GENERATIVE ADVERSARIAL NETWORKS FOR THE SATELLITE DATA SUPER RESOLUTION BASED ON THE TRANSFORMERS WITH ATTENTION

Mykola Lavreniuk[1], Leonid Shumilo[2], Alla Lavreniuk[3]

[1] Space Research Institute NASU-SSAU, Kyiv, Ukraine
[2] Department of Geographical Sciences, University of Maryland, College Park, MD, USA
[3] National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine

ABSTRACT

In recent years, free access to high and medium resolution data has become available, providing researchers with the opportunity to work with low resolution satellite images on a global scale. Sentinel-1 and Sentinel-2 are popular sources of information due to their high spectral and spatial resolution. To obtain a final product with a resolution of 10 meters, we have to use bands with a resolution of 10 meters. Other satellite data with lower resolution, such as Landsat-8 and Landsat-9, can improve the results of land monitoring, but their harmonization requires a process known as super-resolution. In this study, we propose a method for improving the resolution of low-resolution images using advanced deep learning techniques called Generative Adversarial Networks (GANs). The state-of-the-art neural networks, namely transformers, with the combination of channel attention and self-attention blocks were employed at the base of the GANs. Our experiments showed that this approach can effectively increase the resolution of Landsat satellite images and could be used for creating high resolution products.

Index Terms— deep learning, GAN, transformers, attention, super-resolution, Sentinel-2, Landsat-8/9.

1. INTRODUCTION

There is a wide range of satellite data available that can be used to accomplish various tasks, such as land cover classification, yield forecasting, logging detection, and wild fire detection. The choice of satellite data to use depends on the specific task and the area of interest. High-resolution images are generally preferred for small target areas [1], while medium-resolution images are commonly used for larger national and global projects [2] – [4]. Sentinel-1 and Sentinel-2, which provide the highest spatial and spectral resolution among publicly available data, are frequently utilized in research [5], [6]. However, to produce a crop classification map with a resolution of 10 meters, it is necessary to use bands with a resolution of 10 meters [7], [8]. Other satellite data with lower resolution, such as Landsat-8 and Landsat-9, can improve results of land monitoring, but their harmonization requires process known as super-resolution.

Super resolution is a common task in image and video processing for various applied purposes [5]. One of the most basic methods for improving the spatial resolution of an image are interpolation techniques like nearest neighbor, bilinear interpolation, bicubic interpolation, and splines. However, these approaches do not consider the meaning or context of the image or the outlines of objects, resulting in a blurry final image.

As deep learning methods and models have advanced, they have also been applied to the problem of improving image resolution. In 2015, the Super-Resolution Convolutional Neural Network (SRCNN) was proposed as one of the first deep learning methods for this task [10]. The SRCNN method was improved in [11] with the Very Deep Super Resolution (VDSR) model, which consisted of 20 layers compared to the three layers in the SRCNN model. Both of these methods have the drawback of using interpolation at the beginning, leading to a large number of parameters in the models and difficulty learning to increase image resolution.

To overcome these problems, the Fast Super-Resolution Convolutional Neural Networks (FSRCNN) model was introduced in [12]. It takes low-resolution images as input without any preprocessing. Other methods with a similar approach have also been suggested in [13] - [15].

An alternative method for increasing image resolution is to use Generative Adversarial Networks (GAN) [16], which consist of two neural networks: a generator and a discriminator. This approach was first proposed in [17] with the Super Resolution Generative Adversarial Networks (SRGAN). The generator network creates a higher resolution image while the discriminator network tries to determine whether the image is real or generated by the generator. Training of the GAN model is completed when the generator has learned to produce high-resolution images that are indistinguishable from genuine images to the discriminator.

Several papers [18] - [20] have proposed using GAN models for image super resolution tasks and have demonstrated that these models provide the most accurate

results and are the most effective for addressing the super resolution problem. In [21] authors utilized GAN model, namely SRGAN, to increase the resolution of Sentinel-2 bands.

At the same time, neural networks with attention mechanism [22], namely transformers, obtained a lot of popularity in different tasks like image processing, neural language processing, recommendation systems, image generation [23], etc. Mainly it is due to transformers outperformed other common deep learning models mostly based on convolution layers without attention [24]. Therefore, we would like to propose a method for improving the resolution of all available 30 meters Landsat bands to 10 meters based on the GAN model that consists of transformers with attention blocks.

2. STUDY AREA AND MATERIALS DESCRIPTION

An experiment was carried out utilizing Sentinel-2 data and Landsat-9. During the period from June 1, 2022, to June 30, 2022, our experiment concentrated on the part of Kyiv region of Ukraine. To ensure clear and unhindered observations, we employed cloud-free composites of the Sentinel-2 and Landsat-9 data for this specific timeframe. The dataset encompassed a comprehensive representation of major land cover types, including cropland, water bodies, artificial objects, forest, and grassland. Google Earth Engine platform was employed for data preparation. The red, green, and blue bands were utilized for both training and testing purposes. To generate a 30m resolution band from the original 10m resolution bands, cubic interpolation was applied to warp the bands. Subsequently, the training data was divided into small patches measuring 384x384 pixels for the 10m bands and 128x128 pixels for the 30m bands.

3. METHODOLOGY

We propose to improve the resolution of Landsat mission satellite images from 30m to 10m using a GAN model. The GAN consists of two neural networks: a generator and a discriminator. Both models are transformers with the attention blocks. Our approach is based on the Hybrid Attention Transformer (HAT) model [25], which is the state-of-the-are model and has been successful in super resolution tasks for regular images.

First, the GAN model is trained to convert the warped to 30m red, green, and blue bands to original resolution, which is 10m. After this, the trained model can then be used to convert Landsat bands from 30m to 10m resolution. Unlike traditional images, satellite images need to be divided into smaller patches for processing. The input image size is 128x128x1 pixels, while the output image size is 384x384x1 pixels. To avoid cutting off objects at the edges of the image, we generate patches with overlap on all sides, equal to 1/4 of the image size.

The discriminator network is a neural network commonly used for binary classification (determining whether an image is real or generated from a lower resolution) with a sigmoid layer at the end. The most common choices for discriminator are ResNet architecture, but any other backbone could be used as well, such as EfficientNet or Normalizer Free Network. However, taking into account that transformers outperform commonly used convolutional networks, we are focusing on them. Transformers such as VIT, DEIT or VOLO have quadratic dependency on the input image size, on contrarily SWIN [24] has linear dependency, thus, and we have chosen it as a discriminator in our task.

For our purposes, the generator network is comprised of a transformer network with attention blocks. The SWIN Transformer has already shown excellent performance in image super-resolution, but as in [25] declared that the shifted window mechanism (main idea of SWIN) is inefficient to build the cross-window connection. Thus, we sought to enhance it further by incorporating additional attention blocks, specifically cross-channel attention like in HAT model. The overall architecture of a generator network consists of three parts: shallow feature extraction, deep feature extraction, and image reconstruction (Fig. 1).

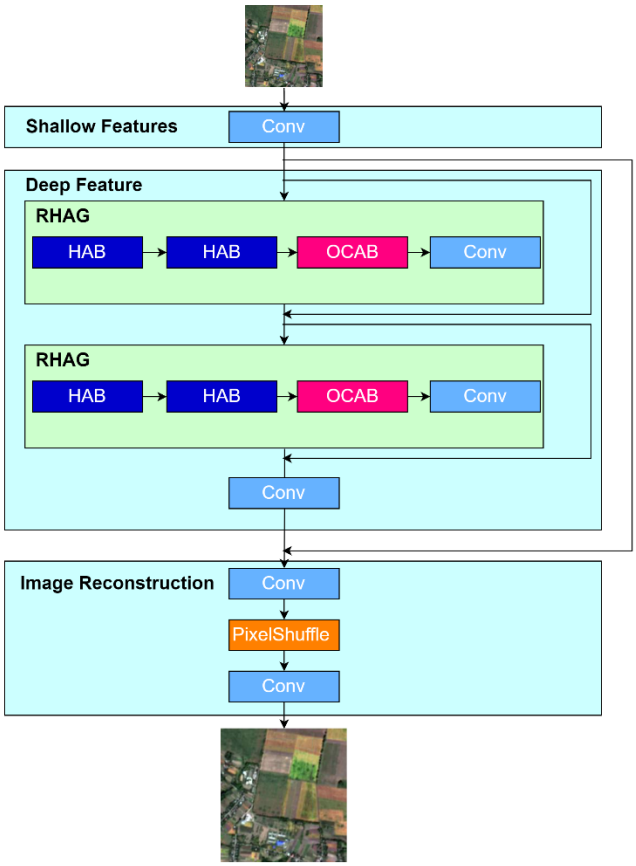

Figure 1. The overall architecture of generator model.

For a given low-resolution image, we first use a single convolutional layer to extract shallow features. The input is transformed from low-dimensional space to high-dimensional space, creating a high-dimensional embedding. Deep features are then extracted through the use of N consequence residual hybrid attention groups (RHAG) and a 3x3 convolutional layer. After deep feature extraction, a global residual connection is used to combine shallow and deep features. Similar to SRGAN the pixel-shuffle technique [17] is used to increase the resolution of the fused feature. Finally, the high-resolution result is reconstructed.

Main innovation block of a generator is the RHAG that consists of M sequential hybrid attention blocks (HAB), an overlapping cross-attention block (OCAB) and a 3×3 convolutional layer. The HUB block integrates a channel attention-based convolution block into the standard Transformer block to further improve the representation capabilities of the network. To be more precise, a channel attention block is added to the standard SWIN Transformer block after the first normalization layer, and running in parallel with the window-based multi-head self-attention module.

4. RESULTS

Previous studies have shown that pre-training is important for low-level tasks. This is because the Transformer model needs a larger amount of data and iterations to learn general knowledge for the task. Thus, we initialized our generator and discriminator models using the pre-trained on ImageNet dataset weights. In our experimental setup, we set the number of RHAGs and HABs to 6 and the channel number of the entire network to 180. The attention head number and window size were also set to 6 and 16, respectively.

We chose three strategies to evaluate the resulting model. The first strategy is evaluation based on the Sentinel-2 data. For this purpose, we are measuring mean average error (MAE) for the spectral bands of generated image (based on the downscaled Sentinel-2 image) and original spectral bands with 10 m spatial resolution. The second strategy uses downscaled Sentinel-2 images to 30 m and Landsat data to 90 m. In this case, we are training model to produce 30 m Landsat multispectral band based on the 90 m band. The third strategy is correlation analysis between generated multispectral bands and original Sentinel-2.

Our model achieved better result based on all 3 strategies in comparison with classical GAN approach, pixel-based deep neural network approach and bilinear regression approach used for the satellite super resolution. MAE for the red channel - 0.0844, MAE for the green channel - 0.0437, MAE for the blue channel - 0.0507, MAE for all channels - 0.05965. Figure 2 illustrates examples of the original Sentinel-2 and Landsat-9 data in two zoom levels, as well as the super resolution results obtained from Landsat-9.

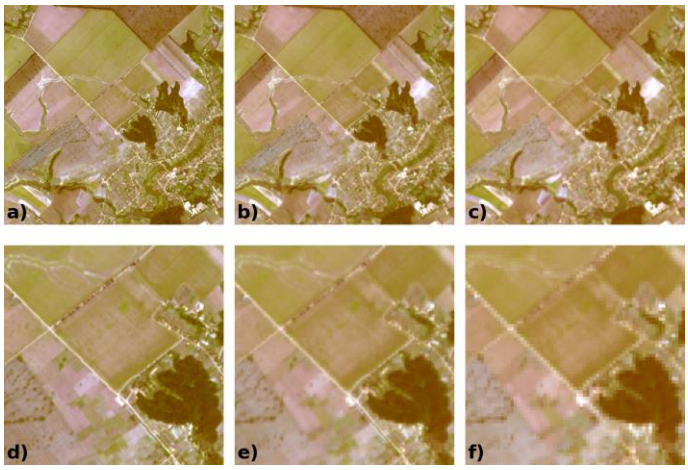

Figure 2. Result of GAN super resolution approach on the image: a) – original 10 m RGB Sentinel-2 image, b) – result of super resolution model, c) – original 30 m RGB Landsat-9 image, d) – zoomed original 10 m RGB Sentinel-2 image, e) – zoomed result of super resolution model, f) – zoomed original 30 m RGB Landsat-9 image.

5. DISCUSSION AND CONCLUSIONS

Landsat mission satellite data collection is one of the largest and significant scientific information sources that can be used for the land monitoring as well as long-range landcover change analysis. However, many studies show that spatial resolution matters and thus multiple research prefer to use Sentinel-2 data. At the same time, the best results in the land monitoring tasks can be achieved only by combination of data from multiple sources. The main bottleneck in the harmonization of Sentinel and Landsat data collections is their spatial resolution. Thus, we propose a new method of satellite data super pixel based on the convolutional GAN with attentions and RHAG blocks that allows to learn dependences between multispectral and textural features and accurately generate 10 m spatial resolution images based on the Landsat data. This method allows to extend possibilities of Landsat-8/9 applications as well as to improve retrospective analysis by upscaling of historical Landsat satellite data.

While our super-resolution method offers significant improvements, its effectiveness in reconstructing images of urban and densely populated areas has limitations. Complex structures and high levels of detail in such areas may pose challenges for optimal results. Further research is needed to explore the applicability and performance of our method in these specific contexts.

In conclusion, our study presents a promising approach for enhancing Landsat satellite data resolution using advanced deep learning techniques. This method has potential benefits for land monitoring and retrospective analysis.

6. ACKNOWLEDGMENT

The authors acknowledge the funding received by the National Research Foundation of Ukraine from the state budget 2020.02/0292 "Deep learning methods and models for applied problems of satellite monitoring" (NRFU Competition "Support for research of leading and young scientists").

7. REFERENCES

[1] Kislov, Dmitry E., et al. "Extending deep learning approaches for forest disturbance segmentation on very high-resolution satellite images." *Remote Sensing in Ecology and Conservation,* vol. 7, no. 3, pp. 355-368, (2021).

[2] Hansen, M. C., et al. “High-resolution global maps of 21st-century forest cover change.” *Science*, 342(6160), 850-853, 2013.

[3] Waldner F., Schucknecht A., Lesiv M. et. al., “Conflation of expert and crowd reference data to validate global binary thematic maps,” *Remote sensing of environment*, vol. 221, pp. 235-246, 2019.

[4] Shelestov A., Lavreniuk M., Kussul N. et. al., “Exploring Google Earth Engine Platform for Big Data Processing: Classification of Multi-Temporal Satellite Imagery for Crop Mapping,” *Frontiers in Earth Science*, vol. 5, 2017.

[5] d'Andrimont, Raphaël, et al. "From parcel to continental scale--A first European crop type map based on Sentinel-1 and LUCAS Copernicus in-situ observations," *arXiv preprint arXiv:2105.09261*, 2021.

[6] Yi Zhiwei, Li Jia, and Qiting Chen, "Crop classification using multi-temporal Sentinel-2 data in the Shiyang River Basin of China," *Remote Sensing,* 12.24, pp 1-21, 2020.

[7] Shelestov A., Lavreniuk M., Vasiliev V., et. al., “Cloud Approach to Automated Crop Classification Using Sentinel-1 Imagery,” *IEEE Transactions on Big Data*, 6(3), pp. 572-582, 2019.

[8] Kussul, N., Lavreniuk, M., Skakun, S., & Shelestov, A. (2017). Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing Letters*, 14(5), 778-782.

[9] Yang, Jianchao, and Thomas Huang. "Image super-resolution: Historical overview and future challenges." *Super-resolution imaging*, pp. 20-34, 2010.

[10] Dong, Chao, et al. "Image super-resolution using deep convolutional networks." *IEEE transactions on pattern analysis and machine intelligence* 38.2 (2015): 295-307.

[11] Kim, Jiwon, Jung Kwon Lee, and Kyoung Mu Lee. "Accurate image super-resolution using very deep convolutional networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[12] Dong, Chao, Chen Change Loy, and Xiaoou Tang. "Accelerating the super-resolution convolutional neural network." *European conference on computer vision*. Springer, Cham, 2016.

[13] Shi, Wenzhe, et al. "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[14] Lim, Bee, et al. "Enhanced deep residual networks for single image super-resolution." *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2017.

[15] Tai, Ying, Jian Yang, and Xiaoming Liu. "Image super-resolution via deep recursive residual network." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[16] Ian, Goodfellow, et al. "Generative adversarial nets.” *In Advances in neural information processing systems*. (2014): 2672-2680.

[17] Ledig, Christian, et al. "Photo-realistic single image super-resolution using a generative adversarial network." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[18] Yuan, Yuan, et al. "Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018.

[19] Wang, Xintao, et al. "Esrgan: Enhanced super-resolution generative adversarial networks." *Proceedings of the European conference on computer vision (ECCV) workshops*. 2018.

[20] Johnson, Justin, Alexandre Alahi, and Li Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution." *European conference on computer vision*. Springer, Cham, 2016.

[21] Lavreniuk, Mykola, et al. "Super Resolution Approach for the Satellite Data Based on the Generative Adversarial Networks." *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2022, pp. 1095-1098.

[22] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*, 30 (2017).

[23] Zhao, Long, et al. "Improved transformer for high-resolution GANs." *Advances in Neural Information Processing Systems,* 34, pp. 18367-18380, 2021.

[24] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

[25] Chen, Xiangyu, et al. "Activating More Pixels in Image Super-Resolution Transformer." *arXiv preprint arXiv:2205.04437*, 2022.