PUSHING THE LIMITS OF SENTINEL-2 FOR BUILDING FOOTPRINT EXTRACTION

C. Ayala[1], *C. Aranda*[1], *M. Galar*[2]

[1] Tracasa Instrumental, Calle Cabárceno, 6, 31621 Sarriguren, Navarra
(cayala, caranda)@itracasa.es
[2] Institute of Smart Cities (ISC), Public University of Navarre,
Arrosadia Campus, 31006, Pamplona, Spain - mikel.galar@unavarra.es

ABSTRACT

Building footprint maps are of high importance nowadays since a wide range of services relies on them to work. However, activities to keep these maps up-to-date are costly and time-consuming due to the great deal of human intervention required. Several automation attempts have been carried out in the last decade aiming at fully automatizing them. However, taking into account the complexity of the task and the current limitations of semantic segmentation deep learning models, the vast majority of approaches rely on aerial imagery ($<$ 1 m). As a result, prohibitive costs and high revisit times prevent the remote sensing community from maintaining up-to-date building maps. This work proposes a novel deep learning architecture to accurately extract building footprints from high resolution satellite imagery (10 m). Accordingly, super-resolution and semantic segmentation techniques have been fused to make it possible not only to improve the building's boundary definition but also to detect buildings with sub-pixel width. As a result, fine-grained building maps at 2.5 m are generated using Sentinel-2 imagery, closing the gap between satellite and aerial semantic segmentation.

Index Terms— Sentinel-2, Remote Sensing, Building Detection, Deep Learning, Convolutional Neural Networks.

1. INTRODUCTION

In this day and age, the detection of objects in remote sensing imagery has many immediate applications. In the last decade, attention has been given to the extraction of building footprints, since they are important inputs for services such as urban planning or rapid mapping. However, a great level of precision and regular updates are needed by these applications to work properly. Traditionally, the extraction of building footprints has been manually performed by remote sensing specialists occasionally assisted by semi-automatic tools. Taking into account not only their complexity but also the great amount of time required by these tasks, many automation attempts have been carried out [1].

Despite initial automation attempts relying on the usage of traditional machine learning algorithms, the rise of deep learning has revolutionized the remote sensing literature [2]. Accordingly, the most promising approaches use deep learning techniques to automatically detect objects such as building footprints [3].

The building detection task requires a great level of detail in the produced masks. However, deep learning models struggle to precisely define building's edges and corners. Due to this limitation, researchers have resorted to other ways for increasing the accuracy of the generated masks. For that purpose, the most straightforward solution is to use very high resolution imagery. Consequently, very high resolution imagery has been traditionally employed to diminish the network's deficiencies [4].

Therefore, the vast majority of works make use of aerial imagery ($< 1\ m$) to extract building footprints [5]. Nevertheless, it must be taken into account that the cost of these products hinders their application on a daily basis. Conversely, few works have assessed the usage of high resolution imagery ($<\ 10\ m$) for building detection tasks, although its greater availability thanks to their lower costs and shorter revisit times [6].

In 2014, the European Space Agency (ESA) in partnership with the European Commission created the Copernicus programme to make remote sensing data more accessible and affordable in Europe. Since then, all the information produced in the framework of Copernicus has been made available free-of-charge to the public. Among the seven missions currently being developed under the Copernicus programme, we have focused on the multi-spectral sensor Sentinel-2 (S2). S2 produces high resolution optical images composed of thirteen bands. In this work, we focus on those bands given at the greatest resolution of 10 m.

Our hypothesis is that using high resolution imagery (10 m) it is possible to extract building footprints precisely. For this purpose, we propose a novel deep learning architecture that fuses super-resolution and semantic segmentation

Christian Ayala was partially supported by the Goverment of Navarra under the industrial PhD program 2020 reference 0011-1408-2020-000008. Mikel Galar was partially supported by Tracasa Instrumental S.L. under projects OTRI 2018-901-073, OTRI 2019-901-091 and OTRI 2020-901-050.

© 2022 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

techniques. Accordingly, this work opens up new possibilities for a wide range of applications, since costs and revisit time are drastically reduced.

The proposed deep learning architecture is based on the U-Net [7] architecture and produces enhanced segmentation masks that quadruple the resolution at the input (2.5 m). In the experimental study, our approach is compared to the standard U-Net that does not alter the output resolution (10 m). Moreover, aiming at assessing the generalization ability of the different approaches against color spectrum variations caused by seasonal rhythms, the study is extended to multiple time-steps thanks to the availability offered by S2. Accordingly, the dataset is composed of 14 cities spread across the Spanish territory. For each city, two trimesters have been considered (2018/12 - 2019/03 and 2019/03 - 2019/06). Moreover, according to the machine learning guidelines [8], the dataset has been divided into training and test sub-sets. The performance of the different architectures has been evaluated using the F-score and the Intersection over Union (IoU) metrics. As a result, experiments show that building footprints can be accurately extracted from high resolution satellite imagery using the proposed deep learning model.

The remainder of this article is organized as follows. The methodology is detailed in Section 2. Thereafter, the experimental framework, experiments, and results are presented and discussed in Section 3. Finally, Section 4 concludes this work and present some future research.

2. METHODOLOGY

2.1. Dataset

The vast majority of open datasets that are focused on building footprint extraction consist of hand-labeled aerial imagery [9]. Therefore, we have opted for creating our own dataset combining S2 imagery with OpenStreetMap (OSM) [10] building labels. It must be noted that OSM building polygons have been rasterized to 2.5 m since the architecture used produces enhanced segmentation masks that quadruple the resolution given at the input (10 m). Although OSM may contain labeling errors, mostly in rural areas, previous works [11] have demonstrated that it is possible to reach a good performance when using large datasets. Accordingly, 14 cities spread across the Spanish territory have been selected and divided into two sets following the machine learning guidelines [8]. Additionally, since the usage of S2 allows us to study the generalization capabilities of the models with respect to different time-steps in the same location, two trimesters (2018/12 - 2019/03 and 2019/03 - 2019/06) have been considered for each city in the dataset. It must be noted that, although in this work we make use of two trimesters, this methodology can be extrapolated to any number of trimesters, even with shorter time intervals (e.g. weeks instead of trimesters). As it is shown in Table 1, each city is assigned to a single set to prevent data leakage.

City	Dimensions	# buildings	Set
A coruña	704 × 576	8554	Train
Albacete	1280 × 1152	5793	Train
Alicante	1216 × 1472	19894	Train
Barcelona N.	1152 × 1728	63783	Test
San Sebastián	512 × 768	5363	Test
Granada	1664 × 1600	10911	Test
Logroño	768 × 960	1996	Train
Madrid N.	1920 × 2688	102750	Train
Murcia	1792 × 1600	7528	Train
Oviedo	960 × 896	11876	Train
Pamplona	1600 × 1536	9489	Test
Santander	1152 × 1216	14148	Train
Valencia	2304 × 1728	30821	Train
Zaragoza	2304 × 2752	10662	Train

Table 1. Summary of the dataset.

2.2. Model

The architecture proposed in this work, as it can be seen in Figure 1, is based on the U-Net model [7].

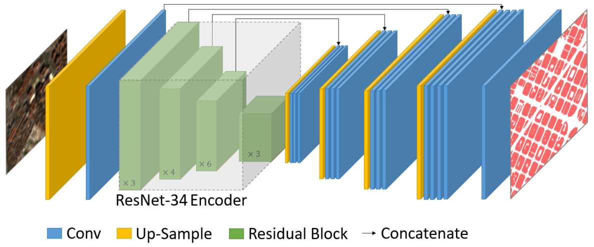

Fig. 1. Proposed U-Net-based architecture.

Two main modifications have been performed to the original U-Net architecture:

1. To include an up-scaling layer prior to the feature extractor in order to take advantage of the U-Net's skip connections and thus, prevent the loss of pattern information. Although in this work we have opted for the classical nearest-neighbor interpolation algorithm to up-scale the input, other algorithms may be considered.
2. To replace the original U-Net's encoder with a ResNet-34 [12], given the capacity residuals models have to exploit the available information whilst reducing the computational cost.

3. EXPERIMENTAL STUDY

3.1. Experimental framework

The architectures proposed in this paper have been implemented using the Keras deep learning framework. A combination of the Dice and Focal losses has been chosen as the loss function, minimized using the Adam optimizer with a fixed learning rate of 1e-3. Models have been trained for 100K iterations, randomly taking batches of 14, 128×128 samples.

3.2. Experiments and discussion

Two experiments have been carried out to answer the following questions:

1. Can high resolution satellite imagery limitations for building footprint detection be overcome by increasing the output resolution?
2. How do the models generalize to varying conditions in different time steps? Can the generalization ability of the models against color spectrum variations be increased using a dataset with multiple time-steps?

Tables 2 and 3 presents the results obtained for experiments 1 and 2, respectively, in terms of relaxed F-score and IoU. Metrics have been individually computed for each city in the test set. Additionally, the overall performance has been computed. It must be noted that the best results for each row and trimester are presented in **boldface**.

3.2.1. Experiment 1: Increasing the output resolution

In this experiment, we compare the original U-Net architecture that preserves the resolution and our approach that quadruples it. That is, when using S2 imagery (10 m), the U-Net gives a 10 m resolution segmentation mask, while 2.5 m masks are provided by the model proposed in this work. It must be noted that in this experiment we have only made use of the first trimester (2018/12 - 2019/03) for both training and testing.

Table 2 quantitatively reports that increasing the resolution at the output results in more accurate segmentation masks. There is a noticeable increase in both average IoU and F-score metrics when generating masks with more resolution than the one given at the input.

	U-Net x1		U-Net + Nearest x4	
City	IoU	F-score	IoU	F-score
Barcelona N.	**0.5878**	**0.7404**	0.5474	0.7075
San Sebastián	0.6207	0.7660	**0.7343**	**0.8468**
Granada	0.6522	0.7895	**0.8119**	**0.8962**
Pamplona	0.5775	0.7322	**0.7490**	**0.8565**
Overall	0.6095	0.7570	**0.7106**	**0.8267**

Table 2. Comparison between the original U-Net architecture and the one proposed in this work that quadruples the resolution at the output.

3.2.2. Experiment 2: Adding multiple time-steps

Here the generalization capability of the model against color spectrum variations mainly caused by adverse atmospheric phenomena and seasonal rhythms is evaluated. In this experiment, we compare the *U-Net + Nearest x4* model trained on the first trimester (2018/12 - 2019/03) with the same model trained on the two trimesters (2018/12 - 2019/03 and 2019/03 - 2019/06).

Results presented in Table 3 show significant differences between both models. That is, when training using only the first trimester, the model (*U-Net + Nearest x4*) does not generalize well to unseen trimesters.However, when the dataset is augmented using multiple time-steps also for training, the model (*U-Net + Nearest x4 + MT*) learns how to overcome color variations, outperforming the previous approach. Moreover, as it can be observed, the results obtained are consistent not only across all the cities within the test set but also are much more stable over time.

	U-Net + Nearest x4				U-Net + Nearest x4 + MT			
	2018/12		2019/03		2018/12		2019/03	
City	IoU	F-score	IoU	F-score	IoU	F-score	IoU	F-score
Barcelona N.	0.5474	0.7075	**0.7901**	**0.8821**	**0.6814**	**0.8105**	0.7678	0.8686
San Sebastián	0.7343	0.8468	0.5992	0.7494	**0.7374**	**0.8488**	**0.6467**	**0.7854**
Granada	**0.8119**	**0.8962**	0.7241	0.8400	0.7945	0.8855	**0.7846**	**0.8793**
Pamplona	**0.7490**	**0.8565**	0.5245	0.6881	0.7172	0.8353	**0.6711**	**0.8032**
Overall	0.7106	0.8267	0.6594	0.7899	**0.7326**	**0.8450**	**0.7175**	**0.8341**

Table 3. Comparison between training using a single trimester and using multiple trimesters.

Figure 2 qualitatively reports the performance of the different architectures tested in this paper, showing several samples randomly taken from the test set. Thoroughly examining the figure, conclusions extracted from Tables 2 and 3 are reinforced. As it can be observed, the original U-Net model struggles to detect small elements, often resulting in coarse segmentation masks. For this reason, when the output resolution is increased, semantic segmentation models cannot only detect objects with sub-pixel width but also precisely define building edges. It must be noted that the benefit of including multiple time-steps (trimesters) is also reflected in the figure. Moreover, the hyper-temporal component makes it possible for the model to deal with complex scenarios and significant color spectrum variations.

4. CONCLUSIONS AND FUTURE WORK

This paper proposes a new deep learning-based architecture to extract building footprints from high resolution satellite (10 m) imagery. As it has been quantitatively and quali-

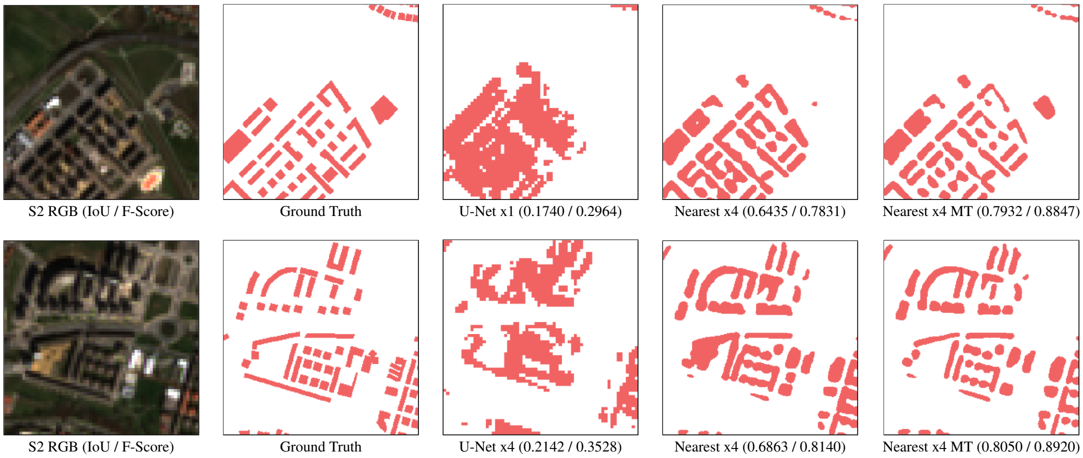
S2 RGB (IoU / F-Score)
Ground Truth
U-Net x1 (0.1740 / 0.2964)
Nearest x4 (0.6435 / 0.7831)
Nearest x4 MT (0.7932 / 0.8847)
S2 RGB (IoU / F-Score)
Ground Truth
U-Net x4 (0.2142 / 0.3528)
Nearest x4 (0.6863 / 0.8140)
Nearest x4 MT (0.8050 / 0.8920)

Fig. 2. Visual comparison between the original U-Net architecture trained using only the first trimester, and the proposed modification with nearest-neighbor input up-scaling trained using both single multi-temporal data.

tatively proved, the proposed model is capable of detecting buildings even if their width reaches sub-pixel size.

However, there is a great deal of research lines that should be addressed in the near future. More areas of interest should be included to make both models and evaluation more robust and fairer. Additionally, it would be interesting to compare the proposed architecture with other state-of-the-art approaches.

5. REFERENCES

[1] John Ball, Derek Anderson, and Chee Seng Chan, "A comprehensive survey of deep learning in remote sensing: Theories, tools and challenges for the community," *Journal of Applied Remote Sensing*, 08 2017.

[2] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, 2016.

[3] X. X. Zhu, D. Tuia, L. Mou, G. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, 2017.

[4] Thorsten Hoeser and Claudia Kuenzer, "Object detection and image segmentation with deep learning on Earth observation data: A review-part I: Evolution and recent trends," *Remote Sensing*, 2020.

[5] Y. Feng, C. Yang, and M. Sester, "Multi-scale building maps from aerial imagery," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLIII-B3-2020, pp. 41–47, 2020.

[6] Jian Li and David P. Roy, "A global analysis of sentinel-2a, sentinel-2b and landsat-8 data revisit intervals and implications for terrestrial monitoring," *Remote Sensing*, vol. 9, no. 9, 2017.

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015.

[8] Christopher M. Bishop, *Pattern Recoginiton and Machine Learning*, Springer, 2006.

[9] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "High-resolution aerial image labeling with convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 12, pp. 7092–7103, 2017.

[10] M. Haklay and P. Weber, "Openstreetmap: User-generated street maps," *IEEE Pervasive Computing*, 2008.

[11] Pascal Kaiser, Jan Dirk Wegner, Aurélien Lucchi, Martin Jaggi, Thomas Hofmann, and Konrad Schindler, "Learning aerial image segmentation from online maps," *CoRR*, 2017.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016.