



## Learning discriminative spatiotemporal features for precise crop classification from multi-temporal satellite images

Shunping Ji, Zhili Zhang, Chi Zhang, Shiqing Wei, Meng Lu & Yulin Duan

To cite this article: Shunping Ji, Zhili Zhang, Chi Zhang, Shiqing Wei, Meng Lu & Yulin Duan (2020) Learning discriminative spatiotemporal features for precise crop classification from multi-temporal satellite images, International Journal of Remote Sensing, 41:8, 3162-3174, DOI: [10.1080/01431161.2019.1699973](https://doi.org/10.1080/01431161.2019.1699973)

To link to this article: <https://doi.org/10.1080/01431161.2019.1699973>



Published online: 26 Dec 2019.



Submit your article to this journal [↗](#)



Article views: 76



View related articles [↗](#)



View Crossmark data [↗](#)



# Learning discriminative spatiotemporal features for precise crop classification from multi-temporal satellite images

Shunping Ji<sup>a</sup>, Zhili Zhang<sup>a</sup>, Chi Zhang<sup>a</sup>, Shiqing Wei<sup>a</sup>, Meng Lu<sup>b</sup> and Yulin Duan<sup>id c</sup>

<sup>a</sup>School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China; <sup>b</sup>Faculty of Geoscience, Department of Physical Geography, Utrecht University, Utrecht, The Netherlands; <sup>c</sup>Chinese Academy of Agricultural Sciences, Institute of Agricultural Resources and Regional Planning, Beijing, China

## ABSTRACT

Precise crop classification from multi-temporal remote sensing images has important applications such as yield estimation and food transportation planning. However, the mainstream convolutional neural networks based on 2D convolution collapse the time series information. In this study, a 3D fully convolutional neural network (FCN) embedded with a global pooling module and channel attention modules is proposed to extract discriminative spatiotemporal presentations of different types of crops from multi-temporal high-resolution satellite images. Firstly, a novel 3D FCN structure is introduced to replace 2D FCNs as well as to improve current 3D convolutional neural networks (CNNs) by providing a mean to learn distinctive spatiotemporal representations of each crop type from the reshaped multi-temporal images. Secondly, to strengthen the learning significance of the spatiotemporal representations, our approach includes 3D channel attention modules, which regulate the between-channel consistency of the features from the encoder and the decoder, and a 3D global pooling module, which selects the most distinctive features at the top of the encoder. Experiments were conducted using two data sets with different types of crops and time spans. Our results show that our method outperformed in both accuracy and efficiency, several mainstream 2D FCNs as well as a recent 3D CNN designed for crop classification. The experimental data and source code are made openly available at <http://study.rsgis.whu.edu.cn/pages/download/>.

## ARTICLE HISTORY

Received 30 January 2019  
Accepted 8 October 2019

## 1. Introduction

Precise crop classification has important applications in yield estimation, grain transportation, crop growing, and health monitoring (Tennakoon, Murty, and Eiumnoh 1992; Soria-Ruiz et al. 2007). Remote sensing satellite imagery provides opportunities to efficiently classify crops on a large scale. Multiple sources of temporal satellite imagery information can improve the accuracy of crop classification. Classical statistical analysis (Boryan et al. 2011), machine learning (Cutler et al. 2007), and physical quantitative inversion (Wang and Zhang 2017) methods have been applied to remote sensing imagery. Among them, the machine learning technique has become one of the most widely-used. The current

machine learning methods can be classified into two categories: 1) conventional methods that rely on empirical feature manufacturing and 2) deep learning-based methods that automatically learn representations of data through multiple neuron layers.

Conventional machine learning methods consist of a handcrafted and empirical feature extractor that extracts task-specific features from original images and a classifier that groups these features into a given label set. Vegetation indices are one kind of typical features used in land cover classification from multi-spectral/multi-temporal images. For example, Xiao et al. (2005) utilized the normalized difference vegetation index (NDVI), the enhanced vegetation index (EVI), and the land surface water index (LSWI), which are derived from the moderate resolution imaging spectroradiometer (MODIS) temporal images to classify rice, cloud, water, and evergreen plants in images. Similarly, Wardlow et al. (2007) used the EVI and NDVI of the MODIS 250 m time-series for crop-related land use/land cover (LULC) classification. Arvor et al. (2011) utilized the MODIS EVI time series to identify five crop classes in agricultural areas. Conrad et al. (2011) divided the MODIS 250 m NDVI time series into several temporal segments as input features for crop classification.

Middle and high-resolution remote sensing sensors are also the main data sources for crop and vegetation classification. For example, Guerschman et al. (2003) distinguished land covers by separately using a NDVI time series and concatenated original image patches from multi-temporal Landsat TM growing seasons' data. Sexton et al. (2013) utilized humidity, luminance, the NDVI, and their changes to classify land covers from multi-temporal Landsat-5 images. Jia et al. (2014) incorporated phenological features extracted from a MODIS NDVI time series in their land cover classification of Landsat-5 data. Dechka et al. (2002) classified multi-temporal IKONOS images according to a wetland habitat class system based on images, NDVI maps and texture measures.

There are several classical classifiers, such as support vector machine (SVM) (Foody and Mathur 2004), *k*-nearest neighbour (*k*-NN) (Blanzieri and Melgani 2008), maximum likelihood classification (MLC) (Foody et al. 1992), and artificial neural network (ANN) (Dreiseitl and Ohno-Machado 2002), for land cover or vegetation classification. For example, Murthy et al. (2003) compared MLC, iterative-MLC, principal component analysis (PCA) based MLC (PCA-MLC), and ANN in wheat extraction from multi-temporal images. Zhu and Liu (2014) utilized a recursive SVM framework to obtain a tree-type map of a forest.

Deep learning is a representation learning method that directly extracts information from original data through multiple layers of abstractions. Convolutional neural networks (CNN) have been widely and successfully applied to natural images and remote sensing images for semantic segmentation and object detection (Cheng, Zhou, and Han 2016; Hu et al. 2015). A variant of CNN called fully convolutional network (FCN) has become a mainstream algorithm for semantic segmentation (Long, Shelhamer, and Darrell 2015; LeCun, Bengio, and Hinton 2015). FCN does not require pixel by pixel labelling and can efficiently extract a segmentation map of the input size through an encoder and a decoder. 3D CNN is another specific case of CNN. Unlike conventional 2D CNN, which operates on 2D images, 3D CNN can learn representations of 3D data and its applications have been extended to video (Tran et al. 2015), light detection and ranging (LiDAR) point cloud (Li 2017), hyperspectral images (Nagasubramanian et al. 2018), and multi-temporal images (Ji et al. 2018).

Ji et al. (2018) introduced a 3D CNN-based segmentation method for crop type classification that can extract the spatiotemporal features of a crop's growth cycle. The method has shown to outperform methods based on 2D CNN (Kussul et al. 2017) where the time dimension was

collapsed. However, the image in their work was segmented pixel-by-pixel, which can cause extremely low computational efficiency. In addition, their simple network structure could be improved to achieve higher classification accuracy. In this paper, we present the development of a new 3D FCN structure that can efficiently extract crop features maps from a whole input image at once in contrast to the pixel-by-pixel approach. In addition, we add a 3D attention mechanism to the 3D FCN to improve the classification accuracy.

Although the encoder-decoder FCN structure has been widely used in semantic segmentation (Ronneberger, Fischer, and Brox 2015; Zhang, Liu, and Wang 2018; Badrinarayanan, Kendall, and Cipolla 2017; Yu et al. 2018b; Tian et al. 2019), its limitation as far as representing the global contextual consistency of features have been shown. First, although each feature map of a layer, especially the top layer, having the highest semantics, has different levels of importance (weights) for a semantic classification problem, all of which are merely concatenated in a channel direction; and second, the lateral connections between an encoder and a decoder at each scale (Çiçek et al. 2016; Han and Ye 2018; Yu et al. 2018b) are implemented with only a concatenation operation, which can result in an imbalance between the features at different semantic levels and subsequently can affect contextual inconsistency. In a 3D FCN, this limitation, in the context of learning spatiotemporal representations, is even more critical not only when global contextual consistency is pursued but temporal consistency as well; for example, in multi-temporal images, a certain crop that may be easily discriminated from other crops during a certain time span may not be distinctive during other time spans. A mechanism that can discriminate the importance of each feature map to reach globally-consistent spatiotemporal representations is necessary.

Global contextual information extraction by global pooling and attention mechanisms has been proven effective in 2D FCN based segmentation. For example, ParseNet (Liu, Rabinovich, and Berg 2015) added global context to deep convolutional networks by using global pooling. BiseNet (Yu et al. 2018a) introduced a global pooling-based feature fusion strategy to combine the features learned from a spatial path and a context path. A discriminative feature network (DFN) (Yu et al. 2018b) introduced a channel attention block and a global average pooling mechanism to select more discriminative features. Hu, Shen, and Sun (2018) utilized an attention mechanism for recalibrating channel-wise feature responses. Inspired by these studies, in this paper we introduce a novel attention mechanism to strengthen the global contextual and temporal consistency of spatiotemporal features that are learned from a 3D FCN. Specifically, this mechanism contains a 3D global pooling module (GPM) at the top layer of the encoder of the 3D FCN to extract the global consistent contextual/temporal representation at the highest semantic level and a 3D channel attention module (CAM) that fuses the feature maps from the encoder and decoder to achieve between-channel consistency. We call our 3D FCN with a Global pooling module and a Channel attention module a 3D FGC.

## 2. Methods

### 2.1. 3D convolution for multi-temporal images

In our method, 3D convolution first is used to extract the crop's spatiotemporal representation from the multi-temporal remote sensing images. Let  $h$  and  $w$  denote the height and width of the input images and  $m$  and  $n$  denote the number of spectral channels and the length of time series, respectively. To extract the spatiotemporal representation of the crop

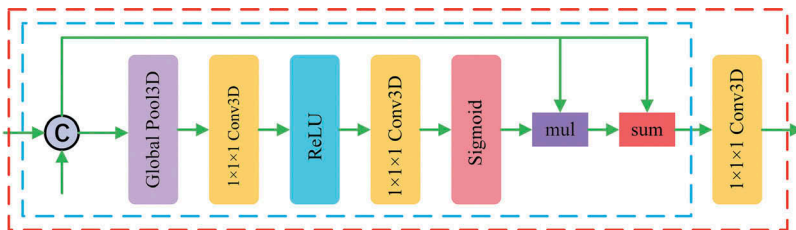
growth cycle, the temporal information must be treated as the third dimension with respect to a 2D image plane. Therefore, the 4D data  $(m, n, w, h)$  are reshaped to  $m$  3D tensors  $(n, w, h)$  in order to use it as the input of a 3D convolution. Equation (1) denotes the calculation of a 3D convolution where  $g$  is the number of input channels;  $x_{c,d,e}$  and  $y_{c,d,e}$  are the input and output at location  $(c, d, e)$ , respectively;  $t, o, f$  is the side lengths of the 3D kernel and  $W$  represents the elements in the kernel; and  $b$  is the bias. An activation function, the rectified linear unit (ReLU), is applied to  $y_{c,d,e}$  to produce the final output of the current point.

$$y_{c,d,e} = \sum_g \sum_{k=1}^t \sum_{i=1}^o \sum_{j=1}^f W_{k,i,j,g} x_{c+i,d+j,e+k,g} + b \quad (1)$$

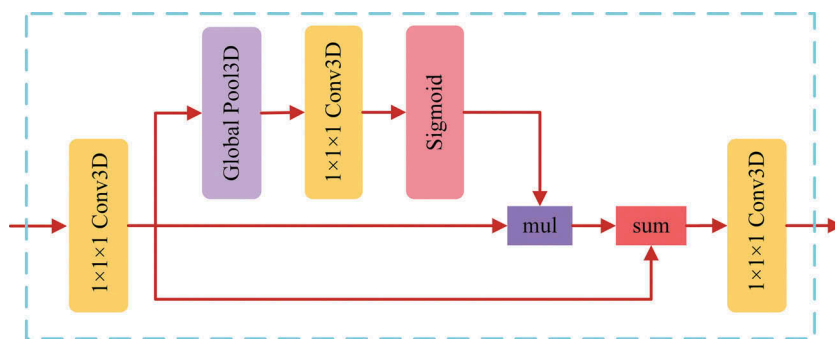
## 2.2. 3D channel attention module and global pooling module

We extend the 2D channel attention block (Li et al. 2018; Yu et al. 2018a, 2018b) used in 2D image segmentation to our 3D channel attention module (CAM) to fuse the spatio-temporal features from the encoder and the decoder through channel-wise re-weighting. Figure 1 shows the structure of our 3D CAM, where the concatenated features from the encoder and decoder are processed by a 3D average global pooling and two  $1 \times 1 \times 1$  convolutions with respective to ReLU and sigmoid activation, to output a weight vector that indicates the relative importance between the feature channels. The vector is then multiplied and summed by the input features and processed by another  $1 \times 1 \times 1$  convolution to produce globally consistent spatiotemporal features. The introduction of the final  $1 \times 1 \times 1$  convolution improves the performance of the original channel attention block (Li et al. 2018), as shown in the blue dotted box.

Yu et al. (2018b) stated that applying a global average pooling on top of an encoder (the highest semantic layer) could assure the highest consistency. We found that an elaborate structure, named the global pooling module (GPM), was able to obtain a better result than single global pooling. Our 3D GPM also utilizes an attention mechanism, which transforms the features at the top of the encoder to the corresponding features of the decoder with global spatiotemporal consistency. In Figure 2, the input 3D feature is processed by the  $1 \times 1 \times 1$  convolution, the 3D global average pooling, and the  $1 \times 1 \times 1$  convolution with sigmoid activation sequentially. The output vector is then multiplied and summed by the input (after the  $1 \times 1 \times 1$  convolution) and processed by another  $1 \times 1 \times 1$  convolution to produce the top layer of the decoder.



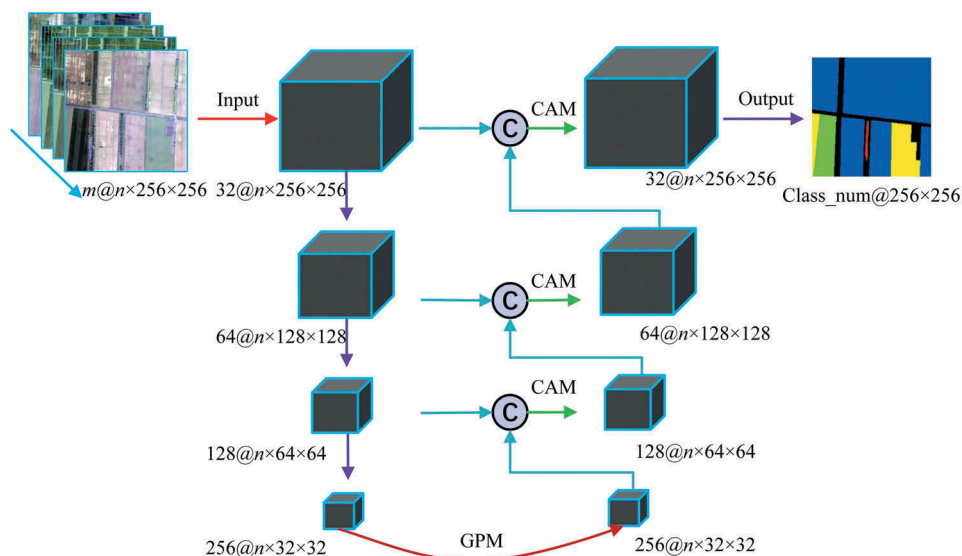
**Figure 1.** CAM to achieve between-channel consistency. The input is the concatenation of features from the encoder and decoder at each scale.



**Figure 2.** GPM, which processes the top layer of the encoder to output a global consistent top layer of the decoder.

### 2.3. Network architecture

Based on the 3D convolution, the GPM, and the CAM, we construct the 3D FGC for crop classification from multi-temporal remote sensing images, as shown in Figure 3. The encoder of the 3D FGC is a visual geometry group (VGG) structure (Simonyan and Zisserman 2014) with 3D convolutional layers. There are four scales of convolutional layers, each of which consists of two identical  $3 \times 3 \times 3$  convolutions both in the encoder and decoder. The number of output channels of each scale is 32, 64, 128, and 256. A max-pooling with the size of (1, 2, 2) is applied, which indicates that the features in the time dimension (the first dimension) is not compressed to preserve the temporal information of the crop's growth cycle. At the top of the encoder, the GPM is used to extract the global consistent contextual/temporal information.



**Figure 3.** Our proposed 3D FGC using a global pooling module and channel attention modules, where 'C' indicates concatenation,  $n$  is the length of time series, and the number before the symbol @ is the channel number.

Then, the CAM is applied to fuse the features from the encoder and decoder, and the features of each layer are sequentially restored up to the input size.

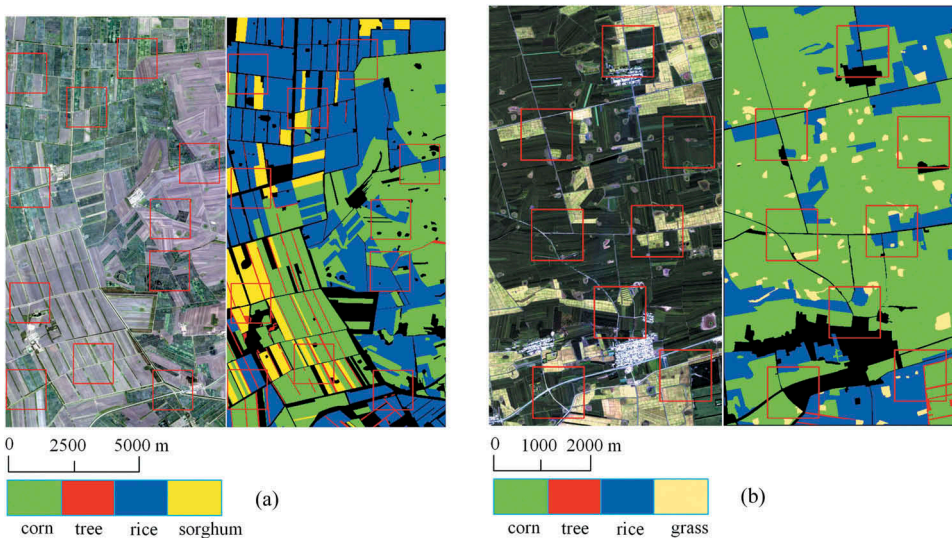
The last 3D feature maps are flattened to 2D segmentation maps, which is achieved by a  $n \times 3 \times 3$  convolution and a  $1 \times 1 \times 1$  convolution. The segmentation maps are then compared to the ground truth for loss computation and backpropagation. Let  $p_{i,j,c}$  be the probability of pixel  $(i, j)$  belonging to class  $c$ , and  $t_{i,j,c}$  the corresponding ground truth. Equation (2) is used to calculate the multi-class cross-entropy of the pixel, which is then summed over the pixel count to produce the loss of the current input image as (3).

$$l_{ij} = - \sum_c t_{ij,c} \ln(p_{ij,c}) \quad (2)$$

$$\text{loss} = \frac{1}{N} \sum_i \sum_j l_{ij} \quad (3)$$

### 3. Experiment and results

We used two GF2 (Gaofen 2 satellite, <http://www.cresda.com/EN/satellite/7157.shtml>) multi-temporal images (Figure 4) for our experiments. The GF2 satellite, which is a part of the Chinese high-resolution earth observation system, was launched in 2014. The GF2 multi-spectral images have four bands (red, green, blue, and near infrared) with 4 m ground resolution. The two data sets were captured in 2015 and 2017. The 2015 data set has a pixel size of  $1417 \times 2652$  and consists of four images captured in June, July, August, and September of 2015. The 2017 data set has a pixel size of  $2102 \times 1163$  and contains seven images captured in June, July, August, September, October, November, and December of



**Figure 4.** GF2 images captured in 2015 (a), and 2017 (b). Black pixels in the shape files indicate lack of label information. Patches in the red rectangles were used for training and the remaining images for prediction.



2017. The two GF2 data sets were preprocessed with the quick atmospheric correction (QUAC) method (Bernstein et al. 2005) and geometrical rectification. Handcrafted shapefiles with land cover information were used as a reference for classification.

3.1. Implementation details

In each data set, a quarter of the area (within the red 256 × 256 rectangles in Figure 4) was used to train the 3D FGC and the remainder was used for assessing prediction accuracy. In the training stage, due to the different number of samples in each crop category, we adopted a class balance strategy (Ronneberger, Fischer, and Brox 2015). We used the Adaptive Moment Estimation (ADAM) optimizer (Kingma and Jimmy 2014) with batch size 1 without batch normalization. The learning rate was set to 1e-4. All the experiments were executed on a single NVIDIA GeForce GTX 1060 GPU with 6 GB RAM. We use overall accuracy (OA), kappa coefficient (*k*), and mean Intersection over Union (mIoU) as evaluation indicators. The IoU is the ratio between the intersection of the pixels of a crop category detected by the algorithm and the true positive pixels and the result of their union. The mIoU is the mean of IoU on all the categories. We established mIoU as the main index because it reflected both the pixel accuracy indicated by the OA and the between-class discriminability indicated by *k*.

3.2. Effectiveness of CAM and GPM

We analysed the effects of the CAM and GPM on a 3D FCN for learning the spatiotemporal features. As shown in Table 1, when the 3D FCN without the CAM and GPM (i.e., replacing the CAM with the concatenation operator and removing the GPM from Figure 3) was applied, an mIoU of 84.5% was obtained. When the CAM was introduced, the mIoU greatly improved to 86.2%, indicating the effectiveness of the attention mechanism in fusing the spatiotemporal features of different semantic levels. The mIoU further improved to 86.5% when the GPM was introduced, which demonstrated that applying the global pooling method on the highest semantic layer promoted global consistency.

We also designed a 2D version of our FGC by replacing the 3D convolution with 2D convolution, which takes the concatenated multi-temporal images as input. The mIoU of the 2D FGC dropped about 3%, which clearly indicated the advantage of a 3D CNN structure over a 2D CNN structure for preserving temporal information. The parameters of a 3D FCN were twice as many as that of the 2D FCN, which resulted in a relatively slower training efficiency of the former.

**Table 1.** The performances of the CAM and GPM on the 3D FCN on the 2015 data set. ‘M’ is the abbreviation of million, the unit of parameter number (short for params), and ‘G’ is the abbreviation of Gillion (thousand million), the unit of floating point operations (short for flops).

Method	params(M)	flops(G)	OA (%)	<i>k</i> (%)	mIoU (%)
2D FGC	2.19	130.23	96.90	94.73	83.48
3D FCN	5.44	104.98	97.11	95.07	84.52
3D FCN+CAM	5.13	83.85	97.51	95.75	86.26
3D FCN+CAM+GPM (FGC)	5.33	84.93	97.56	95.85	<b>86.50</b>



### 3.3. Comparison with other methods

We compared our 3D FGC with other recent methods, including U-Net (Ronneberger, Fischer, and Brox 2015), DeepLabv3+ (Chen et al. 2018) and 3D CNN (Ji et al. 2018) on both datasets. The inputs of the 3D FGC, 3D CNN, and U-Net/DeepLabv3+ were  $m \times n \times 256 \times 256$  3D tensors,  $m \times n \times 8 \times 8$  3D tensors, and  $m \times n \times 256 \times 256$  images, where  $m$  and  $n$  were the number of spectral channels and length of time series, respectively.

Table 2 shows the efficiency of the different algorithms and their prediction results on the two data sets. The efficiency of our method and the two 2D CNN methods were at the same level in training and prediction, but the prediction time of the pixel-wise 3D CNN was hundreds of times slower. Therefore, the consumed time dramatically increases as the input data grows larger, making the pixel-wise 3D CNN unrealistic for large-scale crop classification.

The DeepLabv3+ and U-Net performed the worst with the two data sets, which proved again the poor performance of 2D CNN on extracting spatiotemporal features as the temporal dimension collapsed. The DeepLabv3+ introduced a series of Atrous convolutions, which may not have been very suitable in this case, causing its performance to be worse than that of U-Net. Our method produced the best results on both datasets and outperformed the 3D CNN marginally, which was mainly due to the introduction of the attention mechanism that strengthens the global contextual and temporal consistency of representations both on the highest semantic layer and on each scale.

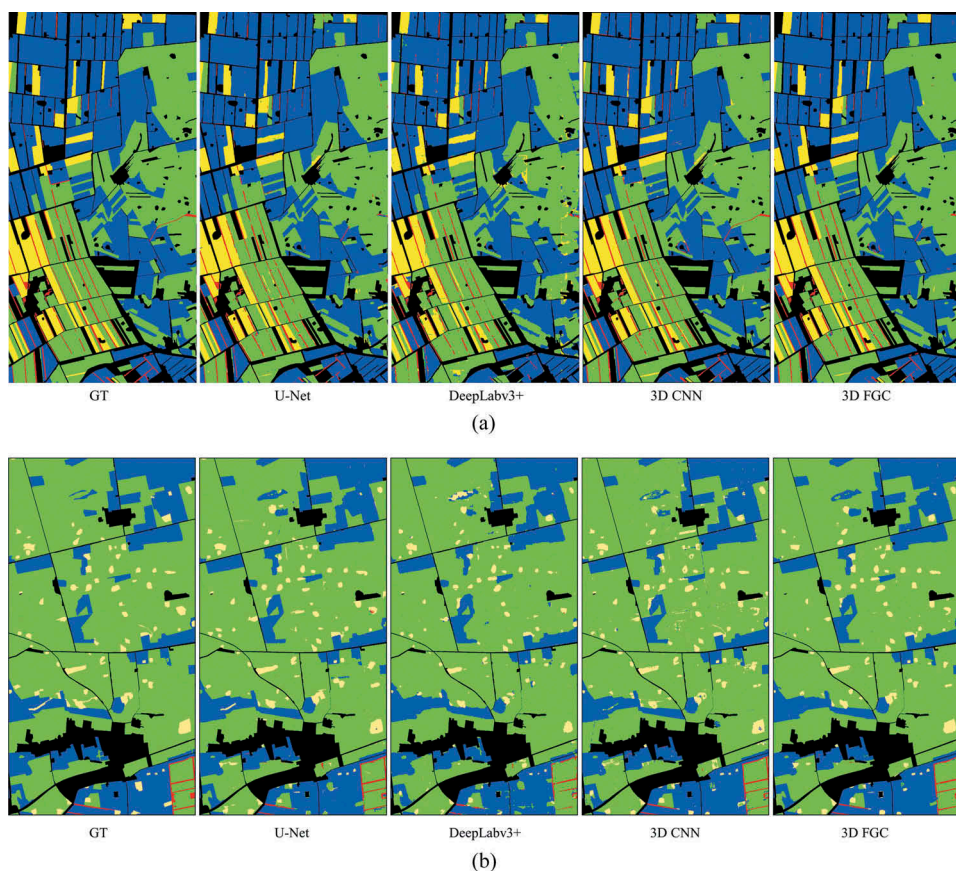
Figure 5 shows the segmentation maps of the U-Net, Deeplabv3+, 3D CNN, and 3D FGC respectively on the two data sets. Compared to the ground truth (the first image of each row), the three methods, except the Deeplabv3+, obtained satisfactory results. At this scale the difference between them was small. For example, in the upper right block of the 2015 image (planted with rice, blue) the 3D FGC performed better than the 3D CNN and U-Net, as shown in the zoomed image of some local regions (Figure 6). The patches of the first two columns are from the 2015 test data and the remaining are from the 2017 data. For all of them, the 3D FGC performed marginally better than the 3D CNN and U-Net.

## 4. Discussion

In addition to the experiments demonstrating that a 3D CNN clearly performed better than a 2D CNN on extracting spatiotemporal representations and that the 3D FGC outperformed the recent 3D CNN both on accuracy and efficiency, we also tested the performance of conventional methods such as  $k$ -NN (Blanzieri and Melgani 2008), SVM

**Table 2.** Comparison of different methods on both datasets.

Method	Train time/epoch (s)	Test time (s)	OA (%)	$k$ (%)	mIoU (%)
2015 dataset					
DeepLabv3+	5.5	11.6	94.81	91.16	65.33
U-Net	2.2	2.5	96.82	94.57	82.78
3D CNN	248.7	8,272.0	97.26	95.35	84.81
3D FGC	11.9	9.3	97.56	95.85	<b>86.50</b>
2017 dataset					
DeepLabv3+	3.4	9.8	95.36	88.53	54.88
U-Net	1.9	2.4	96.38	91.24	74.56
3D CNN	209.8	5,303.3	96.46	91.49	76.08
3D FGC	14.2	10.9	96.81	92.21	<b>76.51</b>



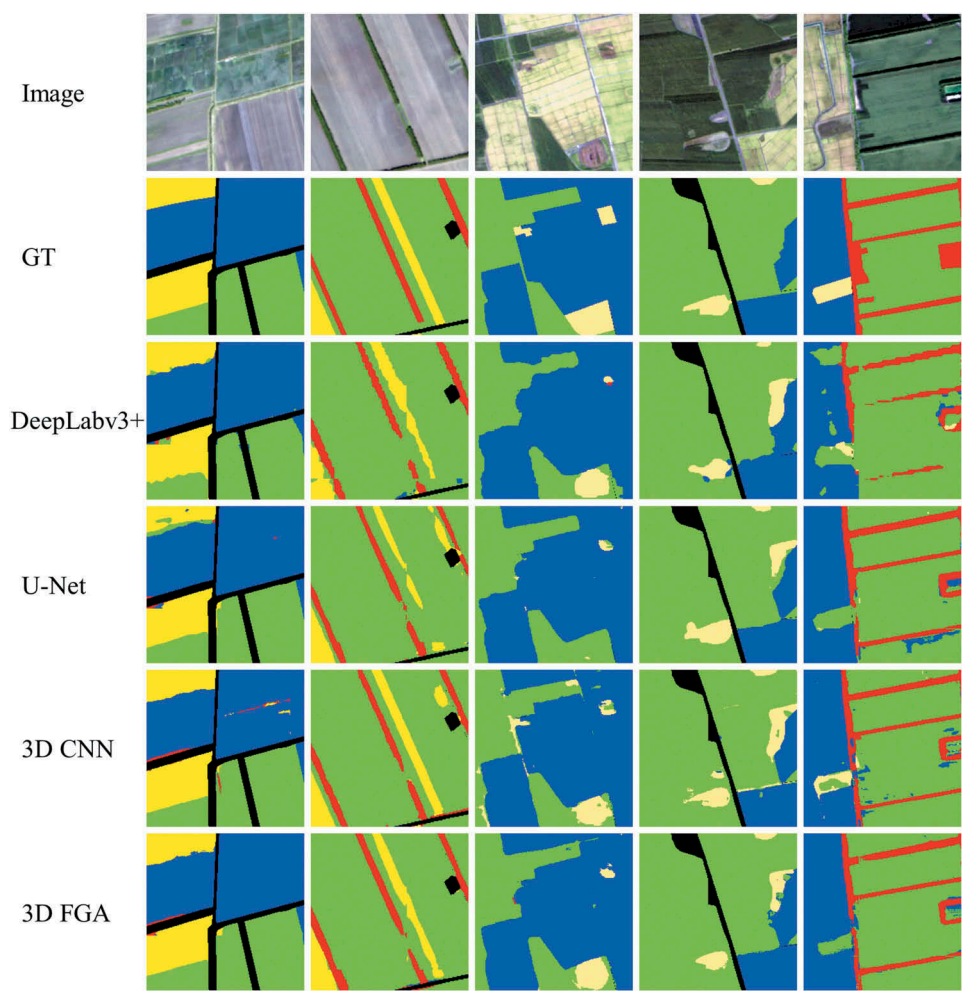
**Figure 5.** Ground truth and the prediction results of the U-Net, Deeplabv3+, 3D CNN and 3D FGC on the 2015 (top) and 2017 (bottom) data sets.

(Suykens and Vandewalle 1999), and  $k$ -NN after dimension reduction with PCA (Wold, Esbensen, and Geladi 1987), with the pixel values of multi-temporal images and a generated NDVI series as input. Our findings were similar to Ji et al. (2018), namely, the conventional methods performed worse than all the CNN-based methods.

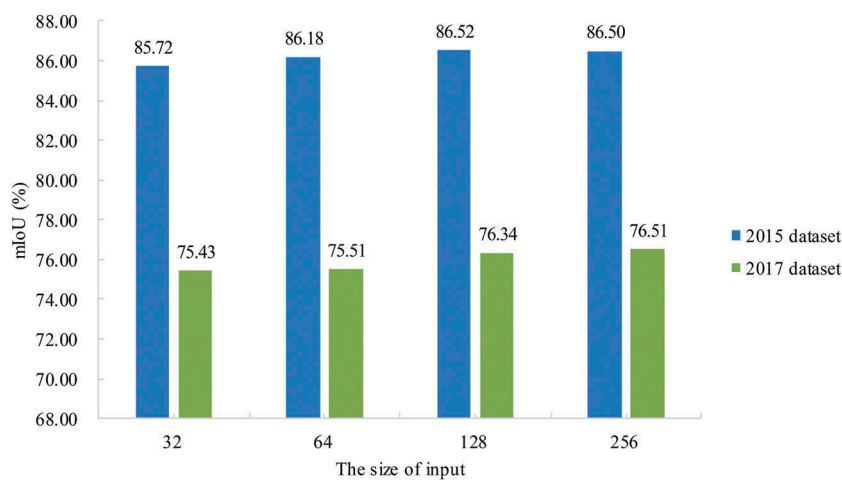
In addition to the network structure changes in the ablation test, the change of input size also affected the performance of our 3D FGC. Figure 7 shows the mIoU scores assessed with different input sizes. When the input size increased, the mIoU increased but only slightly. Using the  $128 \times 128$  input produced almost the same results as did the  $256 \times 256$  input, which confirmed that our network was robust to the input size.

## 5. Conclusion

In this paper we introduced an efficient 3D FCN that includes an attention mechanism that enables learning from spatiotemporal representations with global contextual and temporal consistency for precise crop classification from multi-temporal satellite imagery. The experiments showed that our 3D FCN significantly outperformed two conventional 2D FCN-based methods. Our method also obtained better segmentation accuracy than



**Figure 6.** Zoomed segmentation results of the U-Net, 3D CNN and 3D FGC. The patches of the first two columns are from 2015 dataset, and the rest from 2017 dataset.



**Figure 7.** The effect of different sizes of input on the performance of the 3D FGC on the 2015 and 2017 datasets.

did a current pixel-wise 3D CNN-based segmentation. Our results confirmed that the superior efficiency of our 3D FCN makes it a feasible method in extracting spatiotemporal representation from large-size remote sensing imagery.

## Acknowledgements

This work was supported by the National Key Research and Development Program of China, (2018YFB0505003).

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Funding

This work was supported by the the National Key Research and Development Program of China [2018YFB0505003].

## ORCID

Yulin Duan  <http://orcid.org/0000-0003-4364-3041>

## References

- Arvor, D., and M. Jonathan, M. S. P. Meirelles, V. Dubreuil, and L. Durieux. 2011. "Classification of MODIS EVI time series for crop mapping in the state of Mato Grosso, Brazil". *International Journal of Remote Sensing* 32 (22):7847–7871.
- Badrinarayanan, V., A. Kendall, and R. Cipolla. 2017. "Segnet: A Deep Convolutional Encoder-decoder Architecture for Image Segmentation." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (12): 2481–2495. doi:10.1109/TPAMI.34.
- Bernstein, L. S., S. M. Adler-Golden, R. L. Sundberg, R. Y. Levine, T. C. Perkins, A. Berk, A. J. Ratkowski, G. Felde, and M. L. Hoke. 2005. "Validation of the QUick Atmospheric Correction (QUAC) Algorithm for VNIR-SWIR Multi-and Hyperspectral Imagery." Paper presented at the Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XI, Orlando, USA.
- Blanzieri, E., and F. Melgani. 2008. "Nearest Neighbor Classification of Remote Sensing Images with the Maximal Margin Principle." *IEEE Transactions on Geoscience Remote Sensing* 46 (6): 1804–1811. doi:10.1109/TGRS.2008.916090.
- Boryan, C., Z. Yang, R. Mueller, and M. Craig. 2011. "Monitoring US Agriculture: The US Department of Agriculture, National Agricultural Statistics Service, Cropland Data Layer Program." *Geocarto International* 26 (5): 341–358. doi:10.1080/10106049.2011.562309.
- Chen, L.-C., Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. 2018. "Encoder-decoder with Atrous Separable Convolution for Semantic Image Segmentation." Paper presented at the Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany.
- Cheng, G., P. Zhou, and J. Han. 2016. "Learning Rotation-invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images." *IEEE Transactions on Geoscience and Remote Sensing* 54 (12): 7405–7415. doi:10.1109/TGRS.2016.2601622.
- Çiçek, Ö., A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. 2016. "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation." Paper presented at the International conference on medical image computing and computer-assisted intervention, Athens, Greece.

- Conrad, C., R. R. Colditz, S. Dech, D. Klein, and P. L. G. Vlek. 2011. "Temporal Segmentation Of Modis Time Series for Improving Crop Classification in Central Asian Irrigation Systems." *International Journal of Remote Sensing* 32 (23): 8763–8778.
- Cutler, D. R., C. E. Thomas Jr, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler. 2007. "Random Forests for Classification in Ecology." *Ecology* 88 (11): 2783–2792. doi:10.1890/07-0539.1.
- Dechka, J. A., S. E. Franklin, M. D. Watmough, R. P. Bennett, and D. W. Ingstrup. 2002. "Classification Of Wetland Habitat and Vegetation Communities Using Multi-temporal Ikonos Imagery in Southern Saskatchewan." *Canadian Journal of Remote Sensing* 28 (5): 679–85.
- Dreiseitl, S., and L. Ohno-Machado. 2002. "Logistic Regression and Artificial Neural Network Classification Models: a Methodology Review." *Journal of Biomedical Informatics* 35 (5–6): 352–9.
- Foody, G. M., and A. Mathur. 2004. "Toward Intelligent Training Of Supervised Image Classifications: Directing Training Data Acquisition for Svm Classification." *Remote Sensing of Environment* 93 (1–2): 107–17.
- Foody, G. M., N. A. Campbell, N. M. Trodd, and T. F. Wood. 1992. "Derivation and Applications Of Probabilistic Measures Of Class Membership from The Maximum-likelihood Classification." *Photogrammetric Engineering and Remote Sensing* 58 (9): 1335–41.
- Guerschman, J. P., J. M. Paruelo, C. Dibella, M. C. Giallorenzi, and F. Pacin. 2003. "Land Cover Classification in The Argentine Pampas Using Multi-temporal Landsat Tm Data." *International Journal of Remote Sensing* 24 (17): 3381–402.
- Han, Y., and J. C. Ye. 2018. "Framing U-Net via Deep Convolutional Framelets: Application to Sparse-view CT." *IEEE Transactions on Medical Imaging* 37 (6): 1418–1429. doi:10.1109/TMI.2018.2823768.
- Hu, F., G.-S. Xia, H. Jingwen, and L. Zhang. 2015. "Transferring Deep Convolutional Neural Networks for the Scene Classification of High-resolution Remote Sensing Imagery." *Remote Sensing* 7 (11): 14680–14707. doi:10.3390/rs71114680.
- Hu, J., L. Shen, and G. Sun. 2018. "Squeeze-and-excitation Networks." Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, USA.
- Ji, S., C. Zhang, X. Anjian, Y. Shi, and Y. Duan. 2018. "3D Convolutional Neural Networks for Crop Classification with Multi-temporal Remote Sensing Images." *Remote Sensing* 10 (1): 75. doi:10.3390/rs10010075.
- Jia, K., S. Liang, X. Wei, Y. Yao, Y. Su, B. Jiang, and X. Wang. 2014. "Land Cover Classification Of Landsat Data with Phenological Features Extracted from Time Series Modis Ndvi Data." *Remote Sensing* 6 (11): 11518–32.
- Kingma, D. P., and B. Jimmy. 2014. "Adam: A Method for Stochastic Optimization." *arXiv Preprint* arXiv: 1412.6980.
- Kussul, N., M. Lavreniuk, S. Skakun, and A. Shelestov. 2017. "Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data." *IEEE Geoscience Remote Sensing Letters* 14 (5): 778–782. doi:10.1109/LGRS.2017.2681128.
- LeCun, Y., Y. Bengio, and G. Hinton. 2015. "Deep Learning." *nature* 521 (7553): 436. doi:10.1038/nature14539.
- Li, B. 2017. "3D Fully Convolutional Network for Vehicle Detection in Point Cloud." Paper presented at the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, Canada.
- Li, H., P. Xiong, A. Jie, and L. Wang. 2018. "Pyramid Attention Network for Semantic Segmentation." *arXiv Preprint* arXiv: 1805.10180.
- Liu, W., A. Rabinovich, and A. C. Berg. 2015. "Parsenet: Looking Wider to See Better." *arXiv Preprint* arXiv: 1506.04579.
- Long, J., E. Shelhamer, and T. Darrell. 2015. "Fully Convolutional Networks for Semantic Segmentation." Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, USA.



- Murthy, C. S., P. V. Raju, and K. V. S. Badrinath. 2003. "Classification Of Wheat Crop with Multi-temporal Images: Performance Of Maximum Likelihood and Artificial Neural Networks." *International Journal Of Remote Sensing* 24 (23): 4871-90.
- Nagasubramanian, K., S. Jones, A. K. Singh, A. Singh, B. Ganapathysubramanian, and S. Sarkar. 2018. "Explaining Hyperspectral Imaging Based Plant Disease Identification: 3D CNN and Saliency Maps." *arXiv Preprint arXiv: 1804.08831*.
- Ronneberger, O., P. Fischer, and T. Brox. 2015. "U-net: Convolutional Networks for Biomedical Image Segmentation." Paper presented at the International Conference on Medical image computing and computer-assisted intervention, Munich, Germany.
- Sexton, J. O., D. L. Urban, M. J. Donohue, and C. Song. 2013. "Long-term Land Cover Dynamics by Multi-temporal Classification across The Landsat-5 Record." *Remote Sensing Of Environment* 128: 246-58.
- Simonyan, K., and A. Zisserman. 2014. "Very Deep Convolutional Networks for Large-scale Image Recognition." *arXiv Preprint arXiv: 1409.1556*.
- Soria-Ruiz, J., Y. Fernandez-Ordonez, H. McNairn, and J. Bugden-Storie. 2007. "Corn Monitoring and Crop Yield Using Optical and RADARSAT-2 Images." Paper presented at the 2007 IEEE International Geoscience and Remote Sensing Symposium, Barcelona, Spain.
- Suykens, J. A. K., and J. Vandewalle. 1999. "Least Squares Support Vector Machine Classifiers." *Neural Processing Letters* 9 (3): 293-300. doi:10.1023/A:1018628609742.
- Tennakoon, S. B., V. V. N. Murty, and A. Eiumnoh. 1992. "Estimation of Cropped Area and Grain Yield of Rice Using Remote Sensing Data." *International Journal of Remote Sensing* 13 (3): 427-439. doi:10.1080/01431169208904047.
- Tian, Z., C. Shen, H. Tong, and Y. Yan. 2019. "Decoders Matter for Semantic Segmentation: Data-Dependent Decoding Enables Flexible Feature Aggregation."
- Tran, D., L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. 2015. "Learning Spatiotemporal Features with 3d Convolutional Networks." Paper presented at the Proceedings of the IEEE international conference on computer vision, Santiago, Chile.
- Wang, Q., and J. Zhang. 2017. "Hyperspectral Image Vegetation Change Detection Based on Biochemical Parameters Inversion." Paper presented at the International Conference in Communications, Signal Processing, and Systems, Chongqing, China.
- Wardlow, B. D., S. L. Egbert, and J. H. Kastens. 2007. "Analysis of Time-series Modis 250m Vegetation Index Data for Crop Classification in The U.S. Central Great Plains." *Remote Sensing of Environment* 108 (3): 290-310. .
- Wold, S., K. Esbensen, and P. Geladi. 1987. "Principal Component Analysis." *Chemometrics Intelligent Laboratory Systems* 2 (1-3): 37-52. doi:10.1016/0169-7439(87)80084-9.
- Xiao, X., S. Boles, J. Liu, D. Zhuang, S. Frolking, L. Changsheng, W. Salas, and III. B. Moore. 2005. "Mapping Paddy Rice Agriculture in Southern China Using Multi-temporal Modis Images." *Remote Sensing of Environment* 95 (4): 480-92.
- Yu, C., J. Wang, C. Peng, C. Gao, Y. Gang, and N. Sang. 2018a. "Bisenet: Bilateral Segmentation Network for Real-time Semantic segmentation." Paper presented at the Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany.
- Yu, C., J. Wang, C. Peng, C. Gao, Y. Gang, and N. Sang. 2018b. "Learning a Discriminative Feature Network for Semantic Segmentation." Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA.
- Zhang, Z., Q. Liu, and Y. Wang. 2018. "Road Extraction by Deep Residual U-net." *IEEE Geoscience and Remote Sensing Letters* 15 (5): 749-753. doi:10.1109/LGRS.2018.2802944.
- Zhu, X., and D. Liu. 2014. "Accurate Mapping Of Forest Types Using Dense Seasonal Landsat Time-series." *ISPRS Journal of Photogrammetry and Remote Sensing* 96: 1-11.