Multispectral Remote Sensing Data Enhancement for Automatic Processing Chains - A U-Net- vs Transformer-based Cloud Segmentation and GAN Super Resolution Approach

Anonymous Abstract
Submission 23

1 Introduction

Due to the rapidly increasing amount of satellite Earth observation imagery available in high temporal and spatial resolution, automated data processing chains are highly desired. For this, deep learning plays a vital role throughout all steps in data processing. As remote sensing data is of varying quality, data enhancement is useful to assure processability throughout an automated, multistage data processing chain. For this publication, two processing steps will be presented designed to support the robustness of the downstream systems in the processing chain of the upcoming ******** satellite constellation. An overview of mission and satellite design and its processing chain is given in *.

Both approaches utilize deep learning on almost raw data, only basic geometric and radiometric corrections are applied beforehand. The first neural network supports downstream tasks through false-positive-suppression via cloud masking and the second network is improving delineation of buildings for centroid detection via super resolution. To enable a highly-precise georeferencing of the raw image data, a sufficient amount of ground control points (GCPs) have to be identified within the satellite image. As the ******** mission payload is a line scanner, every image line has its own exterior orientation and needs to be georeferenced seperately. Due to the high satellite velocities of circa 7 $\frac{km}{s}$ and an acquisition rate of 2000 Hz, interpolation between image lines is viable. There are several common types of GCPs and ground control shapes including corner reflectors, buildings or segmented land coverage and roads. For the ******** mission, building centroids detected by a deep neural network will serve as GCPs in a similar way as described in [1]. The two processing steps described for this contribution support the building centroid detection in its robustness in adverse conditions like cloud coverage or blurred imagery.

2 Cloud Segmentation: U-Net and Transformer

To avoid that the neural network for building detection creates false positives in foggy or cloudy areas, a scene segmentation is used to mask out unsuitable areas within the satellite scenes. Additionally, tasks further downstream also rely on cloud masking. For this, we compare two approaches, a U-Net and a Transformer model. Both are adapted to be able to operate on multispectral data encompassing up to nine spectral channels. The U-Net is based on the basic U-Net architecture [2]. Concerning its architecture, the biggest change is made to the first layers to enable an eight-channel input. For the Transformer, two altered versions based on Maskformer [3] with a Swin Transformer [4] backbone are created, one for six and one for nine input channels. The dataset is comprised of PlanetScope [5] scenes both including four and eight spectral bands. Cloud segmentation masks provided by Planet are used as ground truth segmentation masks.

Examplary results of both the U-Net and the Transformer are shown in Fig. 1

Due to data quality, creating reliable ground truth maps for further quantitative evaluation is sometimes challenging even for humans. It is difficult to directly compare different cloud detectors, as differing datasets provide unique radiometric information - e.g. the SWIR bands of Sentinel-2 that are not comprised in PlanetScope data.

In many cases, the Planet ground truth cloud masks are outperformed in a qualitative visual inspection, as the ground truth contains erroneously masked areas itself. Still, the U-Net occasionally misclassifies roads as haze or clouds and the Transformer sometimes introduces artifacts on singular patches. Both drawbacks are currently being addressed and, additionally, the dataset is constantly expanded to further increase reliability for all kinds of biomes. The U-Net achieves an mIoU of .65 and the Transformers a mean of .95. Still, the U-Net presents qualitatively pleasing results.

Overall, the Transformer models provide smoother masks with less false positive details but sometimes lacks in detail. This could result from the U-Net being a pixel-based segmentation.

3 Super Resolution GAN

As the ground sampling distance (GSD) of the ******** satellites will vary around 4 meters, smaller

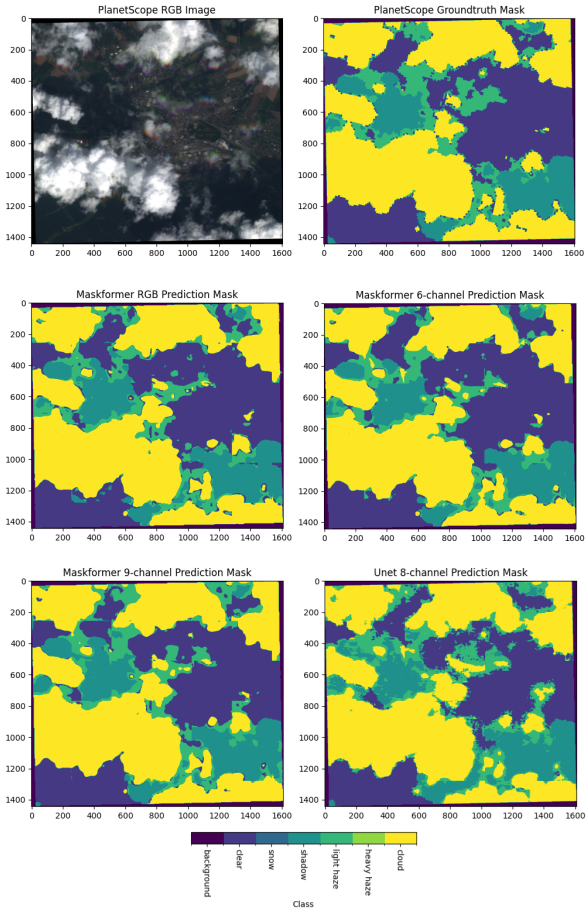

Figure 1. RGB channels of input image, ground truth provided by Planet and our respective results.

buildings might be contained in mixed pixels or not be delineated clearly. To support the neural network in locating as many true positive building centers as precisely as possible, a dual image super resolution GAN is used to sharpen the images before inference. As (very) high resolution imagery of Earth is readily available on a daily basis, it is possible to assume that for each satellite scene captured, a reference scene with equal or better GSD and a maximum temporal shift of one day is available. Our GAN utilizes these as a geometric reference during inference to reduce hallucinations while preserving the radiometric properties of the original scene.

The basic structure of the dual image super resolution GAN is adopted from SRGAN [6]. It is supplemented with the ability to consider the reference scene during inference. The generator contains 16 residual blocks and an upsampling block with in total 1,453,955 parameters, 1,449,731 trainable. The discriminator is built of seven discriminator blocks containing convolutional layers, batch normalization and LeakyReLU. It comprises 107,455,297 parameters, thereof 107,451,585 trainable.

For training, a hallucination-reducing combined adaptive loss function is created and a novel mixed

Figure 2. Example of super resolution results on 4 m GSD PlanetScope data.

pixel approach is introduced to support the GAN in spectral unmixing. Combined adaptive loss in the discriminator encompasses a binary cross-entropy function and a content loss derived from the mean square error of extracted VGG19 [7] features between the high resolution ground truth and the generated image. Content loss reduces hallucination by suppressing the generation of too many new features not present in the reference image. Artificial mixed pixels are fabricated through the generator of the model and support augmenting the training data. These mixed pixels contain the combined radiometric information of a set of pixels in the high resolution image. This supports the learning of spectral unmixing and results in a more stable radiometry in the superresolved image. The dataset itself consists of RGB imagery from the Landsat, Sentinel-2, PlanetScope and SPOT 6 missions. Worldview-3 data is used for quantitative validation as it is not contained in the training data.

An example for the results of the super resolution GAN is shown in Fig. 2. Averaging over the different test combinations, a mean PSNR of 25.30 and SSIM of 0.81 is achieved. These values are good but not outperforming some of the state of the art super resolution GANs listed in [8] concerning these metrics. However, other models are very specific to singular datasets whereas our solution is applicable to a broader range of optical satellite imagery without distorting their unique radiometric properties due to its mixed pixel approach.

4 Conclusion

Our main contribution for the cloud segmentation is to enable the utilization of multispectral data and leveraging its additional information contents compared to RGB imagery.

Our main contribution is a versatile, hallucination-reducing and radiometrically accurate super resolution GAN that is applicable even to satellite datasets whose radiometric properties were not learned during training.

Both processing steps are currently undergoing application tests to evaluate their contribution to processing performance under adverse conditions.

References

[1] M. Greza, L. Hoegner, P.-R. Hirt, R. Roschlaub, and U. Stilla. "Satellite Network Bavaria–Mission and Data Processing". In: *Publikationen der Deutschen Gesellschaft für Photogrammetrie, Fern erkundung und Geoinformation (DGPF)* 43 (2023), pp. 174–182.

[2] O. Ronneberger, P. Fischer, and T. Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer. 2015, pp. 234–241.

[3] B. Cheng, A. Schwing, and A. Kirillov. "Per-Pixel Classification is Not All You Need for Semantic Segmentation". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 17864–17875.

[4] 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: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2021, pp. 10012–10022.

[5] P. L. PBC. *Planet Application Program Interface: In Space for Life on Earth*. Planet, 2018-2024. URL: https://api.planet.com.

[6] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 2017.

[7] K. Simonyan and A. Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. arXiv: 1409.1556 [cs.CV].

[8] K. Karwowska and D. Wierzbicki. "MCWESR-GAN: Improving Enhanced Super-Resolution Generative Adversarial Network for Satellite Images". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 16 (2023), pp. 9459–9479. DOI: 10.1109/JSTARS.2023.3322642.