Transfer Land Cover Maps Across Years: A Time Series-based Semantic Segmentation Approach

Christopher Jabea, Roberto Interdonato, Cassio F Dantas, Dino Ienco, Flavie Cernesson, Eric Barbe, Nadia Guiffant, Christiane Weber

Transfer Land Cover Maps Across Years: A Time Series-based Semantic Segmentation Approach

Christopher Jabea
INRAE, UMR TETIS
University of Montpellier
Montpellier, France
christopher.jabea@inrae.fr

Roberto Interdonato
CIRAD, Inria, UMR TETIS
University of Montpellier
Montpellier, France
roberto.interdonato@cirad.fr

Cassio F. Dantas
INRAE, Inria, UMR TETIS
University of Montpellier
Montpellier, France
cassio.fraga-dantas@inrae.fr

Dino Ienco
INRAE, Inria, UMR TETIS
University of Montpellier
Montpellier, France
dino.ienco@inrae.fr

Flavie Cernesson
AgroParisTech, UMR TETIS
University of Montpellier
Montpellier, France
flavie.cernesson@agroparistech.fr

Eric Barbe
INRAE, UMR TETIS
University of Montpellier
Montpellier, France
eric.barbe@inrae.fr

Nadia Guiffant
INRAE, UMR TETIS
University of Montpellier
Montpellier, France
nadia.guiffant@inrae.fr

Christiane Weber
CNRS, UMR TETIS
University of Montpellier
Montpellier, France
christiane.weber@cnrs.fr

***Abstract*—The widespread availability of satellite imagery data has enabled advancements in Land Use/Land Cover (LULC) and Urban Fabric (UF) mapping through deep learning. However, maintaining up-to-date urban land cover maps is challenged by the high cost and operational constraints of continuous field data collection. This study explores the feasibility of updating urban LULC maps using SITS-based semantic segmentation models trained on historical data, specifically examining a transfer scenario where a model trained on 2015 data is applied to 2020 imagery. We benchmark the performance of two convolution-based architectures (Unet and Unet3D), plus a recent spatio-temporal transformer-based approach (TSViT) and a proposed variant, named TSViT+SW, which incorporates a shifted window attention scheme. Experimental evaluations covering the urban area of Lyon, France, reveal that the proposed TSViT+SW model achieves the best results among transferred models, minimizing performance degradation compared to the ideal in-year training scenario. This work offers insights into the potential and limitations of using historical data to update urban land cover in the absence of fresh labeled data.**

I. INTRODUCTION

In recent years we've seen an increasing availability of Satellite Image Time Series (SITS) data, mainly due to the launch of space programs with open access imagery, such Copernicus and its Sentinel satellites, or NASA's Landsat collections. This vast amount of data, combined with the development of increasingly advanced deep learning methods, has enabled the improvement of methodologies developed for challenging tasks such as Land Use/Land Cover (LULC) [1], [2] and Urban Fabric (UF) [3], [4] mapping. Mapping urban land [5] and urban dynamics [6] is fundamental for modelling urban growth, which is crucial for understanding global environmental change and sustainable urban development. However, while it is well know how deep learning methodologies tend to be data-greed, performing field campaigns on a regular basis in order to get up-to-date labels is often unfeasible, for reasons related to economical costs, time availability and operational constraints.The ability to generate up-to-date LULC maps of urban and peri-urban areas using models trained on previous years' labeled data would support associated tasks with minimal data-collection cost. In this work, we evaluate the possibility to train SITS-based semantic segmentation approaches on annotated maps of urban areas for updating land cover information, focusing on a transfer scenario where a model is trained on 2015 data and then used on 2020 data (*out-of-year* scenario). We compare the obtained classification with that resulting from a model trained and tested on the year 2020 (*in-year* scenario). We test four different methodologies, included a novel extension of the TSViT model [7], where we enhance the original model with the Shifted Window mechanism from the Swin Transformer model [8], namely TSViT+SW. Experimental results on the urban area of Lyon (France) demonstrate how, while transferring a model trained on 2015 data to 2020 data generally results in degraded performance (i.e., compared to a model trained and tested on 2020 data), the proposed TSViT+SW obtains the best absolute performance, also associated to the lower performance drop between the *in-year* and *out-of-year* scenarios.

This work was supported by ANR project GLOURB ANR-22-CE03-0005.

II. DATA

The study area encompasses Lyon, France's second-largest urban area, located northwest of the French Alps and covering a region of 150 km^2. This region features a dense urban area surrounded by agricultural crops, natural vegetation and mountain reliefs. Figure 1 shows the **G**round **T**ruth (GT) data superimposed on a RGB Landsat-8 image (taken the 5th May 2020). We collected SITS of Landsat-8 images covering the year 2015 and the year 2020. For both years, we consider the period from April to October. Satellite data are retrieved via the Microsoft Planetary Computer platform, allowing access

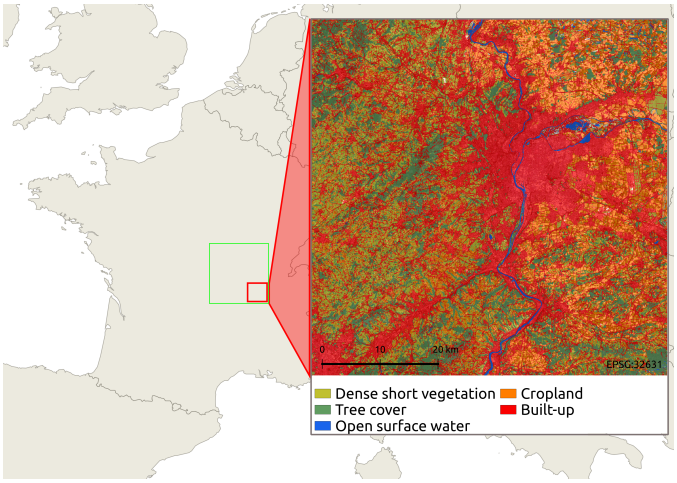

Fig. 1: Area of interest with a focus on Lyon

to Landsat-Collection 2 Level 2[1] products. Only 6 bands were considered in this study: Blue, Green, Red, Near-Infrared and the two Short-wave infrared (swir16 and swir22). All bands are at spatial resolution of 30m. We perform time series gap filling of cloudy pixels, detected via the QA-pixel band, using multi-temporal linear interpolation [9]. Gap-filled images were generated at a regular 16-day frequency, resulting in a sequence of 14 images for each year. Figure 2 shows the histograms of the NDVI (Normalized Difference Vegetation Index) and the IBI (Intelligent Building Index) indices for the year 2015 and 2020. While the former index is commonly used to monitor vegetation and crop , the latter is tailored to characterize and built-up surfaces [10]. We can note some differences between the two years indicating possible distribution shifts for the study area between 2015 and 2020. As GT data we consider the **G**lobal **L**and **A**nalysis and **D**iscovery (GLAD) dataset [11], providing LULC maps each 5 years between the period 2000 and 2020. This dataset provides up to three levels of pixel classification classes from coarser macro classes to a fine quantified estimation of a metric for an intermediate class (e.g. Terra Firma $\rightarrow$ Tree cover $\rightarrow$ Trees height). We choose to rely on the macro classes level except for the *Terra Firma* classes where we choose to take the intermediate level of description. The nomenclature and the statistics associated to our study area for 2015 and 2020 are reported in Table I.

Land Cover	2015		2020	
	%	# Pixels	%	# Pixels
Dense short vegetation	28.42	1.11×10^7	27.06	1.06×10^7
Tree cover	26.13	1.02×10^7	26.04	1.02×10^7
Open surface water	0.92	3.61×10^5	0.85	3.30×10^5
Cropland	14.58	5.70×10^6	15.57	6.08×10^6
Built-up	29.95	1.17×10^7	30.48	1.19×10^7

TABLE I: 2015 and 2020 classes distribution

[1]https://planetarycomputer.microsoft.com/dataset/landsat-c2-l2

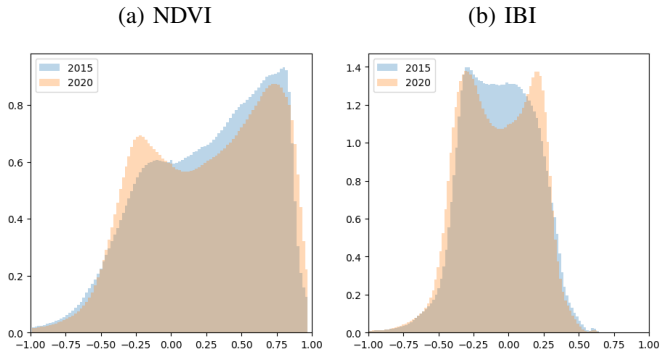

Fig. 2: Distribution histogram of the NDVI and IBI indices on the 2015 and 2020 data.

III. Semantic Segmentation of SITS data

Semantic segmentation has been widely adopted for analyzing single-date Very High spatial Resolution (VHR) remote sensing data [12] with applications to segmentation [13], urban characterization [14], informal settlement identification [15], and more broadly land cover mapping [12]. However, only a few methods have been proposed to address the semantic segmentation of SITS data, primarily evaluated for agricultural and crop type mapping [7].

A first approach for SITS-based semantic segmentation was proposed in [16], where a combination of recurrent and convolutional neural networks was used to handle spatio-temporal information. Subsequent studies [17] have applied the popular UNet model [18] to SITS data, often discarding temporal information and considering the entire multi-temporal radiometric information at once. In [19], a 3D extension of the UNet method (UNet3D) was employed for SITS-based semantic segmentation, using 3D convolutions to explicitly combine spatial and temporal information, thereby modeling temporal dynamics. Recently, a Vision Transformer framework for SITS-based semantic segmentation, referred to as TSViT, was introduced in [7]. This approach leverages the flexibility of the transformer model and attention mechanism to first analyze the temporal information in the SITS data and then spatially combine this information to provide per-pixel dense prediction.

Here, we will briefly review the SITS-based semantic segmentation methods considered in our experimental evaluation:
UNet [18]: A widely adopted semantic segmentation approach used in domains as biomedical imaging, autonomous vehicles, and remote sensing. It is based on a symmetric convolutional neural network autoencoder architecture with skip connections. Originally designed for standard image analysis, it does not model the time dimension.
UNet3D [17]: An extension of the UNet architecture for spatio-temporal data. It uses 3D convolution blocks with small kernels ($3\times3\times3$). The encoder consists of five 3D convolution blocks with a downsampling operator every two blocks to reduce the size of the feature maps. The decoder has a similar structure, employing transposed convolutions to recover the

original spatial resolution. UNet3D also leverages skip connections to preserve spatial information and avoid vanishing gradient issues.

TSViT [7]: A recent transformer-based architecture designed for semantic segmentation of SITS data. TSViT uses attention-based transformers to manage both temporal and spatial information. First, a temporal-transformer module analyzes the temporal information. Then, a spatial-transformer module aggregates the extracted information to generate the final dense classification. The use of a global attention strategy in the spatial aggregation stage permits information that are far away to interact together. This can introduce noise and uncorrelated information, potentially reducing accuracy.

TSViT+SW: An extension of the TSViT model introduced in this work. We enhance the TSViT model with the Shifted Window mechanism from the Swin Transformer model [8]. Instead of using the original spatial-transformer proposed in [7], for the second stage of the TSViT model, we employ a Swin Transformer block without patch merging, with a comparable number of parameters. The Swin Transformer block alternates between global and localized attention strategies, limiting the receptive field of the attention mechanism. The patch merging mechanism is discarded to avoid spatial information down-sampling, which is critical for remote sensing data analysis.

IV. Experiments

Experimental Settings: To evaluate the competing approaches introduced in Section III (UNet, UNet3D, TSViT and TSViT+SW) we consider a transfer scenario where a model is trained on 2015 data and then used on 2020 data (2015 $\rightarrow$ 2020). We name this scenario *out-of-year*. We also consider a scenario where a model is trained and test on the same year, 2020, with the aim of providing a reference for 2020 (2020 $\rightarrow$ 2020). We name this scenario *in-year*. In both scenarios, the same subset of data from 2020 is considered as the test set, corresponding to the 20% of the available data. The values of the SITS data were rescaled between 0 and 1 per year and per band considering the 2nd and 98th percentile of the data distribution as minimum and maximum values. Both 2015 and 2020 data are rescaled using statistics from the former year, for comparability reasons. All the methods are tackling the problem as a semantic segmentation task where dense prediction is provided as outcome. For this reason, we divided the study area in non overlapping patches of 64$\times$64 pixels. For each patch, two information are available: i) the multi-temporal remote sensing data cube covering an area of 64$\times$64 pixels (1 920m$\times$1 920m), 14 timestamps and 6 bands and ii) the GT map with the same spatial extent containing for each pixel the class information. This results in 9 600 patches per year. For the UNet model we used as backbone the efficientnet-b7 [20] convolutional neural network. Additionally, since the UNet model is not designed to explicitly manage the temporal dimension, we concatenate the last two dimensions of each patch to obtain, for each sample, an image with 84 bands (6$\times$14) to fit the required input format. For UNet3D and TSViT we used the implementations provided by [7] while for TSViT+SW, starting from the original TSViT implementation, we introduce the shifted window mechanism as previously mentioned. For TSViT and TSViT+SW we adopt a token dimension of 2$\times$2. All the approaches are trained for 100 epochs using the Adam optimizer, a batch size of 4 and the same learning rate schedule procedure adopted in [7], where a warmup period starting from zero to a maximum value of 10^{-3} is used for the first 10 epochs, followed by a cosine learning rate decay down to 5×10^{-6}. All the results are evaluated by means of Micro F1-Score, hereafter referred to as F1-Score.

MODEL	(2015 $\rightarrow$ 2020)	(2020 $\rightarrow$ 2020)	Drop
UNet	67.7	79.9	12.2
UNet3D	70.2	81.7	11.5
TSViT	69.1	81.3	12.2
TSViT+SW	**75.2**	**81.7**	6.5

TABLE II: Results in terms of F1-Score on the *out-of-year* (2015 $\rightarrow$ 2020) scenario.

Results: Table II reports on the F1-Score performances for all competing approaches on the two scenarios: *out-of-year* (2015 $\rightarrow$ 2020) and *in-year* (2020 $\rightarrow$ 2020). As expected, we can observe a systematic drop in classification performances for all the approaches between *in-year* and *out-of-year* scenarios, with performance drops ranging from 6.5 to 12.2 points. The UNet baseline exhibits the overall lowest performances in both scenarios, underlying the importance to explicitly model the temporal information. The proposed TSViT+SW method turns out to be the best performing method on both the *in-year* (81.7) and *out-of-year* (75.2) scenario, also resulting in the lowest performances loss among all the considered approaches.

Figure 3 depicts the per class F1-score of the four approaches on the *out-of-year* (2015 $\rightarrow$ 2020) scenario. TSViT+SW achieves systematical improvements compared to the original TSViT model. It is also the only approach that outperforms the UNet baseline over all the land cover classes, thus demonstrating its superior ability in leveraging spatio-temporal information simultaneously. Focusing on the built-up land cover class, we can note that methods that explicitly model the time information (UNet3D, TSViT and TSViT+SW) outperform the UNet baseline of at least 7.0 points of F1-Score. An extract of the land cover maps generated by the three models with the highest performances as well as the ground truth is shown in Figure 4. It shows that the UNet3D struggles more than the other competitors with the *Cropland* class and that the TSViT+SW allows a better boundary between the *Dense short vegetation* and *Tree cover* classes. Both of these observations can be linked to the respective performances of the models observed in Figure 3. In the *Built-up* class, the TSViT+SW showed a slight improvement over the TSViT, especially on roads.

V. Conclusion

In this study, we evaluated recent semantic segmentation approaches for SITS data to leverage densely annotated historical maps of urban areas for updating land cover information.

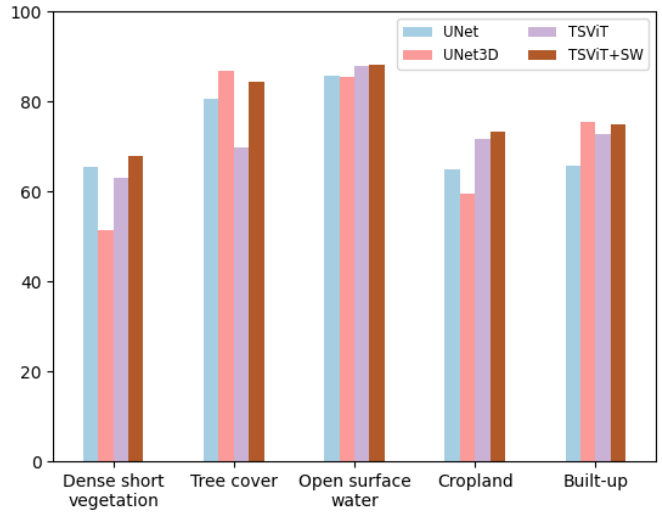

Fig. 3: F1-score per class for the *out-of-year* scenario.

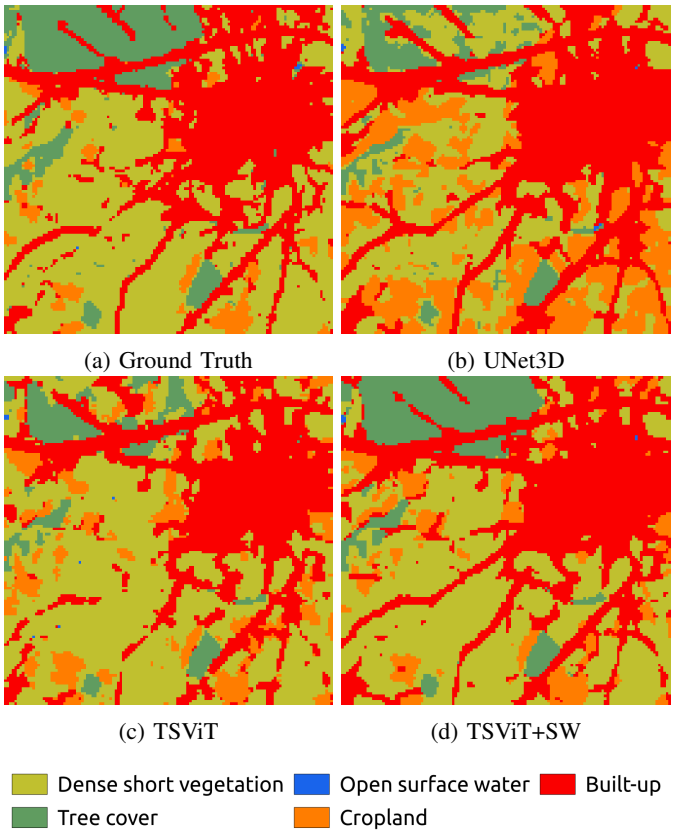

Fig. 4: Detail from the land cover maps obtained with the three best performing methods, for the *out-of-year* scenario.

We introduced TSViT+SW, an extension of the TSViT approach that incorporates the shifted window mechanism from the Swin Transformer model. Our results have shown that transferring a model trained on 2015 data to 2020 data (out-of-year) for the urban area of Lyon, France, results in degraded performances compared to a model trained and tested on 2020 data (in-year). This performance degradation is likely due to data shifts between different years. These results highlight the need for tailored strategies to deal with data distribution shifts when analysing multiple years of remote sensing data.

References

[1] R. Interdonato, D. Ienco, R. Gaetano, and K. Ose. Duplo: A dual view point deep learning architecture for time series classification. *ISPRS J. Photogramm. Remote Sensing*, 149:91–104, 2019.

[2] D. Ienco, R. Gaetano, and R. Interdonato. A constrastive semi-supervised deep learning framework for land cover classification of satellite time series with limited labels. *Neurocomp.*, 567, 2024.

[3] L. El Mendili, A. Puissant, M. Chougrad, and I. Sebari. Towards a multi-temporal deep learning approach for mapping urban fabric using sentinel 2 images. *Remote Sensing*, 12(3), 2020.

[4] R. Wenger, A. Puissant, J. Weber, L. Idoumghar, and G. Forestier. U-net feature fusion for multi-class semantic segmentation of urban fabrics from sentinel-2 imagery: an application on grand est region, france. *Int. J. of Remote Sensing*, 43(6):1983–2011, 2022.

[5] X. Liu, G. Hu, Y. Chen, X. Li, X. Xu, S. Li, F. Pei, and S. Wang. High-resolution multi-temporal mapping of global urban land using landsat images based on the google earth engine platform. *Remote Sensing of Env.*, 209:227–239, 2018.

[6] X. Li, Y. Zhou, Z. Zhu, L. Liang, B. Yu, and W. Cao. Mapping annual urban dynamics (1985–2015) using time series of landsat data. *Remote Sensing of Env.*, 216:674–683, 2018.

[7] M. Tarasiou, E. Chavez, and S. Zafeiriou. Vits for sits: Vision transformers for satellite image time series. In *IEEE/CVF CVPR*, pages 10418–10428, 2023.

[8] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y.Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF CVPR*, pages 10012–10022, 2021.

[9] J. Inglada, M. Arias, B. Tardy, O. Hagolle, S. Valero, D. Morin, G. Dedieu, G. Sepulcre, S. Bontemps, P. Defourny, and B. Koetz. Assessment of an operational system for crop type map production using high temporal and spatial resolution satellite optical imagery. *Remote Sensing*, 7(9), 2015.

[10] R. Estoque and Y. Murayama. Classification and change detection of built-up lands from landsat-7 etm+ and landsat-8 oli/tirs imageries: A comparative assessment of various spectral indices. *Ecological Indicators*, 56:205–217, 2015.

[11] P. Potapov, M. Hansen, A. Pickens, A. Hernandez-Serna, A. Tyukavina, S. Turubanova, V. Zalles, X. Li, A. Khan, F. Stolle, et al. The global 2000-2020 land cover and land use change dataset derived from the landsat archive: first results. *Front. in Rem. Sens.*, 3:856903, 2022.

[12] L. Huang, B. Jiang, S. Lv, Y. Li, and Y. Fu. Deep learning-based semantic segmentation of remote sensing images: A survey. *IEEE JSTARS*, 2023.

[13] J. Li, Y. Hu, and X. Huang. Casaformer: A cross-and self-attention based lightweight network for large-scale building semantic segmentation. *Int. J. Appl. Earth Obs. Geoinf.*, 130:103942, 2024.

[14] A. Maiti, S. 0. Elberink, and G. Vosselman. Uavpal: A new dataset for semantic segmentation in complex urban landscape with efficient multi-scale segmentation. *IEEE JSTARS*, 2023.

[15] T. Stark, M. Wurm, X. Zhu, and H. Taubenböck. Satellite-based mapping of urban poverty with transfer-learned slum morphologies. *IEEE JSTARS*, 13:5251–5263, 2020.

[16] M. Rußwurm and M. Körner. Multi-temporal land cover classification with sequential recurrent encoders. *ISPRS int. j. geo-inf*, 7(4):129, 2018.

[17] R. Rustowicz, R. Cheong, L. Wang, S. Ermon, M. Burke, and D. Lobell. Semantic segmentation of crop type in africa: A novel dataset and analysis of deep learning methods. In *IEEE/CVF CVPR Workshops*, pages 75–82, 2019.

[18] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015.

[19] M. Tarasiou, R. Güler, and S. Zafeiriou. Context-self contrastive pretraining for crop type semantic segmentation. *IEEE TGRS*, 60:1–17, 2022.

[20] M. Tan and Q. V.Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*. PMLR, 2019.