Crop classification from satellite image sequences using a two-stream network with temporal self-attention

Andreas Stergioulas, Kosmas Dimitropoulos, Nikos Grammalidis
Information Technologies Institute, Centre for Research and Technology Hellas,
Thessaloniki, Greece
{andrster, dimitrop, ngramm}@iti.gr

2022 IEEE International Conference on Imaging Systems and Techniques (IST) | 978-1-6654-8102-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/IST55454.2022.9827752

***Abstract*—In recent years the availability of satellite image observations of Earth has been increasing, creating opportunities for automated methods to be applied in tasks with significant economic importance, such as agricultural parcel crop classification. Designing and implementing automated methods that can efficiently interpret satellite images and handle their temporal nature poses a significant challenge in remote sensing. Deep learning models have proven to be able to leverage these type of data, taking into consideration both their spatial as well as temporal nature.**

Building on a state-of-the-art architecture using self-attention to classify crops captured in satellite images time series, we introduce two changes in order to better capture the crop phenology. Specifically, the calculation of the self-attention Query is performed by a Temporal Convolutional Network (TCN), while the TCN output is also taken under consideration for the final classification. Moreover, we utilize the temporal differences between consecutive time steps to create an auxiliary time series that can be employed alongside the original time series, in a two-stream architecture, that proves to be capable of further improving performance. We also conduct a detailed ablation study to assess the impact of these contributions. The proposed model was able to produce results that exceed the state-of-the-art on the publicly available Sentinel2-Agri dataset.

***Index Terms*—Remote sensing, Crop Classification, Deep Learning, Self-Attention, Time Series Classification**

I. Introduction

The classification of farming crops and its continuous monitoring is a significant matter for the agricultural sector at the national and international level. The rising availability of satellite data, has created opportunities to automate the classification process and lower the financial cost. More specifically, the SENTINEL2-A/B and LANDSAT-7/8 satellites, with their high spectral and temporal resolution (13 spectral bands with a revisit rate of five days for SENTINEL2-A/B), have enabled the creation of methods that can properly classify crop phenology i.e. the periodic changes in plant life cycles.

Early works in crop classification relied on machine learning methods such as Random Forest [4], hidden Markov models [19] or Support Vector Machines [1]. Recent state-of-the-art crop classification methods employ Deep Neural Networks (DNNs), since they are capable of handling large amounts of data as well as modeling temporal dependencies efficiently. In order to model the temporal nature of satellite imagery, a variety of temporal architectures have been utilized, from recurrent Neural Networks [10], [16], [18] to convolutional networks such as 1D temporal convolutions (1D CNNs) [22] and temporal convolutional networks (TCN) [14], [20]. Even though TCNs are capable to capture temporal information and handle sequences of arbitrary length, they alone are not suitable for classification of imbalanced data [17]. More recent works in natural language processing and computer vision, have proven that attention based approaches are more efficient for temporal modeling compared to RNNs [21], [23]. As such, self-attention has been adapted for satellite time series classification and methods that employ the original Transformer model or variations of it have shown state-of-the-art results [2], [3], [17]. Self-attention allows time steps that contain more meaningful information regarding the recognition of a crop type to contribute more to the final classification, while downgrading the influence of less informative observations. Furthermore, Transformers yield classification performance that is on par with, and in many cases better than, RNN/LSTM-based models and present the same robustness to cloud-obstructed observations [3], [17].

Motivated by the performance of self-attention, in the current work, we propose a temporal attention module (TAE) that uses a TCN in order to create a self-attention Query that summarizes the multiple temporal steps of the satellite images to a single dimensional output for the entire time series, in contrast to [2], where they define the Query as a learnable parameter, independent of the input data. The TCN output is also taken into consideration during the crops classification step, by combining it with the self-attention output. Additionally, we employ the differences between time steps as extra information regarding the satellite time series, as we argue that this additional representation also aids the temporal modeling. As it has been proven that handling medium resolution satellite images as a pixel set, we utilize the introduced in [3] Pixel Set Encoder (PSE) for the creation of the spatial features. We compare the performance of the proposed PSE-TCN-TAE coupled with the extra differences time series and demonstrate that our method outperforms the state-of-the-art.

Our main contributions are as follows:

- Inspired by Garnot et al. [2], [3] we design a new

Authorized licensed use limited to: Centre for Research and Technology (C.E.R.T.H.). Downloaded on September 27,2023 at 11:00:02 UTC from IEEE Xplore. Restrictions apply.
978-1-6654-8102-1/22/$31.00 ©2022 IEEE

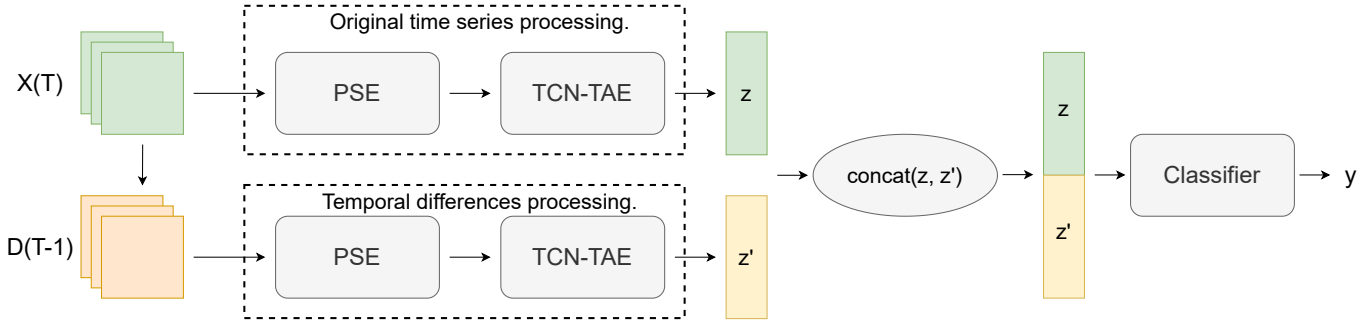

Fig. 1: An overview of the pseudo multimodal approach, using the PSE-TCN-TAE model.

temporal self-attention module that is applied on top of their proposed pixel set encoder.

- We create an additional pseudo modality by taking the temporal differences of the original time series and combine the newly created stream of information with the original input.
- Our model with the proposed temporal module yields state-of-the-art results on the Sentinel2-Agri dataset [3], while with the addition of the extra pseudo modality stream we further improve the classification results.

II. Method

Initially, a certain number of pixels are sampled for each time step of a parcel, creating the input sequence $X(t)$, $t \in [1, ..., T]$. Then, by subtracting samples from two consecutive time steps, a temporal differences time sequence $D(t) = X(t) - X(t-1)$, $t \in [1, ...T-1]$ is constructed. The Pixel-set Encoder (PSE) module defined in [3] is combined with the new proposed TCN-TAE module to encode both time series $X(t), D(t')$. Thus, two PSE-TCN-TAE models are employed to produce two streams of features, z from the original data and z' from the temporal differences pseudo modality. These two streams of features are then concatenated before being fed to an MLP classifier that (decoder). An illustration of the complete approach can be viewed in Figure 1.

In the following subsections, we briefly present the employed spatial module PSE, following [3], before presenting our temporal attention module and the pseudo multimodal approach, where the differences of consecutive time steps are estimated and treated as an extra input modality.

A. Spatial Encoder

Time series of satellite imagery add a temporal dimension to satellite images, which typically contain multiple spectral bands for each pixel that refers to a specific spatial location. In the PSE-TAE [3] implementation, the authors noticed that CNNs are not suitable to handle medium resolution satellite images and instead, inspired by PointNet [15], proposed to handle the spatial part of the time series as pixel sets. The Pixel-Set Encoder (PSE), stochastically samples S pixels from a parcel with N pixels. If the total number of pixels in a parcel is less than S, an arbitrary selected pixel is repeated $N - S$ times. The spectral channels are then processed by a series of fully connected layers, 1D batch normalizations [5] and ReLUs [11] along the channel dimension and the sampled pixels are reduced to their mean and standard deviation values. Finally, a vector of geometrical features containing the perimeter, pixel count, cover ratio and ratio between perimeter and surface of the parcel is concatenated to the PSE outputs, providing the geometrical information that is lost during pixel sampling.

B. Temporal Encoder

In the original Transformer, Vaswani et al. [21] defined a set of three vectors, the $query$ $q(t)$, the $keys$ $k(t)$ and the $values$ $v(t)$ vectors. Each vector is computed by processing the input sequence by three fully connected layers. From a retrieval system perspective, the self-attention mechanism can be viewed as the procedure of mapping a query request against a set of keys describing the entirety of the content, represented by the values vector and retrieving the best matching data. The output of a self-attention module is the sum of previous values, weighted by the probability distribution of the dot product between the query and keys. The computation is done in parallel by H attentional heads, where each head learns to specialize for different temporal positions in a sequence. Moreover, Vaswani et al. introduced positional information to the input sequence by adding a positional sinusoidal encoding tensor to each element, based on an element's positional index. Garnot et al. [2] proposed instead to obtain only the keys with a linear layer, whereas the query was a learnable model parameter, not obtained from the input. The PSE outputs are served directly as the values, since they are learnt alongside the attention module during training. For positional encodings, the number of days since the first observation was used, instead of the sequence index. Finally, they also proposed a channel grouping technique by splitting the input sequence into H groups of size $E' = \frac{E}{H}$.

In our approach, let $E = [e^1, ..., e^T]$ denote the PSE output for $t \in 1, ..., T$ timesteps, $Q = [q_1, ..., q_H]$ the query vector for h attentional heads, $h \in 1, ..., H$ and $K = [k_1^1, ..., k_H^T]$ the keys. For a head h, we calculate the query as follows

$$q_h = G_h(E + POS) \tag{1}$$

Authorized licensed use limited to: Centre for Research and Technology (C.E.R.T.H.). Downloaded on September 27,2023 at 11:00:02 UTC from IEEE Xplore. Restrictions apply.

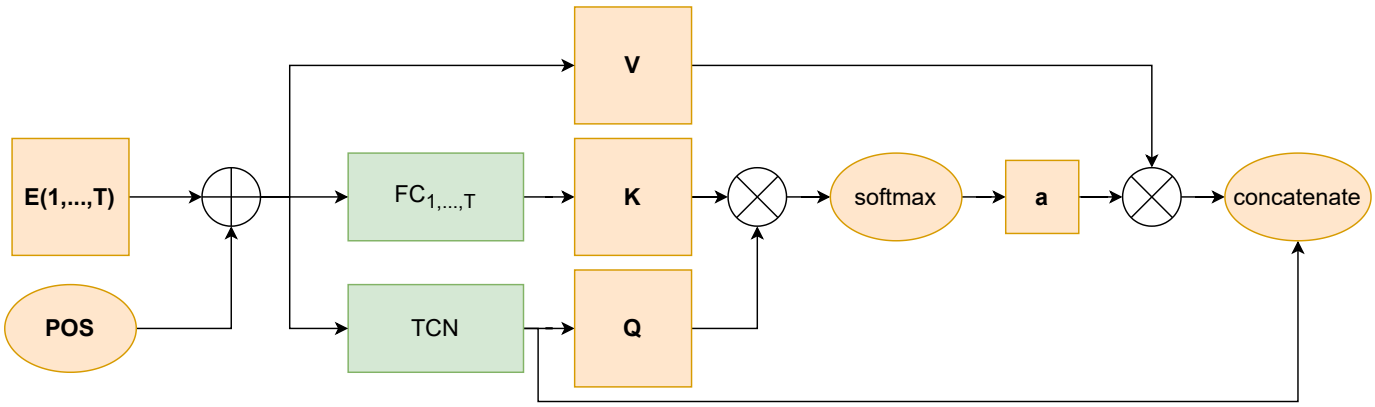

Fig. 2: The proposed TCN-TAE temporal module. The input E are the output PSE features, which are then processed by a linear layer to produce the keys K and a TCN to produce the query Q. The values V are weighted by the attention mask a and the resulting vector is concatenated with the query.

where $POS = [pos(1), ..., pos(T)]$ are the positional encodings, calculated as

$$pos(t) = sin(\frac{day(t)}{T^{\frac{i}{E'}}}), i \in [1, ..., E'] \tag{2}$$

and G_h is a TCN, calculating the query for each head. The keys are as follows:

$$k_h^{1,...,T} = FC_h(E + POS) \tag{3}$$

with FC_h denoting a fully connected layer. The attention score a is then calculated as:

$$a = softmax(\frac{QK^T}{\sqrt{d_k}}), a \in [0, 1]^T \tag{4}$$

where d_k is the dimension of the query and keys.

The self-attention output is then calculated as

$$out_{attn} = aV \tag{5}$$

with V being the values, calculated as

$$V = E + POS \tag{6}$$

The temporal encoder TCN-TAE output, is the concatenated vectors of the self-attention output with the query, obtained from the TCN.

$$out_t = concatenate(out_{attn}, Q) \tag{7}$$

It should be noted that G_h produces a query with a singleton temporal dimension, that when combined with the keys the resulting attention score represents a weighted summary over the values. Thus, the produced output vector describes a parcel over the span of multiple observations during a time period. The PSE-TCN-TAE model can be viewed in Figure 1. Finally, a multi-layer perceptron MLP is used to project $out_{temporal}$ to the classes space, yielding the final classification for each parcel.

C. Temporal Differences

Leveraging extra modalities was shown to improve the model's accuracy in tasks such as image and text classification [6], image to image translation [25] or 3D hand pose estimation [24]. Motivated by this, we defined a pseudo multimodal approach, where an additional input modality is created from the multi-spectral satellite time series, containing the differences between consecutive time steps. For an input satellite images time series $X(t)$, $t = 1, ..., T$, the temporal differences $D(t)$, $t = 2, \ldots, T$ are calculated as

$$D(t) = X(t) – X(t-1) \tag{8}$$

In order to effectively utilize them for the task of satellite images crop classification, the constructed temporal differences are processed by a different PSE-TCN-TAE model, creating a second stream of deep features that are concatenated with the PSE-TCN-TAE output features. The final multimodal output (o_m) to be provided to the decoder (classifier) is then computed as

$$o_m = relu(BN(concatenate(o_t, o_d))) \tag{9}$$

where o_t, o_d are the outputs of the temporal and difference streams respectively and BN stands for 1D Batch Normalization. By combining the original input with their temporal differences, the final model is more capable to recognise the crop phenology.

III. Experimental Results

A. Dataset and Metrics

The experimental evaluation was performed on the publicly available dataset Sentinel2-Agri [3]. Sentinel2-Agri is a dataset comprised of 191,703 temporal sequences of 24 superspectal images for each parcel, with the area of interest (AOI) spanning across a 12,100 km^2 area in southern France. The images were captured from January 2017 until October 2017. As the PSE module forms a tensor from all pixels contained in each parcel, spatial structures are lost, so geometrical features were

Authorized licensed use limited to: Centre for Research and Technology (C.E.R.T.H.). Downloaded on September 27,2023 at 11:00:02 UTC from IEEE Xplore. Restrictions apply.

precomputed and stored beforehand. From the 13 available spectral bands, 10 were used (the atmospheric bands B1, B9 and B10 were discarded). It should be noted that this dataset is highly imbalanced, with four out of the 20 crop classes covering 90% of the samples. We use the same 5-fold cross-validation scheme as the authors of PSE-LTAE [2] and report on Overall Accuracy (OA) and the mean per class Intersection-over-Union (mIoU) as well as the standard deviation between folds. Given the imbalanced nature of Sentinel2-Agri, the mIoU is a more suitable metric for classification comparisons.

B. Implementation Details

All the models presented in this work are implemented in PyTorch [13]. We use the Adam optimizer [7] with learning rate 10^{-3} and batch size 128 and train the models for 150 epochs, using focal loss [9], since there are very dominant classes in the dataset (four classes characterize 90% of the data.

For the temporal experiments, the 1D convolutions module is comprised of an 1D convolutional layer with kernel size 5 and stride 2, followed by batch normalization, ReLU and a max pool layer with kernel size 2 and stride 2. An additional 1D convolution is applied with kernel size 5 and stride 1. The TCN module has kernel size 4 and reduces the temporal dimension gradually by a factor of 2 at each layer. The RNN experiment is implemented by introducing a bidirectional LSTM layer with hidden state size equal to 128. Finally, in order to leverage the temporal differences between consecutive time steps as an extra pseudo modality, we pass the created input from an additional network of the same topology as as the network processing the original data.

C. Ablation Study

The temporal part of satellite image time series contains significant information regarding a crop's type, given the cyclical nature of vegetation life. As such, we perform an ablation study regarding the contribution of the proposed architectural changes to the temporal self-attention mechanism and evaluate the importance of each stream to the crop classification task.

In order to summarize the temporal dimension to a single vector and meaningfully capture the crop phenology across multiple observations of a parcel, the PSE-LTAE architecture defined a master attentional query vector per head (master Query), as an independent model parameter (Query as Parameter) that is not constructed from the input data. Arguably, constructing the master Query from the input PSE features can still be beneficial, therefore we analysed different ways to construct it. More specifically:

- Given that the last hidden state of an RNN contains the information of an entire sequence, an LSTM's last hidden state was employed as the Query.
- The authors of [3] constructed the master Query by taking the mean value across the temporal dimension of the queries. Instead, 1D temporal convolutions and Temporal Convolutional Networks (TCNs) [8] were utilized to construct the master Query, as they can create more coherent temporal features.
- The proposed TCN-TAE module that leverages the TCN output for master Query construction as well at the final temporal output by combining it with the attentional features was evaluated.

Temporal module experiments	Overall Accuracy	Mean IoU
Last LSTM hidden state as master Query	94.14 $\pm$0.11	50.7 $\pm$0.72
1D convolutions for master Query construction	94.04 $\pm$0.05	51.02 $\pm$1.09
TCN for master Query construction	94.2 $\pm$0.16	51.84 $\pm$0.78
PSE-TCN-TAE	94.23 $\pm$0.1	52.52 $\pm$0.45
PSE-1D convolutions	93.2 $\pm$0.21	49.16 $\pm$0.8
PSE-TCN	93.63 $\pm$0.32	47.7 $\pm$0.73
Temporal differences PSE-TCN-TAE	93.46 $\pm$0.22	47.8 $\pm$0.27
Two stream PSE-TCN-TAE	**94.31** $\pm$0.11	**53.66** $\pm$0.5

TABLE I: Ablation study of the different temporal modules on Sentinel2-Agri.

Additionally we considered examining the contribution of each of the proposed modifications by running them in isolation, resulting in the following experiments:

- TCNs consist of dilated causal 1D convolutions, which allows for large receptive fields while restricting access to future time steps. In order to evaluate the performance of the independent TCN output to the crop classification, a TCN architecture was employed for temporal encoding instead of the self-attentional module, while keeping the same spatial module, resulting in a PSE-TCN architecture.
- 1D convolutions have proven to be capable of modeling temporal data [12] and have been recently proposed for the task of crop type mapping [14]. For comparison reasons with the TCN experiment, the PSE module was utilized for spatial encoding while the temporal module was constructed with two 1D convolutions layers, followed by batch normalization and ReLU (PSE-1D convolutions).
- The contribution of the temporal differences was evaluated by passing them from a PSE-TCN-TAE model and performing the classification without employing the original data.
- Finally, the proposed PSE-TCN-TAE model with the auxiliary temporal differences was evaluated in comparison to the above.

In Table I we present a comparison of the temporal modules performance. Regarding the construction of the master Query, we can observe that the TCN is the best suited between the LSTM and the 1D convolutions approaches to summarize the temporal information of the satellite images time series. However, it is still inferior to the "Query as Parameter" of Garnot et al. [2], which is fully capable to summarize the temporal dimension of the satellite time series, even though it is not derived from the input data, as in the original self-attention architecture. Despite that, the proposed use of the

Authorized licensed use limited to: Centre for Research and Technology (C.E.R.T.H.). Downloaded on September 27,2023 at 11:00:02 UTC from IEEE Xplore. Restrictions apply.

TCN output, not only as the master Query, but also to augment the attentional features at the temporal module output level, outperforms the "Query as Parameter" approach, suggesting that the TCN might be able to learn better representations to summarize the temporal sequence.

Additionally, by replacing the temporal module with 1D temporal convolutions or a TCN, there is a significant deterioration in performance. The results indicate that these temporal architectures are not as capable to capture the phenology of the crops compared to the self-attentional approaches. Moreover 1D convolutions and TCNs require a careful selection of kernels and stride size for the receptive fields of the convolution filters, which might be the reason behind their poor performance. The temporal differences alone also seem less efficient in capturing the crop phenology, providing significantly worse results compared to using the original input satellite time series. However, the proposed PSE-TCN-TAE when coupled with the temporal differences stream achieves 94.32 in Overall Accuracy and 53.7 in mean IoU outperforming by a large margin the previous architectures as well as the single stream PSE-TCN-TAE. These results suggest that the temporal differences may carry useful information regarding the spectral bands' rate of change and therefore the rate of change of a crop's phenology, that even though they do not suffice to be used alone for the classification process, they provide valuable extra temporal information that is characteristic for a satellite time series. In this regard, the temporal differences serve a similar role as the extra geometrical features that are provided to the PSE output features, while also being more flexible, as they do not have to be explicitly included in a dataset.

D. Results and Comparison with State-of-the-art

A comparison of the PSE-TCN-TAE approach with current state-of-the-art methods is presented in Table II for the Sentinel2-Agri dataset. The proposed two stream PSE-TCN-TAE outperforms the previous state-of-the-art PSE-LTAE by 1.76 points on the more appropriate mean IoU metric (given the significant class imbalance of the dataset) achieving 53.66 points while being on par in regards to Overall Accuracy. The single stream PSE-TCN-TAE is also able to outperform PSE-LTAE in mean IoU by 0.62 points, achieving 52.52 points while being slightly worse in Overall Accuracy by 0.08 points.

Method	Overall Accuracy	Mean IoU
PSE-TAE [3]	94.22 $\pm$0.04	50.64 $\pm$0.75
PSE-LTAE [2]	94.31 $\pm$0.1	51.9 $\pm$0.65
Rußwurm et al. [17]	92.3 $\pm$0.3	43.1 $\pm$1.1
PSE-TCN-TAE	94.23 $\pm$0.1	52.52 $\pm$0.45
Two-stream PSE-TCN-TAE	**94.31** $\pm$0.11	**53.66** $\pm$0.5

TABLE II: Comparison with state-of-the-art architectures on Sentinel2-Agri.

IV. Conclusion

In this paper, we presented a novel two stream PSE-TCN-TAE method that proposes modifications to the original transformer as well as the previous state-of-the-art in crop classification from satellite time series, while also proposing the addition of a pseudo time series derived from the original data, as an extra modality containing characteristic information regarding the rate of change of a time series. The proposed single stream PSE-TCN-TAE model was employed to create deep features for both the original time series as well as the extra modality and their combination achieved state-of-the-art crop classification performance on the Sentinel2-Agri public dataset. Moreover, we conducted a thorough ablation study and examined various architectural changes regarding the temporal module, in order to access the contribution of each proposed modification to the classification results.

Acknowledgment

This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project codes:T1EDK-01577 ARTEMIS and T2EDK-04396 Smart-BeeKeep).

References

[1] R Devadas, RJ Denham, and M Pringle. Support vector machine classification of object-based data for crop mapping, using multi-temporal landsat imagery. *International archives of the photogrammetry, remote sensing and spatial information sciences*, 39(1):185–190, 2012.

[2] Vivien Sainte Fare Garnot and Loic Landrieu. Lightweight temporal self-attention for classifying satellite images time series. In *International Workshop on Advanced Analytics and Learning on Temporal Data*, pages 171–181. Springer, 2020.

[3] 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.

[4] Jordi Inglada, Marcela Arias, Benjamin Tardy, Olivier Hagolle, Silvia Valero, David Morin, Gérard Dedieu, Guadalupe Sepulcre, Sophie Bontemps, Pierre Defourny, et al. Assessment of an operational system for crop type map production using high temporal and spatial resolution satellite optical imagery. *Remote Sensing*, 7(9):12356–12379, 2015.

[5] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[6] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*, 2019.

[7] Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[8] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017.

[9] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[10] Nando Metzger, Mehmet Ozgur Turkoglu, Stefano D'Aronco, Jan Dirk Wegner, and Konrad Schindler. Crop classification under varying cloud cover with neural ordinary differential equations. *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[11] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.

[12] Ilias Papastratis, Kosmas Dimitropoulos, Dimitrios Konstantinidis, and Petros Daras. Continuous sign language recognition through cross-modal alignment of video and text embeddings in a joint-latent space. *IEEE Access*, 8:91170–91180, 2020.

Authorized licensed use limited to: Centre for Research and Technology (C.E.R.T.H.). Downloaded on September 27,2023 at 11:00:02 UTC from IEEE Xplore. Restrictions apply.

[13] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[14] 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.

[15] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[16] Marc Rußwurm and Marco Korner. Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 11–19, 2017.

[17] Marc Rußwurm and Marco Körner. Self-attention for raw optical satellite time series classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169:421–435, 2020.

[18] Atharva Sharma, Xiuwen Liu, and Xiaojun Yang. Land cover classification from multi-temporal, multi-spectral remotely sensed imagery using patch-based recurrent neural networks. *Neural Networks*, 105:346–355, 2018.

[19] Sofia Siachalou, Giorgos Mallinis, and Maria Tsakiri-Strati. A hidden markov models approach for crop classification: Linking crop phenology to time series of multi-sensor remote sensing data. *Remote Sensing*, 7(4):3633–3650, 2015.

[20] Pengfei Tang, Peijun Du, Junshi Xia, Peng Zhang, and Wei Zhang. Channel attention-based temporal convolutional network for satellite image time series classification. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[22] Liheng Zhong, Lina Hu, and Hang Zhou. Deep learning based multi-temporal crop classification. *Remote sensing of environment*, 221:430–443, 2019.

[23] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8739–8748, 2018.

[24] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5346–5355, 2020.

[25] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 465–476. Curran Associates, Inc., 2017.

Authorized licensed use limited to: Centre for Research and Technology (C.E.R.T.H.). Downloaded on September 27,2023 at 11:00:02 UTC from IEEE Xplore. Restrictions apply.