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

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/

Abstract

Recent constellations of optical satellites are adopting multi-image super-resolution (MISR) from bursts of push-frame images as a way to increase the resolution and reduce the noise of their products while maintaining a lower cost of operation. Most MISR techniques are currently based on the aggregation of samples from registered low resolution images. A promising research trend aimed at incorporating natural image priors in MISR consists in using data-driven neural networks. However, due to the unavailability of ground truth high resolution data, these networks cannot be trained on real satellite images. In this paper, we present a framework for training MISR algorithms from bursts of satellite images without requiring high resolution ground truth. This is achieved by adapting the recently proposed frame-to-frame framework to process bursts of satellite images. In addition we propose an architecture based on feature aggregation that allows to fuse a variable number of frames and is capable of handling degenerate samplings while also reducing noise. On synthetic datasets, the proposed self-supervision strategy attains results on par with those obtained with a supervised training. We applied our framework to real SkySat satellite image bursts leading to results that are more resolved and less noisy than the L1B product from Planet.

1. Introduction

High resolution satellite imagery is key for applications such as monitoring human activity or disaster relief. In recent years, computational super-resolution is being adopted as a cost-effective solution to increase the spatial resolution of satellite images [44, 5].

Super-resolution approaches can be broadly classified into single-image (SISR) and multi-image (MISR). SISR is a severely ill-posed problem. In fact, during the acquisition of the low-resolution (LR) images, some high-frequency components are lost or aliased, hindering their correct reconstruction. As a consequence, SISR methods attempt to generate plausible reconstructions compatible with the LR image, rather than to recover the real high resolution (HR) image. In contrast, MISR aims at exploiting the alias to retrieve the true details in the super-resolved image (SR) by combining the non-redundant information from multiple LR observations.

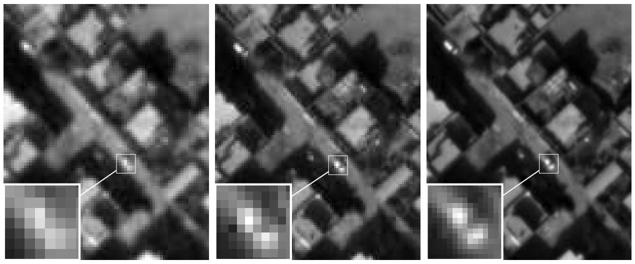

(a) L1A frame (b) L1B (c) Proposed

Figure 1: Super-resolution from a sequence of 15 real low-resolution SkySat L1A frames. (a) Reference L1A frame, (b) Planet L1B product (×1.25), (c) Proposed method (×2).

In this work, we focus on MISR from push-frame satellite sensors such as the SkySat constellation from Planet. The SkySat satellites [44] contain a full-frame sensor capable of capturing bursts of overlapping frames. So the same point on the ground is seen in several consecutive images. Furthermore, thanks to the design of its optical system, the images are aliased, which is an ideal setting for MISR.

In the context of satellite imaging, since the sensor is far from the ground, it is often assumed that the observed scene lies on a plane at infinity. This allows to consider a simplified[1] model [5] for the formation of the low-resolution images I_t^{LR}

$$I_t^{LR} = \Pi A_t(\mathcal{I} * k) + n_t, \tag{1}$$

where $\mathcal{I}$ denotes the infinite-resolution ideal image, k is the Point Spread Function (PSF) modeling jointly optical blur and pixel integration, A_t is a homographic transformation (often approximated by an affine one [5]), Π is the bi-dimensional sampling operator due to the sensor array,

[1] The model should write $(A_t\mathcal{I}) * k$, but in the specific case of rigid transformations, assuming that k is an isotropic kernel, A_t and k commute.

and n_t models the image noise. Because of the spectral decay imposed by the pixel integration and optical blur (k), the image $\mathcal{I}^{bl} := \mathcal{I} * k$ is band limited. For SkySat, the frequency cutoff is at about twice the sampling rate of the LR images. This implies that there is no usable high frequency information beyond the $2\times$ zoom factor. Our goal in this work to increase the resolution by a factor of 2, by estimating I^{HR}, a non-aliased sampling of $\mathcal{I}^{bl}$ from several discrete observations I_t^{LR}. A sharper super-resolved image can also be recovered by partially deconvolving k. Aggregating many frames is also interesting as it allows to greatly reduce the noise.

Lately, deep learning algorithms have proven a success in super-resolution. Data-driven methods can incorporate realistic image priors leading to improved restoration using fewer input images. However, these methods are data-hungry and they heavily rely on the size and quality of the training dataset. The importance of training SISR algorithms with realistic data was highlighted in [12], where it was shown that models trained on synthetic data [2] generalized poorly to a dataset of real pairs of LR/HR images.

MISR datasets with real data are usually small and can only be used as test sets for benchmarking (for example the MDSP dataset [1]). An exception is the PROBA-V dataset, proposed in [39], which allows to train supervised deep-learning MISR methods on real-world satellite images. This is a rare case as the PROBA-V satellite is equipped with two cameras with different resolutions. However, the images in the PROBA-V datasets are unsuitable for training MISR methods for image bursts acquired at a high frame rate, as the LR image sequences are multi-date and present significant content and illumination changes.

Due to this lack of datasets with real LR/HR images, most deep learning MISR algorithms are trained on simulated data [61, 40]. Good results in denoising of real images have been obtained using synthetic datasets [11, 68]. However, this requires a careful modeling of the imaging systems, which is not straightforward for complex satellite sensors.

A similar problem affects other video restoration problems. Recent works [17, 16, 14, 65] have proposed to train video denoising and demosaicking networks with self-supervised learning by exploiting the temporal redundancy in videos. In these works, the network is trained to predict a frame of a noisy sequence using its neighboring frames, eliminating the need for ground truth.

Contributions. In this paper, we propose a framework for self-supervised training of MISR networks without requiring high resolution ground truth images.

Our framework (Sec. 3) can be applied to neural networks that include an explicit motion compensation module. One of the LR frames is set as reference. During training, the reference is only viewed by the motion compensation module (to align the rest of the LR frames) but is withheld from the rest of the network. The network is tasked to predict a super-resolved image which, when downsampled, coincides with the withheld reference frame.

As an additional contribution, we propose a novel MISR architecture, *Deep Shift-and-Add* (DSA), consisting of a shift-and-add fusion of features. Our DSA network accepts a variable number of input frames and is invariant to their order. This allows us to use all available LR frames at test time (including the reference LR frame), which improves the performance.

Experiments conducted on synthetic data (Sec. 4 and 5) show that our DSA network trained with the proposed self-supervision strategy attains state-of-the-art results on par with those obtained with a supervised training. To the best of our knowledge, this is the first method that trains a MISR CNN without supervision. In addition, the proposed method reduces noise, successfully handles degenerate samplings and can integrate the final deconvolution step.

Our framework makes it possible to train a network on datasets of real LR images (Sec. 4). We demonstrate this by training our DSA network on a novel public dataset of real image bursts from SkySat satellites. In a qualitative comparison, we see that the obtained results are more resolved and less noisy than the L1B product from Planet (see Fig. 1).

2. Related works

Video and burst super-resolution. There is a long history of MISR techniques from bursts of images and videos (see [45, 67] for more comprehensive reviews). Most MISR methods are based on two steps: subpixel registration between the LR images and fusion into the super-resolved image. Several fusion strategies have been proposed: local kernel regression [61, 55], variational formulations [57, 38, 18, 49], and fusion in transformed domains [30, 34, 46, 5].

One of the simplest classical strategies is the *shift-and-add* method, in which the pixel values of the low resolution image are shifted according to an estimated motion with respect to a common reference and accumulated in a high resolution image [28]. We incorporate a *feature shift-and-add* module inspired from these methods.

Currently, the state of the art in MISR is dominated by neural networks. Existing approaches can be classified based on the motion compensation strategy. Approaches based on *explicit motion compensation* estimate the motion field between pairs of LR frames and use it to register them. Most methods use backward warping (or pullback) to obtain the registered frames, which requires interpolating the LR frame to be registered [51, 63, 15]. Instead, our DSA architecture uses forward warping (or push forward), where the LR pixels are aggregated into the high resolution grid.

A similar approach is followed in [56], except that the forward warping is applied to the input frames, while we apply it to a feature representation.

Since the motion estimation might be prone to errors, especially with optical flow methods for video, some approaches avoid to explicitly represent motion. Different strategies have been proposed for *implicit motion compensation*: dynamic upsampling filters [27], deformable convolutions [60], progressive fusion residual blocks [64]. Other approaches do not compensate for motion at all and simply present the data to a network, hoping that the motion compensation will be learnt through training [20, 24].

Super-resolution for satellite images. Most MISR methods for satellite images are still based on classic model-based techniques [34, 41, 44, 5, 6]. Obtaining realistic databases with ground truth is the main challenge for training data-driven MISR methods for satellite imagery, as all existing approaches rely on supervised training.

In the case of SISR, some methods resort to simulating realistic data [69] or to combine images acquired from different satellites with different resolutions [48, 52] so as to avoid the synthetic downsampling.

To the best of our knowledge, the only dataset with real LR and HR satellite images is the PROBA-V dataset [39]. This dataset and the associated challenge have triggered research in MISR of satellite imagery [13, 53, 42, 43]. In the PROBA-V dataset, the LR reference (the LR image associated to the HR target) is unknown. This last point was analyzed in [47], where the authors propose PROBA-V-ref, an alternative version of the PROBA-V challenge where the identity of the reference image is provided, a setting which is more relevant to our application.

Learning without ground truth. Lehtinen *et al.* [36] showed that an image denoising network can be trained from pairs of noisy versions N and N' of the same image I with independent noise realizations, by minimizing the following noise-to-noise (N2N) risk:

$$\mathcal{R}_{\text{N2N}}(\textbf{Net}) = \sum_j \ell(\textbf{Net}(N_j), N'_j). \tag{2}$$

Intuitively, since the noise realizations are independent, the noise in N' cannot be predicted from N. Hence, the loss is minimized by estimating the clean image. The optimal estimators for the N2N risk are given by $\mathbb{E}\{N'|N\}$ for the MSE loss, and median$\{N'|N\}$ for the L_1 loss. It can be shown that if the noise in N' preserves the mean, then $\mathbb{E}\{N'|N\} = \mathbb{E}\{I|N\}$, i.e. training with the supervision of the noisy images is equivalent to the one supervised by the clean ones. It was also empirically observed that a similar property holds for the L_1 loss if the noise in N' preserves the median.

Noise-to-noise has inspired several works in self-supervised training of denoising networks. For still images, [33, 8] train a network to predict noisy pixels from their surroundings, thus eliminating the need for the second noisy observation, albeit with a penalty in the quality of the results. In the context of video or bursts of images, the situation is more favorable as a neighboring frame can be used as noisy target (after proper alignment). The frame-to-frame method [17] applied this idea to fine-tune a single frame denoising (and/or demosaicking [16]) network requiring only a single noisy video or burst. Extensions were proposed in [14, 65] for multi-frame denoising networks by withholding the target frame from the inputs to the network. Our self-supervised training draws inspiration from frame-to-frame approaches [17, 16, 14].

Other self-supervision strategies were also explored for SISR. In [54], an algorithm that exploits the internal recurrence of information across scales inside a single image is proposed. The authors of [66, 29] propose to use cycle-consistency and adversarial losses to train a SISR neural network without supervision using unpaired LR and HR images. In [37], an extension of the Deep Image Prior [58] is applied to fine-tune a SISR network on a single image.

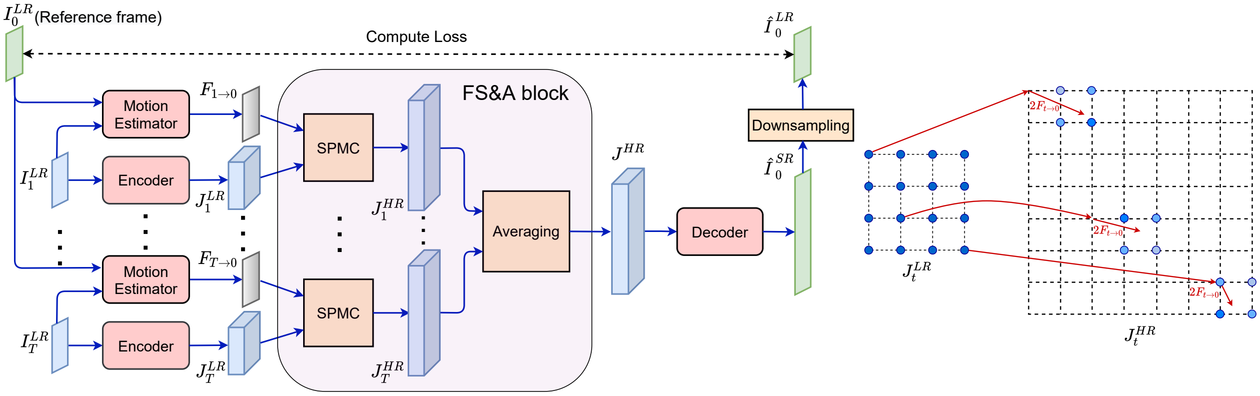
3. Self-supervised multi-image SR

We first present an overview of our proposed *Deep Shift-and-Add* in Sec. 3.1. Then we describe our framework for self-supervised MISR training in Sec. 3.2, and in Sec. 3.3 we provide details about the training.

3.1. Architecture

Our neural network (illustrated in Fig. 2a) takes as input a sequence of LR images $\{I_t^{LR}\}_{t=0}^{T}$ and produces one super-resolved image $\hat{I}_0^{SR}$. The architecture draws inspiration from the traditional shift-and-add MISR algorithms, especially those that perform a weighted average of the aligned LR image samples depending on their subpixel positions [19, 41, 22, 3, 26].

To this aim, the motion fields between all the LR frames in the burst and a reference one I_0^{LR} are first estimated with a trainable motion estimation module. Then, the frames are upscaled and aligned by compensating the motion using a Subpixel Motion Compensation [56] layer (SPMC). The SPMC layer was originally proposed to feed motion compensated frames into a video SR network. However, in our case, we apply it to convolutional features J_t^{LR} extracted from the frames I_t^{LR} as it has been shown that deep feature representations encode at each pixel a rich description of the local neighborhood [9, 13, 62]. The upscaled and aligned features J_t^{HR} are then averaged in a high resolution feature map J^{HR}. The SR image is then obtained by decoding J^{HR}. In summary, the action of the network can be described in three steps: encoding, temporal feature



(a) Overview of our architecture and framework

(b) Illustration of the SPMC layer

Figure 2: (a) Overview of our proposed self-supervised MISR framework at training time. The depicted loss represents the self-supervision term ℓ_{self}, for simplicity the losses concerning the motion estimation module are not illustrated. Note that at inference time the frame I_0^{LR} is also encoded and fed to the FS&A block. (b) SPMC ×2 layer [56]: Splatting LR features onto the HR domain using the flow $F_{t\to0}$.

aggregation, and decoding. The temporal aggregation is done simply by feature averaging, via a *feature shift-and-add* block. This schema allows to aggregate an arbitrary number of frames and is permutation invariant. We will exploit these properties later in Sec. 3.2.

The trainable modules of the proposed architecture (shown in red in Fig. 2a) include the Motion Estimator, the Encoder and the Decoder.

Motion Estimator. We follow the work of [51] to build the network **ME** used to estimate the optical flows between each LR frame $\{I_t^{LR}\}_{t=1}^T$ and the reference frame I_0^{LR}

$$F_{t\to0} = \mathbf{ME}(I_t^{LR}, I_0^{LR}; \Theta_{\mathbf{ME}}) \in [-R, R]^{H\times W\times 2}. \quad (3)$$

The parameters of the Motion Estimator are denoted $\Theta_{\mathbf{ME}}$. A small Gaussian filter ($\sigma = 1$) is applied to the input images to reduce the alias [59]. This network will be trained with a maximum range of motions $[-R, R]^2$ (in this work, $R = 5$ pixels).

The **ME** network follows a simple hourglass style architecture (4 scales with 32, 64, 128 and 256 features, 2 convolutions blocks per scale [51]). More complex methods can be adopted, but since in our application, the apparent motion is mainly caused by the motion of the satellite, a smooth motion estimate suffices.

Encoder. The Encoder module generates relevant features $(J_t^{LR})_{t=1}^T$ for each LR image in the sequence

$$J_t^{LR} = \mathbf{Encoder}(I_t^{LR}; \Theta_{\mathbf{E}}) \in \mathbb{R}^{H\times W\times N}, \quad (4)$$

where $\Theta_{\mathbf{E}}$ is the set of parameters of the encoder and $N = 64$ is the number of produced features. The network comprises 2 convolutional layers at the two ends of a series of 4 residual blocks with 64 features per layer.

Feature Shift-and-Add. A shift-and-add process is used to map and aggregate feature pixels to their positions in the HR grid using the corresponding optical flows. We separate the process in two: first the features of each frame are upscaled by introducing zeros between samples and motion compensated with the SPMC module [56], then a weighted average is computed.

The SPMC module uses the flow $F_{t\to0}$ to compute the positions of the samples from J_t^{LR} in the HR grid

$$J_t^{HR} = \text{SPMC}(J_t^{LR}, \{F_{t\to0}\}) \in \mathbb{R}^{rH\times rW\times N}, \quad (5)$$

where r is the upscaling factor ($r = 2$ in our case). As in [56], every LR pixel is "splatted" on a neighborhood of the computed HR position using bilinear weights (see Fig. 2b). In this way, the operation is differentiable with respect to both the intensities and the optical flows. We perform a weighted aggregation of J_t^{HR}

$$J^{HR} = (\textstyle\sum_t J_t^{HR})(\sum_t W_t^{HR})^{-1}, \quad (6)$$

where $W_t^{HR} = \text{SPMC}(1, \{F_{t\to0}\})$ are the sum of the bilinear weights affecting every pixel. Note that the feature shift-and-add does not have any trainable parameters.

Decoder. The Decoder network reconstructs the SR image $\hat{I}_0^{SR}$ from the fused features

$$\hat{I}_0^{SR} = \mathbf{Decoder}(J^{HR}; \Theta_{\mathbf{D}}) \in \mathbb{R}^{rH\times rW}, \quad (7)$$

where $\Theta_{\mathbf{D}}$ denotes the set of parameters of the decoder. Our decoder comprises 2 convolutional layers at the two ends of a series of 10 residual blocks with 64 features.

3.2. Self-supervised learning

The proposed self-supervised training relies on the minimization of a reconstruction loss in the LR domain plus a motion estimation loss to ensure accurate alignment. Each loss is detailed in the following paragraphs.

Self-supervised SR loss. From the formation model in (1), we see that the LR images I_t^{LR} and the target high resolution image I^{HR} capture the same underlying image $\mathcal{I}^{bl}$, only the sampling and noise differs.

During self-supervised training, LR sequences are randomly selected and for every sequence one frame is set apart as the reference I_0^{LR}. Then, all other LR images in each sequence are registered against I_0^{LR}. Assuming that the registration is perfect, the registered LR images correspond to noisy samples of $\mathcal{I}^{bl}$. Thus, ignoring the noise, I_0^{LR} could be used as target for the fraction of pixels it contains. More specifically, the proposed self-supervised loss writes

$$\ell_{self}(\hat{I}_0^{SR}, I_0^{LR}) = \|D_2(\hat{I}_0^{SR}) - I_0^{LR}\|_1, \tag{8}$$

where $\hat{I}_0^{SR} = \mathbf{Net}(\{I_t^{LR},\}_{t=1}^T, I_0^{LR})$ is the network output and D_2 is the subsampling operator that takes one pixel over two in each direction. That is, the proposed self-supervised loss aims at training the network to produce an image such that when subsampled, it coincides with the target I_0^{LR}. Following noise-to-noise [36], if the noise in the LR frames is independent, the network is unable to predict the noise in I_0^{LR} and it learns to output a noise-free image. The use of the L1 norm in the loss is adapted for frame-independent median preserving noise, as shown in the noise-to-noise framework [36, 17].

Note that in the proposed architecture, the image I_0^{LR} is also an input of **Net** as the super-resolved image $\hat{I}_0^{SR}$ has to be aligned with I_0^{LR}. Usually in self-supervised learning, the target is excluded from the network inputs during training in order to avoid trivial solutions [8, 14]. In our case, the network could achieve zero loss by learning to copy the reference LR frame I_0^{LR} in the subsampled pixels of the super-resolved image $D_2(\hat{I}_0^{SR})$. However, this is not a problem in our framework since the reference I_0^{LR} is only used to estimate the flows and does not enter the encoder path, thus the encoder and the decoder must learn to reproduce I_0^{LR} without having access to it. At test time, since the network has been trained to handle a variable number of input LR frames, the reference frame can be added to the inputs together with the rest of the LR frames.

In conclusion, as long as the network architecture contains an explicit motion estimation module that is decoupled from the fusion module, our framework can be applied to provide self-supervised training.

Motion estimation loss. To ensure a good alignment of the LR frames, we use a motion estimation loss consisting in a photo-consistency term and a regularization term, as the ones used for unsupervised training of optical flow [25]. The loss is computed for each flow $F_{t\to 0} = \mathbf{ME}(I_0^{LR}, I_t^{LR}, \Theta_{\mathbf{ME}})$ estimated by the **ME** module:

$$\ell_{me}(\{F_{t\to 0}\}_{t=1}^T) = \textstyle\sum_t \|I_t^{LR} - \mathbf{Pullback}(I_0^{LR}, F_{t\to 0})\|_1 \\ + \lambda_1 TV(F_{t\to 0}), \tag{9}$$

where **Pullback** computes a bicubic warping of I_0^{LR} according to a flow, TV is the finite difference discretization classic Total Variation [50] regularizer, and λ_1 is a hyperparameter controlling the regularization strength. A small Gaussian filter ($\sigma = 1$) is also applied to the images I_0^{LR}, I_t^{LR} to reduce the alias.

3.3. Training details

We first pre-train the motion estimator on our dataset, and then train the whole system end-to-end. While this is not strictly necessary, it stabilizes the training and accelerates the convergence [56].

To pre-train the motion estimation network we use the motion estimation loss (9). We initialize the weights of the motion estimator with Xavier's initialization [21]. In our experiments, we set λ_1 to 0.01 and batch size to 64, then use Adam [31] with the default Pytorch parameters and a learning rate of 10^{-4} to optimize the loss. The pre-training converges after 20k iterations and takes about 5 hours on one NVIDIA V100 GPU.

We then train the entire system end-to-end using the complete loss:

$$\text{loss} = \ell_{self} + \lambda_2 \ell_{me}. \tag{10}$$

We set $\lambda_2 = 10$ in our experiments. The initial weights are set using He initialization [23], except for the motion estimator whose initial weights are the pre-trained ones.

For our experiments with simulated data, we also train a supervised model which is used as a reference (see Sec. 4.2). In that case, we replace ℓ_{self} in (10) by

$$\ell_{supervised}(\hat{I}_0^{SR}, I^{HR}) = \|\hat{I}_0^{SR} - I^{HR}\|_1, \tag{11}$$

which uses supervision from the high resolution target I^{HR}.

We train both supervised and self-supervised models on LR crops of size 64 × 64 pixels and validate on LR images of size 256 × 256 pixels. During training, our network is fed with a random number of LR input images (from 5 to 30) in each sequence. We set the batch size to 16 and optimize the loss using the Adam optimizer with default parameters. Our learning rates are initialized to 10^{-4} and scaled by a factor of 0.3 when the validation loss plateaus for more than 30 epochs. The training converges after 300 epochs and it takes about 18 hours on one NVIDIA V100 GPU.

4. Experiments

In our experiments, we use real push-frame images acquired by satellites from the SkySat constellation [44]. These images are also used to create a simulated dataset used for a quantitative evaluation.

4.1. Datasets

SkySat imagery. The SkySat satellites contain a full-frame sensor capable of capturing 40 frames per second and is mainly operated in a *push-frame* mode with significant overlap between the frames. As a result, the same point on the ground is seen in at least 15 consecutive images. The individual low-resolution frames are called L1A products. Planet also provides a super-resolved product (called L1B) that corresponds to a ×1.25 zoom of the L1A images and has a resolution between 50 to 70 cm/pixel at nadir. It is important to note that the L1B product has also undergone an unknown sharpening, so it is not easily comparable to the L1A images.

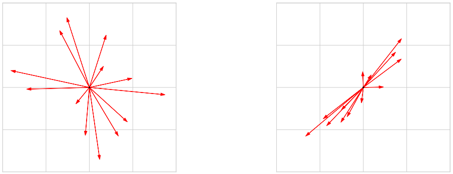
Simulated dataset. A part of our experiments will be conducted on a simulated dataset generated from a set of crops of L1B products. For a given crop B, the ground truth HR image I^{HR} is computed by filtering B with a small Gaussian kernel with $\sigma = 0.3$ so as to simulate a small optical blur. Random shifts (sampled uniformly on a disk) and a $\times 2$ subsampling are then applied to I^{HR} to obtain the set of LR images

$$\begin{aligned} I_0^{LR} &= D_2(I^{HR}) + n_0, \\ I_t^{LR} &= D_2(\text{Shift}_{\Delta_t}(I^{HR})) + n_t, \quad t = 1, \dots, T, \end{aligned} \tag{12}$$

where D_2 is the subsampling operator, Shift_{Δ_t} applies a subpixel translation of Δ_t with Fourier interpolation ($\|\Delta_t\|_1 \leq 2$) to the image and n_t models the noise.

Our simulated data was generated from 370 L1B images of size 3200 × 1350 pixels. We use 320 images for training and 50 for validation. From each image, random crops are extracted to generate bursts of 30 noisy LR frames with additive white Gaussian noise of standard deviation $3/255$. The size of the crops in the training set is 64×64 pixels and in the validation set is 256 × 256 pixels.

The relative position of the samples of the set of LR images is a critical aspect of the MISR problem. When the random shifts are drawn uniformly the restoration problem is usually well-posed. But, due to the motion of the satellite, real sampling configuration can be degenerate, i.e. with all the shifts aligned along the same direction. This is a critical situation for many traditional MISR algorithms that require additional regularization as the problem becomes ill-posed. Ignoring these degenerate configurations during training can result in poor performance in similar cases. Thus, in our main simulated dataset, we simulate a mixture of 80% uniform sampled sequences and 20% degenerate sampled sequences, in which the samples are allocated in a narrow ellipse as shown in Fig. 3. We also generate datasets with 100% uniform and 100% degenerate samplings. We refer to them as *mixed*, *uniform*, and *degenerate*.



(a) Uniform sampling

(b) Degenerate sampling

Figure 3: Uniform and degenerate sampling. The vectors represent the global shifts between the LR frames in a simulated sequence. (a) In the uniform sampling these shifts are uniformly distributed in a disk. (b) In the degenerate sampling these shifts are distributed in a narrow ellipse.

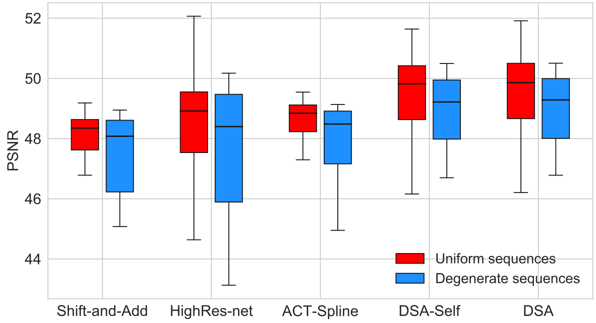


Figure 4: PSNR of different methods over our main validation set with 16 input frames per sequence.

Table 1: Average PSNR (dB) over the validation dataset for different methods with different number of input images per sequence. Our solutions are highlighted in bold.

Method	Shift-and-Add	HighRes-net	ACT-Spline	**DSA-Self-noref**	**DSA-Self**	**DSA**
T = 5	42.99	45.63	45.54	**45.70**	**45.75**	**45.82**
T = 16	47.72	48.17	48.38	**49.18**	**49.27**	**49.33**
T = 30	49.95	49.05	50.15	**50.38**	**50.45**	**50.50**

Dataset of real images. For our experiments on real data, we selected 48 reference SkySat L1A images, and 15 frames overlapping each reference. The stacks of L1A images are pre-aligned to each reference with a discrete translation avoiding any resampling. From each reference image, we randomly crop 20 blocks of size 256×256 pixels, yielding 960 stacks of 15 frames in total, including 60 stacks for the validation set.[2] For each stack, the L1B product from Planet is also extracted, which will only be used for visual comparison as we do not know which sharpening was used.

4.2. Super-resolution on simulated data

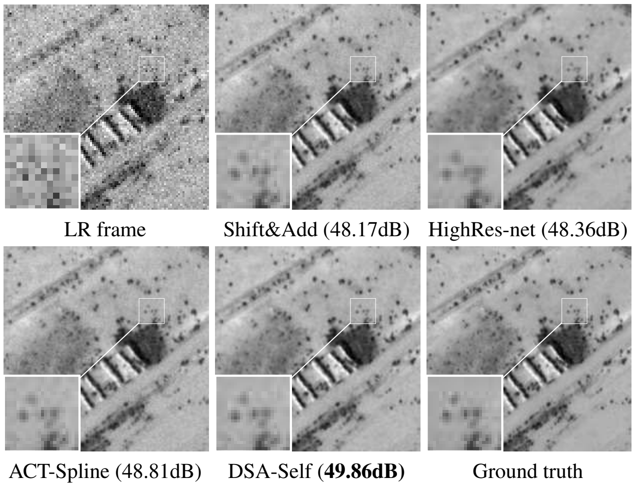
We evaluate the performance of our super-resolution network on the simulated dataset described in Sec. 4.1

[2] This dataset can be downloaded from the project webpage.

and compare against three methods from the literature: *Shift-and-Add*, *ACT-Spline* and *HighRes-net*. The classical *Shift-and-Add* with bilinear splatting will serve as baseline [41, 22, 3, 26]. *ACT-Spline* is a state-of-the-art method based on spline fitting [5]. In Shift-and-Add and ACT-Spline, the LR images are aligned using the inverse compositional method [7, 10]. *HighRes-net* is a MISR CNN with implicit motion estimation trained originally for the PROBA-V challenge [13]. Here we use a variant that was shown in [47] to have a better performance on the PROBA-V-ref dataset. We retrained HighRes-net with supervision on our simulated dataset, following the procedure of [47] but doubling the number of epochs to ensure convergence. As a reference for comparison with SISR methods, we also retrained SRGAN [35] on our dataset. As one could expect from a SISR method, many details are lost. The results are reported in the supplementary material.

Table 1 shows the results of the three methods plus our DSA network (with both supervised and self-supervised training) on the simulated validation set using 5, 16 and 30 input frames. Fig. 4 breaks down the performance of the different methods over the mixed validation dataset for the case with 16 input frames. Our supervised network ranks first, with a significant 0.95dB gain ($T = 16$) over ACT-Spline which was hand-tuned [5] on a dataset of SkySat images. HighRes-net performs 0.23dB worse than ACT-Spline and this gap grows to 1.1dB for $T = 30$. The outputs of HighRes-net are noiseless but tend to be over-smoothed. It seems that for longer bursts, HighRes-net has problems fusing the complementary information of the LR frames. Note that HighRes-net was also trained by varying the number of LR frames. The *DSA-Self-noref* column shows the performance of our self-supervised network when the reference LR image I_0^{LR} is excluded from the fusion step at test time. In this case, we add an additional LR image to maintain the total number of LR images for a fair comparison. This shows that a small gain can be obtained by including I_0^{LR}.

Fig. 5 presents a qualitative comparison between Shift-and-Add, HighRes-net, ACT-Spline and the proposed DSA-Self on a sequence of 16 frames from the validation set with uniform sampling of the shifts. The output of our supervised DSA (49.93dB) was not included as it was indistinguishable to the DSA-Self result. In this example, DSA-Self outperforms the other methods by more than 1dB. On the zoomed area, we can see that our method recovers faithfully the details on the ground. We remark that our network has *never seen any ground truth* HR image during training. It is optimized only by penalizing the loss between the downsampled version of the output and the noisy LR reference frame over a training dataset. On the other hand, the outputs of Shift-and-Add and ACT-Spline are noisy while the one produced by HighRes-net is too blurry and the black spots are barely distinguishable on the field.


LR frame | Shift&Add (48.17dB) | HighRes-net (48.36dB)

ACT-Spline (48.81dB) | DSA-Self (**49.86dB**) | Ground truth

Figure 5: Comparison with other methods in the case of a uniform sequence with 16 LR frames.

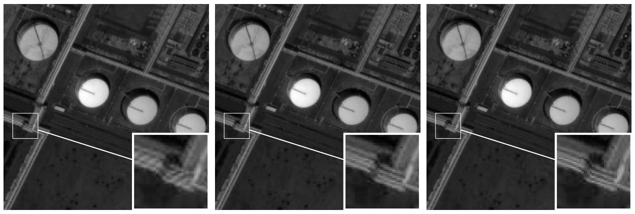

(a) HighRes-net (b) DSA-Self (c) Ground truth

Figure 6: HighRes-net reconstruction from a degenerate sequence of 16 frames presents strong aliasing artifacts.

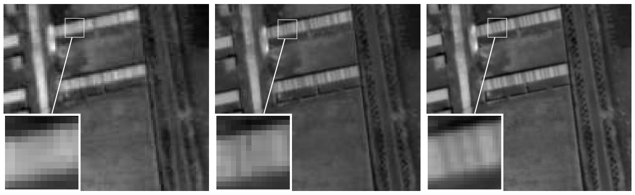

(a) L1A frame (b) Planet L1B (×1.25) (c) DSA-Self (×2)

Figure 7: Super-resolution from a sequence of 15 SkySat L1A frames. (c) was obtained using Eq. (13) as reconstruction loss with deconvolution.

Reproducing the same experiment but with the degenerate samplings, we observe that HighRes-net fails to remove aliasing artifacts of the LR frames (see Fig. 6), despite being trained with such configurations. We argue that the network was not able to exploit the alias in the images failing at increasing the resolution.

4.3. Super-resolution trained on real data

We applied our framework to train our DSA-Self network on the dataset of real SkySat L1A bursts. Since there

Table 2: Evaluation of the impact of training the proposed DSA network with a variable number of input images (rows *variable* or a *fixed* (16) number of inputs) and considering degenerate sampling configurations or not (rows *mixed* or *uniform* datasets).

Training		**Testing on mixed dataset**			**Testing on uniform dataset**			**Testing on degenerate dataset**		
Number of images	**Training dataset**	**Number of images**			**Number of images**			**Number of images**		
		5	16	30	5	16	30	5	16	30
Variable	Mixed	**45.82**	**49.33**	50.50	45.77	49.26	50.41	**45.40**	**48.75**	49.81
Variable	Uniform	45.78	49.27	50.43	**45.79**	**49.31**	50.39	45.34	48.68	49.74
Fixed (16)	Mixed	45.55	49.29	**50.52**	45.52	49.23	50.43	45.15	48.71	**49.83**
Fixed (16)	Uniform	45.32	49.16	**50.52**	45.37	49.20	**50.49**	44.94	48.59	49.81

is no ground truth we conduct a qualitative evaluation comparing with the L1B product from Planet. We recall that our method estimates a high-resolution (but blurry) image sampled from $\mathcal{I}^{bl} := \mathcal{I} * k$, while the L1B product has undergone an unknown sharpening step.

As we do not know the optical characteristics of the SkySat satellites, following [5] we consider a blur kernel k' such that when inverted, the reconstruction is visually well-contrasted. We model our blur kernel in the frequency domain as $\hat{k}'(\boldsymbol{\omega}) = (5|\boldsymbol{\omega}| + 1)^{-1}$. The sharp image could then be obtained by solving a variational non-blind deconvolution problem [4, 32] as in [5]. Instead, we opt for incorporating the deconvolution in the self-supervision loss

$$\ell_{self}(\hat{I}_0^{SR}, I_0^{LR}) = \|D_2(\hat{I}_0^{SR} * k') - I_0^{LR}\|_1. \tag{13}$$

By embedding a deconvolution into the training, the network produces directly a sharp SR image without introducing unwanted high-frequency artifacts (see the supplementary material for a comparison of both techniques).

Fig. 1 and 7 show side-by-side comparisons of results obtained on the validation dataset. As we can see, L1B products present strong stair-casing artifacts. The fine details like the vehicle in the Fig. 1 and the vertical bars in the Fig. 7 are much sharper in the proposed method.

At inference time, our proposed method takes 0.6 seconds to produce a ×2 super-resolved image from a sequence of 15 L1A images (256 × 256 pixels).

5. Ablation Study

To analyze the importance of the different elements of the proposed architecture, we perform experiments in a supervised setting. First, we re-trained our DSA with different number of features produced by the encoder (4, 16 and 64) and evaluated on the 50 validation sequences comprising 16 images. The obtained PSNRs were respectively 48.81, 49.06 and 49.33dB. Thus, 64 features yields the best results. We observed that more features led to diminishing returns. We also tested an architecture without encoder that directly aggregates pixels, but it under-performed with few input images (see the supplementary material for details).

Lastly, we studied the impact of training with a variable number of input images and considering degenerate sampling configurations. The four rows in Table 2 correspond to networks trained: 1. with a *variable* or a *fixed* (16) number of input images; 2. using the *mixed* or the *uniform* dataset. We evaluated the networks on the uniform, degenerate, and mixed test datasets using different number of input images (5, 16, 30). The results show that training with a variable number of input images and with a mixed dataset leads to a network that is more resilient to having less input frames and that can cope with degenerate sampling configurations.

6. Conclusion

We presented a framework for the self-supervised training of multi-image super-resolution networks without requiring ground truth. For our framework to be applicable, the networks need an explicit motion compensation module. In addition, we proposed DSA, a novel MISR architecture consisting of a shift-and-add fusion of features. Our experiments on simulated data showed that the proposed self-supervision strategy attains state-of-the-art results, on par with those obtained with a supervised training. As our framework makes it possible to train a network solely from datasets of real LR images, we trained DSA on real SkySat satellite image bursts, leading to results that are more resolved and less noisy than the L1B product from Planet.

Acknowledgments. 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 2020-AD011011801) 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.

References

[1] MDSP super-resolution and demosaicing datasets. https://users.soe.ucsc.edu/~milanfar/software/sr-datasets.html. Accessed: 2021-03-15. 2

[2] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 126–135, 2017. 2

[3] Mohammad S Alam, John G Bognar, Russell C Hardie, and Brian J Yasuda. Infrared image registration and high-resolution reconstruction using multiple translationally shifted aliased video frames. *IEEE Transactions on instrumentation and measurement*, 49(5):915–923, 2000. 3, 7

[4] Jérémy Anger, Mauricio Delbracio, and Gabriele Facciolo. Efficient blind deblurring under high noise levels. In *IEEE ISPA*, pages 123–128, 2019. 8

[5] Jérémy Anger, Thibaud Ehret, Carlo de Franchis, and Gabriele Facciolo. Fast and accurate multi-frame super-resolution of satellite images. *ISPRS*, 2020. 1, 2, 3, 7, 8

[6] Jérémy Anger, Thibaud Ehret, and Gabriele Facciolo. Parallax estimation for push-frame satellite imagery: application to super-resolution and 3d surface modeling from skysat products. *arXiv preprint arXiv:2102.02301*, 2021. 3

[7] S. Baker and I. Matthews. Equivalence and efficiency of image alignment algorithms. In *CVPR*, 2001. 7

[8] Joshua Batson and Loic Royer. Noise2Self: Blind Denoising by Self-Supervision. *ICML*, 2019. 3, 5

[9] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deep burst super-resolution. *arXiv preprint arXiv:2101.10997*, 2021. 3

[10] Thibaud Briand, Gabriele Facciolo, and Javier Sánchez. Improvements of the Inverse Compositional Algorithm for Parametric Motion Estimation. *IPOL*, 8:435–464, 2018. 7

[11] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11036–11045, 2019. 2

[12] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3086–3095, 2019. 2

[13] Michel Deudon, Alfredo Kalaitzis, Israel Goytom, Md Rifat Arefin, Zhichao Lin, Kris Sankaran, Vincent Michalski, Samira E Kahou, Julien Cornebise, and Yoshua Bengio. Highres-net: Recursive fusion for multi-frame super-resolution of satellite imagery. arxiv 2020. *arXiv preprint arXiv:2002.06460*, 2020. 3, 7

[14] Valéry Dewil, Jérémy Anger, Axel Davy, Thibaud Ehret, Gabriele Facciolo, and Pablo Arias. Self-supervised training for blind multi-frame video denoising. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2724–2734, 2021. 2, 3, 5

[15] Ping Du, Jinhuan Zhang, and Jun Long. Super-sampling by learning-based super-resolution. *International Journal of Computational Science and Engineering*, 21(2):249–257, 2020. 2

[16] Thibaud Ehret, Axel Davy, Pablo Arias, and Gabriele Facciolo. Joint demosaicing and denoising by overfitting of bursts of raw images. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2, 3

[17] Thibaud Ehret, Axel Davy, Jean-Michel Morel, Gabriele Facciolo, and Pablo Arias. Model-blind video denoising via frame-to-frame training. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 3, 5

[18] Gabriele Facciolo, Andrés Almansa, Jean-François Aujol, and Vicent Caselles. Irregular to Regular Sampling, Denoising, and Deconvolution. *Multiscale Modeling & Simulation*, 7(4):1574–1608, jan 2009. 2

[19] Andrew Fruchter and Richard Hook. Drizzle: A method for the linear reconstruction of undersampled images. *Publications of the Astronomical Society of the Pacific*, 114(792):144, 2002. 3

[20] Dario Fuoli, Shuhang Gu, and Radu Timofte. Efficient video super-resolution through recurrent latent space propagation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3476–3485. IEEE, 2019. 3

[21] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010. 5

[22] Thomas J Grycewicz, Stephen A Cota, Terrence S Lomheim, and Linda S Kalman. Focal plane resolution and overlapped array TDI imaging. In *Remote Sensing System Engineering*, volume 7087, page 708704. International Society for Optics and Photonics, 2008. 3, 7

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 5

[24] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. Video super-resolution with recurrent structure-detail network. In *European Conference on Computer Vision*, pages 645–660. Springer, 2020. 3

[25] J Yu Jason, Adam W Harley, and Konstantinos G Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In *European Conference on Computer Vision*, pages 3–10. Springer, 2016. 5

[26] Yunwei Jia. Method and apparatus for super-resolution of images, Nov. 6 2012. US Patent 8,306,121. 3, 7

[27] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3224–3232, 2018. 3

[28] Danny Keren, Shmuel Peleg, and Rafi Brada. Image sequence enhancement using sub-pixel displacements. In *CVPR 88*, pages 742–743, 1988. 2

[29] Gwantae Kim, Jaihyun Park, Kanghyu Lee, Junyeop Lee, Jeongki Min, Bokyeung Lee, David K Han, and Hanseok Ko. Unsupervised real-world super resolution with cycle generative adversarial network and domain discriminator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 456–457, 2020. 3

[30] SP Kim, Nirmal K Bose, and Hector M Valenzuela. Recursive reconstruction of high resolution image from noisy undersampled multiframes. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(6):1013–1027, 1990. 2

[31] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[32] Dilip Krishnan and Rob Fergus. Fast image deconvolution using hyper-laplacian priors. In *NIPS*, pages 1033–1041, 2009. 8

[33] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2Void - Learning Denoising From Single Noisy Images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2129–2137. IEEE, jun 2019. 3

[34] Christophe Latry and Bernard Rougé. Optimized sampling for CCD instruments: the supermode scheme. In *IGARSS*, volume 5, pages 2322–2324. IEEE, 2000. 2, 3

[35] 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. 7

[36] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. *arXiv preprint arXiv:1803.04189*, 2018. 3, 5

[37] Alice Lucas, Santiago Lopez-Tapia, Rafael Molina, and Aggelos K Katsaggelos. Self-supervised fine-tuning for correcting super-resolution convolutional neural networks. *arXiv preprint arXiv:1912.12879*, 2019. 3

[38] Antonio Marquina and Stanley J Osher. Image super-resolution by tv-regularization and bregman iteration. *Journal of Scientific Computing*, 37(3):367–382, 2008. 2

[39] Marcus Märtens, Dario Izzo, Andrej Krzic, and Daniël Cox. Super-resolution of PROBA-V images using convolutional neural networks. *Astrodynamics*, 3(4):387–402, 2019. 2, 3

[40] Evan M Masutani, Naeim Bahrami, and Albert Hsiao. Deep learning single-frame and multiframe super-resolution for cardiac mri. *Radiology*, 295(3):552–561, 2020. 2

[41] Maria Teresa Merino and Jorge Nunez. Super-resolution of remotely sensed images with variable-pixel linear reconstruction. *IEEE TGRS*, 45(5):1446–1457, 2007. 3, 7

[42] Andrea Bordone Molini, Diego Valsesia, Giulia Fracastoro, and Enrico Magli. Deepsum: Deep neural network for super-resolution of unregistered multitemporal images. *IEEE Transactions on Geoscience and Remote Sensing*, 58(5):3644–3656, 2019. 3

[43] Andrea Bordone Molini, Diego Valsesia, Giulia Fracastoro, and Enrico Magli. Deepsum++: Non-local deep neural network for super-resolution of unregistered multitemporal images. In *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, pages 609–612, 2020. 3

[44] Kiran Murthy, Michael Shearn, Byron D Smiley, Alexandra H Chau, Josh Levine, and M Dirk Robinson. SkySat-1: very high-resolution imagery from a small satellite. In *Sensors, Systems, and Next-Generation Satellites XVIII*, volume 9241, page 92411E. International Society for Optics and Photonics, 2014. 1, 3, 5

[45] Kamal Nasrollahi and Thomas B Moeslund. Super-resolution: a comprehensive survey. *Machine vision and applications*, 25(6):1423–1468, 2014. 2

[46] Nhat Nguyen and Peyman Milanfar. A wavelet-based interpolation-restoration method for superresolution (wavelet superresolution). *Circuits, Systems and Signal Processing*, 19(4):321–338, 2000. 2

[47] Ngoc Long Nguyen, Jérémy Anger, Axel Davy, Pablo Arias, and Gabriele Facciolo. Proba-v-ref: Repurposing the proba-v challenge for reference-aware super resolution. *arXiv preprint arXiv:2101.10200*, 2021. 3, 7

[48] Darren Pouliot, Rasim Latifovic, Jon Pasher, and Jason Duffe. Landsat super-resolution enhancement using convolution neural networks and Sentinel-2 for training. *Remote Sensing*, 10(3):394, 2018. 3

[49] Mattia Rossi and Pascal Frossard. Geometry-consistent light field super-resolution via graph-based regularization. *IEEE Transactions on Image Processing*, 27(9):4207–4218, 2018. 2

[50] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992. 5

[51] 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. 2, 4

[52] Luis Salgueiro Romero, Javier Marcello, and Verónica Vilaplana. Super-resolution of Sentinel-2 imagery using generative adversarial networks. *Remote Sensing*, 12(15):2424, 2020. 3

[53] Francesco Salvetti, Vittorio Mazzia, Aleem Khaliq, and Marcello Chiaberge. Multi-image super resolution of remotely sensed images using residual attention deep neural networks. *Remote Sensing*, 12(14):2207, 2020. 3

[54] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3118–3126, 2018. 3

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

[56] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4472–4480, 2017. 3, 4, 5

[57] Brian C Tom and Aggelos K Katsaggelos. Reconstruction of a high-resolution image by simultaneous registration,

restoration, and interpolation of low-resolution images. In *ICIP*, pages 539–542. IEEE, 1995. 2

[58] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9446–9454, 2018. 3

[59] Patrick Vandewalle, Luciano Sbaiz, Joos Vandewalle, and Martin Vetterli. Super-resolution from unregistered and totally aliased signals using subspace methods. *IEEE Transactions on Signal Processing*, 2007. 4

[60] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPR Workshops*, 2019. 3

[61] 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 (TOG)*, 38(4):1–18, 2019. 2

[62] Lei Xiao, Salah Nouri, Matt Chapman, Alexander Fix, Douglas Lanman, and Anton Kaplanyan. Neural supersampling for real-time rendering. *ACM Transactions on Graphics (TOG)*, 39(4):142–1, 2020. 3

[63] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019. 2

[64] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3106–3115, 2019. 3

[65] Songhyun Yu, Bumjun Park, Junwoo Park, and Jechang Jeong. Joint learning of blind video denoising and optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 500–501, 2020. 2, 3

[66] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 701–710, 2018. 3

[67] Linwei Yue, Huanfeng Shen, Jie Li, Qiangqiang Yuan, Hongyan Zhang, and Liangpei Zhang. Image super-resolution: The techniques, applications, and future. *Signal Processing*, 128:389–408, 2016. 2

[68] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Cycleisp: Real image restoration via improved data synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2696–2705, 2020. 2

[69] Xiang Zhu, Hossein Talebi, Xinwei Shi, Feng Yang, and Peyman Milanfar. Super-resolving commercial satellite imagery using realistic training data. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 498–502. IEEE, 2020. 3