POLITECNICO DI TORINO
Repository ISTITUZIONALE

DeepSUM++: Non-local Deep Neural Network for Super-Resolution of Unregistered Multitemporal Images

Original
DeepSUM++: Non-local Deep Neural Network for Super-Resolution of Unregistered Multitemporal Images / Molini, A.B., Valsesia, D., Fracastoro, G., Magli, E.. - (2020), pp. 3644-3656. (International Geoscience and Remote Sensing Symposium, IGARSS 2020) [10.1109/IGARSS39084.2020.9324418].

Availability:
This version is available at: 11583/2844384 since: 2020-09-08T09:49:34Z

Publisher:
IEEE

Published
DOI:10.1109/IGARSS39084.2020.9324418

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
IEEE postprint/Author's Accepted Manuscript

©2020 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

17 September 2026

DEEPSUM++: NON-LOCAL DEEP NEURAL NETWORK FOR SUPER-RESOLUTION OF UNREGISTERED MULTITEMPORAL IMAGES

Andrea Bordone Molini, Diego Valsesia, Giulia Fracastoro, Enrico Magli

Politecnico di Torino, Italy

ABSTRACT

Deep learning methods for super-resolution of a remote sensing scene from multiple unregistered low-resolution images have recently gained attention thanks to a challenge proposed by the European Space Agency. This paper presents an evolution of the winner of the challenge, showing how incorporating non-local information in a convolutional neural network allows to exploit self-similar patterns that provide enhanced regularization of the super-resolution problem. Experiments on the dataset of the challenge show improved performance over the state-of-the-art, which does not exploit non-local information.

***Index Terms*—** Super-resolution, CNN, Non-local

1. INTRODUCTION

The task of generating a high resolution (HR) image from one or more low resolution (LR) images is named super-resolution (SR). Despite the continuous development of ever more advanced optical devices, for some imaging applications acquiring HR images can be cumbersome or impossible due to theoretical and practical limitations. Super-resolution methods are thus required to improve image resolution beyond the sensor capability. The approaches to image super-resolution can be framed into two main categories: single-image SR (SISR) and multi-image SR (MISR). SISR exploits spatial correlation in a single image to recover the HR version. Certain scenarios provide multiple LR versions of the same scene, which can be combined by means of MISR techniques. In MISR, the reconstruction of high spatial-frequency details takes full advantage of the complementary information coming from different observations of the same scene. MISR was first explored by Tsai and Huang [1] who increased the resolution of Landsat TM images in frequency-domain. Many other MISR techniques were proposed over the years [2] mainly in spatial domain to overcome the drawbacks affecting the frequency-domain. Typical spatial-domain methods include non-uniform interpolation [3], iterative back-projection (IBP) [4], projection onto convex sets (POCS) [5], regularized methods [6], and sparse coding [7]. Some of these works have been proposed for remote sensing applications. To improve the effectiveness of MISR methods, non-local techniques have also been introduced by Protter et al. in [8, 9], based on the non-local means filter [10]. The idea of these methods is to exploit non-local structural similarity across spatially distant patches within an image.

Recently, deep learning techniques have spread in various fields of research, due to their ability of extracting high-level features from images. For SISR, deep learning works [11, 12] achieved state-of-the-art results both on traditional natural images and remote sensing imagery [13]. Instead, just a few works propose deep learning solutions addressing MISR for remote sensing imagery [14, 15], which is particularly challenging due to the highly detailed images and complex statistics. Addressing MISR from multiple low-resolution acquisitions of the PROBA-V satellite over the span of a month was the goal of a recent challenge issued by the European Space Agency [16]. The winning approach was DeepSUM [15]: a convolutional neural network (CNN) architecture that can handle registration and fusion of multiple unregistered images from the same scene in way that is robust to content variations. Its main novelty is that it can learn to concurrently solve the reconstruction and registration tasks to obtain a single HR image.

Exploiting non-local information in CNN architectures is currently a hot research topic. Some works in the denoising literature [17, 18] attempt various approaches to define CNNs able to combine both spatially-neighboring as well as distant pixels. The graph convolution in [18] adopts a non-local convolution operator where a similarity graph is constructed connecting pixels whose feature vectors are close to each other.

In this paper, we introduce graph-convolutional layers in the DeepSUM architecture, in order to improve its learning capability and generate non-local feature maps in the hidden layers to solve a MISR problem on remote sensing imagery. The resulting DeepSUM++ architecture shows that the performance of DeepSUM is greatly enhanced by the non-local operations, improving upon the state of the art.

2. PROPOSED METHOD

The proposed method, called DeepSUM++, reconstructs a high-resolution image I^{HR} given a set of N LR images $I^{\text{LR}}_{[0,N-1]}$ representing the same scene. $I^{\text{LR}}_{[0,N-1]}$ and I^{HR} are represented as real-valued tensors with shape $N \times H \times W \times C$

This research has been funded by the Smart-Data@PoliTO center for Big Data and Machine Learning technologies.

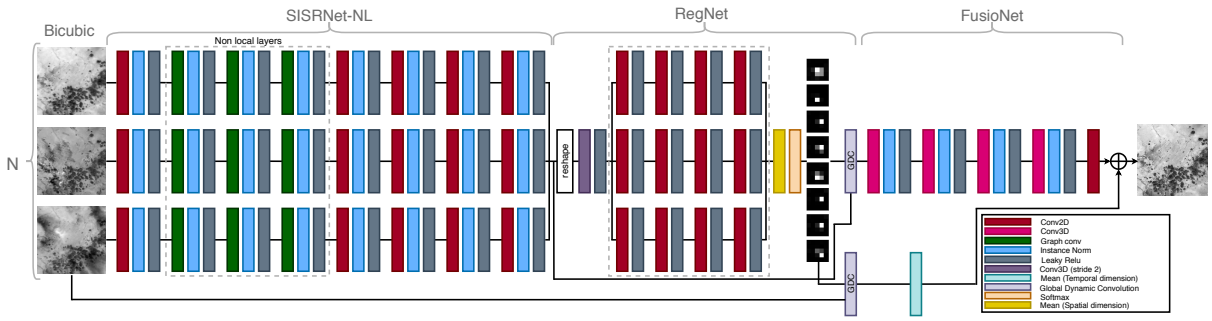

Fig. 1. DeepSUM++ architecture. Graph-convolutional layers are used in SISRNet-NL.

and $1 \times rH \times rW \times C$ respectively, where H and W are the height and the width of the input LR frames, C is the number of channels and r is the scale factor.

DeepSUM++ solves a MISR problem in a setting with multitemporal LR image acquisitions. This is characterized by variations in scene content over multiple acquisitions due to weather or human activities. Moreover, the absolute image brightness may vary among LR images as well as between the reference HR image and the LR set. Finally, the LR images are not registered with each other, and we assume that the geometric disparity is only translational.

DeepSUM addressed this problem with a CNN composed of three main building blocks: feature extraction (SISRNet), registration in feature space (RegNet) and fusion to obtain a single HR reconstruction (FusionNet). All blocks are based on classical 2D or 3D convolutions, so only local information is exploited to obtain the final HR reconstruction. DeepSUM is optimized in an end-to-end fashion allowing the registration and fusion task to leverage the feature learning capabilities of SISRNet that is in fact a crucial component.

DeepSUM++ builds upon the DeepSUM architecture and introduces non-local operations in the SISRNet block, which allows SISRNet to compute more powerful high-dimensional image representations that considerably improve the quality of the subsequent tasks by relying on more informative features. An overview of the network is shown in Fig. 1.

The first block, called SISRNet-NL (non local), is a simple SISR network without the output projection to a single channel, where N bicubically interpolated LR (ILR) images are processed independently. The weights are shared along the temporal dimension, i.e., all the N ILR images go through the same operators. Overall, it acts as a feature extractor exploiting both local and non-local spatial correlations to improve upon the initial bicubic interpolation. In order to exploit non-local spatial correlation, traditional convolution is augmented with a graph convolution operator, which adds to the receptive field of a pixel a predefined number of non-local pixels whose feature vectors are close to the feature vector of the current pixel.

More formally, the graph-convolutional layer takes as input the feature vectors $\mathbf{H}^l \in \mathbb{R}^{F^l \times rHrW}$, i.e., the high-dimensional representations of the image pixels at layer l, and the adjacency matrix of a graph connecting image pixels. The graph structure is constructed as a K-nearest neighbor graph where each pixel is connected to the K pixels whose feature vectors are closest in terms of Euclidean distance, within a search window. The 8 local neighbors are excluded from this search as they already contribute to the local convolution. The graph-convolutional layer is composed of two components both taking as input the feature vectors $\mathbf{H}^l$. A classic 2D convolution aggregates the local neighbors through 3×3 filters and an edge-conditioned convolution (ECC) [19, 18] aggregates the feature vectors of the non-local pixels. For each pixel i, the ECC computes the output feature vector $\mathbf{H}_i^{l+1,\mathrm{NL}} \in \mathbb{R}^{F^{l+1}}$ as follows:

$$\mathbf{H}_i^{l+1,\mathrm{NL}} = \sum_{j \in \mathcal{N}_i^l} \gamma^{l,j \to i} \frac{\mathcal{F}_{w^l}^l(\mathbf{d}^{l,j \to i}) \mathbf{H}_j^l}{|\mathcal{N}_i^l|} = \sum_{j \in \mathcal{N}_i^l} \gamma^{l,j \to i} \frac{\Theta^{l,j \to i} \mathbf{H}_j^l}{|\mathcal{N}_i^l|}$$

where $\mathcal{F}_{w^l}^l : \mathbb{R}^{F^l} \to \mathbb{R}^{F^{l+1} \times F^l}$ is a fully-connected network, parameterized by w^l, that takes as input the differences between feature vectors $\mathbf{d}^{l,j \to i} = \mathbf{H}_j^l - \mathbf{H}_i^l$ and outputs a weight matrix, and $\mathcal{N}_i^l$ is the set of non-local neighbors of pixel i. $\gamma^{l,j \to i}$ is a non-learnable scalar edge-attention term to underweight the edges between pixels with distant feature vectors for training stabilization. This term is computed as:

$$\gamma^{l,j \to i} = \exp(-\|\mathbf{d}^{l,j \to i}\|_2^2 / \delta)$$

where δ is a hyperparameter.

Finally, the local and the non-local contributions are averaged to estimate the output feature vector:

$$\mathbf{H}_i^{l+1} = \frac{\mathbf{H}_i^{l+1,\mathrm{NL}} + \mathbf{H}_i^{l+1,\mathrm{L}}}{2} + \mathbf{b}^l$$

where $\mathbf{H}_i^{l+1,\mathrm{L}}$ is the 2D convolution output and b^l is a bias. We refer the reader to [18] for more details on the advantages of this definition of graph convolution. Depending on the desired computational complexity one may want to use graph convolution for all the layers in SISRNet or just for a subset.

The second network block, called RegNet, dynamically estimates a set of 2D filters using the N higher dimensional image representations produced by the SISRNet-NL block to register them to each other. Handling registration within an

end-to-end trainable network enables the generation of adaptive filters for disparity compensation.

The third block, called FusionNet, merges the registered image representations in the feature space in a "slow" fashion, i.e., by exploiting a sequence of 3D convolutional operations with small kernels. This block allows the network to learn the best space to decouple image features that are relevant to the fusion from irrelevant features and to construct the best function to exploit spatio-temporal correlations. The output of this block is a single super-resolved image.

Finally, the proposed architecture employs a global input-output residual connection. The network estimates only the high frequency details necessary to correct the bicubically-upsampled input. This is an established technique for image restoration problems [11], including SISR. However, with respect to SISR, our proposed network is a many-to-one mapping, so the residual is actually added element-wise to a basic merge of the registered ILR input images $I^{\text{IRLR}}_{[0,N-1]}$ in the form of their average. Notice that registration of the input images $I^{\text{ILR}}_{[0,N-1]}$ is performed before averaging by means of the same filters produced by RegNet as illustrated in Fig. 1. Hence, the output is computed as follows:

$$I^{\text{SR}} = \frac{1}{N} \sum_{i \in [0,N-1]} I^{\text{IRLR}}_i + R.$$

being R the residual estimated by the CNN.

2.1. Loss Function

Model parameters are optimized by minimizing a loss function computed as a modified version of the Euclidean distance between the SR image and the HR target. This modified loss function builds invariance to absolute brightness differences between I^{SR} and I^{HR}. Moreover, since I^{SR} and I^{HR} could be shifted, so the loss embeds a shift correction, resulting in:

$$L = \min_{u,v \in [0,2d]} \|I^{\text{HR}}_{u,v} - (I^{\text{SR}}_{\text{crop}} + b)\|^2, \tag{1}$$

where $I^{\text{SR}}_{\text{crop}}$ is the cropped version of I^{SR}, $I^{\text{HR}}_{u,v}$ a shifted version of I^{HR} and b represents the brightness correction term:

$$b = \frac{1}{(rW-d)(rH-d)} \sum_{x,y} \left(I^{\text{HR}}_{u,v} - I^{\text{SR}}_{\text{crop}}\right).$$

3. RESULTS

In this section, we validate DeepSUM++ by comparing its performance with that of the following methods: averaged bicubic interpolated and registered images (Bicubic+Mean), CNN-based SISR method shared across multiple images followed by registration and averaging (SISR+Mean), IBP [4], BTV [6], dynamic upsampling filters (DUF) network [20] and DeepSUM [15]. For all these methods, we followed the same procedure for the data preparation: bicubic interpolation and registration by phase correlation algorithm, except for DUF and DeepSUM that compute their own registration.

3.1. Experimental setting and training process

In the following experiments, we performed the same preprocessing and data preparation steps as in [15] for both NIR and RED band images. The images used to construct the two datasets have been captured by the PROBA-V satellite and are part the ESA challenge dataset [16]. The size of the collected images is 128×128 and 384×384 for the LR and HR data respectively.

Before training DeepSUM++ as a whole, SISRNet-NL is pretrained by setting up a pure SISR problem (single input image) where an additional projection layer is added at the end, in order to turn the high-dimensional feature space into a single-channel image. SISRNet-NL with the final projection layer is trained with the same loss function in Eq. (1). DeepSUM++ is trained for 300 epochs with a batch size of 4, with separate models for NIR and RED. SISRNet-NL is initialized from the pretraining while FusionNet and RegNet weights are initialized from the model in [15].

We train DeepSUM++ using the minimum number of images available for each scene, i.e., $N = 9$ images.

The exact number of network layers is shown in Fig. 1 and the number of features is 64 everywhere except for the RegNet's first layer, which has 128 filters. Three graph convolutional layers are used in SISRNet-NL.

3.2. Quantitative and qualitative results

The evaluation metric that we consider is a modified version of the PSNR (mPSNR), from which we derived the loss function described in Sec. 2.1:

$$\text{mPSNR} = \max_{u,v \in [0,6]} 20 \log \frac{2^{16}-1}{\| I^{\text{HR}}_{u,v} - (I^{\text{SR}}_{\text{crop}} + b) \|^2}$$

Similarly to the loss function, this metric has been devised to cope with the high sensitivity of the PSNR to biases in brightness and with the relative translation that the reconstructed image might have with respect to the target HR image.

The validation has been performed over the same NIR and RED test sets used in [15], using the sliding window procedure to use 13 images per scene in the testing process. Table 1 reports the results of the comparison. It can be noticed that the proposed method outperforms all the other methods. In particular, DeepSUM++ outperforms DeepSUM by 0.09 dB for NIR and 0.08 dB for RED band. To ensure a fair comparison we have retrained DeepSUM following the same procedure used for DeepSUM++ for the same number of iterations.

These quantitative results are accompanied by a qualitative comparison in Fig. 2. It can be noticed that our proposed method produces visually more detailed images, recovering finer texture and sharper edges.

4. CONCLUSION

In this paper, we demonstrated that non local information can be successfully exploited by neural networks to enhance the reconstruction quality of multitemporal remote sensing images in a MISR problem.

Table 1. Average mPSNR (dB) and SSIM

	Bicubic+Mean	IBP [4]	BTV [6]	SISR+Mean	DUF [20]	DeepSUM [15]	**DeepSUM++**
NIR	45.69/0.97782	45.96/0.97960	45.93/0.97942	46.41/0.98166	47.06/0.98417	47.84/0.98578	**47.93/0.98620**
RED	47.91/0.98507	48.21/0.98648	48.12/0.98606	48.71/0.98787	49.36/0.98948	50.00/0.99075	**50.08/0.99118**

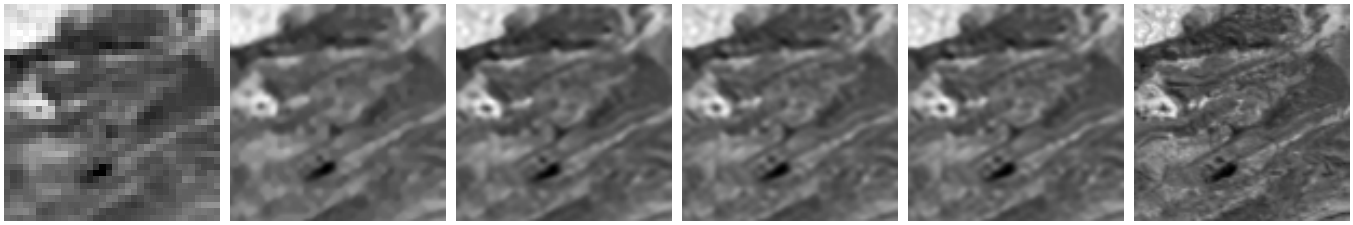

Fig. 2. NIR band images (imgset1144). Left to right: one among the LR images, BTV(47.37 dB / 0.98284), DUF (48.02 dB / 0.98620), DeepSUM (48.74 dB / 0.98844), DeepSUM++ (49.46 dB / 0.98943), HR image.

5. REFERENCES

[1] R.Y. Tsai and T.S. Huang, "Multiframe image restoration and registration," in *Advances in Computer Vision and Image Processing*. JAI Press, 1984, vol. I, pp. 317–339.

[2] M. Elad and Y. Hel-Or, "A fast super-resolution reconstruction algorithm for pure translational motion and common space-invariant blur," *IEEE Transactions on Image Processing (TIP)*, vol. 10, no. 8, pp. 1187–1193, Aug 2001.

[3] S. Lertrattanapanich and N. K. Bose, "High resolution image formation from low resolution frames using delaunay triangulation," *IEEE Transactions on Image Processing (TIP)*, vol. 11, no. 12, pp. 1427–1441, Dec 2002.

[4] Michal Irani and Shmuel Peleg, "Improving resolution by image registration," *Graphical Models and Image Processing (CVGIP)*, vol. 53, no. 3, pp. 231–239, 1991.

[5] Henry Stark and Peyma Oskoui, "High-resolution image recovery from image-plane arrays, using convex projections," *Journal of the Optical Society of America*, vol. 6, no. 11, pp. 1715–1726, Nov 1989.

[6] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE Transactions on Image Processing (TIP)*, vol. 13, no. 10, pp. 1327–1344, Oct 2004.

[7] Kato Toshiyuki, Hino Hideitsu, and Murata Noboru, "Multi-frame image super resolution based on sparse coding," *Neural Networks*, vol. 66, pp. 64 – 78, 2015.

[8] M. Protter, M. Elad, H. Takeda, and P. Milanfar, "Generalizing the nonlocal-means to super-resolution reconstruction," *IEEE Transactions on Image Processing (TIP)*, vol. 18, no. 1, pp. 36–51, Jan 2009.

[9] M. Protter and M. Elad, "Super resolution with probabilistic motion estimation," *IEEE Transactions on Image Processing*, vol. 18, no. 8, pp. 1899–1904, Aug 2009.

[10] A. Buades, B. Coll, and J. . Morel, "A non-local algorithm for image denoising," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, June 2005, vol. 2, pp. 60–65 vol. 2.

[11] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Transactions on Image Processing (TIP)*, vol. PP, 08 2016.

[12] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu, "Residual dense network for image super-resolution," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2472–2481, 2018.

[13] W. Ma, Z. Pan, J. Guo, and B. Lei, "Achieving super-resolution remote sensing images via the wavelet transform combined with the recursive res-net," *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, vol. 57, no. 6, pp. 3512–3527, June 2019.

[14] Michal Kawulok, Pawel Benecki, Krzysztof Hrynczenko, Daniel Kostrzewa, Szymon Piechaczek, Jakub Nalepa, and Bogdan Smolka, "Deep learning for fast super-resolution reconstruction from multiple images," 2019, vol. 10996.

[15] A. B. Molini, D. Valsesia, G. Fracastoro, and E. Magli, "Deepsum: Deep neural network for super-resolution of unregistered multitemporal images," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–13, 2019.

[16] Kelvins - ESA's Advanced Concepts , "PROBA-V Super Resolution," `https://kelvins.esa.int/proba-v-super-resolution`.

[17] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang, "Non-local recurrent network for image restoration," in *Advances in Neural Information Processing Systems*, 2018, pp. 1673–1682.

[18] Diego Valsesia, Giulia Fracastoro, and Enrico Magli, "Deep graph-convolutional image denoising," *arXiv preprint arXiv:1907.08448*, 2019.

[19] Martin Simonovsky and Nikos Komodakis, "Dynamic Edge-Conditioned Filters in Convolutional Neural Networks on Graphs," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 29–38.

[20] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim, "Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.