SPATIAL RESOLUTION ENHANCEMENT OF LAND COVER MAPPING USING DEEP CONVOLUTIONAL NETS

Q. Yu[a], W. Liu[b], J. Li[c, *]

[a] Dept. of Geography and Environmental Management, University of Waterloo, Canada - q45yu@uwaterloo.ca
[b] Virtual Reality and Interactive Techniques Institute, East China Jiaotong University, Jiangxi, China - weiliu@ecjtu.edu.cn
[c] Dept. of Geography and Environmental Management, University of Waterloo, Canada - junli@uwaterloo.ca

KEY WORDS: Spatiotemporal Fusion, Spatial Resolution Enhancement, Land Cover Mapping, Sentinel, MODIS, Deep Learning

ABSTRACT:

Multispectral satellite imagery is the primary data source for monitoring land cover change and characterizing land cover at the global scale. However, the accuracy of land cover classification is often constrained by the spatial and temporal resolutions of the acquired satellite images. This paper proposes a novel spatiotemporal fusion method based on deep convolutional neural networks under the application background of massive remote sensing data, as well as the large spatial resolution gaps between MODIS and Sentinel images. The training was taken on the public SEN12MS dataset, while the validation and testing were conducted using ground truth data from the 2020 IEEE GRSS data fusion contest. As a result of data fusion, the synthesized land cover map was more accurate than the corresponding MODIS-derived land cover map, with an enhanced spatial resolution of 10 meters. The ensemble approach can be implemented for improving data quality when generating a global land cover product from coarse satellite imageries.

1. INTRODUCTION

Remote sensed satellite imagery is the primary data source for monitoring land cover change (LCC) and characterizing land cover at the global scale (Song et al., 2017). However, the accuracy of land cover classification is often constrained by the spatial and temporal resolutions of the acquired satellite images. For instance, Landsat satellites capture images with a moderate spatial resolution of 30 meters but with a long revisit period of 16 days. To the contrary, the Moderate resolution Imaging Spectroradiometer (MODIS) can provide images on a daily basis, 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 in an efficient way as conducting land cover classification. For the purpose of having up-to-date land cover monitoring with fine spatial scale, increasing the spatial resolution of coarse satellite imagery represents a continued advancement in remote sensing research. Global-scale land cover mapping at coarse resolution has been driven by the availability of MODIS dataset, previous researches have conducted spatiotemporal fusion to blend MODIS and Landsat data in order to obtain improved classification results with a higher spatial resolution of 30m (Gevaert and García-Haro, 2015, Wang et al., 2015, Chen et al., 2017). Sentinel-1 and Sentinel-2 are two recently launched satellite constellations which provide higher temporal resolution (3 – 5 days) and higher spatial resolution (5 – 10 meters) than Landsat satellites. These advantages are fundamental in a spatiotemporal fusion process for improving land cover classification.

Recently, deep learning frameworks have enhanced the classification performance by automatic extraction of deep features. Therefore, deep learning-based land cover classification has become a current hotpot in the remote sensing research community. One of the major advantages of using deep learning algorithms is that neutral network is a learning-based method, which automatically learns an end-to-end mapping between coarse resolution images and fine resolution images.

*Corresponding author

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. With the aim of providing enhanced land cover mapping through the fusion of multisource satellite data, this study proposes an end-to-end deep learning method to enhance the spatial resolution of MODIS-derived land cover maps, by integrating the maps (with 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, additional module was embedded in the model, and it automatically updates the coarse labels based on the intermediate predictions on the training sets.

2. METHOD

2.1 DeepLabV3 Plus

The network architecture of the proposed data-fusion model is based on the semantic segmentation framework developed by Chen et al. (2018), namely DeepLabV3 Plus. It is the latest version of DeepLab semantic segmentation architecture, which utilizes a spatial pyramid pooling 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 state-of-the-art mean IOU of 89% on PASCAL VOC 2012 test.

For this study, several modifications were made based on original design of DeepLabV3 Plus. The framework was implemented on TensorFlow, while the proposed model is implemented on PyTorch. The original backbone network Xception+ was replaced

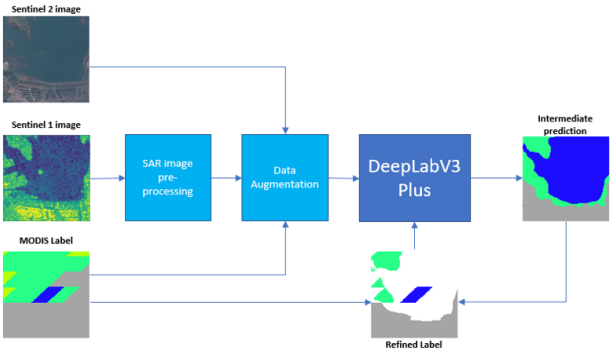

Figure 1: Framework overview of the proposed model

with the deep residual network ResNet-101. For model initialization, the proposed model was pre-trained on ImageNet.

2.2 Pre-processing 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 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 those pixels in the edge directed window. After the speckle filtering, the images were enhanced by 2% linear stretch. With the lowest 2% values 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 4, the noise in the high contrast areas is effectively removed and the edges are enhanced.

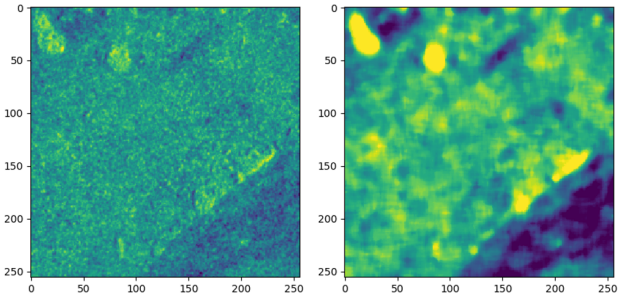

Figure 2: Example of raw Sentinel-1 SAR image (left). The corresponding processed Sentinel-1 SAR image (Right)

2.3 Data Augmentation

To improve the performance by enlarging the training dataset, several augmentation techniques have been added to the data-loader module of the model network. These includes 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).

2.4 Label Refinement

In essence, the major task of this study is semantic segmentation on weakly supervised training, which the annotation (i.e. MODIS labels) is noisy and unreliable. To further improve the performance of the model, additional strategies should be adopted to deal with noisy label specifically. In SEN12MS dataset, 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 5 epochs (an epoch refers to one cycle through the full training dataset). Hence, only for the first five epochs, the model was trained on original MODIS labels. After the fifth epoch, the model outputs the intermediate predictions on all training samples, and then obtain 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 5 epochs.

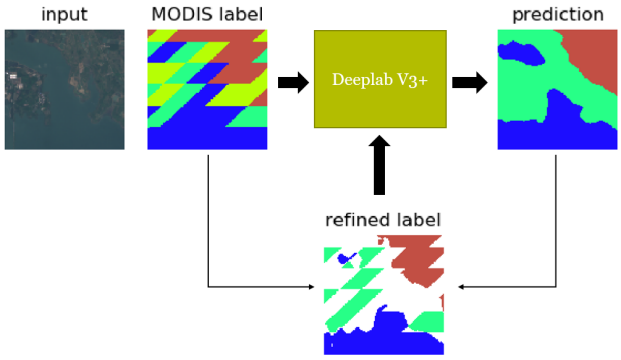

Figure 3: Label refinement process (Ignore mask is in white)

3. EXPERIMENTS

3.1 DATA

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

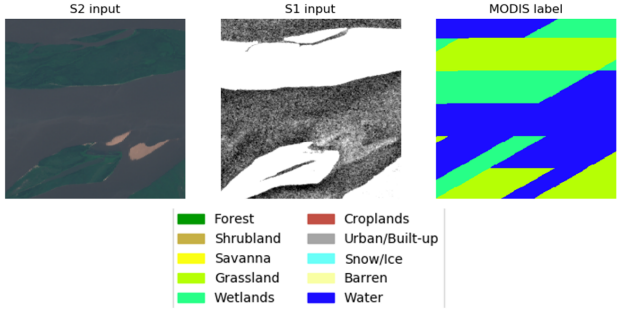

Figure 4: An example of SEN12MS triplets

The Sentinel-1 SAR images are provided in original form with no pre-processing (e.g. speckle filtering). As for the Sentinel-2 mul-

tispectral images, a sophisticated mosaicking workflow was implemented to avoid the impacts of cloud-covered images. On the other hands, the MODIS land cover maps were created based on calibrated MODIS reflectance data in 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 post-processing 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. Detailed information of the chosen classification scheme is presented in the section 2.3.

3.1.2 Validation and Testing datasets To validate and test the performance of presented deep learning spatiotemporal fusion model, the dataset of 2020 IEEE GRSS data fusion contest (DFC2020) is used. The DFC2020 dataset contains scenes with undisclosed geolocation and not contained in the SEN12MS dataset, with semi-manually derived high resolution (10m) 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 dataset (i.e. corresponding Sentinel-1, Sentinel-2, and MODIS labels). The validation set contains 986 quadruplets, and the testing set has 5128 quadruplets.

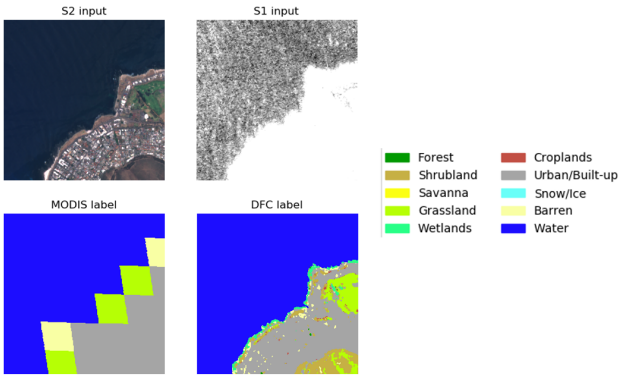

Figure 5: An example of DFC2020 quadruplets

3.1.3 Classification scheme A simplified version of the IGBP classification scheme is used for this project. As shown in the Table 1 below, the original IGBP scheme has 17 classes in total. The simplified scheme has 10 classes.

3.2 Implementation details

In addition to the proposed model, the original DeepLabV3 Plus is used as the baseline model which compares to the proposed model. Both models were trained for 50 epochs. The parameter settings of the baseline model and the proposed model are shown below in Table 2. The implementation details of the proposed model and the baseline are presented in Table 2.

3.3 Experiment results

The results on the validation set and the testing set for the baseline and the proposed models are shown in Table 3 and Table 4,

Table 1: The original and simplified IGBP Land Cover Classification schemes

Simplified Class Number	Simplified Class Name	IGBP Class Name	IGPB Class Number
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

Table 2: Parameters setting of the baseline model and the proposed model.

	Baseline	Proposed
Pretrained on ImageNet	True	True
Label Refinement	False	True
SAR image pre-processing	False	True
Data augmentation	False	True
Backbone network	Xception+	ResNet101
Initial learning rate	0.01	0.001
Batch size	16	16
Output Stride	16	16

respectively. The performances are 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 dataset does not include Savanna (Class #3) and Snow/Ice (Class #8).

From the comparative analysis between the baseline model and the proposed model, it can be observed that the proposed model achieves 51.95% on the validation set and 50.18% on the testing set, which outperforms the baseline model, which obtains 41.34% and 41.12%. Additionally, for each individual class, the proposed model reaches higher performance than the baseline model on most of the 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 reach 5% accuracy. It could be the results of the relatively high textural and spectral similarities between grassland and shrubland, as well as that of urban and barren.

3.4 Visualized 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. As shown in the figures below, some enhanced land cover maps obtained by the proposed model are provided as example 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 the proposed model, and the DFC2020 ground truth label/map. As shown in

Table 3: Performances on DFC2020 Validation set.

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

Table 4: Performances on DFC2020 Testing set.

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

Figure 6, the detection of shorelines and beaches are well recognized on the enhanced land cover map, with smoothed boundaries between land cover parcels.

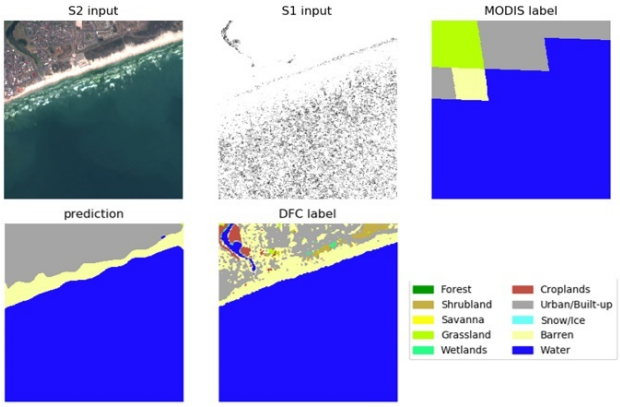

Figure 6: Detection of shoreline and beach

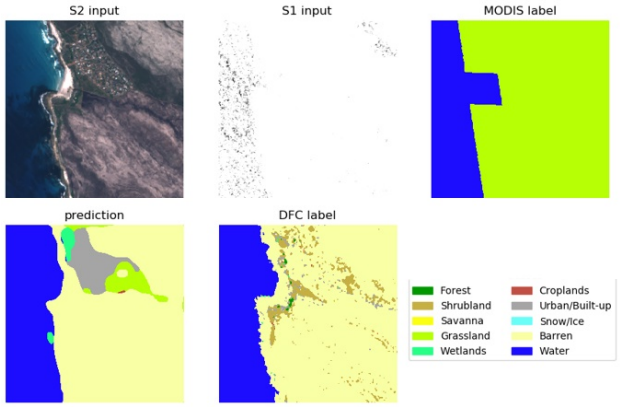

Figure 7: Reduced impact of the weak MODIS label

It can be seen from Figure 7 that the model successfully minimized the impact of misclassified label of grassland on the corresponding MODIS land cover map. Moreover, as 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, while the enhanced map correctly detects the presence of buildings. It indicates that even the ground truth label could still contain minor misclassifications.

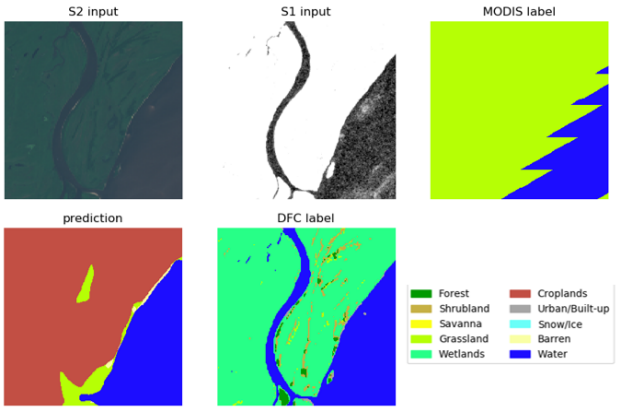

Figure 8: Misclassifications of rivers and wetlands

As shown in the example image in Figure 8, the model performs poorly on identifying narrow rivers or small ponds in despite of the significant spectral differences. Both Figure 8 and Figure 9 show that the proposed model has the tendency to misclassify cropland, wetland and grassland. The incorrect MODIS label certainly mislead the prediction, but the misclassification also could be a result of spectral similarities between the three land covers. For example, 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.

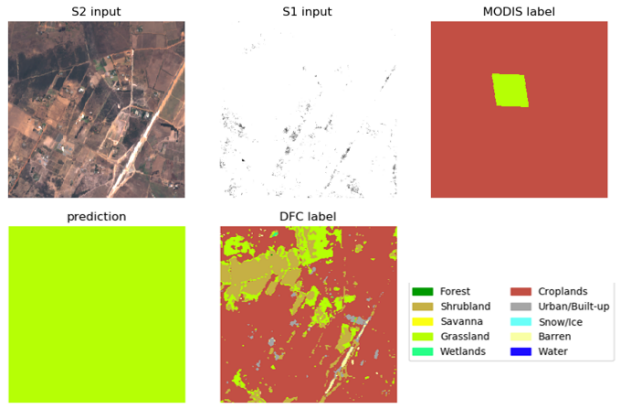

Figure 9: Misclassification of cropland and shrubland

It also can be seen from Figure 9 that the model has difficulties identifying shrubland from cropland or grassland.

To conclude, the 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 a more general textural and spectral characteristics, which could be confusing for the model prediction.

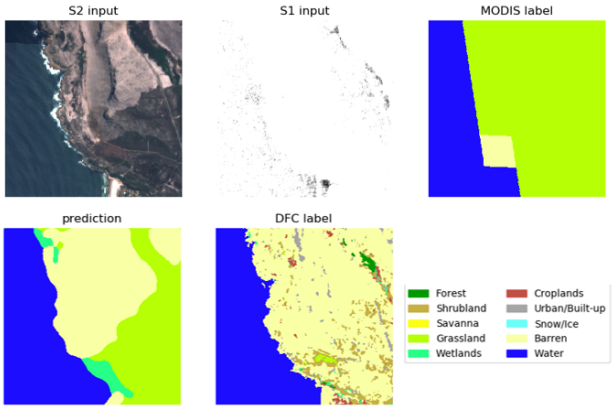

Figure 10: Misclassification of barren

3.5 Foreseeable limitations

Given the computational capacity of the machine platform used for this study, the model has not been trained for sufficient number of epochs. It can be expected that the train loss would continue to decrease after 20 epochs. Therefore, if the model could be trained for more epochs, the result may be further improved. Moreover, there was no fine tuning of the hyperparameters of the model. Most of the hyperparameters were set to be same as the original DeepLabV3 plus model. The performance is expected to be better if the fine tuning of the hyperparameters was carried out.

Furthermore, the proposed label refinement method is relatively simple and naïve comparing to the current state-of-art techniques used in weakly supervised semantic segmentation, such as Expectation Maximization, Multiple Instance Learning, Self Supervised Learning, and Object Proposal Class Inference (Chan et al., 2019). The model could be further improved by adding those techniques as additional modules to deal with noisy annotations.

4. CONCLUSION

In this paper, a deep learning-based model was proposed for the fusion of satellite data at high spatial resolution with satellite-derived land cover maps at high temporal resolution in order to perform enhanced land cover mapping. Experiment results have validated the effectiveness and potential of deep learning-based semantic segmentation architecture in the fusion of multisource satellite data and improving land cover mapping. In the future, the model could be further improved by adding more sophisticated techniques of handling weak annotation.

REFERENCES

Chan, L., Hosseini, M. S. and Plataniotis, K. N., 2019. A comprehensive analysis of weakly-supervised semantic segmentation in different image domains. arXiv preprint arXiv:1912.11186.

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

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. and Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp. 801–818.

Di Gregorio, A., 2005. Land cover classification system: classification concepts and user manual: LCCS. Vol. 2, Food & Agriculture Org.

Gevaert, C. M. and García-Haro, F. J., 2015. A comparison of starfm and an unmixing-based algorithm for landsat and modis data fusion. Remote sensing of Environment 156, pp. 34–44.

Lee, J.-S., 1981. Refined filtering of image noise using local statistics. Computer graphics and image processing 15(4), pp. 380–389.

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

Schmitt, M., Hughes, L. H., Qiu, C. and Zhu, X. X., 2019. Sen12ms–a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. arXiv preprint arXiv:1906.07789.

Song, X.-P., Huang, C. and Townshend, J. R., 2017. Improving global land cover characterization through data fusion. Geospatial information science 20(2), pp. 141–150.

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

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