SPARSELY LABELED LAND COVER CLASSIFICATION WITH OVERSEGMENTATION-BASED GRAPH U-NETS

Johannes Leonhardt
Institute of Geodesy and Geoinformation
University of Bonn
Bonn, Germany
jleonhardt@uni-bonn.de

Ribana Roscher
Institute of Bio- and Geosciences (IBG-2)
Forschungszentrum Jülich GmbH
Jülich, Germany
r.roscher@fz-juelich.de

ABSTRACT

Training neural networks for large-scale land cover classification from satellite imagery requires extensive labels for training and evaluation. While most methods are designed around dense annotations, another promising idea is to rely on sparse labels, such as openly available in-situ data. However, these data pose challenges in terms of model design and training. In this paper, we present a specially designed neural network architecture for sparsely labeled land cover classification from Sentinel-2 images and LUCAS data. Our network is a variant of Graph U-Net which represents images as graphs and uses transformer-inspired graph convolutional layers and pooling layers based on hierarchical image oversegmentations. Additionally, we adapt deep bilateral filtering modules to this architecture. In our experiments, we demonstrate that our network is able to learn from sparse labels more efficiently than traditional approaches, outperforming standard U-Nets.

1 INTRODUCTION

Land cover, i.e., spatial information on the Earth surface's biophysical properties, is essential for a large number of scientific and practical fields, including climate science, regional planning, and disaster management. To obtain large-scale land cover maps, researchers and practitioners have relied on the automated interpretation of satellite images using deep learning methods. In this context, sparsely labeled classification, where the true label of only very few pixels is known, is of particular interest, as dense labels are expensive to obtain for large areas.

In this paper, we use labels from Eurostat's in-situ land use/cover area frame statistical survey (LUCAS) to predict land cover from multispectral Sentinel-2 imagery for the year 2018. LUCAS data has previously been used for training machine learning models for general land cover mapping (Mack et al., 2017; Pflugmacher et al., 2019; Weigand et al., 2020; Mirmazloumi et al., 2022) or with a special focus on crop type mapping (Conrad et al., 2010; d'Andrimont et al., 2021). However, the problem of label sparsity is typically circumvented by using either pixel-wise or segment-wise hand-crafted features as model inputs. Our work, on the other hand, focuses on image-based deep learning methodologies that directly operate on images and are thus able to take into account context on varying spatial scales.

Specifically, we present a variant of Graph U-Net (Gao & Ji, 2019) using transformer-based graph convolutions (Shi et al., 2021) and pooling layers based on hierarchical oversegmentations using the Quickshift algorithm (Vedaldi & Soatto, 2008). By integrating clustering information directly into the network architecture, we incorporate methods from unsupervised learning into our supervised learning pipeline, enabling more efficient learning from sparse labels. A similar approach has previously been taken by Liu et al. (2022), who adapt a Graph U-Net to exploit hierarchical superpixels based on minimum spanning trees for pooling in graph neural networks to perform land cover classification based on hyperspectral imagery. While not the main focus of their work, experiments also demonstrate the model's general capabilities in a setting with artificially reduced labels. Specially focused on the task of sparsely labeled land cover classification, Maggiolo et al. (2022) use fully connected random fields to improve classification results from sparse annotations. Closely related,

Wu et al. (2022) present a deep bilateral filtering module to produce spatially smooth feature maps, which we also adapt to the proposed architecture in this study.

2 METHODOLOGY

2.1 MODEL ARCHITECTURE

We first reframe the original pixel-wise land cover classification task as a node classification task by representing the image as its pixel graph. More technically, the pixel graph of an image $\mathbf{X}$ is represented as a node matrix $\mathbf{V}_X$, which is simply computed by flattening the image's spatial dimensions, and a set of undirected edges, $\mathbf{E}$, which is initialized by connecting each node to itself and its 8 spatial neighbors with respect to the original grid topology of $\mathbf{X}$. The neural network classifier now operates directly on the graph and obtains a node-wise classification result: $\mathbf{V}_{\widehat{Y}} = \mathrm{f}_\theta((\mathbf{V}_X, \mathbf{E}))$. Finally, we can recover the original grid representation by simply rearranging $\mathbf{V}_{\widehat{Y}}$ into its original grid.

To propagate features between our Graph U-Net's layers, graph convolution layers substitute the image-based convolutions in the U-Net. In our model in particular, we use a graph convolution operator based on multi-head attention transformers, where each node's feature $\mathbf{x}_i \in \mathbf{V}_X$ is modulated based on its neighbors $\mathcal{N}(i)$ according to

$$\tilde{\mathbf{x}}_i = \mathbf{W}_r \mathbf{x}_i + \sum_{j \in \mathcal{N}(i)} a_{ij} \mathbf{W}_v \mathbf{x}_j \quad \text{with} \quad a_{ij} = \mathrm{softmax}\left(\frac{(\mathbf{W}_q \mathbf{x}_i)^T (\mathbf{W}_k \mathbf{x}_j)}{\sqrt{d}} \right), \tag{1}$$

where $\mathbf{W}_r$ and $\mathbf{W}_v$ are learnable matrices for the root node and the neighbor nodes and a_{ij} is the attention coefficient, calculated from the query and key matrices $\mathbf{W}_q$ and $\mathbf{W}_k$, as well as number of output dimensions d.

The main advantage of using Graph U-Nets over standard U-Nets in the context of sparsely labeled land cover classification, however, is the increased flexibility with regard to pooling. This is because, unlike in standard U-Nets, the pooling layers in Graph U-Net must not preserve the Euclidian topology of the original images. Thereby, they offer a suitable interface for incorporating clustering information into the network at various levels. In particular, we make use of the Quickshift oversegmentation algorithm, which uses a granularity parameter τ to control the number of segments, where a higher τ leads to fewer segments.

We apply Quickshift with different τ, resulting in K oversegmentations with decreasing granularity $\{\mathrm{Q}(\mathbf{X}, \tau_1), \mathrm{Q}(\mathbf{X}, \tau_2), ..., \mathrm{Q}(\mathbf{X}, \tau_K)\}$. In the network's pooling layers, features are averaged over all nodes within the same segment, such that each segment in $\mathrm{Q}(\mathbf{X}, \tau_k)$ is represented as a node in the pooled result $(\mathbf{V}_X, \mathbf{E})_k$. To perform multiple of these pooling operations consecutively, we make use of the strictly hierarchical nature of Quickshift, meaning that if two pixels are assigned to the same segment at any τ, they are guaranteed to also be assigned to the same segment for any larger τ. This way, $(\mathbf{V}_X, \mathbf{E})_k$ can unambiguously be assigned to new segments based on $\mathrm{Q}(\mathbf{X}, \tau_{k+1})$ for a subsequent pooling operation, and so on. As for the edges, two nodes in the pooled graph $(\mathbf{V}_X, \mathbf{E})_k$ share an edge, if the corresponding segments in $\mathrm{Q}(\mathbf{X}, \tau_k)$ are adjacent with respect to an 8-neighborhood. For unpooling a graph $(\mathbf{V}_X, \mathbf{E})'_k$ in the decoder, we reverse the pooling operation by first initializing a graph $(\mathbf{V}_X, \mathbf{E})'_{k-1}$ based on the topology induced by $\mathrm{Q}(\mathbf{X}, \tau_{k-1})$. Due to the hierarchical nature of the oversegmentations, we can again unambiguously assign a copy of a node's features in $(\mathbf{V}_X, \mathbf{E})'_k$ to each related node in $(\mathbf{V}_X, \mathbf{E})'_{k-1}$.

We furthermore adapt a deep bilateral filtering module to our model architecture. The goal of deep bilateral filtering is to increase homogeneity in intermediate feature maps without imposing significant blurring on edges. To make bilateral filtering, originally known from image processing, feasible for application inside deep learning architectures, the receptive field of the filter is bounded by replacing the Gaussian kernel from the original formulation with a uniform kernel. To perform deep bilateral filtering on a graph in particular, each node's feature $\mathbf{x}_i$ is replaced with a weighted sum of its neighbors' features. The weights are given by a similarity score s_{ij} based on the Euclidian feature distance:

$$\tilde{\mathbf{x}}_i = \frac{1}{\sum_{j \in \mathcal{N}(i)} s_{ij}} \sum_{j \in \mathcal{N}(i)} s_{ij} \mathbf{x}_j \quad \text{with} \quad s_{ij} = \exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|\right). \tag{2}$$

In our implementation, bilateral filtering is applied as part of the skip connection before the subsequent concatenation with the corresponding feature map in the decoder.

A graphical outline of the final architecture is provided in Figure 1.

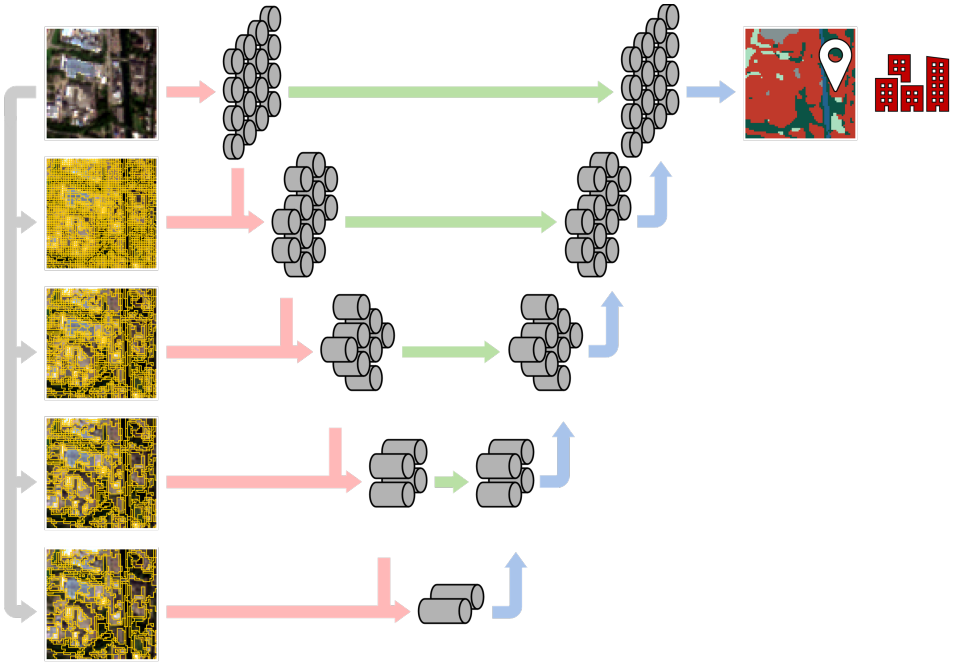

Figure 1: Outline of the proposed Graph U-Net model architecture used to predict land cover from Sentinel-2 images. The labels stem from pointwise LUCAS in-situ data and thus only refer to a single pixel of the predicted land cover map, as indicated by the marker: First, multiple Quickshift oversegmentations of the image with varying levels of granularity are computed (gray arrows). These oversegmentations are used to pool features throughout the image encoder (red arrows). The decoder (blue arrows) mirrors the structure of encoder and uses skip connections including deep bilateral filtering (green arrows).

2.2 Dataset

We conduct our experiments on a dataset based on the LUCAS survey from 2018, which spans across the EU-28 states. The land cover survey broadly differentiates between the classes *artificial land*, *cropland*, *woodland*, *grassland*, *shrubland*, *bare land and lichens/moss*, *water*, and *wetlands*.

To generate a dataset suitable for training an image-based deep neural network, we extract a 64×64 image patch from a Sentinel-2 composite using the red, green, blue and near infrared channels at each LUCAS location. The labeled pixel is at least 5 pixels away from the patch boundary to guarantee sufficient spatial context. In total, there are 337854 data samples, which are geographically divided into training, validation, and test split. In particular, LUCAS data from Austria, Denmark, Finland, France, Ireland, Italy, Latvia, the Netherlands, Poland, Romania, and Slovakia are used for training. Belgium, Bulgaria, Cyprus, Estonia, Germany, Hungary, Malta, Slovenia, Spain are used for validation. Finally, Czech Republic, Croatia, Greece, Lithuania, Luxembourg, Portugal, Sweden, and the United Kingdom are used for testing.

3 Experiments

We implement U-Net and Graph U-Net, as described above, using the same overall parameters: Our networks perform two transformer-based convolutions with 4 attention heads and subsequent batch normalization and activation with ReLU, before each pooling operation. The number of features per pixel/node is increased from 4 to 64 in the first layer and subsequently doubled within each

double convolution. Quickshift segmentations are pre-computed offline, using 300 for the ratio for the spatial-spectral tradeoff and $\tau = \{3, 6, 9, 12\}$ for controlling the hierarchical segmentation granularities.

We train all networks with a batch size of 32 using the Adam optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}. The learning rate is reduced by a factor of 2 once performance on the validation data stagnates for at least two epochs. Training is ultimately concluded when performance on validation data is stagnant for five overall epochs.

We compare the accuracies and F1-scores of U-Net and our Graph U-Net, averaged across the test regions of our data, weighted by the respective number of samples in each region. Like during training, the metrics can only be evaluated at the LUCAS locations. The results are displayed in Table 1. Moreover, a comparsion of two key computational properties of the different models, number of parameters and throughput (including pre-processing) on a single NVIDIA A100 GPU is given in Table 2. Finally, we visualize predicted land cover maps for some of the samples for qualitative comparison in Figure 2.

Table 1: Overall accuracies and F1-scores of the trained model on the test portion of the LUCAS-based test dataset.

	Overall accuracy in %	F1-score in %
U-Net	69.9	48.8
U-Net + DBF	71.2	50.0
Graph U-Net	70.6	49.3
Graph U-Net + DBF	71.7	51.4

Table 2: Computational properties of the models.

	# Parameters	Throughput in # samples per s
U-Net	28.9M	932
U-Net + DBF		737
Graph U-Net	41.6M	76
Graph U-Net + DBF		73

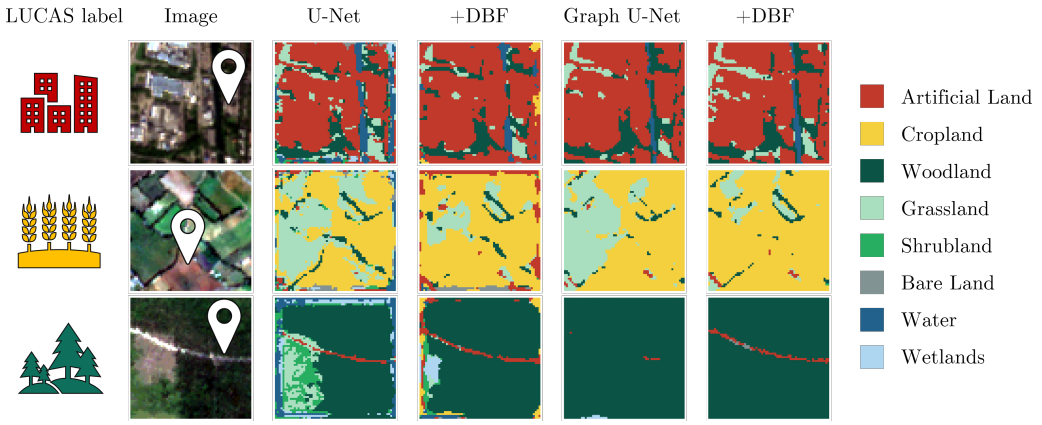

Figure 2: Images, sparse LUCAS labels as indicated by the markers, and predicted land cover maps of the different neural networks for three samples in Hamburg, Germany, Balearic Islands, Spain, and Lapland, Finland (top to bottom). Artifacts within five pixels of the image patch boundary are due to the dataset design and can be ignored.

It can be observed that Graph U-Nets with pooling operations based on Quickshift oversegmentations are able to slightly outperform their standard convolutional counterparts in terms of both overall accuracy and F1-score. In the qualitative comparison we note that Graph U-Nets based on hierarchical oversegmentations, as well as the networks employing deep bilateral filtering modules

tend to produce less noisy results with fewer disjoint regions, which is a desired property of land cover maps. At the same time, the networks are still able to detect land cover details such as the canal in the Germany sample or the road in the Finland sample.

Regarding computational demand, Graph U-Nets' data rate is about 10 times less than that of standard U-Nets due to the segmentation overhead and complex graph-based processing. However, given that training and inference is still feasible on a single GPU within reasonable time, we argue that Graph U-Nets remain a valid option, even for large-scale land cover classification. The additional use of deep bilateral filtering has a comparably small negative effect in this regard and does not introduce any additional learnable parameters to the model.

4 Conclusion

In this work, we present Graph U-Nets with transformer-based graph convolutions, oversegmentation-based pooling, and deep bilateral filtering as a new network architecture for the task of sparsely supervised land cover representations. In initial experiments, we demonstrate both qualitatively and quantitatively that the network is able to learn from pixel-wise labels more efficiently than standard U-Nets and thus achieve better accuracy.

In future work, we plan to conduct more experiments regarding the network's architecture by varying, e.g., the Quickshift segmentation parameters, the graph convolution operator, the perceptive field of the bilateral filtering module, and the dataset size. We also plan to improve the evaluation of the network performance: In addition to the test portion of the LUCAS-based dataset, established high-resolution land cover products may be used or label-independent metrics related to spatial entropy may be considered.

Acknowledgement

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1502/1–2022 - Projektnummer: 450058266

References

Christopher Conrad, Achim Goessl, Sylvia Lex, Annekatrin Metz, Thomas Esch, Christoph Konrad, Gerold Goettlicher, and Stefan Dech. Mapping crop distribution in administrative districts of southwest Germany using multi-sensor remote sensing data. In *Remote Sensing for Agriculture, Ecosystems, and Hydrology XII*, volume 7824, pp. 92–100, 2010. doi: 10.1117/12.865113.

Raphaël d'Andrimont, Astrid Verhegghen, Guido Lemoine, Pieter Kempeneers, Michele Meroni, and Marijn van der Velde. From parcel to continental scale – A first European crop type map based on Sentinel-1 and LUCAS Copernicus in-situ observations. *Remote Sensing of Environment*, 266: 112708, 2021. doi: 10.1016/j.rse.2021.112708.

Hongyang Gao and Shuiwang Ji. Graph U-Nets, 2019. arXiv:1905.05178 [cs, stat].

Qichao Liu, Liang Xiao, Jingxiang Yang, and Zhihui Wei. Multilevel Superpixel Structured Graph U-Nets for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022. doi: 10.1109/TGRS.2021.3112586.

Benjamin Mack, Patrick Leinenkugel, Claudia Kuenzer, and Stefan Dech. A semi-automated approach for the generation of a new land use and land cover product for Germany based on Landsat time-series and Lucas in-situ data. *Remote Sensing Letters*, 8(3):244–253, 2017. doi: 10.1080/2150704X.2016.1249299.

Luca Maggiolo, Diego Marcos, Gabriele Moser, Sebastiano B. Serpico, and Devis Tuia. A Semisupervised CRF Model for CNN-Based Semantic Segmentation With Sparse Ground Truth. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022. doi: 10.1109/TGRS.2021.3095832.

S. Mohammad Mirmazloumi, Mohammad Kakooei, Farzane Mohseni, Arsalan Ghorbanian, Meisam Amani, Michele Crosetto, and Oriol Monserrat. ELULC-10, a 10 m European Land Use and Land Cover Map Using Sentinel and Landsat Data in Google Earth Engine. *Remote Sensing*, 14(13):3041, 2022. doi: 10.3390/rs14133041.

Dirk Pflugmacher, Andreas Rabe, Mathias Peters, and Patrick Hostert. Mapping pan-European land cover using Landsat spectral-temporal metrics and the European LUCAS survey. *Remote Sensing of Environment*, 221:583–595, 2019. doi: 10.1016/j.rse.2018.12.001.

Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjing Wang, and Yu Sun. Masked Label Prediction: Unified Message Passing Model for Semi-Supervised Classification. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, volume 2, pp. 1548–1554, 2021. doi: 10.24963/ijcai.2021/214. ISSN: 1045-0823.

Andrea Vedaldi and Stefano Soatto. Quick Shift and Kernel Methods for Mode Seeking. In David Forsyth, Philip Torr, and Andrew Zisserman (eds.), *Computer Vision – ECCV 2008*, Lecture Notes in Computer Science, pp. 705–718, 2008. doi: 10.1007/978-3-540-88693-8_52.

Matthias Weigand, Jeroen Staab, Michael Wurm, and Hannes Taubenböck. Spatial and semantic effects of LUCAS samples on fully automated land use/land cover classification in high-resolution Sentinel-2 data. *International Journal of Applied Earth Observation and Geoinformation*, 88: 102065, 2020. doi: 10.1016/j.jag.2020.102065.

Linshan Wu, Leyuan Fang, Jun Yue, Bob Zhang, Pedram Ghamisi, and Min He. Deep Bilateral Filtering Network for Point-Supervised Semantic Segmentation in Remote Sensing Images. *IEEE Transactions on Image Processing*, 31:7419–7434, 2022. doi: 10.1109/TIP.2022.3222904.