remote sensing

MDPI

Letter

Uncertainty-Based Human-in-the-Loop Deep Learning for Land Cover Segmentation

Carlos García Rodríguez [1,*], Jordi Vitrià [1] and Oscar Mora [2]

[1] Department of Mathematics and Computer Science, Universitat de Barcelona, 08007 Barcelona, Spain; jordi.vitria@ub.edu

[2] Institut Cartogràfic i Geològic de Catalunya, Parc de Montjuïc, 08038 Barcelona, Spain; oscar.mora@icgc.cat

[*] Correspondence: c.garcia.rodriguez@ub.edu

Received: 19 October 2020; Accepted: 18 November 2020; Published: 23 November 2020

Abstract: In recent years, different deep learning techniques were applied to segment aerial and satellite images. Nevertheless, state of the art techniques for land cover segmentation does not provide accurate results to be used in real applications. This is a problem faced by institutions and companies that want to replace time-consuming and exhausting human work with AI technology. In this work, we propose a method that combines deep learning with a human-in-the-loop strategy to achieve expert-level results at a low cost. We use a neural network to segment the images. In parallel, another network is used to measure uncertainty for predicted pixels. Finally, we combine these neural networks with a human-in-the-loop approach to produce correct predictions as if developed by human photointerpreters. Applying this methodology shows that we can increase the accuracy of land cover segmentation tasks while decreasing human intervention.

Keywords: deep learning; human-in-the-loop; land cover segmentation

1. Introduction

Public cartographic institutions such as the Cartographic and Geological Institute of Catalonia explore how to substitute their strenuous and time-consuming manual tasks for automated processes. To do so, current accuracy in automatic land cover segmentation requires to be improved. These institutions work with images provided by sensors mounted on airplanes, helicopters, and satellite images, which we deal with in this paper.

Since 1959, when the US Explorer 6 made the first satellite image of the earth, many other satellites were placed into orbit to photograph our planet, allowing us to study natural disasters, the evolution of climate change, water resources, or land surfaces. Current satellites, such as Sentinel-2, use sensors to provide large images with a resolution of up to 10 m and different spectral bands. For example, a Sentinel-2 image from Spain (505,990 km^2) would be, roughly, 70,000 × 70,000 pixels for each of its 13 bands. As sensors technology improves over time, images are becoming better in spectral quality and resolution.

Land covers are used in different realms such as forestry for inventory area estimates [1], hydrology regarding microclimatic variables [2], agriculture to improve irrigation [3] or geology in geological hazard, risk identification, and assessment [4], among others.These land covers are provided by institutions that invest time and human resources to fulfill the costly task of producing them. Therefore, knowledge and updated data about land dynamics are essential for territory management with different purposes and multiple fields. Land covers and land classification is one of the main uses of satellite images and the focus of this work.

In cartographic institutions, due to the legal and administrative consequences of their evaluations, land classification is still done manually employing photointerpretation techniques, entailing very high costs in terms of time and human resources. Therefore, cartographic institutes are endeavoring to transit from manual land classification to automation. The transformation towards an automatic solution tends to face a critical point: the low availability of high-resolution land images. However, the new Sentinel-1 and Sentinel-2 missions of the European Spatial Agency, under the Copernicus project, brought about a new paradigm. In the case of multi-spectral sensors, the 2A and 2B satellites provide up to 10-meter pixel resolution every five days. The availability of this detailed information generates the option to explore different artificial intelligence techniques and to exploit all these data.

Deep learning [5] was widely applied to classify aerial and satellite images at the pixel level [6]. In this regard, the results corresponding to different kinds of architectures are diverse in terms of accuracy. The acceptable accuracy for this kind of land cover segmentation applications was set by the Cartographic and Geological Institute of Catalonia at a minimum of 90%. Therefore, the work we present aimed to improve the accuracy of the state of the art results to reach the requested standards.

In this paper, we propose to combine deep learning techniques with human expertise to have an acceptable land cover segmentation, reducing the amount of human time dramatically. To do so, we combined two parallel neural networks, the first one used to classify each pixel, and the second one to determine the confidence of each pixel classification. Next, a committee of experts will decide which amount of the uncertain pixels will be sent to the human photointerpreters who will establish the accuracy/human effort trade-off. Therefore our methodology incorporates a human-in-the-loop approach , where depending on the threshold in the clustering operation, we will reach a different level of accuracy in the final prediction (see Figure 1). This research's main contribution is the development of a combined method of Deep Learning techniques for land cover segmentation and human in the loop, which improves the state of the art accuracy. Furthermore, the method reduces the time and human resources required to carry out this task.

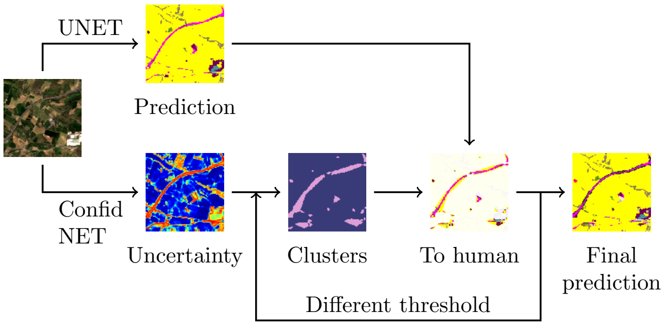

Figure 1. Overview of the proposed method.

2. Related Work

Most aerial and satellite imagery datasets are designed to evaluate the classification problem [7], where each image, or patch, is labeled only with one class. Therefore, segmentation-based datasets, where a class represents each pixel, are scarce. An analysis of related work on semantic segmentation for aerial [8] and satellite [9] images show two key features: small datasets with few classes or big datasets without reliable labeling.

These datasets are usually formed by compiling a set of images from a minimal area, such as a university campus (Pavia University dataset), a small area from an agricultural region in California (Salinas Valley dataset) or a zone in North-western Indiana containing agricultural and forest territories (Indian Pines dataset).

The three datasets mentioned above are the most used for hyperspectral segmentation. In Table 1, we summarized the best scores obtained from different authors [10] for those particular datasets. At first glance, results show high accuracy, although a more in-depth analysis reveals two key aspects: there are few classes, and the scene is tiny.

Table 1. Benchmark for classical datasets for hyperspectral segmentation.

Dataset	Size (Pixels)	Classes	Average Accuracy (%)
Salinas Valley	512 × 217	16	99.33
Pavia Centre	1096 × 1096	9	99.61
Indian Pines	145 × 145	16	98.46

We are not aware of bigger datasets except for the SEN12MS dataset, which consists of 180, 662 patches with Sentinel-1 and Sentinel-2 data. The labeling of this dataset is based on MODIS (Moderate Resolution Imaging Spectroradiometer, an on-board sensor of the Terra satellite, which provides automatic land cover labeling. Although ground truth is not based on human labeling, it is revised frequently and assures an accuracy higher than 81% for land uses). In this study, average accuracy peaked at 43.6% [11], showing a clear difference concerning smaller datasets.

Drawing upon an analysis of related work, we argue that small datasets are relevant from an academic point of view, but if we want to apply new techniques to legal binding land cover classification, state of the art results have to be improved. To reach better results on segmentation, we propose developing a new deep learning methodology with a human-in-the-loop approach.

3. Proposed Method

Current segmentation accuracy for real applications using automatic techniques, as shown in the previous section, is not acceptable for a cartographic institution to produce their land covers. We address this issue by proposing a methodology that can be useful to improve segmentation accuracy in large datasets.

3.1. Segmentation

One of the most used networks in segmentation tasks is the UNET deep network. This architecture *consists of a contracting path to capture contact and a symmetric expanding path that enables precise localization* [12]. The internal part of the UNET consists of an encoder-shaped Resnet18 and a decoder-shaped symmetric Resnet18, which allows an increase of the number of inner layers without losing prediction performance [13]. In its origin, it was designed to work with RGB images (3 channels), but, in our case, we should adapt the input of the encoder to as many channels as our sensors provide [14].

We propose to measure our model's performance using the cross-entropy loss function [15], comparing our predictions with the labeled segmentation. For this purpose, Adam optimizer [16] was used to update the weights of the model.

The representation of the complete model is illustrated in Figure 2, where we can see the input image (ten spectral channels from Sentinel-2) entering the encoder, passing through the bottleneck to then being expanded again into the decoder. All the intermediate connections are used to preserve the resolution of the original image. Finally, the softmax function will return the most probable class, which will be compared with the ground truth to update the model's weights in the training process.

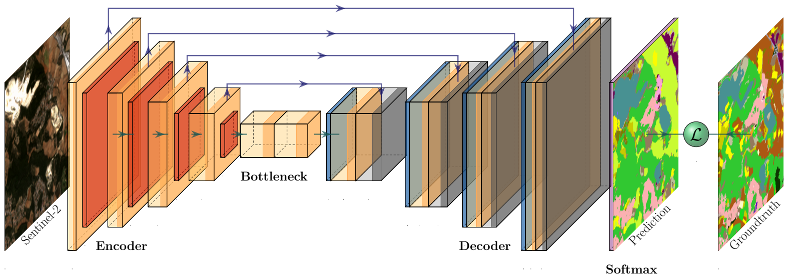

Figure 2. UNET architecture with Sentinel-2 10 bands as input.

3.2. Uncertainty Detection

Corbière et al. [17] pointed out that correct and incorrect predictions correlate in terms of uncertainty when using softmax, meaning that many of the incorrectly predicted images were misclassified with high confidence values. To redirect that issue, they propose to use a confidence network, ConfidNet. In Figure 3 we reproduced this behavior for our dataset. As we can see, wrong predictions decreased in high confident samples.

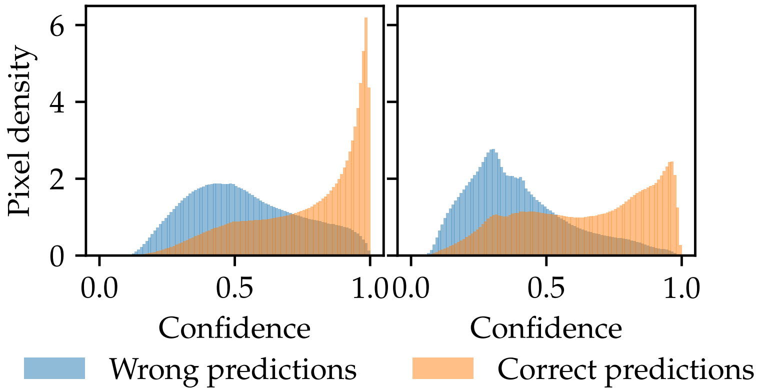

Figure 3. Output confidence using softmax (**left**) and using ConfidNet (**right**). In the second scenario, wrong predictions with high confidence are minimized.

Confidnet is a neural network built on top of a UNET neural network, benefiting from its latent features. We do not intend to improve the model's accuracy; what we want to do is to score the predictions with a confidence value.

Figure 4 shows the change in the intersection area of wrong and correct predictions. Top 60% certain pixels using ConfidNet are better classified than the top 60% using softmax.

ConfidNet takes as input the last decoder layer of the UNET (just before the softmax function), applies a convolutional neural network, and finally, a sigmoid function. The objective is to output an image with values between 0 and 1, where small values mean high uncertainty and vice versa. This convolutional neural network comprises five layers with kernel size at 3×3 and 400, 120, 64, 64, 1 as output layers.

If we want to do a robust human-in-the-loop procedure, we cannot rely on the largest softmax probability because we would misclassify many examples with high confidence. In that case, we will use ConfidNet, which will be used together with the UNET to provide us with the uncertainty for each classified pixel.

To train the confidence network, we first compute the segmentation output of the UNET prediction and label as one those that were correctly predicted and as zero those that were incorrectly predicted. Then, binary cross-entropy loss [15] is used to train the confidence network. It is important to remark that we should freeze the encoder, bottleneck, and decoder layers of the UNET because we want the predictions to remain the same, as our purpose is to improve is the confidence of the model. A simplified scheme of this architecture can be found in Figure 5.

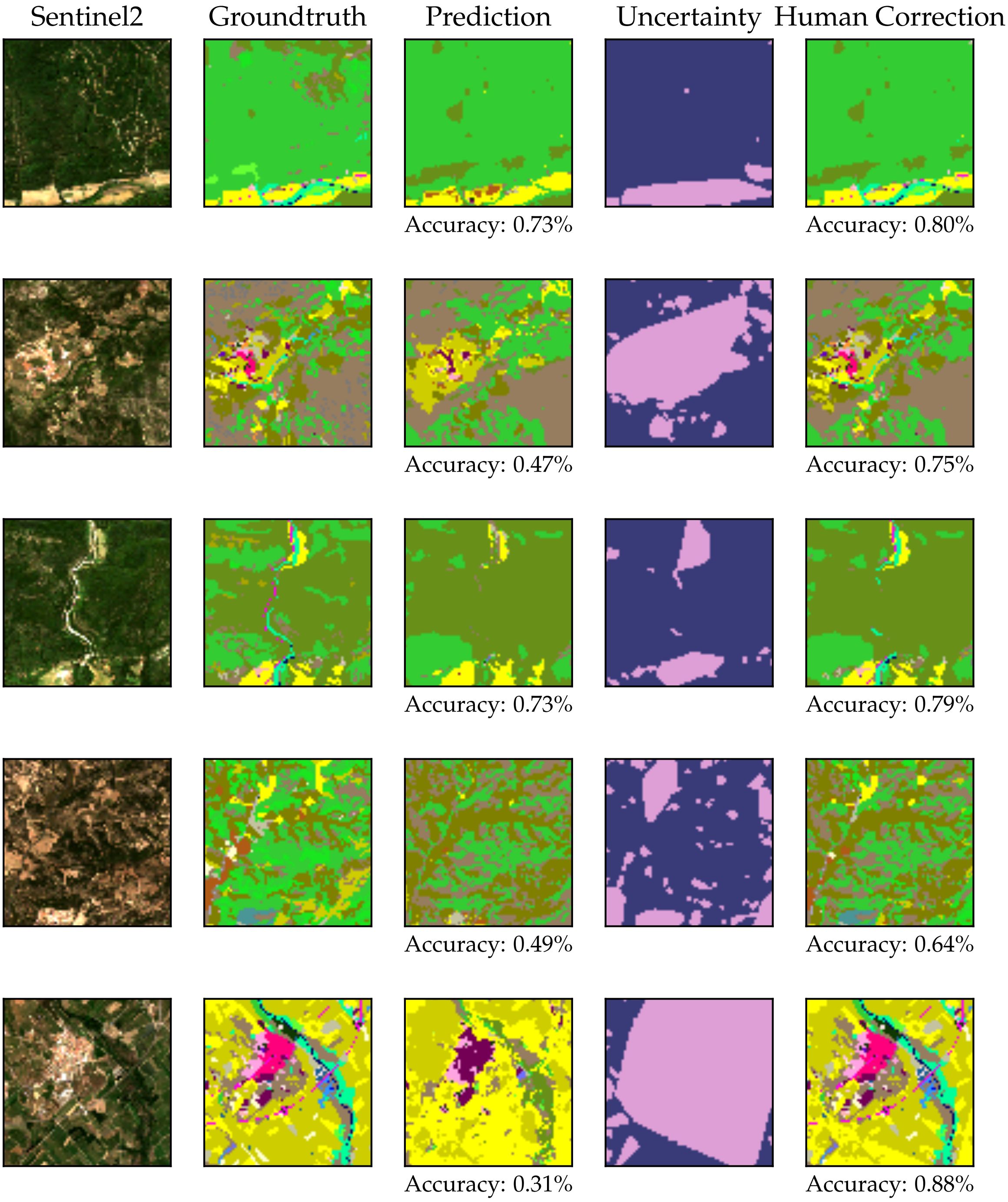

Figure 4. Intersection in confidence between wrong and correct predictions using softmax and ConfidNet (the lower, the better).

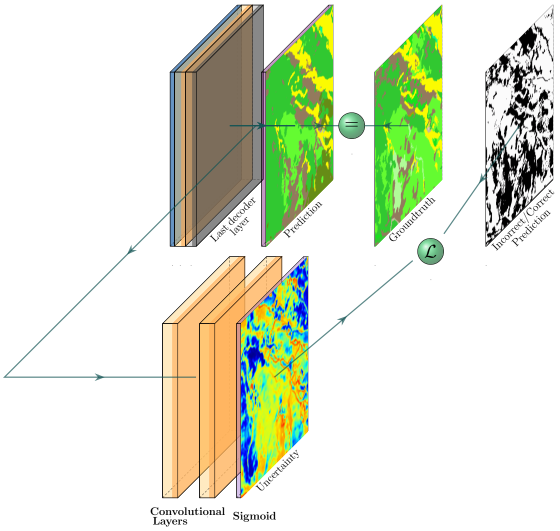

Figure 5. ConfidNet architecture.

3.3. Human-in-the-Loop

Once the model is trained, and segmentation of an image tested, two outputs are predicted: One, a segmented image where each pixel labeled with its most probable class (a value between 0 and the total number of classes minus 1). Two, an image where each pixel will have a value between 0 and 1 (high and low uncertainty). At this point, we propose to incorporate a human-in-the-loop approach to boost the results.

This approach requires an expert committee to analyze the prediction of the neuronal network. After the evaluation, the experts will decide the uncertainty threshold. To do so, they will have to assume the trade-off between the accuracy and the human resources allocated and spent on the task. Once the threshold is set, the complete map predicted by the UNET, with the uncertainty zones below the threshold highlighted, will be sent to the photo interpreters to be revised and corrected if needed. For instance, if the required human effort is set as a 10% of work that would be needed to produce the land cover without automatic methods, 10% of the most uncertain pixels will be sent to the photo interpreters, and 90% will remain as predicted by the network. We propose having a final segmented image done between the presented neural network (UNET) and the photo interpreters.

4. Experimental Results

4.1. Data

The experiments draw upon images obtained by the MSI sensor from the Sentinel-2A and 2B satellites in April 2019. The images obtained from the European Commission Copernicus program were atmospherically corrected using the ESA sen2cor v2.8 software [18] using a 10 m gridded DEM (Digital Elevation Model) generated by Cartographic and Geological Institute of Catalonia from photogrammetric aerial imagery. We cropped a total area of 30,000 × 30,000 pixels that contains the region of Catalonia (Spain). We worked with the following Sentinel-2 bands: 2,3,4 for RGB region; 5,6,7 for the vegetation red edge region; 8,8A for the near-infrared region and 11,12 for the short-wave infrared. Bands 1, 9 and 10 were rejected as they do not provide spectral information for our purpose. All bands were equally converted to the same resolution of 10 m (30 m bands were bilinearly interpolated at 10 m resolution). These images are divided into smaller patches of size 224 × 224 pixels. Sentinel-2 data was scaled down by a factor of 10,000, and clipped between 0 and 1, being 0 the minimal reflectivity and 1 the maximum. All data is normalized between 0 and 1.

The ground truth dataset (This dataset, provided by the ICGC, can be found by using the following Web Map Service address: https://geoserveis.icgc.cat/servei/catalunya/cobertes-sol/wms?) contains 41 different classes, such as herbaceous crops, deciduous forests or lakes and lagoons. Two of those (beach and sea) were deleted due to the small representation in the dataset. The 39 classes to classify can be found in Table 2. It is relevant to note that classes are highly unbalanced. Some classes are over-represented such as "herbaceous crops", while others such as "crops in transformation" are scarce. We decided not to do weighted training because, in segmented-based cases, it is complicated to achieve an accurate weight calibration.

Table 2. 39 Possible classes for the ICGC land cover dataset. The color tag allows class identification in the figures of this paper.

	Agricultural Area		**Urban Area**
0	Herbaceous crops	19	Urban area
1	Orchard	20	"Eixample"
2	Vineyards	21	Lax urban areas
3	Olive groves	22	Isolated buildings
4	Other woody crops	23	Isolated residential areas
5	Crops in transformation	24	Green areas
	Forestal area	25	Industrial or commercial
6	Dense coniferous forests	26	Sports and leisure areas
7	Dense deciduous forests	27	Mining or landfills
8	Dense forests of sclerophylls	28	Areas in transformation
9	Scrub	29	Road network
10	Coniferous forests	30	Urban bare floor
11	Deciduous forests	31	Airport areas
12	Forests of sclerophylls	32	Rail network
13	Meadows and grasslands		**Water mass**
14	Shore forest	33	Reservoirs
15	Bare forest soil	34	Lakes and lagoons
16	Burned areas	35	Water courses
17	Rocky	36	Rafts
18	Wet areas	37	Artificial channels
		38	Sea

4.2. Evaluation Metrics

We trained the UNET model with 10 channels from Sentinel-2 as input for 250 epochs with a batch size of 64 images for training and 32 images for validation. After predicting the segmentation for all territory in a test region, we obtain the confusion matrix shown in Figure 6. Most of the confusions are among classes of the same four super-classes. Next, we froze all UNET weights, connected the ConfidNet, and trained it for 20 epochs with a batch size of 64 images.

To evaluate the results, we calculate the global accuracy concerning the fraction of pixels sent to the photo interpreters (Figure 7). To do so, we calculate how many pixels are in each range of uncertainty and how global accuracy improves as long as we send more pixels to the photo interpreters. In our case, sending half of the pixels to the photo interpreters assures a global accuracy of 90%. To simulate the experts' labeling, we assume that the labels reviewed are correct (i.e., the same as the label in the ground truth). This assumption is made of the staff's inputs that work on land covers production as this is a collaborative research. The experiments presented are based on a database of 900 million pixels (surface of Catalonia, Spain). Considering the database we work with, it was out of the project's scope in which this paper is based on using humans to evaluate the performance. Human validation would require months of photointerpreters devoted to labeling the image's polygons, and even incorporating the expert's variability would not provide an exact measure for the method. Thus, we decided to validate the technique based on a data set verified by quality controls, and therefore it can be considered a good approximation to reality.

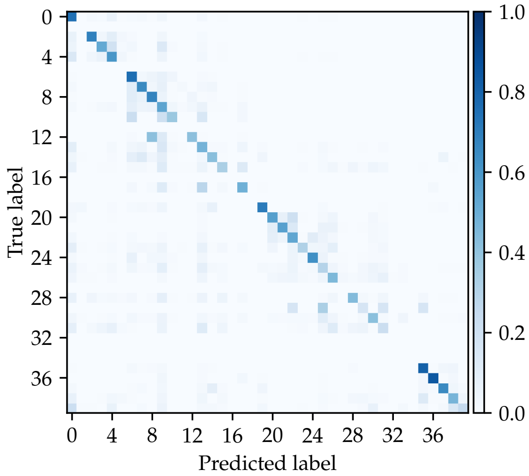

Figure 6. Confusion matrix of the validation set for the ICGC dataset, normalized over the true labels.

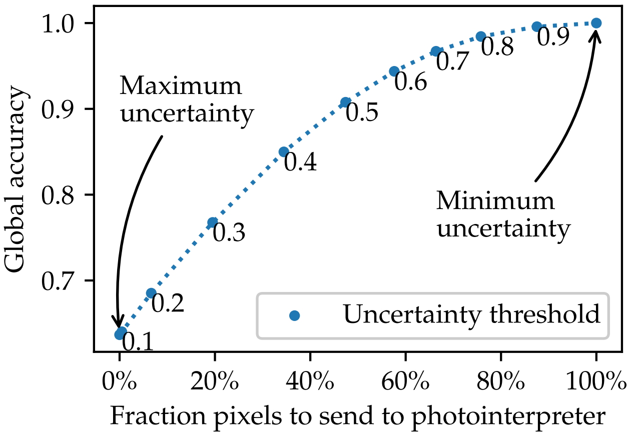

Figure 7. Global accuracy evolution depending on the number of pixels sent to the photointerpreter.

Figure 8 provides an example with all inputs and outputs, where we can see the satellite image, the ground truth, the prediction, the uncertainty by the CondidNet and the softmax output (where red color indicates high uncertainty). If we compare the ConfidNet uncertainty with the softmax output of the UNET network for each class in a qualitative manner, the ConfidNet provides a more detailed uncertainty illustration. Softmax generates a blurred image that makes it harder to differentiate uncertain zones. Paying attention to the outputs of the ConfidNet, a clear example of incorrect classification is the road at the top of the image, which appears as highly uncertain. This is the kind of zone that will be sent to the photo interpreters to be corrected, incorporating a human-in-the-loop approach.

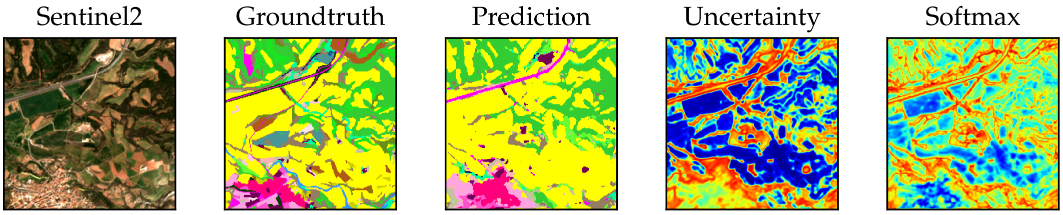

Figure 8. Uncertainty and softmax qualitative comparison.

Figure 9 reports the accuracy evolution for each class in the ConfidNet and the softmax. The 39 classes are located in the axis Y, and the axis X indicates the number of pixels sent to the photo interpreters. The color bar indicates the accuracy for each class; dark means low accuracy and light high accuracy. To obtain similar accuracy values in both cases, we have to send more pixels to the photointerpreter using softmax than using ConfidNet. For instance, in the case of class 35 more than 50% of total pixels are required to be sent for the softmax, while for ConfidNet a 40%.

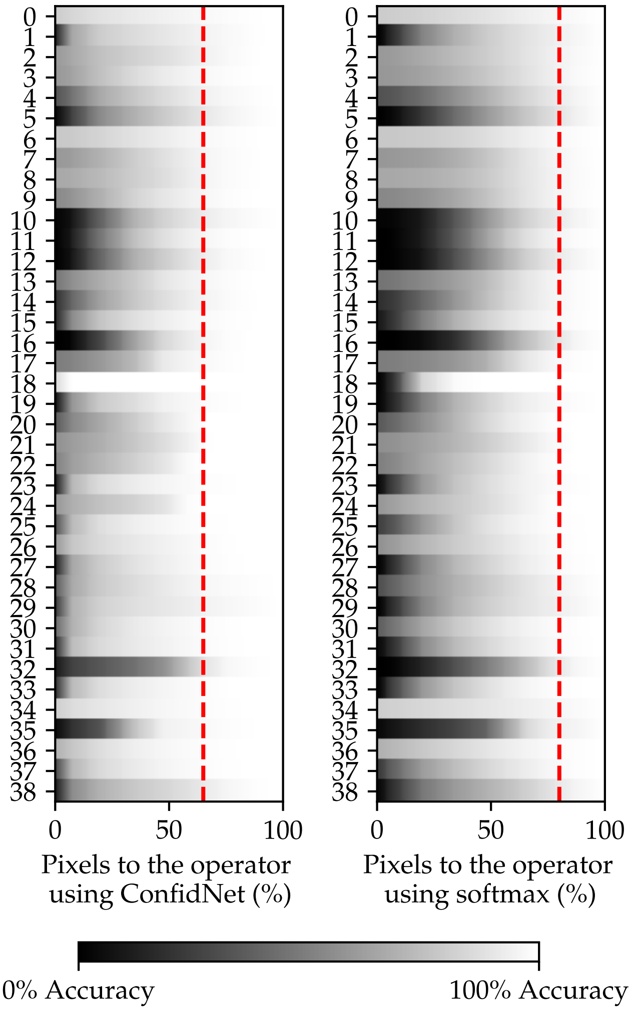

Figure 9. Uncertainty and softmax quantitative comparison per class. Red line indicates all classes with an accuracy of at least 90%.

4.3. Qualitative Results

To generate a more realistic product for the photointerpreters work, a polygon mask was created from uncertainty outputs (Figure 10). Please note that these outputs provide pixel-by-pixel information about the prediction quality, which would be very difficult to analyze by photo interpreters. For this reason, an external algorithm was applied to transform pixel-by-pixel information into polygons, since they are more useful in terms of interpretation. For this purpose, all pixels under a certain threshold were selected (those pixels with low prediction quality) and were grouped in different clusters, applying the DBSCAN clustering method [19], which takes into account a maximum distance between neighbor points and the minimum points required to build up a polygon.

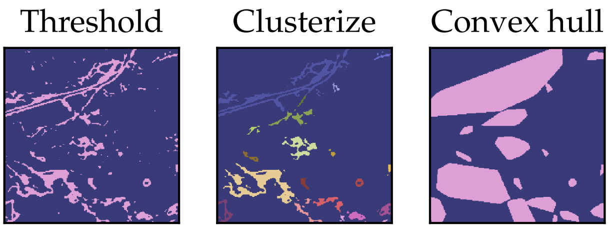

Figure 10. Creation of the polygon mask for the photo interpreters.

Our method makes the photo interpreters work easier by transforming isolated low-quality pixels into a set of polygons.

Overall, the complete human-in-the-loop process is illustrated in Figure 11. First, the original Sentinel-2 image. Next, the prediction made by the UNET neural network. In third place, the uncertainty output made by the ConfidNet and post-produced to obtain the convex clusters. Finally, the prediction corrected by a human photointerpreter considering uncertain regions lower than 20%. We also present the results when the global accuracy has to be of the order of 90% using an uncertainty threshold of 0.45, as can be seen in Figure 12.

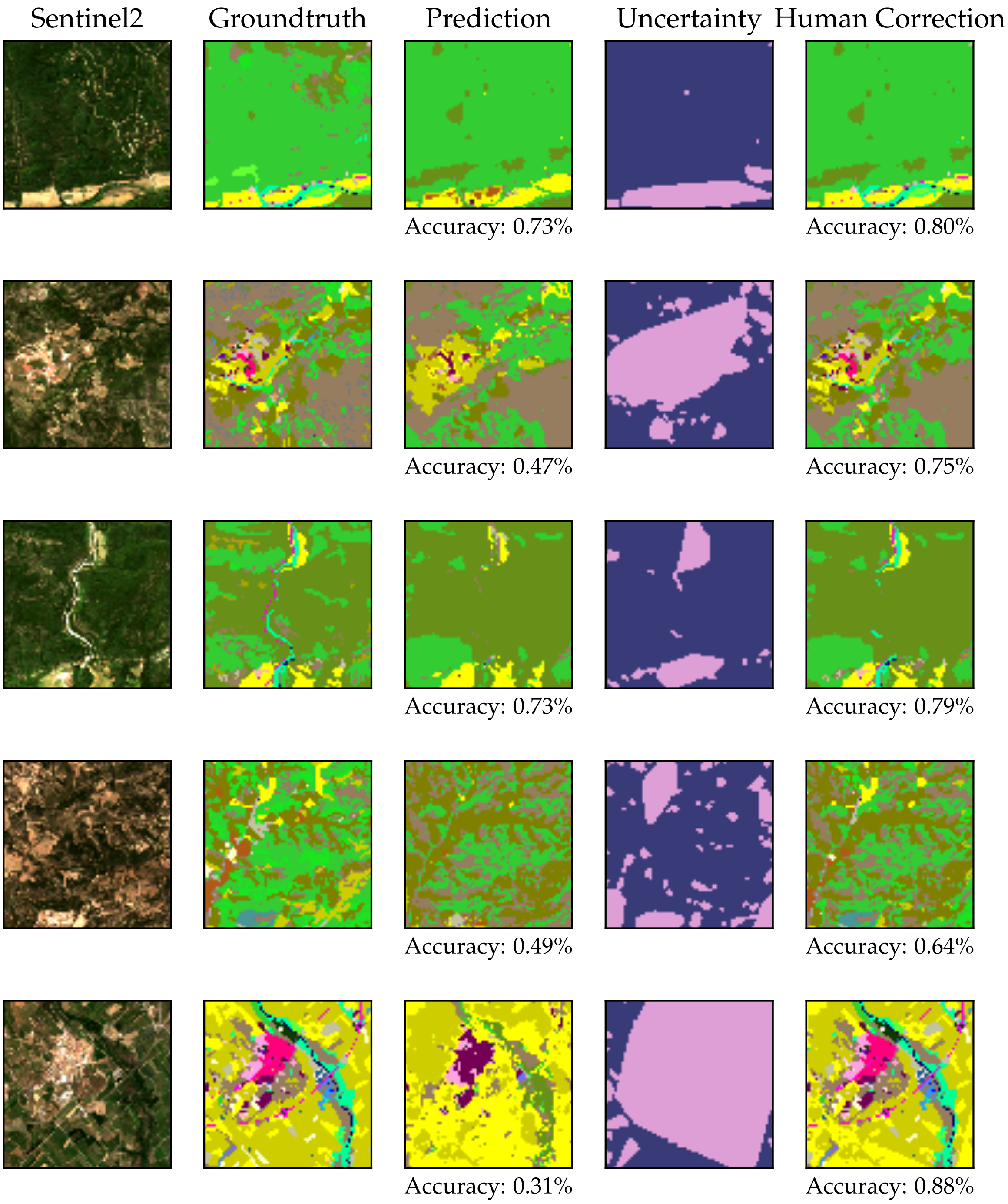

Figure 11. Some results of our human-in-the-loop methodology to segment satellite images. RGB bands represent Sentinel-2 images in the first column.

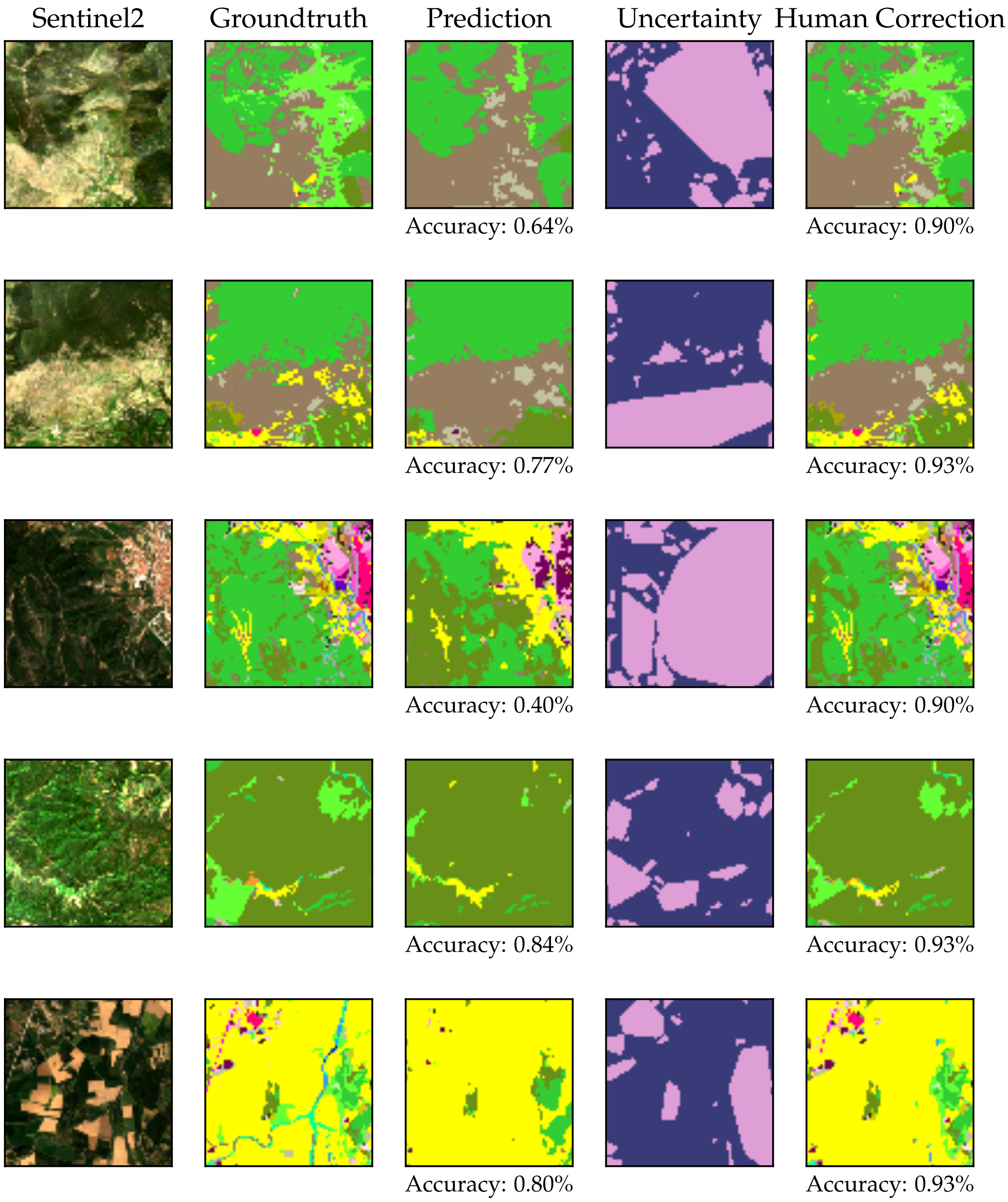

Figure 12. Our human-in-the-loop methodology assumes an uncertainty threshold of 0.45 to reach an accuracy of 90%.

5. Conclusions

We present a human-in-the-loop deep learning methodology to outfit the results in satellite cover land segmentation, aiming to improve current techniques that are not reaching acceptable results for real-world applications in large datasets. We propose to combine artificial intelligence and photo interpreters' work. Two neural networks in parallel are used: a UNET that predicts each pixel and a ConfidNet that outputs the prediction confidence. Following our methodology, areas with low confidence are sent to the photo interpreters to be reviewed and corrected if needed, and areas with high confidence will remain with the UNET prediction.

The proposed methodology reaches a 90% accuracy, reduces to half the human effort, and establishes a methodology for land cover segmentation. Our code will be opened to the scientific community.

Author Contributions: Conceptualization, C.G.R.; methodology, C.G.R., J.V. and O.M.; investigation, C.G.R., J.V. and O.M.; writing—original draft preparation, C.G.R.; writing—review and editing, C.G.R., J.V. and O.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by industrial doctorate grant 2018-DI-0045 of AGAUR between Universitat de Barcelona and Institut Cartogràfic i Geologic de Catalunya and RTI2018-095232-B-C21, 2017SGR1742 grants.

Acknowledgments: In this section you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

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

References

1. McRoberts, R.E.; Wendt, D.G.; Nelson, M.D.; Hansen, M.H. Using a land cover classification based on satellite imagery to improve the precision of forest inventory area estimates. *Remote Sens. Environ.* **2002**, *81*, 36–44.
2. Carlson, T.N.; Arthur, S.T. The impact of land use—Land cover changes due to urbanization on surface microclimate and hydrology: A satellite perspective. *Glob. Planet. Chang.* **2000**, *25*, 49–65.
3. Abou EL-Magd, I.; Tanton, T. Improvements in land use mapping for irrigated agriculture from satellite sensor data using a multi-stage maximum likelihood classification. *Int. J. Remote Sens.* **2003**, *24*, 4197–4206.
4. Haeberlin, Y.; Turberg, P.; Retière, A.; Senegas, O.; Parriaux, A. Validation of Spot-5 satellite imagery for geological hazard identification and risk assessment for landslides, mud and debris flows in Matagalpa, Nicaragua. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci* **2004**, *35*, B1.
5. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [PubMed]
6. Audebert, N.; Le Saux, B.; Lefevre, S. Deep Learning for Classification of Hyperspectral Data: A Comparative Review. *IEEE Geosci. Remote Sens. Mag.* **2019**, *7*, 159–173. [CrossRef]
7. Xia, G.S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981.
8. Marmanis, D.; Wegner, J.D.; Galliani, S.; Schindler, K.; Datcu, M.; Stilla, U. Semantic segmentation of aerial images with an ensemble of CNSS. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *3*, 473–480.
9. Fröhlich, B.; Bach, E.; Walde, I.; Hese, S.; Schmullius, C.; Denzler, J. Land cover classification of satellite images using contextual information. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2013**, *3*, W1.
10. Nalepa, J.; Myller, M.; Kawulok, M. Validating hyperspectral image segmentation. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1264–1268.
11. Schmitt, M.; Hughes, L.H.; Qiu, C.; Zhu, X.X. SEN12MS—A Curated Dataset of Georeferenced Multi-Spectral SENTINEL-1/2 Imagery for Deep Learning and Data Fusion. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2019**, *42*, 153–160. [CrossRef]
12. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
13. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
14. Kemker, R.; Salvaggio, C.; Kanan, C. High-resolution multispectral dataset for semantic segmentation. *arXiv* **2017**, arXiv:1703.01918.
15. Gómez, R. 2018. Understanding Categorical Cross-Entropy Loss, Binary Cross-Entropy Loss, Softmax Loss, Logistic Loss, Focal Loss and All Those Confusing Names. Available online: https://gombru.github.io/2018/05/23/cross_entropy_loss/ (accessed on 12 June 2020)
16. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
17. Corbière, C.; Thome, N.; Bar-Hen, A.; Cord, M.; Pérez, P. Addressing Failure Prediction by Learning Model Confidence. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019; pp. 2898–2909.

18. European Space Agency. Sen2Cor (2.8). 2019. Available online: https://step.esa.int/main/third-party-plugins-2/sen2cor/ (accessed on 23 March 2020)
19. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*; Kdd: Portland, Oregon, 1996, Volume 96, pp. 226–231.

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.