency in the data. CNN+LSTM models CYP as a temporal problem using soil data and surface reflectance TS. The authors processed raw features using 2DCNN and used LSTM to model the sequence embeddings. CYN modeled CYP as a spatiotemporal problem and used soil, and meteorological data along with surface reflectance histograms. All these models work for different locations, and time duration. However, we used the data for the same locations and time duration in all the models for a fair comparison. To the best of our knowledge, there are no existing models working on histograms for the other two applications.

Baseline models: To the best of our knowledge, there is no method that works with SITS for spatiotemporal problems. We applied the proposed PatchNet and TSE models on single modality image time series of Landsat-8 and MODIS, respectively, and compared with FuSITSNet to see the significance of fusing two time series over a single modality.

Generative fusion models: We applied three existing generative fusion models viz. STARFM (Gao et al. 2006), RSFN (Tan et al. 2022), and GAN (Bouabid et al. 2020) to enhance the temporal resolution of Landsat time series. We generated images at every mid-timestamp to get time series of 8-day frequency. We then applied PatchNet for predictions using enhanced SITS and compared them with FuSITSNet.

Details of models for comparison are in Appendix C.

Experiments

We performed experiments using Pytorch 1.11.0 and CUDA 11.7 on an A100 GPU server with 80 GB RAM. A model is trained for 50 epochs with a batch size of 8 using Adam optimizer with a learning rate η. We have trained the model with 5 years of data (2014-2018) and, 2 years (2019 and 2020) for testing. To predict the output for the z^{th} year, the training is conducted until the $(z-1)^{th}$ year. For CYP, $\eta = 0.0005$ for a single modality (TSE and PatchNet) and $\eta = 0.000005$ for FuSITSNet. In case of SCP and SEP, $\eta = 0.00001$ for all three models. We performed each experiment 5 times and observed a standard deviation of less than 0.2 for all proposed models. The evaluation metric used is Root Mean Squared Error (RMSE). Details are given in Appendix D.

Results and Analysis

Significance of using SITS over histograms time series: The first set of experiments are conducted to compare the proposed model PatchNet with existing CYP models using histogram TS. Table 1 presents RMSE (in bu/ac) achieved for corn and soybean yield prediction using various models. It can be observed from the table that for corn yield prediction RMSE reduced by $\approx 12\%$ and 10% with that of CNN and CNN+GP, respectively. These two models use only surface reflectance histograms. The reduction in RMSE is 9% for the CNN+LSTM model which also incorporates meteorological data. CYN uses both meteorological and soil data along with surface reflectance histograms. PatchNet outperforms CYN even without using any other data. However, the error is reduced by an additional 6% when meteorological data is incorporated into PatchNet.

Comparison of FuSITSNet with single modality baselines: Table 2 presents RMSE obtained by FuSITSNet and single modality baselines TSE and PatchNet using MODIS and Landsat-8 time series, respectively. It is evident from the results, that the PatchNet (Landsat-8) performed better than TSE (MODIS) with $\approx 8\%$ and 3.5% lower RMSE in corn and soybean yield prediction, respectively. RMSE reduced from 17.17 to 12.81 for SCP and from 8.86 to 8.54 for SEP, making an improvement of 25% and 3.6%, respectively. This shows the importance of using high spatial resolution data for the applications. The results improved further using FuSITSNet for all three applications. RMSE reduced by $\approx 24\%$ and 30% for corn and soybean yield prediction, respectively when compared with PatchNet (Landsat-8). A similar pattern is observed in SCP with an improvement of 46% from TSE and $\approx 28\%$ from PatchNet (Landsat-8). The maximum improvement is observed in SEP with almost 76%. The huge reduction in error signifies that FuSITSNet exploits high temporal and high spatial features and is thus

Model	CYP		SCP	SEP
	Corn	Soy		
FuSITSNet	16.192	5.038	9.230	2.044
FuSITSNet (no PAM)	17.198	5.286	12.448	2.501
FuSITSNet (no CMA)	17.242	5.847	11.523	2.749

Table 3: Ablation Study

suitable for spatiotemporal applications. The performance of FuSITSNet is improved further by using meteorological data. Details are given in Appendix E. We also compared the models for no. of parameters & running time required. The no. of parameters & training time required for generative fusion models is the sum of parameters & time needed in the generation and prediction process. FuSITSNet has more parameters, but the running time is approx. $1/4^{th}$ of other fusion models as it does not need a generation process.

Our baselines on enhanced SITS: We generated Landsat-8 images at every mid-timestamp. It can be observed from Table 2 that RMSE reduces when PatchNet is applied over enhanced time series in comparison to PatchNet (Landsat-8) with an exception in the case of corn yield prediction. Out of three generative models, GAN generated image time series performed the best with an improvement of ≈16%, 14%, 7%, and 18% for corn, soybean, snow cover, and solar energy prediction, respectively.

Comparison of FuSITSNet with Generative fusion models: Table 2 shows that FuSITSNet outperforms all the scenarios where PatchNet applied on SITS generated by the existing generative fusion models. RMSE reduced by $\approx 20\%$ for CYP in comparison to the pixel-based model STARFM. The reduction in RMSE for FuSITSNet is 29% and 10% in comparison to learning-based models RSFN and GAN, respectively for corn yield prediction and the corresponding reduction is 21% and 19% for soybean. The best results are obtained for solar energy prediction where RMSE is reduced by $\approx 70\%$ using FuSITSNet than that of PatchNet(GAN).

Ablation Study: We carried out an ablation study to show the importance of modules in FuSITSNet. Considering variations in results in Table 3, we observed that RMSE increased significantly without using PAM with a maximum increase of 25% for SCP followed by 18% in solar energy. Similarly, the performance of the model is also degraded without cross-modality attention. This shows that both modules are important to effectively exploit high spatial and high temporal features in the two SITS.

Significance of patch selection mechanism: We conducted experiments without using PSM and NS in PatchNet and instead replaced them with random patch selection. RMSE achieved by random selection is 24.29 bu/ac and 9.98 bu/ac for corn and soybean yield prediction in comparison to 21.47 bu/ac and 7.29 bu/ac, respectively using PatchNet. It is evident that there is a significant improvement in the model performance and PSM and NS collectively work effectively to exploit the required hot spot features in the SITS and eliminate the need to fully process it. Also, it suppresses the noise in the two modalities.

Deciding (1/p)th traversal of SITS: The next set of experiments is performed for PatchNet to know the optimum number of patches for the traversal of Landsat-8 image time series. Figure 4 shows the computation time required and RMSE curve for corn yield prediction by varying p as 5,4,3, and 2 keeping the patch size of $H \times H$, $H = 64$. It can be observed from Figure 3 that, there is an improvement of $\approx 6\%$ in the performance of the model when the traversed region p changes from 5 to 4, and RMSE did not change much after that. However but the computation time required from p = 5 to 4 is almost constant but it increases linearly after that.

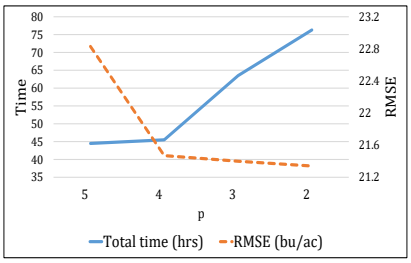

Figure 4: Deciding p for (1/p)th traversal of SITS

We also experimented by changing the **patch size** to $H = 128$ for $p = 4$ and found that RMSE reduced to 21.42 bu/ac which is just 0.2% as compared to that with patch size 64, but the computation time required to process the patch size of 128 increased 1.5 times. To maintain the trade-off between the computation time and RMSE, we chose to carry out the results by taking $p = 4$ and patch size $H = 64$.

Conclusion

Satellite image technology is increasingly being adopted by researchers worldwide for Earth observation to solve problems related to environment and climate change. The democratization of this technology is still marred by the need for processing huge volumes of data and by the unavailability of high spatial and temporal resolution images from a single publicly available satellite system. We proposed two models, PatchNet and FuSITSNet, to overcome these problems. PatchNet makes it feasible to efficiently process high spatial resolution SITS whereas, FuSITSNet fuses two image time series to obtain a joint representation that captures high spatial resolution features of one satellite system and high temporal resolution features of another. We have fused Landsat-8 and MODIS SITS to predict crop yield, snow cover, and solar energy for 2000 US counties and obtained state-of-the-art results. One of the salient features of the models is that high spatial and temporal features are learned without image generation, thereby not increasing the voluminous data further as is the case with generative approaches. Another salient feature is that the fusion module of FuSITSNet suppresses noise. The performance of the proposed approach is improved further by incorporating meteorological data. The proposed approach is applied only on permutation invariant applications for now. In future, we plan to extend our approach to other applications like land use and land cover classification problems.

Acknowledgments

This work is carried out in the Disruptive Technologies lab which is supported by the Department of Science and Technology (DST), Govt. of India in the form of FIST Level-1 grant to the Department of CSIS, BITS Pilani

References

Bouabid, S.; Chernetskiy, M.; Rischard, M.; and Gamper, J. 2020. Predicting landsat reflectance with deep generative fusion. *arXiv preprint arXiv:2011.04762*.

Chen, X.; Zhang, N.; Li, L.; Deng, S.; Tan, C.; Xu, C.; Huang, F.; Si, L.; and Chen, H. 2022. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 904–915.

Choudhary, K.; Pandey, V.; Murthy, C.; and Poddar, M. 2019. Synergetic use of optical, microwave and thermal satellite data for non-parametric estimation of wheat grain yield. *The Int. Archives of Photogrammetry, RS and Spatial Information Sciences*, 42: 195–199.

de Wit, A.; Duveiller, G.; and Defourny, P. 2012. Estimating regional winter wheat yield with WOFOST through the assimilation of green area index retrieved from MODIS observations. *Agricultural and forest meteorology*, 164: 39–52.

European Space Agency Signature. 2017. Sentinel webpage. https://www.netiq.com/documentation/sentinel-82/user/data/bookinfo.html. Accessed: 2023-2-25.

Fan, H.; Zhang, F.; and Gao, Y. 2020. Self-supervised time series representation learning by inter-intra relational reasoning. *arXiv preprint arXiv:2011.13548*.

Fan, J.; Bai, J.; Li, Z.; Ortiz-Bobea, A.; and Gomes, C. P. 2022. A GNN-RNN approach for harnessing geospatial and temporal information: application to crop yield prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 11873–11881.

Gao, F.; Masek, J.; Schwaller, M.; and Hall, F. 2006. On the blending of the Landsat and MODIS surface reflectance: Predicting daily Landsat surface reflectance. *IEEE Transactions on Geoscience and Remote sensing*, 44(8): 2207–2218.

Gavahi, K.; Abbaszadeh, P.; and Moradkhani, H. 2021. DeepYield: A combined convolutional neural network with long short-term memory for crop yield forecasting. *Expert Systems with Applications*, 184: 115511.

Gupta, Y.; Goyal, N.; Varghese, V. J.; and Goyal, P. 2023. Utilizing MODIS Fire Mask for Predicting Forest Fires Using Landsat-9/8 and Meteorological Data. In *2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*, 1–10. IEEE.

Guruprasad, R. B.; Saurav, K.; and Randhawa, S. 2019. Machine learning methodologies for paddy yield estimation in India: a case study. In *IGARSS 2019-2019*, 7254–7257. IEEE.

He, E.; Xie, Y.; Liu, L.; Chen, W.; Jin, Z.; and Jia, X. 2023. Physics Guided Neural Networks for Time-Aware Fairness: An Application in Crop Yield Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 14223–14231.

Hubara, I.; Courbariaux, M.; Soudry, D.; El-Yaniv, R.; and Bengio, Y. 2016. Binarized neural networks. *Advances in neural information processing systems*, 29.

ISRO. 2019. cartosat. https://www.isro.gov.in/CARTOSAT_1.html. Accessed: 2023-5-18.

Jebli, I.; Belouadha, F.-Z.; Kabbaj, M. I.; and Tilioua, A. 2021. Prediction of solar energy guided by pearson correlation using machine learning. *Energy*, 224: 120109.

Ji, Z.; Pan, Y.; Zhu, X.; Zhang, D.; and Wang, J. 2022. A generalized model to predict large-scale crop yields integrating satellite-based vegetation index time series and phenology metrics. *Ecological Indicators*.

Jiang, T.; Huang, M.; Segovia-Dominguez, I.; Newlands, N.; and Gel, Y. R. 2022. Learning space-time crop yield patterns with zigzag persistence-based lstm: Toward more reliable digital agriculture insurance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 12538–12544.

Kaur, A.; Goyal, P.; and Goyal, N. 2023. LSFuseNet: Dual-Fusion of Landsat-8 and Sentinel-2 Multispectral Time Series for Permutation Invariant Applications. In *2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*, 1–10. IEEE.

Kaur, A.; Goyal, P.; Rajhans, R.; Agarwal, L.; and Goyal, N. 2023. Fusion of multivariate time series meteorological and static soil data for multistage crop yield prediction using multi-head self attention network. *Expert Systems with Applications*, 226: 120098.

Kaur, A.; Goyal, P.; Sharma, K.; Sharma, L.; and Goyal, N. 2022. A Generalized Multimodal Deep Learning Model for Early Crop Yield Prediction. In *International Conference on Big Data*, 1272–1279. IEEE.

Khaki, S.; and Wang, L. 2019. Crop yield prediction using deep neural networks. *Frontiers in plant science*, 10: 621.

Måløy, H.; Windju, S.; Bergersen, S.; Alsheikh, M.; and Downing, K. L. 2021. Multimodal performers for genomic selection and crop yield prediction. *Smart Agricultural Technology*, 1: 100017.

Mohla, S.; Pande, S.; Banerjee, B.; and Chaudhuri, S. 2020. Fusatnet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and lidar classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 92–93.

NASA. 2000. MODIS Product. https://developers.google.com/earth-engine/datasets/catalog/MODIS_006_MOD10A1. Accessed: 2022-4-13.

NASA. 2015. MODIS. https://lpdaac.usgs.gov/data/get-started-data/collection-overview/. Accessed: 2022-12-06.

NASA. 2016. Landsatwebpage. https://www.usgs.gov/faqs/what-are-band-designations-landsat-satellites?qt-news_science_products=0#qt-news_science_products. Accessed: 2022-2-16.

Ping, B.; Meng, Y.; and Su, F. 2018. An enhanced linear spatio-temporal fusion method for blending Landsat and

MODIS data to synthesize Landsat-like imagery. *Remote Sensing*, 10(6): 881.

planet. 2019. planetscope. https://www.planet.com/?utm_source=google&utm_medium=paid-search&utm_campaign=discovery&utm_content=pros-leads-responsive-search-0623&utm_source=google&utm_medium=paid-search&gad=1. Accessed: 2023-5-18.

Poonam Goyal. 2022. github. https://github.com/DrPoonamGoyal/FuSITSNet-at-AAAI2024. Accessed: 2023-7-15.

Sakamoto, T. 2020. Incorporating environmental variables into a MODIS-based crop yield estimation method for United States corn and soybeans through the use of a random forest regression algorithm. *ISPRS Journal of Photogrammetry and Remote Sensing*, 160: 208–228.

Shah, A.; Sra, S.; Chellappa, R.; and Cherian, A. 2022. Max-Margin Contrastive Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 8220–8230.

Skakun, S.; Kalecinski, N. I.; Brown, M. G.; Johnson, D. M.; Vermote, E. F.; Roger, J.-C.; and Franch, B. 2021. Assessing within-field corn and soybean yield variability from WorldView-3, Planet, Sentinel-2, and Landsat 8 satellite imagery. *Remote Sensing*, 13(5): 872.

Sun, J.; Di, L.; Sun, Z.; Shen, Y.; and Lai, Z. ???? County-level soybean yield prediction using deep CNN-LSTM model. *Sensors*, 19(20).

Sun, J.; Lai, Z.; Di, L.; Sun, Z.; Tao, J.; and Shen, Y. 2020. Multilevel deep learning network for county-level corn yield estimation in the us corn belt. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13: 5048–5060.

Tan, Z.; Gao, M.; Yuan, J.; Jiang, L.; and Duan, H. 2022. A Robust Model for MODIS and Landsat Image Fusion Considering Input Noise. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–17.

USDA. 2010. USDA/Nass QuickStats AD-hoc query tool. https://quickstats.nass.usda.gov/. Accessed: 2022-7-15.

Verma, U.; Piepho, H.; Goyal, A.; Ogutu, J.; Kalubarme, M.; et al. 2016. Role of climatic variables and crop condition term for mustard yield prediction in Haryana. *Int J Agric Stat Sci*, 12: 45–51.

Wang, Q.; Zhang, Y.; Onojeghuo, A. O.; Zhu, X.; and Atkinson, P. M. 2017. Enhancing spatio-temporal fusion of MODIS and Landsat data by incorporating 250 m MODIS data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(9): 4116–4123.

Wei, J.; Wang, L.; Liu, P.; Chen, X.; Li, W.; and Zomaya, A. Y. 2017. Spatiotemporal fusion of MODIS and Landsat-7 reflectance images via compressed sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 55(12): 7126–7139.

Xiao, X.; He, T.; Liang, S.; Liu, X.; Ma, Y.; Liang, S.; and Chen, X. 2022. Estimating fractional snow cover in vegetated environments using MODIS surface reflectance data. *International Journal of Applied Earth Observation and Geoinformation*, 114: 103030.

You, J.; Li, X.; Low, M.; Lobell, D.; and Ermon, S. 2017. Deep gaussian process for crop yield prediction based on remote sensing data. In *Thirty-First AAAI conference on artificial intelligence*.