applied sciences

MDPI

Article

A Spatio-Temporal Encoding Neural Network for Semantic Segmentation of Satellite Image Time Series

Feifei Zhang †, Yong Wang *,†, Yawen Du and Yijia Zhu

School of Computer Science, China University of Geosciences, Wuhan 430074, China; ffzhang@cug.edu.cn (F.Z.); duyawen@cug.edu.cn (Y.D.); zhuyijia@cug.edu.cn (Y.Z.)

* Correspondence: yongwang@cug.edu.cn; Tel.: +86-1860-273-0899

† These authors contributed equally to this work.

Abstract: Remote sensing image semantic segmentation plays a crucial role in various fields, such as environmental monitoring, urban planning, and agricultural land classification. However, most current research primarily focuses on utilizing the spatial and spectral information of single-temporal remote sensing images, neglecting the valuable temporal information present in historical image sequences. In fact, historical images often contain valuable phenological variations in land features, which exhibit diverse patterns and can significantly benefit from semantic segmentation tasks. This paper introduces a semantic segmentation framework for satellite image time series (SITS) based on dilated convolution and a Transformer encoder. The framework includes spatial encoding and temporal encoding. Spatial encoding, utilizing dilated convolutions exclusively, mitigates the loss of spatial accuracy and the need for up-sampling, while allowing for the extraction of rich multi-scale features through a combination of different dilation rates and dense connections. Temporal encoding leverages a Transformer encoder to extract temporal features for each pixel in the image. To better capture the annual periodic patterns of phenological phenomena in land features, position encoding is calculated based on the image's acquisition date within the year. To assess the performance of this framework, comparative and ablation experiments were conducted using the PASTIS dataset. The experiments indicate that this framework achieves highly competitive performance with relatively low optimization parameters, resulting in an improvement of 8 percentage points in the mean Intersection over Union (mIoU).

Keywords: semantic segmentation; phenology; spatial encoding; temporal encoding; satellite image time series

Citation: Zhang, F.; Wang, Y.; Du, Y.; Zhu, Y. A Spatio-Temporal Encoding Neural Network for Semantic Segmentation of Satellite Image Time Series. *Appl. Sci.* **2023**, *13*, 12658. https://doi.org/10.3390/app132312658

Academic Editors: Ran Tao, Jinfeng Du, Joni A. Downs and Zhaoya Gong

Received: 30 October 2023
Revised: 22 November 2023
Accepted: 22 November 2023
Published: 24 November 2023

1. Introduction

Modern remote sensing satellites possess unprecedented high-frequency access capabilities, thereby obtaining multi-temporal remote sensing imagery that encapsulates vast amounts of information about features on the Earth's surface in both space and time [1]. Fully harnessing this spatio-temporal information will further enhance the applications of remote sensing imagery in environmental monitoring, urban planning, and agricultural land classification.

Food security serves as a crucial foundation for national security, and arable land stands as the lifeline of food production. As a result, governments worldwide have consistently placed great emphasis on the supervision of arable land protection and have implemented strict measures to ensure the sustainable utilization of arable land resources [2]. However, the "non-grainification" [3] phenomenon in arable land, resulting from structural adjustments in land use within agricultural areas, is challenging to capture in single-temporal remote sensing images. This limitation often leads to unsatisfactory outcomes in arable land segmentation. To better monitor arable land protection and effectively manage arable land resources, it is necessary to employ multi-temporal remote sensing imagery

for comprehensive analysis [4,5]. This paper examines satellite image time series (SITS) to acquire information about crop growth, enabling the long-term monitoring of arable land and elevating the level of agricultural intelligence.

In recent years, the use of deep learning methods for semantic segmentation of remote sensing images has become the mainstream approach for land cover classification. Researchers in this field have conducted in-depth investigations on single-temporal remote sensing images and have achieved remarkable results [6,7]. Wang et al. [8] integrated the Atrous Spatial Pyramid Pooling (ASPP) module, which encodes image-level features, into the U-Net [9] network, significantly improving the segmentation accuracy of multi-scale features in the imagery. However, mainstream segmentation networks are unable to fully recover spatial information discarded in the feature extraction stage, which exacerbates the segmentation inaccuracies caused by fuzzy land feature boundaries in the imagery. To address this, Xu et al. [10] introduced the HRNet network, which better preserves spatial information during feature extraction, enhancing the accuracy and precision of segmentation, while also considering global context and multi-scale features. Compared to natural images, lower-resolution remote sensing imagery relies more on contextual information for pixel-level classification. Ding et al. [11] introduced the parcel attention module to enhance the extraction of contextual information through local attention, and the attention embedding module to fuse semantic information at different levels. The remote sensing field has continuously been seeking new methods to enhance the accuracy and efficiency of remote sensing image segmentation. With the tremendous success of Transformer models in the field of computer vision, applying them to remote sensing image segmentation has become a trend. By studying the performance of remote sensing images within Transformer models, we can gain a deeper understanding of how Transformer models work and their characteristics, further promoting their application and development in the remote sensing domain [12]. To fully leverage the capabilities of convolutional neural networks (CNNs) in local feature extraction and the advantages of the Transformer [13] in capturing global contextual information, the fusion of both is a viable choice. Wang et al. [14] proposed the CCTNet, which combines the local detail features from the CNN and the global contextual information from the Transformer, effectively mitigating the misclassifications of small objects and gaps in the imagery. Zhang et al. [15] used the Swin Transformer as an encoder to extract features, while the CNN was employed for multi-scale feature extraction and served as the decoder. Li et al. [16] introduced the dual-encoder structure, which utilizes the Transformer to cross-fuse global multi-scale semantic information and employs ASPP to extract context information with high-level semantics. GLFFNet [17] is another approach that combines two different network architectures, the Transformer and CNN, to extract global high-level interaction features and obtain low-level local features. He et al. [18] embedded the Swin Transformer into the classical U-Net network, creating a dual-encoder structure wherein the extracted global contextual information compensates for the local nature of the convolution.

As described in the aforementioned literature, local information, multi-scale features, and global context all play indispensable roles. However, the diverse phenological phenomena [19–21] of crops should not be overlooked. Their complex temporal patterns can be accurately represented through SITS [22]. Therefore, for the segmentation of agricultural land in remote sensing images, relying solely on spatial and spectral features is not sufficient. It is also crucial to consider the annual cyclic phenological characteristics of crops. Moreover, capturing the phenological characteristics of various land features as they change over time in SITS can effectively mitigate the challenges posed by the "same material, different spectra; different material, same spectra" issue encountered in single-temporal remote sensing imagery. This approach also enhances the utility of remote sensing imagery.

In the early stages of land cover classification using SITS, the conventional approach involved extracting time, statistical, and spectral index information for each pixel from the imagery to create time series. These data were then combined with classical machine learn-

ing algorithms for classification [2,19,23,24]. Garnot et al. [25] introduced a method based on ensemble encoders that successfully extracted statistical information about spectral distribution within the spatial extent of land parcels. They organized this statistical data into time series and applied a time attention encoder for classification. The aforementioned methods, utilizing feature engineering, made use of time information to a certain extent. However, with the continuous development of deep learning technology, an increasing number of studies have begun to explore the use of deep learning models to extract time features from SITS. U-TAE [1] represents the first end-to-end pixel-level segmentation framework for agricultural land within SITS. This framework includes an encoder for image time series, which employs convolutional blocks and a lightweight time attention encoder (L-TAE) to extract rich multi-scale spatio-temporal features. Additionally, they publicly released the first SITS dataset with semantic annotations, known as PASTIS [1]. While the use of L-TAE significantly reduces computational complexity, it fundamentally relies on learned weights for the fusion of feature map time series and does not consider or extract the time series features within SITS.

In this paper, we introduce a novel end-to-end pixel-level semantic segmentation framework for SITS called the spatio-temporal encoding neural network (STENN). The framework is designed to reveal the representation of intrinsic temporal patterns of land features in images, addressing the limitations of previous methods in exploiting the potential information along the temporal dimension of images and, consequently, the underutilization of spatio-temporal information. The framework comprises two encoding modules: a dilated convolution-based spatial encoding module and a Transformer encoder-based temporal encoding module. In the spatial encoding module, we utilize dilated convolutions to rapidly increase the receptive field, capture deep semantic information, and circumvent spatial information loss [26] and up-sampling caused by down-sampling. Within the temporal encoding module, we employ the Transformer encoder to extract temporal features for each pixel in the feature map time series and perform pixel-wise classification based on the spatio-temporal features. Finally, we validate the effectiveness of our proposed model on the PASTIS dataset.

The main contributions of this paper are as follows:

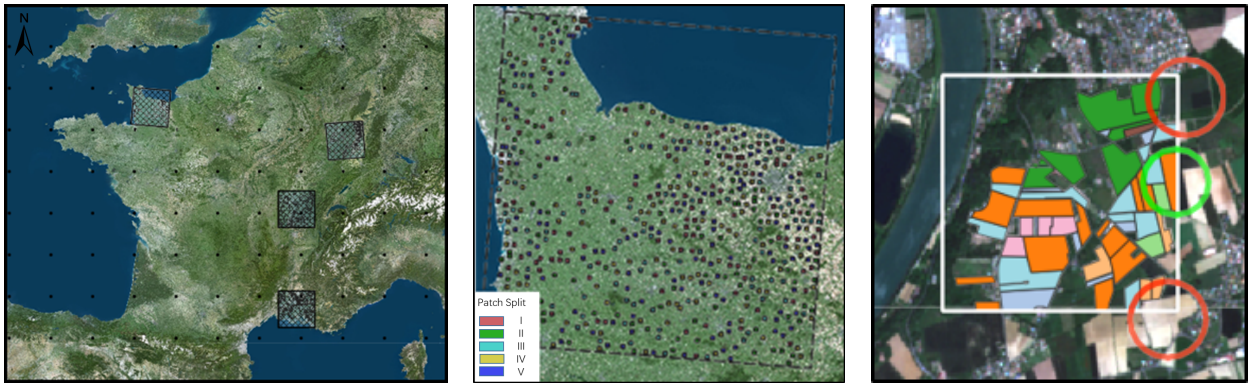
1. We propose a novel backbone network for semantic segmentation tailored for pixel-level classification tasks sensitive to spatial information. This network abandons down-sampling, which may lead to unrecoverable spatial information loss, and instead combines dilated convolutions and dense connections to rapidly expand the receptive field and obtain multi-scale features.
2. For the first time, we utilize a Transformer encoder to extract temporal features for each pixel. To emphasize the annual cyclic patterns of crops, position encoding is calculated based on the position of the acquisition date within that year.
3. We provide an open-source implementation of the model based on PyTorch on GitHub, which can be found at the following URL: https://github.com/ThinkPak/stenn-pytorch (accessed on 25 October 2023).

The subsequent sections of this paper are organized as follows: In Section 2, we will provide a detailed introduction to the PASTIS dataset, including information about the study area and the organizational structure of the dataset. In Section 3, we will outline the research background of the proposed model and provide detailed explanations of the model's two encoders: the spatial encoder and the temporal encoder. Section 4 will cover the evaluation metrics used, experimental details, and the results obtained. In Section 5, we will extensively discuss the experimental results, focusing on key observations and insights gained from analyzing the SITS data. Finally, in Section 6, we will summarize the main contributions of this paper, address its limitations, and provide recommendations for future research.

2. Study Area and Dataset

2.1. Study Area

The SITS data for PASTIS are obtained from four distinct Sentinel-2 tiles situated in various regions within the French metropolitan territory, as illustrated in Figure 1a. These regions exhibit diverse climates and crop distributions. The Sentinel tiles cover an area of more than (4000 km^2) with a spatial resolution of 10 m per pixel. Each pixel is characterized by 13 spectral bands. For the PASTIS dataset, we utilize all bands except the atmospheric bands B01, B09, and B10. Each of these tiles is divided into square patches measuring 1.28 km $\times$ 1.28 km, resulting in a total of approximately 24,000 patches. Only 2433 patches were chosen, constituting 10% of all the patches. The 2433 selected patches were randomly divided into five splits, enabling us to conduct cross-validation, as illustrated in Figure 1b. The selection criteria prioritize patches that include rare crop types to mitigate the severe class imbalance present in the dataset. As shown in Figure 1.



(a) Location of the four tiles. (b) Selected patches. (c) The parcel retention principle.

Figure 1. PASTIS data information. The spatial distribution of the four parcels corresponding to the PASTIS dataset (**a**). A patch distribution within a parcel (**b**). The patch retention principle (**c**); if most of the land features are located outside the patch boundary (red circle), they are labeled as "void". If most of the land features are located inside the patch boundary (green circle), the patch is considered a valid parcel and retained. The legend refers to Figure 2 for reference.

Label and Color	Class Name	Number of parcels
0	Background	-
1	Meadow	31,292
2	Soft winter wheat	8,206
3	Corn	13,123
4	Winter barley	2,766
5	Winter rapeseed	1,769
6	Spring barley	908
7	Sunflower	1,355
8	Grapevine	10,640
9	Beet	871
10	Winter triticale	1,208
11	Winter durum wheat	1,704
12	Fruits, vegetables, flowers	2,619
13	Potatoes	551
14	Leguminous fodder	3,174
15	Soybeans	1,212
16	Orchard	2,998
17	Mixed cereal	848
18	Sorghum	707
19	Void label	35,924

Figure 2. Land class codes and their corresponding colors, along with the number of parcels corresponding to each land class.

2.2. Dataset

Based on the distribution of crop categories contained in the selected parcel samples, we chose 18 categories for the parcel recognition system in France. As shown in Figure 2, parcels that did not belong to these 18 categories were labeled as "void".

In Figure 3, a yearly image time series reveals the annual cyclical changes in crops and also highlights how different land features have similar representations in single-temporal remote sensing images. Based on these observations, this study selects SITS as the focus of research for agricultural land semantic segmentation.

The dataset used in this study is PASTIS, which comprises SITS and labels for the semantic segmentation of agricultural land. The image data were collected using Sentinel-2 and consist of 2433 multi-spectral image sequences. Each image in the sequence comprises 10 channels, and during training, normalization is required for each channel of the image. Each image consists of 128 × 128 pixels. The data were collected from September 2018 to November 2019, with varying observation frequencies ranging from 38 to 61 times and an average interval of 5 days. Each multi-spectral image sequence is organized into a four-dimensional tensor of shape $T \times 10 \times 128 \times 128$, where T represents the number of observations ($T \in [38, 61]$). There are 19 annotated classes in total, including 18 different crop classes and one background class. The class labels were sourced from a publicly available land parcel identification system in France, with an accuracy rate exceeding 98%.

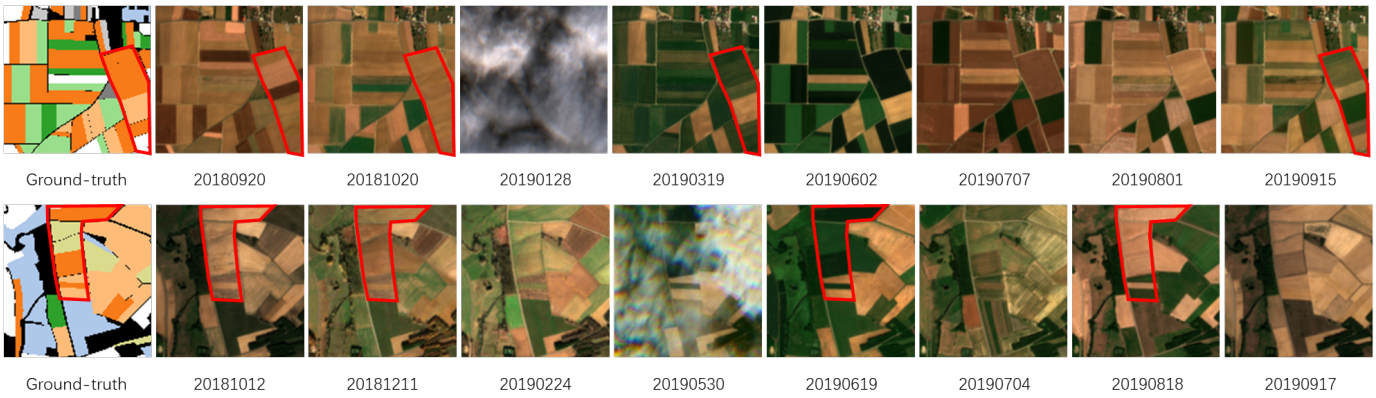


Figure 3. This image excellently illustrates the annual cyclic variations of crops in the SITS dataset. From the image, it is evident that even though different crops may exhibit similar spectral characteristics in single-temporal remote sensing images (as indicated by the red circles), their patterns of change throughout the annual cycle vary significantly. This cyclical variation not only reflects the unique growth habits of various crops but also provides us with rich information for understanding and predicting the dynamics of agricultural production.

3. Methodology

3.1. Related Work

In this section, we have introduced some of the state-of-the-art works related to spatial encoding and temporal encoding, including models based on CNN and models based on Transformers. These models have provided valuable insights and references for us in extracting spatio-temporal features from SITS.

Based on CNN spatial encoding. The conventional approach to semantic segmentation involves using pretrained image classification networks such as VGG [27] and ResNet [28] as backbone networks. These networks are characterized by multiple rounds of convolutional layers and down-sampling layers, which expedite the expansion of the receptive field and, consequently, the extraction of deep but low-resolution semantic information from images. While this semantic information is beneficial for image classification tasks, it is not suitable for dense pixel-level classification tasks. To adapt these networks for semantic segmentation, it is necessary to compensate for the loss of spatial resolution caused by down-sampling. Primarily in the decoding phase, this is achieved by integrating

mid- to high-resolution feature maps generated at various stages of the encoder using skip connections, as seen in typical networks like FCN [29] and U-Net [9]. FCN, as the first network to perform end-to-end segmentation using fully convolutional layers, revolutionized research in semantic segmentation. The U-Net, originally designed for medical image segmentation, employs a symmetric network structure, comprising a contracting pathway to capture contextual information and an expanding pathway symmetric to it to support precise localization. By combining deep, low-resolution semantic information with shallow, high-resolution surface information through skip connections, U-Net is capable of performing dense and fine-grained segmentation tasks. Consequently, convolution has become the foundation for semantic segmentation tasks [30], and the encoder–decoder architecture has become a popular framework for semantic segmentation networks. The encoder is responsible for feature extraction, and the decoder projects the high-level semantic features learned by the encoder into high-resolution pixel space to achieve dense, fine-grained classification. For example, SegNet [27] utilizes the index of max-pooling from the encoding phase in its up-sampling process, reducing the number of parameters and computational workload compared to transposed convolution. However, down-sampling in classification networks results in irreversible spatial information loss. DeepLab V1 [31] introduced dilated convolutions to mitigate the reduction in spatial resolution and spatial insensitivity caused by down-sampling. DeepLab V2 [32] proposed the ASPP module for multi-scale targets, consisting of four parallel dilated convolutions modules with different dilation rates to extract multi-scale features. In addition to adjusting the dilation rates of the ASPP module, DeepLab V3 [33] added a global pooling layer to extract image-level features. DeepLab V3+ [34], building upon DeepLab V3's encoder, introduced depth-wise separable convolution to reduce the number of parameters and included a decoder module to refine object boundaries, significantly enhancing the network's segmentation performance. While skip connections and up-sampling gradually restore deep semantic information to high-resolution space, the loss of spatial information in feature maps due to down-sampling is an irreversible process, and the lost high-resolution spatial information cannot be fully recovered.

Image classification tasks involve making classification decisions by comprehensively considering the information across the entire image. Therefore, down-sampling can be utilized to expedite the enlargement of the receptive field, enabling the extraction of more abstract and robust semantic information without excessive concern for spatial resolution loss. However, for semantic segmentation tasks, which require precise classification for each pixel in an image, spatial precision is of utmost importance. As a result, we believe that improvements should not be limited to enhancing image classification networks but should focus on constructing networks tailored to semantic segmentation, a task highly sensitive to spatial information.

Dilated convolutions were introduced to address image segmentation problems. Unlike the combination of convolutional and pooling layers, dilated convolutions have the advantage of rapidly increasing the receptive field while maintaining the feature map's size, thus avoiding the irreversible loss of spatial accuracy caused by down-sampling [33]. To mitigate gradient vanishing and enhance the utilization of feature maps, the output of each dilated convolution block is directly concatenated with the input of all subsequent dilated convolution blocks, without the need for up-sampling to ensure consistent feature map sizes [35]. Since the input of each dilated convolution block contains the original image, and the dilation rates increase progressively, the network can extract a rich set of multi-scale features to cope with the significant variation in object sizes within remote sensing imagery.

Based on Transformer temporal encoding. Research into time series data classification initially focused on discovering specific patterns within the time series data, such as periodic and trending patterns, and performing classification by identifying these patterns. These methods have been superseded by more effective deep learning techniques that allow neural networks to automatically learn complex features within time series data,

eliminating the need for manual feature engineering and selection. Models like LSTM [36] and Transformer [13] can capture dependencies and dynamic changes within time series data, enabling more accurate classification. Compared to recurrent neural network models, Transformers are more proficient at modeling long-range dependencies between elements in input sequences and support parallel processing of sequences, greatly enhancing model performance and efficiency. In particular, the Transformer encoder, through positional encoding and self-attention calculations, captures correlations between time series data. During self-attention calculations, the data are divided into multi-head, creating various subspaces. This enables the model to capture information from different perspectives, resulting in a more comprehensive understanding of the data. Finally, the information from these subspaces is integrated, and the encoded results are fed into a classifier for classification, as depicted in Figure 4.

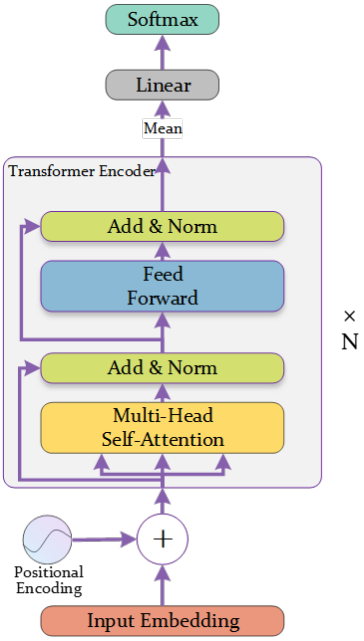


Figure 4. Transformer encoder sequence data classification model includes an embedding layer, positional encoding, multi-head self-attention layer, feedforward fully connected layer, and linear layer.

In this study, we primarily explore the use of dilated convolutions as a substitute for down-sampling operations, facilitating the rapid expansion of the receptive field. We achieve this by employing dense connections and varying dilation rates to obtain multi-scale features. Lastly, we employ the Transformer encoder to extract temporal features for each pixel, which are utilized for pixel-level classification.

3.2. STENN Architecture

In this section, we primarily introduce a shared dilated convolution spatial encoder for SITS and a Transformer encoder for time encoding. The input consists of SITS X, which is a four-dimensional tensor with a shape of $T \times C \times H \times W$. Here, T represents the length of the time series, C is the number of channels in the image, and $H \times W$ denotes the shape of the image.

As shown in Figure 5, our STENN model encodes the sequence X in three steps:

- First, it undergoes spatial encoding. We employ multi-level shared dilated convolutional layers, simultaneously processing each frame of the SITS, generating feature map time series with channel numbers of 16, 32, 32, and 64.
- Next is temporal encoding. We concatenate the results of spatial encoding along the channel dimension, obtaining a feature map time series with 144 channels. Then, we reshape this sequence into a time series for each pixel, and after positional encoding,

it is fed into the Transformer encoder. Subsequently, the encoded results are averaged along the time dimension, ultimately generating a single feature map.

- Finally, through the semantic segmentation head, we map the feature map containing spatio-temporal information into segmentation results.

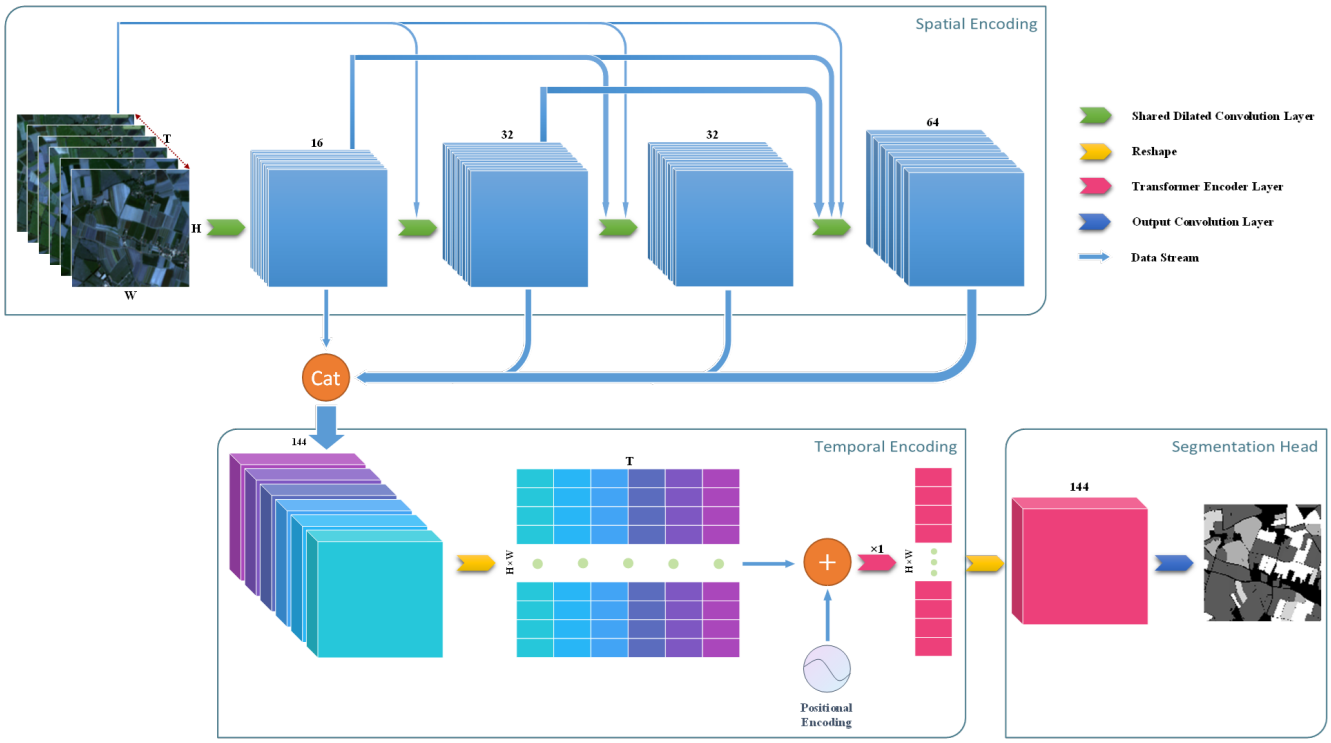


Figure 5. STENN consists of both a spatial encoding and a temporal encoding. The spatial encoding abandons down-sampling, ensuring consistent feature map sizes, and simultaneously enhances feature map utilization through dense connections, enabling the extraction of rich multi-scale features.

3.2.1. Spatial Encoding with Dilated Convolution

Dilated Convolution: Dilated convolutions were introduced for image segmentation tasks, and they ensure that model parameters and feature map resolution remain unchanged while effectively controlling the receptive field of convolution to capture multi-scale information. Compared to the common approach of using pooling layers and convolution layers to increase the receptive field, dilated convolutions avoid the irreversible loss of spatial precision that occurs when feature maps are first reduced and then enlarged. Additionally, by setting different dilation rates and using dense connections, it is possible to capture multi-scale features.

Dense Connection: Dense connections, pioneered by DensNet [22], are a crucial strategy for information exchange between network layers. This strategy primarily involves passing the output of each layer directly to all subsequent layers, fully utilizing the feature maps of each layer and reducing feature redundancy.

Spatial encoding is carried out by the dilated convolution encoder ε, which consists of N layers $(1, \ldots, n, \ldots, N)$. Each layer is composed of a series of dilated convolutions with increasing dilation rates, rectified linear unit (ReLU) activation, and group normalization. As there is no down-sampling operation employed, the feature map size remains consistent for each layer, facilitating dense connections between the dilated convolution layers, as shown in Figure 6.

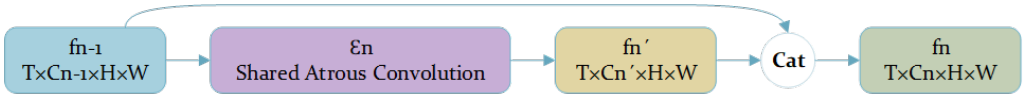


Figure 6. Spatial encoding with dilated convolution. The feature map at each layer is obtained by concatenating the input and output along the channel dimension from the previous layer.

For each image in SITS, the feature map $f_{n-1}^{t}, t \in [1, T]$ serves as the input to the encoder $\varepsilon^{n}, n \in [1, N]$. The output feature map $f_{n}^{t\prime}$ is obtained by concatenating the input feature map f_{n-1}^{t} and the output feature map f_{n}^{t} along the channel dimension, and it becomes the input for the subsequent encoder ε^{n+1}. Its shape is $C_n \times H \times W$, where $C_n = C_{n-1} + C_n'$. Finally, the feature maps from each image are stacked together to create the feature map time series f_n for that layer, with a shape of $T \times C_n \times H \times W$, as represented in the following formula.

$$f_n' = concat[\varepsilon^n(f_{n-1}^t)]_{t=0}^{T}, n \in [1, N] \tag{1}$$

$$f_n = concat[f_{n-1}, f_n'], n \in [1, N-1] \tag{2}$$

where $f_0 = X$. For each layer of the encoder, the concatenated feature map of the input and output from the previous layer is used as the input. Additionally, the feature maps for each layer have the same size as the original image.

3.2.2. Temporal Encoding with Transformer Encoder

After multiple layers of dilated convolution encoding, we obtain a feature map time series with consistent spatial resolution and different depth levels of semantic information, as well as multi-scale information. Concatenating the feature map time series outputs from all layers along the channel dimension results in the feature map time series F, with a shape of $T \times C' \times H \times W$, where $C' = \sum_{n=1}^{N} C_n'$. The concatenated feature map time series is then reshaped into a time series for each pixel P, with a shape of $T \times C'$, totaling $H \times W$ pixels, as expressed in the following formula:

$$F = concat[f_n'], n \in [1, N] \tag{3}$$

Finally, feature extraction is performed on each pixel's time series using the Transformer encoder, as illustrated in Figure 7.

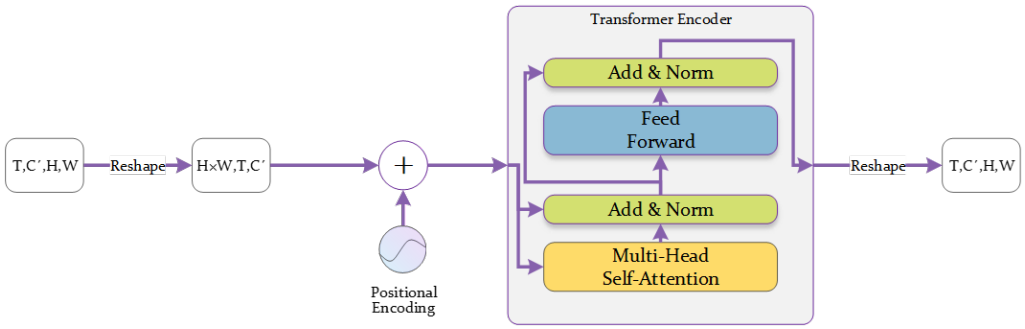


Figure 7. Temporal encoding with Transformer encoder. After spatial encoding, the feature map time series is organized into a feature sequence for each pixel. Following positional encoding, this sequence is fed into the Transformer encoder, and the obtained results are reshaped to the original dimensions.

The images in SITS are arranged in chronological order based on their acquisition dates, and their position indices can directly correspond to their order in the sequence. Alternatively, position indices can be determined by calculating the number of days between each image's acquisition date and a fixed reference date. However, in order to capture the annual phenological patterns of crops within the image time series, the position index *pos*

is calculated based on the image's acquisition date t and the New Year's Day $t_{newyear}$ of the same year. This ensures that the position of the images in the sequence corresponds to the growth stages of the represented crops. In other words, the image sequence is arranged based on the chronological order of crop growth stages, rather than the sequence of collection dates. The formula for calculating pos is as follows:

$$pos = (t - t_{newyear}) \cdot days \tag{4}$$

As shown in Table 1, the process begins with data preprocessing to obtain an SITS $X \in \mathbb{R}^{T \times C \times H \times W}$ for each land parcel, arranged by acquisition time, where T represents the number of image acquisitions, and the lengths may vary. Following this, the feature maps $f_n, n \in [1, N]$ are obtained through the dilated convolution encoder $\varepsilon^n, n \in [1, N]$. The feature maps from each layer are then concatenated along the channel dimension to form the feature map $F \in \mathbb{R}^{T \times C' \times H \times W}$. Subsequently, the feature map F is reshaped into a time series for each pixel, resulting in $P \in \mathbb{R}^{T \times C'}$, where the time series has a length of T and a feature dimension of C'. The pixel-wise time series data P, along with position encoding, is fed into the Transformer encoder, which processes all pixel-wise data to generate spatio-temporal features for the entire image, producing the spatio-temporal feature map $F' \in \mathbb{R}^{T \times C' \times H \times W}$. Finally, the spatio-temporal feature map F' undergoes temporal averaging and convolution to yield the final segmentation result R.

Table 1. STENN training process.

Training STENN for SITS Semantic Segmentation
Input: SITS data X; Ground truth labels Y
Output: Semantic Segmentation R
1. Load the training dataset X, configure model parameters, and initialize the weights of the STENN model.
2. While step $\leq$ Epoch:
3. Iterate through all the data in X, and utilize the STENN model to generate the segmentation result R.
4. Calculate the loss between R and Y using the specified loss function and update the entire model's parameters based on the computed loss.
5. End the training loop once the defined number of training steps (epochs) has been reached.
6. Validate and test the trained STENN model, and save all the results.

4. Experiments and Analysis

4.1. Evaluation Metric

We employed four quantitative metrics to assess the performance, which include model parameter count, inference time (IT) on SITS, overall accuracy (OA), and the mean Intersection over Union (mIoU). The number of parameters is a measure of the model's size, with smaller models being more resource-efficient. Inference time solely considers the time required for the model to process the data, and shorter inference times indicate faster prediction speeds for individual time series. *OA* evaluates the global accuracy of the extraction results, while *mIoU* quantifies the overlap between the predicted results and the ground truth labels. The expressions are as follows:

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$mIoU = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{TP_i}{TP_i + FP_i + FN_i} \right) \tag{6}$$

In which TP (true positives) represents the correct identification of positive instances, TN (true negatives) signifies the correct identification of negative instances, FP (false positives) denotes the erroneous identification of negative instances as positive, and FN (false negatives) indicates the erroneous identification of positive instances as negative. N is the number of categories, TP_i represents the true positives for the i category, FP_i signifies the false positives for the i category, and FN_i denotes the false negatives for the i category.

All experiments were conducted on a desktop system with an Intel(R) Xeon(R) Gold 6130 CPU and NVIDIA Tesla V100-32G GPU.

4.2. Experimental Detail

The parameters for the six comparative algorithms were set according to the referenced papers. For our proposed STENN model, the parameter settings are as shown in the following table:

In Table 2, the number of dilation rates determines the frequency of dilated convolutions. In our experiments, we employed the Adam optimizer with default parameters, set the batch size to two, and defined the learning rate as 0.001. Considering that the images in each sequence were captured at different time points, there may be variations in the sample distribution within each batch. To address this issue, we incorporated group normalization in the encoder, specifying four groups, instead of using batch normalization.

Table 2. STENN model configuration.

Layer Name	Kernel Size	Dilation Rate	Input Channel	Output Channel
Conv Layer-1	3 × 3	{1, 2, 3}	10	16
Conv Layer-2	3 × 3	{2, 4, 6}	26	32
Conv Layer-3	3 × 3	{2, 4, 6}	58	32
Conv Layer-4	3 × 3	{2, 4, 6}	90	64
Concat (dim = channel) & Reshape				
Transformer encoder (input = 144, head = 8)				
Mean (dim = time) & Reshape				
Conv Layer-5	3 × 3	{1, 1}	144	20

To quantitatively and qualitatively validate the effectiveness of the STENN model, we selected U-TAE [1], ConvLSTM [37,38], ConvGRU [39], U-ConvLSTM [40], U-BiConvLSTM [39], and 3D U-Net [40] as comparative algorithms.

- The backbone network of U-TAE [1] is U-Net. At the lowest resolution, the attention-based time encoder generates a set of time attention masks for each pixel. After spatial interpolation, these attention masks fuse the feature map time sequences at all resolutions into a single feature map. In the coding branch, four sets of group normalization are utilized, while batch normalization is employed in the decoding branch. A temporal encoding approach employing an L-TAE with 16 heads and a key-query space dimensionality of four was chosen.
- ConvLSTM [37,38] and ConvGRU [39] are recurrent neural networks, primarily replacing all linear layers in the model with convolutional layers. Their hidden sizes are set to 160 and 188, respectively.
- U-ConvLSTM [40] and U-BiConvLSTM [39] use U-Net as their backbone network, replacing L-TAE in the network with ConvLSTM [37] or bidirectional ConvLSTM. In comparison to the original method, batch normalization in the encoder is replaced with group normalization. Like the U-TAE, the Backbone utilizes the U-Net architecture, simply replacing the L-TAE with ConvLSTM or BiConvLSTM in their respective positions.

- The encoding space of 3D U-Net [40] is three-dimensional, allowing it to simultaneously process spatial and spectral dimensions. Finally, it performs mean fusion in the temporal dimension. The network comprises five consecutive 3D convolutional blocks, conducting spatial down-sampling after the second and fourth blocks. Each convolutional block doubles the channel count of the processed feature maps, with the innermost feature map having a channel size set at 128. Leaky ReLU and 3D batch normalization are employed within the convolutional blocks of this architecture.

4.3. Results

The second part of Table 3 displays the results of ablation studies, where

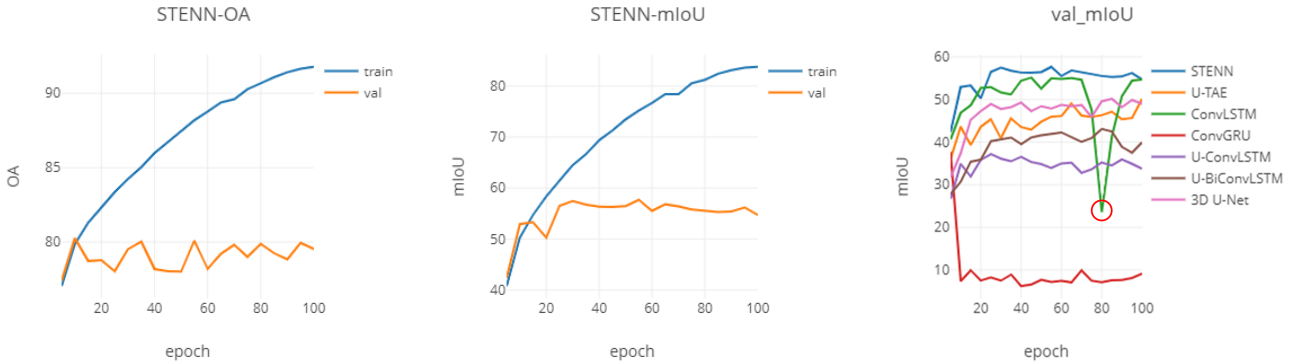
- VGG Backbone: The spatial encoding is replaced with the first 10 layers of VGG-16, and the network width has been adjusted accordingly to maintain a similar parameter count. The feature maps at each stage are up-sampled to match the original image resolution, and after channel-wise concatenation, they undergo temporal encoding.
- No Dense Connection: Remove the dense connections from the spatial encoding, where the input of each layer is only the output of the previous layer.
- No Transformer encoder: Remove the Transformer encoder and directly perform temporal mean fusion on the feature map time series obtained after spatial encoding. Then, obtain the final segmentation result through convolution.
- Single Date (August): Select one image from SITS taken in August for training, and the model produces segmentation results after spatial encoding.
- Single Date (May): Select one image from SITS taken in May for training.

Table 3. Segmentation result of different approach. The best performance is displayed in bold, and the second best performance is underlined.

Model	Param (×1000)	OA	mIoU	IT (ms)
STENN (Ours)	**447**	**79.37**	**55.8**	322
U-TAE	1087	<u>79.31</u>	47.82	**88**
ConvLSTM	1009	77.47	<u>54.06</u>	163
ConvGRU	<u>956</u>	70.31	36.93	141
U-ConvLSTM	1521	75.27	36.19	110
U-BiConvLSTM	1210	76.14	42.32	104
3D U-Net	1554	76.92	47.21	<u>96</u>
VGG Backbone	401	78.18	55.26	193
No dense connection	403	77.26	53.16	245
No Transformer encoder	289	73.41	31.91	153
Single Date (August)	289	58.75	31.12	151
Single Date (May)	289	55.16	30.43	151

5. Discussion

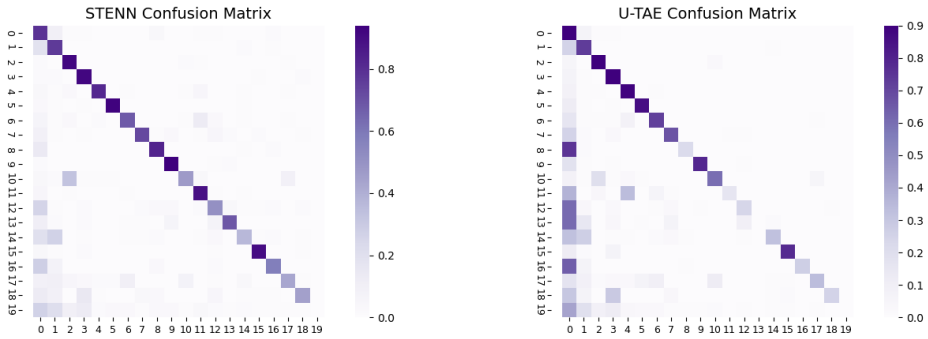
Comparison Analysis. In Table 3 and Figure 8c, the performance of ConvLSTM and ConvGRU demonstrates that LSTM performs better in handling time series data. Furthermore, the combination of using U-Net as the backbone for spatial feature extraction and extracting temporal features through ConvLSTM or Bidirectional ConvLSTM does not yield satisfactory results. The 3D U-Net can simultaneously handle both 2D spatial and 1D spectral data, providing shorter inference times and good performance. Our proposed model has less than half the number of parameters compared to U-TAE. Furthermore, under the same OA conditions, our model's mIoU outperforms the U-TAE by 8 percentage points. This indicates that, when dealing with complex land parcels, our model excels in the delineation of land object boundaries, exhibiting superior robustness and generalization capabilities.



(a) The OA values of the STENN model. (b) The mIoU values of the STENN model. (c) The mIoU values for all models.

Figure 8. During the training and validation process, the OA (**a**) and mIoU (**b**) of the STENN model. The change in mIoU for all models on the validation set during the training process (**c**). All models using the same learning rate may lead to oscillations during the training process for some models (red circles).

Based on the comparison in Figure 9, it is evident that our model has shown significant improvements in recognizing many land cover categories, especially Grapevine (8), Winter durum wheat (11), Fruits, vegetables, flowers (12), Potatoes (13), and Sorghum (18). This improvement primarily stems from U-TAE harnessing spatial features predominantly, while training attention weights solely on the temporal dimension. Instances of misclassification occur when the target land cover shares visual similarities with the background in crucial temporal segments, leading to erroneous categorization into the background.



(a) Confusion matrix of STENN (b) Confusion matrix of U-TAE

Figure 9. The confusion matrices for STENN (**a**) and U-TAE (**b**) display the classification accuracy of these two models when predicting crop categories. Specifically, each entry in the matrix at the *i* row and *j* column represents the proportion of samples belonging to class *j* that were classified as class *i*. The darker the color, the larger the proportion, indicating better model performance. Figure 2 provides a correspondence between the crop type numbers and their specific crop names.

According to Figure 10, it is visually evident that our model's segmentation results outperform other methods significantly. Other methods exhibit clear instances of small object omissions, large object edge dissolution, and missed areas (highlighted in red circles). Additionally, comparing the patch with the ground truth, it is apparent that the accuracy of the PASTIS dataset requires improvement.

Ablation Analysis. First, we replaced the spatial encoding with VGG, and the different resolution feature maps were simply up-sampled without the gradual resolution restoration used in FCN or U-Net. We observed a 1% drop in OA. We also verified the effectiveness of the dense connections. The experimental results show that models without dense connections decreased by 2% in both OA and mIoU. Extracting temporal features proved to be necessary, as removing the Transformer encoder resulted in a significant performance

drop, particularly a 23% drop in mIoU. Additionally, we trained our proposed spatial encoding module on single temporal images from August and May. We observed a significant decrease in both OA and mIoU, confirming the importance of the annual phenological cycle of crops in agricultural land classification.

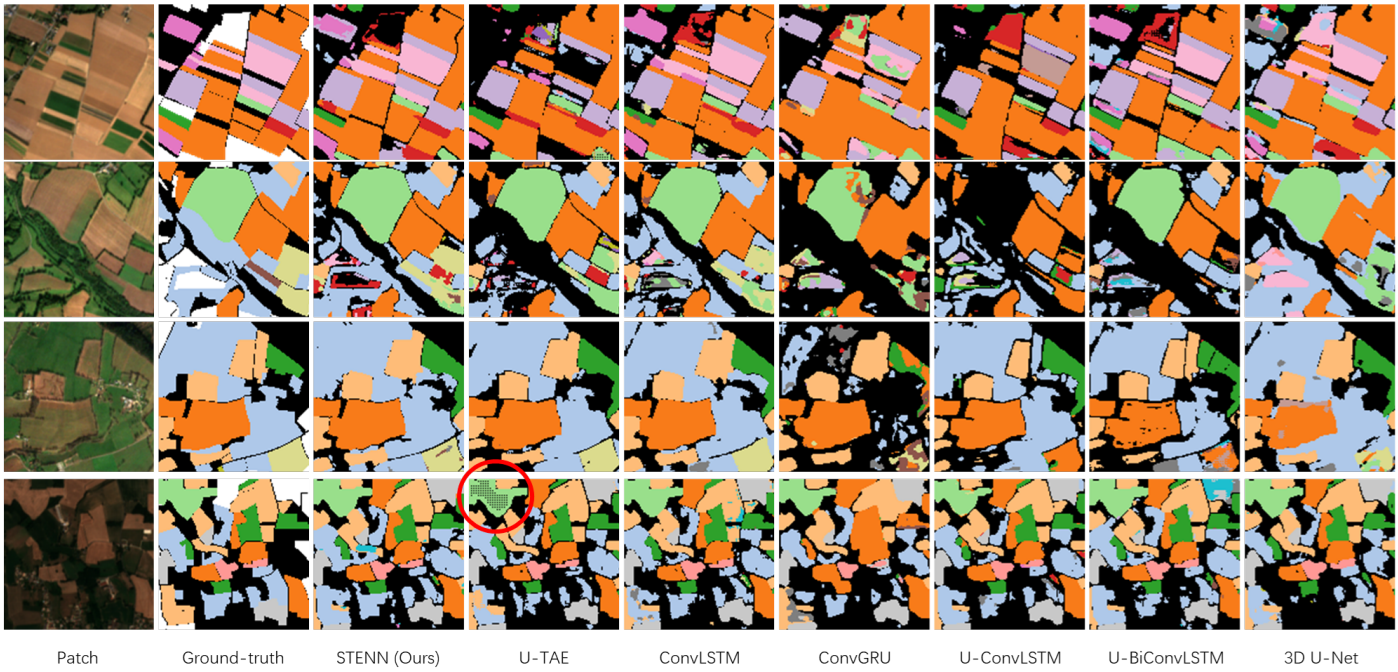


Figure 10. Qualitative semantic segmentation results. We begin with a single image from the sequence represented using the RGB channels, and for which we have knowledge of the ground truth parcel's boundaries and crop type. Subsequently, we showcase the pixelwise predictions generated by our approach, as well as those of six other comparative algorithms. The legend refers to Figure 2 for reference. Gridding effect due to downsampling (highlighted in red circles).

6. Conclusions

Historical images encompass the phenological variations of crops over time, which hold crucial significance for agricultural land segmentation. Based on this observation, this paper introduces a novel spatio-temporal encoding neural network for semantic segmentation in SITS. Specifically, the network comprises two critical encoding modules. The first one is the spatial encoding module, which constructs a high-resolution spatial feature encoding module using dilated convolutions and dense connections. This module swiftly increases the receptive field, extracts deep semantic information, and acquires rich multi-scale features without compromising spatial information. Subsequently, the temporal encoding module employs a Transformer encoder to extract the temporal features of each pixel within the feature map time series. Furthermore, we use unique position encoding to better capture the annual cyclic characteristics of crops. Extensive experiments conducted on the PASTIS dataset demonstrate that our STENN model not only significantly reduces the required training parameters but also enhances feature extraction capabilities, resulting in more satisfactory segmentation performance.

Our team's next steps include the following two aspects:

- Further optimization of time series models: This involves improvements to the Transformer encoder or the exploration of other architectures suitable for time series modeling. This can contribute to reducing computational requirements and memory demands.
- Integration across diverse data sources: Consider incorporating more data sources into the framework to further enhance the accuracy of semantic segmentation. For in-

stance, meteorological data, land-use data, and other sources may provide valuable information for interpreting land cover phenology.

Author Contributions: Conceptualization, F.Z. and Y.W.; methodology, F.Z. and Y.W.; validation, F.Z.; writing—original draft preparation, F.Z. and Y.W.; writing—review and editing, Y.D. and Y.Z.; funding acquisition, Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China Joint Fund Key Project (grant number U21A2013); and Open Research Project of The Hubei Key Laboratory of Intelligent Geo-Information Processing (grant number KLIGIP-2021B05).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: PASTIS is a benchmark dataset designed for the panoptic and semantic segmentation of agricultural parcels using satellite time series. This dataset comprises 2433 patches located within the French metropolitan territory, and it comes with panoptic annotations, which means it provides instance indices along with semantic labels for every pixel. Each patch consists of a Sentinel-2 multi-spectral image time series of varying lengths. This dataset, along with additional information regarding its composition, can be publicly accessed at https://github.com/VSainteuf/pastis-benchmark (accessed on 21 November 2022).

Acknowledgments: We appreciate the constructive comments and suggestions from the reviewers that helped improve the quality of this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Garnot, V.S.F.; Landrieu, L. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 4872–4881.
2. Abad, M.S.J.; Abkar, A.A.; Mojaradi, B. Effect of the temporal gradient of vegetation indices on early-season wheat classification using the random forest classifier. *Appl. Sci.* **2018**, *8*, 1216. [CrossRef]
3. Chen, Y.; Li, M.; Zhang, Z. Does the Rural Land Transfer Promote the Non-Grain Production of Cultivated Land in China? *Land* **2023**, *12*, 688. [CrossRef]
4. Pluto-Kossakowska, J. Review on multitemporal classification methods of satellite images for crop and arable land recognition. *Agriculture* **2021**, *11*, 999. [CrossRef]
5. Pandey, P.C.; Koutsias, N.; Petropoulos, G.P.; Srivastava, P.K.; Ben Dor, E. Land use/land cover in view of earth observation: Data sources, input dimensions, and classifiers—A review of the state of the art. *Geocarto Int.* **2021**, *36*, 957–988. [CrossRef]
6. Yuan, X.; Shi, J.; Gu, L. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Syst. Appl.* **2021**, *169*, 114417. [CrossRef]
7. Wang, L.; Yan, J.; Mu, L.; Huang, L. Knowledge discovery from remote sensing images: A review. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2020**, *10*, e1371. [CrossRef]
8. Wang, Y.; Zhang, D.; Dai, G. Classification of high resolution satellite images using improved U-Net. *Int. J. Appl. Math. Comput. Sci.* **2020**, *30*, 399–413.
9. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; pp. 234–241.
10. Xu, Z.; Zhang, W.; Zhang, T.; Li, J. HRCNet: High-resolution context extraction network for semantic segmentation of remote sensing images. *Remote Sens.* **2020**, *13*, 71. [CrossRef]
11. Ding, L.; Tang, H.; Bruzzone, L. LANet: Local attention embedding to improve the semantic segmentation of remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 426–435. [CrossRef]
12. Yu, M.; Qin, F. Research on the Applicability of Transformer Model in Remote-Sensing Image Segmentation. *Appl. Sci.* **2023**, *13*, 2261. [CrossRef]
13. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
14. Wang, H.; Chen, X.; Zhang, T.; Xu, Z.; Li, J. CCTNet: Coupled CNN and transformer network for crop segmentation of remote sensing images. *Remote Sens.* **2022**, *14*, 1956. [CrossRef]
15. Zhang, C.; Jiang, W.; Zhang, Y.; Wang, W.; Zhao, Q.; Wang, C. Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–20. [CrossRef]

16. Li, Y.; Cheng, Z.; Wang, C.; Zhao, J.; Huang, L. RCCT-ASPPNet: Dual-Encoder Remote Image Segmentation Based on Transformer and ASPP. *Remote Sens.* **2023**, *15*, 379. [CrossRef]
17. Tian, Q.; Zhao, F.; Zhang, Z.; Qu, H. GLFFNet: A Global and Local Features Fusion Network with Biencoder for Remote Sensing Image Segmentation. *Appl. Sci.* **2023**, *13*, 8725. [CrossRef]
18. He, X.; Zhou, Y.; Zhao, J.; Zhang, D.; Yao, R.; Xue, Y. Swin transformer embedding UNet for remote sensing image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
19. Bolton, D.K.; Friedl, M.A. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. *Agric. For. Meteorol.* **2013**, *173*, 74–84. [CrossRef]
20. Pan, L.; Xia, H.; Zhao, X.; Guo, Y.; Qin, Y. Mapping winter crops using a phenology algorithm, time-series Sentinel-2 and Landsat-7/8 images, and Google Earth Engine. *Remote Sens.* **2021**, *13*, 2510. [CrossRef]
21. Meroni, M.; d'Andrimont, R.; Vrieling, A.; Fasbender, D.; Lemoine, G.; Rembold, F.; Seguini, L.; Verhegghen, A. Comparing land surface phenology of major European crops as derived from SAR and multispectral data of Sentinel-1 and-2. *Remote Sens. Environ.* **2021**, *253*, 112232. [CrossRef]
22. Moskolaï, W.R.; Abdou, W.; Dipanda, A.; Kolyang. Application of deep learning architectures for satellite image time series prediction: A review. *Remote Sens.* **2021**, *13*, 4822. [CrossRef]
23. Sakamoto, T.; Yokozawa, M.; Toritani, H.; Shibayama, M.; Ishitsuka, N.; Ohno, H. A crop phenology detection method using time-series MODIS data. *Remote Sens. Environ.* **2005**, *96*, 366–374. [CrossRef]
24. Sun, C.; Bian, Y.; Zhou, T.; Pan, J. Using of multi-source and multi-temporal remote sensing data improves crop-type mapping in the subtropical agriculture region. *Sensors* **2019**, *19*, 2401. [CrossRef]
25. Garnot, V.S.F.; Landrieu, L.; Giordano, S.; Chehata, N. 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, Seattle, WA, USA, 13–19 June 2020; pp. 12325–12334.
26. Chen, G.; Li, C.; Wei, W.; Jing, W.; Woźniak, M.; Blažauskas, T.; Damaševičius, R. Fully convolutional neural network with augmented atrous spatial pyramid pool and fully connected fusion path for high resolution remote sensing image segmentation. *Appl. Sci.* **2019**, *9*, 1816. [CrossRef]
27. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
28. Diakogiannis, F.I.; Waldner, F.; Caccetta, P.; Wu, C. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *162*, 94–114. [CrossRef]
29. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
30. Kattenborn, T.; Leitloff, J.; Schiefer, F.; Hinz, S. Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. *ISPRS J. Photogramm. Remote Sens.* **2021**, *173*, 24–49. [CrossRef]
31. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv* **2014**, arXiv:1412.7062.
32. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
33. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
34. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
35. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
36. Dou, G.; Zhao, K.; Guo, M.; Mou, J. Memristor-based LSTM network for text classification. *Fractals* **2023**, *31*, 2340040. [CrossRef]
37. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; Woo, W.-C. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Adv. Neural Inf. Process. Syst.* **2015**, *28*.
38. Rußwurm, M.; Körner, M. Convolutional LSTMs for cloud-robust segmentation of remote sensing imagery. *arXiv* **2018**, arXiv:1811.02471.
39. Ballas, N.; Yao, L.; Pal, C.; Courville, A. Delving deeper into convolutional networks for learning video representations. *arXiv* **2015**, arXiv:1511.06432.
40. Rustowicz, R.M.; Cheong, R.; Wang, L.; Ermon, S.; Burke, M.; Lobell, D. Semantic segmentation of crop type in Africa: A novel dataset and analysis of deep learning methods. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019; pp. 75–82.