Single Image Super-Resolution for Optical Satellite Scenes Using Deep Deconvolutional Network

Sumedh Pendurkar[1(✉)], Biplab Banerjee[2], Sudipan Saha[3,4], and Francesca Bovolo[3]

[1] College of Engineering Pune, Pune, India
sumedh.pendurkar@gmail.com
[2] Indian Institute of Technology Bombay, Mumbai, India
bbanerjee@iitb.ac.in
[3] Fondazione Bruno Kessler, Trento, Italy
{saha,bovolo}@fbk.eu
[4] University of Trento, Trento, Italy

Abstract. In this paper, we deal with the problem of super-resolution (SR) imaging and propose a deep deconvolutional network based model for the same. In principle, the SR problem considers the construction of the high-resolution (HR) version of a scene given a number of so-called low-level image instances of the respective scene. Moreover, if there is a single low-resolution (LR) image available, the problem becomes even difficult and ill-posed. We deal with such a scenario and show how the popular deconvolutional network can effectively reconstruct the HR image by learning the functional mapping at the patch level. We evaluate the proposed model on a number of optical remote sensing (RS) images obtained from the UC-Merced dataset. Experimental results suggest that the proposed model consistently outperforms the existing deep and shallow models for single image SR for the RS images.

Keywords: Satellite imaging · Deconvolutional neural networks · Image super resolution · Deep learning

1 Introduction

Rapid developments in RS technologies have contributed to the availability of large quantity of visual data pertaining to the Earth's surface. Satellite images are used in variety of applications ranging from environmental monitoring to homeland security since they reveal a vast amount of intricate details regarding the different geographical locations on ground.

For the sake of extracting accurate information from these images, the quality of the satellite images must be as pristine as possible. Satellite images obtained from sensors are generally affected by different degradation factors and sophisticated image enhancement techniques are needed in order to improve their spatial

E. Ricci et al. (Eds.): ICIAP 2019, LNCS 11751, pp. 410–420, 2019.
https://doi.org/10.1007/978-3-030-30642-7_37

resolution. Among the different approaches, the spatial resolution of an imaging system can be improved using a class of image enhancement algorithms known as SR imaging [16]. Particularly in RS applications including image classification, having higher spatial resolution helps to extract minute features from the respective scenes, thus significantly enhancing the classification results. However, sensors with high spatial resolution are required at the hardware level for obtaining high quality images which is not always feasible. Another challenge in this regard is due to the down-linking of the HR satellite images to ground stations which is often difficult and expensive. All such factors invariably degrade the quality of the satellite images to a considerable extent. As a remedy, SR techniques have become much popular to convert LR satellite images to the corresponding HR versions.

In this regard, the forward model [16] for imaging and motion process can be formulated as

$$Y_k = DB_kM_kX + n_k \tag{1}$$

given the HR scene X, warp matrix M, blur matrix B, down-sampling matrix D, noise vector n and the k^{th} LR image Y_k, respectively. As can be understood, we obtain the LR images because of the degradation caused by warping, blurring and sub-sampling performed to the captured HR scenes due to limitations of cameras. From Eq. 1 it can be affirmed that the process of obtaining the HR images from the LR counterparts is ill-posed nature. Please note that, in this paper, we consider HR scenes as the upscaling of the resolution of available LR images by a factor of 2.

Initially, multi-image SR [17] techniques were followed to generate the HR image from multiple LR observations. As expected, these techniques often face difficulties in registering the LR scenes on the HR grid. This subsequently instigated the research focus on single-image SR. However, the key problem in this respect is the absence of prior knowledge regarding the high frequency details from the images. In this regard, the learning based single image SR techniques such as sparse coding [2,6] are based on an assumption that the sparse representation of the LR image patch over the LR dictionary is same as the corresponding HR patch over the HR dictionary. However, this assumption does not always hold true which leads to restricted performance by these models.

On the other hand, a number of recently introduced deep learning strategies find their application to RS image analysis [3,25]. Recently, deep learning algorithms [5,8,10,19] are used to tackle the SR problem for natural scenes as well as on RS applications. Following the same, we propose a deconvolutional network model for the purpose of single image SR from optical RS data.

The proposed model learns an end-to-end mapping between the LR image and HR image pairs at the patch level. In particular, the images are divided into patches of size 32×32 and then forward-propagated through the network, following which, the reconstruction error is calculated and is subsequently back-propagated. For testing, we consider the standard simulated scenario where the images are upscaled by a factor of 2 and then forward propagated through our network to obtain the predicted HR image.

2 Related Work

2.1 Image Super Resolution

As aforementioned, based on the availability of LR images to be deployed for the SR process, the existing SR algorithms can broadly be classified into two families [1,16]: (i) single image SR, and (ii) multi-image SR.

For multi-image SR, the basic premise is the availability of multiple LR images representing a given scene. These LR images provide different views belonging to the same scene in terms of sub-pixel level shifts. Multi-image SR techniques are broadly classified into: non-uniform interpolation approaches, frequency domain approaches, regularized image reconstruction approaches. Non-uniform interpolation based methods [4] register the LR images on the HR grid. The main problem with registration is the motion estimation with reference to any of the LR images that is required to account for these sub-pixel shifts. Restoration methods such as de-blurring, modeled as spatial averaging operator are used to smoothen the obtained HR image. In contrast, frequency based approaches use the aliasing relationship between continuous Fourier transform of HR image and the discrete Fourier transform of the captured LR images to reconstruct the HR image. Regularization based reconstruction methods are usually used when plenty of LR images are available. Prior knowledge of the solution is used to stabilize the inversion of this ill-posed problem. Either of the deterministic approach or stochastic approaches like Maximum-a-Posteriori (MAP) [17] are used for this purpose.

On the other hand, single-image SR presents more challenging scenario as it involves prediction of the high frequency image details. Some of the early works on single-image SR are documented in [7,22]. Single-image SR techniques are classified into four categories - prediction models, edge based models, image statistical models and exemplar based models [21]. Among them, exemplar based models haven shown to outperform the rest for images of different modalities. Most of these approaches focus on learning a mapping between the LR and HR patch. SR using sparse coding (SCSR) [2,6] are based on regularizing the dictionaries for the HR and LR patches so as to make the dictionary atoms coherent.

2.2 Deep Learning for Image Super Resolution

Convolutional Neural Networks (CNN) have shown high accuracy in image classification [12], object detection [15] and many more. On the other hand, SRCNN [5] is arguably the most popular model for SR from natural images. They propose a 3 layer network consisting of 3 conv layers while the pooling layers are eliminated to avoid loss of pixel information during the reconstruction process.

2.3 Deconvolutional Networks

Likewise, deconvolutional neural networks (deconv-net) are extensively deployed for image denoising, feature extraction, and semantic segmentation [14].

By design, deconv-net follows the encoder-decoder architecture and they have enabled production of highly diverse set of filters beyond the edge primitives [24].

In this paper, deconv-net is used to obtain the HR image from the features extracted by the conv layers in the network for satellite imaging applications. Although it is observed that the deeper networks are proved to be beneficial, however in case of SRCNN the results have saturated at three layers even though the layers are increased. On the other hand, given their ability in efficiently reconstructing images in the decoder stage, deconv-net can incorporate both the deeper structure and learn invariant features which is expected to output better HR versions of the underlying scenes.

3 Deconv-Nets for Single-Image SR

Different stages of the proposed model include pre-processing the image, formulation of the model and training the deconv-net, as detailed in the following:

3.1 Pre-processing

We convert all images into YCbCr color space. All the three channels are upscaled by factor of 2 using bicubic interpolation and the proposed model is applied on the luminance channel following the setup of majority of the existing single-image SR models [18]. Once we obtain the resultant 'Y' channel from the model, the upscaled 'Cb' and 'Cr' are directly stacked to it to obtain the final HR image. For training, we obtain sub-images of 32×32 with a stride of 14 as proposed in [5]. This method is adopted so that we would have training images of fixed sizes for the simplicity of programming. Let us denote the luminance channel after upscaling as Y (not to be confused with 'Y') and the original image sample as X, which is the objective image to be generating by propagating Y through the network.

While deploying the proposed model, we pass the luminance channel of the image without dividing it into patches. This is done to avoid incorporating other methods to stitch the obtained results from patches to form the eventual HR image and handle cases like borders of image-patch, which might result in the poor quality of the obtained image.

3.2 Description of the Proposed Model

The proposed model uses conv layers, each followed by an activation function in order to introduce non-linearities. ReLU [13] is chosen as the activation function since it speeds up the computation and performs relatively good. The deconvolution layers are subsequently used for the reconstruction of the respective HR image. Note that pooling and un-pooling are not incorporated in order to reduce possible information loss as they would reduce the dimensions which is unsuitable for our task. Besides, in case for image SR, feature maps do not require any scale invariance which is generally required for many deep learning tasks.

The block diagram of the proposed deconv neural network based model is shown in Fig. 1. The deconv layers are exactly mirror-like reflection of the conv layers, with same number of layers as in the convolutional part and same filter sizes as that of conv layers.

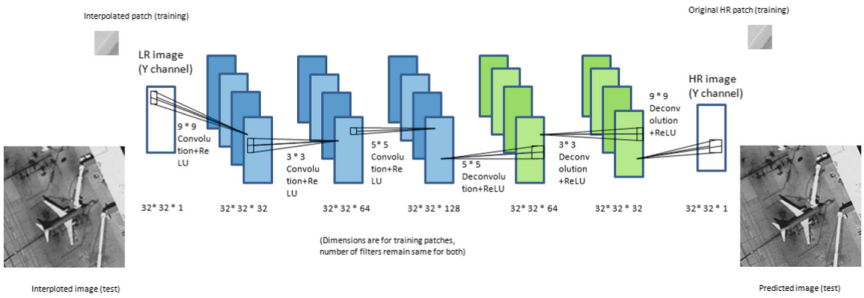

Fig. 1. An illustration of proposed model showing the different layers of the deconvolutional network for image SR.

To summarize, the proposed SR model consists of three stages:

- **Patch extraction:** The first conv layer is used for patch extraction. Larger filters are used to extract patches as well as the basic feature maps from input LR image Y.
- **Feature extraction and Mapping:** The next two conv layers are used to extract high level features and map the LR feature maps into the corresponding HR feature maps.
- **Reconstruction:** The last three deconv layers are used for the construction of the HR image from the feature maps obtained from the conv layers. We choose deconv layers with a stride of 1 over the conv layers as deconv layers are basically transposed conv layers, that work like a backward pass operation which allow reconstruction of original images from the learnt feature maps.

3.3 Training

Using the definitions mentioned in Sect. 3.1, X can be represented as a function of Y given the network parameters θ:

$$X = F(Y; \theta) \tag{2}$$

The standard mean squared error (MSE) over n LR-HR patch pairs given by Eq. 3 is used as the loss function for the proposed model.

$$MSE = \frac{1}{n}(\sum_{i=1}^{n}(F(Y_i; \theta) - X_i)^2 \tag{3}$$

For optimizing MSE, we rely on the Adam's optimizer [11]. The parameter update rule followed in this case is given by:

$$\theta_t = \theta_{t-1} - \frac{\alpha_t \cdot m_t}{(\sqrt{\nu_t} + \hat{\epsilon})} \tag{4}$$

where m_t is the gradient of MSE with respect to θ, ν_t is the squared gradient, β_1 and β_2 are hyper parameters controlling the moving values of the gradient. On the other hand, a small constant $\hat{\epsilon}$ is used for numerical stability. α_t is the learning rate, which is tuned based on Eq. 5.

$$\alpha_t = \alpha \cdot \frac{\sqrt{1 - \beta_2^t}}{(1 - \beta_1^t)} \tag{5}$$

3.4 Implementation Details

Given the proposed architecture, the size of filters in the conv layers are 9×9, 3×3 and 5×5 whereas the number of filters considered in each of these layers are 32, 64 and 128, respectively. Note that the number of filters are increased progressively considering that they yield more high-level features, apart from restricting much loss of image details. On the other hand, the deconv layer filters are constructed in opposite fashion compared to the conv layer filters (Fig. 1). In total, the proposed network has $451,969$ parameters. We initialize the weights of the network as per the He uniform initialization [9] as they consider the distribution of outputs after ReLU activation while deciding the variance of the uniform distribution of the weights which makes it easier to train.

We set $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for the Adam's optimizer, inspired by [11]. Learning rate (α) is set to 0.001 with decay of 10^{-6}.

We also pad the output of each layer by zeros for handling the pixels that lie on border. Therefore height and width of feature maps of each layer remain identical (in our case, it is 32×32 for all layers). This is in contrast to SRCNN, which explicitly requires to strip the border pixels for preserving the resolution of the feature maps.

The output of the network is the luminance channel of obtained high resolution image. We interpolate the Cb, Cr channels of the low resolution image and stack the obtained luminance channel on top of it. We convert this resultant YCbCr image, into RGB format to get the final image.

4 Results and Experiments

4.1 Data Set

The model is deployed on the popular UC-Merced optical RS dataset [23] which is extensively used for different RS applications including classification etc. This dataset consists of 21 different scene themes. Each class has a total of 100 images of size 256×256 pixels providing us with total of 2100 images.

50 randomly selected images from each class are used for training while 10 images per class are deployed for cross-validation. The model is tested on 4 images per category. This subsequently generates a total of 205, 520 image patches for mapping the LR to HR patches.

4.2 Metrics

The goodness of the proposed SR model is tested using the standard signal to noise ratio (PSNR) as mentioned in the following:

$$PSNR = 10 \times \log_{10}(255/MSE) \tag{6}$$

where MSE is obtained according to Eq. 3. Besides, we use the Structural Similarity (SSIM) [20] for measuring the visual similarity at the patch level (between LR and HR patches)

$$SSIM(x, y) = \frac{(2 \times \mu_x \times \mu_y + c_1)(2 \times \sigma_{xy} + c_2)}{(\mu_x + \mu_y + c_1)(\sigma_x + \sigma_y + c_2)} \tag{7}$$

where x and y represent the LR and HR patches, μ is average value of the luminance channel, σ is standard deviation, σ_{xy} is covariance. Further, $c_1 = (0.01L)^2$, $c_2 = (0.03L)^2$ where L is the dynamic range of the pixel values: $2^{bitsperpixel} - 1$, e.g., in this case $L = 127$.

4.3 Discussions

Fixation of the Network Structure. In order to obtain the optimal architecture, we initially repeat the experiments with varied network structures (in terms of the number of deployed conv and deconv layers). Different combinations used include the 2conv-2deconv model, 3conv-3deconv model and 4conv-4deconv models where 2conv-2deconv implies a model with 2 conv layers followed by 2 deconv layers and so on. From Fig. 2, which is a plot of validation error against epochs, we conclude that 3conv-3deconv layered network performs the best and this architecture is subsequently finalized. From Fig. 3 we conclude that the 2conv-2deconv model underfits the data and fails to establish a relationship between LR and HR images effectively. Whereas, the 4conv-4deconv model's accuracy averaged on the test data is slightly worse as that of 3conv-3deconv model though it performs slightly better on some of the test images. Moreover, it is computationally slower as compared to 3conv-3deconv model as it has more trainable parameters due to addition of more layers. Therefore, the superiority of the 3conv-3deconv model can be validated over the others based on the quality of the obtained HR images in terms of the PNSR measure as well as the computational efficiency.

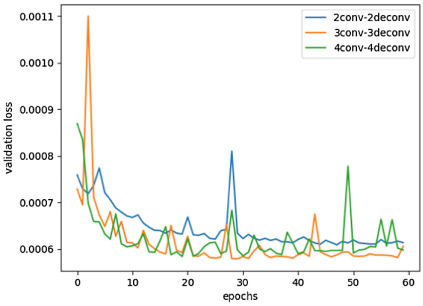

Fig. 2. Validation error versus epochs.

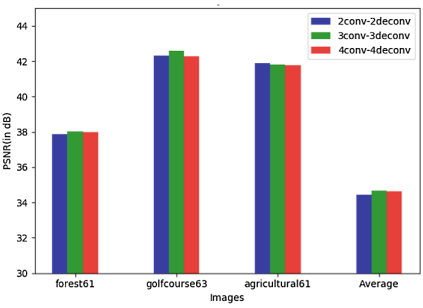

Fig. 3. Comparison of PSNR for different layers.

Empirical Study. The 3conv-3deconv model is also compared with a number of the recent state-of-the-art methods ScSR [21], SRCNN [5] and bicubic interpolation. We have retrained ScSR and SRCNN on the same data set and split as we did for our model to have a fair comparison. Figure 4 shows the HR image generated by state-of-the-art models, our proposed model and the original HR image, respectively for qualitative assessment. On the other hand, Table 1 depicts the accuracy of models based on PSNR and SSIM. From Table 1 it is clear that our proposed model outperforms than state-of-the-art methods for SR on satellite images based on both the considered metrics. Also, from Fig. 4 we can infer that our proposed model recovers more details of HR image as compared to other models.

Table 1. Comparison between Bicubic, ScSR, SRCNN and proposed model

	Bicubic		ScSR		SRCNN		Proposed model	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
airplane60	33.014	0.923	36.431	0.954	36.503	0.953	37.128	0.955
forest60	32.742	0.952	35.660	0.974	35.369	0.973	35.633	0.975
harbor60	24.044	0.900	26.223	0.940	26.351	0.939	27.652	0.957
parkinglot60	25.800	0.856	27.188	0.898	27.320	0.898	28.121	0.911
Average	31.837	0.883	33.964	0.919	34.095	0.918	34.642	0.924

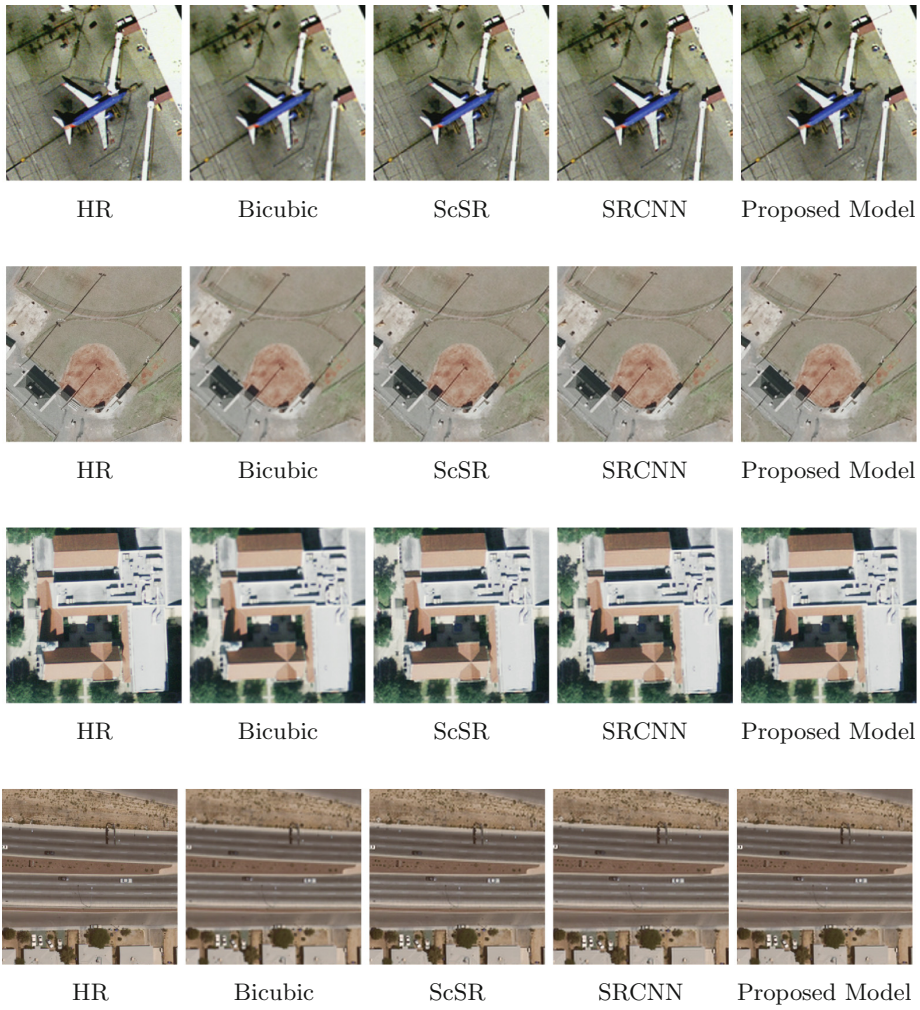

Fig. 4. Qualitative results comparing the images obtained from different algorithms.

5 Conclusions

In this paper, we present an end-to-end deep deconvolutional network based single-image SR model for optical satellite images which is trained on image patches. This is one of the preliminary study in remote sensing regarding the use of deconvolutional network for image SR. Our model produces comparable and even better performance as compared to the existing ad-hoc and deep image SR techniques. Currently, we are interested in exploring the paradigm of zero-shot SR based on deconv-net.

Acknowledgements. Biplab Banerjee was supported by Early Career Research Award from SERB India (File No: ECR/2017/000365). We gratefully acknowledge the support of Intel Corporation for giving access to the Intel®AI DevCloud platform used for this work.

References

1. Bevilacqua, M.: Algorithms for super-resolution of images and videos based on learning methods. Theses, Université Rennes 1, June 2014
2. Candès, E.: Compressive sampling. In: Proceedings of the International Congress of Mathematicians Madrid, pp. 1433–1452. European Mathematical Society Publishing House, August 2006
3. Castelluccio, M., Poggi, G., Sansone, C., Verdoliva, L.: Training convolutional neural networks for semantic classification of remote sensing imagery. In: 2017 Joint Urban Remote Sensing Event (JURSE). IEEE, March 2017
4. Clark, J., Palmer, M., Lawrence, P.: A transformation method for the reconstruction of functions from nonuniformly spaced samples. IEEE Trans. Acoust. Speech Signal Process. **33**(5), 1151–1165 (1985)
5. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 295–307 (2016). https://doi.org/10.1109/tpami.2015.2439281
6. Donoho, D.: Compressed sensing. IEEE Trans. Inform. Theory **52**(4), 1289–1306 (2006). https://doi.org/10.1109/tit.2006.871582
7. Glasner, D., Bagon, S., Irani, M.: Super-resolution from a single image. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 349–356, September 2009
8. Haut, J.M., Fernandez-Beltran, R., Paoletti, M.E., Plaza, J., Plaza, A., Pla, F.: A new deep generative network for unsupervised remote sensing single-image super-resolution. IEEE Trans. Geosci. Remote Sens. **56**(11), 6792–6810 (2018). https://doi.org/10.1109/TGRS.2018.2843525
9. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), ICCV 2015, pp. 1026–1034. IEEE Computer Society, Washington (2015). https://doi.org/10.1109/ICCV.2015.123
10. Jiang, K., Wang, Z., Yi, P., Jiang, J., Xiao, J., Yao, Y.: Deep distillation recursive network for remote sensing imagery super-resolution. Remote Sens. **10**(11), 1700 (2018). https://doi.org/10.3390/rs10111700
11. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. CoRR abs/1412.6980 (2014)
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017)
13. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML 2010, Omnipress, USA, pp. 807–814 (2010)
14. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: 2015 IEEE International Conference on Computer Vision (ICCV). IEEE, December 2015. https://doi.org/10.1109/iccv.2015.178

15. Ouyang, W., et al.: DeepID-net: deformable deep convolutional neural networks for object detection. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, June 2015
16. Park, S.C., Park, M.K., Kang, M.G.: Super-resolution image reconstruction: a technical overview. IEEE Signal Process. Mag. **20**(3), 21–36 (2003)
17. Shen, H., Li, P., Zhang, L., Zhao, Y.: A MAP algorithm to super-resolution image reconstruction. In: Third International Conference on Image and Graphics (ICIG), pp. 544–547. IEEE (2004). https://doi.org/10.1109/icig.2004.8
18. Timofte, R., De, V., Gool, L.V.: Anchored neighborhood regression for fast example-based super-resolution. In: 2013 IEEE International Conference on Computer Vision. IEEE, December 2013. https://doi.org/10.1109/iccv.2013.241
19. Tuna, C., Unal, G., Sertel, E.: Single-frame super resolution of remote-sensing images by convolutional neural networks. Int. J. Remote Sens. **39**(8), 2463–2479 (2018). https://doi.org/10.1080/01431161.2018.1425561
20. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004). https://doi.org/10.1109/tip.2003.819861
21. Yang, C.-Y., Ma, C., Yang, M.-H.: Single-image super-resolution: a benchmark. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 372–386. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_25
22. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution via sparse representation. IEEE Trans. Image Process. **19**(11), 2861–2873 (2010)
23. Yang, Y., Newsam, S.: Bag-of-visual-words and spatial extensions for land-use classification. In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS, pp. 270–279. ACM Press (2010). https://doi.org/10.1145/1869790.1869829
24. Zeiler, M.D., Krishnan, D., Taylor, G.W., Fergus, R.: Deconvolutional networks. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, June 2010. https://doi.org/10.1109/cvpr.2010.5539957
25. Zhu, X.X., et al.: Deep learning in remote sensing: a comprehensive review and list of resources. IEEE Geosci. Remote Sens. Mag. **5**(4), 8–36 (2017)