DeepVibes: Correcting Micro-Vibrations in Satellite Imaging with Pushbroom Cameras

Minh Hai Nguyen, François de Vieilleville, Pierre Weiss

DeepVibes: Correcting Micro-Vibrations in Satellite Imaging with Pushbroom Cameras

Minh Hai Nguyen, François de Vieilleville and Pierre Weiss

***Abstract*—In this paper, we propose new algorithms for estimating micro-vibrations and correcting its effects in satellite imaging with linear pushbroom camera. We first design an accurate model of the acquisition process with linear pushbroom camera, that incorporates the satellite attitude as parameters. We then propose a two stage reconstruction method based on an identification neural network to identify the attitude, followed by a deep unrolled network to correct the micro-vibrations. We then evaluate the proposed framework on synthetic and real data, showing promising results for this challenging problem. Our results highlight the critical role of the focal plane's geometry, to improve the micro-vibrations identifiability and therefore, the reconstruction quality.**

***Index Terms*—Pushbroom, micro-vibration, jitter, satellite imaging, convolutional neural network, unrolled network.**

I. Introduction

PUSHBROOM cameras are widely used for Earth observation as they provide high-resolution images and large surveillance area, up to 700 000 square kilometers daily [1]. Many commercial satellites use this kind of camera, such as Pléiades 1A and 1B, Sentinel-2, WorldView 1, 2 and 3 or QuickBird. Linear pushbroom camera consists of one or multiple linear arrays of thousands of small sensors, that work collectively to deliver satellite images with large swath. They are acquired by scanning the Earth surface line by line, as the satellite is moving on its orbit. The image quality can however be compromised by camera vibrations, which can originate from onboard mechanical equipment, such as gyroscopic actuators. They affect the camera orientation, leading to geometric distortions or misalignment of lines, see Fig. 3 and [2]. The resulting geometric distortions are often called jitter. The frequencies and amplitudes of these vibrations vary across different satellites. In this paper, we are particularly interested by vibrations with small amplitudes, often referred to as *micro-vibrations*.

Minh Hai Nguyen is with Institut de Recherche en Informatique de Toulouse and Centre de Biologie Intégrative, Laboratoire MCD, CNRS & Université de Toulouse, France, and also with Agenium Space, Toulouse, France (e-mail: minh-hai.nguyen@univ-tlse3.fr).

François de Vieilleville is with Agenium Space, Toulouse, France (e-mail: francois.devieilleville@agenium.com).

Pierre Weiss is with Institut de Recherche en Informatique de Toulouse and Centre de Biologie Intégrative, Laboratoire MCD, CNRS & Université de Toulouse, France (e-mail: pierre.armand.weiss@gmail.com).

A. Related works

Different methods have been proposed for detecting and correcting the vibration effects in satellite images. Hardware components or extra information can be used to determine the vibrations, such as high-performance attitude and orbit control system (AOCS) [3] or ground control points (GCPs) [4]. Methods based on AOCS can be inaccurate for small satellites [5]. Furthermore, the frequency of onboard AOCS or inertial measurement units [6] can be lower than the frequency at which images are sampled. For instance, the Sodern star trackers are currently limited to 30Hz, and higher frequency vibrations cannot be corrected with this approach. In particular it is worth noting that for LEO sun synchronous satellite providing very high resolution imagery at 1.5 meter GSD, 0.2ms are required per line acquisition. To provide an attitude for each acquired line would require a star tracker system providing a sampling rate fairly above 1kHz. This discrepancy leads to a situation where the attitude information is not fully available, requiring the development of post-processing methods. In [7], the authors proposed an approach to correct jitter effects based on Tikhonov regularization and a simple linearized forward model. Alternatively, [8] proposed an approach based on the acquisition of a couple of images with different viewing angles.

In this study, we focus the feasibility of correcting single multi-spectral images. We consider a self-calibrated approach, where the satellite's attitude is retrieved using only the vibrated image. One of the pioneering approaches in this category is the work by Perrier et al [7]. There, a Bayesian approach with quite simple image priors was developed. More recently, supervised learning methods have been proposed [9], [10]. The general ideas is to use neural networks to identify and correct the satellite vibrations. These works are close in spirit to the ones we propose, but differ by a few important aspects, as argued below.

B. Contributions

The main contributions of this paper are as follows:

- We design an accurate mathematical model of the acquisition process of linear pushbroom camera under micro-vibrations, while previous works considered simple interpolation models.
- We design a specific neural network architecture that adapts to the size of the input image.
- We are primarily interested in micro-vibrations, which are significantly harder to detect and correct.
- We demonstrate the performance of our method on synthetic and real data.
- The analysis reveals that the attitude can be efficiently estimated and corrected, if the distance between the different spectral bands on the focal plane is sufficiently large, further strengthening the results of [7].

Overall, this computational imaging study suggests that the open problem of micro-vibration correction can be addressed satisfactorily with purely numerical methods, given that some care is given when arranging the different CCD sensors on the camera.

II. Methods

The problem of micro-vibration estimation and correction can be seen as a blind inverse problem. From a mathematical perspective, we can view the acquisition process as a linear operator $A(\xi)$:

$$\begin{array}{rccc} A(\xi): & \mathcal{U} & \longrightarrow & \mathcal{V} \\ & u & \mapsto & A(\xi)u. \end{array}$$

The space $\mathcal{U}$ is the vector space of "ideal" images, that we would like to reconstruct. A good candidate is $\mathcal{U} = L^2(\mathbb{R}^2)$, the space of finite energy signals, for which convolution operators are well defined. This space will be discretized for numerical computation. The space $\mathcal{V} = \mathbb{R}^{N_x \times N_y}$ is the space of discrete images acquired by the satellite when sampling N_x lines with N_y sensors on each array. The vector ξ describes the satellite's attitude at each sampling time. In what follows, we let v denote the vibrated image. It can be related to the ground truth image u using the following model:

$$v = \mathcal{P}\left(A(\xi)u\right), \tag{1}$$

where $\mathcal{P}$ describes some sensors induced perturbations. Our objective in this paper can be formulated as follows.

Problem 1 (micro-vibration detection and correction)**.** *Given an observed image $v \in \mathcal{V}$, find a sharp image $\hat{u} \in \mathcal{U}$ and some attitude $\hat{\xi}$ such that $A(\hat{\xi})\hat{v} \approx u$.*

In this section, we properly define the operator $A(\xi)$, the degradation process $\mathcal{P}$ and then propose numerical algorithms to recover ξ and u from v only.

A. Modeling the acquisition process

Proper physical models are deemed essential for accurate satellite image correction [11]. The proposed model shares similarities with the one in [4]. It accounts for the observation geometry, for an accurate optical model and the noise model.

Assumption 1 (Main assumptions)**.** *Our work is based on the following assumptions:*

- *The satellite is traveling along a straight line at a constant velocity with respect to the ground frame.*
- *The region observed by the camera is planar.*

1) The observation geometry: In this paragraph, we describe how a point on the ground relates to a point on the focal plane of the camera.

a) Camera attitude: Without vibrations, the camera is oriented along the local orbital frame: its origin is the satellite's center of mass. The $z-$axis points towards the center of the Earth and the $x-$axis points towards the satellite moving direction. The $y-$axis is determined by requiring that the axes form an orthonormal right-handed coordinate frame, see Fig. 1. Micro-vibrations make the camera orientation (or attitude) change over time. This can be modeled by camera rotations around the three axes: roll ($x-$axis), pitch ($y-$axis) and yaw ($z-$axis) with the corresponding angles (ω, ϕ, κ). In what follows, we will refer to this triplet as attitude or rotation angles. The three elementary rotations can be applied using Givens rotation matrices. The final rotation is obtained by composing them:

$$R(\omega,\phi,\kappa) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & c_\omega & -s_\omega \\ 0 & s_\omega & c_\omega \end{bmatrix} \begin{bmatrix} c_\phi & 0 & s_\phi \\ 0 & 1 & 0 \\ -s_\phi & 0 & c_\phi \end{bmatrix} \begin{bmatrix} c_\kappa & -s_\kappa & 0 \\ s_\kappa & c_\kappa & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

where c_a and s_a denote $\cos a$ and $\sin a$, respectively. The rotation matrix allows us to pass from the local orbital frame to the camera frame, as shown in Fig. 1.

b) Localization function: At time t, we assume that the satellite is at position $(x_t, 0, h)$ in the ground frame. We let f denote the focal length of the camera, $\xi_t = (\omega(t), \phi(t), \kappa(t)) \in \mathbb{R}^3$ its attitude and $R_t = R(\omega(t), \phi(t), \kappa(t))$ the associated rotation matrix.

Definition 1 (Localization function)**.** *The function that maps a location on the focal plane of the camera with a location on the ground is called* localization function*. It is illustrated in Fig. 2 and defined by*

$$L_{\xi_t} : (x, y) \in \mathbb{R}^2 \mapsto \left(x_t + h\frac{x_t}{z_t}, -h\frac{y_t}{z_t} \right)$$

where $(x_t, y_t, z_t)^T = R_t \cdot (x, y, f)^T$.

2) The sampling process: In this paragraph, we describe the precise forward acquisition model. We focus on the multi-spectral sensor of the Pléiades satellite.

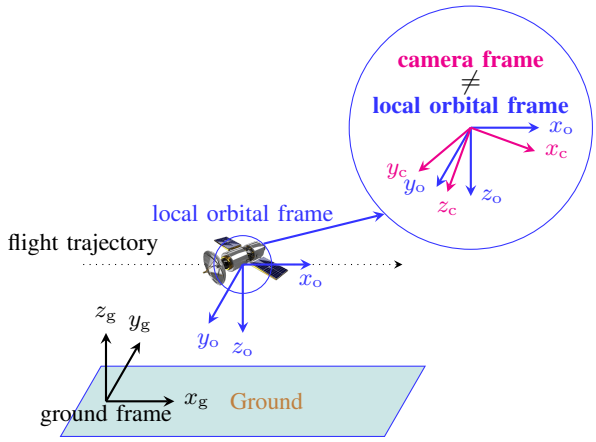


Fig. 1: The orientation of the ground and local orbital frames are fixed, while the orientation of the camera frame may change over time and differ from the local orbital frame.

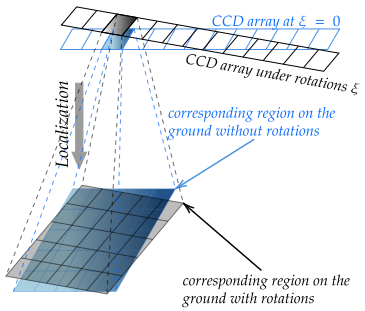


Fig. 2: Acquisition under rotations. A pixel on the CCD array is discretized as a uniform grid on the ground for sampling and integration.

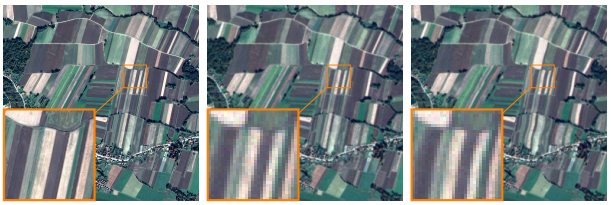
a) Optical Point Spread Function (PSF): The camera suffers from diffraction and therefore produces band-limited images. We model the diffraction by a convolution with a point spread function (PSF) $k \in L^2(\mathbb{R}^2)$. Since the latter is hardly accessible in practice [12], we simply approximate it by a centered Gaussian kernel with standard deviation $\sigma = 0.27$. This model was proposed and calibrated in [13] for Pléiades satellites.

b) Sensor geometry: The light is integrated on each CCD detector of the camera. We assume that they are all squares of identical edge-length δ. We let Ω_j denote the support of the j-th sensor in the camera frame:

$$\Omega_j = [-\delta/2, \delta/2] \times [y_j - \delta/2, y_j + \delta/2],$$

where $y_j = \delta j$ for $j \in \{-N_y/2, \dots, N_y/2 - 1\}$. In Pléiades satellites, we have $\delta = 52$ µm.

c) Forward model: We let $(t_1, \dots, t_{N_x})$ denote the sampling times. At each instant t_i, the satellite acquires one "line" on the ground. We let $\xi = (\xi_{t_1}, \dots, \xi_{t_{N_x}})$ denote the associated satellite's attitude. The operators



Ground image | Acquired image with $\xi = 0$ | Acquired image with $\xi \neq 0$

Fig. 3: Example of simulated images with 1 pixel error.

$A(\xi)$ can now be defined pixel-wise as:

$$[A(\xi)u][i,j] = \int_{\Omega_{i,j}} \int_{\mathbb{R}^2} u\left(L_{\xi_{t_i}}(x - s_x, y - s_y)\right) \quad (3)$$
$$k(s_x, s_y)\, ds_x\, ds_y\, dx\, dy$$

where $\Omega_{i,j} = \Omega_j - (x_{t_i}, 0)$. In words, the above equation indicates that we distort the image u using the localization function, we then convolve it with the PSF k of the camera and then integrate the light for every CCD sensor.

From a numerical viewpoint, the above integrals are replaced by discretized versions. We use a cubic convolution algorithm to interpolate the discretized image u and replace the integrals by Riemann sums. We therefore need to discretize the CCD sensor as illustrated on Fig. 2. In our numerical experiments, each CCD sensor is replaced by a 5×5 grid.

3) The noise model: The complete forward model reads $v = \mathcal{P}(A(\xi)u)$ where $\mathcal{P} : \mathcal{V} \to \mathcal{V}$ is a random perturbation. It takes the form $\mathcal{P}(v_0) = Q_{12}(v_0 + \epsilon)$, where Q_{12} denotes a 12-bits quantifier and ϵ is additive noise.

The noise distribution on CCD sensors can be well approximated by Poisson-Gaussian distributions. In this paper, we approximate the Poisson distribution by an additive heteroskedastic (space varying variance) Gaussian distribution. The noise $\epsilon \in \mathbb{R}^{N_x \times N_y}$ is drawn with the distribution $\mathcal{N}(0, a + b \odot u)$, where $\odot$ denotes the element-wise product. The values a and b were calibrated in [14] as $a = 3.24, b = 0.037$. The constant a models a noise with constant standard deviation modeling the amplification chain noise and darkness noise. The constant b models the Poisson noise amplitude.

4) Yaw versus Pitch and Roll: The camera rotations induce displacements with respect to a reference frame on the ground. Following [4], [15], it turns out that the displacements due to the yaw are negligible compared to the ones related to the pitch and roll when the camera has nadir view. For example, in Pléiades satellites, an error of 4×10^{-6} rad of roll or pitch results in a displacement of 1 pixel at nadir. An error of yaw with the same magnitude produces a displacement of only 6.76×10^{-5} pixel.

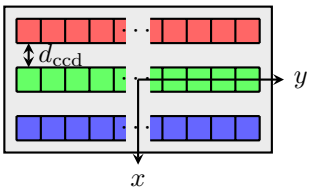


Fig. 4: Multiple linear CCD arrays on the focal plane. d_{ccd} denotes the intra-distance.

5) Multi-sensor arrays: In most of satellites, several linear pushbroom cameras are placed in parallel on the focal plane of the satellite. Those cameras record different spectral bands. For simplicity, we focus on three linear CCD arrays corresponding to red, green and blue band of RGB format, as in Fig. 4. We can associate an operator A^r, A^g, A^b as in (3) for each of them. The only difference is that we need to replace the sensor locations Ω_j by Ω_j^r, Ω_j^g, Ω_j^b. Each of them is shifted in the x direction by a different amount d_{ccd}.

Due to the geometry of the focal plane, the three cameras are not observing the same part of the scene on the ground, thus the information captured in different channel is shifted along the $x-$axis.

The space of discrete images observed by satellite is $\mathcal{V} = \mathbb{R}^{C\times N_x\times N_y}$, where $C = 3$ is the number of spectral bands.

6) Simulating the satellite's attitude: The attitude can be represented by a sum of several sinusoidal and cosinusoidal functions of certain frequencies [7], [9]. We supposed that

$$\omega_i = \sum_{k=k_1}^{k_2} a_{i,k}\sin\left(\frac{2\pi i}{N_x}k\right) + b_{i,k}\cos\left(\frac{2\pi i}{N_x}k\right)$$
$$\phi_i = \sum_{k=k_1}^{k_2} a'_{i,k}\sin\left(\frac{2\pi i}{N_x}k\right) + b'_{i,k}\cos\left(\frac{2\pi i}{N_x}k\right)$$

where $a_{n,k}, a'_{n,k}, b_{n,k}, b'_{n,k}$ are taken independently at random with a Gaussian distribution. The frequency range $[k_1, k_2]$ should be chosen depending on the typical vibration frequencies met on satellites. A review of the typical ranges for various satellites is provided in [16]. For a Pléiades satellite, the vibrations typically occur between 70 and 80Hz. As for the the coefficients $a_{i,k}$ and $b_{i,k}$, their amplitudes drive the pixel shifts extents. For Pléiades satellites, micro-vibrations typically induce shifts lower than 0.5 pixel. For the range of frequencies we chose, this means that the random coefficients should not be larger than 4×10^{-6} rad. An example of synthetic vibrated image is given in Fig. 3. A typical image produced by the complete pipeline described above is given in Fig. 3.

B. Proposed method

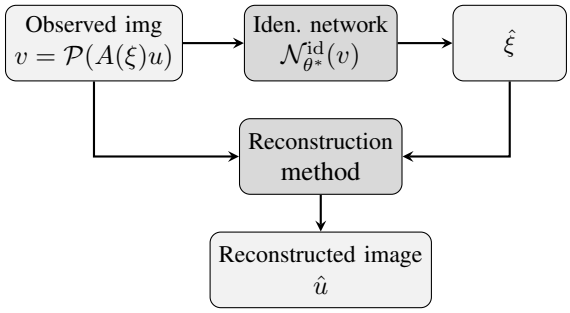


Fig. 5: Proposed method: an identification network is first trained to precisely estimate the attitude $\hat{\xi}$ from the observed image v. This estimated attitude $\hat{\xi}$ and the observed image v are then used to design a reconstruction network to recover the ground image $\hat{u}$.

We propose a two-stage method to correct the vibrations, as illustrated in Fig. 5. A similar principle was already proposed for blind image deblurring tasks in [17], [18].

1) Step 1 – Identifying the micro-vibrations: The micro-vibrations create geometric distortions, which may lead to out-of-distribution measured images. In this paper, we explore the use of convolutional neural networks (CNNs) to identify this distribution shift and retrieve the satellite's attitude from a single input image. This first CNN will be referred to as the *identification network*. Let $\mathcal{N}^{\text{id}}_\theta$ denote our identification network. It depends on weights θ and reads:

$$\mathcal{N}^{\text{id}}_\theta : \mathcal{V} \longrightarrow \mathcal{E}$$
$$v \mapsto \mathcal{N}^{\text{id}}_\theta(v) = \hat{\xi},$$

where $\mathcal{E} = \mathbb{R}^{N_x\times 3}$ describes the space of rotation angles for each sampling time.

The identification network is trained to minimize the loss function:

$$\theta^* = \underset{\theta}{\text{argmin}}\, \mathbb{E}_{v,\xi}\left[\|\mathcal{N}^{\text{id}}_\theta(v) - \gamma\xi\|_1\right]. \qquad (4)$$

In this equation, the vibrated image v and its attitude ξ are seen as a random vectors. The ground truth image u is drawn uniformly at random within a dataset. The attitude ξ is drawn at random following the model described in paragraph II-A6. The vibrated image v is related to u and ξ by the relationship (1) which involves a random perturbation.

The value γ is a normalization constant. It plays an important role for numerical stability and convergence, when training the identification network. In practice, we set it equal to the inverse of an upper-bound on the maximal angle ξ used to generate the vibrated images. In the numerical tests, for the 0.5 pixel shifts, it is chosen as $\gamma = \frac{1}{4\cdot10^{-6}}$.

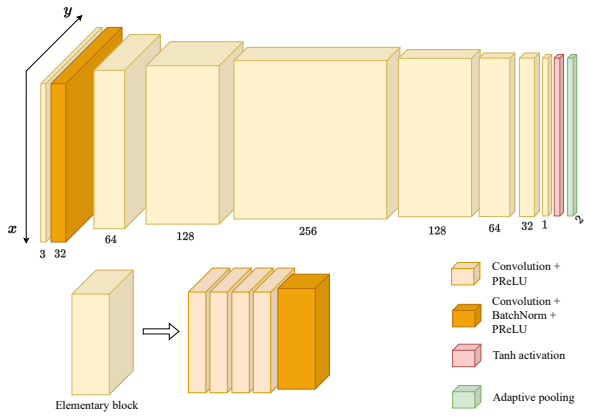


Fig. 6: Idenfication network architecture. The network is composed of 6 elementary blocks with the same structure. Each elementary block is composed of 4 consecutive convolutional layers and PReLU activations that keep the spatial dimension unchanged. It ends by a convolutional layer which reduces the $y-$dimension by a factor 2, a batch normalization and a PReLU. The number of channels from the largest spatial dimension to the smallest spatial dimension are $32, 64, 128, 256, 128, 64, 32, 1$. The total number of parameters of the identification network is about 4.65 million parameters.

a) Identification architecture: We propose an original identification network designed to estimate the rotation angles for each line of an input image. The network architecture is illustrated in Fig. 6. An original aspect of this network lies in its gradual reduction of spatial dimension along the $y-$direction only (CCD line direction) while maintaining the size along the $x-$direction (flight direction) constant. This architecture is tailored to a line-by-line estimation.

Once the identification network has been trained, we can estimate the camera attitude from a single vibrated image v as:

$$\hat{\xi} \overset{\text{def.}}{=} \mathcal{N}^{\text{id}}_{\theta^*}(v) \tag{5}$$

2) Step 2 – Reconstruction: The image reconstruction task is ill-posed and prior information is needed to restrain the solution space. First, the space of ground truth images $\mathcal{U}$ has to be discretized. We use a fine discretization $\mathcal{U}'$ of $\mathcal{U}$, allowing us to simultaneously reduce the vibrations and super-resolve the vibrated images v. In our experiments, $\mathcal{U}'$ is a set of images super-resolved by a factor 4 in each dimension compared to $\mathcal{V}$.

For the reconstruction process, we developed specific *unrolled neural networks* (e.g. [19], [20]). They regularly rank among the top competitors in recent image reconstruction challenges (see e.g. the FastMRI challenge [21] or the Finish inverse problem challenges). To explain their principle, we can draw a parallel with the Maximum A Posteriori estimators (MAP). In our setting, they would typically write:

$$\begin{aligned}\hat{u} &= \underset{u\in\mathcal{U}'}{\operatorname{argmax}}\, p(u|v,\hat{\xi}) \\ &= \underset{u\in\mathcal{U}'}{\operatorname{argmax}}\, \log p(v|u,\hat{\xi}) + \log p(u).\end{aligned}$$

Assuming that $p(u) \propto \exp(-R(u))$ for some function $R : \mathcal{U}' \to \mathbb{R} \cup \{+\infty\}$, and under an additive Gaussian noise assumption, this yields:

$$\hat{u} = \underset{u\in\mathcal{U}'}{\operatorname{argmin}} \underbrace{\frac{1}{2}\|A(\hat{\xi})u - v\|_2^2}_{F(u)} + R(u) \tag{6}$$

Many hand-crafted regularizers R have been developed, including ℓ^2 penalization (Tikhonov) or ℓ^1 penalization (sparsity-based regularizers, total variation). To solve the resulting optimization problem, different iterative optimization algorithms have been developed (6). Many of them rely on proximal operators defined by

$$\operatorname{prox}_R(u_0) \overset{\text{def.}}{=} \underset{u\in\mathcal{U}'}{\operatorname{argmin}}\, R(u) + \frac{1}{2}\|u - u_0\|_{\ell^2}^2. \tag{7}$$

The unrolled networks mimick the iterative algorithms. They consist in replacing the proximal operator prox_R by a learned mapping $\mathcal{D}_\theta$. This mapping can be chosen as a CNN and can be interpreted as a denoiser, specifically tailored to the residual artifacts after inversion of $A(\hat{\xi})$. In this paper, we propose to use Douglas-Rachford splitting-like algorithm (e.g., see [22]). It is detailed in Algorithm 1. The initialization z_0 is obtained by the pseudo-inverse $A(\hat{\xi})^+$ of $A(\hat{\xi})$. It can be computed using a conjugate gradient algorithm. In our experiments, we set $\gamma = 0.5$. To reduce the number of trainable parameters, we adopted a strategy where the same weights θ are shared across all iterations. The proximal operator $\operatorname{prox}_{\gamma F}$ is defined as

$$\operatorname{prox}_{\gamma F}(z) = (\gamma A(\hat{\xi})^* A(\hat{\xi}) + I)^{-1}(\gamma A(\hat{\xi})^* v + z)$$

It can be computed efficiently using conjugate gradient for 10 iterations. The adjoint operator $A(\hat{\xi})^*$ is obtained

Algorithm 1 DR Net for correcting micro-vibrations

Require: vibrated image v, estimated attitude $\hat{\xi}$, number of iterations K, regularization parameters $\gamma, \lambda > 0$

Initialize $z_0 = A(\hat{\xi})^+ v$
for $k \in \{0, 1, ..., K-1\}$ **do**
  $u_{k+1} \leftarrow \mathcal{D}_\theta(z_k)$
  $w_{k+1} \leftarrow \operatorname{prox}_{\gamma F}(2u_{k+1} - z_k) - u_{k+1}$
  $z_{k+1} \leftarrow z_k + w_{k+1} - u_{k+1}$
return u_{K-1}

through automatic-differentiation. The denoiser $\mathcal{D}_\theta$ is trained to minimize the following risk:

$$\min_\theta \mathbb{E}_{v,u,\xi} \left[\|u_{K-1} - u\|_1 \right], \tag{8}$$

where v, u and ξ are seen as random vectors.

The denoising network $\mathcal{D}_\theta$ architecture is a DRUNET [23], [24]. It is among the best competitors in plug-and-play models. We used pre-trained weights for the initialization. The total number of parameters to be estimated is about 34.4 millions. In our numerical experiments, we use a 48GB GPU from NVidia RTX8000 for training. We set $K = 5$ iterations, which appears as a good balance between memory consumption and reconstruction performance in related problems.

Remark II.1. *Notice that we worked under an additive white Gaussian noise assumption to derive Algorithm 1. This may seem paradoxal since we use a Poisson-Gaussian noise approximation when synthetizing the images. A possibility to handle this mismatch would be to use the Anscombe transform [25], which makes it possible to change a Poisson distribution to a Gaussian one. We did not explore this solution in this paper. Indeed, it would be a common misconception to see the unrolled network as a MAP estimator, as was first suggested in [26]. At the training stage,* our objective is not to construct a MAP estimator, but rather the MMSE. *Indeed, we minimize an approximation of the average risk* $\mathbb{E}[\|\hat{u} - u\|_2^2 | v, \hat{\xi}]$*, which is known to be the conditional expectation* $\mathbb{E}[u|v,\hat{\xi}]$ *[27]. Hence, the parallel to a MAP approach is used only to design the network architecture. We train the network* $\mathcal{D}_\theta$ *to learn how to handle noise with a specific distribution, related to the nature of the perturbation* $\mathcal{P}$ *and the inversion of* $A(\xi)$.

III. Numerical experiments and Results

A. Dataset and metrics

We trained the neural networks on the 2017 MS COCO training dataset, using RGB bands only. Images from MS COCO are resized to 1024×1024 pixels and used as the reference images to reconstruct. We focus on the modeling of Pléiades satellites, with micro-vibrations producing errors between 0.25 and 0.5 pixel on drifts of $25 - 75$ lines. At each iteration, we simulate a random micro-vibration for a randomly picked image with a random amplitude (so that our model can handle multiple levels of micro-vibrations). We also synthesize vibrated images of size 256×256 from the high resolution images observed using the model Eq. (1). We used the whole training set of MS COCO for training, i.e. approximately 120000 images.

To quantify the discrepancy between the true signal x – which can be an image or the attitude – and the estimated signal $\hat{x}$, we use the Signal-to-Noise Ratio (SNR). We also use the Structural Similarity Index Measure (SSIM) to evaluate the quality of images. The SNR is defined as

$$\mathrm{SNR}(\hat{x}, x) \overset{\text{def.}}{=} 10 \log_{10} \left(\frac{\|x\|_2^2}{\|\hat{x} - x\|_2^2} \right). \tag{9}$$

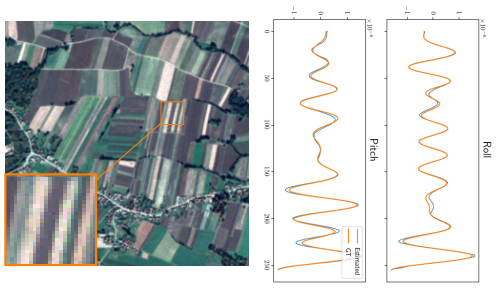


Fig. 7: Example of estimated attitude by identification network. Left: observed image with a maximal pixel shift of 0.5 pixel. Right: estimation of the attitude (orange: ground truth, blue: estimation).

B. Micro-vibration identification

We studied the identifiability of micro-vibrations as a function of the focal plane's geometry. We supposed that the linear CCD arrays of different wavelength are set in parallel (Fig. 4). We then varied the distance d_{ccd} between the CCD lines from 0 to 6 pixels (or equivalently CCD side-length). Grayscale means that we used a single linear CCD array. For each value d_{ccd}, we have trained an identification network $\mathcal{N}^{\mathrm{id}}_{\theta_{d_{\mathrm{ccd}}}}$ with the same training procedure and dataset. We then evaluated its performance on a test set of 1000 images picked at random in the MS COCO evaluation dataset. The results are reported in Table I. A numerical result obtained with acquisition parameters corresponding to a realistic setting is given in Fig. 7.

It reveals noteworthy performance gaps between single-channel (grayscale) and multi-channel (RGB) images. The identification performance increases as a function of the distance d_{ccd} and seems to stagnate for distances larger than 4 pixels. A plausible explanation for this phenomenon is that when the distance increases, the different CCD lines acquire distant regions on the ground. Hence a similar region of space is acquired at different times with a different attitude for the different bands. This yields additional information, as in a stereoscopic system, which can be exploited by the identification network. This principle was used explicitly in [7], while we let the network learn it in our approach.

This outcome provides valuable insights into the optimal construction of the focal plane to make it possible to identify and correct micro-vibrations. It should be carefully calibrated with the specifications of each satellite.

TABLE I: EVALUATION OF THE ATTITUDE IDENTIFICATION PERFORMANCE. THE SNR ARE AN AVERAGE EVALUATED ON 1000 VIBRATED IMAGES. OBSERVE HOW THE PERFORMANCE INCREASES WHEN THE DISTANCE BETWEEN THE DIFFERENT WAVELENGTHS CCDS INCREASES.

d_{ccd}	Grayscale	0	1	2	3	4	5	6
SNR	11.76 ± 2.27	19.36 ± 1.21	22.04 ± 1.48	24.16 ± 1.75	25.34 ± 1.63	26.26 ± 1.71	26.26 ± 1.49	$\mathbf{26.56 \pm 1.45}$

C. An intrinsic identification issue

Vibrations induce distortions in the geometry of the observed images. However, there exists adversarial situations where the geometry of the ground image makes the identification impossible. A simple example is scenes with an homogeneous albedo which can occur in deserts, seas, glaciers. More complex situations can occur. An example is shown in Fig. 8. There, we show that depending on some texture orientation, either the pitch or the roll cannot be identified. However, in this situation, acquired image does not suffer from the pitch (resp. roll) effects, and there is therefore no need for a restoration.

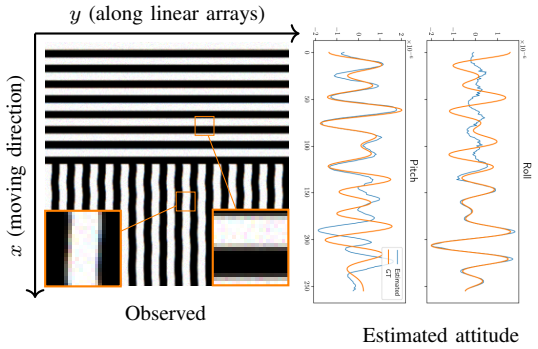


Fig. 8: An example of ambiguity. This illustration showcases that some geometries / textures can make it impossible to identify the micro-vibrations. This can be the case either in $x-$direction (aligned with the flight direction) or $y-$direction (parallel to the CCD arrays). On the right: observe how the roll and pitch are badly estimated depending on the flight position.

D. Reconstruction on synthetic data

We performed an empirical comparison between classical methods: Tikhonov (ℓ_2), total variation minimization (TV), a plug and play method [23] (PnP) and the proposed unrolled neural network (DR Net). The results are presented in Table II together with the associated computational costs.

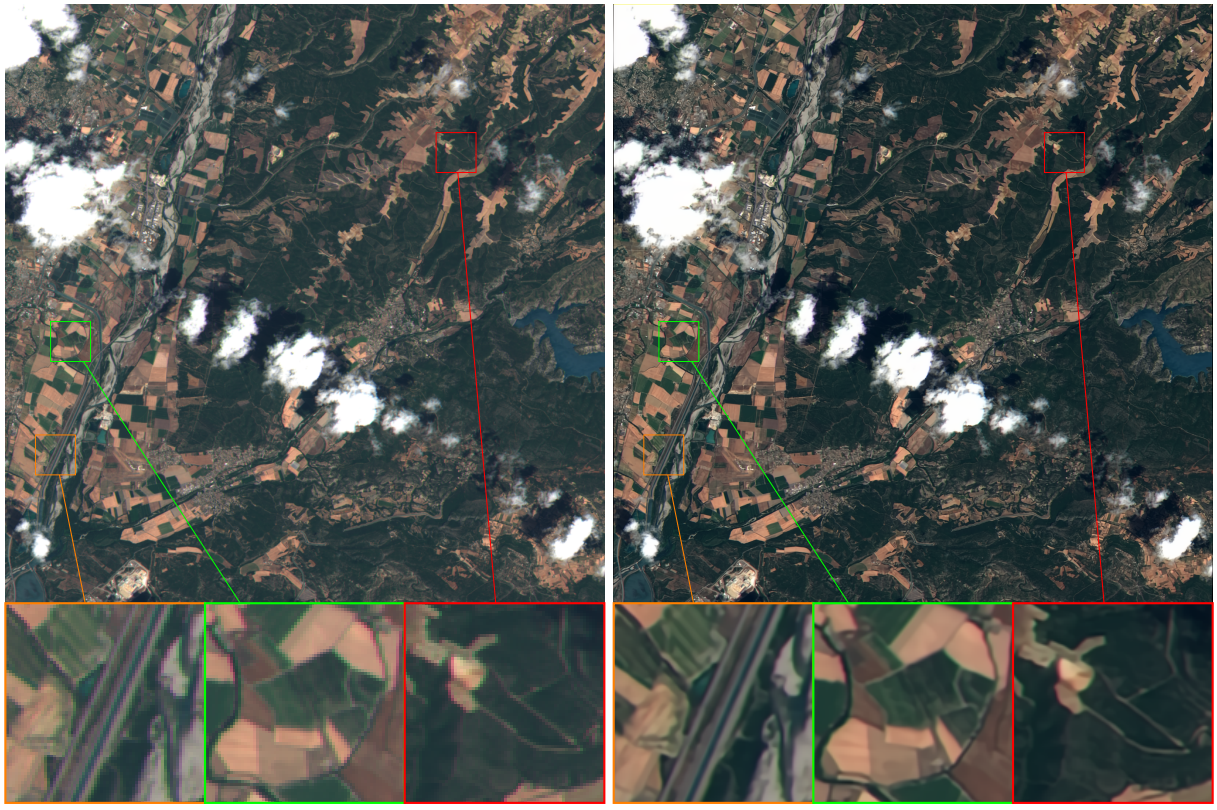
[1]The hyperparameter λ was optimized by grid search. The computational benchmark procedure was executed using PyTorch's internal benchmark tool. It was conducted on a server equipped with an Intel Xeon W-2275 processor (comprising 28 CPU cores with a clock speed of up to 3.3 GHz) and an Nvidia Quadro RTX 5000 16GB GPU. The image size was fixed to 256×256, RGB. The CPU benchmark is conducted using 8 CPU threads. We used the DPIR implementation of DeepInv library

The proposed method demonstrates superior performance both in terms of SNR and SSIM. The closest contender is the state-of-the-art PnP method, but the unrolled network still provides results with a SNR larger than 2dB in average. The method is also advantageous in terms of computational costs, especially compared to methods which require more than a hundred iterations to stabilize.

A notable feature of our model is its capacity to seamlessly achieve a "super-resolution" task. This is done by designing the forward model so that its input lives in a space of higher resolution images. However, notice that the high frequency contents only comes from an implicit prior encoded in the unrolled network and is should not be present in the vibrated images if they are well sampled. In Fig. 10, we show qualitative comparisons of different methods. For handcrafted priors such as Tikhonov and total variation regularization, there are still some residual artifacts. Plug-and-play method with a deep neural network prior improve over handcrafted priors, but still provide results with an accuracy 1dB to 2dB lower than the proposed supervised approach.

E. Reconstruction on real Sentinel-2B data

In this section, we validate the effectiveness of our proposed method on real data. Our primary motivation for this work is Pléiades data, but the raw sensor data needed for our algorithms is not available. For this experiment, we therefore turn to a real Sentinel-2B image, which also works with a linear pushbroom camera. We adapted our method to the Sentinel-2B satellite, by taking the camera parameters from Sentinel Online. As can be seen on the Fig. 9a, false color at object boundaries are visible. This is due to non-uniform shifts between the different channels. Their origin is satellite vibrations. Notice that it is a raw image from Sentinel-2B, without any post-processing steps (Level-0 product). In that example, the displacements are large (up to 5 pixels) and are rather low frequency. They cannot be considered as micro-vibrations. However, the methodology can still be applied. We focus on the RGB bands. In Fig. 9b, we applied our algorithm and observe that the color mismatches are siginificantly reduced and that the reconstructed image is more pleasant visually. Since no ground-truth is available, the improvement can however not be quantified. Yet, this preliminary result is promising for real applications.



(a) Real raw data from Sentinel-2B, cropped to 1024×1024 (b) Processed 2048×2048

Fig. 9: Our proposed method applied on a real raw data from Sentinel-2B.

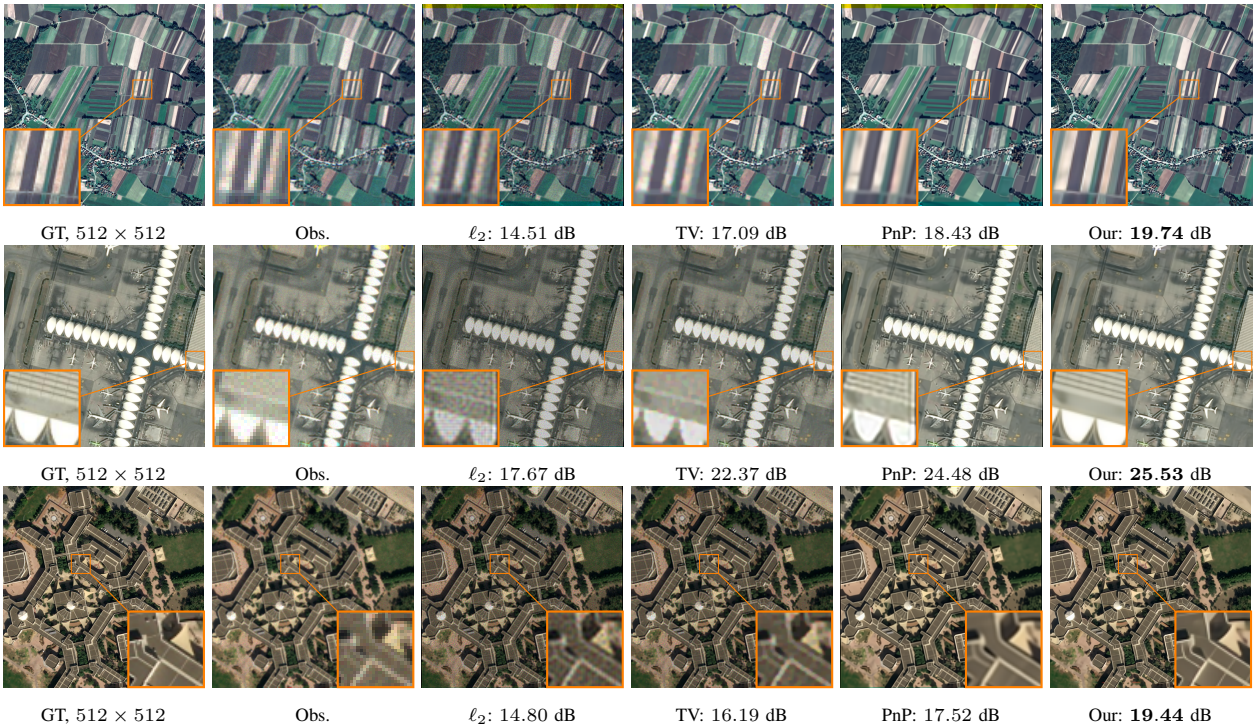


Fig. 10: Reconstruction and 2x super-resolution. From left to right: ground truth images, observation (of size 256×256), 2x super-resolution reconstruction results using Tikhonov, total variation, plug-and-play and our proposed method.

TABLE II: QUANTITATIVE EVALUATION OF THE RECONSTRUCTION PERFORMANCE ON THE TEST SET. [1].

	ℓ_2	TV	PnP [23]	DR Net
Reconstruction and 2x super resolution				
SNR	16.56 ± 0.5	24.48 ± 3.46	25.37 ± 2.62	$\mathbf{27.40 \pm 3.92}$
SSIM	0.78 ± 0.08	0.86 ± 0.06	0.91 ± 0.05	$\mathbf{0.92 \pm 0.04}$
GPU time (s)	2.85 ± 0.12	14.95 ± 0.01	8.39 ± 0.09	$\mathbf{1.80 \pm 0.07}$
CPU time (s)	91.44 ± 0.5	418.40 ± 1.2	220.39 ± 1.3	$\mathbf{46.93 \pm 0.09}$
N. iter	100	100	**8**	**5**

IV. DISCUSSION AND CONCLUSION

We introduced a novel approach to restore images suffering from a specific type of jitter. In our setting, micro-vibrations of a pushbroom camera induce pixel shifts with an amplitude typically smaller than half a pixel. Detecting and correcting them is challenging, but important to improve the geometric quality of the products for various applications such as image interpretation or for generating digital elevation models from stereoscopic pairs. We designed a two-step approach: first a tailored neural network performs a line by line identification of the satellite's attitude. This information is then provided to a second neural network, which is in charge of correcting the pixel shifts and – possibly – creating a higher resolution image.

Surprisingly, the identification network is able to identify the attitude convincingly in this extremely challenging setting. A noteworthy finding of this study is the critical role of the spacing between the different spectral bands on the focal plane. A higher spacing can improve the attitude's estimation performance from about 12dB to more than 26dB. The proposed unrolled network is capable of restoring the images with a quality significantly higher than more standard methods (e.g. Tikhonov or total variation), but also state-of-the art PnP methods, in shorter computing times.

At the current stage, the method is based on the assumption that the scene is planar. This simplifies the forward model, which is a key element of the proposed methodology. The transition from flat scenes to models incorporating elevation is not straightforward. The forward operator A, which is currently dependent on the satellite's attitude alone, would also be contingent on the elevation. In the event that a digital elevation model (DEM) is available, the proposed approach remains valid, albeit with the necessity for modifying the operator's definition. In the absence of a DEM, the correction should depend on the unknown elevation, seen as a latent variable. It therefore looks questionable that this approach can work. Yet, as illustrated on real data, the proposed correction is visually pleasant, sugggesting that the approximation is acceptable, given that the elevation is varying sufficiently smoothly. However, for stereo applications, where the challenge is to reconstruct a DEM from image pairs, this approximation is not acceptable. For this important application type however, the stereo pair contain the elevation implicitly and could resolve the ambiguity. A preliminary study was already proposed in [8]. The present study is an encouraging step towards the feasibility of solving this problem with artificial intelligence tools. The results suggest that they may lead to breakthroughs in the reconstruction performance.

ACKNOWLEDGMENT

The authors would like to thank Fabrice Buffe and Christophe Latry from the CNES for fruitful discussions and advice. P. Weiss acknowledges a support from ANR-3IA Artificial and Natural Intelligence Toulouse Institute ANR-19-PI3A-0004 and from the ANR Micro-Blind ANR-21-CE48-0008. This work was performed using HPC resources from GENCI-IDRIS (Grant 2021-AD011012210R1).

REFERENCES

[1] A. Defence and Space, "Pléaides imagery user guide," 2021.

[2] M. Wang, Y. Zhu, S. Jin, J. Pan, and Q. Zhu, "Correction of zy-3 image distortion caused by satellite jitter via virtual steady reimaging using attitude data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 119, pp. 108–123, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271616301058

[3] J. Takaku and T. Tadono, "High resolution dsm generation from alos prism - processing status and influence of attitude fluctuation -," in *2010 IEEE International Geoscience and Remote Sensing Symposium*, 2010, pp. 4228–4231.

[4] C. de Franchis, E. Meinhardt-Llopis, D. Greslou, and G. Facciolo, "Attitude Refinement for Orbiting Pushbroom Cameras: a Simple Polynomial Fitting Method," *Image Processing On Line*, vol. 5, pp. 328–361, 2015, https://doi.org/10.5201/ipol.2015.146.

[5] J. Straub, M. Swartwout, M. Nunes, and V. Lappas, "Cubesats and small satellites," *International Journal of Aerospace Engineering*, vol. 2019, pp. 1–3, 2019.

[6] V. Amberg, C. Dechoz, L. Bernard, D. Greslou, F. De Lussy, and L. Lebegue, "In-flight attitude perturbances estimation: Application to pleiades-hr satellites," in *Earth Observing Systems XVIII*, vol. 8866. SPIE, 2013, pp. 327–335.

[7] R. Perrier, E. Arnaud, P. Sturm, and M. Ortner, "Estimation of an observation satellite's attitude using multimodal pushbroom cameras," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 5, pp. 987–1000, 2015.

[8] F. de Lussy, D. Greslou, and L. Cross-Colzy, "Process line for geometrical image correction of disruptive microvibrations," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 17, no. 14, 2008.
[9] Z. Zhaoxiang, A. Iwasaki, and G. Xu, "Attitude jitter compensation for remote sensing images using convolutional neural network," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 9, pp. 1358–1362, 2019.
[10] Z. Wang, Z. Zhang, L. Dong, and G. Xu, "Jitter detection and image restoration based on generative adversarial networks in satellite images," *Sensors*, vol. 21, no. 14, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/14/4693
[11] T. Toutin, "Geometric processing of remote sensing images: models, algorithms and methods," *International journal of remote sensing*, vol. 25, no. 10, pp. 1893–1924, 2004.
[12] A. Meygret, G. Blanchet, C. Latry, A. Kelbert, and L. Gross-Colzy, "On-orbit star-based calibration and modulation transfer function measurements for pleiades high-resolution optical sensors," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 8, pp. 5525–5534, 2019.
[13] L. Lebègue, D. Greslou, G. Blanchet, F. de Lussy, S. Fourest, V. Martin, C. Latry, P. Kubik, J.-M. Delvit, C. Dechoz, and V. Amberg, "Pléiades-hr satellites image quality commissioning," *Revue Française de Photogrammétrie et de Télédétection*, no. 209, p. 5–10, janv. 2015. [Online]. Available: https://rfpt.sfpt.fr/index.php/RFPT/article/view/137
[14] C. Latry, S. Fourest, and C. Thiebaut, "Restoration technique for pleiades-hr panchromatic images," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XXXIX-B1, pp. 555–560, 2012. [Online]. Available: https://isprs-archives.copernicus.org/articles/XXXIX-B1/555/2012/
[15] Y. Teshima and A. Iwasaki, "Correction of attitude fluctuation of terra spacecraft using aster/swir imagery with parallax observation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 1, pp. 222–227, 2008.
[16] X. Tang, J. Xie, H. Zhu, and F. Mo, "Overview of earth observation satellite platform microvibration detection methods," *Sensors*, vol. 20, no. 3, p. 736, 2020.
[17] A. Shajkofci and M. Liebling, "Spatially-variant cnn-based point spread function estimation for blind deconvolution and depth estimation in optical microscopy," *IEEE Transactions on Image Processing*, vol. 29, pp. 5848–5861, 2020.
[18] V. Debarnot and P. Weiss, "Deep-blur: Blind identification and deblurring with convolutional neural networks," 2022.
[19] J. Adler and O. Öktem, "Learned primal-dual reconstruction," *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1322–1332, 2018.
[20] V. Monga, Y. Li, and Y. C. Eldar, "Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing," 2020.
[21] M. J. Muckley, B. Riemenschneider, A. Radmanesh, S. Kim, G. Jeong, J. Ko, Y. Jun, H. Shin, D. Hwang, M. Mostapha *et al.*, "Results of the 2020 fastmri challenge for machine learning mr image reconstruction," *IEEE transactions on medical imaging*, vol. 40, no. 9, pp. 2306–2317, 2021.
[22] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," 2010.
[23] K. Zhang, Y. Li, W. Zuo, L. Zhang, L. Van Gool, and R. Timofte, "Plug-and-play image restoration with deep denoiser prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6360–6376, 2021.
[24] S. Hurault, A. Leclaire, and N. Papadakis, "Proximal denoiser for convergent plug-and-play optimization with nonconvex regularization," in *International Conference on Machine Learning*. PMLR, 2022, pp. 9483–9505.
[25] B. Zhang, J. M. Fadili, and J.-L. Starck, "Wavelets, ridgelets, and curvelets for poisson noise removal," *IEEE Transactions on image processing*, vol. 17, no. 7, pp. 1093–1108, 2008.
[26] S. V. Venkatakrishnan, C. A. Bouman, and B. Wohlberg, "Plug-and-play priors for model based reconstruction," in *2013 IEEE global conference on signal and information processing*. IEEE, 2013, pp. 945–948.
[27] A. Gossard and P. Weiss, "Training adaptive reconstruction networks for inverse problems," *SIAM Imaging Science*, 2024.

Minh Hai Nguyen received the M.Sc degree in applied mathematics from INSA Toulouse, France in 2023. He is currently PhD student in applied mathematics and image processing at the Université de Toulouse, France. He is under the supervision of Pierre Weiss and Edouard Pauwels.

François de Vieilleville received the PhD in Computer Science at the Université de Bordeaux, France in 2007. He is currently the CTO of Agenium Space, a French company working on development of egde-AI techniques for satellite imaging and embedding AI to on-board satellites.

Pierre Weiss is a French researcher hired by the CNRS since 2009. He is currently pursuing his research at the Université de Toulouse, France, in the computer science (IRIT), biology (CBI, MCD) and mathematics (IMT) laboratories. He is specialized in the theoretical and applied aspects of imaging with a special emphasis in techniques involving optimization and artificial intelligence. He leads the MAMBO team, aiming at improving the understanding of life, through microscopic imaging and mathematical modelling.