SURE-ERGAS: UNSUPERVISED DEEP LEARNING MULTISPECTRAL AND HYPERSPECTRAL IMAGE FUSION

Han V. Nguyen, Magnus O. Ulfarsson*, Johannes R. Sveinsson*, and Mauro Dalla Mura†*

*Faculty of Electrical and Computer Engineering, University of Iceland, Reykjavik, Iceland
†GIPSA-Lab, Grenoble Institute of Technology, Saint Martin d'Hères, France.

ABSTRACT

This paper proposes a new loss function to train a convolutional neural network (CNN) for multispectral and hyperspectral (MS-HS) image fusion. The loss function is based on the relative dimensionless global error synthesis (ERGAS), where we exchange the mean squared error (MSE) for its unbiased estimate using Stein's risk unbiased estimate (SURE). The loss function has a good balance between the spectral and spatial information implied by the weighted MSE, therefore it does not need a parameter to balance the spectral and spatial terms as in MSE loss function, and it also converges faster than the MSE one. Additionally, the loss function enables unsupervised training and avoids overfitting, since it is derived by using SURE. Experimental results show that the proposed method yields good results and outperforms the competitive methods. **Codes are available at https://github.com/hvn2/SURE-ERGAS**

Index Terms— Hyperspectral and multispectral image fusion, Stein's unbiased risk estimate (SURE), ERGAS, unsupervised CNN.

1. INTRODUCTION

Hyperspectral image (HSI) provides rich spectral information containing in several hundreds of contiguous spectral bands. However, it is hard to observe high spatial resolution (HR) HSI because of the sensor limitation. One solution to enhance the spatial resolution of an HSI is fusion of a low spatial resolution (LR) HSI with an HR multispectral image (MSI). The fused image has both high spatial and spectral resolution, and it usually benefits subsequent applications, such as classification, environmental monitoring and crop mapping, etc. [1].

The most straightforward MSI and HSI (MS-HS) fusion approach is to extend the traditional component substitution (CS) and multi-resolution analysis (MRA) methods in pansharpening for MS-HS fusion [2]. This approach is simple and fast, but it suffers from spectral and spatial distortion. Another MS-HS fusion approach is the model-based methods [3], where the MS-HS fusion is formulated as an inverse problem under an imaging model. The inverse problem is ill-posed, and it requires regularization (image prior) to be solved. The challenge is that it is hard to design the regularization and to fine-tune the algorithm's hyperparameters. Recently, deep learning (DL)-based MS-HS fusion approach has been proposed [4]. The DL-based methods implicitly learns the image prior using a deep neural network, e.g., a convolutional neural network (CNN), by training it with a large dataset. The DL-based methods usually outperform the traditional CS, MRA and the model-based methods in MS-HS fusion. However, there are two factors that limit the applications of those methods in practice. First, the HR HSIs (labels) are costly to produce, and training a CNN using synthesized reduced-resolution data may not be applicable in real application. Second, the deep models are sensitive to imaging model changing, i.e., the CNN must be retrained if the observed imaging model changes.

To bridge the gap between the model-based and the DL-based methods in MS-HS fusion, hybrid methods have been introduced [5, 6]. Those methods rely on the deep image prior (DIP) induced by an untrained CNN [7], which allow one to train a CNN in a similar manner as in the model-based methods, where DIP is the regularizer and a DL optimizer is the optimization algorithm. The main limitation of the DIP-based methods is overfitting for fusion of noisy HSI and MSI. One way to overcome this limitation has been proposed in the paper [8]. The main idea of this method is to train an unsupervised CNN using a loss function based on the Stein's unbiased risk estimate (SURE) [9, 10]. Applying the SURE method [8] for MS-HS image fusion needs a SURE term for the HSI (spectral) and another SURE term for the MSI (spatial). To obtain good results, the trade off between the spatial and the spectral terms should be carefully chosen. In this paper, we proposed a loss function based on SURE and the relative dimensionless error synthesis (ERGAS), which is called SURE-ERGAS. We will demonstrate that SURE-ERGAS loss function posses a good balance between the spectral and spatial terms. Therefore, training an CNN for MS-HS image fusion using the SURE-ERGAS loss function is parameter-free and yields better results than the SURE loss function [8].

This work was supported by the Icelandic Research Fund under Grant 207233-051.

2. THE SURE-ERGAS METHOD

MS-HS image fusion is the fusion of co-registered MSI and LR HSI to obtain an HR HSI. Usually, the MSI and LR HSI are assumed as the results of spectral and spatial degradation applied to the HR HSI [3], and are mathematically given by

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \boldsymbol{\epsilon},$$
$$\mathbf{G} = \mathbf{X}\mathbf{R} + \mathbf{W},$$

where $\mathbf{y}, \mathbf{x}(\mathbf{X})$ and $\mathbf{G}$ are the LR HSI, (unknown) HR HSI and MSI, respectively (note that for the ease of mathematical expression, the images are written in either vector, e.g., $\mathbf{x}$ or matrix forms, e.g., $\mathbf{X}$). The noise added to HSI is assumed to be Gaussian with zero mean and covariance matrix $\boldsymbol{\Omega}$, i.e, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$. The noise added to MSI is represented by a matrix, $\mathbf{W}$, and also assumed to be Gaussian. The spatial and spectral degradation are characterized by $\mathbf{H}$ and $\mathbf{R}$, respectively. The spatial degradation $\mathbf{H}$ is built using the point spread function and is a block-circulant-circulant-block matrix under the assumption of circulant boundary. The spectral degradation $\mathbf{R}$ contains the spectrum response that integrates the spectrum of an HR HSI to the one of MSI. In this paper, we assume that both $\mathbf{H}$ and $\mathbf{R}$ are known.

Recently, the MS-HS image fusion methods based on the DIP [7] have been given [5, 6]. DIP is implied by the fact that a CNN tends to fit a natural looking image faster than noise. Therefore, terminating training a CNN at a proper point before it overfitting gives good reconstructed image. In practice, it is hard to choose an optimal stopping point for the DIP-based methods since there are no criteria to do it. To tackle this problem, the paper [8] proposed a loss function based on SURE [9, 10] (called SURE-MSE). This loss function estimates the MSE of the fused and ground truth images, and is computed without the ground truth image. The SURE-MSE loss function [8] is given by

$$\mathcal{L}_{\text{SURE-MSE}}(\boldsymbol{\theta}) = \mathcal{L}_{\text{MSE-MSI}}(\boldsymbol{\theta}) + \lambda \mathcal{L}_{\text{MSE-HSI}}(\boldsymbol{\theta}), \qquad (1)$$

where $\boldsymbol{\theta}$ is the network parameters. The term $\mathcal{L}_{\text{MSE-HSI}}(\boldsymbol{\theta})$ is the SURE term for HSI (the spectral term), the term $\mathcal{L}_{\text{MSE-MSI}}(\boldsymbol{\theta})$ is the SURE term for MSI (spatial term), and λ is a trade-off parameter. To obtain good results, λ must be fine-tuned. For example, Fig. 1a shows the tuning results for λ using grid-search in a space of $\lambda \in \{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}\}$. The evaluation metrics are PSNR and ERGAS [11] and the dataset is the simulated Pavia University (PU), which is described in Section 3 below. Best results for the high noise and low noise cases are $\lambda = 0.001$ and $\lambda = 0.01$, respectively.

To avoid the dependency on the tuning parameter λ and to retain the advantage of the SURE-MSE loss mentioned above, we proposed a loss function based on ERGAS and SURE. ERGAS measures the global spectral and spatial distortion as

$$\text{ERGAS} = \frac{100}{r}\sqrt{\frac{1}{d}\sum_{i=1}^{d}\frac{\text{MSE}(\mathbf{x}_i, \hat{\mathbf{x}}_i)}{\mu^2_{\mathbf{x}_i}}},$$

where $\hat{\mathbf{x}}$ is the fused image obtained at the output of a CNN. $\text{MSE}(\mathbf{x}_i, \hat{\mathbf{x}}_i)$ is the MSE between $\mathbf{x}_i$ and $\hat{\mathbf{x}}_i$ for the ith band, d is the number of HSI bands, r is the resolution ratio between the HR and LR images, and $\mu_{\mathbf{x}_i}$ is the mean of the ith band of the reference image (HR HSI). Using SURE to estimate the MSE, we have

$$\text{SURE-ERGAS} = \frac{100}{r}\sqrt{\frac{1}{d}\sum_{i=1}^{d}\frac{\widehat{\text{MSE}}(\mathbf{y}_i, \hat{\mathbf{x}}_i)}{\mu^2_{\mathbf{x}_i}}}, \qquad (2)$$

where $\widehat{\text{MSE}}(\mathbf{y}_i, \hat{\mathbf{x}}_i)$ is the unbiased estimate of $\text{MSE}(\mathbf{x}_i, \hat{\mathbf{x}}_i)$, and is computed using SURE [8] as

$$\begin{aligned}\widehat{\text{MSE}}(\mathbf{y}_i, \hat{\mathbf{x}}_i) &= \|\mathbf{P}_i(\mathbf{y}_i - \mathbf{H}_i f_{\boldsymbol{\theta}}(\mathbf{z})_i)\|_2^2 \\ &\quad + 2\text{tr}\big(\sigma^2_{(p)i}\mathbf{P}_i\mathbf{H}_i\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{z})_i}{\partial \mathbf{z}}\big) - N_i\sigma^2_{(p)i},\end{aligned} \qquad (3)$$

where $\mathbf{P}_i = \mathbf{H}_i^{\dagger} = \mathbf{H}_i^T(\mathbf{H}_i\mathbf{H}_i^T)^{-1}$ is the back-projection, $\sigma_{(p)i}$ is the standard deviation of the noise of the band after applying $\mathbf{P}_i$ to the LR HSI band $\mathbf{y}_i$, and N_i is the number of pixels of $\mathbf{y}_i$. In general, (3) is not exactly the unbiased estimate of $\text{MSE}(\mathbf{x}_i, \hat{\mathbf{x}}_i)$, but (3) and $\text{MSE}(\mathbf{x}_i, \hat{\mathbf{x}}_i)$ are different up to a constant that does not depend on the network parameters [12]. In a special case, e.g., denoising, $\mathbf{H} = \mathbf{I}$, (3) becomes the unbiased estimate of $\text{MSE}(\mathbf{x}_i, \hat{\mathbf{x}}_i)$.

Similar to (1), the ERGAS-based loss function using SURE should involve a spectral and a spatial term. Using (3) and (2), we propose a SURE-ERGAS loss function as follows

$$\mathcal{L}_{\text{SURE-ERGAS}}(\boldsymbol{\theta}) = \mathcal{L}_{\text{ERGAS-MSI}}(\boldsymbol{\theta}) + \mathcal{L}_{\text{ERGAS-HSI}}(\boldsymbol{\theta}),$$

where

$$\mathcal{L}_{\text{ERGAS-HSI}}(\boldsymbol{\theta}) = \frac{100}{r}\sqrt{\frac{1}{d}\sum_{i=1}^{d}\frac{\widehat{\text{MSE}}(\mathbf{y}_i, \hat{\mathbf{x}}_i)}{\mu^2_{\mathbf{y}_i}}}, \qquad (4)$$

and

$$\mathcal{L}_{\text{ERGAS-MSI}}(\boldsymbol{\theta}) = 100\sqrt{\frac{1}{D}\sum_{i=1}^{D}\frac{\widehat{\text{MSE}}(\mathbf{g}_i, \hat{\mathbf{g}}_i)}{\mu^2_{\mathbf{g}_i}}}. \qquad (5)$$

Here, $\widehat{\text{MSE}}(\cdot)$ is computed using SURE as in (3). Note that, in (4), since $\mathbf{x}$ is unknown, we empirically replaced $\mu_{\mathbf{x}_i}$ by $\mu_{\mathbf{y}_i}$ and the results are not significantly affected. In (5), the estimated MSI, $\hat{\mathbf{g}}$, is obtained by applying the spectral degradation to the fused image, i.e., $\hat{\mathbf{G}} = \hat{\mathbf{X}}\mathbf{R}$, and the number of MSI bands is D. We would like to emphasize that the SURE-ERGAS loss function does not need a tuning parameters (i.e., λ) to control the spectral and spatial terms as in the SURE-MSE loss, and this will be analyzed in the following sections, more specifically in Fig. 1b.

3. EXPERIMENTAL RESULTS

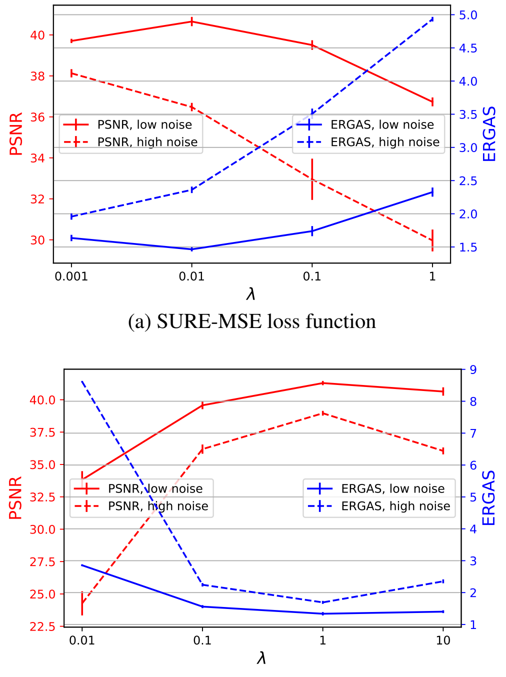
To evaluate the proposed method, we follow the experimental strategy in [8]. The simulated dataset is the PU dataset (MSI: $200 \times 200 \times 4$, HSI: $50 \times 50 \times 93$) with two cases of noise: (1) High noise: Anisotropic Gaussian noise added to HSI obtaining signal to noise ratio (SNR) of SNR (HSI) $= 11.56$ dB, (2) Low noise: Isotropic Gaussian noise added to HSI obtaining SNR $= 25$ dB, and in both cases SNR (MSI) $= 40$ dB. We use the peak SNR (PSNR) in decibels (dB), the ERGAS and the spectral angle mapper (SAM) in degrees (°) [8] to compare the results quantitatively.

We compare the SURE-ERGAS method with the SURE-MSE method [8] and a model-based method, HySure [3]. Here, the SURE-ERGAS and the SURE-MSE use the same network structure and optimization algorithm as in [8].

Fig. 1 shows the means and standard deviations of the PSNR and ERGAS over 10 runs using the SURE-MSE and SURE-ERGAS loss functions. For SURE-MSE, $\lambda = 0.001$ and $\lambda = 0.01$ give the best results for the high noise and low noise scenarios, respectively (see Fig. 1a). For SURE-ERGAS $\lambda = 1$ gives best results in both noise cases (see Fig. 1b), which means that SURE-ERGAS loss function does not need a parameter to control the spectral and spatial terms as in SURE-MSE one. The reason is most likely that the SURE-ERGAS loss function weights the MSE by the mean of each band and the resolution ratio of the HR and LR bands, therefore the spectral and spatial terms contribute equally [11].

Fig. 2 shows the ERGAS (in logarithm scale) during training obtained by the same network [8] trained with the BP-DIP (i.e., the SURE-MSE loss function without the trace term), SURE-MSE ($\lambda = 0.001$) and SURE-ERGAS ($\lambda = 1$) loss functions for the high noise case of PU dataset. The ERGAS curve of BP-DIP reaches an optimum at about 1800 iterations and raises up rapidly, which indicates overfitting. Both SURE-MSE and SURE-ERGAS avoid overfitting, since ERGAS keeps the same level after certain number of iterations. It also notices that the SURE-ERGAS loss function seems to converge faster than the SURE-MSE one.

To ensure convergence, we run the SURE-ERGAS and SURE-MSE methods for 3000 iterations in both noise cases and take the running average of the outputs as the fused images [7]. Alternatively, one can stop training based on monitoring the training loss (e.g., stop training if the loss does not decrease after a certain of iterations) to obtain the results with a trade off between running time and performance. Quantitative and qualitative results are shown in Fig. 3, where the numbers below each image are PSNR (dB), ERGAS and SAM(°), respectively. The SURE-ERGAS method outperforms competitive methods in all metrics and in both high noise and low noise cases. Fig. 3 shows the fused images along with the root mean square error based residual images for the high noise case. All methods give good results where the missing detail (high frequencies) of the LR HSI are recovered in the fused images. SURE-ERGAS yields the best fused image (see the zooming red box, where the small red dots are clearly appeared), and that is verified by the smallest error in the residual image.



(a) SURE-MSE loss function

(b) SURE-ERGAS loss function

Fig. 1: Means and standard deviations of PSNR (dB) and ERGAS over 10 runs using SURE-MSE and SURE-ERGAS loss functions with different values of λ.

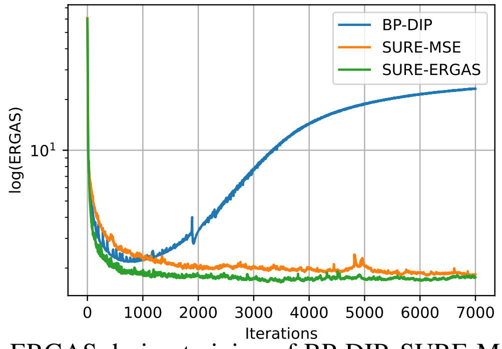


Fig. 2: ERGAS during training of BP-DIP, SURE-MSE and SURE-ERGAS loss functions using PU dataset (high noise).

4. CONCLUSION

An unsupervised DL-based method for MS-HS image fusion has been presented in this paper. We proposed a loss function based on SURE and ERGAS, which eliminate the dependency on a hyperparameter controlling the trade off between the spatial and spectral terms in the loss function based on SURE and MSE. The SURE-ERGAS loss function converges faster than the SURE-MSE loss function. Also, the

SURE-ERGAS loss function successfully avoided overfitting. Experimental results have demonstrated the efficiency of the SURE-ERGAS method that gave good fusion results and outperformed the competitive methods. The method can be applied to the general image reconstruction problems such as denoising, inpainting, and super-resolution.

5. REFERENCES

[1] L. Alparone, B. Aiazzi, S. Baronti, and A. Garzelli, *Remote Sensing Image Fusion*. CRC Press, 2015.

[2] G. Vivone, R. Restaino, G. Licciardi, M. Dalla Mura, and J. Chanussot, "Multiresolution analysis and component substitution techniques for hyperspectral pansharpening," in *2014 IEEE Geoscience and Remote Sensing Symposium*, 2014, pp. 2649–2652.

[3] M. Simoes, J. Bioucas-Dias, L. B. Almeida, and J. Chanussot, "A convex formulation for hyperspectral image superresolution via subspace-based regularization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 6, pp. 3373–3388, 2014.

[4] Y. Li, J. Hu, X. Zhao, W. Xie, and J. Li, "Hyperspectral image super-resolution using deep convolutional neural network," *Neurocomputing*, vol. 266, pp. 29–41, 2017.

[5] T. Uezato, D. Hong, N. Yokoya, and W. He, "Guided deep decoder: Unsupervised image pair fusion," in *European Conference on Computer Vision*. Springer, 2020, pp. 87–102.

[6] J. Yao, D. Hong, J. Chanussot, D. Meng, X. Zhu, and Z. Xu, "Cross-attention in coupled unmixing nets for unsupervised hyperspectral super-resolution," in *European Conference on Computer Vision*. Springer, 2020, pp. 208–224.

[7] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9446–9454.

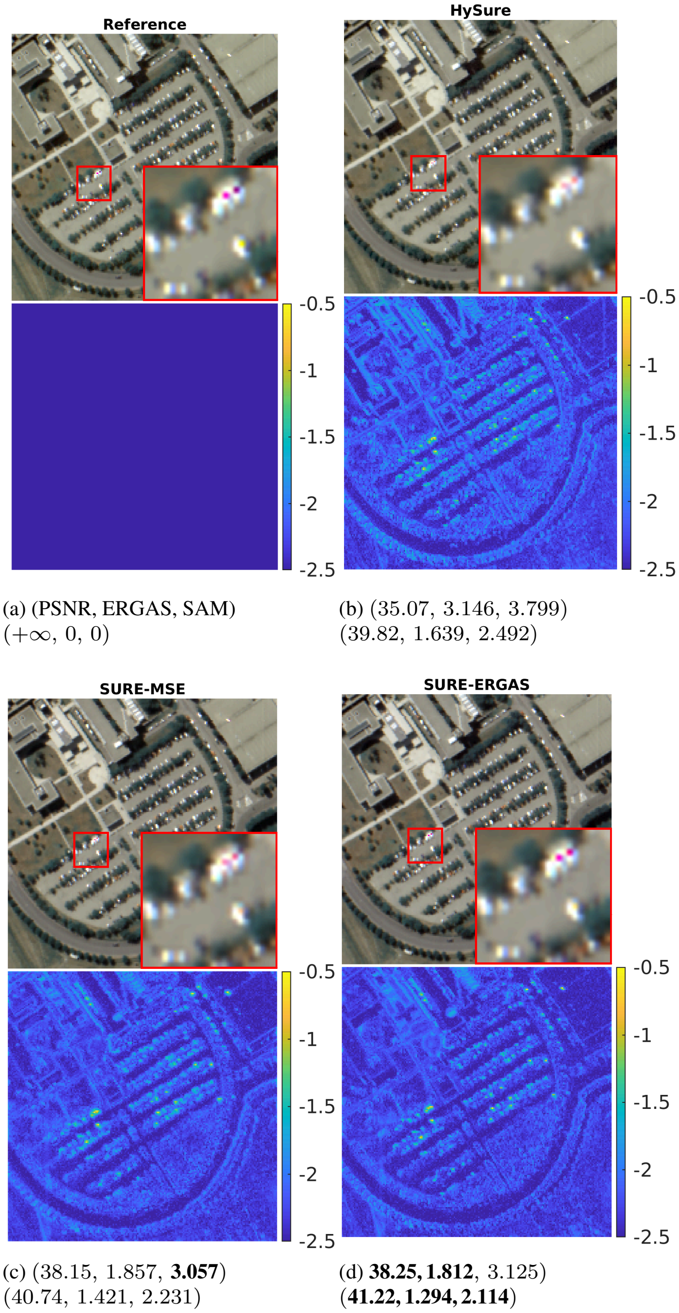
[8] H. V. Nguyen, M. O. Ulfarsson, J. R. Sveinsson, and M. Dalla Mura, "Deep SURE for unsupervised remote sensing image fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[9] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *The Annals of Statistics*, pp. 1135–1151, 1981.

[10] V. Solo, "A sure-fired way to choose smoothing parameters in ill-conditioned inverse problems," in *Proceedings of 3rd IEEE International Conference on Image Processing*, vol. 3, 1996, pp. 89–92.

[11] 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.

[12] Y. C. Eldar, "Generalized SURE for exponential families: Applications to regularization," *IEEE Transactions on Signal Processing*, vol. 57, no. 2, pp. 471–481, 2008.



(a) (PSNR, ERGAS, SAM)
$(+\infty, 0, 0)$

(b) (35.07, 3.146, 3.799)
(39.82, 1.639, 2.492)

(c) (38.15, 1.857, **3.057**)
(40.74, 1.421, 2.231)

(d) **38.25, 1.812**, 3.125)
(**41.22, 1.294, 2.114**)

Fig. 3: Reference and fused images (high noise) shown in false color images and RMSE-based residual images for all methods. The numbers below each images are PSNR (dB), ERGAS and SAM (°) for the high noise (first rows) and low noise (second rows).