What Data are needed for Semantic Segmentation in Earth Observation?

J. Castillo-Navarro, N. Audebert, A. Boulch, B. Le Saux
DTIS, ONERA, University Paris Saclay
F-91123 Palaiseau - France.
javiera.castillo_navarro@onera.fr, bertrand.le_saux@onera.fr

S. Lefèvre
Univ. Bretagne Sud, UMR 6074, IRISA
F-56000 Vannes, France
sebastien.lefevre@irisa.fr

***Abstract*—This paper explores different aspects of semantic segmentation of remote sensing data using deep neural networks. Learning with deep neural networks was revolutionized by the creation of ImageNet. Remote sensing benefited of these new techniques, however Earth Observation (EO) datasets remain small in comparison. In this work, we investigate how we can progress towards the ImageNet of remote sensing. In particular, two questions are addressed in this paper. First, how robust are existing supervised learning strategies with respect to data volume? Second, which properties are expected from a large-scale EO dataset? The main contributions of this work are: (i) a strong robustness analysis of existing supervised learning strategies with respect to remote sensing data, (ii) the introduction of a new, large-scale dataset named MiniFrance.**

***Index Terms*—Deep Learning, Supervised Learning, Semantic Segmentation, Land Use/Land Cover Mapping.**

I. Introduction

The ever-growing amount of remote sensing imagery data in the last two decades has allowed new developments in the fields of ecology, urban planning or natural disaster response. Those data are also more easily available, even openly as in the Copernicus or Landsat programs. However, data exploitation still remains a bottleneck. It requires human interprets, for example to identify tree species and study deforestation in a local ecosystem, or to find new buildings and measure growth of urban areas.

Thanks to the new deep learning methods developed for processing multimedia images in recent years, it is now possible to automate most of these tasks for Earth Observation (EO) data. Indeed, many state-of-the-art algorithms for object detection and image segmentation or classification [2], [13] have been successfully transferred to aerial and satellite images. It allows to produce quickly and without human intervention precise semantic maps, in both urban and rural contexts.

However, these learning algorithms rely heavily on the availability of annotated image databases. Even if collaborative cartographic resources such as OpenStreetMap can be used as annotations [8], these are restricted in terms of semantics (only roads, buildings, etc.). Thus, the question of quantifying the influence of existing datasets on the models we learn arises. Moreover, we need to define what it requires to make a good dataset for training EO data classification algorithms, and thus make a first step towards the ImageNet (a massive, annotated image dataset) of remote sensing.

Thus, the contributions of this paper are twofold: (i) An experimental analysis of the amount of data necessary to successfully achieve supervised learning, and moreover, of the required data variability; (ii) The constitution of a large-scale dataset of various remote sensing images and annotations from different sources, which overcomes limitations of standard reference datasets.

The rest of the paper is organized as follows. First we explain the problem statement and explore some related work in Sec. II. We then discuss the robustness of supervised learning in two cases: on current, small-scale datasets in Sec. III and at large-scale in Sec. IV. Finally, we conclude and draw perspectives in Sec. V.

II. Problem statement and related work

Convolutional Neural Networks (CNNs) have greatly benefited to computer vision and remote sensing for different tasks including image segmentation such as land use and land cover mapping. Furthermore, the pixel-wise predictions of Fully Convolutional Networks (FCN) and derivatives offer an appealing approach for Earth Observation data analysis. These models have reached the state-of-the-art for remote sensing data analysis including: land cover mapping in urban areas [10], [12], object segmentation [1] or semantic segmentation for multi-modal and multi-scale remote sensing data [2].

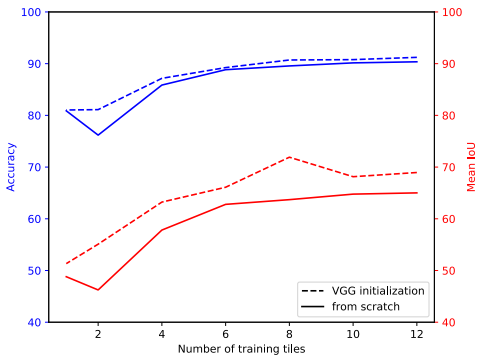
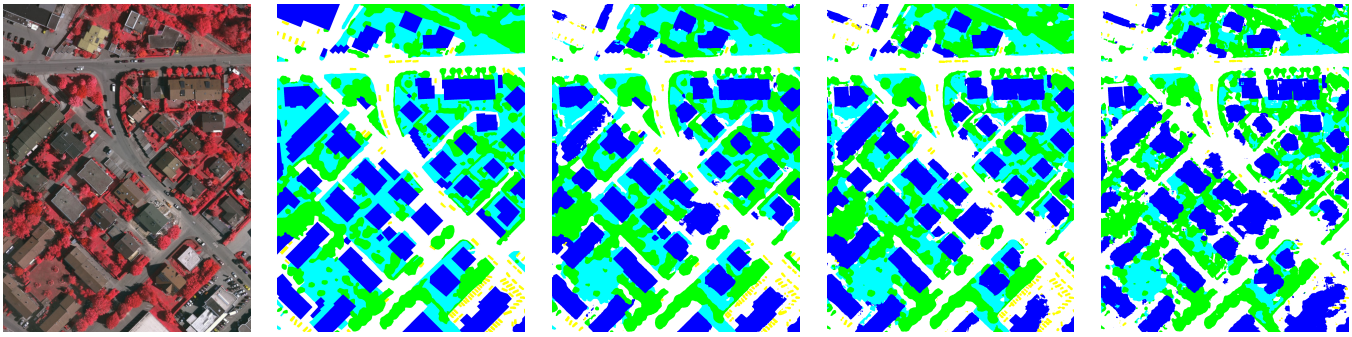


Fig. 1: Influence of the training set size (number of tiles) on the network performances, with and without weight initialization.

J. Castillo-Navarro's work is partially funded by a CNES grant.
N. Audebert's work is partially funded by ONERA-Total Naomi project.

978-1-7281-0009-8/19/$31.00 ©2019 IEEE



(a) Tile 30 (b) Ground truth (c) 10 tiles training (d) 6 tiles training (e) 1 tile training

Fig. 2: Visualization of results of the different training phases over the Vaihingen dataset.

Deep networks achieve top results on most of the public benchmark EO datasets [4], [8], [11], [14], with excellent performances over 80 to 90% accuracy. However, these methods are supervised, and still require a large amount of dense pixel-wise annotated data during the training phase. Most of the time, there exists only a small number of annotations on which to train models. In this context, we aim to answer the following question: how much data are needed to train a supervised neural network for semantic segmentation in Earth Observation?

This problem can be tackled in three ways: a first category tries to reduce the number of parameters of a model, while keeping data unchanged [9], a second one trains models with large, incomplete ground truth [7] and a third one is to focus on a small amount of well annotated data. In this paper, we consider the third option. In addition, applying a model trained on one dataset to another one yields in a serious drop of accuracy. This is known as the domain adaptation problem [16], and leads us to wonder what are the desired properties of a good training dataset.

To this aim, we set up several experiments to investigate the behaviour of network training with respect to the data. First, we aim at quantifying the amount of data required to obtain good performances in a standard set-up. Second, we renew the experiment with a more large-scale dataset to understand the relationship between training and genericity of the resulting models. In all experiments, we use as a baseline an efficient and versatile FCN: SegNet [2], [3]. It presents an encoder-decoder architecture, its encoder part is similar to VGG16 [15] and thus can be initialized with pre-trained weights. Training is achieved with Stochastic Gradient Descent and the standard cross-entropy loss function weighted over pixels.

III. Analysis of supervised learning on small-scale datasets

The ISPRS 2D Semantic Labeling [14] Vaihingen dataset consists of 33 infrared-red-green tiles with a spatial resolution of 9cm/px and an average size of 2000×1500px. Dense annotations are available on 16 tiles for 6 classes of interest: impervious surfaces, buildings, low vegetation, trees, cars and clutter. The associated benchmark being now closed, we perform experiments using 12 annotated tiles for training (*train*) and 4 tiles for validation (denoted by *val*).

In our first experiment, we aim to test the sensitivity of supervised learning to training. So we reduce the amount of annotated images used for training, from 12 tiles to only one, while *val* remains unchanged. Additionally, two different training settings are compared on the course of this experiment. First, we train our network by initializing the encoder with the pre-trained weights from VGG-16 on ImageNet. Secondly, we repeat the experiment, but weights are randomly initialized using the policy from [6]. This allows us to study if we can benefit from transfer learning. Results are presented in Fig. 1.

The outcomes of this experiment are somehow surprising. However, going from 90% to 80% of overall accuracy and 65% to 50% of mIoU when reducing the number of training tiles from 12 to 1 is much better than one would expect. Indeed, we supposed that reducing the number of training tiles would seriously impact the performance of the network. One possible reason is that all the images in the Vaihingen dataset are alike, thus, to generalize on them is a relatively easy task. On the other hand, the mean IoU decreases faster than the accuracy showing that one must be careful when interpreting these scores.

To better understand the quantitative scores from Fig. 1 in terms of segmentation quality, Fig. 2 shows the different predictions obtained for tile 30. We can observe that the quality of the segmentation map decreases notably when less annotated tiles are used during the training phase. It is interesting to note that there is not a considerable difference between training with 10 tiles and with 6 tiles, however there is a greater difference when training with 1 tile: borders are less smooth and little objects (such as cars) are not well learned, which explains why the mean IoU decreases faster on Fig. 1.

To assess the idea that the Vaihingen dataset has much redundancy, we observed its statistical distribution. In Fig. 4(a) and (b), we compare the color histograms over the 3 channels for *train* and *val*. Indeed, they are almost identical, which indicates that learning on a single location might not be so challenging. Actually, this is even promising in terms of practical business applications, since mapping an area can be achieved after labeling only a few images.

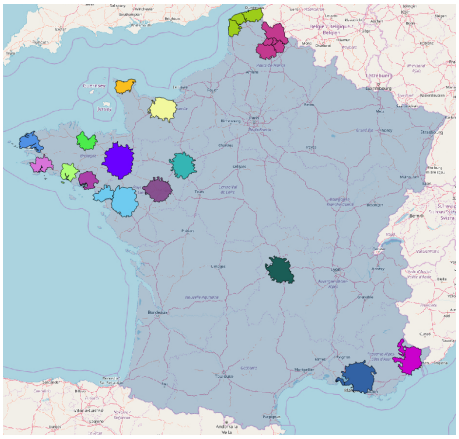


Fig. 3: MiniFrance dataset overview.

IV. Supervised learning at large-scale

The previous section stressed out a limitation of standard datasets for semantic mapping. If some already take into account multiple locations, they are devoted to a single class (such as buildings [8], [11]) or to land cover classes [16], but do not offer generic land use classes. Consequently, we introduce the MiniFrance dataset with the purpose of adding variety to the existing datasets for semantic segmentation.

MiniFrance consists of aerial images of 16 cities or conurbations from different regions in France (see Fig. 3). It is built using the 2 following data sources:

- Orthophotos from the BD ORTHO® free access database from the IGN (French National Mapping Agency) [1], acquired between 2012 and 2014. Each tile size is $10,000 \times 10,000$ pixels and the ground sampling distance of these images is 50cm/px.
- Labelled ground truth obtained from the Copernicus Urban Atlas 2012 database[2]. In this paper we focus on the second level of the semantic hierarchy, and 15 land use classes are considered, including urban fabric, industrial and transport units, arable lands, pastures, forests, open spaces, waterlands, and water.

We also propose a fixed partition for coherency of comparisons: 8 cities are used for training and the remaining 8 ones for test, keeping diversity in terms of architecture and urban design in both subsets. All in all, it is constituted of 2121 images, each of them of size $10,000 \times 10,000$ pixels. Therefore, MiniFrance dataset is 2719 times larger than Vaihingen.

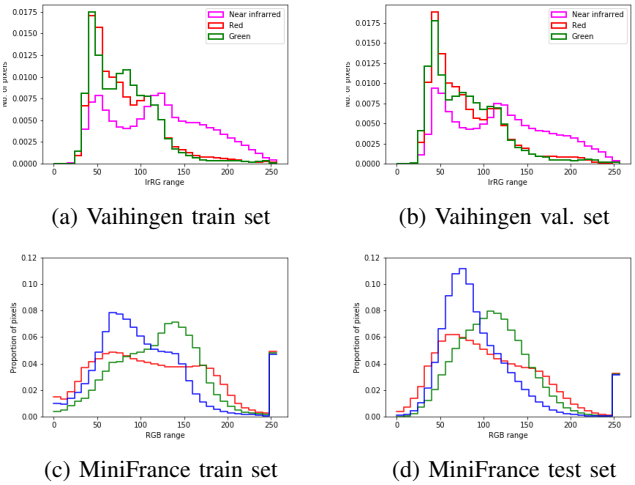
Similarly to Sec. III, we first test the influence of the amount of training data over the classification. However, due to computational times [3], we conduct more focused experiments. We train with the whole dataset, then only consider 10% of images on the dataset (we make sure to pick 10% of images from each conurbation to conserve the diversity of the dataset), and finally use only one city for training (the seaside town of Caen, which represents a similar amount of data: 12.5%). Test set remains the same. Following results of previous experiments (see Fig. 1), weights are initialized as the pre-trained weights for the VGG-16 network.

TABLE I: Classification performances with respect to amount of data.

Train set	*OA*	*mIoU*
100 %	52.40	15.79
10 %	50.14	15.25
Caen only ($\sim 12.5\%$)	42.09	10.05

Hence, results are shown in Table I [4]. Performances are not reaching the same level than on Vaihingen, which could be expected since working with this dataset is still at early stages and the land use classes are more abstract and difficult than land cover ones. However, considering our current issue, it is worth noting that training with all and 10% data leads to similar scores, both in accuracy and IoU. By picking our 10% sample images all over the dataset, we preserved the diversity of the training set and did not degrade the results too much (even if more data is better). On the contrary, training with a single location implies a 10% loss in accuracy and 5% less of mIoU. Clearly, the training dataset does not then offer enough variety to encompass all the potential images of the test set.

In a second experiment, we apply the model trained on the whole dataset to each city or conurbation of the test set. Results are shown in Table II. It is interesting to observe that the performance of the network varies notably between some conurbations, revealing differences between cities and a lack of generalization capacity from the model. Indeed, we observe in



(a) Vaihingen train set (b) Vaihingen val. set

(c) MiniFrance train set (d) MiniFrance test set

Fig. 4: Per channel color histograms over the Vaihingen and the MiniFrance datasets, comparing train and validation/test subsets.

[1] http://professionnels.ign.fr/bdortho

[2] https://land.copernicus.eu/local/urban-atlas/urban-atlas-2012/view

[3] Using a Titan X GPU, training over MiniFrance takes 40 hours, while testing takes 25 hours

[4] On this table and for the rest of the document, *OA* stands for Overall Accuracy and *mIoU* is Mean IoU.

Fig. 4 that the statistical distribution of the pixel colors differs from train to test. Thus MiniFrance is a much more diverse dataset than many others, and offers exciting challenges to overcome.

TABLE II: Results by conurbation, over the entire train set.

Score	Marseille	Rennes	Angers	Quimper	Vannes	Clermont	Lille	Cherbourg
OA	46.13	51.56	44.85	50.82	49.51	46.51	61.35	67.54
mIoU	12.77	15.05	13.15	13.93	12.66	11.40	16.93	15.82

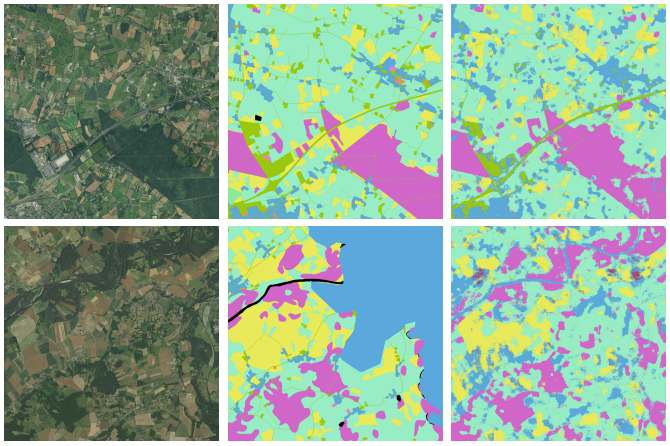


Fig. 5: MiniFrance semantic segmentation results. From left to right: RGB image, ground truth and prediction. Legend of main classes: Urban fabric , Industrial, commercial, public, military, private and transport units , Complex and mixed cultivation patterns , Herbaceous vegetation associations .

To better understand these numbers, Fig. 5 presents semantic maps obtained during testing. The first row shows a quite accurate prediction. Indeed, the model correctly identifies most classes present on the image. However, the predicted map appears more fragmented than the ground-truth, which shows that the network is sensitive to color variations of the image, and sometimes misses some abstract semantic classes. In the second row, we can observe an image where ground-truth is not reliable and for which the prediction obtained by our method seems actually more accurate.

V. Analysis and conclusion

Experiments in sections III and IV aimed at shedding some light on the relationship between data and semantic segmentation networks. On small-scale, single-location datasets such as Vaihingen, these networks are robust to a drastic decrease of the amount of training data when testing on similar data. This is promising as practical applications can be solved with less labeled data than what is usually thought, but meanwhile is limited because models are likely to overfit. Going large-scale raises new issues. Indeed, images are more diverse, corresponding to various geographical idiosyncrasies and various acquisition conditions and times. Also, large-scale annotations are often approximate, due to crowd-sourcing or semi-automated processes. Additionally, a new challenge is the higher semantic level of the classes with respect to pre-existing datasets which focus on structural classes (building, roads, etc.). Indeed, in crowdsourced maps such as OSM, classes are more symbolic and the classes are more land use oriented, related to the parcel and based on the interpretation of structural classes (e.g. building density in an area).

To promote research on large-scale EO analysis, we constituted MiniFrance, a new dataset for urban semantic segmentation. By using different cities for training and testing, we augment the variety of the dataset, which will require learning algorithms with good generalization capacities and the right bias/variance compromise [5]. MiniFrance raises new challenges including predicting high level semantics classes from approximate ground truth or structured learning. In addition, in this paper we also proposed a first FCN baseline for further experimental comparison.

References

[1] N. Audebert, B. Le Saux, and S. Lefèvre. Segment-before-Detect: Vehicle Detection and Classification through Semantic Segmentation of Aerial Images. *Remote Sensing*, 9(4):1–18, Apr. 2017.

[2] N. Audebert, B. Le Saux, and S. Lefevre. Beyond RGB: Very High Resolution Urban Remote Sensing With Multimodal Deep Networks. *ISPRS J. of Photogrammetry and Remote Sensing*, 140:20–32, 2018.

[3] V. Badrinarayanan, A. Kendall, and R. Cipolla. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. on Patt. Anal. and Mach. Intel.*, 39(12):2481–2495, 2017.

[4] M. Campos-Taberner et al. Processing of extremely high-resolution LiDAR and RGB data: Outcome of the 2015 IEEE GRSS Data Fusion Contest-Part A: 2-D Contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(12):5547–5559, 2016.

[5] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. *CoRR*, abs/1502.01852, 2015.

[7] L. Maggiolo, D. Marcos, G. Moser, and D. Tuia. Improving maps from CNNs trained with sparse, scribbled ground truths using fully connected CRFs. In *IGARSS*, 2018.

[8] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez. Can semantic labeling methods generalize to any city? the INRIA aerial image labeling benchmark. In *IGARSS*, 2017.

[9] D. Marcos, M. Volpi, B. Kellenberger, and D. Tuia. Land cover mapping at very high resolution with rotation equivariant cnns: Towards small yet accurate models. *ISPRS J. of Photogrammetry and Remote Sensing*, 145:96–107, 2018.

[10] D. Marmanis, J. D. Wegner, S. Galliani, K. Schindler, M. Datcu, and U. Stilla. Semantic segmentation of aerial images with an ensemble of CNNs. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 3:473, 2016.

[11] L. Mou et al. Multitemporal Very High Resolution from Space: Outcome of the 2016 IEEE GRSS Data Fusion Contest. *IEEE J. of Sel. Topics in Applied Earth Obs. and Remote Sensing*, 10(8):3435–3447, 2017.

[12] S. Paisitkriangkrai et al. Effective semantic pixel labelling with convolutional networks and conditional random fields. In *Proceedings of the IEEE CVPR Workshops*, pages 36–43, 2015.

[13] N. Rey, M. Volpi, S. Joost, and D. Tuia. Detecting animals in African Savanna with UAVs and the crowds. *Remote Sensing of Environment*, 200:341–351, 2017.

[14] F. Rottensteiner, G. Sohn, J. Jung, M. Gerke, C. Bailard, S. Benitez, and U. Breitkopf. The ISPRS benchmark on urban object classification and 3D building reconstruction. In *ISPRS Congress*, pages 293–298, 2012.

[15] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR International Conference on Learning Representations*, 2015.

[16] N. Yokoya et al. Open data for global multimodal land use classification: Outcome of the 2017 IEEE GRSS Data Fusion Contest. *IEEE J. of Sel. Top. in Applied Earth Obs. and Rem. Sens.*, 11(5):1363–1377, 2018.