Satellite image super-resolution for forest localization

Eleftherios Lymperopoulos *, Paraskevi Tzouveli †and Stefanos Kollias ‡
School of Electrical and Computer Engineering,
National Technical University of Athens
Athens, Greece
Email: *lefterislib@gmail.com, †tpar@image.ntua.gr, ‡stefanos@cs.ntua.gr

***Abstract*—With the ever-increasing amount of satellite missions in orbit, Earth observation and remote sensing have advanced rapidly and are applied in a multitude of fields, such as ecosystem monitoring and natural disaster prevention. At the same time, more and more machine-learning architectures are developed, which attempt to map and detect pixel-scale changes on the Earth's surface. Such an application is forest monitoring, which is deemed nowadays highly critical, considering the threat of deforestation on a global scale. However, the spatial resolution of available open-source satellite imagery is limited. One way to address this issue is to augment the remote sensing imagery resolution using Image Super Resolution Networks.**

The goal of this work is to perform the Semantic Segmentation task with Sentinel-2 satellite images in order to detect forest areas. Three Deep Neural Network architectures are trained and tested. Moreover, we use a Super Resolution Deep Network to augment the resolution of the images. We create two datasets, one with original Sentinel-2 images and one with super-resolved images. We train the three architectures on the two datasets independently and we test them on a test set which combines original and super-resolved images. We show that Swin Transformer produces the best results, while training on the super-resolved images enhances the modeling power of all architectures.

***Index Terms*—Deep Learning Networks, Forest Localization, Semantic Segmentation, Super Resolution**

I. Introduction

Forests cover about one third of the Earth's land surface and store about 45% of the world's terrestrial carbon. [1] Moreover, apart from regulating the world's climate, forests play a major role to people's lives socially, financially and aesthetically. It is commonly accepted that forests provide the greatest protection against soil erosion from rain, wind and coastal waves [2], while providing energy sources, like timber and charcoal. [3]

However, human intervention clears large forest areas for the sake of agricultural activity and urban expansion. Deforestation is a major threat, not only because it causes loss of biodiversity, but also because it leads to greenhouse gas emissions and, consequently, global warming. [4] Monitoring forest areas helps tracking changes. Regular and methodical monitoring of forest cover provides crucial information in order to preserve and sustainably manage forest areas.

Remote sensing is of great significance to forest monitoring. More and more satellite missions are dedicated to Earth observation and land cover change tracking. Meanwhile, many machine learning works have been developed recently in the fields of object and surface change detection from remote sensing images. One common problem, though, is that publicly available remote sensing images have limited spatial resolution. For example, the largest spatial resolution we can get from Sentinel-2 images is 10 metres per pixel. One way to mitigate this problem is to augment the resolution from these images by using Image Super Resolution neural network architectures.

This work aims to the development of a Machine Learning system, which processes satellite images in order to detect forest areas with the Semantic Segmentation method. Initially, a dataset is created which comprises of remote sensing images from the Sentinel-2 satellite mission along with open-source ground truth labels, which are collected from OpenStreetMap. After this, the ESRGAN Image Super Resolution architecture [8] is used, which increases the resolution of the images and labels, resulting in having two datasets, one with super-resolved images and one with the original images. Then 3 Semantic Segmentation networks are trained on the two datasets and tested on a mutual test set which includes both super-resolved and original images. The performance of the networks compared to one another, as well as the effect of the super-resolved dataset on the performance of the networks, are examined. Results show that the super resolved dataset improves the ability of the models to extract the desired forest features and can form an important preprocessing stage in similar remote sensing segmentation applications.

A. Related Work

In recent years, Neural Network architectures have been developed for the purpose of monitoring and mapping of forest areas. In [9], a UNet-like architecture performs multi-layer semantic segmentation in order to map endangered areas in the Amazon rainforests. This approach creates accurate mappings within a much smaller temporal baseline than current operational forest monitoring systems and allows a direct monitoring of dynamic changes in forest coverage. In [10] three architectures are compared on the TanDEM-X dataset. Namely, ResNet, DenseNet and UNet are tested on the segmentation of pixels belonging to forest areas, with UNet achieving the best results. In a more recent application, in [11] implements a UNet-style architecture on Sentinel-2 imagery in order to map forest areas in Africa and also estimate their

vegetation height. This method has such low computational cost that it allows for almost real time satellite data processing, even at a continental scale.

II. Dataset Preparation

A. Ground truth collection

In this work, we use Sentinel-2 images which cover all of the surface of Peloponnese, Greece. In order to identify the regions in which forests are located, we use the open-source database OpenStreetMap [12] and its API, Overpass API. OpenStreetMap data can be accessed by queries and the query we used to extract the forest areas is displayed below.

Listing 1. Our Overpass API query

```
[out:json][timeout:1000];
area["name:en"="Peloponnese"]->.searchArea;
(
  way["natural"="wood"](area.searchArea);
  relation["natural"="wood"](area.searchArea);
  );
out body;
>;
out body;
```

Every forest area is represented as a polygon, and all polygons are recorded in a .GeoJSON file. We then created a grid of size which, just like EEA reference grids [13], split the surface of Peloponnese into $2km \times 2km$ regions. Out of this grid, we only kept the regions which contain a polygon included in the aforementioned .GeoJSON file. These regions were needed to download the Sentinel-2 images.

B. Image collection

In this work, we perform semantic segmentation on Sentinel-2 images. In order to collect them, we used the Google Earth Engine API. [14] In particular, for each $2km \times 2km$ region we kept through the ground truth collection process, we get the images available for the time interval from 01-06-2018 until 01-09-2018. We chose summer months, because it is less likely that severe weather conditions weaken the quality of the images. Out of this series of images, we filtered out the images in which the cloud pixel percentage is over 20%. After this, we chose the spectral bands we would use, which in this work, are the three visible bands, blue, green and red. The median of the remaining images was calculated and downloaded. The dataset eventually comprises of the medians of the images for each $2km \times 2km$ region, along with the ground truths, which were converted into grayscale images. In total, 3204 images and ground truths were collected. Because the spatial resolution of the images is 10 metres per pixel and the regions we chose had size $2km \times 2km$, this means that the images and ground truths have size 200×200 pixels.

III. Methodology

A. Image Super Resolution

In order to alleviate the issue of the low spatial resolution of the images, we used an Image Super Resolution architecture, the ESRGAN. [8] This architecture follows the Generative Adversarial Network [15] paradigm. In short, there are two neural networks, the generator and the discriminator. The generator recreates a high resolution image from a low resolution image and the discriminator calculates the probability that this image is not "synthetic", i.e. the image is not constructed by the generator. The generator tries to "trick" the discriminator, that is to construct images that are as realistic as possible so that the discriminator calculates a high probability. The ESRGAN model is an improvement on the SRGAN model by having residual-in-residual blocks in its generator and replacing the classic SRGAN disciminator with a relativistic one. Those modifications have proven to construct images with finer details. In this work, we use a pretrained ESRGAN generator implemented in [16] in order to upscale the images by a factor of 4.

B. Semantic Segmentation

We use three Semantic Segmentation Neural Networks for identification of the pixels belonging to forest :

- The ResUNet-a [5], an encoder-decoder style architecture which was designed for segmentation on remote sensing images. Each encoder and decoder block is a residual block in which multiple parallel atrous convolutions are performed with different dilation rates. With this approach, the network is able to extract features from different receptive fields. In addition, after the encoder and at the second to last layer before the creation of the segmentation mask, a Pyramid Scene Parsing Pooling Module (PSPP) [17] is used, which splits the input into 4 equal partitions and performs max pooling in each of them. With the PSPP module, the network converges faster to optimality.
- The DeepLabv3+ [6], which is an encoder-decoder style architecture. DeepLabv3+ improves the DeepLabv3 architecture by using it or other backbone networks, e.g. ResNet [18] as an encoder and adding a simple and efficient decoder network. By using atrous, pointwise and depthwise convolutions, the decoder is able to process the rich semantic information provided by the encoder network in an efficient way and produce an even better segmentation mask.
- The Swin Transformer [7], which is a vision transformer architecture. The Swin Transformer is a general-purpose backbone for computer vision and is used for tasks such as segmentation and object detection. The Swin Transformer creates a hierarchical representation of the input and calculates self-attention on shifted windows. This approach is more efficient because it limits self-attention computation to non-overlapping windows, it creates connections between them and has linear complexity with respect to image size. This leads to greater modeling power at various scales and it proves very effective in image classification.

C. Experimental Procedure

In this work, we create two datasets. The first is the original dataset which we created with the method described in section II. The second is the super-resolved dataset. To create it, we apply the ESRGAN generator to the images and ground truths of the original dataset in order to augment the resolution from 200×200 to 800×800 pixels. After this, we cut each 800×800 image into 16 200×200 images and we only keep those images in which over 10% of pixels belong to the forest class, according to the ground truths. This is the "filtering" stage of the image below. Note that the "filtering" stage is performed on the original dataset as well, as the distribution of the forest and non-forest classes is highly imbalanced. In the original dataset, we only keep the images in which over 5% of pixels belong to the forest class.

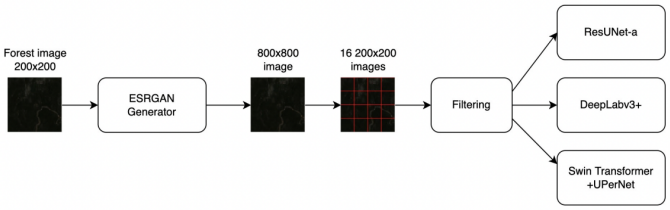

Fig. 1. Procedure for the super-resolved dataset.

At this point the original dataset has 861 images in the train set and 331 images in the validation set. The super-resolved dataset has 4296 images in the train set and 1281 images in the validation set. We then train the three semantic segmentation architectures on the two datasets independently and we test all of them on a mutual test set, which contains 250 images and 250 super-resolved images.

The experiments were conducted on 2 Nvidia GeForce GTX 1080 GPUs with 8GB memory each.

For the training, due to memory constraints, a batch size equal to 2 was selected. The number of epochs was 100, and the Adam optimizer with learning rate 0.001 was selected. ResUNet-a and DeepLabv3+ were implemetned on the Keras library [19], while regarding the Swin Transformer, MMSegmentation framework [20] was used.

For ResUNet-a, 5 downsampling layers, followed by an Atrous Spatial Pyramid Pooling (ASPP) Module [21] with 256 filters, 5 upsampling layers and an ASPP module with 128 filters, were selected. Dilation rates were 1, 3, 15, 31, ReLU activation function and sigmoid output activation function were used.

The DeepLabv3+ network that was used had a ResNet50 backbone pretrained on ImageNet.

The Swin Transformer had an embedding dimension C=96, the number of Swin Transformer blocks were 2, 2, 6, 2 per level respectively, which is the Swin-T version of the backbone. It also has window size equal to 7 and the number of attention heads were 3, 6, 12, 24 per level respectively. In this work, Swin Transformer is backbone to a UPerNet model [22].

All models were trained with 4 different loss functions. Those were Binary Cross Entropy, Dice Loss, the Hybrid Loss described in [23] without the $L_{MS-SSIM}$ component, as it was unstable in training, and Tanimoto loss, which was introduced as an improvement to Dice Loss in the ResUNet-a paper [5].

IV. Results

In the tables below the mean Dice Coefficient, mean Intersection over Union and Recall metrics of the models are presented. We can see that Swin Transformer achieves the best performance among the three architectures, regardless of the dataset on which they trained. Furthermore, the training on the super-resolution dataset improves the performance of the models and their ability to extract the forest features, as imferred by their high recall scores.

There is an important observation at this point. After a thorough examination of the ground truth labels, we discovered that in many of them, although the pixels that were marked as forest areas were evidently correct, there were other areas in the images that could visibly classified as forests, however they were not marked as such. Because of this, the precision metric was not deemed reliable, because it punishes false positives which could actually be true with more accurate ground truths.

Loss Function	Res/a n.s.	Res/a w.s.	DLv3+ n.s.	DLv3+ w.s.	Swin T n.s.	Swin T w.s.
C. Entropy	0.301	**0.405**	0.281	**0.388**	0.387	**0.462**
Dice	0.244	**0.391**	0.321	**0.332**	0.377	**0.526**
Hybrid	0.279	**0.397**	0.312	0.312	0.378	**0.54**
Tanimoto	0.265	**0.406**	0.219	**0.386**	0.25	**0.514**

TABLE I
DICE COEFFICIENT SCORES OF ALL MODELS IN THE MUTUAL TEST SET. WITH N.S. ARE DESIGNATED THE MODELS WHICH WERE TRAINED IN THE NO-SUPER-RESOLVED DATASET AND WITH W.S. ARE DESIGNATED THE MODELS THAT WERE TRAINED IN THE SUPER-RESOLVED DATASET.

Loss Function	Res/a n.s.	Res/a w.s.	DLv3+ n.s.	DLv3+ w.s.	Swin T n.s.	Swin T w.s.
C. Entropy	0.258	**0.38**	0.242	**0.352**	0.24	**0.393**
Dice	0.255	**0.329**	**0.295**	0.27	0.232	**0.357**
Hybrid	0.227	**0.34**	**0.26**	0.232	0.233	**0.369**
Tanimoto	0.203	**0.353**	0.164	**0.326**	0.143	**0.346**

TABLE II
INTERSECTION-OVER-UNION SCORES OF ALL MODELS IN THE MUTUAL TEST SET.

Loss Function	Res/a n.s.	Res/a w.s.	DLv3+ n.s.	DLv3+ w.s.	Swin T n.s.	Swin T w.s.
C. Entropy	0.388	**0.776**	0.314	**0.793**	0.314	**0.782**
Dice	0.313	**0.899**	0.668	**0.828**	0.314	**0.766**
Hybrid	0.319	**0.847**	0.82	**0.873**	0.306	**0.735**
Tanimoto	0.27	**0.822**	0.229	**0.76**	0.165	**0.758**

TABLE III
RECALL SCORES OF ALL MODELS IN THE MUTUAL TEST SET.

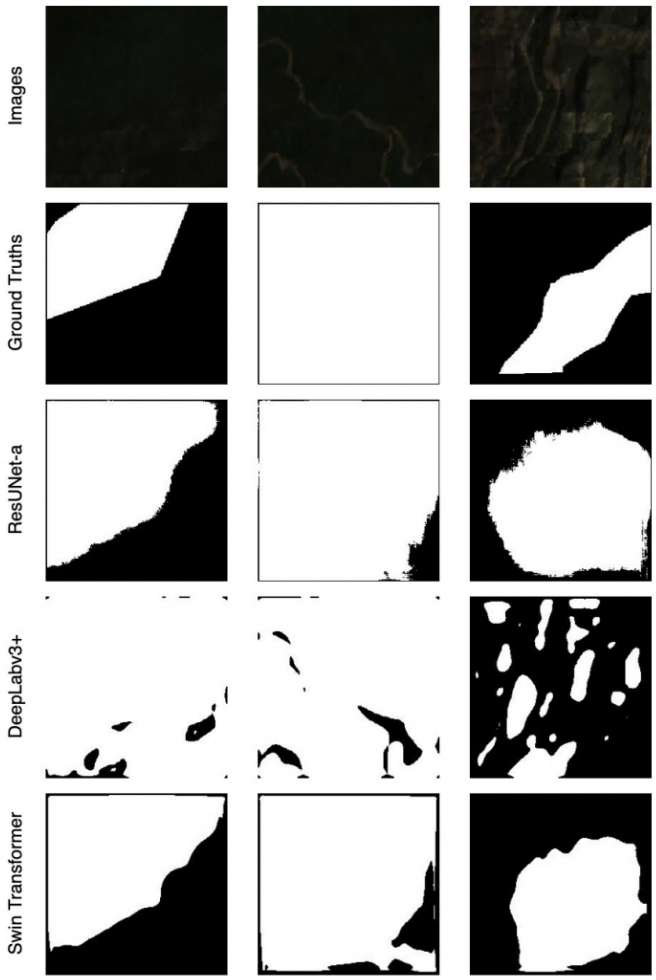

Fig. 2. Predictions of the models trained on the super-resolved dataset

V. Conclusions

Based on the above results, we can draw the conclusion that the Swin Transformer had the best performance among the three archtitectures, with ResUNet-a having the second best performance. Training with Dice Loss and Hybrid Loss had the best results. However, the most significant observation is that training on the super-resolved dataset increased the modeling ability of all architectures. The networks trained on the super-resolved dataset outperformed the models trained on the original dataset and if we take into account especially the recall metric, we can see that the models trained on the super-resolved dataset had a much better ability to identify the pixels marked in the ground truths. This means that by training on the super-resolved dataset, the models were much more able to extract the semantic features of the forest areas. From this, we can deduce that super-resolution can be a significant preprocessing step, especially for low-resolution remote sensing imagery, in order to achieve higher modeling ability. In closing, it is worth noting that this dataset was created to show that these experiments can be conducted to detect forests in any region of the world. The experiments can be adjusted to feature standard datasets as well, and are to be covered in a future work.

References

[1] Erika Romijn et al., Assessing change in national forest monitoring capacities of 99 tropical countries, Forest Ecology and Management, Vol. 352, 2015, pg. 109-123, ISSN 0378-1127, https://doi.org/10.1016/j.foreco.2015.06.003

[2] Satoru Miura et al., Protective functions and ecosystem services of global forests in the past quarter-century, Forest Ecology and Management, Vol. 352, 2015, pg. 35-46, ISSN 0378-1127, https://doi.org/10.1016/j.foreco.2015.03.039

[3] Daisy Núñez, Laura Nahuelhual, Carlos Oyarzún, Forests and water: The value of native temperate forests in supplying water for human consumption, Ecological Economics, Vol. 58, Issue 3, 2006, pg. 606-616, ISSN 0921-8009, https://doi.org/10.1016/j.ecolecon.2005.08.010

[4] Cramer Wolfgang et al.2004 Tropical forests and the global carbon cycle: impacts of atmospheric carbon dioxide, climate change and rate of deforestation Phil. Trans. R. Soc. Lond. B359331–343 http://doi.org/10.1098/rstb.2003.1428

[5] Diakogiannis, Foivos and Waldner, Francois and Caccetta, Peter and Wu, Chen. (2020). ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS Journal of Photogrammetry and Remote Sensing. 16. 94-114. 10.1016/j.isprsjprs.2020.01.013.

[6] Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.. (2018). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. https://doi.org/10.48550/arxiv.1802.02611

[7] Liu, Z. et al. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. https://doi.org/10.48550/arxiv.2103.14030

[8] Wang, X. et al. (2018). ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks https://doi.org/10.48550/arxiv.1809.00219

[9] Dal Molin R Jr., Rizzoli P. Potential of Convolutional Neural Networks for Forest Mapping Using Sentinel-1 Interferometric Short Time Series. Remote Sensing. 2022; 14(6):1381. https://doi.org/10.3390/rs14061381

[10] Mazza A, Sica F, Rizzoli P, Scarpa G. TanDEM-X Forest Mapping Using Convolutional Neural Networks. Remote Sensing. 2019; 11(24):2980. https://doi.org/10.3390/rs11242980

[11] Anders U. Waldeland, Øivind Due Trier, Arnt-Børre Salberg, Forest mapping and monitoring in Africa using Sentinel-2 data and deep learning, International Journal of Applied Earth Observation and Geoinformation, Vol. 111, 2022, 102840, ISSN 1569-8432, https://doi.org/10.1016/j.jag.2022.102840

[12] OpenStreetMap contributors. (2017). Planet dump retrieved from https://planet.osm.org, https://www.openstreetmap.org

[13] https://www.eea.europa.eu/data-and-maps/data/eea-reference-grids-2

[14] Gorelick, N. et al. (2017). Google Earth Engine: Planetary scale geospatial analysis for everyone. Remote Sensing of Environment. https://doi.org/10.1016/j.rse.2017.06.031

[15] Goodfellow, I. J et al. (2014). Generative Adversarial Networks. arXiv. https://doi.org/10.48550/ARXIV.1406.2661

[16] Francesco Cardinale et al., F. C. (2018) ISR. https://github.com/idealo/image-super-resolution

[17] Zhao, H., Shi, J., Qi, X., Wang, X, Jia, J. (2016). Pyramid Scene Parsing Network. arXiv. https://doi.org/10.48550/ARXIV.1612.01105

[18] He, K., Zhang, X., Ren, S. and Sun, J. (2015). Deep Residual Learning for Image Recognition. arXiv. https://doi.org/10.48550/ARXIV.1512.03385

[19] Chollet, F. et al. (2015). Keras. https://keras.io

[20] Contributors, Mms. (2020). MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark. https://github.com/open-mmlab/mmsegmentation

[21] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A. L. (2016). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. arXiv. https://doi.org/10.48550/ARXIV.1606.00915

[22] Xiao, T., Liu, Y., Zhou, B., Jiang, Y. and Sun, J. (2018). Unified Perceptual Parsing for Scene Understanding. arXiv. https://doi.org/10.48550/ARXIV.1807.10221

[23] H. Huang et al., "UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 1055-1059, doi: 10.1109/ICASSP40776.2020.9053405