Self-supervised multi-image super-resolution for push-frame satellite images Supplementary material

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
https://cmla.github.io/DSA-Self/

1 DSA architecture

Our DSA architecture (see Figure 1) has 4 major modules: Motion estimator, Encoder, Feature Shift-and-add, and Decoder. The Feature Shift-and-add block does not have any trainable parameters. Our motion estimator follows the work of [5]. Our encoder and our decoder are inspired from the SRResNet architecture [4], and built from the residual blocks (see Table 1). Convolutions of the encoder and decoder are performed using reflection padding. In total, our networks have 2853411 trainable parameters (Table 2).

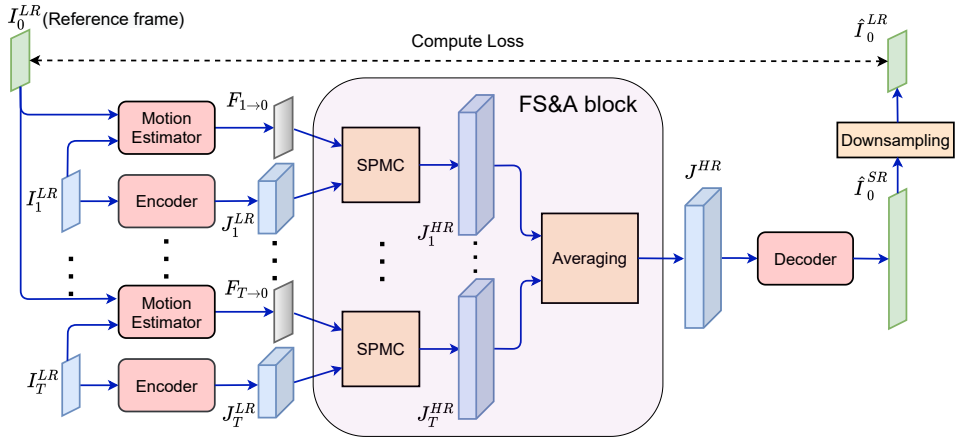


Figure 1: Overview of the DSA framework at training time.

2 The encoder matters

In this experiment, we study the role of the feature extraction in our proposed DSA architecture. By removing the encoder from the DSA architecture, we obtain a different method that performs pixel fusion instead of feature fusion. We observe that the network which performs feature fusion is better at removing the residual noise in the SR image. This aspect is especially noticeable when we have few LR input images. With 5 input images per sequence, the network with encoder performs 0.22dB better than the one without encoder (see Table 3).

Table 1: `ResBlock(N)`

Input	Tensor N channels
Layer 1	`Conv2d(in=N, out=N, k=3, s=1, p=1)`
	`ReLU`
Layer 2	`Conv2d(in=N, out=N, k=3, s=1, p=1)`
Output	Layer 2 + Input

Table 2: DSA Architecture

Modules	**Layers**	**Number of parameters**
Motion Estimator	`FNet` [5]	1744354

Encoder	`Conv2d(in=1, out=64, k=3, s=1, p=1)`
	`ResBlock(64)` ×4
	`ReLU`
	`Conv2d(in=64, out=64, k=3, s=1, p=1)`
		332992

FS&A		0

Decoder	`Conv2d(in=64, out=64, k=3, s=1, p=1)`
	`ResBlock(64)` ×10
	`ReLU`
	`Conv2d(in=64, out=1, k=3, s=1, p=1)`
		776065

		Total: 2853411

Table 3: Fusion in feature space: Average PSNR (dB) over the main validation dataset for our supervised DSA networks with and without Encoder.

DSA	Number of images		
	5	16	30
With encoder	45.82	49.33	50.50
Without encoder	45.60	49.33	50.51

3 Comparison with SISR methods

In this section, we compare our MISR method with a SISR method. For that, we retrained the SRGAN [4] method on our synthetic dataset. According to the Nyquist–Shannon sampling theorem, high-frequency details cannot be fully recovered from a single aliased LR image. Therefore, SISR techniques often introduce "hallucinations" in the reconstruction, which are inappropriate for most remote sensing applications. In contrast, MISR methods aim at increasing the true optical resolution in the SR images. The advantage of MISR over SISR is demonstrated in the Table 4 and in the Figures 2, 3.

4 Image sharpening

In this section, we examine in details the two formulas for the self-supervision loss

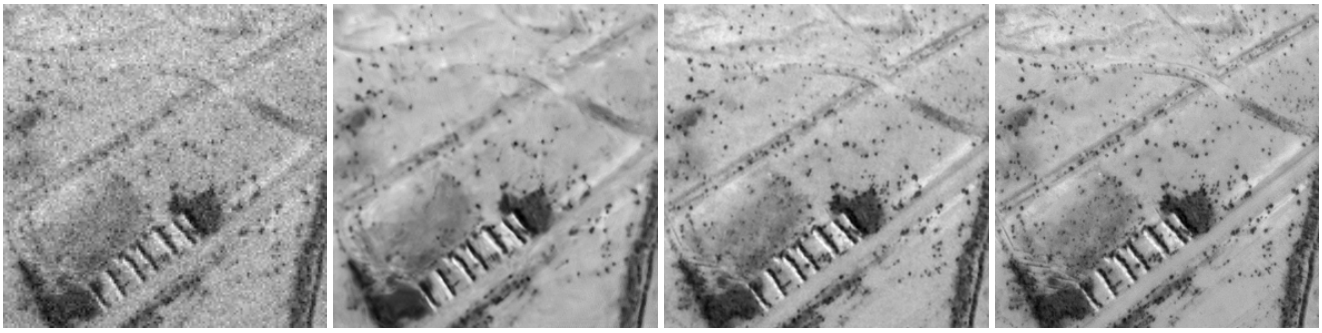
$$\ell_{self}(\hat{I}_0^{SR}, I_0^{LR}) = \|D_2(\hat{I}_0^{SR}) - I_0^{LR}\|_1, \tag{1}$$

$$\ell_{self}(\hat{I}_0^{SR}, I_0^{LR}) = \|D_2(\hat{I}_0^{SR} * k) - I_0^{LR}\|_1. \tag{2}$$

In Equation (1), the network is trained to produce an SR image such that after subsampling, it coincides with the LR reference. According to our image formation model, the output of this network is a filtered high-resolution image (blurry).

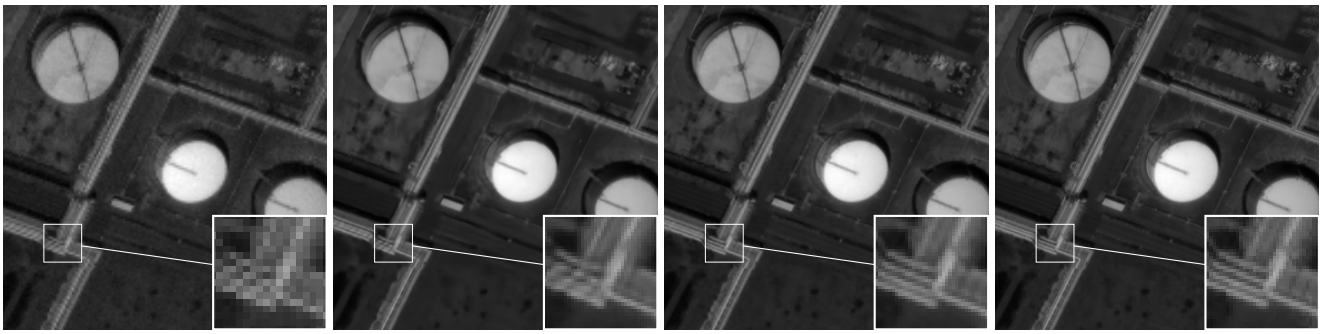
Table 4: Quantitative comparison with the SISR method SRGAN. T is the number of imput images.

Methods	SRGAN ($T = 1$)	DSA ($T = 5$)	DSA ($T = 16$)	DSA ($T = 30$)
PSNR	43.92	45.82	49.33	50.50



(a) LR (b) SRGAN (43.97dB) (c) DSA-Self (**49.86dB**) (d) Ground truth

Figure 2: Visual comparison with the SISR method SRGAN. In this example, SRGAN confuses the black spots on the field with noise, and thus cannot recover correctly these details.



(a) LR (b) SRGAN (39.25dB) (c) DSA-Self (**44.50dB**) (d) Ground truth

Figure 3: Visual comparison with the SISR method SRGAN. In this example, SRGAN fails to remove the alias present in the LR image.

After restoring a high-resolution – but blurry – image with a network trained using the loss (1), we can then restore a sharp image I from $\hat{I}_0^{SR}$ by solving a non-blind deconvolution problem

$$\arg\min_I \|I * k - \hat{I}_0^{SR}\|_2^2 + \lambda \|\nabla I\|_1, \tag{3}$$

where the blur kernel k is defined in the Fourier domain as $\hat{k}(\boldsymbol{\omega}) = (5|\boldsymbol{\omega}| + 1)^{-1}$ [2], and the the regularization weight λ can be set to a very low value thanks to the low noise level of the SR results. This inverse problem can be solved efficiently using a half-quadratic splitting method as in [3, 1].

As an alternative to using this variational method, we can also integrate the deconvolution (with the same blur kernel k) into the self-supervision loss as in the Equation (2). In this way, the network is trained to produce directly a sharp SR image (one that once blurred matches the observed blurry samples). The Figures 4 and 5 show visual comparisons between our loss-based method and the variational method. Figures 4a and 5a are the outputs of the network trained with the self-supervision loss in the Equation (1). Figures 4b and 5b show the deconvolved results obtained from these blurry images by the variational method. Figures 4c and 5c are the outputs of the network trained with the self-supervision loss (2). As we can see, the loss-based result is as sharp as the variational result while avoiding unwanted high-frequency artifacts and with less noise. Moreover, our loss-based method is simple, efficient and does not require any regularization.

References

[1] Jérémy Anger, Mauricio Delbracio, and Gabriele Facciolo. Efficient blind deblurring under high noise levels. In *IEEE ISPA*, pages 123–128, 2019. 3

[2] Jérémy Anger, Thibaud Ehret, Carlo de Franchis, and Gabriele Facciolo. Fast and accurate multi-frame super-resolution of satellite images. *ISPRS*, 2020. 3

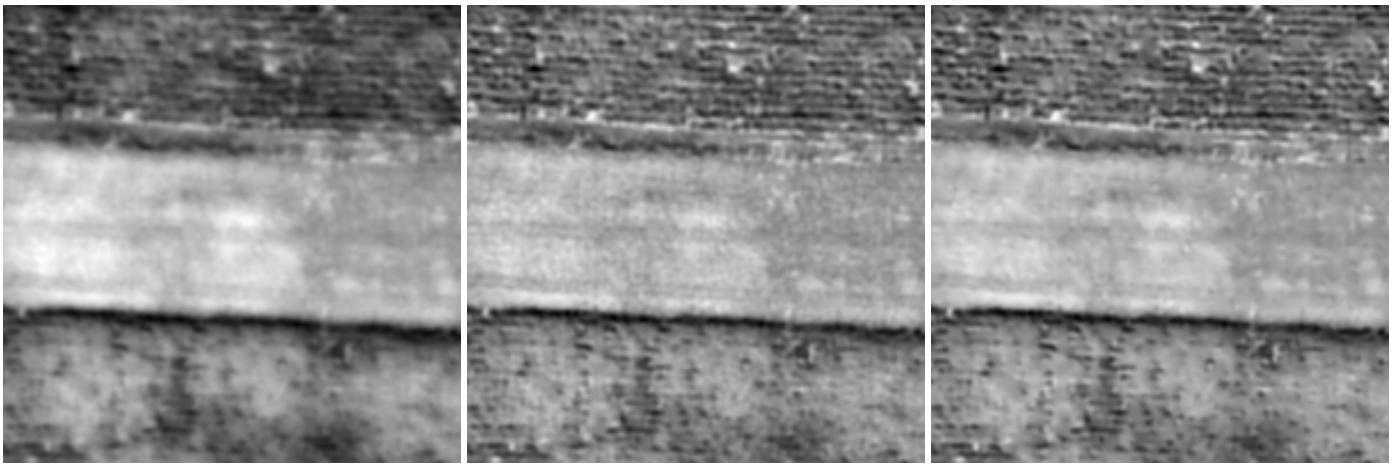
[3] Dilip Krishnan and Rob Fergus. Fast image deconvolution using hyper-laplacian priors. In *NIPS*, pages 1033–1041, 2009. 3

[4] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 1, 2

[5] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6626–6634, 2018. 1, 2

(a) Blurry SR result (b) Variational deconvolution (c) Loss-based deconvolution

Figure 4: Loss-based result contains less noise and no ringing artifacts (on the top of the building).



(a) Blurry SR result (b) Variational deconvolution (c) Loss-based deconvolution

Figure 5: Even with a regularization term, variational method has unwanted high-frequency artifacts. Our loss-based method produces a clean, sharp image without the need of any explicit regularization.