LEARNING FROM NOISY SAMPLES FOR MAN-MADE IMPERVIOUS SURFACE MAPPING

Chunping Qiu[1], Paolo Gamba[2], Michael Schmitt[1], Xiao Xiang Zhu[1,3,*]

[1]Signal Processing in Earth Observation, Technical University of Munich (TUM), Munich, Germany - (chunping.qiu, m.schmitt)@tum.de
[2]Telecommunications and Remote Sensing Laboratory, University of Pavia, Italy - paolo.gamba@unipv.it
[3]Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany - Xiaoxiang.Zhu@dlr.de

KEY WORDS: Classification, Fully convolutional networks (FCNs), Impervious surface mapping, Noisy samples, Robust loss function, Sentinel-2

ABSTRACT:

Man-made impervious surfaces, indicating the human footprint on Earth, are an environmental concern because it leads to a chain of events that modifies urban air and water resources. To better map man-made impervious surfaces in any region of interest (ROI), we propose a framework for learning to map impervious areas in any ROIs from Sentinel-2 images with noisy reference data, using a pre-trained fully convolutional network (FCN). The FCN is first trained with reference data only available in Europe, which is able to provide reasonable mapping results even in areas outside of Europe. The proposed framework, aiming to achieve an improvement over the preliminary predictions for a specific ROI, consists of two steps: noisy training data pre-processing and model fine-tuning with robust loss functions. The framework is validated over four test areas located in different continents with a measurable improvement over several baseline results. It has been shown that a better impervious mapping result can be achieved through a simple fine-tuning with noisy training data, and label updating through robust loss functions allows to further enhance the performances. In addition, by analyzing and comparing the mapping results to baselines, it can be highlighted that the improvement is mainly coming from a decreased omission error. This study can also provide insights for similar tasks, such as large-scale land cover/land use classification when accurate reference data is not available for training.

1. INTRODUCTION

The global man-made impervious surface, consisting of buildings, roads, and other man-made structures, is a crucial indicator of the human footprint on Earth. It provides a possibility for evidence-based decision making with respect to various applications and challenges such as climate change, disaster management, and sustainable development (Pesaresi et al., 2016). Like any other classification tasks, impervious area mapping can be performed using deep learning (DL) approaches, that have proved to be very powerful tools (Zhu et al., 2017, Lang et al., 2019).

Despite the success of deep neural networks in the remote sensing (RS) field for various supervised learning tasks such as land cover classification and change detection, its superior performance highly depends on the availability of massive training data with accurate annotations. Without them, the performance of DL would inevitably suffer because deep neural networks can overfit to the noise in the training data (Zhang et al., 2016). Even though DL is shown to be robust to non-adversarial label noise, the required amount of clean data increases as there are more noisy labels (Rolnick et al., 2017). This aspect is crucial for RS, because collecting reliable training labels for extended areas or even on a global scale is a costly and error-prone task. On the other hand, there is a large amount of geospatial data products available from previous efforts. In the case of built-up area/human settlement/impervious area mapping, examples include the Global Urban Footprint (GUF) (Esch et al., 2012, Esch et al., 2013), the Global Human Settlement Layer (GHSL) (Corbane et al., 2017), the GlobeLand30 land cover map (Chen et al., 2017), and the Global Human Built-up And Settlement Extent (HBASE) Dataset (Wang et al., 2017). To fully exploit such datasets for training, however, one has to take into account the errors they, as predictions of machine learning approaches, may contain, due to temporal gaps or inaccuracy in the original processing chain. In addition, the classes in the reference data might have a slightly different definition from those to be considered for the task at hand. Therefore, how to robustly learn a superior model from potentially noisy reference data is a problem of great importance, especially in deep learning applied to remote sensing.

The idea of learning from noisy samples is based on the desire to better exploit the reliable samples of the training set, while being less impacted by the unreliable ones. For this purpose, it is critical to distinguish the reliable or clean samples from the others (Chen et al., 2019, Han et al., 2018, Northcutt et al., 2017).

A first way to deal with this problem is resorting to different forms of regularization algorithms and avoiding the overfitting to noisy labels (Damodaran et al., 2019, Han et al., 2018, Ma et al., 2018). Another approach is to explicitly or implicitly model the noise using a noise transition matrix that is either based on prior knowledge or learned using e.g. directed graphical models or conditional random fields (Patrini et al., 2017, Sukhbaatar, Fergus, 2014). A third idea is to use noise-tolerant loss functions such as mean square error and mean absolute error, which theoretically guarantee a good results according to statistical learning principles (Zhang, Sabuncu, 2018, Reed et al., 2014). In a different way, inaccurate labels can be also corrected and updated during the iterative training process with a bootstrapping scheme. This can be carried out either by alterna-

*Corresponding author

tively updating network parameters and correcting labels during the training process (Reed et al., 2014, Tanaka et al., 2018) or by means of an independent prediction step based on selected reliable samples.

Inspired by the related work about learning from noisy samples, in this study, we propose a framework to exploit noisy training samples from existing maps (called reference data from hereon in order to distinguish it from actual ground truth) to improve baseline mapping results achieved by means of attention-based FCNs. In this framework, robust loss functions for fine-tuning of the pre-trained models are used so that the employed original reference data labels can be modified during the fine-tuning of the pre-trained model. This way, the adverse effect of incorrect labels can be alleviated.

The remainder of this study is organized as follows. Section 2 contains a description of our proposed framework including the general idea and details of the specific implementation used in this study. In Section 3, we introduce the experimental design as well as the used data and experimental setup. We compare our achieved results from different approaches with each other and with the state-of-the-art in Section 3.2 and 3.3. An interpretation and discussion of the outcome of our experiments is also presented along with the results. Finally, Section 4 summarizes the main findings and contributions of this study.

2. A FRAMEWORK FOR LEARNING FROM NOISY REFERENCE DATA

The proposed framework is illustrated in Fig. 1. The focus of this study is highlighted by colorization. The main part is to improve the preliminary mapping results of specific areas via fine-tuning of the pre-trained model, where we propose to consider using robust loss functions because the employed reference data is not absolutely accurate with respect to our target task. Instead of using the default loss function for classification problems, i.e. binary cross-entropy, we propose to use two kinds of robust loss functions, which are able to consider noises within the reference labels during training. To compare the performance of different approaches, we chose to map test areas that are far from the training sites and as a result might be subject to domain shift.

The network architecture for the first preliminary prediction and the subsequent finetuning, our choice and preparation of reference data used for fine-tuning, and the employed robust loss functions will be presented in detail in the following sections.

2.1 Attention-based FCN for MIS mapping

The adapted attention-based FCN-ResNet architecture is illustrated in Fig. 2. The main part is a ResNet-based FCN-8s, which is chosen to capture spectral and spatial local features by multi-scale feature fusion (Long et al., June 8–10, 2015). In addition to this ResNet-based FCN-8s, we employ attention modules: a position attention module (PAM), and a channel attention module (CAM). PAM and CAM together, in a parallel manner is called dual attention module. These attention modules, on the one hand, are similar to a "feature selection" process where salient features are being assigned larger weights. On the other hand, they are assumed to learn additional features by taking into account the long-range contextual information over both the channel dimension and the spatial dimension (Fu

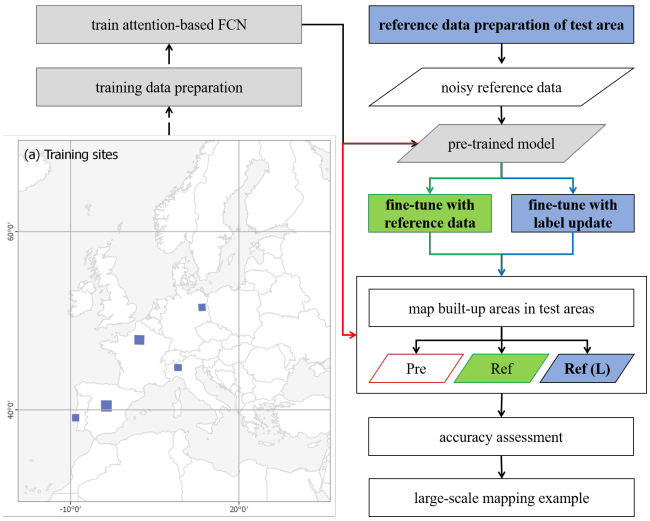

Figure 1. The framework for learning from noisy reference data for impervious area mapping, with the main content colorized.

et al., 2018). In this way, the feature representation can be further improved and enhanced. The sum of the output features of the modules is subsequently exploited for the prediction of the HSE. Additionally, this architecture is adapted for HSE mapping by outputting a final prediction of the half-size, instead of the same size as the input images, as illustrated in Fig. 2. This is realized by removing one up-sampling layer from the original FCN-8s. This architecture results in a HSE classification map with 20 m GSD, which is the same as the GSD of the used reference data during the preliminary training.

2.2 Reference data preparation for fine-tuning

When ground truth data collecting is not desired for training models, data from existing maps (the so called reference data) can be used. To this aim, in this study a combination of points extracted from the GUF map and from the one obtained by a pre-trained model is used. Specifically, if in one location the label is non-built-up according to GUF and non-urban according to the DL pre-trained model, that location will be added to the non-impervious training set. In all other cases, it will be added to the impervious training set. This decision is based on the observation that GUF is strong at detecting sparsely built-up areas in villages or suburban areas and the predictions from the DL model include roads and other impervious surfaces (Esch et al., 2017, Qiu et al., 2020).

Considering the errors in the original data, it is certain that there is some noise in the labels. Accordingly, the use of these data to refine DL predictions for a specific test area should be performed carefully. This is the reason why robust loss functions has been applied to the fine-tuning of the pre-trained models.

2.3 Model Fine-tuning via Robust Loss Functions

CNN training is based on updating the network weights to minimize a loss function that expresses the divergence between the model predictions and the reference labels. If the labels are noisy, the weights update can be sub-optimal, thus hindering model convergence or even worse leading to overfitting to the noisy input data. Loss functions that are robust against label noise are helpful because they rely less on the labels in the reference data. In this study, we propose to use two kinds of

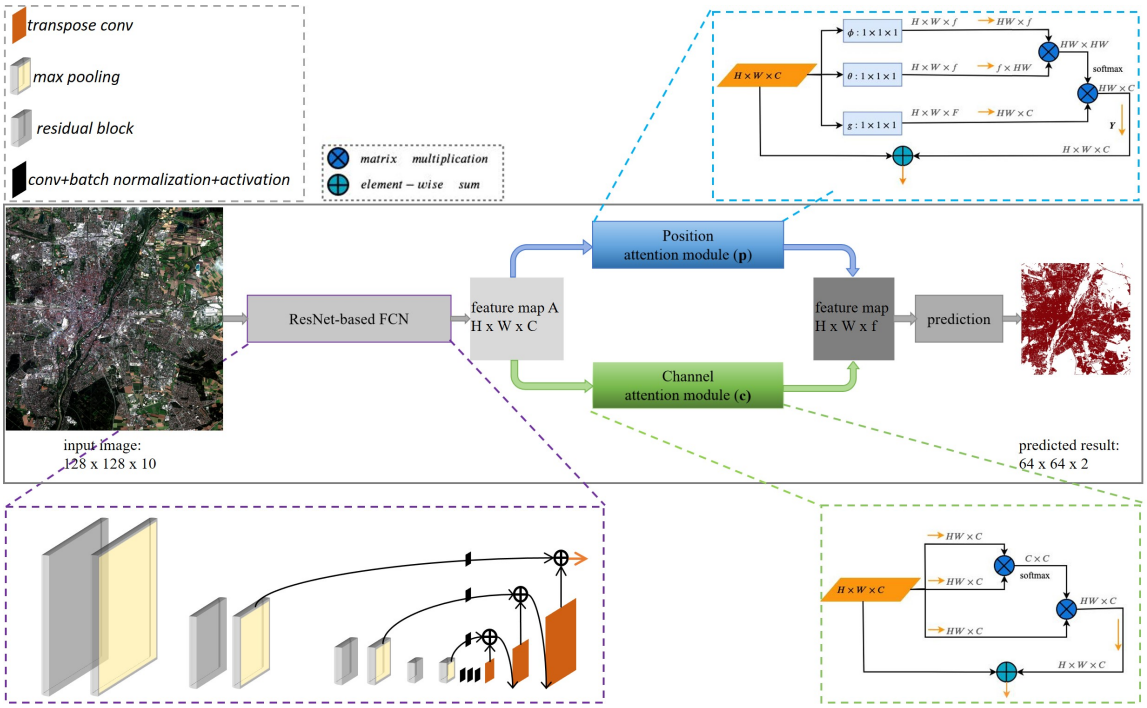

Figure 2. Attention-based FCN-ResNet architecture. "H", "W", and "C" denote height, width, and the channel number of the feature maps, respectively. The size of the final prediction is half of the input patch, which is decided based on the GSD of the used reference data during the preliminary training.

robust loss functions that are both modifications of the categorical cross-entropy (CCE) loss commonly used for classification. The original CCE loss is defined as

$$L_{\text{cce}} = -\sum_{k=1}^{K} y_k log(\hat{y}_k), \tag{1}$$

where y_k is the k-th element of the target label represented by a one-hot encoded vector. $\hat{y}_k$ is the k-th element of the predicted class probabilities, and K is the number of classes. The case of impervious area mapping is a binary classification problem, so the loss is simplified into binary cross-entropy (BCE):

$$L_{\text{bce}} = -(y_k log(\hat{y}_k) + (1 - y_k) log(1 - \hat{y}_k)) \tag{2}$$

Instead of directly using the original labels, the L_{soft} loss function dynamically changes the target labels based on the current state of the model:

$$L_{\text{soft}} = -\sum_{k=1}^{K} [\beta y_k + (1 - \beta)\hat{y}_k] log(\hat{y}_k) \quad \beta \in [0, 1] \tag{3}$$

where β is a parameter to be selected according to the confidence in the reference data. Specifically, instead of directly using the label in the reference data for loss calculation, L_{soft} updates the label by combing it with the current prediction from the model. In this way, both the predictions and the reference data are used for loss calculation, a procedure mentioned as soft bootstrapping in (Reed et al., 2014). As a result, the potentially noisy samples are less heavily relied during fine-tuning.

A second option is the L_q loss function, a generalization of CCE and mean absolute error (MAE) proposed in (Zhang, Sabuncu, 2018):

$$L_q = -\frac{1 - (\sum_{k=1}^{K} y_k \hat{y}_k)^q}{q} \quad q \in (0, 1] \tag{4}$$

where q is a hyper-parameter that controls how CCE and MAE are combined: L_q is equivalent to CCE when $q \to 0$, and becomes MAE when $q = 1$. The idea is to take advantage of the benefits of both CCE and MAE, as successfully shown in (Fonseca et al., 2019).

3. EXPERIMENTAL RESULTS AND DISCUSSION

To validate the proposed framework, a number of experiments with Sentinel-2 images have been performed.

3.1 Experimental Setup

The original reference data for the preliminary training comes from five European scenes, Berlin, Lisbon, Madrid, Milan, and Paris, as indicated in Fig. 1, which is with a GSD of 20 m (Langanke, 2016). The test areas have been selected across the world to better assess the potential of the framework. Specifically, in this study four sites were selected for test: Beijing, Jakarta, Nairobi, and Tehran. For each test scene, checking points for accuracy assessment (manually labeled grid-based checking point, MLGCPs) were prepared, as presented in Fig. 3. Only the Sentinel-2 bands with 10 and 20 meter spatial resolution were considered. More details of reference and Sentinel-2 data pre-processing, and MLGCPs can be found in the previous work (Qiu et al., 2020).

In the first stage, the input images and their corresponding reference labels (from the five European scenes) are used to train the network with the Nesterov Adam optimizer implementation of Keras (Chollet et al., 2015). We used a minibatch size of 8

Table 1. OA and Kappa values for different approaches in the four test cities. Boldface indicates there is an improvement over results from pre-trained classifiers (Pre), and red color indicates an improvement over the default fine-tuning approach (BCE).

approach			OA					Kappa				
			Beijing	Nairobi	Tehran	Jakarta	Mean	Beijing	Nairobi	Tehran	Jakarta	Mean
baseline		GUF	*82.1*	*85.5*	*88.0*	*81.7*	*84.3*	*0.64*	*0.70*	*0.76*	*0.60*	*0.68*
		pre	*86.3*	*84.2*	*89.5*	*86.9*	*86.7*	*0.73*	*0.68*	*0.79*	*0.71*	*0.73*
fine-tuning	*BCE*		*86.0*	***89.1***	*89.0*	*87.9*	***88.0***	*0.72*	***0.78***	*0.78*	***0.72***	***0.75***
	L_soft	0.1	86.0	**89.1**	**90.9**	**87.5**	**88.4**	0.72	***0.78***	**0.82**	0.71	**0.76**
		0.3	**87.1**	**87.9**	**90.0**	**87.7**	**88.2**	**0.74**	**0.75**	**0.80**	**0.72**	**0.75**
		0.5	**87.4**	**88.5**	**90.0**	**87.1**	**88.2**	**0.75**	**0.77**	**0.80**	0.70	**0.75**
		0.7	**87.4**	**91.5**	**90.4**	**87.9**	**89.3**	**0.75**	**0.83**	**0.81**	**0.72**	**0.78**
	L_q	0.3	86.3	**90.9**	**90.0**	**87.5**	**88.7**	0.73	**0.82**	**0.80**	0.71	**0.76**
		0.5	86.3	**90.9**	**90.4**	86.5	**88.5**	0.73	**0.82**	**0.81**	0.69	**0.76**
		0.7	86.3	**89.7**	**89.5**	86.3	**87.9**	0.73	**0.79**	0.79	0.68	**0.75**
		0.9	**87.1**	**89.1**	**90.4**	85.4	**88.0**	**0.74**	**0.78**	**0.81**	0.66	**0.75**

Table 2. Omission and commission error percentages for different approaches in the four test cities. Boldface indicates there is an improvement over results from pre-trained classifiers (Pre), and red color indicates an improvement over the default fine-tuning approach (BCE).

approach			omission error					commission error				
			Beijing	Nairobi	Tehran	Jakarta	Mean	Beijing	Nairobi	Tehran	Jakarta	Mean
baseline		GUF	*22.1*	*28.2*	*16.8*	*16.0*	*20.8*	*16.3*	*3.4*	*6.0*	*12.7*	*9.6*
		pre	*11.0*	*29.5*	*7.1*	*9.6*	*14.3*	*16.4*	*5.2*	*11.8*	*10.5*	*11.0*
fine-tuning	BCE		***8.7***	***17.9***	***4.4***	***4.2***	***8.8***	*18.2*	*5.9*	*14.3*	*13.1*	*12.9*
	L_soft	0.1	**9.9**	**16.7**	**4.4**	**4.8**	**8.9**	17.6	*7.1*	**11.5**	13.2	12.3
		0.3	**8.7**	**21.8**	**4.4**	**4.5**	**9.9**	16.5	**4.7**	12.9	13.1	11.8
		0.5	**8.7**	**17.9**	**3.5**	**4.8**	**8.8**	**16.0**	7.2	13.5	13.7	12.6
		0.7	**7.0**	**12.8**	**2.7**	**4.5**	**6.7**	17.1	5.6	13.4	12.9	12.2
	L_q	0.3	**8.7**	**15.4**	**4.4**	**5.1**	**8.4**	17.8	**4.3**	12.9	12.9	12.0
		0.5	**8.7**	**15.4**	**3.5**	**5.1**	**8.2**	17.8	**4.3**	12.8	14.2	12.3
		0.7	**9.3**	**16.7**	**4.4**	**5.8**	**9.0**	17.5	5.8	13.6	14.0	12.7
		0.9	**8.7**	**16.7**	**3.5**	**5.4**	**8.6**	16.5	7.1	12.8	15.2	12.9

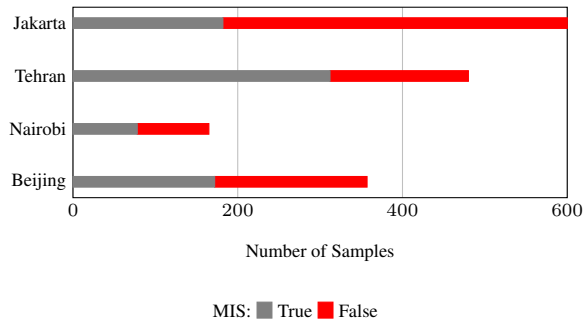

Figure 3. Number of MLGCPs for MIS mapping assessment.

images. The learning rate is 2×10^{-3}. To control the training time and avoid overfitting, early stopping was used, and the monitored metric is the validation loss with patience of 10 epochs. After getting the pre-trained model from the first stage, a preliminary prediction can be obtained by feeding the images to the model. Subsequently, fine-tuning is carried out for each ROI for 10 epochs. With the fine-tuned models, the final predictions can be obtained.

3.2 Accuracy Assessment

The accuracy of the mapping results is presented in Tab. 1 and 2, and is based on the independent MLGCPs. To show the improvement from this study, we also show some baseline maps, including the GUF layer, the maps obtained by DL, i.e., without fine-tuning, as well as the results after fine-tuning without considering the sample noise. Four parameters are used for both L_{soft} loss (β) and L_q (q) loss, respectively. Tab. 1 and 2 show an improvement from the default fine-tuning (using BCE), with the average Kappa value increasing from 0.73 to 0.75, and the average Omission error decreasing from 14.3% to 8.8%. Additionally, there is a further improvement from fine-tuning to fine-tuning with robust loss functions, with the average Kappa value increasing to 0.78, and the average Omission error decreasing to 6.7%. Finally, within the employed robust loss functions, L_{soft} is better than L_q in general, and L_{soft} with $\beta = 0.7$ provides the best results. Accordingly, the best setup has been used in all the following reported tests.

Please note that all these improvements are consistent among the four test areas. By comparing OA and Kappa values, as well as omission and commission errors, it is clear that this improvement is mainly coming from a decreased omission error while the commission error may increase. One possible explanation is that the above mentioned pre-processing procedure is prone to include more errors due to commission. As a result, areas containing even a small proportion of buildings or man-made structures tend to be recognized as impervious areas after the fine-tuning process.

3.3 Discussion

To better understand the results, in this section we visualize the achieved improvements by means of colored maps. We first

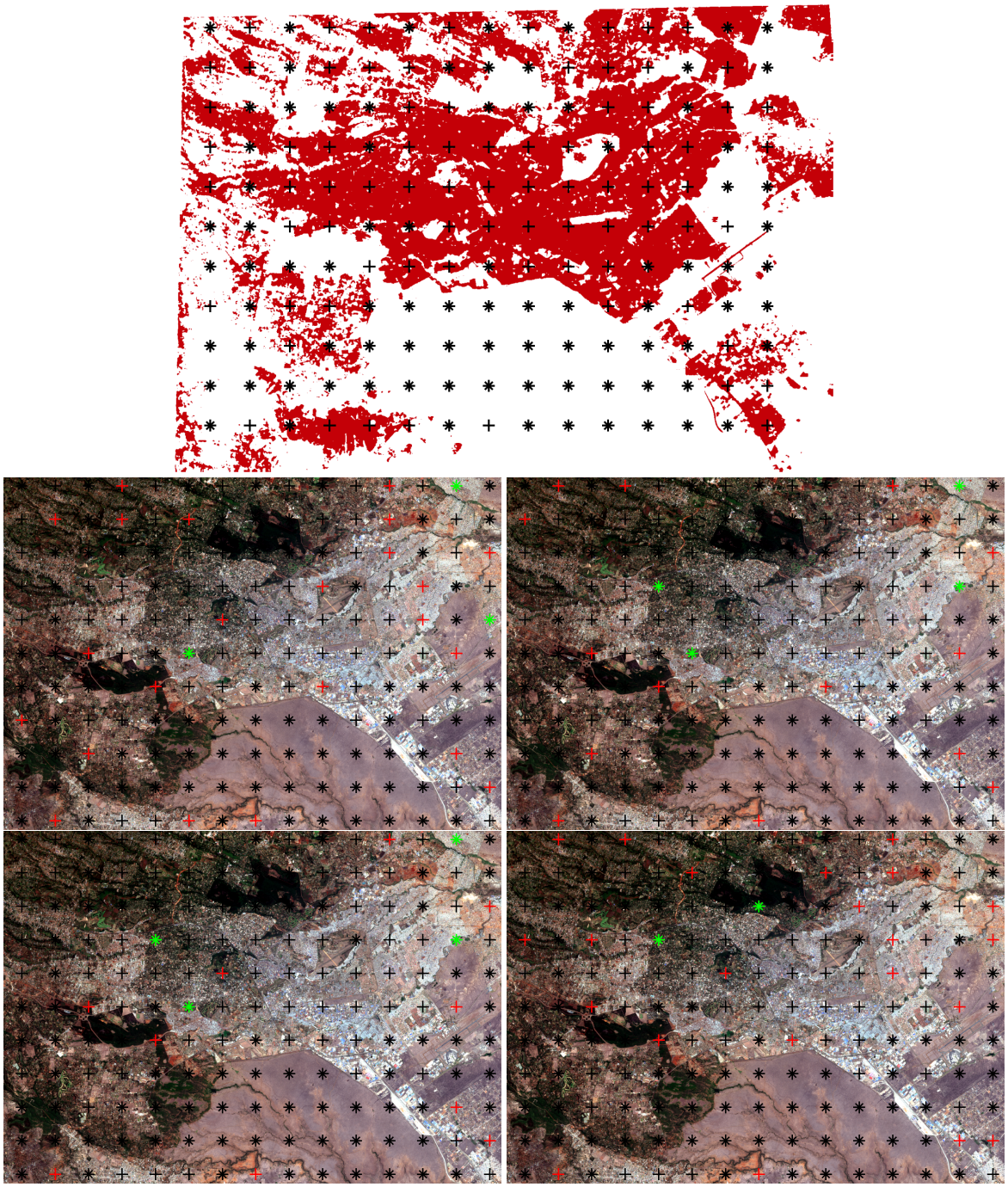

Figure 4. Comparison of mapping results to baseline maps using MLGCPs in the city of Nairobi. Top: final mapping result overlaid by MLGCPs: $*$ pervious, $+$ impervious. Bottom: Sentinel-2 images overlaid with MLGCPs, where colored ones indicate mistakes, either false positives (green) or false negatives (red). Specifically, top-left is based on the results before refinement, top-right based on those after refinement with CCE, bottom-left based on those after refinement with robust loss function, and bottom-right based on the GUF layer.

compare the results to baseline maps looking at the classification of the MLGCPs, and then present a detailed comparison of mapping results on a city scale as well as in zoomed-in areas.

3.3.1 Analysis of the causes for omission and commission errors Figure 4 presents the comparisons with test areas in Nairobi as an example, representing pervious points as "$*$" and impervious points as "$+$". The mapping result after fine-tuning overlaid with MLGCPs is first presented on the top, followed by Sentinel-2 images overlaid with MLGCPs, where false positives and false negatives, corresponding to commission and omission errors of different approaches, are colored in green and red, respectively.

It is clear from Fig. 4 that the main urban areas as well as some big roads are correctly mapped. By comparing with the results from baseline approaches, the decreased omission error is visualized by the decreased number of red crosses. Additionally, it can be seen that the remaining mistakes of the fine-tuning approach, both false positives and false negatives, are in areas

such as roads, sparsely built-up areas such as small villages, as well as at the urban-rural fringe. In all these areas it is difficult to discriminate between built-up areas and non built-up areas using images with 10 meter spatial resolution. This can also be partly explained from the fact that the remaining false negatives (omission errors, red crosses), are missing in the GUF layer, too (Esch et al., 2017). We need to mention that the roads and other non-vertical impervious surfaces are not included in the GUF layer.

3.3.2 Spatial analysis of the improvements Figure 5 visualizes the difference between the final result of the procedure presented in this work with respect to some baseline maps, i.e. the results before fine-tuning (Pre), the GUF layer, and the used reference (Ref), using the whole city of Nairobi as an example. The figure shows additionally mapped areas in green as well as those areas that are eventually removed in red. While most of the map remains the same, there is a clear difference between the mapping results after robust-loss-based fine-tuning and those from other approaches. Compared to Pre, there are clearly more impervious areas, which is consistent with the decreased omission error observed in Tab.2. Compared to GUF, some roads are added as impervious areas, which again is expected according to the experimental setup as well as the definitions of the GUF layer and the mapped impervious surface in this study. Compared to the chosen reference, there is mainly a removal of impervious areas, which is expected as pervious areas might be mistaken as impervious ones in the reference preparation procedure, and these mistakes are corrected during fine-tuning, even using robust loss functions.

A more detailed analysis is provided in Fig. 6, where three zoomed-in areas for Beijing, Jakarta, and Nairobi, respectively, are shown. Some correctly removed and added areas in the mapping results with respect to Pre, GUF, as well as Ref are clearly visible.

Finally, Figure 7 compares the mapped impervious areas obtained by means of different approaches as well as in the original baseline maps in subset areas of Nairobi. In line with previous observations, the improvement is mainly resulting from a decreased omission error. Possible explanations for the remaining mapping mistakes, mainly due to commission errors, include the small data size for the model fine-tuning, as well as the about pixel-level geo-location accuracy of the Sentinel-2 images (Drusch et al., 2012).

4. CONCLUSIONS AND OUTLOOK

MIS mapping provide key information in support of the development of local governments and world-wide collaborations to address issues such as climate change and air pollution. To better map man-made impervious surfaces on a large scale without relying on massive amounts of training data, and more specifically to decrease the omission errors for impervious area mapping, this paper proposes a framework to exploit possibly noisy reference data that are already globally available. The main idea is fine-tuning pre-trained DL models using available reference data in a specific ROI. To this aim, robust loss functions are used to mitigate the effect of potential errors in the reference data. This framework is validated using four test areas across the world, and improvements over baseline results have been obtained. Future research effort is aimed at an improved verson of the proposed framework for a better understanding and practical applications, e.g., by further investigating into the influences of different FCN architectures and comprehensively testing the potential of this approach in more test sites.

Figure 5. Comparison of the produced results in Nairobi to three kinds of baseline maps: Pre (top), GUF (middle), and Ref (bottom), respectively.

REFERENCES

Chen, J., Cao, X., Peng, S., Ren, H., 2017. Analysis and applications of GlobeLand30: a review. *ISPRS International Journal of Geo-Information*, 6(8), 230.

Chen, P., Liao, B., Chen, G., Zhang, S., 2019. Under-

Figure 6. Comparison of the produced result to three baseline maps in three subsets of Beijing, Nairobi, and Jakarta, respectively. The color legend is the same as in Fig. 5.

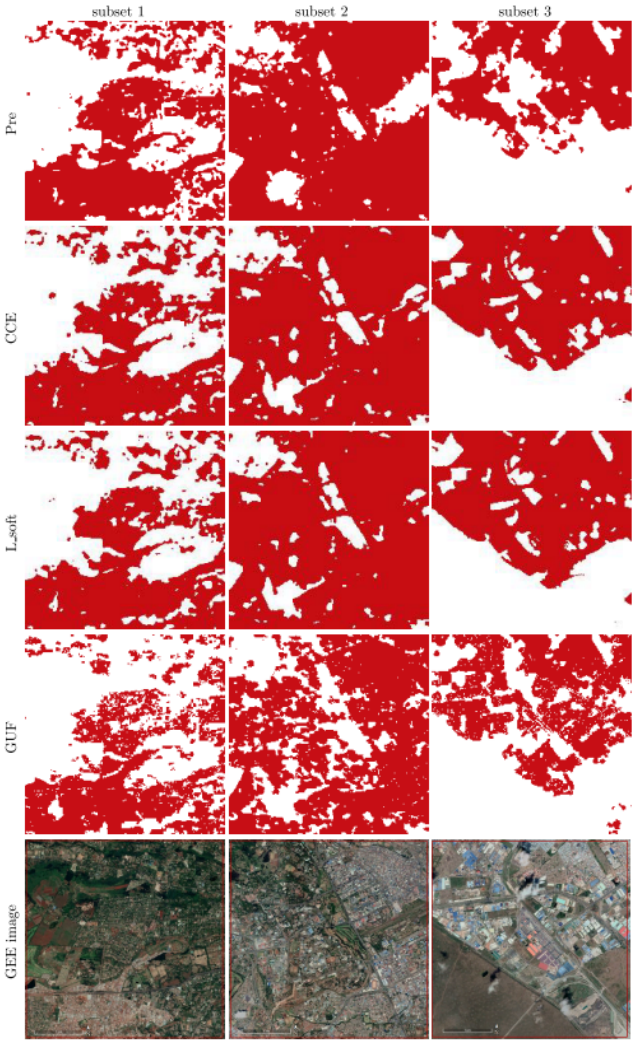

Figure 7. Comparison of the produced results with different methods with baseline maps in Nairobi.

standing and Utilizing Deep Neural Networks Trained with Noisy Labels. *arXiv preprint arXiv:1905.05040*.

Chollet, F. et al., 2015. Keras. https://keras.io.

Corbane, C., Pesaresi, M., Politis, P., Syrris, V., Florczyk, A. J., Soille, P., Maffenini, L., Burger, A., Vasilev, V., Rodriguez, D. et al., 2017. Big earth data analytics on Sentinel-1 and Landsat imagery in support to global human settlements mapping. *Big Earth Data*, 1(1-2), 118–144.

Damodaran, B. B., Fatras, K., Lobry, S., Flamary, R., Tuia, D., Courty, N., 2019. Pushing the right boundaries matters! Wasserstein Adversarial Training for Label Noise. *arXiv preprint arXiv:1904.03936*.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P. et al., 2012. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote sensing of Environment*, 120, 25–36.

Esch, T., Heldens, W., Hirner, A., Keil, M., Marconcini, M., Roth, A., Zeidler, J., Dech, S., Strano, E., 2017. Breaking new ground in mapping human settlements from space–The Global Urban Footprint. *ISPRS Journal of Photogrammetry and Remote Sensing*, 134, 30–42.

Esch, T., Marconcini, M., Felbier, A., Roth, A., Heldens, W., Huber, M., Schwinger, M., Taubenböck, H., Müller, A., Dech, S., 2013. Urban footprint processorFully automated processing chain generating settlement masks from global data of the TanDEM-X mission. *IEEE Geoscience and Remote Sensing Letters*, 10(6), 1617–1621.

Esch, T., Taubenböck, H., Roth, A., Heldens, W., Felbier, A., Schmidt, M., Mueller, A. A., Thiel, M., Dech, S. W., 2012. TanDEM-X mission-new perspectives for the inventory and monitoring of global settlement patterns. *Journal of Applied Remote Sensing*, 6(1), 061702.

Fonseca, E., Plakal, M., Ellis, D. P., Font, F., Favory, X., Serra, X., 2019. Learning sound event classifiers from web audio with noisy labels. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 21–25.

Fu, J., Liu, J., Tian, H., Fang, Z., Lu, H., 2018. Dual attention network for scene segmentation. *arXiv preprint arXiv:1809.02983*.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M., 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 8527–8537.

Lang, N., Schindler, K., Wegner, J. D., 2019. Country-wide

high-resolution vegetation height mapping with Sentinel-2. *arXiv preprint arXiv:1904.13270*.

Langanke, T., 2016. Copernicus Land Monitoring Service High Resolution Layer Imperviousness: Product Specifications Document. *Copernicus team at EEA*.

Long, J., Shelhamer, E., Darrell, T., June 8–10, 2015. Fully convolutional networks for semantic segmentation. *Proc. IEEE conference on computer vision and pattern recognition*, Boston, Massachusetts, 3431–3440.

Ma, X., Wang, Y., Houle, M. E., Zhou, S., Erfani, S. M., Xia, S.-T., Wijewickrema, S., Bailey, J., 2018. Dimensionality-driven learning with noisy labels. *arXiv preprint arXiv:1806.02612*.

Northcutt, C. G., Wu, T., Chuang, I. L., 2017. Learning with confident examples: Rank pruning for robust classification with noisy labels. *arXiv preprint arXiv:1705.01936*.

Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., Qu, L., 2017. Making deep neural networks robust to label noise: A loss correction approach. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1944–1952.

Pesaresi, M., Ehrlich, D., Ferri, S., Florczyk, A., Freire, S., Halkia, M., Julea, A., Kemper, T., Soille, P., Syrris, V., 2016. Operating procedure for the production of the Global Human Settlement Layer from Landsat data of the epochs 1975, 1990, 2000, and 2014. *Publications Office of the European Union*.

Qiu, C., Schmitt, M., Gei, C., Chen, T.-H. K., Zhu, X. X., 2020. A framework for large-scale mapping of human settlement extent from Sentinel-2 images via fully convolutional neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 163, 152 - 170.

Reed, S., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., Rabinovich, A., 2014. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*.

Rolnick, D., Veit, A., Belongie, S., Shavit, N., 2017. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*.

Sukhbaatar, S., Fergus, R., 2014. Learning from noisy labels with deep neural networks. *arXiv preprint arXiv:1406.2080*, 2(3), 4.

Tanaka, D., Ikami, D., Yamasaki, T., Aizawa, K., 2018. Joint optimization framework for learning with noisy labels. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5552–5560.

Wang, P., Huang, C., Brown de Colstoun, E. C., Tilton, J. C., Tan, B., 2017. Documentation for the Global Human Built-up And Settlement Extent (HBASE) Dataset From Landsat. *Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC)*.

Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O., 2016. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.

Zhang, Z., Sabuncu, M., 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 8778–8788.

Zhu, X., Tuia, D., Mou, L., Xia, G., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: a comprehensive review and list of resources. *IEEE Geosci. Remote Sens. Mag.*, 5(4), 8–36.