ZERO-SHOT SENTINEL-2 SHARPENING USING A SYMMETRIC SKIPPED CONNECTION CONVOLUTIONAL NEURAL NETWORK

Han V. Nguyen, Magnus O. Ulfarsson, Johannes R. Sveinsson, and Jakob Sigurdsson

Faculty of Electrical and Computer Engineering
University of Iceland

ABSTRACT

Sentinel-2 (S2) satellite constellations can provide multispectral images of 10 m, 20 m, and 60 m resolution for visible, near-infrared (NIR) and short-wave infrared (SWIR) in the electromagnetic spectrum. In this paper, we present a sharpening method based on a symmetric skipped connection convolutional neural network, called SSC-CNN, to sharpen 20 m bands using 10 m bands. The main advantage of SSC-CNN architecture is that it brings the features of the input branch to the output, thus improving convergence without using too many deep layers. The proposed method uses the reduced-scale combination of 10 m bands and 20 m bands, and the observed 20 m bands as the training pairs. The experimental results using two Sentinel-2 datasets show that our method outperforms competitive methods in quantitative metrics and visualization[1].

***Index Terms*—** Sentinel-2, image fusion, image sharpening, super resolution, convolutional neural network.

1. INTRODUCTION

In remote sensing, the demand for high-resolution images is apparent for many applications such as classification [1], mapping, and monitoring of the environment [2]. To maintain high image quality (e.g., signal to noise ratio), an optical remote sensing system that acquires co-registered images may need to vary the spatial resolution of the bands. This means that panchromatic (PAN) images have a higher spatial resolution than narrow wavelength multispectral (MS) or hyperspectral (HS) images. Therefore, researchers are interested in fusing high spatial and low spectral resolution images with low spatial and high spectral resolution images. This fusion results fused images that have both high spatial and spectral resolutions. Examples of such fusion techniques are pansharpening, hypersharpening, and multi-hyperspectral image fusion where the image pairs used for fusion are (PAN, MS), (PAN, HS) and (MS, HS), respectively.

S2 constellation satellites acquire 13 MS bands at 10 m, 20 m, and 60 m resolution without PAN. The S2 sharpening is therefore different from pansharpening, where the fine spatial resolution bands assume the role of PAN to sharpen the coarse resolution bands. But S2 sharpening is more difficult than pansharpening since there is less spectral correlation between coarse and fine resolution bands than for PAN and MS. However, traditional pansharpening methods, such as component substitution (CS) and multiresolution analysis (MRA) [3], can be used in S2 sharpening. The basic idea of CS and MRA is to inject the detail obtained from PAN to MS. A more advanced S2 sharpening method is the area-to-point regression kriging (ATPRK) algorithm [4]. In [5], the S2 sharpening process was formulated as a solution to an inverse problem in lower-dimensional subspace and solved using a cyclic descent based method.

Recently, advances in deep learning have introduced new directions for S2 image sharpening. The deep learning-based S2 sharpening methods can be roughly divided into two approaches. The first approach proposed in [6] is to train a CNN on a large number of different S2 data so the network can generalize well enough. As a result, it could work well on any S2 data. The second approach given in [7, 8] presented a single image learning (zero-shot) approach in which the training and testing were performed on the same image. The network was trained using many small patches extracted from the S2 image and tested using the same given image to sharpen the coarse resolution images. The zero-shot approach does not require a huge amount of S2 data, and the training is quick. However, the network must be re-trained for different images.

In this paper, we propose a zero-shot SSC-CNN to sharpen S2 images. SSC-CNN uses symmetric skipped connection layers, which improves convergence without using too deep network. Experiments with real S2 datasets show that our proposed method outperforms competitive methods in both quantitative evaluation and sharpened images quality.

2. THE METHOD

In this paper, we denote the observed fine resolution bands (10 m bands) as $\mathbf{X}_f \in \mathbb{R}^{d_1 \times d_2 \times L_1}$, and the observed coarse

This work was supported in part by the Icelandic Research Fund under Grant 174075-05 and Grant 207233-051, and University of Iceland Doctoral Fund.

[1] Project codes: https://github.com/hvn2/S2-SSC-CNN

resolution bands (20 m bands) as $\mathbf{X}_c \in \mathbb{R}^{\frac{d_1}{r}\times\frac{d_2}{r}\times L_2}$, where r is the fine-coarse resolution ratio, $d_1 \times d_2$ is the spatial size of fine resolution bands, and L_1 and L_2 are the band numbers of the fine and the coarse bands, respectively. The S2 sharpening problem is to estimate the coarse resolution bands at a finer scale, i.e., $\widehat{\mathbf{X}}_C = f_\theta(\mathbf{X}_f, \mathbf{X}_c)$, where $\widehat{\mathbf{X}}_C \in \mathbb{R}^{d_1\times d_2\times L_2}$ and $f_\theta(\cdot)$ is the output of the SSC-CNN. The training needs a high resolution (reference) $\mathbf{X}_C \in \mathbb{R}^{d_1\times d_2\times L_2}$ which is not available at full-scale resolution. We overcome this problem by using the Wald protocol [9], which is based on working with a down-sampled version of $\mathbf{X}_f$ and $\mathbf{X}_c$. Thus, the observed coarse resolution bands $\mathbf{X}_c$ can be used as the reference. The proposed S2 sharpening method at reduced-scale resolution is described in Algorithm 1.

Algorithm 1: S2 sharpening using SSC-CNN at reduced-scale resolution

```
1. Input: $\mathbf{X}_f, \mathbf{X}_c$
2. Down-sample $\mathbf{X}_f, \mathbf{X}_c$ by $r$: $\mathbf{X}_c^D \overset{r\downarrow}{=} \mathbf{X}_c, \mathbf{X}_f^D \overset{r\downarrow}{=} \mathbf{X}_f$,
   where $\mathbf{X}_c^D \in \mathbb{R}^{\frac{d_1}{r^2}\times\frac{d_2}{r^2}\times L_2}, \mathbf{X}_f^D \in \mathbb{R}^{\frac{d_1}{r}\times\frac{d_2}{r}\times L_1}$
3. Create $\mathbf{Y}^D \in \mathbb{R}^{\frac{d_1}{r}\times\frac{d_2}{r}\times(L_1+L_2)}$ by:
     • Up-sample $\mathbf{X}_c^D$: $\mathbf{X}_c^{DU} \overset{r\uparrow}{=} \mathbf{X}_c^D$
     • Concatenate $\mathbf{X}_f^D$ and $\mathbf{X}_c^{DU}$ along spectral axis:
       $\mathbf{Y}^D \overset{\text{cat}}{=} [\mathbf{X}_c^{DU}, \mathbf{X}_f^D]$
4. Extract training patch pairs $\{\mathbf{y}, \mathbf{x}\}$ from the pair
   $\{\mathbf{Y}^D, \mathbf{X}_c\}$
5. Train SSC-CNN using training patch pairs $\{\mathbf{y}, \mathbf{x}\}$
6. Obtain sharpened images at reduce-scale resolution:
   $\widehat{\mathbf{X}}_c = f_\theta(\mathbf{Y}^D)$, where $f_\theta(\cdot)$ is the pre-trained network
   estimation function, and $\theta$ are network parameters
7. Output: $\widehat{\mathbf{X}}_c$
```

We can use the pre-trained network at a reduced-scale resolution to estimate the full-scale resolution of coarse bands by

$$\widehat{\mathbf{X}}_C = f_\theta(\mathbf{Y}), \tag{1}$$

where $\mathbf{Y}$ is the concatenation of up-sampled $\mathbf{X}_c$ and $\mathbf{X}_f$, i.e., $\mathbf{Y} \overset{\text{cat}}{=} [\mathbf{X}_c^U, \mathbf{X}_f]$, where $\mathbf{Y} \in \mathbb{R}^{d_1\times d_2\times(L_1+L_2)}$, and $\mathbf{X}_c^U \overset{r\uparrow}{=} \mathbf{X}_c$ is the up-sampling of $\mathbf{X}_c$ by r.

2.1. Network Architecture

To demonstrate the idea of the proposed S2 sharpening, we use a skipped connection CNN which is inspired by the U-net architecture [10] with some modifications. Since the down-sampling operator can lead to information loss, we do not use any down-sampling and up-sampling layers. Instead the spatial size of the input images is kept the same throughout the network. The batch normalization layers in [10] are removed in our network. The network architecture is shown in Fig. 1.

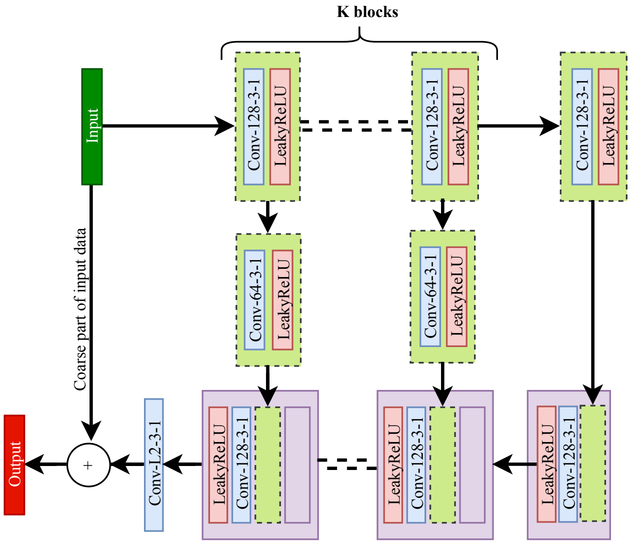

Fig. 1: Network structure. Conv-128-3-1 represents a convolutional layer with 128 filters of kernel size 3 and stride 1. The number of skipped connection blocks is $K = 3$.

It consists of an input branch and an output branch. The backbone of input and output branches are K blocks of convolutional and *LeakyReLU* (conv-relu) layers. In each input block, a skipped connection is generated by using conv-relu layers. It is then concatenated with the corresponding output block except for the last layer. The sharpened image is the result of element-wise adding of the output of the last convolution layer with L_2 filters and the coarse parts of the input.

2.2. Training and Testing

Here, we use a single image learning approach that allows training and testing using a single image. As described in Algorithm 1, the network is trained by using small overlapping patch pairs $\{\mathbf{y}, \mathbf{x}\}$ created from the pair $\{\mathbf{Y}^D, \mathbf{X}_c\}$. The training patches have a spatial size of 16×16 and a stride of 8 in the spatial axes. In each training batch, M patch pairs are randomly selected. The loss function is given by the mean squared error between the network output and the training target.

In the testing, at the reduced-scale resolution, the same image as used in training of size $\frac{d_1}{r} \times \frac{d_2}{r} \times (L_1 + L_2)$ is the input of the network. As a result, the sharpened image is the output of the network, i.e., $\widehat{\mathbf{X}}_\mathbf{c} = f_\theta(\mathbf{Y}^D)$. At full-scale resolution, the estimated high resolution of coarse bands can be obtained by using the same pre-trained network which is given in (1). However, we will omit the full-scale evaluation in this paper. The remaining parameters of training and testing are chosen as follows: The number of skipped connection blocks is $K = 3$; the optimizer is Adam with learning rate is 0.005; the number of training patches is 1024; the batch size is 64; and the number of training epochs is 200.

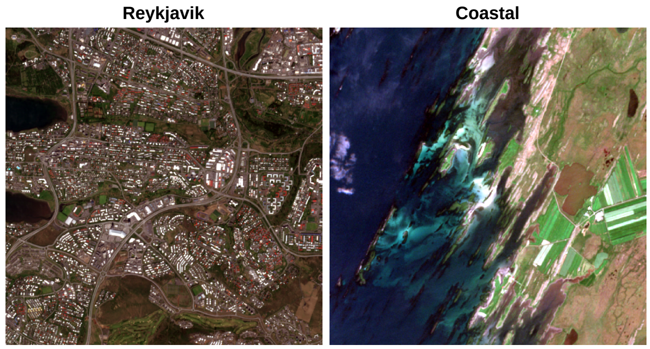

Fig. 2: The RGB color images created using bands 2, 3, and 4 of the datasets used in experiments.

3. EXPERIMENTAL RESULTS

The data used in the experiments are the real S2 data of the level-1C product obtained in June 2019, in Iceland. Each S2 tile captures an area of 100×100 km^2 in 10 m, 20 m and 60 m resolution bands. We use two smaller datasets that are cropped from the S2 tile data for the experiments. The first dataset, called Reykjavik, shows a part of Reykjavik. The second dataset, called Coastal, shows a coastal area in the northwest part of Iceland. The spatial sizes of the 10 m resolution bands of Reykjavik and Coastal datasets are 512×512 and 408×408, respectively. They are shown in Fig. 2 as RGB color images using bands 2, 3, and 4.

The performance is evaluated at a reduced-scale resolution using the following quantiative metrics [5]: The signal to reconstructed error (SRE), the mean structural similarity (MSSIM) index, and the spectral angle mapper (SAM).

The proposed SSC-CNN method is compared with a model-based method and a deep learning-based method that are ATPRK [4] and ResNet [7], respectively. ATPRK and ResNet used the default parameters from the above-mentioned papers. SSC-CNN uses the network structure and parameters described in Sections 2.1 and 2.2. Table 1 shows the quantitative evaluation of different sharpening methods for both datasets. In this table, B5-B12 are the band-wise SREs in decibels, MSRE is the mean SRE in decibels, MSSIM is the mean SSIM, and SAM in degrees. Our proposed method outperforms ATPRK and ResNet in most metrics for both datasets. SSC-CNN gives higher mean SRE than the second best, ResNet, by approximately 2 dB and 3 dB for the Reykjavik and Coastal datasets, respectively. While ATPRK is the worst method in all metrics for all datasets. ATPRK is outperformed by SSC-CNN by roughly 0.2 in term of MSSIM and 3 degrees in term SAM, for both datasets. The visual quality of the sharpened image for band 5 and band 6 of the Coastal dataset are pictured in Fig. 3 and Fig. 4, respectively. SSC-CNN and ResNet can accurately reconstruct bands 5 and 6, since they are almost indistinguishable from the reference images, as shown in Fig. 4. ATPRK gives blurry image as can be seen when zooming in the image in Fig. 4. This is clearly verified by the residual images ($\mathbf{X}_c - \widehat{\mathbf{X}}_c$) shown in Fig. 3 where ATPRK has the highest residual structure and SSC-CNN has the smallest residual structure.

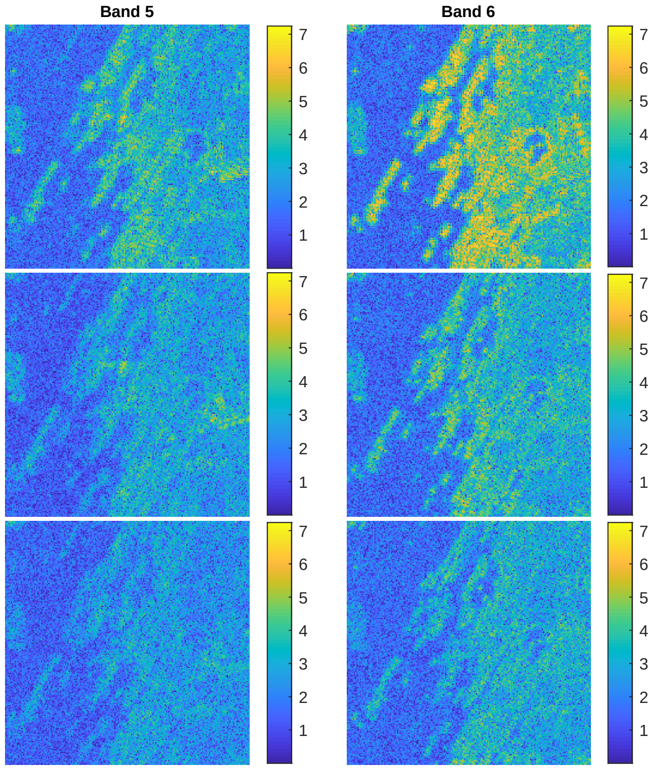

Fig. 3: Residual images in log-scale ($\log(1 + Residual)$) of band 5 and band 6 of Coastal dataset. The images are shown from top to bottom for the methods: ATPRK, ResNet, and SSC-CNN.

4. CONCLUSION

In this paper, we addressed the problem of sharpening the 20 m resolution bands to 10 m resolution bands of the real S2 data. Our proposed method using an SSC-CNN can accurately recover the sharpened image at a reduced-scale. The experimental results on two S2 datasets, where one shows the urban area and other shows the coastal area, demonstrate that our proposed method considerably outperforms competitive methods.

5. REFERENCES

[1] F. Palsson, J. R. Sveinsson, J. A. Benediktsson, and H. Aanaes, "Classification of pansharpened urban satellite images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 1, pp. 281–297, 2011.

[2] Y. Xie, Z. Sha, and M. Yu, "Remote sensing imagery in

Table 1: Performance at a reduced-scale resolution for Reykjavik and Coastal datasets using different methods. B5-B12 are band-wise SREs in dB, MSRE is the mean SRE in dB, MSSIM is the mean SSIM, and SAM is in degree. Best results are given in boldface type.

Dataset	Method	B5	B6	B7	B8a	B11	B12	MSRE	MSSIM	SAM
Coastal	ATPRK	25.35	21.72	21.13	21.40	21.61	20.59	21.97	0.7493	5.4764
	ResNet	29.41	31.32	30.85	32.62	30.89	27.27	30.40	**0.8466**	2.0755
	SSC-CNN	**32.77**	**34.65**	**35.08**	**36.08**	**34.30**	**30.66**	**33.92**	0.8361	**1.5166**
Reykjavik	ATPRK	14.02	16.63	16.38	16.37	16.00	13.65	15.51	0.7551	4.0273
	ResNet	24.30	27.63	27.78	28.65	27.71	23.92	26.67	0.9563	1.8524
	SSC-CNN	**25.98**	**28.76**	**28.72**	**29.55**	**29.20**	**25.95**	**28.03**	**0.9647**	**1.6765**

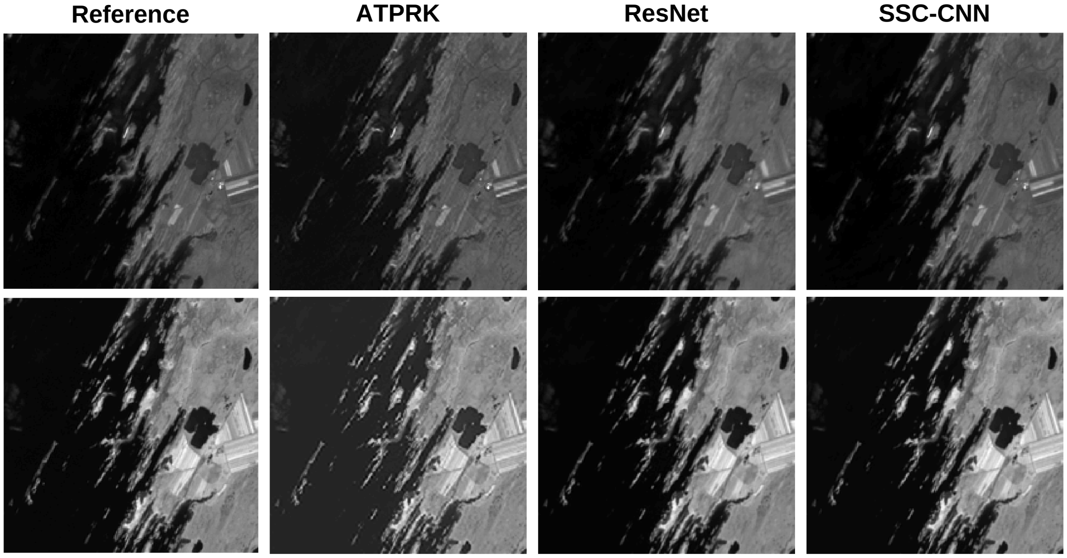

Fig. 4: Sharpening results at a reduced-resolution for Coastal dataset for all methods. The images are shown in gray-scale. First row is band 5, second row is band 6.

vegetation mapping: a review," *Journal of Plant Ecology*, vol. 1, no. 1, pp. 9–23, 2008.

[3] G. Vivone, L. Alparone, J. Chanussot, M. Dalla Mura, A. Garzelli, G. A. Licciardi, R. Restaino, and L. Wald, "A critical comparison among pansharpening algorithms," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 5, pp. 2565–2586, 2014.

[4] Q. Wang, W. Shi, Z. Li, and P. M. Atkinson, "Fusion of Sentinel-2 images," *Remote Sensing of Environment*, vol. 187, pp. 241–252, 2016.

[5] M. O. Ulfarsson, F. Palsson, M. Dalla Mura, and J. R. Sveinsson, "Sentinel-2 sharpening using a reduced-rank method," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6408–6420, 2019.

[6] C. Lanaras, J. Bioucas-Dias, S. Galliani, E. Baltsavias, and K. Schindler, "Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 146, pp. 305–319, 2018.

[7] F. Palsson, J. Sveinsson, and M. Ulfarsson, "Sentinel-2 image fusion using a deep residual network," *Remote Sensing*, vol. 10, no. 8, p. 1290, 2018.

[8] A. Shocher, N. Cohen, and M. Irani, "“Zero-Shot” super-resolution using deep internal learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3118–3126.

[9] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogrammetric Engineering and Remote Sensing*, vol. 63, no. 6, pp. 691–699, 1997.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proceeding of International Conference on Medical Image Computing and Computer-assisted Intervention*, 2015, pp. 234–241.