

LAND COVER SEGMENTATION WITH SPARSE ANNOTATIONS FROM SENTINEL-2 IMAGERY

Marco Galatola¹, Edoardo Arnaudo^{1,2}, Luca Barco¹, Claudio Rossi¹, Fabrizio Dominici¹

1. LINKS Foundation, *AI, Data & Space (ADS)*, Torino (TO), Italy
2. Politecnico di Torino, *Dipartimento di Automatica e Informatica (DAUIN)*, Torino (TO), Italy

ABSTRACT

Land cover (LC) segmentation plays a critical role in various applications, including environmental analysis and natural disaster management. However, generating accurate LC maps is a complex and time-consuming task that requires the expertise of multiple annotators and regular updates to account for environmental changes. In this work, we introduce SPADA, a framework for fuel map delineation that addresses the challenges associated with LC segmentation using sparse annotations and domain adaptation techniques for semantic segmentation. Performance evaluations using reliable ground truths, such as LUCAS and Urban Atlas, demonstrate the technique’s effectiveness. SPADA outperforms state-of-the-art semantic segmentation approaches as well as third-party products, achieving a mean Intersection over Union (IoU) score of 42.86 and an F1 score of 67.93 on Urban Atlas and LUCAS, respectively.

Index Terms— machine learning, computer vision.

1. INTRODUCTION

Land Cover (LC) segmentation plays a crucial role in various applications, including urban analysis and natural disaster management [1]. However, the manual production of accurate maps is a time-consuming activity that often requires several expert annotators. Additionally, regular updates are necessary to account for environmental changes. In natural disaster management, such information is crucial for studying the propagation and impact of disasters such as wildfires and floods, requiring the differentiation of flammable areas (forests, shrubs) from urban borders (buildings, roads).

However, generating efficient and reliable LC maps introduces several unique challenges that needs to be addressed to achieve accurate results. Existing EU-wide open LC datasets typically exhibit lower spatial resolution than required by some applications, while high-resolution products are often sparse and limited in terms of classification taxonomy.

To address these challenges, we selectively combine existing Copernicus¹ datasets, namely Corine Land Cover (CLC), Urban Atlas (UA), and Land Use and Coverage Area frame Survey (LUCAS), taking advantage of the strength of each data source. To cope with the sparse ground truth, we propose a novel framework called *SParse Annotations with DAformer* (SPADA). Leveraging Unsupervised Domain Adaptation (UDA) techniques, we propose a *teacher-student* framework, where the teacher model generates robust pseudo-labels to expand the annotations across the full input space. Similar to DAFormer[2], we mix the pseudo-labels, filtered and weighted by their prediction confidence, with the processed sparse ground truth to overcome feedback loops during self-training. We compare our solution with standard semantic segmentation approaches and third-party products, including the Sentinel-2 Global Land Cover (S2GLC), achieving a mean IoU score of 42.86 and F1 Score of 67.93 on UA and LUCAS, respectively. SPADA, without multi-temporal source images or post-processing, outperforms the strongest baselines by +7.88 mean IoU and +3.86 F1 on LUCAS.

In short, our contributions can be summarized as follows: first, we propose and evaluate the capability of SPADA, a novel framework for creating fuel maps from Sentinel-2 using vision transformers and exploiting a UDA technique leveraging on labelled and unlabeled pixels during training. Second, we release the dataset and the code used in this work, comprising the input data and sparse annotations used to train the segmentation model. ²

2. RELATED WORK

In aerial and remote sensing, semantic segmentation is applied to various target environments such as urban areas [1], land cover [3], and agricultural scenarios [4]. Supervised semantic segmentation from remotely sensed imagery presents several challenges, including the high number of input bands, the image size, the top-down viewpoint, and limited ground truth availability. To increase the segmentation performances, extra bands are typically included by introducing multiple encoders, or by expanding the input layers [4], while the

This publication is part of the project NODES which has received funding from the MUR – M4C2 1.5 of PNRR with grant agreement no. ECS00000036. This work is also part of the H2020 projects: SAFERS (GA n.869353) and OVERWATCH (GA n.101082320).

¹www.copernicus.eu

²Code and dataset: <https://github.com/links-ads/spada>

large input dimensions and the top-down viewpoint can be exploited to implement additional regularization, considering multiscale regularization [5] or invariance to rotation [4], both at training or test time. In critical tasks, the presence of sparsely annotated dataset poses additional challenges to achieving acceptable performances and can be addressed with weakly-supervised approaches, which rely on less precise annotations. In this context, Class Activation Maps (CAM) or attention maps have been effectively used [6], propagating labels from discriminative image regions. Also, approaches like AffinityNet [6] and its variants [7] predict semantic affinities between adjacent image portions, penalizing regions with different semantics. Sparse annotations are often addressed using scribbles, which provide efficient but approximate labels [8]. Similar to weak supervision, the goal is to expand the sparse ground truth to every pixel of the object, ensuring semantic consistency. Approaches like ScribbleSup [8] employ graph propagation. Tree Energy Loss [9] utilizes a minimum spanning tree among pixels for pairwise affinities, while FESTA [1] leverages an unsupervised neighbourhood loss. In this work, we adapt the self-training UDA methods [2], substituting the concept of source inputs with pairs of images and sparse labels, and target inputs with the same images with scribbles mixed with pseudo-labels, as detailed in Sec. 4.

3. DATASET

We train and validate our framework using a combination of several datasets. Specifically, we merge different Copernicus datasets, as detailed next. First, Sentinel-2 L2A cloud-free mosaics, using the whole 12-band input. Second, Corine Land Cover (CLC), which is a pan-European land cover classification dataset that provides information on land cover and land use. Third, the Land Use and Coverage Area frame Survey (LUCAS), a European-wide survey that collects data on land use and land cover across the continent. LUCAS provides only point-wise annotations, but the available areas have been manually validated, and therefore it is useful for both training and validation purposes. Fourth, Urban Atlas (UA), another European land cover and land use dataset localized around large urban areas, featuring higher spatial resolution, albeit with a reduced number of classes. Given the precise delineations offered by this source, we exploit it for training and validation purposes. Fifth, the Dominant Leaf Type High-Resolution Layer (HRL), which provides information about the dominant leaf type across Europe. Because our focus is on the production of fuel maps, we focus our study on the Mediterranean area, where wildfires are more frequent and intense. The ground truth consists of two annotations for each training region: a *scribble* label, comprising a sparse fuel map derived from CLC, and a *point-wise* label, derived from LUCAS. These annotations are the result of the following pre-processing pipelines. Both the *scribble* and the *point-wise*

labels are obtained by mapping the original classes into our fuel map taxonomy, as described in Table 1.

Following the methodology used to produce the S2GLC dataset [3], we apply to the remapped CLC classes a filtering process using Normalized Difference Vegetation Index (NDVI) and Normalized Difference Water Index (NDWI) thresholds, eliminating potentially mislabeled pixels. Next, we transform the filtered CLC classes into scribbles through morphological skeletonization followed by a small buffering of 5 pixels to increase their thickness. Finally, urban category labels from CLC are replaced with more accurate UA labels, while we use the HRL dataset to differentiate wooded areas into coniferous and broadleaf forests, providing a more detailed categorization. To generate the *point-wise*, we rasterize the LUCAS points within the considered areas, assigning the LUCAS class to the closest fuel class. These preprocessing steps generate a sparse ground truth with detailed fuel type information.

4. METHOD

4.1. Problem statement

We investigate a semantic segmentation task for fuel mapping in the presence of sparse annotations, a situation where only a subset of the pixels in an image are annotated with their corresponding class label, and the rest are left unmarked. Let us define as \mathcal{X} the set of multi-spectral input images, where each image x is constituted by a set of pixels \mathcal{I} , and as \mathcal{Y} the set of semantic annotations associating a class from the label set \mathcal{C} to each pixel $j \in \mathcal{I}$, where $|\mathcal{J}| \ll |\mathcal{I}|$. As described in Sec. 3, we have two sets of sparsely-annotated maps: (i) a set of *scribble* annotations, denoted as Y_S , and (ii) a set of *point-wise* annotations, denoted as Y_P . The goal is to find a parametric function f_θ that maps a multi-spectral image to a pixel-wise probability, i.e., $f_\theta : X \rightarrow \mathcal{R}^{|\mathcal{I}| \times |\mathcal{C}|}$, and evaluate it on unseen images. The parameters of the model θ are tuned to minimize a sum of two different categorical cross-entropy losses, namely $L_{seg} = L_S(\hat{y}, y_M) + \lambda L_P(\hat{y}, y_P)$, where \hat{y} is the predicted label, λ represents a weighting factor, while y_M represents the ground truth derived from mixing scribbles with expanded pseudo-labels, as detailed in Sec. 4.2.

4.2. Framework

SPADA is based on DAFormer, a self-training UDA framework which consists of a transformer-based encoder with a multilevel context-aware decoder. We adapt this UDA framework by substituting the concept of target domain with sparse labels. In this case, source and target data belong to the same image, where a small portion is provided with ground truth labels, and the remaining pixels remain unlabeled. We simplify the original framework by substituting RCS with simple class weights without performance loss, and by removing FD due

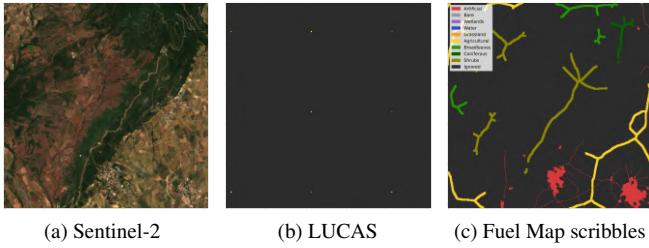


Fig. 1: Input image and corresponding annotations extracted from the final fuel map dataset.

Fuel Class	CLC Class ID	LUCAS Class ID	Color
Artificial	111 112 121 122 123 124 131 132 133 142	7	[Red]
Bare	331 332 335	6	
Wetlands	411 412 421 422 423	8 9	[Blue]
Water	511 512 521 522 523		
Grassland	211 231 321	3	[Yellow]
Agricultural	212 213 221 222 223 241 242 243 244	1 2	[Orange]
Broadleaves	311	4	[Green]
Coniferous	312	4	[Light Green]
Shrubs	322 323 324 333	5	[Light Blue]
Ignored	141 313 334 999		[Grey]

Table 1: Mapping and aggregation carried out for CLC and LUCAS class IDs.

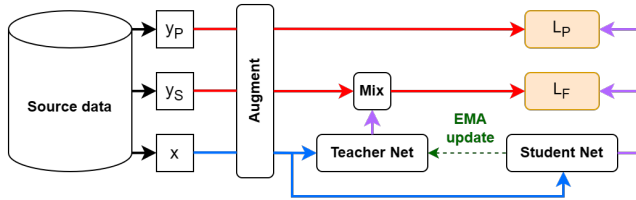


Fig. 2: Overview of SPADA framework.

to its inapplicability to land cover classes. The SPADA framework is composed of three main blocks: (i) a *student* network, trained on a mix of ground truth scribbles and dense pseudo-labels, and regularized using the *point-wise* annotations, (ii) a *teacher* network, obtained as EMA from the student model, that generates pseudo-labels from Sentinel-2 inputs in a robust and consistent way. Lastly, (iii) a *label mixing* strategy between scribbles and these pseudo-labels.

The final training labels are obtained in two steps: first, the pseudo-labels are filtered based on a fixed confidence threshold, second the scribbles are fused on top of the remaining labels for consistency. Formally, each mixed label y_M is obtained as composition of $\hat{y}_T \odot y_S$, where the \hat{y}_T represents the pseudo-labels inferred by the teacher model, and y_S identifies the *scribble* annotations. In order to avoid overconfident predictions, we include a weight on the mixed labels $w_i \in [0, 1]$ for each pixel i , where $w_i = 1$ if the pixel belongs to y_S , or $w_i = |\hat{\mathcal{I}}|/|\mathcal{I}|$ if i belongs to \hat{y}_T . Here, $\hat{\mathcal{I}}$ is the set of pseudo-label pixels above a given threshold τ .

5. EXPERIMENTS

To cope with limited computational resources, we perform our evaluation over a total surface of $472,717 \text{ km}^2$, encompassing the south European countries that have been most affected by wildfires in the last three years (i.e., Portugal, Spain, France, Italian regions, Greece, and Balkans). We use as input the Sentinel-2 cloudless mosaics, computed from April to September 2018. We select this interval to match the ground truth used (i.e., CLC, UA, VHR, LUCAS) and to enable a meaningful performance comparison with state-of-

the-art products such as S2GLC. We split the dataset into 12 equivalently sized areas, selecting 8 areas for training, and 4 for testing. Training areas are further split into training and validation sections of size $2,048 \times 2,048$ pixels, keeping 90% and 10% for training and validation respectively. All data is then tiled into 512×512 chips. The final set consists of 20,398 tiles for training, 5100 for validation, and 394 full sections for testing that are divided into tiles at runtime. We test our solution against semantic segmentation baselines and third-party products, namely the S2GLC [3] and the original Corine Land Cover. All the baseline models are trained on the fuel maps without sparse annotations, while CLC and S2GLC are only remapped to match the considered fuel classes. Given the lack of dense annotations, we assess the performance of our solution on test areas using the two most reliable ground truths: LUCAS, using F1 score, and Urban Atlas, by means of the Intersection over Union (IoU) metric. Each model is trained for 160,000 iterations with an AdamW optimizer and a polynomial scheduler with a linear warm-up. We further augment the inputs using horizontal and vertical flips, affine transforms, Gaussian blur, and we exploit test-time augmentations to improve the inference quality further. We first evaluate the classification abilities of our system, comparing it to the manually validated LUCAS points. The results, listed in Table 2, show a consistent improvement of our model over all tested baselines, including Segformer, achieving +3.86 increment in terms of F1 score against S2GLC. In Table 3, we report the results in terms of IoU over the considered UA regions available in the test set. While specific categories (e.g., *bare*, *grassland*) are slightly underperforming, SPADA obtains on average comparable or substantially higher results on the other classes, with an IoU increment of +7.88 w.r.t. S2GLC. Additionally, we include in both tables the performances computed on the raw CLC layers, which is the reference land cover product in Europe, showing that SPADA achieves better performances by a large margin.

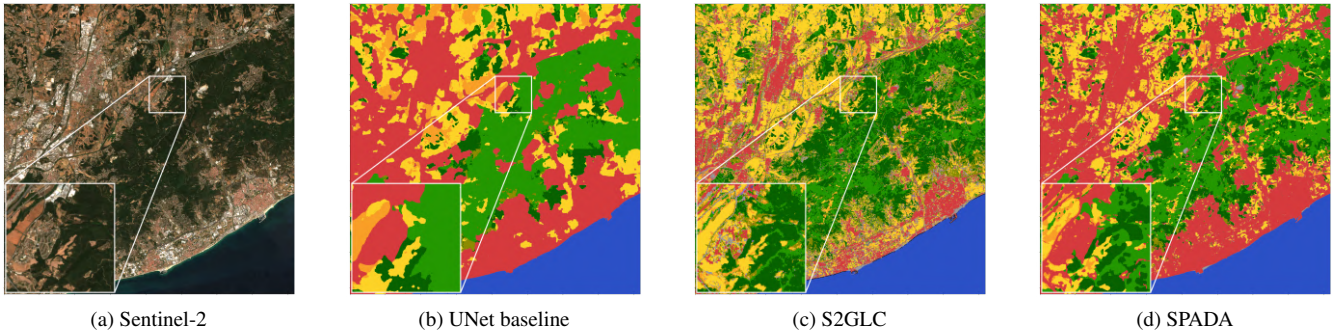


Fig. 3: Qualitative comparison between outputs, from left to right: S2 input, UNet baseline, S2GLC, SPADA. Best viewed zoomed in.

Method	Agricultural	Grassland	Broadleaves	Coniferous	Shrubs	Bare	Artificial	Water	Avg	Acc
CLC	55.72	24.71	66.49	68.51	33.55	60.63	52.71	63.40	55.53	52.54
UNet	57.94	28.40	72.30	68.33	34.72	34.47	55.98	73.01	58.00	55.62
OCRNet	59.39	26.25	70.71	65.17	33.77	33.81	53.88	68.73	56.87	54.38
PSPNet	58.51	18.69	66.38	64.56	31.66	52.87	52.88	68.18	55.04	52.10
DeepLabV3Plus	60.64	17.30	67.56	63.62	32.25	55.45	54.39	67.06	55.84	53.42
SegFormer	64.09	14.83	72.65	68.38	33.56	56.08	49.45	64.59	58.63	57.25
S2GLC	69.39	36.06	75.22	70.60	34.15	61.62	55.71	75.99	64.07	62.14
SPADA (Ours)	77.72	39.36	76.78	74.19	38.08	63.64	58.54	70.49	67.93	66.99

Table 2: Experiments on LUCAS test set (F1 score).

Method	Artificial	Bare	Wetlands	Water	Grassland	Agricultural	Forest	mIoU	mAcc
CLC	58.86	14.78	32.44	58.68	14.62	36.56	45.69	37.37	49.81
UNet	51.06	1.83	25.10	32.00	11.77	39.88	53.67	30.76	49.04
OCRNet	53.15	2.39	34.36	30.79	12.57	39.34	50.59	31.88	48.15
PSPNet	55.12	3.85	34.12	30.68	11.81	38.72	46.57	31.55	47.64
DeepLabV3Plus	54.09	5.6	30.06	31.25	10.74	41.31	48.45	31.64	47.55
SegFormer	59.06	13.01	24.46	52.9	8.52	43.56	52.17	36.24	53.82
S2GLC	40.55	7.12	4.00	66.97	14.19	46.09	65.95	34.98	52.49
SPADA (Ours)	64.36	13.56	27.27	65.95	10.01	54.3	64.56	42.86	58.11

Table 3: Experiments on Urban Atlas test set (IoU).

6. CONCLUSIONS

Exploiting a curated set of sparse annotations, we build an ad-hoc dataset for fuel map segmentation. We then propose SPADA, a framework for sparsely annotated semantic segmentation inspired by UDA techniques. We perform an extensive performance evaluation of our framework over a wide area in Europe, showing that our solution outperforms both semantic segmentation baselines and existing land cover products such as CLC and S2GLC. Future works will focus on expanding the available data with a wider range of geographical areas and modalities, as well as improving the methodology with ad-hoc refinements over the pseudo-label generation.

7. REFERENCES

- [1] Yuansheng Hua, Diego Marcos, Lichao Mou, Xiao Xiang Zhu, and Devis Tuia, “Semantic segmentation of remote sensing images with sparse annotations,” *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.
- [2] Lukas Hoyer, Dengxin Dai, and Luc Van Gool, “Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9924–9935.
- [3] Radek Malinowski, Stanisław Lewiński, Marcin Rybicki, Ewa Gromny, Małgorzata Jenerowicz, Michał Krupiński, Artur Nowakowski, Cezary Wojtkowski, Marcin Krupiński, Elke Krätzschar, and Peter Schauer, “Automated production of a land cover/use map of europe based on sentinel-2 imagery,” *Remote Sensing*, vol. 12, no. 21, 2020.
- [4] Antonio Tavera, Edoardo Arnaudo, Carlo Masone, and Barbara Caputo, “Augmentation invariance and adaptive sampling in semantic segmentation of agricultural aerial images,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1656–1665.
- [5] Wuyang Chen, Ziyu Jiang, Zhangyang Wang, Kexin Cui, and Xiaoning Qian, “Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8924–8933.
- [6] Jiwoon Ahn and Suha Kwak, “Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4981–4990.
- [7] Adrien Nivaggioli and Hicham Randrianarivo, “Weakly supervised semantic segmentation of satellite images,” in *2019 Joint urban remote sensing event (JURSE)*. IEEE, 2019, pp. 1–4.
- [8] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun, “Scribblesup: Scribble-supervised convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3159–3167.
- [9] Zhiyuan Liang, Tiancai Wang, Xiangyu Zhang, Jian Sun, and Jianbing Shen, “Tree energy loss: Towards sparsely annotated semantic segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16907–16916.