Handheld Burst Super-Resolution Meets Multi-Exposure Satellite Imagery

Jamy Lafenetre Ngoc Long Nguyen Gabriele Facciolo Thomas Eboli

Université Paris-Saclay, CNRS, ENS Paris-Saclay, Centre Borelli, France

https://github.com/Jamy-L/handheld-sr-satellite

Abstract

Image resolution is an important criterion for many applications based on satellite imagery. In this work, we adapt a state-of-the-art kernel regression technique for smartphone camera burst super-resolution to satellites. This technique leverages the local structure of the image to optimally steer the fusion kernels, limiting blur in the final high-resolution prediction, denoising the image, and recovering details up to a zoom factor of 2. We extend this approach to the multi-exposure case to predict from a sequence of multi-exposure low-resolution frames a high-resolution and noise-free one. Experiments on both single and multi-exposure scenarios show the merits of the approach. Since the fusion is learning-free, the proposed method is ensured to not hallucinate *details, which is crucial for many remote sensing applications.*

1. Introduction

Remote sensing is an important research field on which are based practical applications such as natural disaster detection or ecological evaluations. For each application, image resolution and the signal-to-noise ratio (SNR) are two important criteria in practice for visual inspection, with further consequences on downstream tasks, *e.g*. object detection [19]. However, the resolution is limited by the aperture of the telescope, and additional noise further reduces the image quality, resulting in low-resolution (LR) and noisy frames. To circumvent these issues, multi-image super-resolution (SR) algorithms reconstruct the underlying high-frequencies spanned in the aliasing artifacts [21]. Furthermore, combining multiple images leverages both spatial and temporal denoising [6].

Nowadays, the best visual accuracy is achieved by deep learning algorithms [17, 18]. They significantly outperform traditional image fusion strategies based on classical kernel regression [1] or inverse problem solvers [10] in both speed and visual accuracy. Notwithstanding, they are not silver-bullet solutions. First, efficient training of neural networks (NNs) requires very large amounts of carefully collected supervisory data. For image processing tasks, it translates to perfectly-aligned LR burst/HR noise-free target pairs, whose collection is extremely challenging, especially in remote sensing. Self-supervised learning (SSL) [18] have recently circumvented this issue, but NNs are known to underperform when they are deployed on tasks not seen during training. Since synthetic data are limited (hard to model parallax or occlusions for instance), networks that perform well on the evaluation data may fall in practice because of imperceptible details in the images of the real-world scenario. Second, NNs are notorious for *hallucinating* details in the HR prediction. Such details that may sneak from the training data during evaluation are for many applications out-of-question artifacts, which limit the domains where NNs can be safely deployed.

In this work, we follow a different trend by adapting to remote sensing the recent learning-free burst SR approach of [24], proposed for personal photography. It consists first in aligning the raw frames of a burst to a reference one with block-matching and Lucas-Kanade iterations [16], and second in merging the frames into a HR and denoised image using kernel regression. The kernels are steered with respect to a structure tensor [20] that retains the details next to the edges and corners and denoises the flat regions. Structure-adeptness of the structure tensor is particularly suited for remote sensing since many objects such as buildings have regular details such as salient edges that must be restored differently from flat areas such as fields or the sea.

However, since the sequences taken by satellites like Planet's SkySat may have various exposures, with jitter in the exposure coefficients [18], we cannot expect the approach of [24] to be a drop-in replacement of the existing art for remote sensing SR. First, multi-exposure frame registration is an especially challenging problem [15,23], for which the motion model of [24] designed for single-exposure is not adapted. Second, the jitter in the exposure measurements leads to artifacts in classical kernel regression if no correction is applied [18].

We address these issues with two fixes: (i) we plug the

NN of [18] to estimate the optical flow from variously-exposed frames and show it is accurate enough when combined with the robustness weight of [24], and (ii) we follow the base-detail (BD) decomposition strategy of [18] and apply the kernel regression strategy of [24] on the detail layers of the LR frames, handling the jitter to the bases. We show that the fusion technique is flexible enough to incorporate such decomposition. Note that with this approach, the final reconstruction is achieved by a learning-free module, ensuring that possible hallucinations in the predicted optical flow barely have consequences in the HR and noise-free prediction. The proposed method combines the advantages of learned robust alignment for both single and multi-exposure cases, and hallucination-free high-quality reconstruction of an HR image, all with GPU-accelerated implementations. This practical hybrid technique is suitable for a wide range of remote sensing applications.

Our contributions are summarized as follows:

- We present and adapt the handheld burst SR algorithm of [24] to satellite SR imagery for zoom factors between 1 and 2 and possibly important noise levels.
- We include the flow estimators and base-detail strategy of [18], and add a new weight penalizing exposure to adapt the technique to the multi-exposure case.
- We evaluate the method on both synthetic and real data to illustrate its flexibility and merits. In particular, it copes with the NN-based technique from [18] and exceed the performance of the kernel regression of [1].

2. Related work

Most approaches for multi-frame SR focus on the single-exposure case. The idea behind combining multiple frames is to detect and leverage the aliasing caused by the integration on the sensor that contains fragments of the original high-frequency content [21]. This has been historically solved by accumulating frames in a shift-and-add (SA) strategy [10, 13], by solving inverse problems [2, 9], or via kernel regression [1, 5, 14, 20, 24]. These approaches reconstruct a signal with a higher pixel density, and thus containing details beyond the Nyquist rate of the sensor. However, they are blurry due to the blurring inherent to interpolators, and the reconstruction of the lens point-spread function (PSF) [2]. As a result, a subsequent deblurring [1, 9] or sharpening [8] is performed to predict the final image.

Despite being successful in many applied fields, including remote sensing, the state-of-the-art is nowadays dominated by deep learning, *e.g.* [4]. These approaches are notorious for the large amounts of high-quality supervisory data they require for training, yet such high-quality LR/HR image pairs are hardly obtainable in the context of satellites. To overcome this issue, Nguyen *et al.* [17] train a CNN with a self-supervised loss. However, and despite improved visual accuracy over the handcrafted counterparts, these methods may hallucinate details, which is not compatible with many remote sensing applications. In contrast, we propose to adapt the kernel regression technique from [24], which is: efficiently parallelized on a GPU, signal-adaptive, robust to motion and noise, and learning-free, thus ensuring no hallucination while providing high-quality results.

Multi-exposure imagery is another important family of multi-image methods that are highly relevant for remote sensing. In high-dynamic range (HDR) imagery, taking sequences of images at different exposures with limited dynamic range, and fusing them together results in a new HDR one [7]. Neural networks may be trained on bursts of bracketed LR frames to jointly address HDR and SR reconstruction [15]. Nguyen *et al.* [18] propose such an approach for remote sensing, again trained in a self-supervised manner. In particular, they train a CNN to predict the optical flow between two frames differently exposed, a very challenging problem in the HDR literature [23, 25]. In this work, we adapt the kernel regression approach of [24] to satellite bracketed bursts by plugging the optical flow CNN of [18] to align the bracketed frame, and incorporate classic HDR weights [11, 22] to these kernels to reconstruct high-quality HDR and SR satellite images.

3. Approach

3.1. Multi-Exposure Kernel Regression

The data of the problem are the N LR raw frames J_n and corresponding exposures t_n ($n = 1, \dots, N$). Our goal is to predict a single HR frame I aligned with a reference in the LR sequence. In the single-exposure case, any frame is equally valid. However, in the multi-exposure case, each frame may have a different SNR depending on the exposure [12], and there may be saturated areas that are impossible to align. We consider the frames for which the pixel saturation rate is below a hand-fixed threshold (such as 5%), and chose the most exposed among them. If no frame is below the threshold, then the least exposed frame is chosen.

The proposed approach is shown in Figure 1. It illustrates the several stages that we describe in this section. First, we normalize the raw LR images by the exposure coefficients, for all n:

$$H_n = J_n / t_n, \tag{1}$$

and compute the corresponding saturation masks via a threshold arbitrary set to 0.99 as in [15]:

$$M_n = J_n > 0.99. \tag{2}$$

From these normalized images, we first compute the optical flows between the frames $H_2, \dots, H_n$ and H_1 that will be

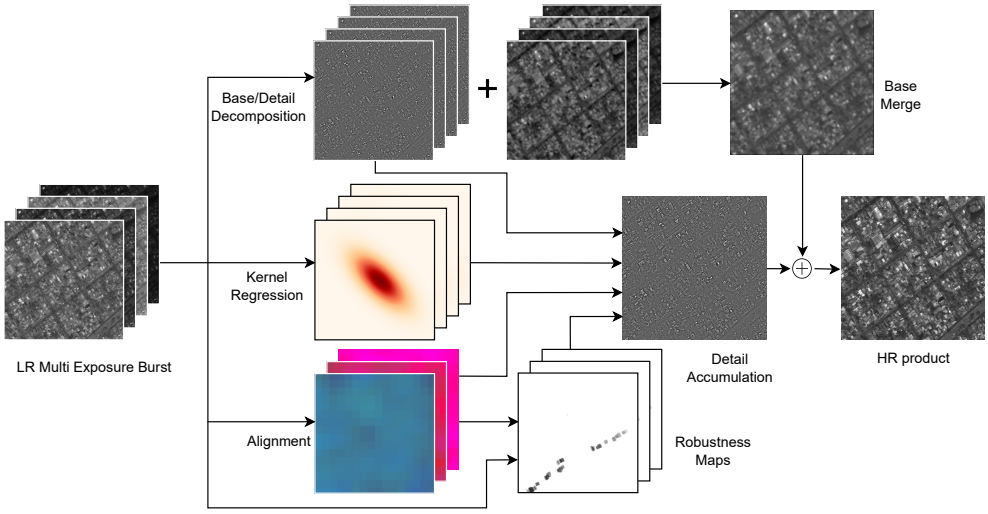


Figure 1. Our pipeline generates one HR image from a burst of LR images. Our method first aligns and identifies outliers using robustness maps, then it estimates local geometric structures using kernel regression. The pipeline further decomposes the burst into base and detail components, where the details are accumulated using Eq. (6). The bases are merged, upscaled to HR space, and combined with the HR detail to produce the HR output.

used during the fusion. In theory any method may work, but in Sec. 4 we show that the FNet model of [18] is the most reliable one for remote sensing, significantly outperforming the patchwise inverse compositional algorithm (ICA) iterations [3] used in [14], which is more suited to the personal photography. Let us name FNet f. The n $(n = 1, \dots, N)$ optical flows obtained as:

$$F_n = f(H_1, H_n), \tag{3}$$

with the convention that F_1 is 0 (no motion). The "Alignment" module in Fig. 1 thus returns N flows $F_1, \dots, F_N$. Following Nguyen *et al.* [18], we proceed by decomposing the n LR frames H_n into base and detail layers to be robust to exposure jitter:

$$B_n = H_n * G_\sigma, \tag{4a}$$
$$D_n = H_n - B_n, \tag{4b}$$

with G_σ a Gaussian filter of variance σ^2 (σ is set to 1 in our experiments). The HR estimated base is simply the accumulation of the LR base images upsampled by bilinear interpolation, and registered to the reference as:

$$B = b\left(\frac{1}{N}\sum_{n=1}^{N} W(H_n, F_n)\right), \tag{5}$$

where W is the warp function that pulls back the frame H_n according to the flow F_n, and b is the bilinear upscaling operation by a factor s. This is valid because the base images only contain the low-frequencies up to the cut-off of the Gaussian filter G_σ. Equation (5) corresponds to the "Base Merge" module in Fig. 1.

More attention is given to recovering the HR details that contain frequencies beyond the Nyquist rate. It is achieved with the kernel regression strategy of [24]:

$$D(x, y) = \frac{\sum_n \sum_{(p,q)\in\mathcal{N}} k_n(p, q) D_n(p, q)}{\sum_n \sum_{(p,q)\in\mathcal{N}} k_n(p, q)}, \tag{6}$$

where $\mathcal{N}$ is the 3×3 neighborhood of pixels in each LR frame that are the closest to the location (x, y) on the HR grid (details in [14]). Equation (6) is accounted for by the module "Detail Accumulation" in Fig. 1. The k_n's are computed as the multiplication of three weights:

$$k_n(p, q) = w_n(p, q) r_n(p, q) h_n(p, q), \tag{7}$$

where w_n is a geometric weight, r_n is a robustness weight that rejects mobile objects and artifacts, and h_n is an HDR weight that gives more importance to frames with better exposure. The first two weights come from [24] and are those corresponding to burst SR. The latter, dubbed h_n, is a contribution of this work to handle the multi-exposed frames. The final image is the summation of the HR predicted base and detail layers:

$$I = B + D. \tag{8}$$

In what follows we explain all the intermediate results to obtain the predicted detail HR layers.

3.2. Description of the weights

An overview of these weights is presented in [24], and implementation details are disclosed in [14].

Geometric weight The geometric weight barely differs from the original paper. It corresponds to the "kernel regression" module in Fig. 1. It consists in steerable kernels adapted to the local geometry, *e.g.* corners or edges. It reads

$$w_n(p,q) = \exp\left(-\frac{1}{2}\begin{bmatrix}x-p\\y-q\end{bmatrix}^\top \Omega^{-1}\begin{bmatrix}x-p\\y-q\end{bmatrix}\right), \quad (9)$$

where Ω is the locally adaptive covariance matrix. Ω is shaped by the hyperparameters k_{detail} and $k_{denoise}$, and $D \in [0,1]$, measuring the amount of details (presence of noise and/or textures). It is estimated using the local structure tensor [20, 24]. We build Ω in the eigen basis P of the local structure tensor as follows:

$$\Omega = P\begin{bmatrix}k_1^2 & 0\\0 & k_2^2\end{bmatrix}P^\top. \quad (10)$$

In the case where both direction have equal energy in the structure tensor like on a corner or a flat region, *i.e.* isotropic behavior, k_1 is equal to k_2, and reads:

$$k_1 = (1-D)k_{detail} + Dk_{denoise}. \quad (11)$$

Conversely, when on an edge, where a direction has much more energy than the other one, k_1 and k_2 are different and stretch the kernel along the edge:

$$k_1 = (1-D)[0.5k_{detail}] + Dk_{denoise}, \quad (12a)$$
$$k_2 = (1-D)[4k_{detail}] + Dk_{denoise}. \quad (12b)$$

In this anisotropic case, k_{detail} is made smaller for the normal direction to the edge to avoid collecting pixels across it, thus reducing blur, and is enlarged along the edge direction to increase denoising without blurring. This is a unique feature of the steerable kernels, not implemented in [1].

In both cases, following the value of D, the kernel w_n is made larger to denoise, or smaller to prevent blurring of the corners and edges. The amount of spatial denoising and detail conservation is controlled by $k_{denoise}$ and k_{detail} that are two hyper-parameters automatically set by the estimated SNR score. More details on k_1 and k_2 can be found in [14, 24]. Overall, this approach is a data-adaptive way to combine images, a merit of CNNs but in a learning-free manner. We show examples of steered kernels for a flat area, an edge and a corner in a real image in Figure 2.

Robustness weight Robustness is tailored to make plausible natural images from everyday life scenes with deformable objects and numerous occlusions. This is far more challenging than the satellite imagery case where the common assumption is to assume static scenes. In this paragraph, all the quantities are pixelwise but we omit this dependency for the sake of conciseness. Robustness is based

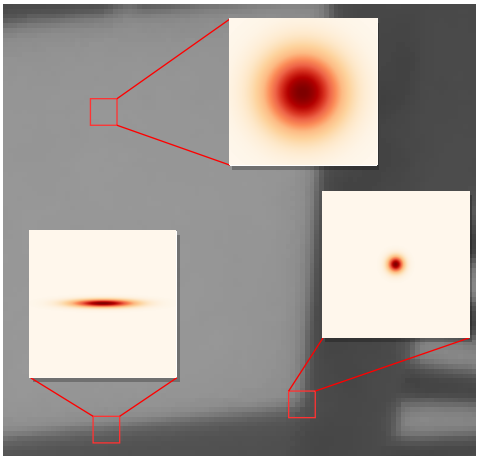


Figure 2. Illustration of the adaptive kernels of our method. A large isotropic kernel is used for areas without details, and a narrow isotropic kernel is used for areas such as corners. A stretched kernel is used for edges.

on the ratio d/σ at a LR pixel location (p,q), where d measures the difference between a pixel and its matching position on the reference frame, and σ represents the local variance [24]. Let p_1 and p_n be two corresponding patches in the images H_1 and $W(H_n, F_n)$. A preliminary step consists of computing the mean color difference $d_n = \|\mu(p_1) - \mu(p_n)\|_2$ with μ the mean function, and the standard deviation of p_1, dubbed σ_1. Since these statistics are computed from few pixels in 3×3 neighborhoods, they might be too noisy to be used directly. Instead, they are corrected with the simulated values d_s and σ_s as:

$$d = \frac{d^2}{d^2 + d_s^2}d \quad \text{and} \quad \sigma = \max(\sigma_s, \sigma_1). \quad (13)$$

Wronski *et al.* [24] simulate the expected values σ_s, d_s by performing Monte-Carlo on the clipped Poisson-Gaussian noise model, for every ISO and binned brightness levels, since p_1 and p_n have the same brightness in their framework. After this correction, higher values of the ratio indicate that an area may be prone to artifacts, whereas low ratios characterize safe to merge pixels: the aliased details necessary for SR, and the noisy flat areas important for efficient temporal denoising. In the single-exposure setting, the robustness coefficient is then obtained using the ratio of the corrected values:

$$r_n(p,q) = \text{clip}\left(s\exp\left(-\frac{d^2}{\sigma^2}\right) - t, 0, 1\right). \quad (14)$$

The scaling factor s is a function of the magnitude of the local optical flow. If the flow is too large, the risk of misalignment artifacts is more important. Therefore the patch is deemed unreliable, and s is set to 2, otherwise to the much larger value of 12. The threshold t is set to 0.12 as in [14].

	SE	ME-0%	ME-5%	ME-20%
ICA-P	52.24	52.46	52.29	50.87
ICA-G	**53.78**	**53.58**	53.34	51.76
FNet	53.57	53.41	**53.40**	**53.34**

Table 1. Comparison of different registration techniques plugged to the steerable kernel regression module. We report the average PSNR for 200 synthetic bursts of 15 256×256 LR frames in the single-exposure (SE) and multi-exposure (ME) settings. For the latter, we follow [18] and inject jitter in $\{0, 5, 20\}\%$ to the exposures to measure robustness of flow estimation to such practical artifacts. For ME, we use BD decomposition for the merge, in order to restrain the effect of jitter to the flow estimation.

In our case, it also prevents to aggregate patches for which FNet may have returned a false prediction, *e.g.* hallucination. Lastly, the robustness weights in (14) are pooled on a 5×5 local neighborhood to share the worst-case confidence. This pessimistic strategy further prevents the accumulation of possible artifact-prone patches.

However, in the multi-exposure case the normalized images H_1 and H_n, and thus the patches p_1 and p_n may have roughly the same brightness but their SNRs remain different. This penalizes the ratio d/σ even if the two patches are visually similar, thus unnecessarily discarding important frames for denoising and SR. Therefore, we would need to simulate d_s and σ_s for every exposure ratio, every ISO and every binned brightness. However, we do not need to adapt to many camera settings as in [24], and can therefore simulate the curves for the single ISO gain of the satellite, thus making this approach tractable. If we had access to the exact noise characteristics of the SkySat satellite, we could also include in the simulation of d_s and σ_s the dependency on the exposure of the noise profile [11, 12]. This would further improve the quality of the robustness. An example of robustness map detecting mobile objects in a real sequence is shown in Fig. 5, in the experiments section.

HDR weight The additional HDR weight compared to [24], gives more importance to the well-exposed frames, and filters out the saturated areas and the darker regions with the lowest SNRs. This weight differs from the robustness weight in many ways. First, it relies only on a single frame, whereas the robustness is defined for image pairs via d. Second, the HDR weight may discard the reference if better-exposed frames are available, especially on saturated areas in the reference. We use the weight of [22], since a reliable estimate of the noise standard deviation σ_s has already been computed for the robustness weight. This weight reads

$$h_n(p,q) = \frac{t_n}{\sigma_s(p,q)} M_n(p,q), \tag{15}$$

where $\sigma_s(p,s)$ is the noise standard deviation estimated for the mean brightness of the pixel located in (p,q). Ideally, it should be obtained using the noise curve specifically estimated for the exposure t_n, but we use the same curve for all frames since from the available data only a single noise profile could be determined. Note that we could have also used the local statistics of the robustness stage to compute HDR weights based on local estimates of the SNR [11, 12], a common practice in the HDR community. We have noted in our experiments that those of (15) were enough to achieve satisfactory results.

4. Experiments

We base our implementation on the official codes from [14] and [18]. Quantitative metrics are computed over a synthetic dataset generated from a set of Skysat L1B satellite images, which has a dynamic range spanning 4096DN.

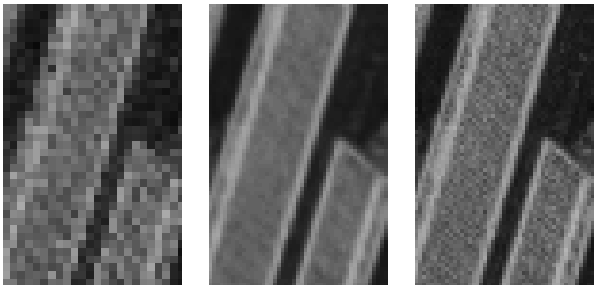
4.1. Alignment

We first evaluate the quality of the output image using three different registration methods. We compare the patchwise ICA algorithm shipped with the code of [14] and adapted for personal photography, the global ICA algorithm used in [1], adapted to satellite imagery since most motions across images are that of the satellite itself, and the CNN dubbed FNet from [18] trained in an end-to-end manner via self-supervised learning on multi-exposure synthetic bursts.

We report in Table 1 the average PSNR estimated on 200 simulated bursts of size $15 \times 256 \times 256$ for both single-exposure (SE) and multi-exposure (ME) with 3 jitter rates as in [18]. The patchwise ICA (dubbed ICA-P) can suffer from instabilities and does not perform as well as ICA global (ICA-G) and FNet. For low jitter rates, global ICA performs better since the synthetic dataset was generated using a global translation model. Yet, FNet falls shortly behind, and performs consistently as the jitter rate rises, contrary to both ICA method for which the PSNR drops. Note a drop of about 2dB for ICA-G compared to FNet for the most severe jitter on the exposure ratios. This suggets that FNet is more robust for general ME scenarios. However, in the case of SE, the three methods are in the same ballpark but we note that FNet and ICA-P better handle mobile objects than ICA-G in practice. Because the kernel regression methods are compatible with most registration techniques, we show that the choice of the alignment is problem-dependent.

4.2. Impact of noise

Raw measurements are degraded by noise coming from both the nature of light and the electronics [12]. A super-resolution method should thus be robust to several signal-to-noise (SNR) ratios to deliver high-quality results. In [24], the parameters k_{detail} and $k_{denoise}$ are automatically set



(a) Reference. (b) Low. (c) High.

Figure 3. Illustration of the difference between narrow and wide kernels for noisy bursts (standard deviation of 50 DN). Wide kernels allow for a better denoising while narrow kernels allow for a better recovery of high frequencies.

Noise std.	8 DN	16 DN	32 DN	50 DN
Low	51.69	51.66	50.57	**48.73**
Medium	53.08	52.82	**50.68**	47.98
High	**55.07**	**53.78**	49.08	45.35

Table 2. Comparison of the mean PSNR estimated for 200 bursts of 15 256×256 LR frames in the single-exposure (SE) setting, for different noise std. Three sets of kernel parameters $(k_{detail}, k_{denoise})$ are considered : Low $(0.33px, 1.65px)$, Medium $(0.24px, 0.96px)$ and High $(0.15px, 0.45px)$. Each parameter set outperforms the 2 others for at least one noise level.

from the measured SNR: the lower the SNR, the larger the kernels, yielding better spatial denoising.

We show in Table 2 the performance of the SR approach for single-exposure for 4 sets of 200 synthetic bursts of size $15 \times 256 \times 256$, each degraded with white Gaussian noise of standard deviation in $\{8, 16, 32, 50\}$DN. For the sake of illustration we test three sets of $(k_{detail}, k_{denoise})$ to monitor their efficiency on satellite images: Low $(0.33px, 1.65px)$ (the default parameters in [14]), Medium $(0.24px, 0.96px)$ and High $(0.15px, 0.45px)$. Each set of parameters is particularly adapted to certain SNRs. Medium is an all-purpose setting to handle both small and important noise instances whereas the default parameter Low is only adapted to the least favorable case, as expected. We illustrate in Figure 3 the impact of choosing between the "Low" and "High" sets of parameters. The one specialized for low SNRs tends to overblur the image, whereas that for higher SNRs may reconstruct correlated noise when achieving SR. Notwithstanding, note that the method may retrieve very fine details such as the stripes on the building with the "High" setting.

We show in Table 3 a comparison with an implementation of shift-and-add (SA), and two state-of-the-art approaches: ACT-spline (dubbed ACTS) [1] and DSP, the CNN from [18]. We set the Gaussian noise level to 50 DN, leading to low SNR. Our approach, which therefore automatically runs with the "Low" setting after leveraging the noise model, better handles such instances of noise. Note that the CNN was not retrained, and may therefore underperform. However, this illustrates the flexibility of our method, which automatically adapts to a broad range of noise levels, contrary to a deepnet that generally excels for one specific noise level. Figure 4 shows a qualitative comparison of the same panel of methods for a sequence with low SNR. ACTS [1] is not designed to jointly address denoising and SR, and thus correlates the noise. DSP [18] also correlates the noise but restores sharp details. Our approach, on the other hand, may efficiently remove important noise while recovering a lot of details.

	PSNR	SSIM
SA	46.81	0.995
ACTS [1]	45.46	0.993
DSP [18]	42.52	0.985
Ours	**48.73**	**0.996**

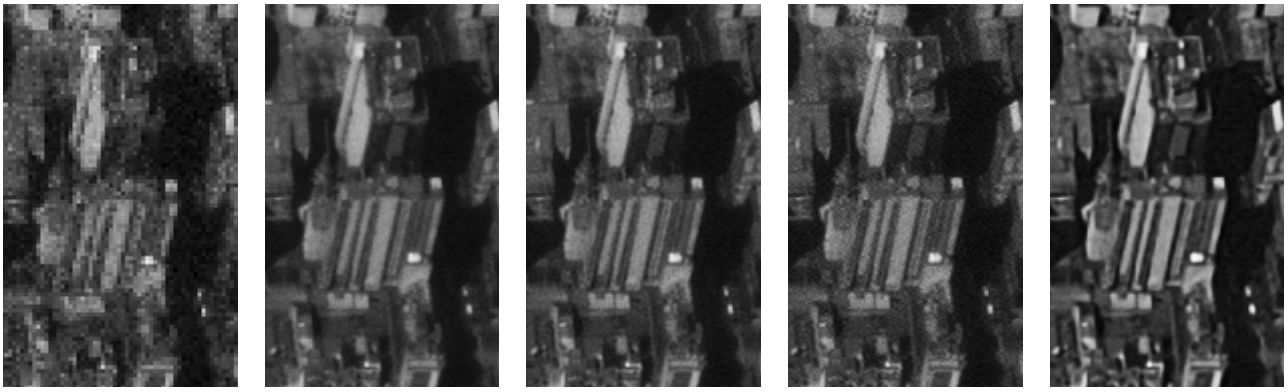
Table 3. Single-exposure SR $\times 2$ with stack size $N = 15$. Average PSNR on 200 bursts with 50 DN Gaussian noise.

4.3. Handling mobile objects

In this section, we illustrate the robustness of our method for mobile objects. Since there are no method to generate synthetic bursts with mobile object, we focus on qualitative results. We show in Figure 5 a crop from a real sequence featuring moving cars on a road. The prediction of DSP [18] splatters the car along the road , whereas our result retains the car. This is crucial for several applications of remote sensing such as surveillance. This difference is explained by the fact that our method includes a robustness mask that attributes a map of confidence to each frame: when every frame but the reference is rejected, our method is a mere upsampling technique. Yet, it retains the moving details in contrast to the CNN. We also show in Figure 5 the accumulation of the confidence score that we call "robustness mask". This image gives hints on the mobile objects and misalignments, and can be used in practice to explain the behavior of our approach, whereas the black-box CNN cannot be diagnosed in case of failure.

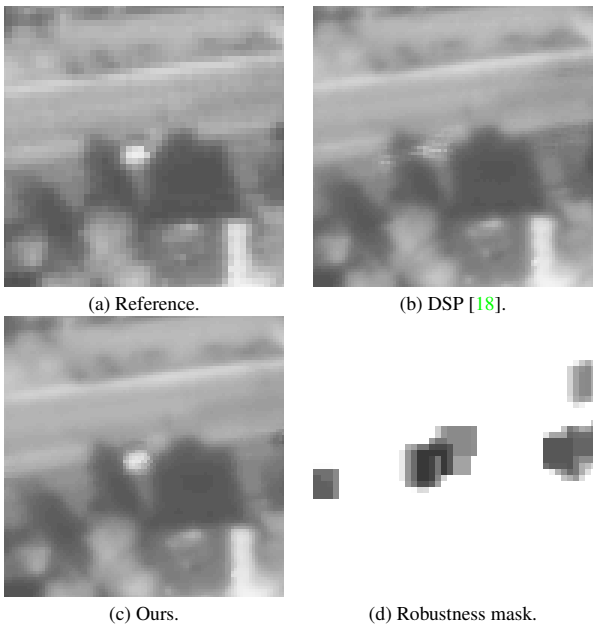
4.4. Single-exposure validation

We quantitatively evaluate our approach with the panel composed of SA, ACTS [1], and DSP [18]. We generate 200 bursts of size $15 \times 256 \times 256$ from HR crops of the Skysat L1B satellite dataset with the protocol detailed in [17]: we blur the image with a Gaussian filter with standard deviation of 0.3, translate the other frames than the reference with a subpixel shift in the Fourier domain, decimate by 2 with nearest neighbor interpolation, and lastly



(a) Reference. (b) SA. (c) ACT [1]. (d) DSP [18]. (e) Ours.

Figure 4. Joint denoising and SR for $N = 15$ simulated frames and 50 DN Gaussian noise. Our approach automatically steers the kernels to produce a HR noise-free image. In contrast, DSP [18] returns a HR image with correlated noise. Better seen on a computer screen.



(a) Reference. (b) DSP [18]. (c) Ours. (d) Robustness mask.

Figure 5. Illustration of the robustness mask of our model on a set of real images. The white dot is a car moving forward on a road, and partially occluded by trees. The dark points on the accumulated robustness mask are areas where frames are rejected due to scene motion, and where the accumulation mostly relies on the reference frame.

add Gaussian noise of standard deviation of 16 DN. Since the noise level of this test set is low, we select the "High" set of values for $(k_{detail}, k_{denoise})$ as previously discussed.

Table 4 shows the average PSNR scores for the panel of methods we consider, and so for three burst size: N=5, 10, and 15. It can be seen that for $N = 5$ images, we fall short by less than 1dB to DSP [18]. We rank second for this burst size, and above the other kernel regression technique

	$N = 5$	$N = 10$	$N = 15$
SA	49.14	51.83	53.11
ACTS [1]	48.88	51.64	52.93
DSP [18]	**51.21**	<u>52.61</u>	<u>53.49</u>
Ours	<u>50.79</u>	**52.74**	**53.78**

Table 4. Single-exposure SR $\times 2$ with varying stack size N. Average PSNR on 200 bursts of size N varying in $\{5, 10, 15\}$, and noise of standard deviation of 16 DN.

	Time (ms/burst)	Peak mem. (GB)
SA	49.5 ± 2.7	3.1
DSP [18]	548.3 ± 22.8	10.8
Ours (ICA)	129.7 ± 14.4	2.0
Ours (FNet)	118.6 ± 10.1	2.4

Table 5. Execution time per burst (s/burst) of size $15 \times 256 \times 256$ pixels on a single NVIDIA RTX 3090 graphic card. We benchmark our method for the patchwise ICA alignment, since an efficient GPU implementation had already been designed for [14], as well as with FNet flows.

ACTS [1] by a margin of 1.6dB. However for larger burst sizes, we rank first with margins of 0.13dB and 0.29dB over DSP. We are thus in the same ballpark as deep learning for these more practical values of N, validating our approach. We also keep important margins of about 1dB over ACTS. Our adaptive kernels better preserve the details such as the edges and corners whereas ACTS is equivalent to kernel regression with isotropic kernels [5], thus blurring details. We have also noted during our experiments that our method is mostly as good as the deepnet to handle instances of parallax next to skyscrapers in real-world images. This shows the merits of our method to urban scenes.

Lastly, we report in Table 5 the average running time for 200 bursts of size $15 \times 256 \times 256$. Our approach, while relying on a non-official reimplementation of the hand-held method [14], is faster than the CNN from [18], and much more memory-efficient, showcasing its practicality. This showcases that the bulk of the computations in [18] are attributed to the fusion stage, that is as accurate with our learning-free, but for a much smaller computational cost. Since Wronski *et al*. [24] claim to process in 100ms on a 2018's smartphones a dozen of 12 Megapixel raw photographs with their own non-released implementation, speed improvements are expected with a better engineered implementation than the non-official one of [14]. Note that the original method merges the frames sequentially in order to fit a low memory device; the runtime could therefore be improved further at the cost of a heavier memory usage by merging frames simultaneously. We do not report the running time for ACTS [1] as only a CPU code was available; It is slower by several order of magnitudes compared to the other methods. We remark that SA has a larger peak memory usage despite being much simpler than the steerable kernel strategy. This is because the code from [18] parallelizes image processing whereas we proceed sequentially, yielding lower memory usage.

4.5. Multi-exposure validation

We compare the performance of the proposed approach for SR of multi-exposed sequences by following the protocol of [18]. The protocol is similar to the single exposure one, but with randomized exposure ratios and heteroscedastic noise. To model real-world imprecision in the ratios, we introduce additional noise jitter ranging from 0% to 20%.

Table 6 shows the performance of different algorithms, evaluated on the synthetic dataset. We evaluated in our panel the kernel regression based techniques, *i.e*. [1] and ours, with and without the BD decomposition proposed in [18]. DSP is run with the BD decomposition. We observe that the BD decomposition is mostly beneficial for high jitter rates, but ensures a consistent PSNR over a wide range of jitter values. Our method ranks second behind DSP, with the advantage of a smaller memory footprint and computational cost. It is consistently better than ACTS, confirming that data-adaptivity leads to better accuracy. Note that this evaluation setting is the most favorable for DSP since the network was trained for this exact noise profile. It was shown in Section 4.2 that our method remains competitive for varying noise levels, and can therefore outperform NNs on images with different SNRs.

4.6. Limitations

We have observed during evaluation on certain real images that some details are not as well restored as ACTS [1] and DSP [18]. When switching to structure tensor-independent narrow isotropic kernels, these pixel-thin details are restored. We posit it comes from the computation of the gradients since [24] average the gradients of several neighboring pixels to be robust to noise. We have observed in practice that for these small details, the dominant eigenvalue of the tensor was not as high as expected, thus favoring too-large kernels where details should be retained instead. This is acceptable in [24] because of the image resolution, whereas the details in our LR frames may be tiny and lead to incorrect gradients, resulting in overblurring of certain details. Better tuning of the hyper-parameters or task-specific image gradients should alleviate this issue.

	Exp. 0%	Exp. 5%	Exp. 20%
ACTS [1]	53.19	51.35	44.78
ACTS (B.D) [18]	52.79	52.71	50.97
SA	53.42	53.00	49.60
DSP [18]	**55.54**	**55.54**	**55.49**
Ours	<u>54.53</u>	52.98	46.07
Ours (B.D)	53.41	<u>53.40</u>	<u>53.34</u>

Table 6. Multi-exposure SR $\times 2$. Average PSNR on 200 bursts of $N = 15$ frames and exposure ratio jitter in $\{0, 5, 20\}\%$.

5. Conclusion

We embedded the fast and efficient steerable kernel regression approach from [24] for single-exposure burst SR into a hybrid scheme for multi-exposure SR in the remote sensing context. We combined this image-fusion module with two modules from [18]: the neural network for optical flow, trained on real images to handle the challenging problem of varying exposures, and the base-detail decomposition strategy to handle jitter in the exposure coefficient. This combination is a fast and interpretable blend of learnable flow and handcrafted image fusion, taking the best of the two worlds for estimating robust complex motions, and merging the frames with the guarantee to never *hallucinate* details. The latter is a key criterion for many practical remote sensing applications, and we thus believe that the proposed approach is perfectly suited for both academia and industry. Experiments for both single-exposure and multi-exposure frames empirically validate our approach.

Acknowledgements This work was partly financed by the DGA Astrid Maturation project SURECAVI ANR-21-ASM3-0002, the Office of Naval research grant N00014-17-1-2552, and the ANR project IMPROVED ANR-22-CE39-0006-04. This work was performed using HPC resources from GENCI–IDRIS (grants 2023-AD011012453R2, 2023-AD011012458R2). Centre Borelli is also with Université Paris Cité, SSA and INSERM.

References

[1] Jérémy Anger, Thibaud Ehret, Carlo de Franchis, and Gabriele Facciolo. Fast and accurate multi-frame super-resolution of satellite images. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, V-1-2020:57–64, 2020. 1, 2, 4, 5, 6, 7, 8

[2] Simon Baker and Takeo Kanade. Limits on super-resolution and how to break them. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1167–1183, 2002. 2

[3] Simon Baker and Iain A. Matthews. Equivalence and efficiency of image alignment algorithms. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 1090–1097, 2001. 3

[4] Goutam Bhat, Martin Danelljan, Fisher Yu, Luc Van Gool, and Radu Timofte. Deep reparametrization of multi-frame super-resolution and denoising. In *Proceedings of the International Conference on Computer Vision*, pages 2440–2450, 2021. 2

[5] Thibaud Briand. Low memory image reconstruction algorithm from RAW images. In *Proceedings of the Image, Video, and Multidimensional Signal Processing Workshop*, pages 1–5, 2018. 2, 7

[6] Kostadin Dabov, Alessandro Foi, and Karen O. Egiazarian. Video denoising by sparse 3d transform-domain collaborative filtering. In *Proceedings of the European Signal Processing Conference*, pages 145–149, 2007. 1

[7] Paul E. Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *Proceedings of SIGGRAPH*, pages 369–378. ACM, 1997. 2

[8] Thomas Eboli, Jean-Michel Morel, and Gabriele Facciolo. Breaking down polyblur: Fast blind correction of small anisotropic blurs. *Image Processing On Line*, 12:435–456, 2022. 2

[9] Sina Farsiu, Michael Elad, and Peyman Milanfar. Multiframe demosaicing and super-resolution of color images. *IEEE Transactions on Image Processing*, 15(1):141–159, 2006. 2

[10] Sina Farsiu, M. Dirk Robinson, Michael Elad, and Peyman Milanfar. Fast and robust multiframe super resolution. *IEEE Transactions on Image Processing*, 13(10):1327–1344, 2004. 1, 2

[11] Miguel Granados, Boris Ajdin, Michael Wand, Christian Theobalt, Hans-Peter Seidel, and Hendrik P. A. Lensch. Optimal HDR reconstruction with linear digital cameras. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 215–222, 2010. 2, 5

[12] Samuel W. Hasinoff, Frédo Durand, and William T. Freeman. Noise-optimal capture for high dynamic range photography. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 553–560, 2010. 2, 5

[13] Danny Keren, Shmuel Peleg, and Rafi Brada. Image sequence enhancement using sub-pixel displacements. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 742–746, 1988. 2

[14] Jamy Lafenetre, Gabriele Facciolo, and Thomas Eboli. Implementing handheld burst super-resolution. *Image Processing On Line*, 2023. 2, 3, 4, 5, 6, 7, 8

[15] Bruno Lecouat, Thomas Eboli, Jean Ponce, and Julien Mairal. High dynamic range and super-resolution from raw image bursts. *ACM Transactions on Graphics*, 41(4):38:1–38:21, 2022. 1, 2

[16] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 674–679, 1981. 1

[17] Ngoc Long Nguyen, Jérémy Anger, Axel Davy, Pablo Arias, and Gabriele Facciolo. Self-supervised multi-image super-resolution for push-frame satellite images. In *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops*, pages 1121–1131, 2021. 1, 2, 6

[18] Ngoc Long Nguyen, Jérémy Anger, Axel Davy, Pablo Arias, and Gabriele Facciolo. Self-supervised super-resolution for multi-exposure push-frame satellites. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 1848–1858, 2022. 1, 2, 3, 5, 6, 7, 8

[19] Jacob Shermeyer and Adam Van Etten. The effects of super-resolution on object detection performance in satellite imagery. In *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1432–1441, 2019. 1

[20] Hiroyuki Takeda, Sina Farsiu, and Peyman Milanfar. Kernel regression for image processing and reconstruction. *IEEE Transactions on Image Processing*, 16(2):349–366, 2007. 1, 2, 4

[21] R. Y. Tsai and Thomas S. Huang. Multiframe image restoration and registration. *Advance Computer Visual and Image Processing*, 1:317–339, 1984. 1, 2

[22] Yanghai Tsin, Visvanathan Ramesh, and Takeo Kanade. Statistical calibration of the CCD imaging process. In *Proceedings of the International Conference on Computer Vision*, pages 480–487, 2001. 2, 5

[23] Greg Ward. Fast, robust image registration for compositing high dynamic range photographs from hand-held exposures. *Journal on Graphics, GPU, & Game Tools*, 8(2):17–30, 2003. 1, 2

[24] Bartlomiej Wronski, Ignacio Garcia-Dorado, Manfred Ernst, Damien Kelly, Michael Krainin, Chia-Kai Liang, Marc Levoy, and Peyman Milanfar. Handheld multi-frame super-resolution. *ACM Transactions on Graphics*, 38(4):28:1–28:18, 2019. 1, 2, 3, 4, 5, 8

[25] Henning Zimmer, Andrés Bruhn, and Joachim Weickert. Freehand HDR imaging of moving scenes with simultaneous resolution enhancement. *Computer Graphics Forum*, 30(2):405–414, 2011. 2