MULTI-TEMPORAL DATA AUGMENTATION FOR HIGH FREQUENCY SATELLITE IMAGERY: A CASE STUDY IN SENTINEL-1 AND SENTINEL-2 BUILDING AND ROAD SEGMENTATION

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), Department of Statistics, Computer Science and Mathematics, Public University of Navarre (UPNA), Arrosadia Campus, 31006, Pamplona, Spain - mikel.galar@unavarra.es

KEY WORDS: Sentinel-2, Sentinel-1, Multi-temporal, Remote Sensing, Road Network Extraction, Building Footprint Detection, Deep Learning, Convolutional Neural Networks.

ABSTRACT:

Semantic segmentation of remote sensing images has many practical applications such as urban planning or disaster assessment. Deep learning-based approaches have shown their usefulness in automatically segmenting large remote sensing images, helping to automatize these tasks. However, deep learning models require large amounts of labeled data to generalize well to unseen scenarios. The generation of global-scale remote sensing datasets with high intraclass variability presents a major challenge. For this reason, data augmentation techniques have been widely applied to artificially increase the size of the datasets. Among them, photometric data augmentation techniques such as random brightness, contrast, saturation, and hue have been traditionally applied aiming at improving the generalization against color spectrum variations, but they can have a negative effect on the model due to their synthetic nature. To solve this issue, sensors with high revisit times such as Sentinel-1 and Sentinel-2 can be exploited to realistically augment the dataset. Accordingly, this paper sets out a novel realistic multi-temporal color data augmentation technique. The proposed methodology has been evaluated in the building and road semantic segmentation tasks, considering a dataset composed of 38 Spanish cities. As a result, the experimental study shows the usefulness of the proposed multi-temporal data augmentation technique, which can be further improved with traditional photometric transformations.

1. INTRODUCTION

In the last decade, the remote sensing community has rapidly grown, mainly due to the great deal of potential applications that have emerged. In fact, insights derived from the foto-interpretation of earth observation products can be used for urban planning (Guo et al., 2021) or disaster assessment (Ghaffarian and Emtehani, 2021) among other use cases.

Traditionally, the foto-interpretation of large remote sensing images has been manually performed by experts, demanding a great deal of human effort and thus, entailing high costs. However, recent advances in deep learning, especially with Convolutional Neural Networks (CNNs), have made it possible to process vast amounts of remote sensing data, reducing costs and saving time (Zhu et al., 2017).

Deep learning models are data-hungry since they require large amounts of labeled data to generalize to unseen scenarios. Furthermore, this problem may be even more evident in remote sensing imagery than in problems involving natural images, since earth observation images are subject to color spectrum variations caused by the sun's position, adverse atmospheric conditions, etc. (Guo et al., 2020). Therefore, it is very costly and time-consuming to develop deep learning models that generalize well even to cases where spatial and temporal shifts occur.

In deep learning, Data Augmentation (DA) is commonly used to face the lack of labeled data by artificially introducing small changes to the inputs without altering the outputs, giving the models more variety without increasing the size of the dataset. DA techniques may be seen as a powerful tool to face the lack of labeled data, artificially expanding the dataset (Taylor and Nitschke, 2018). When working with images, geometric DA techniques such as random rotations, flips, crops, and scale transformations are commonly applied to prevent over-fitting and improve generalization (Dieleman et al., 2015). However, geometric DA techniques do not make the model robust against color spectrum variations. In this regard, photometric DA techniques such as color jittering, random brightness, contrast, hue, and saturation transformations are applied (Wu et al., 2015).

Despite photometric DA techniques have been proved beneficial in a wide range of remote sensing tasks, the resulting images may contain synthetic artifacts such as saturated pixels or null values, losing valuable spectral information. Furthermore, the parameters of photometric DA techniques are difficult to tune, since they depend on each specific problem. Moreover, there are events such as shadows casted by near buildings, seasonal rhythms, crop cycles, etc., that can not be simulated through photometric DA techniques.

To address these problems, this paper proposes a simple methodology that takes advantage of the high revisit times provided by sensors such as Sentinel-1 (S1) and Sentinel-2 (S2) to perform a realistic multi-temporal color data augmentation (multi-temporal DA). The idea is to consider multiple observations for the same area of interest to have a variety of color spectrums coming from real images. In this regard, for a given area of interest, multiple observations are considered, varying the color spectrum without including synthetic artifacts.

To assess the usefulness of the proposed approach both building and road semantic segmentation problems have been considered following the experimental framework in (Ayala et al., 2021).

[*] Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIII-B3-2022-25-2022

It must be noted that buildings and roads have different degrees of variations in their shapes and colors, which makes them ideal for this study. For evaluating the proposed approach, a dataset composed of 38 Spanish cities has been considered. Moreover, for each city, four observations have been chosen corresponding to the four seasons of a year. The experiments, which have been evaluated using the Intersection over Union (IoU) and F-score metrics, showed that the proposed methodology improves the results from traditional DA techniques in the two scenarios. Furthermore, when the proposed multi-temporal DA technique is combined with the traditional photometric DA transformations, the results are further enhanced.

2. RELATED WORKS

DA techniques such as geometric and photometric prevent overfitting artificially increasing the variety of the dataset. Generally, geometric transformations lead to larger improvements in the model's performance than photometric transformations (Taylor and Nitschke, 2018). Furthermore, the former is easier to implement and computationally more efficient compared to the latter. In remote sensing, distortions of rigid-shape objects are commonly avoided. Hence, geometric transformations such as the dihedral DA technique which combines 90-degree rotations along with vertical and horizontal flips are used, which do not alter the image content (Iglovikov et al., 2017).

When photometric transformations are applied it is easy to lose spectral information, resulting in unrealistic images. In spite of this, in remote sensing, the models need to learn how to deal with color spectrum variations such as seasonal rhythms, shadows, etc., which are typical in every use case related to earth observation.

There are more complex DA techniques that use domain-specific synthesis to expand the dataset. These techniques generate richer data compared to the generic geometric and photometric augmentations (Peng et al., 2014). For example, Yan et al., proposed a novel data augmentation method that simulates remote sensing images combining background images and 3D ship models for tackling the insufficient number of training samples in the ship detection task (Yan et al., 2019a). Thereafter, they extrapolate the methodology to aircraft detection tasks, employing 3D aircraft models to form simulated images (Yan et al., 2019b). Illarionova et al. proposed an object-based augmentation technique that exploits segmentation masks to generate new training samples copy-pasting objects in label-free backgrounds (Illarionova et al., 2021), outperforming standard geometric and photometric DA techniques. Generative Adversarial Networks (GANs) have been also used to generate plausible synthetic data along with their corresponding segmentation masks (Howe et al., 2019). However, the development of complex DA approaches requires domain-specific knowledge, making them not applicable to different problems.

In this paper, we focus on exploiting multi-temporal data for data augmentation. Multi-temporal data is useful for a wide range of applications. Multiple observations of the same area can be used to learn transferable representations leveraging temporal information (Mañas et al., 2021). Furthermore, multi-scale spatio-temporal features can be extracted by making use of complex deep learning architectures that combine CNNs with Recurrent Neural Networks (RNNs) (Garnot and Landrieu, 2021). However, to the best of our knowledge, no previous work takes advantage of the high revisit times provided by sensors such as S1 and S2 to realistically augment the dataset, making models robust against color spectrum variations.

3. PROPOSAL

Photometric DA techniques such as random transformations of the brightness, contrast, saturation, and hue, may produce undesired synthetic artifacts, having a negative effect on the model performance. Moreover, the application of these techniques may result in unrealistic images, since the spectral information is arbitrarily altered. Furthermore, setting the proper hyper-parameters for these DAs is not straightforward, since they need to be adapted to each problem. For this reason, this paper proposes a novel easy-to-implement color DA technique, taking advantage of the high revisit times provided by S1 and S2 sensors.

Rather than applying standard photometric DA techniques that alter the original image, multiple observations can be considered for the same area, preserving their original color information and hence, avoiding creating undesired synthetic artifacts. Our hypothesis is that this approach can be more effective than the traditional photometric DA since there are events such as seasonal rhythms, sun position, or shadows casted by buildings that can not be easily simulated. Figures 1 and 2 can help understanding the differences between photometric DA and the usage of multiple observations. Figure 1 shows the differences between three observations (O_1, O_2, and O_3) and their corresponding augmented versions applying brightness, saturation and contrast photometric DA transformations to the RGB channels. This figure shows the fact that events such as harvesting cannot be easily simulated with standard photometric DA transformations (e.g., O_1 cannot be obtained from O_2 or O_3). Moreover, Figure 2 shows the differences between three (O_1, O_2, and O_3) S1 observations. As it can be seen in the figure, the nature of radar data makes the application of photometric DA transformations complex and meaningless.

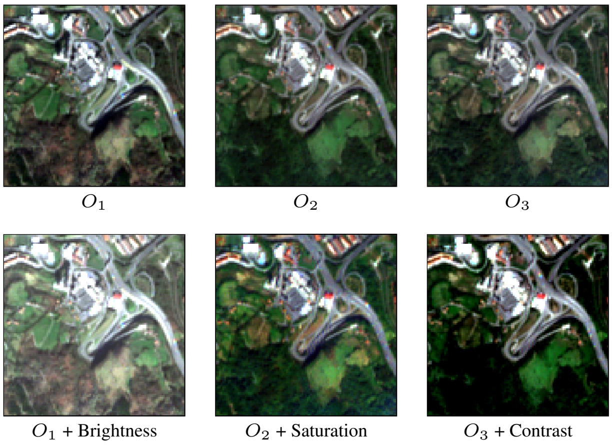


Figure 1. Visual comparison of the proposed multi-temporal color DA technique based on multiple observations and the corresponding altered versions using standard photometric transformations.

In (Ayala et al., 2021) multiple observations were used to augment the dataset, however, the experimental setup did not assess the contribution of using multiple observations. Therefore, this paper aims to deeply study the effect that the proposed multi-temporal DA technique has on the robustness of semantic seg-

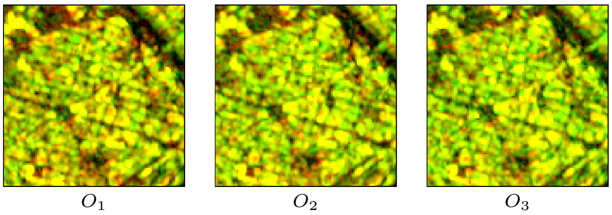


Figure 2. Visual comparison of multiple observations for S1's VV and VH backscatter Red-Green composition.

mentation remote sensing deep learning models. For this purpose four trimesters have been considered following the dataset described in (Ayala et al., 2021), leaving the last one for assessing the performance of the models.

4. EXPERIMENTAL STUDY

In this section, the experimental study carried out to assess the usefulness of the proposed multi-temporal DA technique is presented. First, the dataset generation pipeline is described in Section 4.1. Then, details regarding the experimental framework are given in Section 4.2. Thereafter, experiments carried out are outlined in Section 4.3. Finally, Section 4.4 summarizes the results and the conclusions extracted from the experiments.

4.1 Dataset

In this work, we have made use of the dataset described in (Ayala et al., 2021). The dataset has been generated by combining high-resolution S1 and S2 satellite imagery along with OpenStreetMap (OSM) building and road annotations. Moreover, given the high revisit times provided by S1 and S2 sensors, multiple observations have been considered. Specifically, we have considered data from the four seasons of a year: 2018-06/2018-09, 2018-09/2018-12, 2018-12/2019-03, and 2019-03/2019/06. It must be noted that the number of observations could be arbitrarily increased up to 70 on the equator per year.

Figure 3 depicts the overall pipeline for a generic region of interest. First, S2 products are downloaded from the Sentinels Scientific Data Hub (SciHub). The 10 m GSD bands from S2 are selected (Red, Green, Blue, and Near Infrared). Furthermore, the Normalized Difference Vegetation Index (NDVI) is also calculated and combined with the other bands.

In the case of S1, we used the Level-1 GRD product in the Interferometric Wide (IW) swath mode. This product has a swath width of 250 kilometers, a resolution of 20 × 22 m (depending on the beam id), and could be provided in four polarization modes (VV, VH, HH, HV). However, because dual horizontal polarization (HH, HV) is limited to polar regions, only dual vertical polarization (VV, VH) has been considered. The SciHub has been used to download S1 raw products, queried by a time interval of 7 days ± the mean of the ingestion times of the S2 products considered in the preceding stage. After that, raw S1 products were pre-processed using the Sentinel application platform (SNAP). Firstly, in the radiometric calibration stage, backscatter intensities were estimated using the GRD metadata. Then, in the terrain correction step, the Digital Elevation Model (DEM) from the Shuttle Radar Topography Mission (SRTM) has been used to address the side-looking effects. Finally, backscatter intensities were log-scaled and converted to decibels.

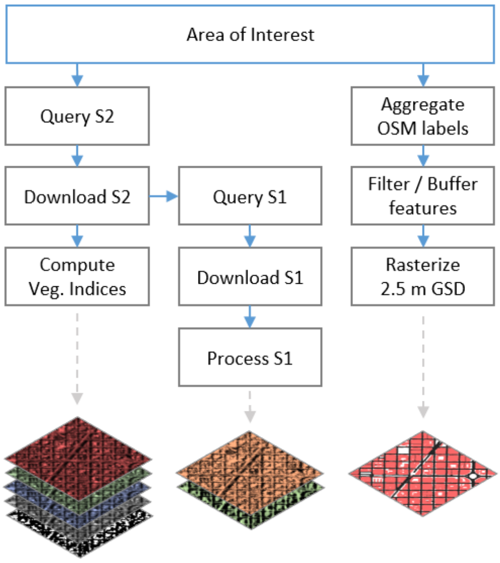


Figure 3. dataset generation pipeline for a generic area of interest.

On the other hand, OSM has been proved useful for a great deal of remote sensing tasks (Kaiser et al., 2017). However, OSM should be reclassified beforehand due to the large number of layers it contains. In this regard, different types of roads have been aggregated to construct the road label, whereas the building polygon outlines constitute the building label. The selected OSM codes are presented in Table 1. It must be noted that, since OSM only contains roads' centerlines, line-strings were buffered to match S2's spatial resolution (10 m GSD). Moreover, due to the limited spatial resolution of S1 and S2 sensors, buildings with an area inferior to 50 m^2 have been filtered out. Finally, building and road vector features have been rasterized to 2.5 m GSD.

Code	Fclass	Description
5111	motorway	Motorway/freeway
5112	trunk	Important roads, typically divided
5113	primary	Primary roads, typically national
5114	secondary	Secondary roads, typically regional
5115	tertiary	Tertiary roads, typically local
5121	unclassified	Smaller local roads
5122	residential	Roads in residential areas
5123	living_street	Streets where pedestrians have priority
5124	pedestrian	Pedestrian only streets
5131	motorway_link	Roads connections (same of lower category)
5132	trunk_link	Roads connections (same of lower category)
5133	primary_link	Roads connections (same of lower category)
5134	secondary_link	Roads connections (same of lower category)
1500		Building outlines

Table 1. Reclassification of OSM vector features into the road and building labels.

It must be noted that, as suggested in (Ayala et al., 2021), sensor and label-specific validation masks have been taken into account to handle sensing noise and labeling errors, respectively. Accordingly, validation masks have been used at both training and testing times to filter out low-quality samples.

The final dataset comprises 38 Spanish cities, which have been separated into two sub-sets following the machine learning standards. That is, in order to prevent data leakage, each city is assigned to either the training set or the test set, as shown in Table 2. It must be noted that this dataset is the same used in (Ayala et al., 2021), discarding cities with missing observations.

City	Dimensions	Set
La Coruña	704 × 576	Train
Albacete	1280 × 1152	Train
Alicante	1216 × 1472	Train
Barcelona N.	1152 × 1728	Test
Barcelona S.	896 × 1088	Test
Bilbao	576 × 832	Train
Burgos	512 × 704	Train
Cáceres	1024 × 896	Test
Cartagena	768 × 1216	Train
Castellón	1024 × 1024	Train
Córdoba	1088 × 1792	Train
Denia	640 × 768	Train
Ferrol	384 × 704	Test
Gijón	704 × 832	Test
Granada	1664 × 1600	Test
León	1216 × 768	Train
Logroño	768 × 960	Train
Lugo	768 × 576	Test
Madrid S.	1280 × 2624	Train
Majadahonda	1472 × 1344	Test
Mérida	512 × 640	Train
Murcia	1792 × 1600	Train
Ourense	960 × 704	Train
Oviedo	960 × 896	Train
Palma	1024 × 1344	Test
Pamplona	1600 × 1536	Test
Pontevedra	384 × 512	Train
Rivas-vacía	1088 × 1088	Train
Salamanca	832 × 960	Train
Santander	1152 × 1216	Train
Sevilla	2176 × 2368	Train
Teruel	640 × 768	Test
Valencia	2304 × 1728	Test
Valladolid	1408 × 1024	Test
Vigo	704 × 1024	Train
Vitoria	576 × 896	Train
Zamora	512 × 576	Train
Zaragoza	2304 × 2752	Train

Table 2. Summary of the dataset. Overall, the training set is composed of 25 zones ($\approx$ 66%) whereas the test set consists of 13 zones ($\approx$ 34%).

4.2 Experimental framework

The experimental framework also follows the specifications described in (Ayala et al., 2021). Regarding the deep learning network itself, a U-Net architecture (Ronneberger et al., 2015) has been considered. As it can be seen in Figure 4, the vanilla U-Net architecture has been modified including a bicubic upscaling layer prior to the feature extractor and replacing the base encoder with a ResNet-34 (He et al., 2016). As a result, semantic segmentation masks that quadruple the input spatial resolution are generated, making it possible to detect elements with sub-pixel width.

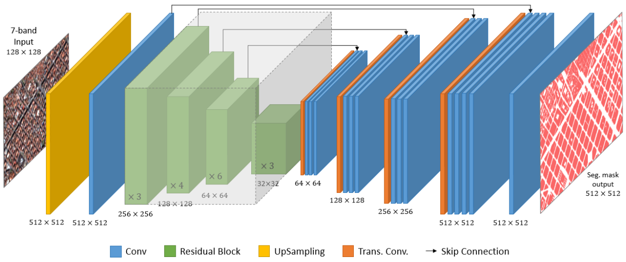


Figure 4. Network architecture.

Considering the large number of experiments we plan to run in order to contrast the usage of photometric DA with the proposed multi-temporal DA technique, we have opted for reducing the number of epochs from 1,000 to 200 in comparison with (Ayala et al., 2021). That is, all the models have been trained for 200 epochs consisting of 1,000 gradient updates. It must be noted that this modification does not alter the conclusions derived, since there is little margin for improvement after this epoch as contrasted in our preliminary experiments. The batch size has been set to 32 samples of 128 × 128 pixels. Furthermore, samples have been randomly taken, considering only those with at least 10% of pixels corresponding to the positive class (either road or building). Finally, since no validation set has been used, the last epoch model is taken.

Regarding the loss function, a combination of the Binary Cross-entropy and the Dice Loss has been chosen, to better control the trade-off between false positives and false negatives:

$$\begin{gathered}\mathcal{L}(y,\hat{y}) = \alpha \times \mathcal{L}_{BCE}(y,\hat{y}) + (1-\alpha) \times \mathcal{L}_{DICE}(y,\hat{y}),\\ \mathcal{L}_{BCE}(y,\hat{y}) = -(y\log(\hat{y}) + (1-y)\log(1-\hat{y})),\\ \mathcal{L}_{DICE}(y,\hat{y}) = \frac{2y\hat{y}+1}{y+\hat{y}+1}\end{gathered} \tag{1}$$

where $\hat{y}$ denotes the predicted segmentation mask, y the corresponding ground-truth mask, and the α parameter weights the contribution of the $\mathcal{L}_{DICE}$ loss (0.5 in these experiments). The loss function has been minimized, using the Adam optimizer with a fixed learning rate of $1e^{-3}$.

The Intersection over Union (IoU) and F-score metrics have been chosen to evaluate the performance of the models:

$$\text{IoU}(y,\hat{y}) = \frac{y \cap \hat{y}}{y \cup \hat{y}}, \qquad \text{F-score}(y,\hat{y}) = \frac{2y\hat{y}}{y+\hat{y}} \tag{2}$$

Additionally, both metrics are also calculated following a precision relaxation strategy (Mnih and Hinton, 2010, Zhang et al., 2018) aiming at reducing the impact of the low spatial resolution on the metrics. That is, doubtful pixels located on the edges of the roads and buildings are disregarded.

The experiments have been run on a computing node with a 2 × Intel Xeon E5-2609 v4 @ 1.70 GHz processor with 128 GB of RAM and 4 × NVIDIA RTX2080Ti GPUs (11 GB of RAM).

4.3 Experiments

Several experiments have been run to compare the proposed multi-temporal DA technique with the traditional photometric DA transformations. To make the evaluation fair, out of the 4 observations available in this dataset, the last one has been left out for testing purposes, whereas the remaining ones have been used to train the models.

First, the effect of including more observations has on the performance has been studied. In this regard, models have been trained considering 1, 2, and 3 observations, and tested using the 4th one. It must be noted that for 1 and 2 observations all their possible combinations have been run and averaged whereas, in the case of using 3 observations, the results of three executions have been averaged.

Thereafter, the proposed multi-temporal DA technique has been compared with the traditional photometric DA transformations, not only to determine which technique performs better but also to assess if they further improve the generalization capability when used together.

Despite the aforementioned color DA techniques, geometric DA techniques that have been widely used as a de facto augmentation in remote sensing are also applied. In this regard, the dihedral transformation, which consists of combinations of horizontal and vertical flips along with 90-degree rotations have been considered as a base for all experiments.

It must be noted that the same experiments have been run for building footprint detection and road network extraction tasks. Considering these two tasks, the usefulness of the proposed multi-temporal DA can be better assessed.

4.4 Results and discussion

Tables 3 and 4 summarize the quantitative results in terms of IoU and F-score obtained for the building footprint detection and road network extraction tasks, respectively. Additionally, a relaxed version of both metrics (Rlx. IoU and Rlx. F-score, respectively) is also calculated. Finally, the best results achieved in each task are presented in **boldface**.

Overall, increasing the number of observations with multi-temporal DA improves the generalization capability of the models. In fact, it is more beneficial to increase the number of observations than to apply photometric DA. Nevertheless, applying both DAs together provides the best performance. In the following, we analyze these findings in detail.

When working with mono-temporal imagery (a single observation) one can benefit from standard photometric DA techniques making models more robust to color spectrum variations (0.5051 ± 0.0107 vs. 0.4845 ± 0.0351 for buildings, and 0.5049 ± 0.0035 vs. 0.5008 ± 0.0072 for roads, in terms of IoU).

Nevertheless, if two observations are available, one can apply the proposed multi-temporal DA technique outperforming the standard photometric DA transformations applied over a single observation (0.5091 ± 0.0397 vs. 0.5051 ± 0.0107 for buildings, and 0.5125 ± 0.0044 vs. 0.5049 ± 0.0035 for roads, in terms of IoU).

Furthermore, there is a great increase in performance when considering 3 observations instead of only 2 (0.5635 ± 0.0101 vs. 0.5091 ± 0.0397 for buildings, and 0.5286 ± 0.0067 vs. 0.5125 ± 0.0044 for roads, in terms of IoU). In fact, the more the number of observations is, the better the generalization capability of the models becomes.

Finally, for all number of observations tested (1, 2, and 3), the standard photometric transformations help making models more robust. Furthermore, when photometric transformations are combined with the proposed multi-temporal DA technique with 3 observations the best results are achieved (0.5741 ± 0.0166 and, 0.5295 ± 0.0025, in terms of IoU for the building and road extraction tasks, respectively). It must be noted that both color DA techniques, in general, have a greater impact on building metrics than road ones, which is due to the higher variance of buildings shapes and colors compared to roads.

To complement the quantitative analysis, Figures 5 and 6 visually compare the performance of the proposed approaches in terms of visual IoU. That is, True Positives (TP) are presented in green, False Positives (FP) in blue, False Positives (FP) in red and True Negatives (TN) in white. According to these figures, one draws the same conclusions as those looking at Tables 3 and 4, respectively, with some extra information. Augmenting the dataset including multiple observations makes the model more robust against color spectrum variations. In this regard, the proposed multi-temporal DA technique is able to reduce the number of FP and FN. However, there are still some FP caused by labeling errors inherent to OSM.

5. CONCLUSIONS AND FUTURE WORK

In this paper, a novel color DA technique has been proposed, taking advantage of the high revisit times provided by sensors such as S1 and S2. Accordingly, multiple observations for the same area of interest are considered to have a variety of color spectrums coming from real images rather than augmenting the dataset synthetically using photometric DA transformations. The usefulness of the proposed method has been shown in two semantic segmentation tasks with different degrees of variation in their target's shapes and colors, outperforming standard photometric DA techniques. The multi-temporal DA technique requires no hyper-parameter tuning, which makes it easier to apply than traditional photometric DA transformations. Additionally, it can be directly applied to any sensor, including radar imagery such as S1, which is a limitation of photometric DA techniques. Finally, when the multi-temporal DA technique is combined with standard photometric DA techniques the best results are achieved.

Nonetheless, there are still several research lines on this subject that should be pursued in the future. Regarding the dataset, more observations should be considered to further assess the effect that increasing the number of observations has on the generalization capability of the model. Moreover, it would be interesting to extrapolate the analysis to other sensors different from S1 and S2 (e.g. hyperspectral, thermal, microwave, ...). Finally, other tasks such as land use and land cover semantic segmentation or classification of remote sensing images may be considered to gain valuable insights regarding not only the usefulness but also the limitations of the proposed approach.

6. ACKNOWLEDGMENTS

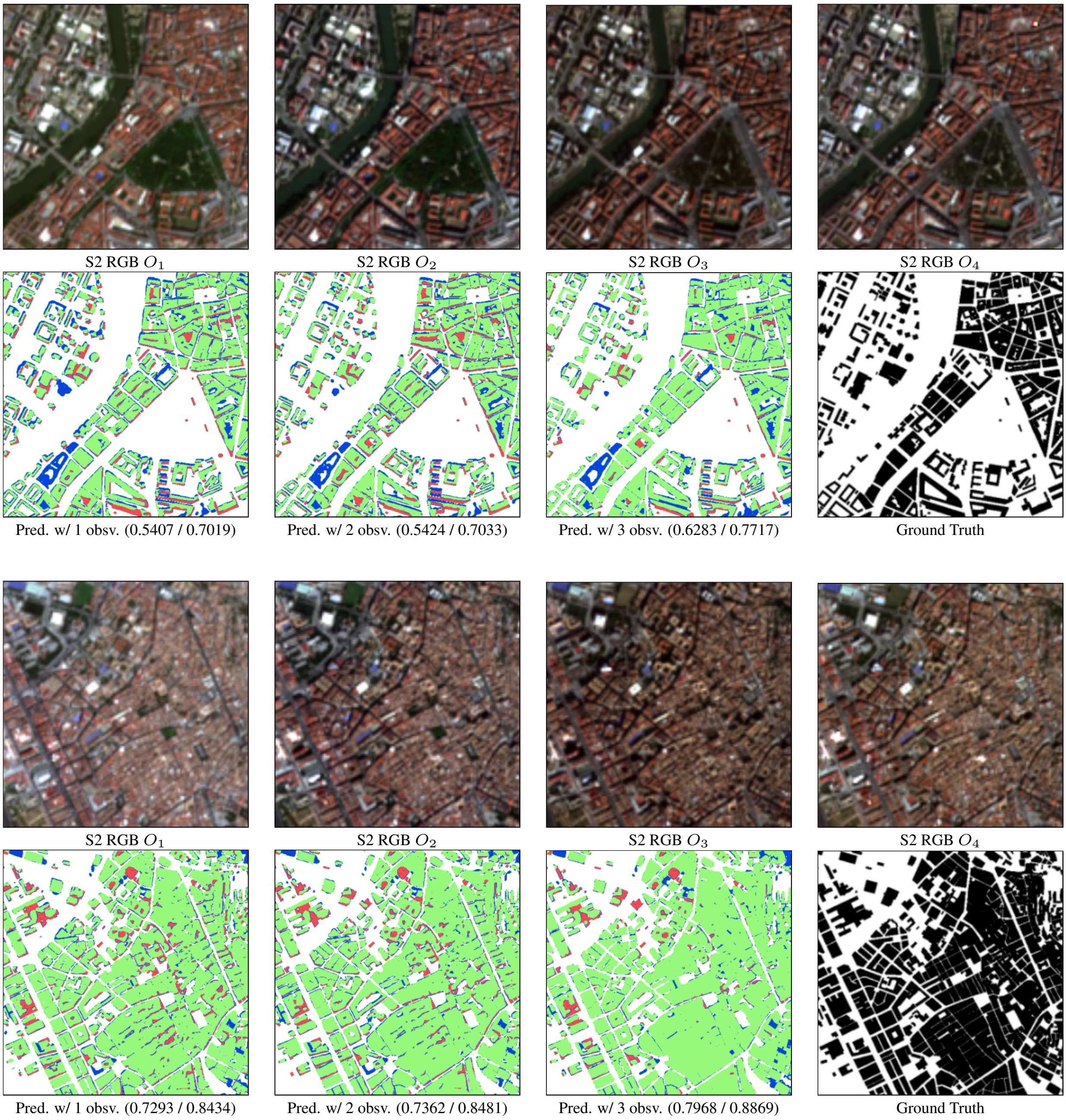
Christian Ayala was partially supported by the Government of Navarra under the industrial Ph.D. program 2020 reference 0011-1408-2020-000008. Mikel Galar was partially supported by Tracasa Instrumental S.L. under the project OTRI 2020-901-156, and by the Spanish MICIN (PID2019-108392GB-I00 / AEI / 10.13039/501100011033).

REFERENCES

Ayala, C., Sesma, R., Aranda, C., Galar, M., 2021. A Deep Learning Approach to an Enhanced Building Footprint and Road Detection in High-Resolution Satellite Imagery. *Remote Sensing*, 13(16).

Dieleman, S., Willett, K. W., Dambre, J., 2015. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly Notices of the Royal Astronomical Society*, 450(2), 1441–1459.

Garnot, V. S. F., Landrieu, L., 2021. Panoptic Segmentation of Satellite Image Time Series with Convolutional Temporal Attention Networks. *CoRR*, abs/2107.07933.



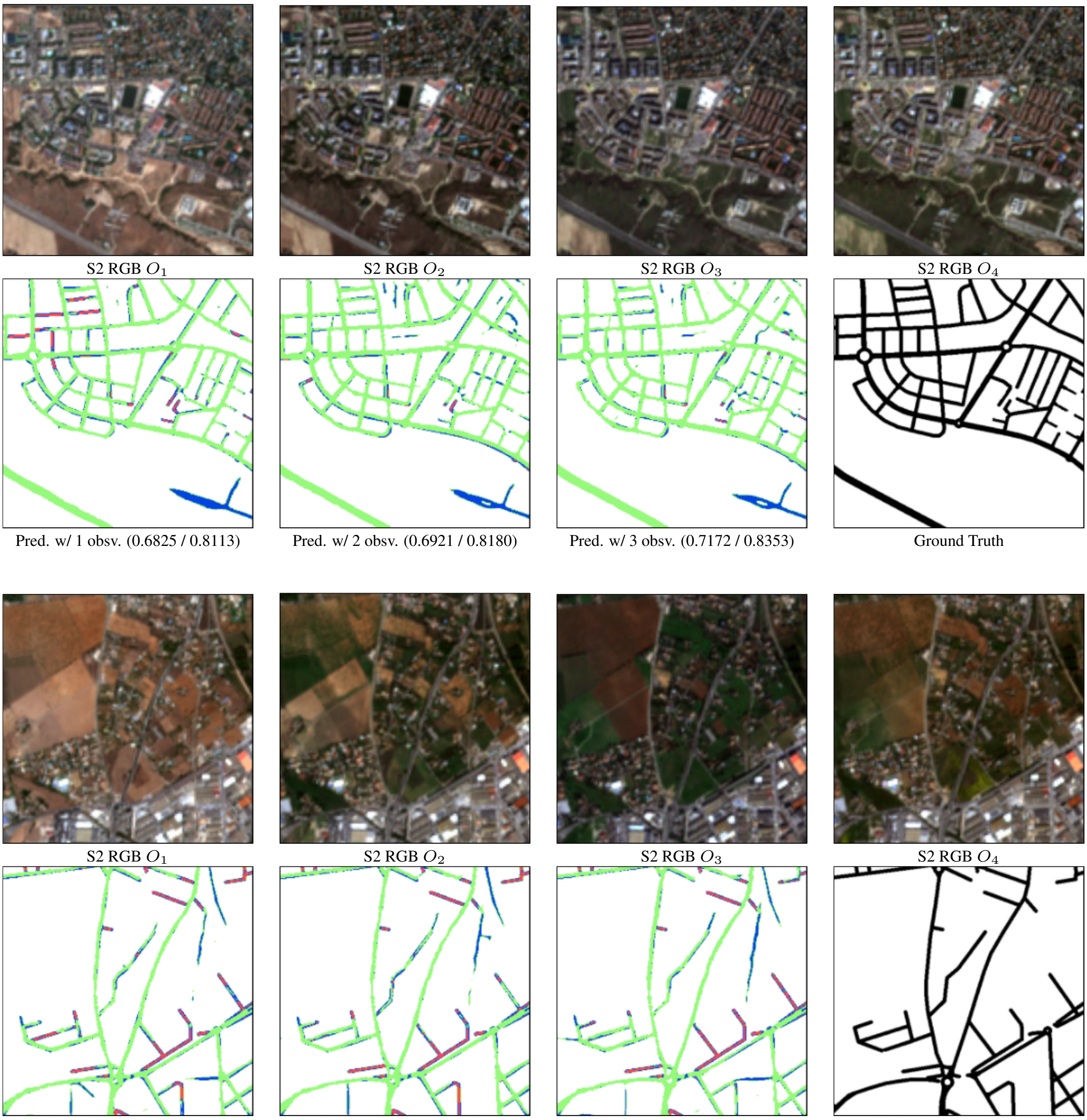
S2 RGB O_1 S2 RGB O_2 S2 RGB O_3 S2 RGB O_4

Pred. w/ 1 obsv. (0.5407 / 0.7019) Pred. w/ 2 obsv. (0.5424 / 0.7033) Pred. w/ 3 obsv. (0.6283 / 0.7717) Ground Truth

S2 RGB O_1 S2 RGB O_2 S2 RGB O_3 S2 RGB O_4

Pred. w/ 1 obsv. (0.7293 / 0.8434) Pred. w/ 2 obsv. (0.7362 / 0.8481) Pred. w/ 3 obsv. (0.7968 / 0.8869) Ground Truth

Figure 5. Visual comparison of the results obtained for the building footprint extraction task. Predictions are performed for the 4th observation (O_4) when trained using 1, 2, and 3 observations for two zones randomly taken from the test set. True Positives (TP) are presented in green, False Positives (FP) in blue, False Positives (FP) in red, and True Negatives (TN) in white. Moreover, the IoU and F-score metrics have been included.



S2 RGB O_1 | S2 RGB O_2 | S2 RGB O_3 | S2 RGB O_4

Pred. w/ 1 obsv. (0.6825 / 0.8113) | Pred. w/ 2 obsv. (0.6921 / 0.8180) | Pred. w/ 3 obsv. (0.7172 / 0.8353) | Ground Truth

S2 RGB O_1 | S2 RGB O_2 | S2 RGB O_3 | S2 RGB O_4

Pred. w/ 1 obsv. (0.5356 / 0.6976) | Pred. w/ 2 obsv. (0.5506 / 0.7101) | Pred. w/ 3 obsv. (0.5562 / 0.7148) | Ground Truth

Figure 6. Visual comparison of the results obtained for the road network extraction task. Predictions are performed for the 4th observation (O_4) when trained using 1, 2, and 3 observations for two zones randomly taken from the test set. True Positives (TP) are presented in green, False Positives (FP) in blue, False Positives (FP) in red, and True Negatives (TN) in white. Moreover, the IoU and F-score metrics have been included.

https://doi.org/10.5194/isprs-archives-XLIII-B3-2022-25-2022

#Observations	photometric DA	IoU	F-score	Rlx. IoU	Rlx. F-score
1	✗	0.4845 ± 0.0351	0.6321 ± 0.0422	0.5562 ± 0.0463	0.6872 ± 0.0529
	✓	0.5051 ± 0.0107	0.6518 ± 0.0109	0.5843 ± 0.0136	0.7116 ± 0.0137
2	✗	0.5091 ± 0.0397	0.6590 ± 0.0416	0.5831 ± 0.0478	0.7159 ± 0.0473
	✓	0.5457 ± 0.0125	0.6972 ± 0.0128	0.6314 ± 0.0110	0.7621 ± 0.0114
3	✗	0.5635 ± 0.0101	0.7133 ± 0.0077	0.6493 ± 0.0165	0.7771 ± 0.0118
	✓	**0.5741 ± 0.0166**	**0.7231 ± 0.0146**	**0.6658 ± 0.0228**	**0.7911 ± 0.0182**

Table 3. Results obtained in test set for the building extraction task.

#Observations	photometric DA	IoU	F-score	Rlx. IoU	Rlx. F-score
1	✗	0.5008 ± 0.0072	0.6645 ± 0.0065	0.5734 ± 0.0098	0.7255 ± 0.0080
	✓	0.5049 ± 0.0035	0.6680 ± 0.0030	0.5534 ± 0.0427	0.7084 ± 0.0357
2	✗	0.5125 ± 0.0044	0.6750 ± 0.0040	0.5882 ± 0.0052	0.7376 ± 0.0042
	✓	0.5173 ± 0.0058	0.6795 ± 0.0051	0.5902 ± 0.0087	0.7395 ± 0.0068
3	✗	0.5286 ± 0.0067	0.6895 ± 0.0058	0.6027 ± 0.0074	0.7494 ± 0.0057
	✓	**0.5295 ± 0.0025**	**0.6903 ± 0.0020**	**0.6058 ± 0.0019**	**0.7518 ± 0.0014**

Table 4. Results obtained in test set for the road extraction task.

Ghaffarian, S., Emtehani, S., 2021. Monitoring Urban Deprived Areas with Remote Sensing and Machine Learning in Case of Disaster Recovery. *Climate*, 9(4).

Guo, H., Shi, Q., Du, B., Zhang, L., Wang, D., Ding, H., 2020. Scene-Driven Multitask Parallel Attention Network for Building Extraction in High-Resolution Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 1–20.

Guo, H., Su, X., Tang, S., Du, B., Zhang, L., 2021. Scale-Robust Deep-Supervision Network for Mapping Building Footprints From High-Resolution Remote Sensing Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 10091–10100.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Howe, J., Pula, K., Reite, A. A., 2019. Conditional Generative Adversarial Networks for Data Augmentation and Adaptation in Remotely Sensed Imagery. *CoRR*, abs/1908.03809.

Iglovikov, V., Mushinskiy, S., Osin, V., 2017. Satellite Imagery Feature Detection using Deep Convolutional Neural Network: A Kaggle Competition. *CoRR*, abs/1706.06169.

Illarionova, S., Nesteruk, S., Shadrin, D., Ignatiev, V., Pukalchik, M., Oseledets, I. V., 2021. Object-Based Augmentation Improves Quality of Remote SensingSemantic Segmentation. *CoRR*, abs/2105.05516.

Kaiser, P., Wegner, J. D., Lucchi, A., Jaggi, M., Hofmann, T., Schindler, K., 2017. Learning Aerial Image Segmentation from Online Maps. *IEEE Transactions on Geoscience and Remote Sensing*.

Mañas, O., Lacoste, A., Giró-i-Nieto, X., Vázquez, D., Rodríguez, P., 2021. Seasonal Contrast: Unsupervised Pre-Training from Uncurated Remote Sensing Data. *CoRR*, abs/2103.16607.

Mnih, V., Hinton, G., 2010. Learning to detect roads in high-resolution aerial images. *Computer Vision – ECCV 2010*, 6316, 210–223.

Peng, X., Sun, B., Ali, K., Saenko, K., 2014. Exploring Invariances in Deep Convolutional Neural Networks Using Synthetic Images. *CoRR*, abs/1412.7122.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

Taylor, L., Nitschke, G., 2018. Improving deep learning with generic data augmentation. *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1542–1547.

Wu, R., Yan, S., Shan, Y., Dang, Q., Sun, G., 2015. Deep Image: Scaling up Image Recognition.

Yan, Y., Tan, Z., Su, N., 2019a. A Data Augmentation Strategy Based on Simulated Samples for Ship Detection in RGB Remote Sensing Images. *ISPRS International Journal of Geo-Information*, 8(6).

Yan, Y., Zhang, Y., Su, N., 2019b. A Novel Data Augmentation Method for Detection of Specific Aircraft in Remote Sensing RGB Images. *IEEE Access*, 7, 56051–56061.

Zhang, Z., Liu, Q., Wang, Y., 2018. Road Extraction by Deep Residual U-Net. *IEEE Geoscience and Remote Sensing Letters*, 15(5), 749–753.

Zhu, X. 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 Geoscience and Remote Sensing Magazine*.