Spatial Resolution Enhancement for Large-Scale Land Cover Mapping via Weakly Supervised Deep Learning

Qiutong Yu, Wei Liu, Wesley Nunes Gonçalves, José Marcato Junior, and Jonathan Li

Abstract

Multispectral satellite imagery is the primary data source for monitoring land cover change and characterizing land cover globally. However, the consistency of land cover monitoring is limited by the spatial and temporal resolutions of the acquired satellite images. The public availability of daily high-resolution images is still scarce. This paper aims to fill this gap by proposing a novel spatiotemporal fusion method to enhance daily low spatial resolution land cover mapping using a weakly supervised deep convolutional neural network. We merge Sentinel images and moderate resolution imaging spectroradiometer (MODIS)-derived thematic land cover maps under the application background of massive remote sensing data and the large spatial resolution gaps between MODIS data and Sentinel images. The neural network training was conducted on the public data set SEN12MS, while the validation and testing used ground truth data from the 2020 IEEE Geoscience and Remote Sensing Society data fusion contest. The proposed data fusion method shows that the synthesized land cover map has significantly higher spatial resolution than the corresponding MODIS-derived land cover map. The ensemble approach can be implemented for generating high-resolution time series of satellite images by fusing fine images from Sentinel-1 and -2 and daily coarse images from MODIS.

Introduction

Remotely sensed satellite imagery is the primary data source for monitoring land cover change and characterizing land cover on a global scale (Song *et al.* 2017). Satellite images with daily coverage and fine spatial resolution are highly desired for Earth observation and related environmental applications (Sun and Zhang 2019). However, the consistency of daily land cover monitoring is often constrained by the spatial and temporal resolutions of the acquired satellite images freely available. For instance, Landsat satellites capture images with a moderate spatial resolution of 30 meters but with a long revisit period of 16 days. On the contrary, the moderate resolution imaging spectroradiometer (MODIS) can provide images daily, with coarser spatial resolutions of 250 m, 500 m, and 1 km. Hence, it is important to understand how to jointly leverage complementary data sources efficiently to conduct land cover classification. To have up-to-date land cover monitoring with a fine spatial scale, increasing the spatial resolution of coarse satellite imagery represents a continued advancement in remote sensing research. The availability of the MODIS data set has driven global-scale land cover mapping at coarse resolution. Previous works have conducted spatiotemporal fusion to blend MODIS and Landsat data to obtain improved classification results with a higher spatial resolution of 30 m (Gevaert and García-Haro 2015; Wang *et al.* 2015; Chen *et al.* 2017). *Sentinel-1* and *Sentinel-2* today can provide higher temporal resolution (three to five days) and higher spatial resolution (10 to 20 m) than Landsat satellites. However, these images are frequently unavailable for land cover mapping due to the presence of clouds. Hence, it is necessary to develop a feasible method to integrate remote sensing data from different sensors and time phases to acquire geospatial data with high spatial and temporal resolutions.

Recently, deep learning frameworks have enhanced the classification performance by automatic extraction of in-depth features. Therefore, deep learning-based land cover classification has become a current hotspot in the remote sensing research community. One of the significant advantages of using deep learning algorithms is that it is a learning-based method, which automatically learns an end-to-end mapping between coarse resolution images and fine resolution images. Previous research indicates that semantic segmentation classification with deep learning methods at the pixel level is promising in land cover mapping (Huang *et al.* 2018; Kemker *et al.* 2018).

To the best of our knowledge, no deep learning-based model has yet been introduced to conduct spatiotemporal fusion to blend MODIS data and Sentinel satellite images. The novelty is emphasized by the proposal of a weakly supervised approach. With the aim of providing enhanced land cover mapping through the fusion of multi-source satellite data, this paper extends one of the current state-of-the-art semantic segmentation networks, DeepLabV3+ (Chen *et al.* 2018), and then employ it to enhance the spatial resolution of MODIS-derived land cover maps, by integrating the maps (with an original spatial resolution of 500 m), synthetic-aperture radar (SAR) images derived from *Sentinel-1*, and multispectral images derived from *Sentinel-2*. The outputs of the model are high-resolution (10 m) land cover thematic maps. Technically, this is a task of supervised semantic segmentation of the Sentinel images since the MODIS maps are utilized as the target ground truth labels, and the model assigns one of the label classes to each pixel in the Sentinel images. However, due to the coarse resolution of MODIS maps, the Sentinel images only contain partial observations of the target ground truth labels, which makes the task become a weakly supervised semantic segmentation. To deal with weakly annotated ground truth labels, an

Qiutong Yu, Wei Liu, and Jonathan Li are with the Department of Geography and Environmental Management, University of Waterloo, Canada (junli@uwaterloo.ca).

Wesley Nunes Gonçalves and José Marcato Junior are with the Faculty of Engineering, Architecture and Urbanism and Geography, Federal University of Mato Grosso do Sul (UFMS), Campo Grande, Brazil.

Contributed by Dongdong Wang, August 11, 2020. (sent for review November 23, 2020; reviewed by Ava Vali, Jiangsan Zhao)

Photogrammetric Engineering & Remote Sensing
Vol. 87, No. 6, June 2021, pp. 405–412.
0099-1112/21/405–412

doi: 10.14358/PERS.87.6.405

additional module was embedded in the model, and it automatically updates the coarse labels based on the intermediate predictions on the training sets. The contributions of this paper can be summarized as follows.

First, a deep learning-based method was developed for an effective fusion of the MODIS and Sentinel data. Second, a deep learning semantic segmentation network, DeepLabV3+, was comprehensively evaluated given that the ground truth labels are noisy and unreliable. Third, more challenging land cover types can be classified using the proposed method.

The rest of this paper is organized as follows. Section "Related Works" reviews some previous studies regarding spatiotemporal fusion and the use of DeepLabV3+ on remote sensing images. Section "Method" introduces the technical information about DeepLabV3+ and the modifications added to its original architecture. The section "Experiments and Discussion" describes the experimental setup, including the implementation detail of our proposed model and the data set used in this study. Finally, the last section concludes the paper with remarks and expectations.

Related Works

Spatiotemporal Fusion of Remote Sensing Images

In the field of remote sensing, many key application domains stand to benefit from data fusion techniques. For example, the increase of spatial resolution contributes to land cover classification and ground object identification. Recently, many remote sensed data fusion methods have been proposed to deal with the specific problems that arise from the trade-off between spatial resolution and temporal frequency. In general, they can be categorized into image pair-based and spatial unmixing-based methods (Ghamisi *et al.* 2019).

The image pair-based method utilizes the relationship between the available coarse/fine image pairs to guide the prediction of fine images from coarse images on other days. Image paired-based methods can be further classified into a filter-based algorithm and learning-based algorithm (Song *et al.* 2018)

Among existing image pair-based spatiotemporal data fusion algorithms, the spatial and temporal adaptive reflectance fusion model (STARFM) (Gao *et al.* 2006) was the first model developed. It has been widely applied for fusing Landsat and MODIS to monitor environmental changes (Chen *et al.* 2015; Gevaert and García-Haro 2015). It uses one known pair of Landsat and MODIS images and one MODIS image at the prediction date. STARFM assumes that for a pure coarse pixel where only one land cover type exists, the changes in fine pixels within that coarse pixel can be implied directly by the coarse pixel changes. For heterogeneous coarse pixels with two or more land cover types, a weighted function is used for prediction, which assigns higher weights to the neighboring fine pixels where they are physically closer and spectrally similar to the coarse pixels (Gao *et al.* 2006).

Since STARFM assumes that the temporal changes of all land cover classes within a coarse pixel are consistent, it is thereby suitable for homogeneous landscapes (Ghamisi *et al.* 2019). However, it is sensitive to high heterogeneity and abrupt land cover changes (Sun and Zhang 2019). Subsequently, several algorithms have been developed to improve the accuracy of STARFM. For instance, Hilker *et al.* (2009) proposed the spatial-temporal adaptive algorithm for mapping reflectance change, designed to detect reflectance changes using Tasseled Cap transformations of both Landsat and MODIS data. Zhu *et al.* (2010) developed enhanced spatial and temporal adaptive reflectance fusion model to deal with heterogeneous landscapes specifically. It requires two coarse/fine image pairs to estimate the temporal change rate of each land cover class separately and assumes the change rates to be consistent. To summarize, these methods are different mainly in modeling the relationship between the paired pixels. These methods, including STARFM, can be considered filter-based methods because each pixel is predicted from a filtering model, a weighted sum of spectrally similar neighboring pixels from the input images (Song *et al.* 2018).

Recently, some learning-based spatiotemporal fusion algorithms have been proposed, such as support vector machine (Wang *et al.* 2018), Hopfield neural networks (Fung *et al.* 2019), and deep convolutional neural network (Song *et al.* 2018). These models directly take image pairs as inputs and automatically learn the relationship between coarse/fine image pairs. The results indicate that learning-based algorithms are more robust than the traditional spatiotemporal fusion algorithm (Sun and Zhang 2019). However, it usually requires abundant data for training the mapping relationship between fine and coarse satellite images.

The spatial unmixing-based methods are applied to compute the endmember (i.e., label) of coarse pixels and estimate the fined pixels using weighted endmembers (Zurita-Milla *et al.* 2008). According to Gevaert and García-Haro (2015), there are four steps in a spatial unmixing-based fusion model: (1) clustering the high-resolution data set to define the endmembers, (2) calculating the fractions of each endmember within each coarse spatial resolution pixel, (3) unmixing the medium-resolution pixel, and (4) assigning reflectance spectra to the high-resolution pixels. The unmixing can be applied using only one land cover thematic map with a fine spatial resolution (i.e., prior classification results). The thematic map can be produced by interpreting the available fine spatial resolution data (e.g., land use database). For example, Zurita-Milla *et al.* (2008) produced a 30 m Landsat-like time series by integrating one 30 m thematic map obtained by the classification of an available Landsat image and 300 m medium resolution imaging spectrometer time series. Furthermore, recent research illustrates that image pair-based and spatial unmixing-based methods can be combined (Zhu *et al.* 2016; Xie *et al.* 2016). Gevaert and García-Haro (2015) combined the advantages of STARFM and unmixing-based algorithms to propose a novel spatial and temporal reflectance unmixing model, which directly estimates the land cover changes between two coarse images.

In summary, traditional spatiotemporal fusion methods are mostly based on fusing each fine-coarse image pair in a pixel-wise process, which is not suitable for large-scale remote sensing data sets as the prediction is very time-consuming. In recent years, a variety of deep learning networks and large-scale remote sensing data sets have been published. The potential of deep learning-based spatiotemporal fusion methods needs to be further investigated, and novel methods should be proposed, mainly based on weak supervision.

Semantic Segmentation of Remote Sensing Image Using DeepLabV3+

Several studies have been carried out on using DeepLabV3+ for land cover classification tasks for aerial images. In a comparison study by Pashaei *et al.* (2020), the authors evaluated the performances of multiple semantic segmentation architectures on unmanned aircraft vehicle images for efficient land cover mapping. The experimental results demonstrate that DeepLabV3+ has a great potential for accurate land cover prediction tasks on a limited labeled image. On the other hand, some researchers extend the original DeepLabV3+ network to be more applicable for land observation images. For instance, Chen *et al.* (2019) proposed an improved network-based on DeepLabV3+ for semantic segmentation of high-resolution remote sensing images. The authors adopt dilated convolution by adding an augmented atrous spatial pyramid pool layer and a fully connected fusion path layer. As a result, dilated convolution enlarges the receptive field of feature points while the feature map resolution remains unchanged.

In addition to general land cover classification, DeepLabV3+ has been utilized for specific land cover mapping applications such as agricultural mapping (Du *et al.* 2019) and vegetation mapping (Ayhan and Kwan 2020). Here, we proposed its usage in a weakly supervised deep learning-based data fusion method.

Method

The workflow of the proposed approach to weakly supervised deep learning-based data fusion is shown in Figure 1. Details about each stage are presented in the next subsections.

Semantic Segmentation

The basic framework of our data-fusion model is the semantic segmentation network developed by Chen *et al.* (2018), namely DeepLabV3+. It is the latest version of DeepLab semantic segmentation architecture, which utilizes an atrous spatial pyramid pooling (ASPP) module. It extends the previous version (DeepLabV3) by adding a decoder module to refine the segmentation results, especially along object boundaries (Chen *et al.* 2018). The framework achieves a state-of-the-art mean intersection-over-union of 89% on the PASCAL VOC 2012 test. The ASPP mechanism improves the segmentation performance by exploiting the multi-scale contextual information of the features. The encoder part of the network structure enables DeepLabV3+ to reduce the feature maps and capture semantic information, while the decoder part recovers the spatial information.

Preprocessing of *Sentinel-1* SAR Images

The presence of speckle noise in the *Sentinel-1* SAR images makes the interpretation of the contents difficult, thereby degrading the quality of the image. Therefore, an efficient speckle noise removal technique needs to be applied to the *Sentinel-1* SAR images. In this study, SAR images are processed by the Enhanced Lee Filter (Lee 1981) to deal with the common problem of noisy edge boundaries. The filter algorithm operates by using edge directed windows. The local mean and local variance are computed using only the pixels in the edge directed window. After the speckle filtering, the images are enhanced by 2% of the linear stretch. The lowest and the highest 2% values are set to 0 and 255, respectively. Values in between are distributed from 0 to 255. As shown in Figure 2, the noise in the high contrast areas is effectively removed, and the edges are enhanced.

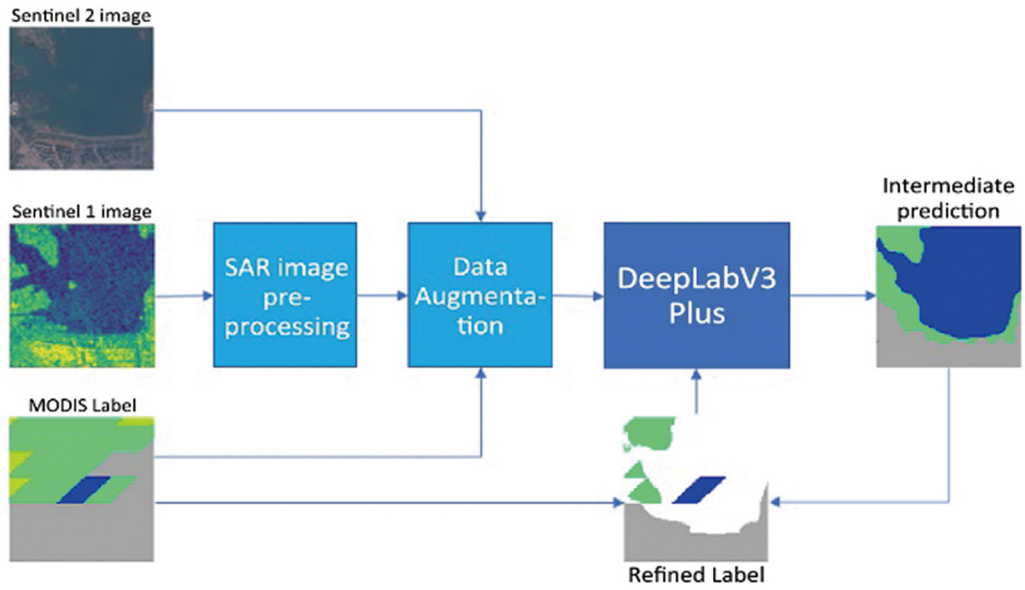

Figure 1. The workflow of the proposed approach.

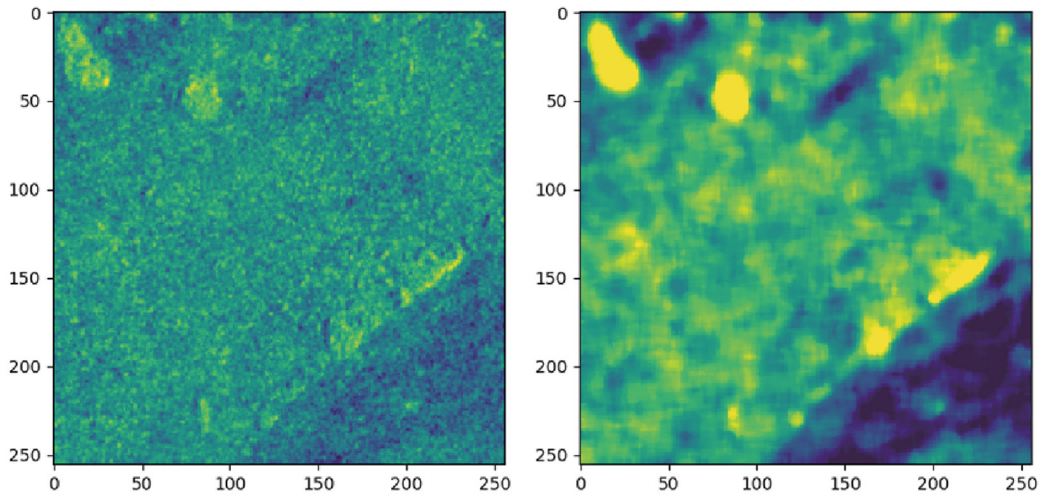

Figure 2. Example of raw Sentinel-1 SAR image (left) and the processed counterpart (right).

Data Augmentation

Several augmentation techniques have been added to the data-loader module of the model network to improve the performance by enlarging the training data set. These include geometric transformations (e.g., flip, rotation, warp) and linear transformations (e.g., 2%–98% contrast stretch). All geometric transformations are randomly selected and applied to images, each with a probability of 0.5. The linear stretch is assumed to be useful as applying to images with low contrast (e.g., image taken during nighttime).

Label Refinement

In essence, the major task of this study is semantic segmentation on weakly-supervised training, in which the annotation (i.e., MODIS labels) is noisy and unreliable. To further improve the performance of the model, additional strategies were adopted to deal with noisy labels specifically. In the SEN12MS data set, images of each scene were selected and cropped to be relatively homogenous. Noises (or incorrect labels) normally exist at the edges of land cover parcels. For example, shorelines are not clearly shown on the MODIS maps. For that matter, an additional module was added to the model, which updates the labels every five epochs (an epoch refers to one cycle through the full training data set). Hence, only for the first five epochs, the model is trained on original MODIS labels. After the fifth epoch, the model outputs the intermediate predictions on all training samples and then obtains the updated labels by comparing the intermediate predictions with the original MODIS labels. The differences would be covered with an ignore mask, and only the intersection of the MODIS labels and the predictions are used for the next five epochs. Figure 3 shows the label refinement steps.

Implementation Details

Our model was implemented on PyTorch and worked on one graphics processing unit (NVIDA 2070-super). The weights of a pretrained model on the ImageNet data set are used for the initialization of our model. It is worth mentioning that the number of land covers in the training data set is different from the number of classes in the ImageNet data set, so the logit weights in the pretrained model are excluded. In this work, several modifications were made to the original DeepLabV3+ network. Preprocessing of *Sentinel-1* SAR images and data augmentation were added to the data-loader module of the network, and the structure of the network was altered to update the label during the training process. In addition, the original DeepLabV3+ is used as the baseline model to compare with our model. Both models were trained for 50 epochs, and the average time per epoch is around one hour. The training parameters of our model and the baseline are presented in Table 1.

Table 1. Training parameters used for the baseline model and our proposed model.

	Baseline	Ours
Pretrained on ImageNet	True	True
SAR image preprocessing	False	True
Data augmentation	True	True
Label refinement	False	True
Backbone network	ResNet-101	
Momentum	0.9	
Initial learning rate	0.001	
Number of epochs	50	
Batch size	16	
Output stride	16	
Weight decay	0.00005	

Experiments and Discussion

Data Set

The model is trained on a public satellite imagery data set, SEN12MS, which was published by Schmitt *et al.* (2019). This data set contains globally distributed scenes, covering inhabited continents during all meteorological seasons. SEN12MS includes 180 662 triplets of Sentinel land cover maps (see Figure 4), dual-polarized (VV and VH) SAR *Sentinel-1* image patches, and *Sentinel-2* multispectral image patches. Each image is cropped to a size of 256 × 256 pixels. While all data are oversampled to be at a ground sample distance of 10 m, the Sentinel images have a native resolution of about 10 to 60 m per pixel, and the MODIS-derived land cover has a native resolution of 500 m per pixel.

The *Sentinel-1* SAR images were provided in the original form with no preprocessing (e.g., speckle filtering). For the *Sentinel-2* multispectral images, a sophisticated mosaicking workflow was implemented to avoid the impacts of cloud-covered images. On the other hand, the MODIS land cover maps were created based on calibrated MODIS reflectance data in

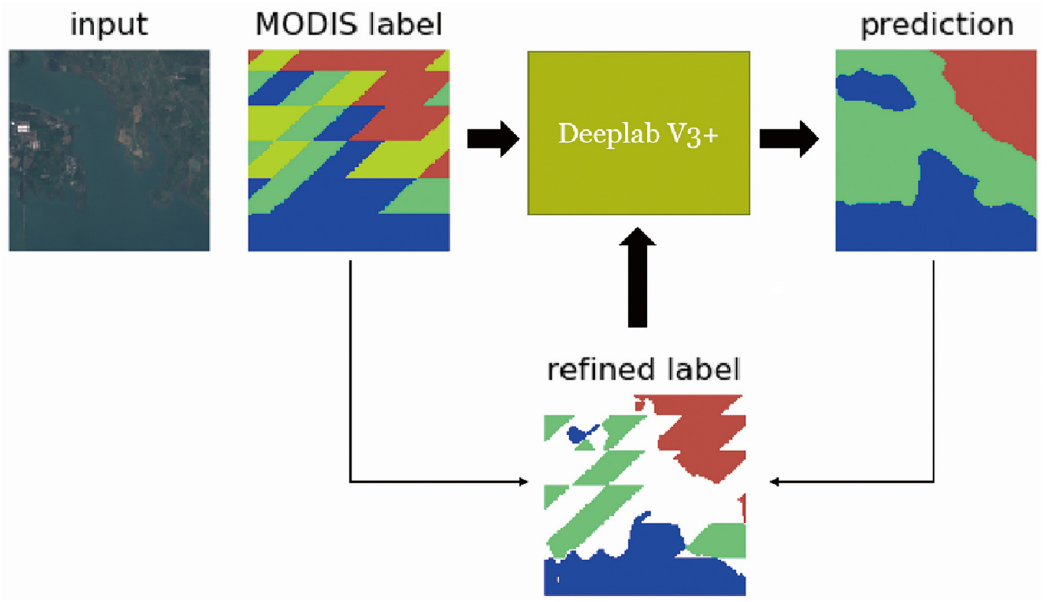

Figure 3. Label refinement process (the ignored mask is in white).

2016. The raw reflectance data was classified following the International Geosphere-Biosphere Programme (IGBP) classification scheme (Loveland and Belward 1997) and land cover classification system (LCCS) scheme (Di Gregorio 2005). Moreover, sophisticated postprocessing is carried out for class-specific refinement, which integrates prior knowledge, auxiliary information, and temporal regularization based on a Markov random field (Schmitt *et al.* 2019). For different classification schemes, the provided MODIS maps have overall accuracies of approximately 67% under IGBP, 74% under LCCS land cover, and 81% under LCCS land use (Sulla-Menashe *et al.* 2019). For this study, a simplified version of IGBP was chosen to be the classification scheme. It means that the coarse label used in this study can only correctly annotate at most 67% of the image pixels.

The data set of the 2020 IEEE Geoscience and Remote Sensing Society Data Fusion Contest (DFC2020) was used to validate and test the performance of our deep learning spatiotemporal fusion model. The DFC2020 data set contains scenes with undisclosed geolocation and not contained in the SEN12MS data set, with semimanually derived high resolution (10 m) land cover maps as the ground truth labels. In addition to the high-resolution ground truth labels, the validation and testing images are provided in the same triplet format as the training data set (i.e., corresponding *Sentinel-1*, *Sentinel-2*, and MODIS labels). The validation set contains 986 quadruplets, and the testing set has 5128 quadruplets (see Figure 5).

Classification Scheme and Evaluation Metric

A simplified version of the IGBP classification scheme is used for this project. As shown in Table 2, the original IGBP scheme has 17 classes in total. The simplified scheme has 10 classes.

The fusion results were evaluated using the classification accuracy as the quantitative indicator. The higher the accuracy is, the training model has a stronger ability to classify land cover features. The accuracy is defined by

$$\text{Accuracy} = \frac{1}{m}\sum_{i=0}^{m}[f_i == y_i], \tag{1}$$

where m is the number of samples, f_i and y_i are the true and predicted pixel label values, and [.] is the Iverson bracket operator, which evaluates to 1 when the labels match, and to 0 when labels mismatch.

Quantitative Results

It took about 25 minutes for the trained model to predict the 5128 images of the testing set. The results on the validation set and the testing set for the baseline and the proposed models are shown in Tables 3 and 4, respectively. The overall performances were assessed using average class accuracy (AA), which indicates the mean of the accuracies of all land cover classes in the simplified IGBP scheme. It is worth mentioning that the validation and testing data set does not include savanna (Class 3) and snow/ice (Class 8).

Table 2. Original and simplified IGBP land cover classification schemes.

Simplified Class No.	Simplified Class Name	IGBP Class Name	IGPB Class No.
1	Forest	Evergreen Needleleaf Forest	1
		Evergreen Broadleaf Forest	2
		Deciduous Needleleaf Forest	3
		Deciduous Broadleaf Forest	4
		Mixed Forest	5
2	Shrubland	Closed Shrublands	6
		Open Shrublands	7
3	Savanna	Woody Savannas	8
		Savannas	9
4	Grassland	Grasslands	10
5	Wetlands	Permanent Wetlands	11
6	Croplands	Croplands	12
		Cropland/Natural Vegetation Mosaics	14
7	Urban/ Built-Up	Urban/Built-Up	13
8	Snow/Ice	Permanent Snow and Ice	15
9	Barren	Barren	16
10	Water	Water Bodies	17

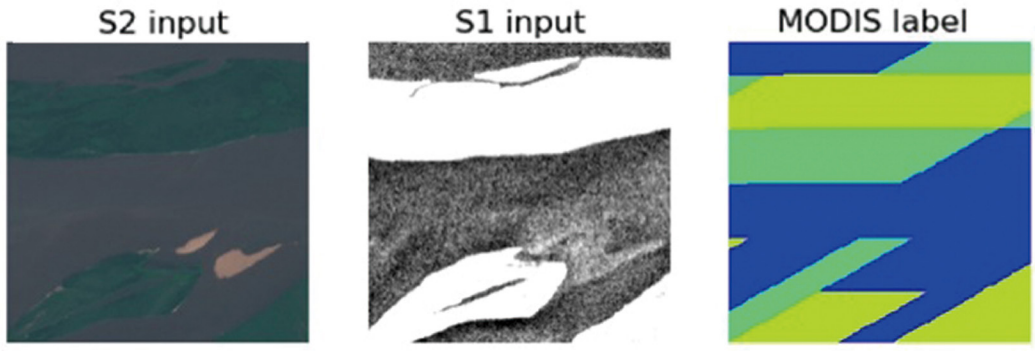

Figure 4. An example of SEN12MS triplets.

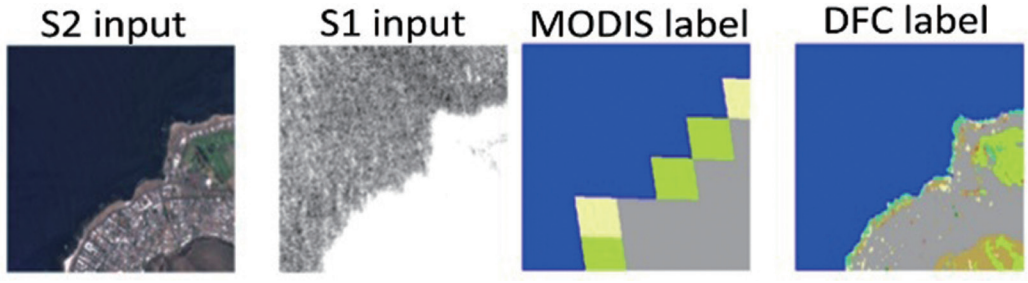

Figure 5. An example of DFC2020 quadruplets.

Table 3. Performances on the validation set.

	Baseline (%)	Ours (%)
Average class accuracy (AA)	41.34	51.95
Pixel-wise accuracy (PA)	50.17	62.99
Forest (Class 1)	62.61	85.67
Shrubland (Class 2)	1.07	13.58
Grassland (Class 4)	48.21	26.23
Wetlands (Class 5)	14.02	29.98
Croplands (Class 6)	43.54	75.04
Urban/built-up (Class 7)	66.35	84.51
Barren (Class 9)	0.21	3.74
Water (Class 10)	94.72	96.87

Table 4. Performances on the testing set.

	Baseline (%)	Ours (%)
Average class accuracy (AA)	41.12	50.18
Pixel-wise accuracy (PA)	49.93	62.57
Forest (Class 1)	60.04	74.53
Shrubland (Class 2)	2.31	14.17
Grassland (Class 4)	50.05	46.75
Wetlands (Class 5)	12.45	28.48
Croplands (Class 6)	41.59	64.29
Urban/built-up (Class 7)	69.26	77.78
Barren (Class 9)	0.37	1.20
Water (Class 10)	92.92	94.22

From the comparative analysis between the baseline model and our model, it can be observed that our model achieves 51.95% on the validation set and 50.18% on the testing set, which outperforms the baseline model that obtains 41.34% and 41.12%, respectively. Additionally, for each class, the proposed model reaches higher performance than the baseline model on most land cover classes, except grassland (Class 4). The highest accuracies are related to forest (Class 1), urban/built-up (Class 7), and water (Class 10). Considering the spectral characteristics of these classes, the high accuracies are the results of the effectiveness of the model to extract distinct pixel values. On the contrary, the model performs poorly on identifying shrubland (Class 2) and barren (Class 9). Neither of the two classes reaches 5% accuracy. It could be the results of the relatively high textural and spectral similarities between grassland and shrubland and that of urban and barren.

According to the normalized confusion matrix shown in Table 5, the deep network was successful in predicting pixels belonging to the forest, urban, and water. Compared to the other classes, water represents the least confused class. Only a small portion of water pixels was misclassified as a wetland. However, real wetland pixels are mostly confused with water pixels, while barren pixels are most likely confused with urban. Pixels belonging to shrubland, grassland, and cropland are more likely to be confused with each other. It is noticeable that these confused classes exhibit very high interclass similarities. The heterogeneity of urban areas also resulted in confusing urban pixels and all other classes.

Qualitative Comparison

In addition to the accuracy evaluations, the visualization of the predicted maps was also presented for a qualitative overview of the spatial resolution enhancement of the land cover mapping. Enhanced land cover maps obtained by our model are shown in Figures 6a and 6b to demonstrate how the model performs on predicting different land covers. Each example includes the input *Sentinel-2* multispectral image, the input *Sentinel-1* SAR image, the original MODIS label/map, the enhanced map from the prediction of our model, and the DFC2020 ground truth label/map. As shown in Figure 6, the detection of shorelines and beaches are well recognized on the enhanced land cover map, with smoothed boundaries between land cover parcels.

Figure 6b shows that the model successfully reduced the impact of the misclassified label of grassland on the corresponding MODIS land cover map. Moreover, by visually analyzing the input image and the DFC ground-truth label, we can find that the DFC map underestimates the area of urban/built-up in this image. In contrast, the enhanced map correctly detects the presence of buildings. It indicates that even the ground truth label could still contain minor misclassifications. Additionally, Figure 6c shows that the model poorly identifies narrow rivers or small ponds despite the significant spectral differences. Both Figure 6c and 6d show that our model tends to misclassify cropland, wetland, and grassland.

The incorrect MODIS label certainly misleads the prediction, but the misclassification could also result from spectral similarities between the three land covers. For example, the paddy field is one type of cropland, but it is very similar to wetland (a mix of water and vegetation) as it contains a lot of water. Additionally, irregular cropland can also be confused with natural grassland. Our study demonstrated that the proposed model has difficulties in separating shrubland from cropland or grassland, see Figure 6d, and difficulties in the segmentation of barren, see Figure 6e.

In summary, the proposed model tends to be biased toward high represented classes such as forest, grassland, and urban. This is probably related to the fact that those classes exhibit more general textural and spectral characteristics, confusing the model prediction. In any case, our model presented results superior to the baseline with a significant margin.

Conclusion

In this paper, a weakly supervised deep learning-based approach was proposed for the fusion of satellite data at high spatial resolution (*Sentinel-1* and *Sentinel-2*) with satellite-derived land cover maps at high temporal resolution (MODIS) to perform the enhanced land cover mapping. Considering the large spatial resolution gap between Sentinel and MODIS images, the fusion was conducted through a weakly supervised semantic segmentation. We modified the original DeepLabV3+ segmentation architecture by adding a label-update module to update the coarse label throughout the training automatically.

Table 5. Normalized confusion matrix for our model on DFC2020 testing set.

Class	Forest	Shrubland	Grassland	Wetland	Cropland	Urban/Built-Up	Barren	Water
Forest	0.75	0.11	0.07	0.01	0.05	0.01	0.00	0.00
Shrubland	0.23	0.14	0.45	0.02	0.14	0.01	0.01	0.00
Grassland	0.12	0.15	0.47	0.01	0.23	0.02	0.00	0.00
Wetlands	0.05	0.03	0.03	0.28	0.14	0.00	0.00	0.47
Croplands	0.01	0.07	0.13	0.10	0.64	0.00	0.00	0.05
Urban/built-up	0.06	0.02	0.04	0.02	0.03	0.78	0.01	0.04
Barren	0.07	0.14	0.05	0.00	0.05	0.62	0.01	0.06
Water	0.00	0.00	0.00	0.05	0.01	0.00	0.00	0.94

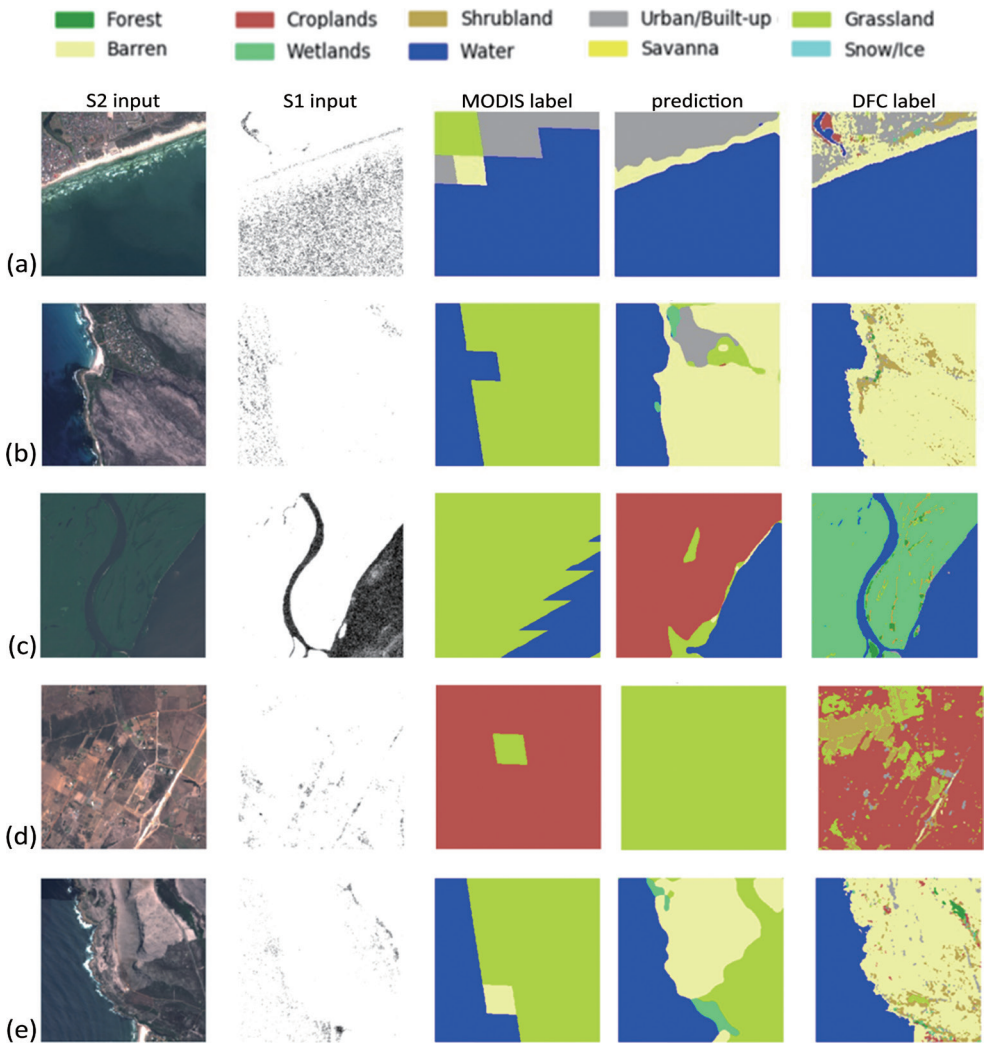

Figure 6. Five excerpts of classification results obtained by our model. (a) detection of shoreline and beach, (b) reduced impact of the weak MODIS label, (c) misclassifications of rivers and and wetlands, (d) misclassification of cropland and shrubland, and (e) misclassification of barren land.

The experiment results have validated the effectiveness and potential of deep learning-based semantic segmentation architecture in the fusion of multi-source satellite data, improving land cover mapping.

References

Ayhan, B. and C. Kwan. 2020. Tree, shrub, and grass classification using only RGB images. *Remote Sensing* 12:1333. doi: 10.3390/rs12081333.

Chan, L., M. S. Hosseini and K. N. Plataniotis. 2020. A comprehensive analysis of weakly-supervised semantic segmentation in different image domains. *International Journal of Computer Vision*. doi: 10.1007/s11263-020-01373-4.

Chen, B., B, Huang and B. Xu. 2017. Multi-source remotely sensed data fusion for improving land cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing* 124:27–39.

Chen, G., C. Li, W. Wei, M. Woźniak, T. Blažauskas and R. Damaševičius. 2019. Fully convolutional neural network with augmented atrous spatial pyramid pool and fully connected fusion path for high resolution remote sensing image segmentation. *Applied Sciences* 9 (9):1816. doi: 10.3390/app9091816.

Chen, L. C., Y. Zhu, G. Papandreou, F. Schroff and H. Adam. 2018a. Encoder-decoder with atrous separable convolution for semantic image segmentation. Pages 801–818 in *Proceedings of the European Conference on Computer Vision (ECCV)*, held in Munich, Germany.

Chen, Y., D. Ming, L. Zhao, B. Lv, K. Zhou and Y. Qing. 2018b. Review on high spatial resolution remote sensing image segmentation evaluation. *Photogrammetric Engineering and Remote Sensing* 84 (10):629–646.

Di Gregorio, A. 2005. *Land Cover Classification System: Classification Concepts and User Manual*, *Software Version 2*. Rome, Italy: Food and Agriculture Organization of the United Nations (FAO), ISBN: 92-5-105327-8.

Du, Z., J. Yang, C. Ou and T. Zhang. 2019. Smallholder crop area mapped with a semantic segmentation deep learning method. *Remote Sensing* 11 (7):888. doi: 10.3390/rs11070888.

Fung, C. H., M. S. Wong and P. W. Chan. 2019. Spatio-temporal data fusion for satellite images using hopfield neural network. *Remote Sensing* 11 (18):2077. doi: 10.3390/rs11182077.

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

Gevaert, C. M. and F. J. García-Haro. 2015. A comparison of STARFM and an unmixing-based algorithm for Landsat and MODIS data fusion. *Remote Sensing of Environment* 156:34–44.

Ghamisi, P., B. Rasti, N. Yokoya, Q. Wang, B. Höfle, L. Bruzzone, F. Bovolo, M. Chi, K. Anders, R. Gloaguen, P. Atkinson and J. Benediktsson. 2019. Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art. *IEEE Geoscience and Remote Sensing Magazine* 7 (1):6–39.

Hilker, T., M. A. Wulder, N. C. Coops, J. Linke, G. McDermid, J. G. Masek, F. Gao and J. C. White. 2009. A new data fusion model for high spatial-and temporal-resolution mapping of forest disturbance based on Landsat and MODIS. *Remote Sensing of Environment* 113 (8):1613–1627.

Huang, B., B. Zhao and Y. Song. 2018. Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery. *Remote Sensing of Environment* 214:73–86.

Kemker, R., C. Salvaggio and C. Kanan. 2018. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing* 145:60–77.

Lee, J. S. 1981. Refined filtering of image noise using local statistics. *Computer Graphics and Image Processing* 15 (4):380–389.

Loveland, T. R. and A. Belward. 1997. The international geosphere biosphere programme data and information system global land cover data set (discover). *Acta Astronautica* 41 (4–10):681–689.

Schmitt, M., L. H. Hughes, C. Qiu and X. X. Zhu. 2019. Sen12ms—A curated dataset of georeferenced multispectral Sentinel-1/2 imagery for deep learning and data fusion. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences.* doi: 10.5194/isprs-annals-IV-2-W7-153-2019.

Song, H., Q. Liu, G. Wang, R. Hang and B. Huang. 2018. Spatiotemporal satellite image fusion using deep convolutional neural networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11 (3):821–829.

Song, X. P., C. Huang and J. R. Townshend. 2017. Improving global land cover characterization through data fusion. *Geospatial Information Science* 20 (2):141–150.

Sulla-Menashe, D., J. M. Gray, S. P. Abercrombie and M. A. Friedl. 2019. Hierarchical mapping of annual global land cover 2001 to present: The MODIS collection 6 land cover product. *Remote Sensing of Environment* 222:183–194.

Sun, Y. and H. Zhang. 2019. A two-stage spatiotemporal fusion method for remote sensing images. *Photogrammetric Engineering and Remote Sensing* 85 (12):907–914.

Wang, B., J. Che, B. Wang and S. Feng. 2018. A solar power prediction using support vector machines based on multi-source data fusion. Pages 4573–4577 in *Proceedings of 2018 International Conference on Power System Technology (POWERCON)*, held in Guangzhou, China.

Wang, J., C. Li and P. Gong. 2015. Adaptively weighted decision fusion in 30 m land-cover mapping with Landsat and MODIS data. *International Journal of Remote Sensing* 36 (14):3659–3674.

Xie, D., J. Zhang, X. Zhu, Y. Pan, H. Liu, Z. Yuan and Y. Yun. 2016. An improved STARFM with help of an unmixing-based method to generate high spatial and temporal resolution remote sensing data in complex heterogeneous regions. *Sensors* 16 (2):207. doi: 10.3390/s16020207.

Zhu, X., J. Chen, F. Gao, X. Chen and J. G. Masek. 2010. An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions. *Remote Sensing of Environment* 114 (11):2610–2623.

Zhu, X., E. H. Helmer, F. Gao, D. Liu, J. Chen and M. A. Lefsky. 2016. A flexible spatiotemporal method for fusing satellite images with different resolutions. *Remote Sensing of Environment* 172:165–177.

Zurita-Milla, R., J. G. Clevers and M. E. Schaepman. 2008. Unmixing based Landsat TM and Meris FR data fusion. *IEEE Geoscience and Remote Sensing Letters* 5 (3):453–457.