# Joint Compression and Demosaicking for Satellite Images

Pascal Bacchus, Renaud Fraisse, Aline Roumy, Christine Guillemot

# JOINT COMPRESSION AND DEMOSAICKING FOR SATELLITE IMAGES

*Pascal Bacchus*[1]*, Renaud Fraisse*[2]*, Aline Roumy*[1]*, Christine Guillemot*[1]

[1]INRIA, Rennes, France, [2]Airbus Defence and Space, Toulouse, France
[1] *name.surname*@inria.fr, [2] renaud.fraisse@airbus.com

## ABSTRACT

Image sensors used in real camera systems are equipped with colour filter arrays which sample the light rays in different spectral bands. Each colour channel can thus be obtained separately by considering the corresponding colour filter. While existing compression solutions mostly assume that the captured raw data has been demosaicked prior to compression, in this paper, we describe an end-to-end trainable neural network for joint compression and demosaicking of satellite images. We first introduce a training loss combining a perceptual loss with the classical mean square error, which is shown to better preserve the high-frequency details present in satellite images. We then present a multi-loss balancing strategy which significantly improves the performance of the proposed joint demosaicking-compression solution.

***Index Terms*—** Deep Image Compression, Neural Networks, Demosaicking

## 1. INTRODUCTION

With the new generation of on-board satellite cameras, images with increased spatial and spectral resolution can be acquired, leading to huge amount of data that needs to be transmitted on ground. Therefore, efficient algorithms need to be designed to compress these remote-sensing images.

Designing efficient compression algorithms for satellite images must take into account several constraints. First, (i) it must be well *adapted to the raw data format*. In particular, we consider in this work, the cameras for the Lion satellite constellation with ultra-high spatial resolution at the price of a lower spectral resolution. More precisely, the three spectral bands (RGB) are acquired with a single sensor with an in-built filter array. Second, (ii) it should be *adapted to the image statistics*. Indeed, satellite mages contain very high-frequency details with small objects spread over very few pixels only. Finally, (iii) the compression must be quasi-lossless to allow accurate on-ground interpretation. In this work, we propose a compression algorithm that can efficiently deal with these three constraints.

Regarding adaptation to raw data format (i), each pixel only measures the intensity of one colour band, according to the Bayer pattern. The usual image acquisition pipeline consists of two steps [1]. First, the three colour channels for each pixel of the colour images are reconstructed, using demosaicking algorithms [2]. Then, the colour image is compressed. In this paper instead, we propose to perform the two steps (demosaicking and compression) jointly. The potential gain is twofold. The processing is more efficient. Indeed, the joint processing avoids adding redundancy (demosaicking) first and removing it in the compression step. Moreover, learning-based demosaicking algorithms tend to add high-frequency details [3], which may increase the data rate. The joint processing will instead add these details when reconstructing the data, which will not impact the data rate. Such joint processing already applies to compression and denoising operations with combined results superior to sequential results [4, 5, 6].

Another key challenge is to be able to adapt to the statistics of satellite images, which differ from natural images acquired with a handheld perspective camera. This adaptation can be made thanks to variational auto-encoders VAE. Indeed, VAE have first been introduced to learn end-to-end compression algorithms for natural images [7, 8, 9, 10] and eventually outperform traditional codecs [11, 12]. Adaptation to the high frequency details of image satellite images has been proposed in [13] based on attention module. Here, we further improve our results, by designing a new loss based on perceptual metrics. We then present a multi-loss balancing strategy, which improves the overall performance. Note that the use of AEs for satellite image compression has also been explored in [14] to reduce the complexity of AE based architectures such as [9]. Finally, to meet the last constraint (iii) of quasi-lossless compression, we perform the learning at rather high bit rate.

## 2. RAW DATA COMPRESSION

In this section, we present our global architecture that performs joint compression and demosaicking to be able to adapt to the raw data format. We first review the state-of-the-art compression algorithm also called hyper-prior architecture [9]. We finally describe our proposed loss.

The hyper-prior architecture [9] is composed of two AE networks as shown in Figure 1. The first AE produces a latent representation $y$ of the input data $x$. Standard compression

This work was supported by BPI France Lichie project.

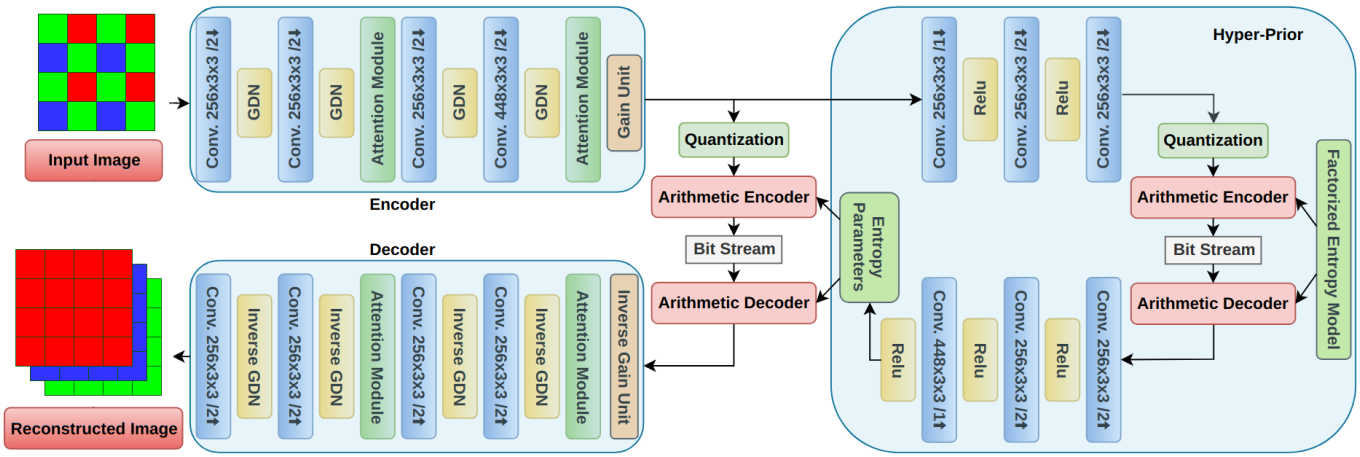



**Fig. 1**. Joint compression-demosaicking architecture with raw data as input [13]

operations such as quantization and entropy coding are performed on this latent representation to produce a bit-stream, which is then decoded by the entropy decoder as $\hat{y}$. The decoder reconstructs the signal $\hat{x}$ with inverse transforms. The other AE (the hyper-prior) models the parameters of the latent representation distribution to enhance the entropy model. This shared entropy model is adapted to the characteristics of this input data as the entropy parameters are re-estimated at each input.

We employ the learning capabilities of neural networks to exploit correlation in raw data since joint processing already yields better results than separate demosaicking and denoising [3, 15]. The training is performed with a pair of ground-truth and Bayer filter images (GRBG pattern) with the ground truth as a reference for loss computation.

This model is designed to target high bit rate due to high quality reconstruction requirements. To not have performance drops when increasing the bit rate with very detailed images, the number of filters is set to 448 in the bottleneck layer. The resulting number of features in the latent representation is of the same order of magnitude as in the input image to preserve details and keep necessary information for demosaicking.

We apply a scaling [16] before quantization which acts as a quality parameter at run time so that the model can perform well within a small bit rate range around the target bit rate. It allows for more flexibility as trained models are no longer blocked to a fixed rate-distortion point as it is the case with most [9, 10, 11].

## 3. SATELLITE CHARACTERISTICS AWARE COMPRESSION

### 3.1. Loss to preserve high frequency details

The classic rate-distortion trade-off for compression optimization problems is used to create a loss function from which a gradient descent algorithm is derived. As the derivative of the quantization function is zero or undefined we replace it with uniform noise for training.

The main source of error in our reconstructed images comes from high-frequency stripped patterns [13]. They have a spatial frequency at the pixel size and disappear due to the blur generated by the distortion metric, the $l2$ Euclidean norm. To better fit the data characteristics and better preserve high-frequency details during compression we have incorporated perceptual metrics in the loss function. This metric differs from pixel-based metrics as it aims at minimizing an error over some extracted features. This perceptual loss is used in a wide range of applications [17, 18] to generate more realistic textures and sharper edges in image processing problems.

We define a loss function based on VGG [19] to extract structures inside our features and guide the learning towards better high-frequency reconstruction.

$$P(x,\hat{x}) = \frac{1}{nm}(VGG_{0:2}(x,\hat{x})^2 + VGG_{0:4}(x,\hat{x})^2) \quad (1)$$

We decide to use early layers of VGG as they oversee the learning of low-level spatial features [18] while deeper layers focus on more abstract features. Since our problem is more detail-oriented we use the first four layers of VGG to extract two sets of features. We compute the $l2$ norm between the ground truth and reconstructed features.

$$\mathcal{L} = \lambda_a D(x,\hat{x}) + \lambda_b P(x,\hat{x}) + \alpha R(\hat{y}) \quad (2)$$

$\alpha$ is set to target a bit rate and controls the rate-distortion trade-off.

### 3.2. Multi-loss balancing

Learning for multiple tasks, with their respective loss functions, can lead to a better result for all tasks than learning for

each task individually as shown in [20] where semantic classification and depth estimation induce better performances together than separately. However it implies adding more balancing terms in the loss function and since hyper-parameters are troublesome to tune, it becomes harder to optimise the network.

A solution to set the different loss parameters is to jointly tune all parameters inside the loss function [21] with an automatically controlled scheme. It removes the need for manual tuning and ensures an optimal trade-off between all loss terms [22]. The loss function becomes:

$$\mathcal{L} = \lambda_1 D'(x, \hat{x}) + \lambda_2 P'(x, \hat{x}) + \alpha R(\hat{y}) \quad (3)$$

with $D' = \lambda_a.D; P' = \lambda_b.P$

To automatically evaluate the $\lambda_k$ we are following the dynamic weight average approach [23] to compute at each epoch a new $\lambda_k$ based on previous loss measures for the distortion and perceptual metric:

$$\lambda_k = K.\frac{exp(\frac{w_k(t-1)}{T})}{\sum_i exp(\frac{w_i(t-1)}{T})}, w_k = \frac{L_k(t-1)}{L_k(t-2)} \quad (4)$$

Each Loss $L_k$ is linked to its corresponding $\lambda_k$. $T$ measures the softness of the process with the analogy with the annealing temperature, and controls how close can the different values $\lambda_k$.

## 4. EXPERIMENTS

### 4.1. Training details

The data set used includes 300 12-bits RGB satellite images (2000x2000) with 50cm geometric resolution as in [13], 5% are used for testing, the rest for training. Raw data are obtained with the Bayer filter applied to ground-truth images to form the necessary training pair for supervised learning. Every batch of images is cropped into patches and randomly augmented with rotation to provide rotational invariance. The networks have been designed using the CompressAI [24] Python library, a PyTorch overlay for neural network compression models.

We use reference methods close to the performance of onboard satellites. For compression, we consider JPEG 2000 as it is similar to the standard used for RGB images [25] using DCT transforms. For demosaicking, we use the linear filter proposed by Malvar [26] which gives good results while being simple. It gives the highest PSNR compared to other traditional demosaicking algorithms [2] while being visually sharp, which diminishes the amount of blur added to the processed image.

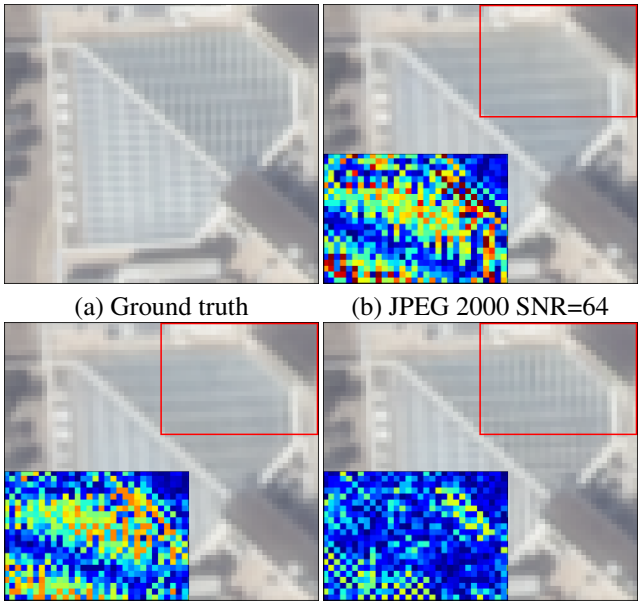
For the joint compression-demosaicking scheme with both the MSE distortion and the VGG perceptual loss, both $\lambda_a$ and $\lambda_b$ are set to have the distortion and perceptual metrics at the same order of magnitude.

$$\lambda_a = 2.6 * 10^6; \lambda_b = 10^4$$

The relationship between $\alpha$ and the target bit rate is empirical. $\alpha$ is set to 0.6 for all experiments to target 2bpp for the reconstruction to be of sufficient quality for satellite applications. Experiments were conducted on NVIDIA A40 GPUs for 200 epochs. Inference time is around 1s for the encoder and 1.5s for the decoder.

### 4.2. Qualitative results

Figure 2 shows visual results obtained with different methods in comparison with the ground truth, for a 50cm geometric resolution satellite image of a city landscape. The ground truth image is compressed at 2bpp with the reference baseline JPEG 2000 with Malvar demosaicking, the joint compression/demosaicking network with multi-loss balancing and the joint compression/demosaicking network without the VGG perceptual loss. All images are well reconstructed since we target a high bit rate. Nevertheless, high-frequency details such as the stripped patterns on this rooftop's building have disappeared for models (b) and (c). Model (d) with the perceptual loss balanced with MSE is close to recovering all those details as early layers of VGG brought more weight to the structure reconstruction. When zooming and analyzing the pixel value difference to the ground truth, we see the impact of a higher SNR for learned models on the image quality. The pixel difference to the ground truth is much lower even if this is hard to perceive at that bit rate.



(a) Ground truth (b) JPEG 2000 SNR=64

(c) MSE model SNR=85 (d) Multi-Loss model SNR=105

**Fig. 2**. Visual comparison of compressed images at 2 bpp with the ground truth. An error map shows the relative difference with the ground truth at a pixel level (range [0;32]).

### 4.3. Quantitative results

We first evaluate the efficiency of the joint processing model with sequential models in Figure 3. The sequential processing used is close to satellite imaging standards with JPEG 2000 as codec [25]. The joint model achieves huge bit rate gain at a constant quality and outperforms both sequential models. Those data-driven models excel at extracting information from irregular data. The joint model can also reach reconstruction quality not feasible for any sequential models.

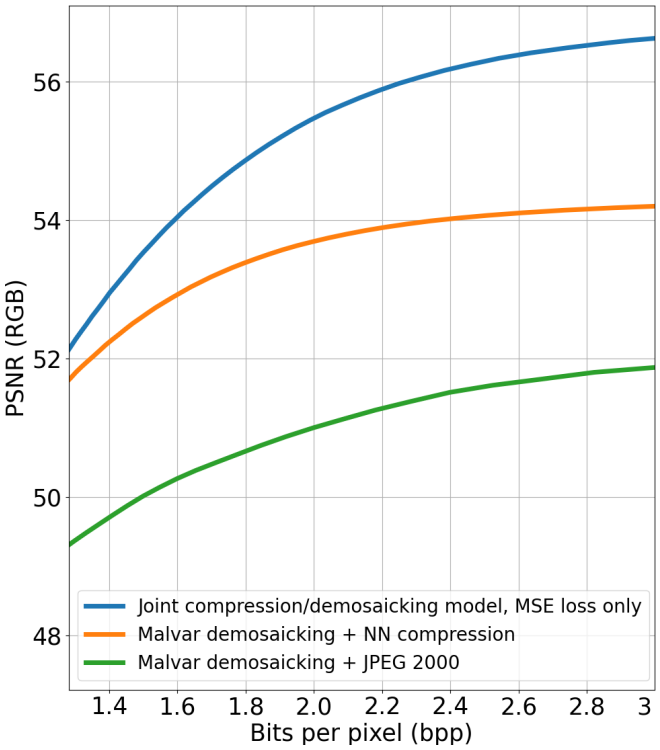



**Fig. 3**. Effect of the joint processing of compression and demosaicking compared to sequential processing. The joint model does not use perceptual loss and multi-loss balancing.

We then assess the performance gain that perceptual loss and multi-loss balancing bring to the joint model in Figure 4. We compare our model for different loss functions based on MSE, VGG or both losses and with multi-loss balancing when performed during training. VGG alone still performs decently with the SNR metric even though it is not tailored to the MSE distortion. When trained only with MSE, the network has unsurprisingly better results when evaluated using the SNR metric. Both metrics combined lead to even better performances, particularly at high bit rates. This combination between an optimized metric on MSE and a metric focused on extracting structure reduces the blurring effects induced by the compression scheme. The multi-loss balancing scheme brings the previous model to a better rate-distortion trade-off over the whole bit rate range. During training, the parameters $\lambda_k$ adapt to the relative importance given to their respective task in previous epochs. This enables the network to escape some local minima as the main $\lambda_k$ leading the gradient changes over time.

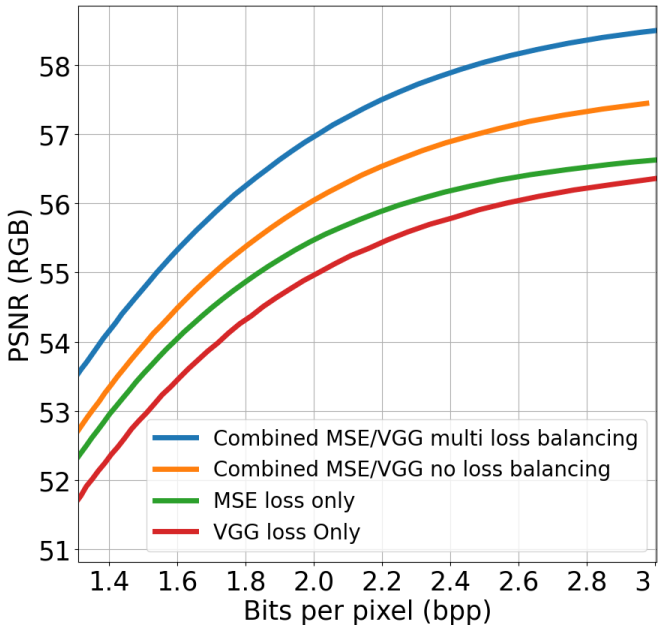



**Fig. 4**. Effect of the loss functions and multi-loss balancing on the joint compression/demosaicking model performances.

## 5. CONCLUSION

In this work, we have proposed a joint compression and demosaicking model designed for raw RGB satellite images with increased rate-distortion performance compared to traditional sequential processing. The reconstruction is further improved with the addition of a perceptual metric to extract high-frequency structures, and the multi-loss strategy to tune each loss function parameter. The next step is to adapt this type of joint processing to other colour filter arrays than the standard Bayer filter and to add other processing tasks such as denoising for an extended processing pipeline.

## 6. REFERENCES


[1] T. Wiegand and H. Schwarz, *Video Coding: Part II of Fundamentals of Source and Video Coding*, 01 2016.

[2] P. Getreuer, "Malvar-he-cutler linear image demosaicking," *Image Process. Line*, vol. 1, 2011.

[3] M. Gharbi, , G. Chaurasia, S. Paris, and F. Durand, "Deep joint demosaicking and denoising," vol. 35, no. 6, dec 2016.

[4] S. R. Alvar, M. Mateen, H. Choi, and I. V. Bajić, "Joint image compression and denoising via latent-space scalability," 2022.

[5] K. L. Cheng, Y. Xie, and Q. Chen, "Optimizing image compression via joint learning with denoising," in *Proceedings of the European Conference on Computer Vision*, 2022.

[6] V. Alves de Oliveira, M. Chabert, T. Oberlin, C. Poulliat, M. Bruno, C. Latry, M. Carlavan, S. Henrot, F. Falzon, and R. Camarero, "Satellite image compression and denoising with neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, 2022.

[7] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *ICLR*, 2017.

[8] L. Theis, W. Shi, A. Cunningham, and F. Huszár, "Lossy image compression with compressive autoencoders," in *ICLR*, 2017.

[9] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *ICLR*, 2018.

[10] D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," in *NeurIPS*, 2018.

[11] D. Minnen and S. Singh, "Channel-wise autoregressive entropy models for learned image compression," in *ICIP*, 2020.

[12] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned image compression with discretized gaussian mixture likelihoods and attention modules," in *CVPR*, 2020.

[13] P. Bacchus, R. Fraisse, A. Roumy, and C. Guillemot, "Quasi lossless satellite image compression," in *IGARSS 2022*, 2022, pp. 1532–1535.

[14] V. Alves de Oliveira, M. Chabert, T. Oberlin, C. Poulliat, M. Bruno, C. Latry, M. Carlavan, S. Henrot, F. Falzon, and R. Camarero, "Reduced-complexity end-to-end variational autoencoder for on board satellite image compression," *Remote Sensing*, vol. 13, no. 3, 2021.

[15] F. Kokkinos and S. Lefkimmiatis, "Iterative residual network for deep joint image demosaicking and denoising," *CoRR*, vol. abs/1807.06403, 2018.

[16] T. Dumas, A. Roumy, and C. Guillemot, "Autoencoder based image compression: Can the learning be quantization independent?," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 1188–1192.

[17] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," *CoRR*, vol. abs/1801.03924, 2018.

[18] M. Saeed Rad, B. Bozorgtabar, U. V. Marti, M. Basler, H. Kemal Ekenel, and J. P. Thiran, "SROBB: targeted perceptual loss for single image super-resolution," *CoRR*, vol. abs/1908.07222, 2019.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014.

[20] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7482–7491, 2018.

[21] M. Crawshaw, "Multi-task learning with deep neural networks: A survey," *CoRR*, vol. abs/2009.09796, 2020.

[22] R. Bischof and M. Kraus, "Multi-objective loss balancing for physics-informed deep learning," *CoRR*, vol. abs/2110.09813, 2021.

[23] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1871–1880, 2019.

[24] J. Bégaint, F. Racapé, S. Feltman, and A. Pushparaja, "Compressai: a pytorch library and evaluation platform for end-to-end compression research," *arXiv preprint arXiv:2011.03029*, 2020.

[25] Consultative Committee for Space Data Systems (CCSDS), *Image data compression CCSDS 122.0-B-1*, CCSDS, 2005.

[26] H.S. Malvar, Li wei He, and R. Cutler, "High-quality linear interpolation for demosaicing of bayer-patterned color images," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004, vol. 3, pp. iii–485.