SELF-SUPERVISED PUSH-FRAME SUPER-RESOLUTION WITH DETAIL-PRESERVING CONTROL AND OUTLIER DETECTION

Ngoc Long Nguyen[1] *Jérémy Anger*[1,2] *Axel Davy*[1] *Pablo Arias*[1] *Gabriele Facciolo*[1]

[1] Université Paris-Saclay, CNRS, ENS Paris-Saclay, Centre Borelli, France [2] Kayrros SAS

ABSTRACT

Self-supervised training enables the application of deep-learning based methods for multi-image super-resolution of satellite imagery. In this work we propose two improvements on the self-supervised Deep-Shift-and-Add (DSA) method introduced by Nguyen *et al.* First, we demonstrate how the self-supervised loss of DSA can be extended to provide the image interpreter with a spatially varying parameter to control the trade-off between detail preservation and noise removal at test time. Second, we endow the DSA architecture with a mechanism that enables the network to be robust to outliers produced for example by dead pixels, reflections or registration errors.

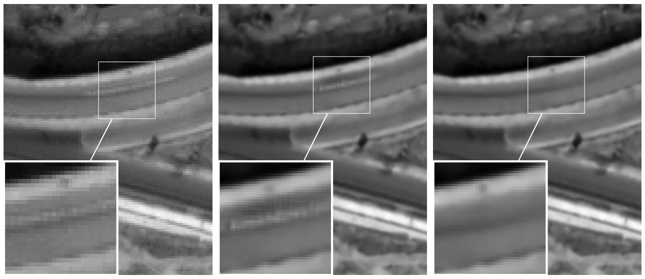
***Index Terms*—** push-frame burst satellite imaging, multi-image super-resolution, deep-learning, self-supervised learning, robust estimators

1. INTRODUCTION

Multi-image super-resolution (MISR) by numerical methods has recently been adopted as way to increase the resolution of push-frame satellites [1, 2]. By leveraging high framerate low-resolution acquisitions, low-cost constellations can be effective competitors to traditional high-cost satellites.

Satellite images are usually noisy. This noise is often considered as an unwanted artifact and is subject to removal. In doing so, it is inevitable that some details will also be lost. In this work, our goal is to perform robust joint super-resolution and denoising from a burst of satellite images, with control over the trade-off between noise removal and detail preservation. We focus on push-frame satellite sensors such as the SkySat constellation from Planet.

Most MISR methods for satellite images are still based on classic model-based techniques [1, 2, 4]. For example, Anger *et al.* [2] propose solving a least-squares problem that fits spline polynomials to the observed samples (we denote it *ACT*, as its early trigonometric polynomial formulation [5]). In [6], Farsiu *et al.* propose a variational method that extends the classic shift-and-add MISR to be robust to outliers, which amounts to applying pixel-wise medians.



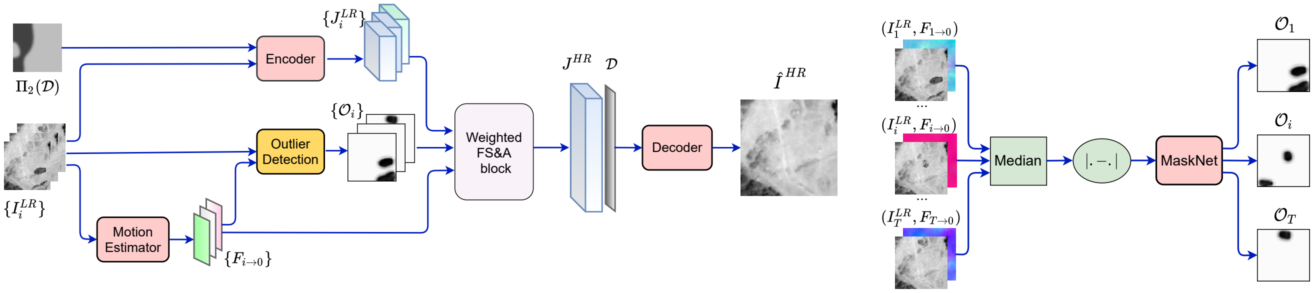
(a) L1B (b) DSA (c) DSA + MaskNet

Fig. 1: Super-resolution from a sequence of 15 real low-resolution SkySat L1A frames. (a) L1B from Planet, (b) DSA [3], (c) Our improvement with an additional CNN to detect the outliers.

On the other hand, training supervised deep-learning based MISR methods is challenging because it is difficult to obtain real training satellite data with ground truth. To overcome this problem, Nguyen *et al.* [3] propose *Deep Shift-and-Add* (DSA), a neural network for MISR from satellite image bursts that can be trained in a *self-supervised* manner, i.e. it does not require ground truth HR data. This work was recently extended in [7] to support multi-exposure bursts. Notwithstanding its advantages, DSA does not handle outliers. For satellite imagery, moving objects misregistration, reflections or dead pixels are examples of such outliers.

Besides, like most deep-learning based image restoration algorithms, DSA tends to smooth out textures or details that have a low contrast relative to the noise level. This is a known problem for restoration methods based on minimizing a distortion measure (such as the MSE or L_1 loss) [8]. As these results have a bad perceptual quality, several works combine distortion losses with adversarial losses that aim at reducing the distance between the distribution of restored images and that of real HR images [8, 9]. However, this requires the net-

This work was supported by a grant from Région Île-de-France. It was also partly financed by IDEX Paris-Saclay IDI 2016, ANR-11-IDEX-0003-02, Office of Naval research grant N00014-17-1-2552, DGA Astrid project « filmer la Terre » n° ANR-17-ASTR-0013-01, MENRT. This work was performed using HPC resources from GENCI–IDRIS (grant 2021-AD011012458) and from the "Mésocentre" computing center of CentraleSupélec and ENS Paris-Saclay supported by CNRS and Région Île-de-France (`http://mesocentre.centralesupelec.fr/`). We thank Planet for providing the L1A SkySat images.



(a) Our architecture

(b) Outlier detection block

Fig. 2: Overview of our method, which builds upon the Deep Shift-and-Add (DSA) architecture [3].

work to "invent" plausible information in the regions where the original content cannot be recovered. This behavior is desired for applications where the goal is to create an aesthetically pleasing natural looking image, but it is unacceptable in cases where critical decisions are made based on the data.

In this paper we propose two contributions to address these limitations of DSA.

1. A data-fitting term that allows controlling the amount of detail in the solution with a spatially varying map that is given to the network as input at test time. In this way a user can control this noise reduction and detail preservation trade-off depending on the application. This is inspired by a common photography trick for recovering low contrast details lost to image denoising, which consists in adding back to the output a fraction of the noisy input image.

2. We also propose a robust version of DSA, thanks to the introduction of outlier masks computed by the network itself (see Fig. 1). Our proposed framework contains DSA, ACT and ACT-robust (a robust variant of the ACT method formulated as an L1 fitting) as particular cases (without the need to solve a computationally complex optimization problem at test time) and yields state-of-the-art results.

2. PROPOSED METHOD

Our network (Fig. 2) is built upon the Deep Shift-and-Add (DSA) architecture [3] with four major modules: Motion Estimator, Encoder, Feature Shift-and-Add block (FS&A) and Decoder. The self-supervised DSA loss [3] drives the network to produce a super-resolved image $\widehat{\mathcal{I}}^{HR}$ such that when subsampled, it coincides with the reference frame I_0^{LR}

$$\text{DSA loss} = \left\| \Pi_2(\widehat{I}^{HR}) - I_0^{LR} \right\|_1, \tag{1}$$

where $\widehat{I}^{HR}$ is the network output and Π_2 is the subsampling operator. Since the reference frame is withheld from fusion during training, the network cannot learn to reproduce the noise in the reference. Thus the training converges to produce a noise-free high-resolution images. This self-supervised loss is based on the minimization of a distortion measure with respect to a target. It is a known fact that these type of losses tend to smooth fine details whose magnitude is comparable with that of the noise [8].

We propose a loss that permits to control the trade-off between noise reduction and detail preservation. For that we incorporate a multi-frame data fitting term controlled by a spatial map $\mathcal{D}$ (Sec. 2.1). In addition we introduce in the DSA network architecture MaskNet, a new trainable module (Fig. 2), whose purpose is to produce outlier masks $\mathcal{O}$ that indicate the presence of outliers (Sec. 2.2).

2.1. Noise reduction – detail preservation trade-off

The self-supervised DSA loss imposes a data-driven prior that favors smooth reconstruction in regions with lower contrast. The cost of producing a noiseless image is that some details might be lost. To counteract this, we add the following loss, which corresponds to ACT [2, 5] when $p = 2$:

$$\text{LL}_p \text{ loss} = \frac{1}{T} \sum_{i=1}^{T} \left\| \Pi_2 \left(\text{Warp}(\widehat{I}^{HR}, 2F_{i\to 0}) \right) - I_i^{LR} \right\|_p^p, \tag{2}$$

with $1 \leq p \leq 2$ (in this work we only consider $p = 1$ and $p = 2$) and where Warp is an operator that warps its input according to the estimated motion field $F_{i\to 0}$ (see [3] for details). This loss corresponds to the likelihood of the data under a generalized Gaussian noise model [6].

To understand the rationale behind this loss, consider a case in which the images can be aligned with an integer shift (so that no interpolation is needed). Then the minimizer of the LL_p loss is obtained by aligning and aggregating the LR images on the HR grid. For $p = 2$ the aggregation is the average, and for $p = 1$ is the median. For Gaussian noise, these solutions are unbiased estimators of the high-resolution image. Thus no details are lost and the noise is reduced via the temporal aggregation. Of course some noise will remain, with variance depending on the number of images in the sequence. Note that these solutions do not require any data-driven learning: i.e. they do not depend on any priors learned from the data; they only depend on the set of LR images. This is because the target frames I_i^{LR} are all part of the input.

Since the least-squares (LS) solution ($p = 2$) is sensitive to outliers [6], we propose using the $p = 1$, which we call least absolute value (LAV) loss.

Complete training loss. Most of the time, our priority is to produce a noise-free HR image. Nevertheless keeping details might be preferred when we have few images or when we want to detect very high-frequency objects such as crosswalks, solar panels, etc. To control the trade-off between noise removal and detail conservation, we introduce a noise-detail map (denoted $\mathcal{D}$) as a parameter to balance the losses (1) and (2). This map is spatially varying and takes values between 0 and 1. Values closer to 1 indicate that we want to keep details (and noise), whereas small values imply that the corresponding region should be denoised. The training loss is defined as the balance between the DSA and the LAV loss per pixel

$$\text{loss} = \left\| \left(\Pi_2(\widehat{I}^{HR}) - I_0^{LR} \right) \cdot (1 - \Pi_2(\mathcal{D})) \right\|_1 + \frac{1}{T} \sum_{i=1}^{T} \left\| \left(\Pi_2(\text{Warp}(\widehat{I}^{HR}, 2F_{i\to 0})) - I_i^{LR} \right) \cdot \Pi_2(\mathcal{D}) \right\|_1, \quad (3)$$

where $\widehat{I}^{HR} = \textbf{Net}(I_{i=1,\ldots,T}^{LR}, \mathcal{D})$ is the network output and "$\cdot$" denotes the element-wise multiplication. To simplify, we assume that $\mathcal{D}$ is smooth and that the images are coarsely prealigned so that the $\mathcal{D}$ does not have to be warped in the loss.

2.2. Outlier handling

In DSA [3], the features computed by the encoder are averaged by the Feature Shift-and-Add module. Because of this averaging, outliers have a strong impact that the decoder cannot entirely mitigate. For this reason, we propose removing them from the averaging by incorporating a submodule MaskNet to the DSA architecture to predict outlier masks. We take inspiration from a video denoising application [10] where a similar mask predicting network is used for removing misaligned areas in a recursive frame fusion method. We define outliers as regions that are inconsistent with the majority of frames in the sequence, and masks allow to exclude them from fusion. To estimate such masks, we first approximate a low-resolution outlier-free image using a temporal median of the LR frames aligned to the reference, which we denote by M^{LR}. Then the absolute difference [10] between the warped median image and each image is used as input for MaskNet

$$\mathcal{O}_i = \text{MaskNet}\left(|\, \text{Warp}(M^{LR}, F_{i\to 0}) - I_i^{LR} | \right). \quad (4)$$

We also impose the smoothness of the produced masks by adding a TV regularization term in the loss. The outlier masks are then used as weights in the weighted FS&A block

$$J^{HR} = \frac{\sum_{i=1}^{T} \text{SPMC}(J_i^{LR} \cdot \mathcal{O}_i, F_{i\to 0})}{\sum_{i=1}^{T} \text{Max}(\text{SPMC}(\mathcal{O}_i, F_{i\to 0}), \epsilon)}, \quad (5)$$

where ϵ is a threshold to avoid division by 0, $\{J_i^{LR}\}$ are the features computed by the Encoder, $\{F_{i\to 0}\}$ are the optical flows estimated by the Motion Estimator, and the SPMC module [3,11] maps the LR features onto a common HR grid. The outliers will be assigned negligible weights in the outlier masks so that they do not contribute in the fusion.

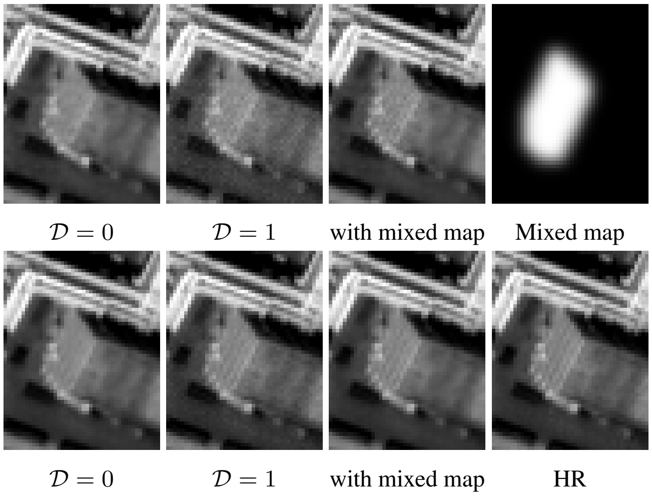


Fig. 3: Super-resolution from a stack of 4 (first row) and 14 (second row) noisy synthetic images. From left to right: Reconstruction with $\mathcal{D} = 0$ (without detail preservation), with $\mathcal{D} = 1$, and with a mixed $\mathcal{D}$ map.

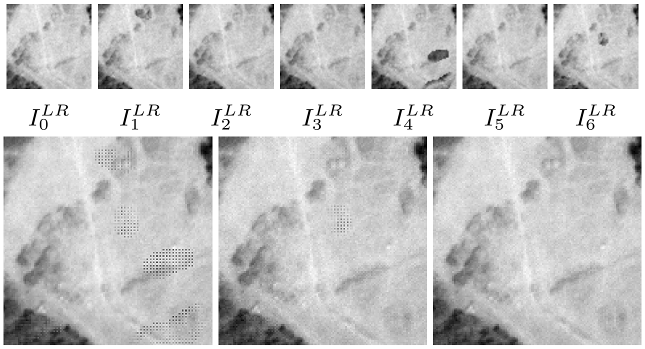
3. EXPERIMENTS

In our experiments, we first demonstrate the trade-off between denoising and detail restoration using the noise-detail map. Then we justify our choice of architecture and losses in order to handle outliers.

3.1. Examining detail preservation map $\mathcal{D}$

In order to train the network, we prepare sets of training input data $\{I_{i=0,\ldots,T}^{LR}; \mathcal{D}\}$. The random spatially-varying noise-detail maps $\mathcal{D}$ are generated first by thresholding a filtered Gaussian noise image ($\sigma = 40$) (the filter itself is a Gaussian filter with $\sigma = 28$), then the resulting binary image is then smoothed with a small Gaussian filter ($\sigma = 3$).

Fig. 3 illustrates the trade-off between noise reduction and detail preservation when we change $\mathcal{D}$ for the cases of 4 and 14 input frames. As expected, with $\mathcal{D} = 0$ the network removes noise while smoothing out the textures as it cannot distinguish high frequencies from noise. On the other hand, with $\mathcal{D} = 1$ the output is noisier but it better preserves the high frequency details. The difference between the two behaviors is particularly noticeable when using few input frames. Moreover, we can use a spatially varying map to reduce noise in uniform regions and preserve details in textured regions as



LS loss, with Mask LAV loss, w/o Mask LAV loss, with Mask

Fig. 4: Effect of the architecture and the loss in robustness to outliers. First line: 7 LR with outliers. Second line: Reconstruction with the LS or LAV loss, with and without MaskNet.

Table 1: Average PSNR on the synthetic test set with outliers.

	$\mathcal{D} = 1$	$\mathcal{D} = 0$	$(\mathcal{D} = 1)$ + Mask	$(\mathcal{D} = 0)$ + Mask
$T = 4$	31.10	36.87	32.65	37.16
$T = 14$	36.52	40.45	37.25	40.77

shown in Fig. 3. Since the map $\mathcal{D}$ is a network input provided at test time, it lends itself to applications where a user interactively edits the map.

3.2. Robustness to outliers

To train MaskNet we add synthetic outliers in the LR images during training. To this aim, we first generate random blobs in an image and then substitute the pixels of these regions with data from a different stack.

As the LL$_p$ loss optimization is not data-driven, it is strongly affected by outliers. Here, we justify two key features in our framework that enable its robustness to outliers: The MaskNet and the L1 norm in the LAV loss.

The authors of [6] show that the optimal solution of the LS (resp. LAV) problem is the pixelwise average (resp. median) of the LR images. Consequently, the LS problem is not robust to outliers. We experimentally observed (Fig. 4) that using the L2 norm, MaskNet learns to produce only constant maps and the network produces artifacts in the SR result. Conversely, with the L1 norm, the MaskNet is able to detect outliers in the LR images and impose a negligible weight to these regions.

SR on synthetic data with outliers. Table 1 highlights the usefulness of the MaskNet when the image stacks contain outliers. We can notice that with few or many frames, both for $\mathcal{D} = 0$ or $\mathcal{D} = 1$, MaskNet helps to increase significantly the PSNR by 0.3 - 1.5dB.

SR on real satellite data with moving objects. Moving objects that are not correctly aligned can be considered as outliers. Fig. 1 illustrates how our architecture with and without MaskNet handles moving objects. As expected, the motion estimator of L1B [1] and DSA predicts smooth optical flows and ignores small moving objects. Consequently, without MaskNet we observe a blur trait on the highway. On the other hand, when we use MaskNet, the network is able to filter out the motion of the car, leading to a better reconstruction.

4. CONCLUSION

We extended the self-supervised DSA method [3] by providing a spatially varying parameter to control the trade-off between detail preservation and noise removal at test time. In addition we endow the DSA architecture with a mechanism that enables the network to be robust to outliers produced for example by dead pixels, reflections or registration errors. All within a self-supervised framework. These improvements lead to state-of-the-art results.

In this work we assume that outliers are everything not consistently visible in most of the frames. However, if we want to preserve the content of the reference image some modifications might be necessary. This will be the subject of future work.

5. REFERENCES

[1] K. Murthy, M. Shearn, B. D. Smiley, A. H. Chau, J. Levine, and M. D. Robinson, “Skysat-1: very high-resolution imagery from a small satellite,” in *Sens. Syst. Next-Gener. Satell.*, 2014.

[2] J. Anger, T. Ehret, C. de Franchis, and G. Facciolo, “Fast and accurate multi-frame super-resolution of satellite images,” *ISPRS*, 2020.

[3] N. L. Nguyen, J. Anger, A. Davy, P. Arias, and G. Facciolo, “Self-supervised multi-image super-resolution for push-frame satellite images,” in *CVPRW*, 2021.

[4] J. Anger, T. Ehret, and G. Facciolo, “Parallax estimation for push-frame satellite imagery: application to super-resolution and 3d surface modeling from skysat products,” *IGARSS*, 2021.

[5] H. G. Feichtinger, K. Gr, T. Strohmer, et al., “Efficient numerical methods in non-uniform sampling theory,” *Numerische Mathematik*, vol. 69, no. 4, pp. 423–440, 1995.

[6] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar, “Robust shift and add approach to superresolution,” in *Appl. digit. image process. XXVI*, 2003.

[7] N. L. Nguyen, J. Anger, A. Davy, P. Arias, and G. Facciolo, “Self-supervised super-resolution for multi-exposure push-frame satellites,” in *CVPR*, 2022.

[8] Y. Blau and T. Michaeli, “The perception-distortion tradeoff,” in *CVPR*, 2018.

[9] C. Ledig et al., “Photo-realistic single image super-resolution using a generative adversarial network,” in *CVPR*, 2017.

[10] M. Maggioni et al., “Efficient Multi-Stage Video Denoising with Recurrent Spatio-Temporal Fusion,” in *CVPR*, 2021.

[11] X. Tao, H. Gao, R. Liao, J. Wang, and J. Jia, “Detail-revealing deep video super-resolution,” in *ICCV*, 2017.