université
PARIS-SACLAY

High-Definition from Above: Advancing Satellite Imagery Super-Resolution via Self-Supervision

Thèse de doctorat de l'Université Paris-Saclay

École doctorale n°574 : mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques aux interfaces
Unité de recherche : Université Paris-Saclay, CNRS, ENS Paris-Saclay,
Centre Borelli, 91190, Gif-sur-Yvette, France
Référent : ENS Paris-Saclay

Thèse présentée et soutenue à l'ENS Paris-Saclay, le 20 Oct 2023, par

Ngoc Long NGUYEN

Au vu des rapports de :

Composition du jury :

Thèse de doctorat

NNT:

Link for the newest version of the thesis HERE

Remerciements

Contents

1 Introduction

1.1 Motivation

Earth Observation (EO) satellites play a critical role in monitoring and understanding the dynamics of our planet. They provide invaluable data for numerous applications, including but not limited to, weather forecasting, disaster management, environmental monitoring, and urban planning [CVTGC+11, LKC15]. However, the efficacy of these applications is often limited by the resolution of the satellite imagery [YLXC15]. This is where super-resolution (SR) comes into play.

As an image processing technique, SR permits to overcome some limitations of low-resolution satellite imagery by enhancing fine image details and structures that are barely visible in the original images. This augmentation of resolution improves the accuracy of object detection, segmentation, classification, or refining land cover mapping and change detection processes, the enhanced resolution offered by SR empowers more informed and precise decision-making [SVE19, TAH06].

So, why is this application of SR crucial? The answer lies in the inherent challenges faced by remote sensing. Several factors influence the spatial resolution of satellite images. The Point Spread Function (PSF) denotes the system's blurring effect, impacted by factors like diffraction and lens aberrations [AdFF19]. Noise, stemming from factors such as system calibration errors, defective sensors, or other types of noise including photonic, thermal, electronic contribute to image degradation. Importantly, both the diffraction limit that affects the maximum resolution attainable and the photonic noise that impedes image clarity are dependent on the aperture size. A larger aperture mitigates these challenges by reducing the effects of diffraction and allowing more light to reach the sensor, thereby enhancing the signal-to-noise ratio. Nevertheless, the benefits of a larger aperture come with a cost, as it significantly increases the size, weight, and overall expense of the satellite.

Despite these complexities, the demand for high-resolution images is continually on the rise, driven by the need for detailed and accurate information across various fields. From urban planning to environmental monitoring, high-resolution images are integral for precise analyses. SR, by enhancing the resolution of these images, bridges this gap between necessity and limitation.

The economic feasibility of SR is another compelling factor. The deployment of new, high-resolution satellites entails substantial complexity and cost. Although such satellites do exist (e.g. GeoEye-1, WorldView-3), access to their data is often expensive and can be limited. Conversely, SR techniques improve the resolution of images from existing satellites (Planet SkySat, Satellogic Aleph-1, Sentinel-2), making it a more economically viable

option [MSS+14, AEdFF20]. By leveraging the data we already possess, SR optimizes the return on investment in satellite technology.

Recognizing these immense possibilities, this thesis explores the application of deep learning for advancing super-resolution of EO satellite imagery. By harnessing the capabilities of these complex and efficient processing methods, we seek to enhance the quality and cost-effectiveness of satellite technology. The work presented in this thesis endeavors to make a valuable contribution to the field of remote sensing and beyond.

1.2 About optical satellite imaging

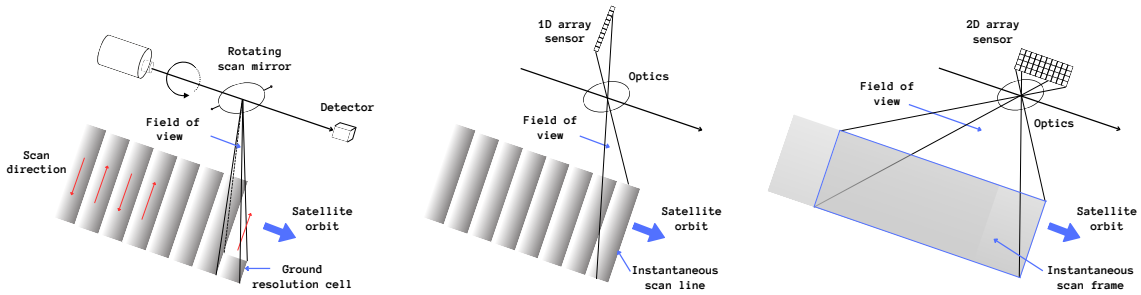
Optical satellite imaging has significantly evolved since its inception (see Fig. 1.1), providing an increasingly comprehensive view of Earth's terrestrial phenomena.

In the 1970s, spaceborne remote sensing was dominated by across-track or whisk-broom scanners, exemplified by the Return Beam Vidicon (RBV) on Landsat 1 (1972) and the Advanced Very High Resolution Radiometer (AVHRR) on NOAA's Polar Operational Environmental Satellites (POES) launched in 1978. These systems relied on a mechanically rotating mirror, scanning one pixel at a time to assemble the image line by line. However, the short exposure time for each pixel led to limited light-gathering capacity, and continuous mirror movements resulted in mechanical wear. These issues catalyzed a transition towards more efficient push-broom imaging technology by the late 1980s.

Push-broom or along-track scanning is adopted by satellites like the French SPOT series (1986 onwards), the European Space Agency's PROBA-V (2013) and Sentinel-2 (2015), and commercial satellites such as IKONOS (1999), QuickBird (2001), GeoEye-1 (2008), and the WorldView series (from 2007). This technique uses an array of detectors arranged in a line, capturing a swath of Earth's surface as the satellite moves along its orbital path. The resultant image exhibits enhanced quality, thanks to the increased dwell time of the scene on the sensor. However, the push-broom technique demands advanced satellite stabilization systems to prevent image smearing and distortions from platform vibrations or fluctuations in satellite's attitude or altitude.

The advent of the 2010s witnessed the emergence of push-frame imaging, epitomized by small, cost-effective satellites like Planet's SkySat (first launched in 2013) and Satellogic's Aleph-1 (2016). These CubeSats use two-dimensional Complementary Metal-Oxide-Semiconductor (CMOS) sensors, capturing a full frame or two-dimensional array of pixels at each shot, creating a series of overlapping images. This overlap not only allows for redundancy against minor errors but also facilitates advanced computational imaging techniques such as burst denoising and super-resolution, despite the need for high-bandwidth data links and complex computational processing.

Processing optical satellite imagery stands in stark contrast consumer cameras due to the unique challenges from data acquisition to the global-scale analysis. Initially, data correction is required, a stage where high-altitude imaging demands mitigating sensor noise, atmospheric distortions, and misalignments due to Earth's curvature. These issues are rarely encountered for consumer cameras. Subsequently, orthorectification and image registration take place. Orthorectification counters perspective and terrain distortions to simulate a nadir viewpoint, while image registration aligns different spectral bands from the same scene, a step not required for typical cameras with Bayer sensors. The concluding phase involves post-processing to amplify image quality for later analysis. Unlike



(a) Whisk-broom instrument (b) Push-broom instrument (c) Push-frame instrument

Figure 1.1: A depiction of different satellite imaging technologies, each illustrating the detector of one spectral band: from whisk-broom (1.1a) scanning pixel by pixel through push-broom (1.1b) scanning line by line to push-frame (1.1c) scanning frame by frame.

enhancing consumer images that often prioritize aesthetics, satellite image post-processing is directed towards scientific interpretability.

This thesis engages with multiple image super-resolution (MISR) and single image super-resolution (SISR) using data from a variety of satellites, each presenting unique challenges. We utilize push-frame SkySat for burst SR, push-broom PROBA-V for multi-date SR, and Sentinel-2 for SISR, noting an increasing degree of complexity across these tasks. While SkySat Burst SR benefits from multiple quick-succession images, PROBA-V Multi-date SR faces more variability due to time-lapse between images. Sentinel-2 SISR, though typically an ill-posed problem, is mitigated slightly due to each spectral band viewing the scene from different perspectives prior to registration, introducing a MISR-like component with varying spectral content. By addressing these nuances, we strive to enhance satellite image super-resolution methodologies.

1.3 About image super-resolution

Based on the number of input images, super-resolution techniques can be broadly categorized into two groups: Multi-Image Super-Resolution (MISR) and Single-Image Super-Resolution (SISR).

1.3.1 Multi-image super-resolution (MISR)

MISR is a technique that merges information from multiple LR images $I_t^{LR}, t \in [1, \dots, K]$ of the same scene to yield a HR output I^{HR} [FREM04b]. These LR images may exhibit slight shifts, or alterations in exposure times, capturing the object from various perspectives or at different instances.

The image formation model can be mathematically described using a pinhole camera [HZ03] along with the processes of geometric transformation, blurring, downsampling, and noise degradation [Mil17]:

$$I_t^{LR} = \Pi\left((F_t \circ \mathcal{I}) * k\right) + n_t, \qquad t \in [1, \dots, K], \tag{1.1}$$

where $\mathcal{I}$ denotes the infinite-resolution ideal image, k is the Point Spread Function (PSF) that jointly models optical blur and pixel integration, F_t represents the motion corresponding to frame t, Π is the bi-dimensional sampling operator due to the sensor array that introduces aliasing, and n_t models the image noise.

We can frame multi-image super-resolution as an inverse problem, employing a mathematical model to understand and invert the image formation process (1.1). Given a set of aliased and noisy LR images, MISR works to recover an HR image which is compatible with the observed LR images. This can be mathematically expressed as follows:

$$I^{HR} = \arg\min_{u} \sum_{t=1}^{K} \left\| \mathbf{Warp}^{\downarrow}((u * k), F_t) - I_t^{LR} \right\|_p , \tag{1.2}$$

where $\mathbf{Warp}^{\downarrow}(., F_t)$ warps and downsamples a HR image with a factor $z \geq 1$, according to the motion F_t

$$\mathbf{Warp}^{\downarrow}(u, F_t)(x) = u(x + zF_t(x)), \tag{1.3}$$

where an interpolation scheme is required as $x + zF_t(x)$ lies outside the HR pixel grid. This expression results from assuming that the blur kernel commutes with the motion F_t, which is a reasonable approximation if the motion is roughly translational within the filter's support. In the absence of noise, and with sufficient LR images the problem can be analyzed in the context of sampling theory and shown to be well-posed [Tsa84] (assuming their relative motion is known). The result is a stable solution capable of revealing the true details of the scene. In most practical cases however, the problem is ill-posed, and there might be many high-frequency recontructions compatible with the available observations. In these cases, some prior or regularization is used to select one among all the posible reconstructions [MO08, ACHR06, FAAC09, PJ07].

MISR techniques can be broadly divided into two main categories: classical methods, which primarily rely on mathematical models and optimization, and more recent approaches based on deep learning.

Classical methods: The existing literature on classical super-resolution methods is vast, encompassing a multitude of strategies. These include frequency domain approaches [KBV90,NM00,RK99] and spatial domain approaches [FREM04a,MSS+14,MN07,TMPE09, WGDE+19, TOS92, AEdFF20].

Shift-and-Add Methods: These works build a HR image by registering multiple LR images and integrating their pixel-level information into the HR grid [KPB88, FREM04a, MSS+14, MN07]. Once registration is completed, every LR pixel is assigned to its nearest HR neighbor or "splatted" across an area of HR pixels according to some interpolation methods, such as bilinear or bicubic. Following this, a final step typically involves weighted aggregation, with the weights correlating to the interpolation used in the preceding step.

For an upscaling factor $z \geq 1$, this process can be mathematically expressed as:

$$\begin{aligned} \mathbf{Warp}^{\uparrow}(I_t^{LR}, F_t) &= \sum_x I_t^{LR}(x)\delta(x + zF_t(x)), \\ J_t^{HR} &= \mathbf{Warp}^{\uparrow}(I_t^{LR}, F_t) * \mathcal{K}, \\ W_t^{HR} &= \mathbf{Warp}^{\uparrow}(\mathbf{1}^{LR}, F_t) * \mathcal{K}, \\ I^{HR} &= \frac{\sum_{t=1}^{K} J_t^{HR}}{\sum_{t=1}^{K} W_t^{HR}}. \end{aligned} \tag{1.4}$$

Here $\mathbf{Warp}^{\uparrow}(I_t^{LR}, F_t)$ registers frame I_t^{LR} onto the HR grid using motion F_t. Since this motion might be sub-pixel, this warped image is represented as an irregular Dirac's comb. $\mathcal{K}$ denotes the interpolation kernel in the HR space, which "splats" the Dirac's pixels

onto the integer HR grid, and W_t^{HR} are the aggregation weights used for normalization, computed by "splatting" an image of ones $\mathbf{1}^{LR}$. The last division is meant element-wise.

In [FREM04c] the authors point out that the shift-and-add algorithm can be seen as the solution of a weighted least squares problem, where the weights are given by splatting kernel $\mathcal{K}$. Based on this observation, they derive a robust version that handles outliers by minizing a sum of absolute errors. This result in an algorithm which computes weighted medians across pixels [FREM04c]. Note that these direct methods can leave holes in the output if the samples are insufficient or degenerate [KPB88]. To deal with potential gaps and outliers, regularizers based on Total Variation (TV) are often incorporated into an energy minimization post-processing step [FREM04a]. It should be noted that the HR output obtained at this stage is inherently blurred due to the impact of the Point Spread Function (PSF) and pixel integration (denoted as k) [MSS+14]. To rectify this, a final deblurring step is routinely performed. This can be achieved by solving the following optimization problem:

$$I_{\text{sharp}}^{HR} = \arg\min_u \left\| u * k - I^{HR} \right\|_p + \mathcal{R}(u), \tag{1.5}$$

where $\mathcal{R}$ is a regularization function, which can include Total Variation (TV), Tikhonov regularization, or a combination of both. The norm p is typically set to either 1 or 2, adapting to the specific characteristics of the blur and the desired sharpness level in the output.

Shift-and-add techniques are among the earliest super-resolution techniques, with results that are far away from the current state-of-the-art. Yet, they still remain relevant due to their simplicity and interpretability. These characteristics, together with the fact that it is a differentiable operation, allow for their seamless integration into deep learning models, something that we leverage in our work.

Deep-learning methods: The field of MISR has experienced a significant transformation with the advent of deep learning techniques. These new methods, emerging from diverse sources such as video and burst denoising, deblurring, and super-resolution, carry great potential [TDV20, MBC+18, SDW+17, TGL+17, SVB18, CWY+21].

Deep learning methods for MISR fall broadly into two categories. The first group tends to favor the development of complex architectures and novel methodologies, often sacrificing interpretability in the process [SCH+16, KLNK18, JWOKJK18, LLTMK19a, DKG+20, LHD+19, MVFM19, SMKC20, AMSC+20]. These techniques typically process a stack of low-resolution images as input and generate a high-resolution output, without explicit motion estimation [SCH+16, JWOKJK18, KLNK18, LLTMK19a, LHD+19, DKG+20, AMSC+20, IJG+20a] or other traditional techniques [MVFM19, SMKC20]. Despite their proven effectiveness, the lack of interpretability makes it challenging to understand why they work and how to adapt them for specific use-cases [CWY+21].

The second category of methods incorporates more transparent and explicit operations, providing clearer insights into their approach [TGL+17, HSU19, BRE19, KBP+19, LPM21, CLK21, LPME22, SJC+22]. While these techniques are more complex and challenging to develop, they often yield superior results due to their explicit incorporation of well-understood principles such as shift-and-add [TGL+17, KBP+19], plug-and-play [VBW13, BRE19, LPM21, LPME22], wavelet transform [CLK21], or optimal transport [PC+19, SJC+22]. These techniques usually require explicit motion estimation, a computationally

intensive task. However in the context of satellite imaging, the high-altitude perspective and simplicity of scene dynamics often render this task more tractable.

Our work falls into the second category, with an aim to blend the advanced neural networks with the interpretability and differentiability of the traditional Shift-and-Add method. This combination allows us to harness the benefits of modern deep learning techniques while preserving the simplicity and robustness of classical methods [FREM04c].

The type of loss is a key element in training MISR methods. Deep learning methods for MISR primarily leverage pixel-based loss functions such as Mean Squared Error (MSE), L1 loss, and Charbonnier loss for optimization [KYDK16, TGL+17, SVB18, CBFAB97, MVFM19, DKG+20, CWY+21, LCF+22, ZGFK16]. These objective functions effectively drive the learning of high-resolution reconstructions but their inherent averaging tendencies can result in a blur effect. However, the degree of this blurring is heavily dependent on the presence of aliasing in the LR images: when the LR images are significantly aliased, the high-frequency information can be better preserved and thus a sharper HR image can be recovered. In scenarios where aliasing is minimal and high-frequency details are not well preserved, a combination of Generative Adversarial Network (GAN) loss and perceptual loss may be employed to generate more visually pleasing and detailed results [CXLT18, LLTMK19a]. However, one should bear in mind that the utilization of these losses can lead to the hallucination of textures and details that are not present in the ground truth, thereby negatively impacting the reconstruction of true details due to the perception-distortion tradeoff [BM18].

1.3.2 Single-image super-resolution (SISR)

SISR is inherently an ill-posed problem since there are multiple HR images that can correspond to a single LR image. The main goal in SISR is to generate high-frequency details that are plausible and convincing to human observers. SISR is widely used in a variety of fields, such as medical imaging [Gre09], surveillance [ZZSL10], and remote sensing [HFBP+18]. In the next section, we provide a brief overview of classical SISR approaches and discuss the recent transition towards deep learning SISR methodologies.

Classical methods: Classical SISR methods can be broadly classified into four main groups: interpolation-based, example-based, sparse representation-based, and variational methods. These techniques are rooted in the same mathematical model as MISR (eq. (1.1) with $K = 1$), but since they rely on a single image, priors are indispensable to mitigate the ill-posedness of the problem.

Interpolation-based techniques, such as bilinear and bicubic interpolation, are straightforward in their approach, using predefined mathematical functions to fill in the gaps between pixels [LGZ13]. Example-based methods, on the other hand, elevate image quality by identifying similar patches within a database or the image itself, then using these to enhance resolution [FJP02, GBI09, HSA15]. Taking a different approach, sparse representation-based methods operate on the presumption that image patches can be expressed as sparse linear combinations of elements from an over-complete dictionary [YWHM10, ZEP12, MBP+09]. Variational methods present SISR as an optimization problem, applying variational inference to find the most plausible solution [MO08, UPWB10, CDL18].

Notwithstanding the array of available methods, each one bears its own set of limitations. The overall quality of results can be significantly hampered by the priors imposed, and

the true high-frequency details are often difficult, if not impossible, to accurately restore from the aliased ones.

Deep-learning methods: The state-of-the-art in SISR is largely dominated by deep learning-based methods. These approaches leverage the power of neural networks, trained on extensive datasets, to model the complex mapping from LR to HR images. Unlike traditional methods, which often rely on handcrafted features and priors, DL methods are capable of automatically learning hierarchical features from data, making them more flexible and powerful in capturing complex image patterns and structures.

The architectural evolution of SISR networks has been fast-paced and transformative. Starting from the simple three-layered SRCNN model [DLHT14], we have seen the adoption of advanced structures that enhance model performance and training efficiency. Residual learning was one such impactful innovation, bypassing the vanishing gradient problem to enable the training of much deeper networks [HZRS16, KKLML16, LSK+17]. Another significant development was the introduction of dense connections, facilitating the extraction of richer hierarchical features and more effective information flow across the network layers [HLVDMW17, TLLG17, ZTK+18]. Most notably, the advent of GANs [LTH+17, WYW+19] and transformer architectures [YYF+20, LCS+21] marked a turning point, offering unique capabilities like the generation of perceptually pleasing images and the ability to model long-range dependencies, respectively.

In terms of training losses, the quality of super-resolution outputs is influenced by the chosen objective function. While traditional loss functions like L_1, L_2, and Charbonnier loss are often utilized, the introduction of GANs has fundamentally shifted the approach in SISR. The adversarial setting of GANs, with a generator producing HR images and a discriminator discerning between real and generated images, has become the mainstream approach in SISR. This paradigm, exemplified by models like SRGAN [LTH+17] and ESRGAN [WYW+19], emphasizes generating HR images that are perceptually closer to real images, focusing less on pixel-wise accuracy and more on superior visual quality. This approach is particularly relevant and effective for most use cases.

However, in satellite image super-resolution, the primary aim is typically to restore true high-frequency details rather than achieving visually pleasing results. Consequently, despite the mainstream success of GANs in image super-resolution, their application is not as suitable for our specific context. Therefore, we choose not to use GANs in our SISR framework. Satellite images, uniquely, possess certain advantageous characteristics for SISR. Specifically, the offset between spectral detectors on satellites ensures each spectral band views the scene from a slightly different perspective, thereby containing additional information, much like in MISR. In our work, we demonstrate the feasibility of applying SISR to recover true high-frequency details in Sentinel-2 multi-spectral images without resorting to GANs, effectively leveraging these distinctive properties of satellite imagery.

1.4 About self-supervised super-resolution

The performance of a deep learning model is heavily dependent on the quality and abundance of the training dataset. In the context of supervised learning for super-resolution algorithms, the necessity for training with realistic data cannot be overstated. As exemplified in the study by Cai et al. [CZY+19], models trained on a dataset comprised of real pairs of low-resolution (LR) and high-resolution (HR) images outperformed those trained

on synthetic data [AT17]. This underscores the importance of using genuine datasets that faithfully represent the nuances of real-world scenarios.

However, the creation and use of real-world, large-scale datasets come with their own set of challenges. Existing datasets for multi-image super-resolution (MISR) such as RBSR [BDVGT21] for smartphone raw burst super-resolution, WorldStrat [COK22] for Sentinel-2 multi-date super-resolution, and the dataset by Martens et al. [MIKC19] for PROBA-V multi-date super-resolution, all necessitate laborious dataset creation and meticulous preprocessing. Ensuring alignment in spatial and spectral content can be particularly demanding and time-consuming.

Having highlighted these considerations, we now turn our attention to self-supervised learning. In contrast to supervised learning, self-supervised learning techniques eliminate the need for ground truth labels, potentially sidestepping some of the issues related to data preprocessing and annotation. In the following sections, we will study some foundational self-supervised methods that underpin our work.

Self-supervised image restoration methods are generally divided into two categories: Intra-image learning methods and inter-image learning methods.

Intra-image learning methods are grounded in the concept of cross-scale internal redundancies and self-similarity inherent within a single image. These methods hinge on the principle that patterns within a single image often exhibit significant resemblance or repetition. Such self-similarity within the image is exploited for tasks like denoising or super-resolution. Traditional denoising methods like Non-Local Means (NLM) [BCM05] and BM3D (Block-Matching and 3D filtering) [DFKE07], as well as example-based super-resolution [FJP02], fall into this category. Recent deep learning approaches such as Noise2Self [BR19] and ZSSR (Zero-Shot Super-Resolution) [SCI18] also belong to this category as they learn from a single noisy or LR image to deduce the clean output.

Inter-image learning methods, on the other hand, utilize two or several observations of the same scene as the target to guide the learning process. These observations could be various degraded versions of the same image or different views of the same scene. In contexts such as burst and video restoration, these observations often contain geometric transformations, but still share a common underlying content. A groundbreaking contribution to this category is the Noise2Noise image denoising method [LMH+18], which learns to denoise an image by comparing two noisy realizations of the same scene. It is important to note that in inter-image learning methods, the reconstruction network must not have access to the degraded target to avoid trivial solutions.

To the best of our knowledge all existing self-supervised super-resolution techniques [YLZ+18, SCI18, BKSI19, KJK20, EPC21, BP22] fall within the realm of intra-image learning. These approaches typically use the input image as the target for the super-resolution of its degraded LR counterpart. However, this approach is not devoid of limitations. The most pressing concern is that the input image itself may contain aliasing, making it less than ideal to be used as the target for super-resolution. Moreover, the LR version of the input image may exhibit different noise, blur, and particularly aliasing patterns, compared to the original input image. This discrepancy could significantly impact the model's learning performance. Additionally, this approach only uses internal information from the LR input, overlooking a substantial amount of external information. This results in their struggle to separate and accurately recover high-frequency details from aliased ones. Therefore, our focus pivots to inter-image learning methods. In the subsequent paragraphs, we study the

two self-supervised denoising methods that inspire our work: Noise2noise [LMH+18] and Frame-to-frame [EDM+19] methods.

In Noise2noise [LMH+18] Lehtinen et al. 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:

$$\mathcal{R}_{\text{N2N}}(\mathbf{Net}) = \sum_j \ell(\mathbf{Net}(N_j), N'_j). \tag{1.6}$$

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 by 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.

While Noise2Noise provides a robust way of denoising still images, its application broadens when considered in the context of video or burst images. Here, a neighboring frame can serve as a noisy target once properly aligned. This concept underpins the frame-to-frame method [EDM+19], which aims to fine-tune a network **Net** to output a single denoised frame $\hat{I}_t$ from a noisy frame t using one or more noisy frames around t.

In this setup, we describe the set of input frames as the input stack $\mathcal{S}_t = [I_{t-m}, ..., I_{t+m}]$, with $2m + 1$ being the total number of input frames. Prior to computing $\hat{I}_r$, the target frame I_r is excluded from the stack S_t. The network training then proceeds by minimizing the frame-to-frame loss:

$$\ell_p^{\text{MF2F}}\left(\mathbf{Net}(\mathcal{S}_{t\setminus r}), I_r\right) = \|O_{t\to r} \circ (\mathbf{Warp}_{t\to r}(\mathbf{Net}(\mathcal{S}_{t\setminus r})) - I_r)\|_p^p. \tag{1.7}$$

In this equation, $p \in \{1, 2\}$, "$\circ$" denotes the element-wise product, $\mathbf{Warp}_{t\to r}$ aligns frame t to r using the estimated motion from r to t, and $O_{t\to r}$ is a misalignment mask removing regions from the loss that are not correctly aligned. Misalignments might occur due to factors such as occlusions, illumination changes, or errors in the optical flow. A core assumption of this framework is that neighboring frames share clean image content subject to a geometric transformation:

$$I_r = \mathbf{Warp}_{t\to r}(\hat{I}_t) + n_r, \tag{1.8}$$

where $\hat{I}_t$ is the desired output frame and n_r models the noise.

The frame-to-frame concept can also extend to other video and burst restoration tasks. For instance, joint burst demosaicing and denoising can be effectively managed by including an additional moisaicking operator in the F2F loss [EDAF19]. Similarly, video denoising tasks can be greatly enhanced by harnessing the potential of the inter-image learning [DAD+21, YPPJ20]. This wide applicability makes it a promising tool in the continuing quest for higher-quality image and video processing.

1.5 A short outline of the thesis

This thesis is divided into two parts, each dedicated to a distinct type of super-resolution: multi-image and single-image. The contributions of this thesis primarily revolve around

the development and application of self-supervised super-resolution techniques for satellite imagery. It starts with the present introductory chapter that presents the motivation for this research and provides a general background in the field.

Part I, titled "Multi-image super-resolution in satellite imagery", begins by evaluating and refining the PROBA-V multi-date super-resolution dataset, resulting in the practical PROBA-V-REF variant. It proceeds to introduce a novel self-supervised multi-image super-resolution approach, adept at processing bursts of satellite images without the need for high-resolution ground truth. To amplify the utility and resilience of this approach, subsequent chapters provide enhancements such as detail-preserving control, outlier detection and a pioneering extension to multi-exposure sequences.

In Part II, "Single-image super-resolution in satellite imagery", the discourse turns to single-image super-resolution (SISR). The section opens with an analysis of the SwinIR method, currently leading in SISR, and its potential applications to satellite imagery. Next, the focus narrows on harnessing Sentinel-2's unique sensor specifications for deep learning-based super-resolution. Lastly, the thesis presents L1BSR, a self-supervised deep learning method developed for SISR and band alignment of Sentinel-2 L1B 10m bands, leveraging overlapping areas in L1B images to eliminate the need for high-resolution ground truth.

Concluding the thesis, a final chapter synthesizes the presented research and considers potential avenues for future explorations, offering a springboard for continued advancement in this promising field.

In the next section, I'll provide a brief overview for each chapter in this thesis.

Chapter 3: Repurposing the Proba-V challenge for reference-aware super-resolution

This chapter takes a fresh look at the 2019 PROBA-V satellite multi-temporal image super-resolution challenge. The PROBA-V dataset comprises series of LR images, each captured on a different date and exhibiting significant variations, and their HR target counterparts for supervised training. Notably, one of the LR images corresponds to the HR image capture time; however, this reference frame's identity is not provided, posing a challenge for the training.

Given the significance of reference-aware multi-image super-resolution, we question the seemingly random assignment of reference frames in this challenge. We posit that this lack of transparency could introduce bias and noise into the benchmarking results. To address this concern, we propose the PROBA-V-REF, a variant of the PROBA-V dataset, that explicitly provides the true LR references. This eliminates the need for heuristic guesswork in the reference selection, thereby offering a more efficient approach to super-resolution (See Fig. 1.2).

This chapter outlines a simple method to identify the LR reference for each training sequence. This is done by aligning the LR frames with the downscaled HR and computing the pixel-wise root-mean-square errors (RMSE) between them. The LR image with the smallest RMSE, indicating the highest similarity to the HR, is chosen as the true reference. The reference selection can be mathematically represented as follows:

$$ref = \underset{t=1,\dots,K}{\arg\min} \|\mathbf{Register}(I_t^{LR}, I^{HR}\downarrow) - I^{HR}\downarrow\|_2 \tag{1.9}$$

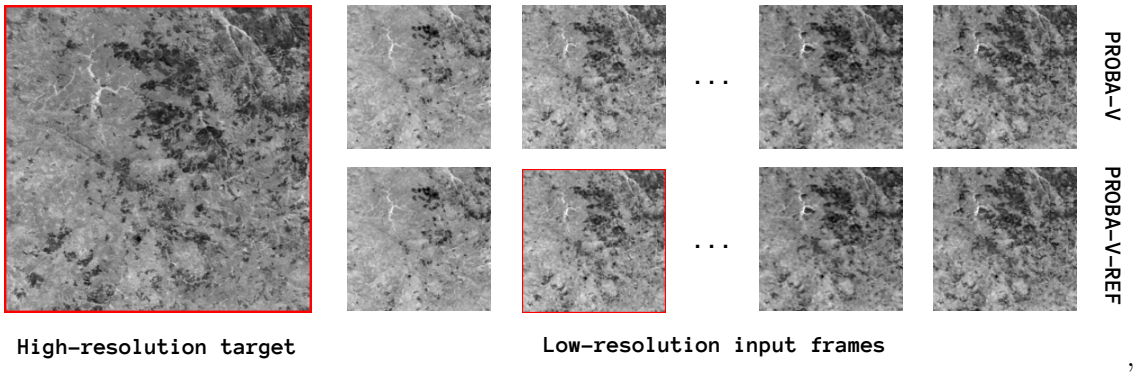


Figure 1.2: The top row presents a LR sequence from the PROBA-V dataset, where the reference frame remains unknown. In contrast, our proposed PROBA-V-REF dataset, shown in the second row, clearly identifies the reference frame, optimizing super-resolution applications.

where $I^{HR}\downarrow$ is the downsampled version of the HR target, and **Register** aligns each LR frame to the downscaled HR.

Interestingly, upon training these methods on the PROBA-V-REF dataset, we observe a reversal in their original ranking. This change underscores the crucial role of the correct reference image selection and its significant impact on method ranking, influenced largely by the reference choice heuristic. Importantly, both quantitative and qualitative evaluations demonstrate that models trained on PROBA-V-REF outperform those trained on PROBA-V, providing a more robust and practical solution for real-world applications.

Chapter 4: Self-supervised multi-image super-resolution for push-frame satellite

In this chapter, we shift our focus to the unique challenges and opportunities offered by MISR from push-frame satellite sensors such as the SkySat constellation from Planet. While these sensors provide an ideal setting for MISR, leveraging their potential is a complex task inasmuch as the ground truth HR targets are not available.

Real large-scale MISR datasets are scarce, with the exception of the PROBA-V dataset we discussed in the previous chapter. However, the PROBA-V images exhibit significant content and illumination changes over time due to their multi-date nature, which makes them unsuitable for training SR algorithms for image bursts captured in quick succession. Therefore, most current burst SR and VSR algorithms end up relying on simulated data, which leads to sub-optimal performance when applied in real-world scenarios.

In light of these challenges, we propose a framework for self-supervised training of MISR networks without requiring high-resolution ground truth images. We also introduce a novel MISR architecture, named Deep Shift-and-Add (DSA), which incorporates a shift-and-add fusion in the feature space. Moreover, DSA is permutation-invariant and capable of handling a variable number of frames.

Within this framework, we randomly select a frame from the input burst S_t to serve as the reference frame I_r^{LR}. We compute the motions between the reference and other frames to perform fusion. The HR output is directly aligned to the reference frame, allowing us

to train the model using a self-supervised loss, defined as follows:

$$\ell^{\mathrm{DSA}}\left(\mathbf{Net}(\mathcal{S}_{t\setminus r}), I_r\right) = \|\mathbf{Net}(\mathcal{S}_{t\setminus r}) \downarrow - I_r^{LR})\|_1. \tag{1.10}$$

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

Figure 1.3: 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).

The efficacy of our self-supervised learning strategy and DSA network is evidenced by experiments on synthetic data, where the results attained compete with those achieved through supervised training. The true potential of this approach is demonstrated in its performance on a public dataset of real image bursts from SkySat satellites. The DSA network effectively reduces noise and adeptly handles degenerate samplings, producing more resolved and less noisy images compared to the L1B product from Planet (See Fig. 1.3).

Chapter 5: Adding detail-preserving control and outlier detection

This chapter extends our exploration of MISR within the context of push-frame satellites. Our primary objective is to enhance the process of robust joint super-resolution and denoising, whilst introducing effective strategies for outlier management.

While super-resolution is usually coupled with denoising, this process can potentially interfere with high-frequency detail retrieval. We address this challenge by considering an additional self-supervised loss function specifically designed for detail recovery:

$$\ell^{\mathrm{LAV}}(\mathbf{Net}(\mathcal{S}_t)) = \frac{1}{K}\sum_{t=1}^{K}\left|\left(\mathbf{Warp}^{\downarrow}(\mathbf{Net}(\mathcal{S}_t), F_{t\to r})\right) - I_t^{LR}\right|, \tag{1.11}$$

Balancing the trade-off between detail-preservation and denoising, we employ a noise-detail map as a network parameter. This map is also incorporated into the total training loss to equilibrate between the denoising effect by ℓ^{DSA} (1.10) and detail-preservation effect by ℓ^{LAV} (1.11).

In order to manage outliers, we deploy a network, termed as MaskNet, which produces a weight mask $\mathcal{O}_t$ for each frame t. These masks are utilized in the weighted fusion module.

From an intuitive perspective, an outlier in any frame can create residual traces post-fusion. Therefore, to minimize the loss, MaskNet should assign smaller weights to these outliers. Fig. 1.4 showcases the efficacy of our outlier detection module in the case of a moving car.

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

Figure 1.4: Super-resolution from a sequence of 15 real low-resolution SkySat L1A frames. (a) L1B from Planet, (b) DSA, (c) Our improvement with an additional CNN to detect the outliers.

Chapter 6: Extension to multi-exposure sequences and improved feature fusion

This chapter elaborates on the extension of the DSA framework for handling multi-exposure sequences, with the aim of jointly performing super-resolution and High Dynamic Range (HDR) processing from a time series of bracketed satellite images.

Addressing this task presents a unique set of challenges: First, the noise in images is signal-dependent, and second, the exposure times reported by small push-frame satellites can be unreliable with errors reaching up to 20%. To tackle these issues, we utilize a noise-level-aware encoder network and adopt a base-detail decomposition strategy, respectively.

The base-detail decomposition provides a robust way to handle inaccuracies in exposure times. The detail components, robust to exposure time errors, is pertinent for super-resolution, while the base components, devoid of high-frequency information, can be up-scaled without the risk from aliasing. The final output is simply the sum of the upsampled base and detail components. This can be briefly summarized by:

$$\begin{aligned} B_t^{LR} &= I_t^{LR} * G, \\ D_t^{LR} &= I_t^{LR} - B_i^{LR}, \quad t = 1, \dots, K \\ B_r^{HR} &= \text{Zoom}(\overline{\{B_t^{LR}\}}) \\ D_r^{HR} &= \mathbf{Net}(\{D_t^{LR}\}) \\ I_r^{HR} &= B_r^{HR} + D_r^{HR} \end{aligned} \tag{1.12}$$

Here G is a Gaussian kernel of standard deviation 1 and $\overline{\{B_t^{LR}\}}$ designates a weighted average of $\{B_t^{LR}\}$ for $t = 1, \dots, K$.

Additionally, we introduce a novel temporal pooling fusion strategy, surpassing the performance of the Feature Shift-and-Add (FS&A) in the DSA. Through quantitative and qualitative experiments (see Fig. 1.5), we demonstrate that our method delivers state-of-the-art performance in satellite MISR.

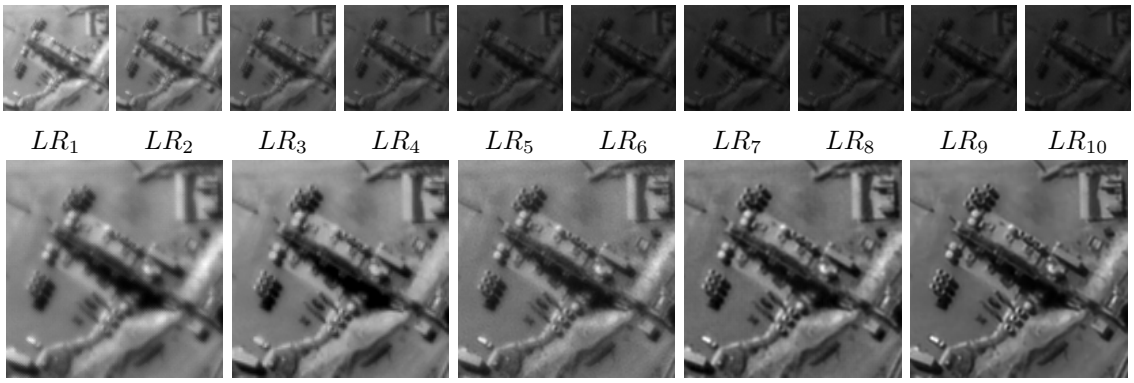


Figure 1.5: Super-resolution from a real multi-exposure sequence of 10 SkySat images. Top row: Bracketed LR sequence. Bottom row: Reconstructions from five methods, including ours trained with self-supervision (right).

Chapter 7: A brief analysis of the SwinIR super-resolution method

In Chapter 7, we take a fresh look at SwinIR, a state-of-the-art image restoration model based on the Swin Transformer architecture. Unlike traditional convolutional neural networks, SwinIR adeptly captures intricate attention patterns among image patches, leading to superior results.

Central to our discussion is SwinIR's proficiency in single image super-resolution. We conduct an insightful examination of the self-attention mechanism's impact, particularly in the context of self-similarity, an important factor in image restoration.

Leveraging the Urban100 dataset, abundant with auto-similar structures, we illuminate the merit of SwinIR's self-attention mechanism in exploiting these similarities to boost super-resolution reconstruction, particularly in low-contrast or aliased areas (Fig. 1.6). However, when applied to satellite imagery, the model's performance diminishes due to domain-specific challenges.

Chapter 8: On the role of alias and band-shift for super-resolution of L1C products

Contrary to conventional SISR methods that depend on perceptual restoration or GANs to tackle the ill-posedness of SISR, this chapter unveils the potential for genuine high-frequency detail recovery in Sentinel-2 imagery. This success is thanks to Sentinel-2's distinct characteristics, specifically aliasing and inter-band shift.

Aliasing arises from a low spatial sampling relative to the instrument's modulation transfer function, while inter-band shifts are attributed to time delays in the acquisition lines of different spectral bands. The synergistic effect of these characteristics transforms the SISR problem into a better-posed scenario (like MISR), paving the way for actual high-frequency information recovery.

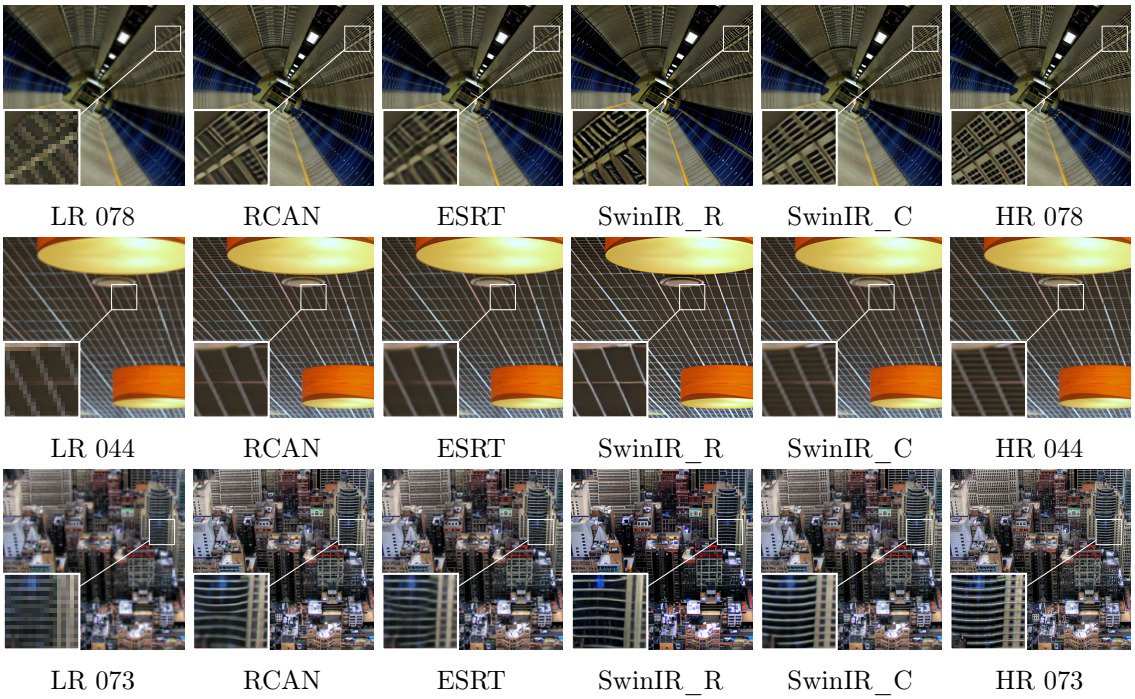


Figure 1.6: Qualitative comparison between the two SwinIR models, RCAN, and ESRT on the Urban100 dataset. Super-resolution by factor of 4.

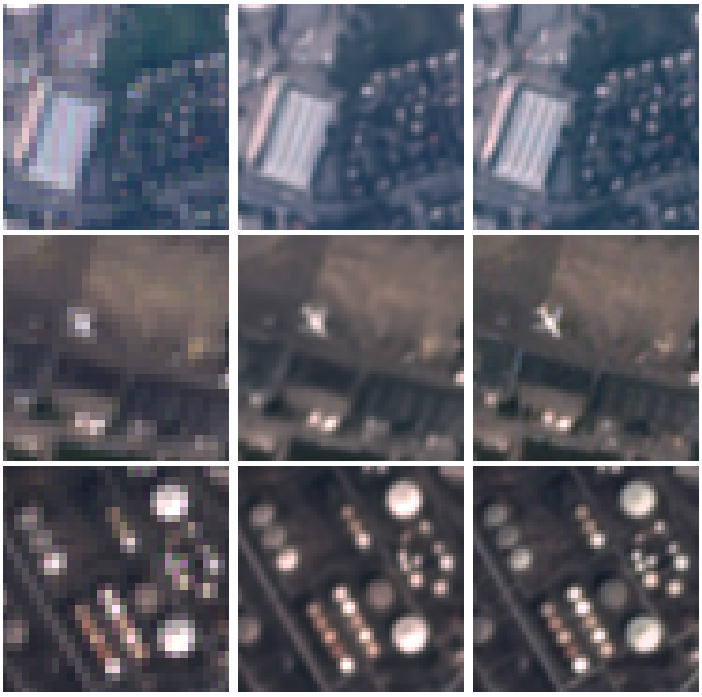


Figure 1.7: SISR results obtained with the L_1 loss. We argue that the characteristic alias and band-shift are key for x2 SR of Sentinel-2 imagery.

Our study is corroborated through experiments involving both synthetic and real data, with the latter utilizing PlanetScope for supervision of Sentinel-2's super-resolution. The

performance on real Sentinel-2 L1C data is shown in Fig. 1.7.

Chapter 9: Exploiting detector overlap for self-supervised super-resolution of L1B products

The previous chapter revealed the potential of SISR for Sentinel-2 Level 1C (L1C) images. This chapter introduces a unique approach for self-supervised joint SISR and band-alignment of Sentinel-2 Level 1B (L1B) products. The focus transitions from L1C products to the raw, early stage L1B products, primarily due to an inherent feature of Sentinel-2's sensor design - detector overlap.

This detector overlap results in overlapping L1B images, which enables us to train our network through self-supervised learning. The strategy is simple yet effective: one overlapping image is used as input and another as the target; the network is tasked to output a band-aligned high-resolution (HR) image such that, after warping and downsampling, it matches the hidden target. Mathematically, this strategy is equivalent to minimizing the self-supervised loss as expressed in the equation:

$$\ell^{\text{L1BSR}}(\mathbf{Net}(I_0^{LR}), I_1^{LR}) = \left\| \mathbf{Warp}^{\downarrow}(\mathbf{Net}(I_0^{LR}), F_{1\to 0}) - I_1^{LR} \right\|_1, \tag{1.13}$$

Nevertheless, L1B products present their own set of challenges: they contain significant misalignment between their spectral bands, which complicates the computation of the motion between the two overlapping images (i.e., $F_{1\to 0}$ in (1.13)). To address this, we introduce a novel cross-spectral registration (CSR) module, which is also trained with self-supervision. It is important to note that the CSR module is instrumental to our framework, but only used during training to guide the reconstruction. At the inference stage, our network directly produces a band-aligned HR output from a band misaligned L1B image. In the experimental section, we conduct comprehensive experiments to demonstrate the

Figure 1.8: L1BSR produces a 5m high-resolution (HR) output with all bands correctly registered from a single 10m low-resolution (LR) Sentinel-2 L1B image with misaligned bands. Note that our method is trained on real data with self-supervision, i.e. without any ground truth HR targets.

effectiveness of our self-supervised framework. We also show that the performance of various modules within our framework are on par with those of the supervised ones, both quantitatively and qualitatively. Fig. 1.8 presents the HR reconstruction by our method on real L1B images.

1.6 List of publications

Presented in this thesis

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. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 3881–3884. IEEE, jul 2021

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1121–1131, 2021

Ngoc Long Nguyen, Jérémy Anger, Axel Davy, Pablo Arias, and Gabriele Facciolo. Self-supervised push-frame super-resolution with detail-preserving control and outlier detection. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 131–134. IEEE, 2022

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1858–1868, 2022

Ngoc-Long Nguyen. A brief analysis of the swinir image super-resolution. *Image Processing On Line*, 12:582–589, 2022

Ngoc Long Nguyen, Jérémy Anger, Lara Raad, Bruno Galerne, and Gabriele Facciolo. On the role of alias and band-shift for sentinel-2 super-resolution. *arXiv preprint arXiv:2302.11494*, 2023

Ngoc Long Nguyen, Jérémy Anger, Axel Davy, Pablo Arias, and Gabriele Facciolo. L1bsr: Exploiting detector overlap for self-supervised single-image super-resolution of sentinel-2 l1b imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2012–2022, 2023

Others

Jamy Lafenetre, Ngoc Long Nguyen, Gabriele Facciolo, and Thomas Eboli. Handheld burst super-resolution meets multi-exposure satellite imagery. *arXiv preprint arXiv:2303.05879*, 2023

2 Introduction (en français)

2.1 Motivation

Les satellites d'Observation de la Terre (OT) jouent un rôle crucial dans la surveillance et la compréhension de la dynamique de notre planète. Ils fournissent des données inestimables pour de nombreuses applications, incluant mais ne se limitant pas à, la prévision météorologique, la gestion des catastrophes, la surveillance environnementale et l'urbanisme [CVTGC+11, LKC15]. Cependant, l'efficacité de ces applications est souvent limitée par la résolution des images satellite [YLXC15]. C'est ici que la super-résolution (SR) entre en jeu.

En tant que technique de traitement d'image, la SR permet de surmonter certaines limitations des images satellites à basse résolution en améliorant les détails et les structures d'image fines qui sont à peine visibles dans les images originales. Cette augmentation de résolution améliore la précision de la détection d'objets, la segmentation, la classification ou le raffinement des processus de cartographie et de détection des changements de couverture terrestre, la résolution améliorée offerte par la SR permet une prise de décision plus informée et précise [SVE19, TAH06].

Alors, pourquoi cette application de la SR est-elle cruciale? La réponse réside dans les défis inhérents auxquels la télédétection est confrontée. Plusieurs facteurs influencent la résolution spatiale des images satellites. La fonction d'étalement du point (ou PSF en anglais) dénote l'effet de flou du système, impacté par des facteurs tels que la diffraction et les aberrations de l'objectif [AdFF19]. Le bruit, provenant de facteurs tels que les erreurs de calibration du système, les capteurs défectueux, ou d'autres types de bruit incluant photonique, thermique, électronique contribue à la dégradation de l'image. Notamment, à la fois la limite de diffraction qui affecte la résolution maximale atteignable et le bruit photonique qui entrave la clarté de l'image dépendent de la taille de l'ouverture. Une ouverture plus grande atténue ces défis en réduisant les effets de la diffraction et en permettant à plus de lumière d'atteindre le capteur, améliorant ainsi le rapport signal-bruit. Néanmoins, les avantages d'une ouverture plus grande ont un coût, car cela augmente considérablement la taille, le poids et le coût global du satellite.

Malgré ces complexités, la demande d'images haute résolution est constamment à la hausse, poussée par le besoin d'informations détaillées et précises dans divers domaines. De l'urbanisme à la surveillance environnementale, les images haute résolution sont essentielles pour des analyses précises. La SR, en améliorant la résolution de ces images, comble ce fossé entre la nécessité et la limitation.

La faisabilité économique de la SR est un autre facteur convaincant. Le déploiement de

nouveaux satellites à haute résolution entraîne une complexité et un coût substantiels. Bien que de tels satellites existent (par exemple GeoEye-1, WorldView-3), l'accès à leurs données est souvent coûteux et peut être limité. À l'inverse, les techniques de SR améliorent la résolution des images des satellites existants (Planet SkySat, Satellogic Aleph-1, Sentinel-2), ce qui en fait une option plus économiquement viable [MSS+14, AEdFF20]. En exploitant les données que nous possédons déjà, la SR optimise le retour sur investissement dans la technologie satellite.

Reconnaissant ces immenses possibilités, cette thèse explore l'application de l'apprentissage profond pour faire avancer la super-résolution des images de satellites OT. En exploitant les capacités de ces méthodes de traitement complexes et efficaces, nous cherchons à améliorer la qualité et le rapport coût-efficacité de la technologie satellite. Le travail présenté dans cette thèse s'efforce d'apporter une contribution précieuse au domaine de la télédétection et au-delà.

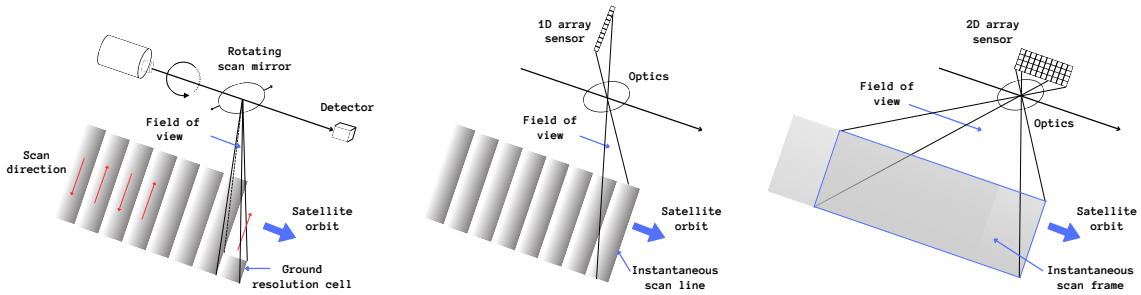
2.2 À propos de l'imagerie satellite optique

L'imagerie satellite optique a considérablement évolué depuis sa création (voir Fig. 2.1), offrant une vue de plus en plus complète des phénomènes terrestres de la Terre.

Dans les années 1970, la télédétection spatiale était dominée par les scanners across-track ou whisk-broom, illustrés par le Return Beam Vidicon (RBV) sur Landsat 1 (1972) et l'Advanced Very High Resolution Radiometer (AVHRR) sur les Polar Operational Environmental Satellites (POES) de NOAA lancés en 1978. Ces systèmes reposaient sur un miroir rotatif mécanique, balayant un pixel à la fois pour assembler l'image ligne par ligne. Cependant, le court temps d'exposition pour chaque pixel a conduit à une capacité de collecte de lumière limitée, et les mouvements continus du miroir ont entraîné une usure mécanique. Ces problèmes ont catalysé une transition vers une technologie d'imagerie push-broom plus efficace à la fin des années 1980.

Le balayage push-broom ou along-track est adopté par des satellites comme la série SPOT française (à partir de 1986), PROBA-V de l'Agence spatiale européenne (2013) et Sentinel-2 (2015), et des satellites commerciaux tels qu'IKONOS (1999), QuickBird (2001), GeoEye-1 (2008), et la série WorldView (à partir de 2007). Cette technique utilise un réseau de détecteurs disposés en ligne, capturant une bande de la surface de la Terre alors que le satellite se déplace le long de sa trajectoire orbitale. L'image résultante présente une qualité améliorée, grâce au temps de séjour accru de la scène sur le capteur. Cependant, la technique push-broom nécessite des systèmes de stabilisation satellite avancés pour prévenir le flou d'image et les distorsions causées par les vibrations de la plateforme ou les fluctuations de l'attitude ou de l'altitude du satellite.

L'avènement des années 2010 a vu l'émergence de l'imagerie push-frame, incarnée par de petits satellites rentables comme le SkySat de Planet (lancé pour la première fois en 2013) et l'Aleph-1 de Satellogic (2016). Ces CubeSats utilisent des capteurs Complementary Metal-Oxide-Semiconductor (CMOS) bidimensionnels, capturant un cadre complet ou un tableau bidimensionnel de pixels à chaque prise, créant une série d'images superposées. Cette superposition permet non seulement une redondance contre les erreurs mineures, mais facilite également des techniques d'imagerie computationnelles avancées telles que le débruitage en rafale et la super-résolution, malgré la nécessité de liaisons de données à large bande et de traitements computationnels complexes.



(a) Whisk-broom (b) Push-broom (c) Push-frame

Figure 2.1: Une représentation des différentes technologies d'imagerie satellite, chacune illustrant le détecteur d'une bande spectrale: du balayage pixel par pixel avec la technologie "whisk-broom" (2.1a), au balayage ligne par ligne avec la technologie "push-broom" (2.1b), jusqu'au balayage image par image avec la technologie "push-frame" (2.1c).

Le traitement de l'imagerie satellite optique contraste fortement avec les appareils photo grand public en raison des défis uniques de l'acquisition de données à l'analyse à l'échelle mondiale. Initialement, une correction des données est requise, une étape où l'imagerie en haute altitude exige la mitigation du bruit du capteur, des distorsions atmosphériques et des désalignements dus à la courbure de la Terre. Ces problèmes sont rarement rencontrés pour les appareils photo grand public. Par la suite, l'orthorectification et le recalage d'images ont lieu. L'orthorectification contrebalance les distorsions de perspective et de terrain pour simuler un point de vue nadir, tandis que le recalage d'images aligne différentes bandes spectrales de la même scène, une étape non nécessaire pour les appareils photo typiques avec des capteurs Bayer. La phase finale implique un post-traitement pour amplifier la qualité de l'image pour une analyse ultérieure. Contrairement à l'amélioration des images grand public qui privilégie souvent l'esthétique, le post-traitement des images satellite est orienté vers l'interprétabilité scientifique.

Cette thèse s'engage avec la super-résolution multi-image (MISR) et la super-résolution single-image (SISR) en utilisant des données provenant de divers satellites, chacun présentant des défis uniques. Nous utilisons le SkySat en mode push-frame pour la SR en rafal, le PROBA-V en mode push-broom pour la SR multi-date, et Sentinel-2 pour la SISR, notant un degré croissant de complexité à travers ces tâches. Alors que la SR en rafale de SkySat bénéficie de plusieurs images en succession rapide, la SR multi-dates de PROBA-V est confrontée à plus de variabilité en raison du laps de temps entre les images. La SISR de Sentinel-2, bien qu'étant généralement un problème mal posé, est légèrement atténuée car chaque bande spectrale visualise la scène sous différents angles avant le recalage, introduisant un composant de type MISR avec un contenu spectral variable. En abordant ces nuances, nous nous efforçons d'améliorer les méthodologies de super-résolution d'images satellites.

2.3 À propos de la super-résolution

Selon le nombre d'images en entrée, les techniques de super-résolution peuvent être largement catégorisées en deux groupes: Super-Résolution Multi-Image (MISR) et Super-Résolution Single-Image (SISR).

2.3.1 Super-résolution multi-image (MISR)

La MISR est une technique qui fusionne des informations provenant de multiples images à basse résolution $I_t^{LR}, t \in [1, \dots, K]$ de la même scène pour produire une sortie à haute résolution I^{HR} [FREM04b]. Ces images à basse résolution peuvent présenter de légères décalages, ou des modifications dans les temps d'exposition, capturant l'objet sous diverses perspectives ou à différents moments.

Le modèle de formation d'image peut être mathématiquement décrit en utilisant une caméra à sténopé [HZ03] avec les processus de transformation géométrique, de flou, de sous-échantillonnage et de dégradation par le bruit [Mil17]:

$$I_t^{LR} = \Pi\left((F_t \circ \mathcal{I}) * k\right) + n_t, \qquad t \in [1, \dots, K], \tag{2.1}$$

où $\mathcal{I}$ désigne l'image idéale à résolution infinie, k est la Point Spread Function (PSF) qui modélise conjointement le flou optique et l'intégration des pixels, F_t représente le mouvement correspondant au cadre t, Π est l'opérateur d'échantillonnage bidimensionnel dû au réseau de capteurs qui introduit l'aliasing, et n_t modélise le bruit de l'image.

Nous pouvons présenter la super-résolution multi-image comme un problème inverse, en utilisant un modèle mathématique pour comprendre et inverser le processus de formation de l'image (2.1). Étant donné un ensemble d'images à basse résolution, aliasées et bruitées, la MISR s'efforce de récupérer une image à haute résolution qui est compatible avec les images à basse résolution observées. Ceci peut être mathématiquement exprimé comme suit:

$$I^{HR} = \arg\min_u \sum_{t=1}^{K} \left\| \mathbf{Warp}^{\downarrow}((u * k), F_t) - I_t^{LR} \right\|_p, \tag{2.2}$$

où $\mathbf{Warp}^{\downarrow}(., F_t)$ déforme et sous-échantillonne une image à haute résolution avec un facteur $z \geq 1$, selon le mouvement F_t

$$\mathbf{Warp}^{\downarrow}(u, F_t)(x) = u(x + zF_t(x)), \tag{2.3}$$

où un schéma d'interpolation est nécessaire car $x + zF_t(x)$ se situe en dehors de la grille de pixels à haute résolution. Cette expression résulte de l'hypothèse que le noyau de flou commute avec le mouvement F_t, ce qui est une approximation raisonnable si le mouvement est approximativement translationnel à l'intérieur du support du filtre. En l'absence de bruit, et avec suffisamment d'images à basse résolution, le problème peut être analysé dans le contexte de la théorie de l'échantillonnage et montré comme bien posé [Tsa84] (en supposant que leur mouvement relatif est connu). Le résultat est une solution stable capable de révéler les vrais détails de la scène. Dans la plupart des cas pratiques cependant, le problème est mal posé, et il peut y avoir de nombreuses reconstitutions à haute fréquence compatibles avec les observations disponibles. Dans ces cas, une certaine priorité ou régularisation est utilisée pour sélectionner une parmi toutes les reconstructions possibles [MO08, ACHR06, FAAC09, PJ07].

Les techniques de MISR peuvent être largement divisées en deux grandes catégories: les méthodes classiques, qui s'appuient principalement sur des modèles mathématiques et l'optimisation, et des approches plus récentes basées sur l'apprentissage profond.

Méthodes classiques: La littérature existante sur les méthodes classiques de super-résolution est vaste, englobant une multitude de stratégies. Celles-ci incluent des approches

dans le domaine fréquentiel [KBV90,NM00,RK99] et des approches dans le domaine spatial [FREM04a, MSS+14, MN07, TMPE09, WGDE+19, TOS92, AEdFF20].

Méthodes Shift-and-Add: Ces travaux construisent une image HR en enregistrant plusieurs images LR et en intégrant leurs informations au niveau des pixels dans la grille HR [KPB88, FREM04a, MSS+14, MN07]. Une fois l'enregistrement terminé, chaque pixel LR est attribué à son voisin HR le plus proche ou "éclaboussé" sur une zone de pixels HR selon certaines méthodes d'interpolation, telles que bilinéaire ou bicubique. Après cela, une dernière étape implique généralement une agrégation pondérée, les poids étant corrélés à l'interpolation utilisée à l'étape précédente.

Pour un facteur d'agrandissement $z \geq 1$, ce processus peut être mathématiquement exprimé comme suit:

$$\begin{aligned} \mathbf{Warp}^{\uparrow}(I_t^{LR}, F_t) &= \sum_x I_t^{LR}(x)\delta(x + zF_t(x)), \\ J_t^{HR} &= \mathbf{Warp}^{\uparrow}(I_t^{LR}, F_t) * \mathcal{K}, \\ W_t^{HR} &= \mathbf{Warp}^{\uparrow}(\mathbf{1}^{LR}, F_t) * \mathcal{K}, \\ I^{HR} &= \frac{\sum_{t=1}^{K} J_t^{HR}}{\sum_{t=1}^{K} W_t^{HR}}. \end{aligned} \tag{2.4}$$

Ici $\mathbf{Warp}^{\uparrow}(I_t^{LR}, F_t)$ enregistre l'image I_t^{LR} sur la grille HR en utilisant le mouvement F_t. Comme ce mouvement peut être sous-pixel, cette image déformée est représentée comme un peigne de Dirac irrégulier. $\mathcal{K}$ désigne le noyau d'interpolation dans l'espace HR, qui "éclabousse" les pixels de Dirac sur la grille HR entière, et W_t^{HR} sont les poids d'agrégation utilisés pour la normalisation, calculés en "éclaboussant" une image de uns $\mathbf{1}^{LR}$. La dernière division est censée être élément par élément.

Dans [FREM04c] les auteurs soulignent que l'algorithme shift-and-add peut être vu comme la solution d'un problème de moindres carrés pondérés, où les poids sont donnés par le noyau d'éclaboussement $\mathcal{K}$. Sur la base de cette observation, ils dérivent une version robuste qui gère les valeurs aberrantes en minimisant une somme d'erreurs absolues. Cela aboutit à un algorithme qui calcule les médianes pondérées à travers les pixels [FREM04c]. Notez que ces méthodes directes peuvent laisser des trous dans la sortie si les échantillons sont insuffisants ou dégénérés [KPB88]. Pour faire face aux éventuels écarts et valeurs aberrantes, des régularisateurs basés sur la Variation Totale (TV) sont souvent incorporés dans une étape de post-traitement de minimisation de l'énergie [FREM04a]. Il faut noter que le résultat HR obtenu à ce stade est intrinsèquement flou en raison de l'impact de la Fonction de Point d'Étalement (PSF) et de l'intégration de pixels (dénommée k) [MSS+14]. Pour rectifier cela, une dernière étape de défloutage est régulièrement effectuée. Cela peut être réalisé en résolvant le problème d'optimisation suivant:

$$I_{\text{sharp}}^{HR} = \arg\min_u \left\| u * k - I^{HR} \right\|_p + \mathcal{R}(u), \tag{2.5}$$

où $\mathcal{R}$ est une fonction de régularisation, qui peut inclure la Variation Totale (TV), la régularisation de Tikhonov, ou une combinaison des deux. La norme p est généralement fixée à 1 ou 2, s'adaptant aux caractéristiques spécifiques du flou et au niveau de netteté souhaité dans la sortie.

Les techniques Shift-and-Add sont parmi les plus anciennes techniques de super-résolution, avec des résultats qui sont loin de l'état actuel de l'art. Pourtant, elles restent pertinentes

en raison de leur simplicité et de leur interprétabilité. Ces caractéristiques, associées au fait qu'il s'agit d'une opération différentiable, permettent leur intégration transparente dans les modèles d'apprentissage profond, ce que nous exploitons dans notre travail.

Méthodes d'apprentissage profond: Le domaine du MISR a connu une transformation significative avec l'avènement des techniques d'apprentissage profond. Ces nouvelles méthodes, émanant de diverses sources telles que la débruisation vidéo et en rafale, le défloutage et la super-résolution, présentent un grand potentiel [TDV20, MBC+18, SDW+17, TGL+17, SVB18, CWY+21].

Les méthodes d'apprentissage profond pour le MISR se divisent en deux grandes catégories. Le premier groupe a tendance à favoriser le développement d'architectures complexes et de nouvelles méthodologies, sacrifiant souvent l'interprétabilité dans le processus [SCH+16, KLNK18, JWOKJK18, LLTMK19a, DKG+20, LHD+19, MVFM19, SMKC20, AMSC+20]. Ces techniques traitent généralement une pile d'images à basse résolution en entrée et génèrent une sortie à haute résolution, sans estimation explicite du mouvement [SCH+16, JWOKJK18, KLNK18, LLTMK19a, LHD+19, DKG+20, AMSC+20, IJG+20a] ou d'autres techniques traditionnelles [MVFM19, SMKC20]. Malgré leur efficacité prouvée, le manque d'interprétabilité rend difficile de comprendre pourquoi elles fonctionnent et comment les adapter pour des cas d'utilisation spécifiques [CWY+21].

La seconde catégorie de méthodes intègre des opérations plus transparentes et explicites, offrant une meilleure compréhension de leur approche [TGL+17, HSU19, BRE19, KBP+19, LPM21, CLK21, LPME22, SJC+22]. Bien que ces techniques soient plus complexes et difficiles à développer, elles donnent souvent de meilleurs résultats en raison de leur incorporation explicite de principes bien compris tels que le shift-and-add [TGL+17, KBP+19], plug-and-play [VBW13, BRE19, LPM21, LPME22], la transformation en ondelettes [CLK21], ou le transport optimal [PC+19, SJC+22]. Ces techniques nécessitent généralement une estimation explicite du mouvement, une tâche informatiquement intensive. Cependant, dans le contexte de l'imagerie par satellite, la perspective en haute altitude et la simplicité de la dynamique de la scène rendent souvent cette tâche plus gérable.

Notre travail s'inscrit dans la seconde catégorie, avec pour objectif de combiner les réseaux neuronaux avancés avec l'interprétabilité et la différentiabilité de la méthode traditionnelle Shift-and-Add. Cette combinaison nous permet de tirer parti des avantages des techniques modernes d'apprentissage profond tout en préservant la simplicité et la robustesse des méthodes classiques [FREM04c].

Le type de perte est un élément clé dans l'entraînement des méthodes MISR. Les méthodes d'apprentissage profond pour le MISR s'appuient principalement sur des fonctions de perte basées sur les pixels comme l'erreur quadratique moyenne (MSE), la perte L1, et la perte de Charbonnier pour l'optimisation [KYDK16, TGL+17, SVB18, CBFAB97, MVFM19, DKG+20, CWY+21, LCF+22, ZGFK16]. Ces fonctions objectif guident efficacement l'apprentissage des reconstructions haute résolution mais leurs tendances intrinsèques à la moyenne peuvent entraîner un effet de flou. Cependant, le degré de ce flou dépend fortement de la présence d'aliasing dans les images LR: lorsque les images LR sont fortement aliasé, l'information haute fréquence peut être mieux conservée et ainsi une image HR plus nette peut être récupérée. Dans des scénarios où l'aliasing est minimal et les détails haute fréquence ne sont pas bien conservés, une combinaison de perte de réseau antagoniste génératif (GAN) et de perte perceptive peut être utilisée pour générer des résultats plus visuellement agréables et détaillés [CXLT18, LLTMK19a]. Cependant, il faut

garder à l'esprit que l'utilisation de ces pertes peut conduire à l'hallucination de textures et de détails qui ne sont pas présents dans la vérité terrain, impactant négativement la reconstruction des vrais détails en raison du compromis perception-distorsion [BM18].

2.3.2 super-résolution single-image (SISR)

La SISR est intrinsèquement un problème mal posé car il peut y avoir plusieurs images HR qui correspondent à une seule image LR. L'objectif principal en SISR est de générer des détails haute fréquence qui sont plausibles et convaincants pour les observateurs humains. La SISR est largement utilisée dans divers domaines, tels que l'imagerie médicale [Gre09], la surveillance [ZZSL10], et la télédétection [HFBP+18]. Dans la section suivante, nous donnons un bref aperçu des approches classiques de SISR et discutons de la transition récente vers les méthodologies de SISR basées sur l'apprentissage profond.

Méthodes classiques: Les méthodes classiques de SISR peuvent être largement classées en quatre groupes principaux: les méthodes basées sur l'interpolation, les méthodes basées sur des exemples, les méthodes basées sur une représentation sparse, et les méthodes variationnelles. Ces techniques sont ancrées dans le même modèle mathématique que le MISR (eq. (2.1) avec $K = 1$), mais comme elles s'appuient sur une seule image, des a priori sont indispensables pour atténuer le caractère mal posé du problème.

Les techniques basées sur l'interpolation, telles que l'interpolation bilinéaire et bicubique, sont simples dans leur approche, utilisant des fonctions mathématiques prédéfinies pour combler les lacunes entre les pixels [LGZ13]. Les méthodes basées sur des exemples, en revanche, améliorent la qualité de l'image en identifiant des patches similaires dans une base de données ou l'image elle-même, puis en les utilisant pour améliorer la résolution [FJP02, GBI09, HSA15]. Prenant une approche différente, les méthodes basées sur une représentation sparse fonctionnent sur la présomption que les patches d'images peuvent être exprimés comme des combinaisons linéaires sparse d'éléments provenant d'un dictionnaire surcomplet [YWHM10, ZEP12, MBP+09]. Les méthodes variationnelles présentent le SISR comme un problème d'optimisation, appliquant l'inférence variationnelle pour trouver la solution la plus plausible [MO08, UPWB10, CDL18].

Nonobstant la panoplie de méthodes disponibles, chacune a ses propres limites. La qualité globale des résultats peut être significativement entravée par les a priori imposés, et les vrais détails haute fréquence sont souvent difficiles, voire impossibles, à restaurer avec précision à partir de ceux aliasé.

Méthodes d'apprentissage profond: L'état de l'art en SISR est largement dominé par les méthodes basées sur l'apprentissage profond. Ces approches exploitent la puissance des réseaux neuronaux, formés sur des jeux de données volumineux, pour modéliser le mappage complexe des images LR vers HR. Contrairement aux méthodes traditionnelles, qui reposent souvent sur des caractéristiques et des a priori fabriqués à la main, les méthodes d'apprentissage profond sont capables d'apprendre automatiquement des caractéristiques hiérarchiques à partir de données, les rendant plus flexibles et puissantes pour capturer des modèles et des structures d'images complexes.

L'évolution architecturale des réseaux SISR a été rapide et transformative. À partir du simple modèle SRCNN à trois couches [DLHT14], nous avons vu l'adoption de structures avancées qui améliorent la performance du modèle et l'efficacité de la formation.

L'apprentissage résiduel a été une innovation marquante, contournant le problème du gradient évanescent pour permettre la formation de réseaux beaucoup plus profonds [HZRS16, KKLML16, LSK+17]. Un autre développement significatif a été l'introduction de connexions denses, facilitant l'extraction de caractéristiques hiérarchiques plus riches et un flux d'information plus efficace à travers les couches du réseau [HLVDMW17, TLLG17, ZTK+18]. Plus particulièrement, l'avènement des GAN [LTH+17, WYW+19] et des architectures de transformer [YYF+20, LCS+21] a marqué un tournant, offrant des capacités uniques comme la génération d'images perceptuellement agréables et la capacité de modéliser des dépendances à longue distance, respectivement.

En termes de pertes de formation, la qualité des résultats de super-résolution est influencée par la fonction objective choisie. Alors que des fonctions de perte traditionnelles comme L_1, L_2, et la perte de Charbonnier sont souvent utilisées, l'introduction des GAN a fondamentalement modifié l'approche en SISR. Le cadre adversarial des GAN, avec un générateur produisant des images HR et un discriminateur discernant entre les images réelles et générées, est devenu l'approche dominante en SISR. Ce paradigme, illustré par des modèles comme SRGAN [LTH+17] et ESRGAN [WYW+19], met l'accent sur la génération d'images HR qui sont perceptuellement plus proches des images réelles, se concentrant moins sur la précision au niveau des pixels et plus sur une qualité visuelle supérieure. Cette approche est particulièrement pertinente et efficace pour la plupart des cas d'utilisation.

Cependant, en super-résolution d'images satellites, l'objectif principal est généralement de restaurer les vrais détails haute fréquence plutôt que d'obtenir des résultats visuellement plaisants. Par conséquent, malgré le succès généralisé des GAN en super-résolution d'images, leur application n'est pas aussi appropriée pour notre contexte spécifique. Nous choisissons donc de ne pas utiliser de GAN dans notre cadre SISR. Les images satellites possèdent certaines caractéristiques avantageuses pour le SISR. Plus précisément, le décalage entre les détecteurs spectraux sur les satellites garantit que chaque bande spectrale visualise la scène sous un angle légèrement différent, contenant ainsi des informations supplémentaires, à l'instar du MISR. Dans notre travail, nous démontrons la faisabilité d'appliquer le SISR pour récupérer de vrais détails haute fréquence dans les images multispectrales Sentinel-2 sans recourir aux GAN, exploitant efficacement ces propriétés distinctives de l'imagerie satellite.

2.4 À propos de la super-résolution auto-supervisée

La performance d'un modèle d'apprentissage profond dépend fortement de la qualité et de l'abondance de l'ensemble de données d'entraînement. Dans le contexte de l'apprentissage supervisé pour les algorithmes de super-résolution, la nécessité de s'entraîner avec des données réalistes ne saurait être trop soulignée. Comme l'a montré l'étude de Cai et al. [CZY+19], les modèles formés sur un jeu de données composé de vraies paires d'images à basse résolution (LR) et haute résolution (HR) ont surpassé ceux formés sur des données synthétiques [AT17]. Cela souligne l'importance d'utiliser de véritables ensembles de données qui représentent fidèlement les nuances des scénarios du monde réel.

Cependant, la création et l'utilisation de jeux de données à grande échelle du monde réel comportent leur propre ensemble de défis. Les jeux de données existants pour la super-résolution multi-image (MISR) tels que RBSR [BDVGT21] pour la super-résolution de rafales brutes de smartphones, WorldStrat [COK22] pour la super-résolution multi-dates

de Sentinel-2, et le jeu de données de Martens et al. [MIKC19] pour la super-résolution multi-dates de PROBA-V, nécessitent tous une création de jeu de données laborieuse et un prétraitement méticuleux. Assurer l'alignement du contenu spatial et spectral peut être particulièrement exigeant et chronophage.

Après avoir souligné ces considérations, nous nous tournons maintenant vers l'apprentissage auto-supervisé. Contrairement à l'apprentissage supervisé, les techniques d'apprentissage auto-supervisé éliminent le besoin d'étiquettes de vérité terrain, contournant potentiellement certains des problèmes liés au prétraitement et à l'annotation des données. Dans les sections suivantes, nous étudierons certaines méthodes auto-supervisées fondamentales qui sous-tendent notre travail.

Les méthodes de restauration d'images auto-supervisées sont généralement divisées en deux catégories: les méthodes d'apprentissage intra-image et les méthodes d'apprentissage inter-image.

Les méthodes d'apprentissage intra-image sont basées sur le concept de redondances internes à l'échelle croisée et l'auto-similarité inhérente à une seule image. Ces méthodes reposent sur le principe que les motifs dans une seule image présentent souvent une ressemblance ou une répétition significative. Cette auto-similarité au sein de l'image est exploitée pour des tâches comme le débruitage ou la super-résolution. Des méthodes de débruitage traditionnelles comme les moyennes non locales (NLM) [BCM05] et BM3D (Bloc-Matching et filtrage 3D) [DFKE07], ainsi que la super-résolution basée sur des exemples [FJP02], entrent dans cette catégorie. Des approches récentes d'apprentissage profond comme Noise2Self [BR19] et ZSSR (Zero-Shot Super-Resolution) [SCI18] appartiennent également à cette catégorie car elles apprennent à partir d'une seule image bruyante ou LR pour déduire la sortie propre.

D'autre part, les méthodes d'apprentissage inter-image utilisent deux ou plusieurs observations de la même scène comme cible pour guider le processus d'apprentissage. Ces observations pourraient être diverses versions dégradées de la même image ou différentes vues de la même scène. Dans des contextes tels que la restauration de rafales et de vidéos, ces observations contiennent souvent des transformations géométriques, mais partagent toujours un contenu sous-jacent commun. Une contribution révolutionnaire à cette catégorie est la méthode de débruitage d'image Noise2Noise [LMH+18], qui apprend à débruiter une image en comparant deux réalisations bruyantes de la même scène. Il est important de noter que dans les méthodes d'apprentissage inter-image, le réseau de reconstruction ne doit pas avoir accès à la cible dégradée pour éviter des solutions triviales.

À notre connaissance, toutes les techniques existantes de super-résolution auto-supervisée [YLZ+18, SCI18, BKSI19, KJK20, EPC21, BP22] relèvent de l'apprentissage intra-image. Ces approches utilisent généralement l'image d'entrée comme cible pour la super-résolution de son équivalent LR dégradé. Cependant, cette approche n'est pas dénuée de limites. La préoccupation la plus pressante est que l'image d'entrée elle-même peut contenir de l'aliasing, ce qui la rend moins qu'idéale pour être utilisée comme cible pour la super-résolution. De plus, la version LR de l'image d'entrée peut présenter différents bruits, flous, et en particulier des motifs d'aliasing, par rapport à l'image d'entrée originale. Cette différence pourrait avoir un impact significatif sur les performances d'apprentissage du modèle. De plus, cette approche n'utilise que des informations internes de l'entrée LR, négligeant une grande quantité d'informations externes. Cela se traduit par leur difficulté à séparer et à récupérer avec précision les détails de haute fréquence des aliasés. Par conséquent, notre attention se tourne vers les méthodes d'apprentissage inter-image. Dans les paragraphes

suivants, nous étudions les deux méthodes de débruitage auto-supervisées qui inspirent notre travail: Noise2noise [LMH+18] et les méthodes de trame à trame [EDM+19].

Dans Noise2noise [LMH+18], Lehtinen et al. ont montré qu'un réseau de débruitage d'images peut être formé à partir de paires de versions bruyantes N et N' de la même image I avec des réalisations de bruit indépendantes, en minimisant le risque suivant de bruit à bruit (N2N):

$$\mathcal{R}_{\text{N2N}}(\mathbf{Net}) = \sum_j \ell(\mathbf{Net}(N_j), N'_j). \tag{2.6}$$

Intuitivement, puisque les réalisations du bruit sont indépendantes, le bruit dans N' ne peut être prédit à partir de N. Par conséquent, la perte est minimisée en estimant l'image propre. Les estimateurs optimaux pour le risque N2N sont donnés par $\mathbb{E}N'|N$ pour la perte MSE, et médiane$N'|N$ pour la perte L_1. Il peut être montré que si le bruit dans N' préserve la moyenne, alors $\mathbb{E}N'|N = \mathbb{E}I|N$, c'est-à-dire *l'entraînement avec la supervision par les images bruyantes est équivalent à celui supervisé par les images propres.* Il a également été observé empiriquement qu'une propriété similaire s'applique à la perte L_1 si le bruit dans N' préserve la médiane.

Bien que Noise2Noise offre une manière robuste de débruiter les images fixes, son application se généralise lorsqu'on la considère dans le contexte de vidéos ou d'images en rafale. Ici, une trame voisine peut servir de cible bruyante une fois correctement alignée. Ce concept sous-tend la méthode de trame à trame [EDM+19], qui vise à affiner un réseau **Net** pour produire une seule trame débruitée $\hat{I}_t$ à partir d'une trame bruyante t en utilisant une ou plusieurs trames bruyantes autour de t.

Dans cette configuration, nous décrivons l'ensemble des trames d'entrée comme la pile d'entrée $\mathcal{S}_t = [I_{t-m}, ..., I_{t+m}]$, avec $2m+1$ étant le nombre total de trames d'entrée. Avant de calculer $\hat{I}_r$, la trame cible I_r est exclue de la pile S_t. L'entraînement du réseau se poursuit alors en minimisant la perte de trame à trame:

$$\ell_p^{\text{MF2F}}\left(\mathbf{Net}(\mathcal{S}_{t\setminus r}), I_r\right) = \|O_{t\to r} \circ (\mathbf{Warp}_{t\to r}(\mathbf{Net}(\mathcal{S}_{t\setminus r})) - I_r)\|_p^p. \tag{2.7}$$

Dans cette équation, $p \in 1, 2$, " $\circ$ " désigne le produit terme à terme, $\mathbf{Warp}t \to r$ aligne la trame t sur r en utilisant le mouvement estimé de r à t, et $Ot \to r$ est un masque de désalignement qui élimine les régions de la perte qui ne sont pas correctement alignées. Des désalignements peuvent se produire en raison de facteurs tels que les occlusions, les changements d'éclairage ou les erreurs dans le flux optique. Une hypothèse fondamentale de ce cadre est que les trames voisines partagent le contenu de l'image propre soumis à une transformation géométrique:

$$I_r = \mathbf{Warp}t \to r(\hat{I}_t) + nr, \tag{2.8}$$

où $\hat{I}_t$ est la trame de sortie souhaitée et n_r modélise le bruit.

Le concept de trame à trame peut également être étendu à d'autres tâches de restauration de vidéos et d'images en rafale. Par exemple, le démosaïcage et le débruitage de rafales conjointes peuvent être efficacement gérés en incluant un opérateur de mosaïque supplémentaire dans la perte F2F [EDAF19]. De même, les tâches de débruitage de vidéos peuvent être grandement améliorées en exploitant le potentiel de l'apprentissage inter-image [DAD+21, YPPJ20]. Cette large applicabilité en fait un outil prometteur dans la quête continue d'une meilleure qualité de traitement des images et des vidéos.

2.5 Un bref aperçu de la thèse

Cette thèse est divisée en deux parties, chacune consacrée à un type distinct de super-résolution: multi-image et mono-image. Les contributions de cette thèse tournent principalement autour du développement et de l'application de techniques de super-résolution auto-supervisées pour l'imagerie satellite. Elle commence par le présent chapitre d'introduction qui présente la motivation de cette recherche et fournit un contexte général dans le domaine.

La première partie, intitulée "Super-résolution multi-image de l'imagerie satellite", commence par évaluer et affiner l'ensemble de données de super-résolution multi-dates PROBA-V, aboutissant à la variante pratique PROBA-V-REF. Elle continue à introduire une nouvelle approche de super-résolution multi-image auto-supervisée, capable de traiter des rafales d'images satellites sans avoir besoin de vérité terrain en haute résolution. Pour amplifier l'utilité et la résilience de cette approche, les chapitres suivants proposent des améliorations telles que le contrôle préservant les détails, la détection des valeurs aberrantes et une extension pionnière aux séquences à exposition multiple.

Dans la deuxième partie, intitulée "super-résolution single-image de l'imagerie satellite", le discours se tourne vers la super-résolution single-image (SISR). Cette section commence par une analyse de la méthode SwinIR, actuellement leader en SISR, et de ses applications potentielles à l'imagerie satellite. Ensuite, l'accent est mis sur l'exploitation des spécifications uniques du capteur Sentinel-2 pour la super-résolution basée sur l'apprentissage profond. Enfin, la thèse présente L1BSR, une méthode d'apprentissage profond auto-supervisée développée pour le SISR et l'alignement de bandes de Sentinel-2 L1B 10m, en exploitant les zones de chevauchement dans les images L1B pour éliminer le besoin d'une vérité terrain en haute résolution.

En conclusion de la thèse, un dernier chapitre synthétise la recherche présentée et envisage des voies potentielles pour de futures explorations, offrant un tremplin pour une avancée continue dans ce domaine prometteur.

Dans la section suivante, je fournirai un bref aperçu de chaque chapitre de cette thèse.

Chapitre 3: Réutilisation du défi Proba-V pour la super-résolution avec référence

Ce chapitre jette un regard neuf sur le défi de super-résolution d'images multi-temporelles du satellite PROBA-V de 2019. L'ensemble de données PROBA-V comprend des séries d'images LR, chacune capturée à une date différente et présentant des variations significatives, et leurs homologues HR cibles pour l'entraînement supervisé. Notamment, une des images LR correspond au moment de la capture de l'image HR ; cependant, l'identité de cette image de référence n'est pas fournie, ce qui pose un défi pour l'entraînement.

Étant donné l'importance de la super-résolution multi-image avec référence, nous remettons en question l'attribution apparemment aléatoire des images de référence dans ce défi. Nous supposons que ce manque de transparence pourrait introduire un biais et du bruit dans les résultats de l'évaluation. Pour répondre à cette préoccupation, nous proposons le PROBA-V-REF, une variante de l'ensemble de données PROBA-V, qui fournit explicitement les véritables références LR. Cela élimine le besoin de devinettes heuristiques dans la sélection des références, offrant ainsi une approche plus efficace pour la super-résolution (voir Fig. 2.2).

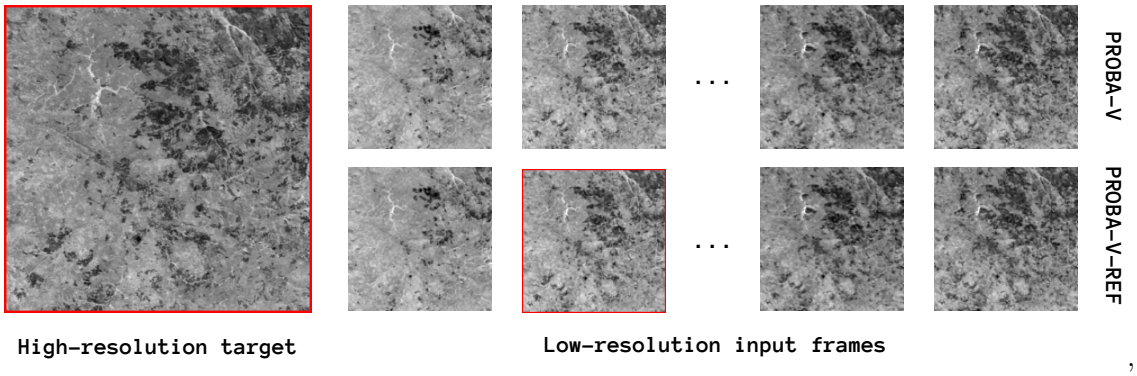


Figure 2.2: La rangée du haut présente une séquence LR du jeu de données PROBA-V, où le cadre de référence reste inconnu. En revanche, notre jeu de données proposé PROBA-V-REF, présenté dans la deuxième rangée, identifie clairement le cadre de référence, optimisant les applications de super-résolution.

Ce chapitre décrit une méthode simple pour identifier la référence LR de chaque séquence d'entraînement. Ceci est réalisé en alignant les images LR avec la version réduite de l'image HR et en calculant les erreurs quadratiques moyennes pixel par pixel entre elles. L'image LR avec la plus petite erreur quadratique moyenne, indiquant la plus grande similarité avec l'image HR, est choisie comme référence véritable. La sélection de la référence peut être représentée mathématiquement comme suit:

$$ref = \underset{t=1,\dots,K}{\arg\min} \|\mathbf{Register}(I_t^{LR}, I^{HR}\downarrow) - I^{HR}\downarrow\|_2 \tag{2.9}$$

où $I^{HR}\downarrow$ est la version sous-échantillonnée de la cible HR, et **Register** aligne chaque image LR sur la HR sous-échantillonnée.

De façon intéressante, lors de l'entraînement de ces méthodes sur le jeu de données PROBA-V-REF, nous observons une inversion dans leur classement original. Ce changement souligne le rôle crucial de la bonne sélection de l'image de référence et son impact significatif sur le classement des méthodes, largement influencé par l'heuristique de choix de la référence. Il est important de noter que tant les évaluations quantitatives que qualitatives montrent que les modèles formés sur PROBA-V-REF surpassent ceux formés sur PROBA-V, offrant une solution plus robuste et pratique pour des applications en conditions réelles.

Chapitre 4: Super-résolution multi-image auto-supervisée pour satellite push-frame

Dans ce chapitre, nous orientons notre attention vers les défis et opportunités uniques offerts par la MISR à partir de capteurs satellites push-frame tels que la constellation SkySat de Planet. Bien que ces capteurs fournissent un cadre idéal pour la MISR, exploiter leur potentiel est une tâche complexe dans la mesure où les cibles HR de vérité terrain ne sont pas disponibles.

Les ensembles de données MISR à grande échelle réels sont rares, à l'exception de l'ensemble de données PROBA-V dont nous avons discuté dans le chapitre précédent. Cependant, les images PROBA-V présentent des variations significatives de contenu et d'éclairage au

fil du temps en raison de leur nature multi-dates, ce qui les rend inadaptées à la formation d'algorithmes SR pour des rafales d'images capturées en succession rapide. Par conséquent, la plupart des algorithmes actuels de SR et VSR en rafale finissent par se reposer sur des données simulées, ce qui conduit à des performances sous-optimales lorsqu'ils sont appliqués dans des scénarios réels.

Compte tenu de ces défis, nous proposons un cadre pour la formation auto-supervisée de réseaux MISR sans nécessiter d'images de vérité terrain à haute résolution. Nous introduisons également une nouvelle architecture MISR, nommée Deep Shift-and-Add (DSA), qui incorpore une fusion shift-and-add dans l'espace des caractéristiques. De plus, DSA est invariante à la permutation et capable de gérer un nombre variable de cadres.

Dans ce cadre, nous sélectionnons aléatoirement un cadre de la rafale d'entrée S_t pour servir de cadre de référence I_r^{LR}. Nous calculons les mouvements entre la référence et les autres cadres pour effectuer la fusion. La sortie HR est directement alignée sur le cadre de référence, ce qui nous permet de former le modèle à l'aide d'une perte auto-supervisée, définie comme suit:

$$\ell^{\text{DSA}}\left(\mathbf{Net}(\mathcal{S}_{t\setminus r}), I_r\right) = \|\mathbf{Net}(\mathcal{S}_{t\setminus r}) \downarrow - I_r^{LR})\|_1. \tag{2.10}$$

(a) L1A (b) L1B (c) DSA

Figure 2.3: Super-résolution à partir d'une séquence de 15 trames SkySat L1A de basse résolution. (a) Trame L1A, (b) Planet L1B product (×1.25), (c) Méthode proposé (×2).

L'efficacité de notre stratégie d'apprentissage auto-supervisée et du réseau DSA est mise en évidence par des expériences sur des données synthétiques, où les résultats obtenus sont en concurrence avec ceux obtenus par une formation supervisée. Le véritable potentiel de cette approche est démontré par sa performance sur un ensemble de données publiques de vraies rafales d'images provenant de satellites SkySat. Le réseau DSA réduit efficacement le bruit et gère habilement les échantillonnages dégénérés, produisant des images plus résolues et moins bruitées par rapport au produit L1B de Planet (Voir Fig. 2.3).

Chapitre 5: Ajout de contrôle de préservation des détails et détection d'outliers

Ce chapitre prolonge notre exploration de la MISR dans le contexte des satellites push-frame. Notre objectif principal est d'améliorer le processus de super-résolution et de

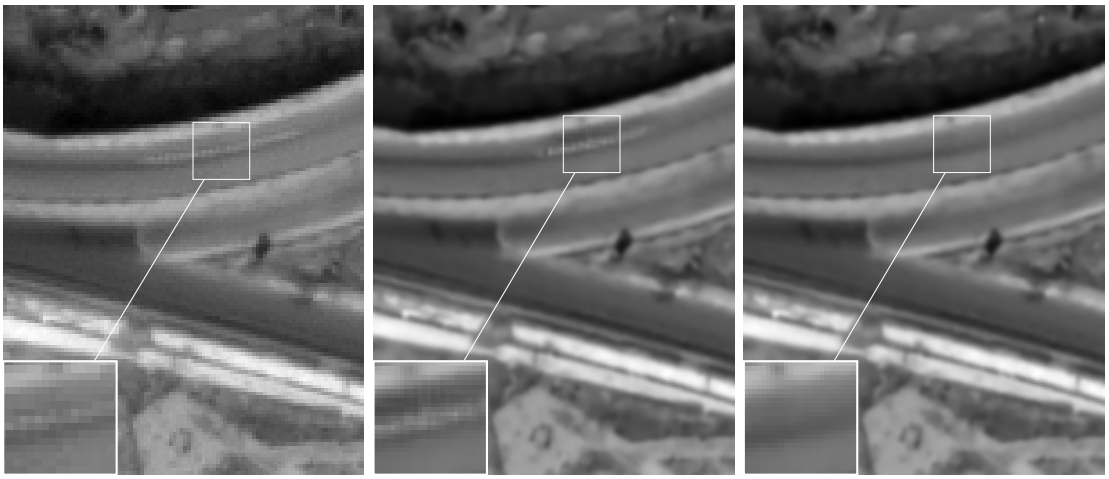
débruitage conjoints robustes, tout en introduisant des stratégies efficaces pour la gestion des outliers.

Bien que la super-résolution soit généralement couplée au débruitage, ce processus peut potentiellement interférer avec la récupération des détails à haute fréquence. Nous relevons ce défi en considérant une fonction de perte auto-supervisée supplémentaire spécifiquement conçue pour la récupération des détails:

$$\ell^{\mathrm{LAV}}(\mathbf{Net}(\mathcal{S}_t)) = \frac{1}{K}\sum_{t=1}^{K}\left|\left(\mathbf{Warp}^{\downarrow}(\mathbf{Net}(\mathcal{S}_t), F_{t\to r})\right) - I_t^{LR}\right|, \tag{2.11}$$

Pour équilibrer le compromis entre la préservation des détails et le débruitage, nous employons une carte de détail-bruit comme paramètre de réseau. Cette carte est également incorporée à la perte totale d'entraînement pour équilibrer entre l'effet de débruitage par ℓ^{DSA}(2.10) et l'effet de préservation des détails par ℓ^{LAV}(2.11).

Afin de gérer les outliers, nous déployons un réseau, appelé MaskNet, qui produit un masque de poids $\mathcal{O}_t$ pour chaque cadre t. Ces masques sont utilisés dans le module de fusion pondérée. D'un point de vue intuitif, un outlier dans n'importe quel cadre peut créer des traces résiduelles après la fusion. Par conséquent, pour minimiser la perte, MaskNet devrait attribuer des poids plus petits à ces outliers. La Fig. 2.4 montre l'efficacité de notre module de détection d'outliers dans le cas d'une voiture en mouvement.



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

Figure 2.4: Super-résolution à partir d'une séquence de 15 trames SkySat L1A. (a) L1B de Planet, (b) DSA, (c) Notre méthode avec réseau à détecter les outliers.

Chapitre 6: Extension aux séquences multi-expositions et amélioration de la fusion des caractéristiques

Ce chapitre détaille l'extension du cadre DSA pour traiter les séquences multi-expositions, dans le but d'effectuer conjointement la super-résolution et le traitement Haute Gamme Dynamique (HDR) à partir d'une série temporelle d'images de satellites avec différentes expositions.

Ce travail présente un ensemble unique de défis: Tout d'abord, le bruit dans les images dépend du signal, et ensuite, les temps d'exposition rapportés par les petits satellites push-frame peuvent être peu fiables avec des erreurs pouvant atteindre jusqu'à 20

La décomposition base-détail fournit une manière robuste de gérer les imprécisions dans les temps d'exposition. Les composantes de détail, robustes aux erreurs de temps d'exposition, sont pertinentes pour la super-résolution, tandis que les composantes de base, dépourvues d'informations à haute fréquence, peuvent être mises à l'échelle sans risque d'aliasing. La sortie finale est simplement la somme des composantes de base et de détail mises à l'échelle. Cela peut être brièvement résumé par:

$$\begin{aligned} B_t^{LR} &= I_t^{LR} * G, \\ D_t^{LR} &= I_t^{LR} - B_i^{LR}, \quad t = 1, \dots, K \\ B_r^{HR} &= \text{Zoom}(\overline{\{B_t^{LR}\}}) \\ D_r^{HR} &= \mathbf{Net}(\{D_t^{LR}\}) \\ I_r^{HR} &= B_r^{HR} + D_r^{HR} \end{aligned} \tag{2.12}$$

Ici G est un noyau Gaussien de déviation standard 1 et $\overline{B_t^{LR}}$ désigne une moyenne pondérée de B_t^{LR} pour $t = 1, \dots, K$.

De plus, nous introduisons une nouvelle stratégie de fusion par regroupement temporel, surpassant les performances du Feature Shift-and-Add (FS&A) dans le DSA. À travers des expériences quantitatives et qualitatives (voir Fig. 2.5), nous démontrons que notre méthode offre des performances de pointe en MISR satellite.

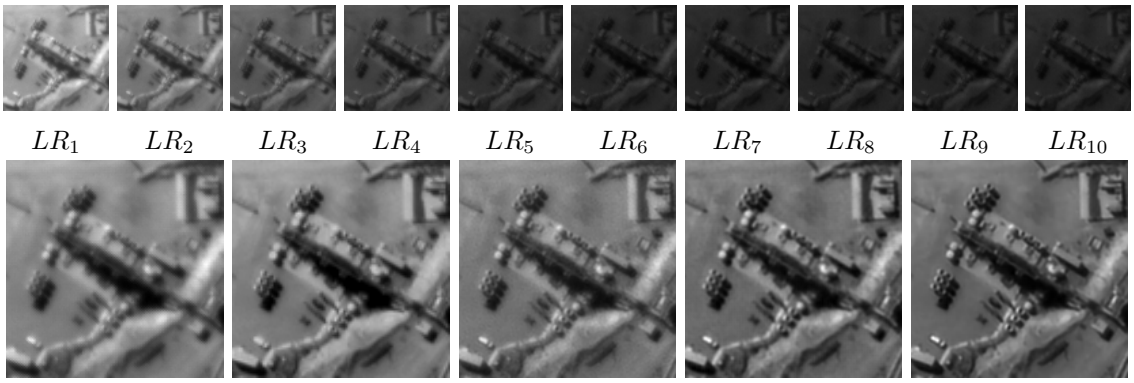


Figure 2.5: Super-résolution à partir d'une séquence multi-expositions de 10 images SkySat. Rangée supérieure: séquence LR avec différentes expositions. Rangée inférieure: Reconstructions à partir de cinq méthodes, y compris la nôtre formée avec auto-supervision (à droite).

Chapitre 7: Une brève analyse de la méthode de super-résolution SwinIR

Dans le Chapitre 7, nous jetons un regard neuf sur SwinIR, un modèle de restauration d'image de pointe basé sur l'architecture Swin Transformer. Contrairement aux réseaux neuronaux convolutionnels traditionnels, SwinIR capte habilement les motifs d'attention complexes parmi les patches d'image, ce qui donne des résultats supérieurs.

Au centre de notre discussion se trouve la compétence de SwinIR en matière de super-résolution single-image. Nous menons une analyse approfondie de l'impact du mécanisme d'auto-attention, notamment dans le contexte de l'auto-similarité, un facteur important dans la restauration d'image.

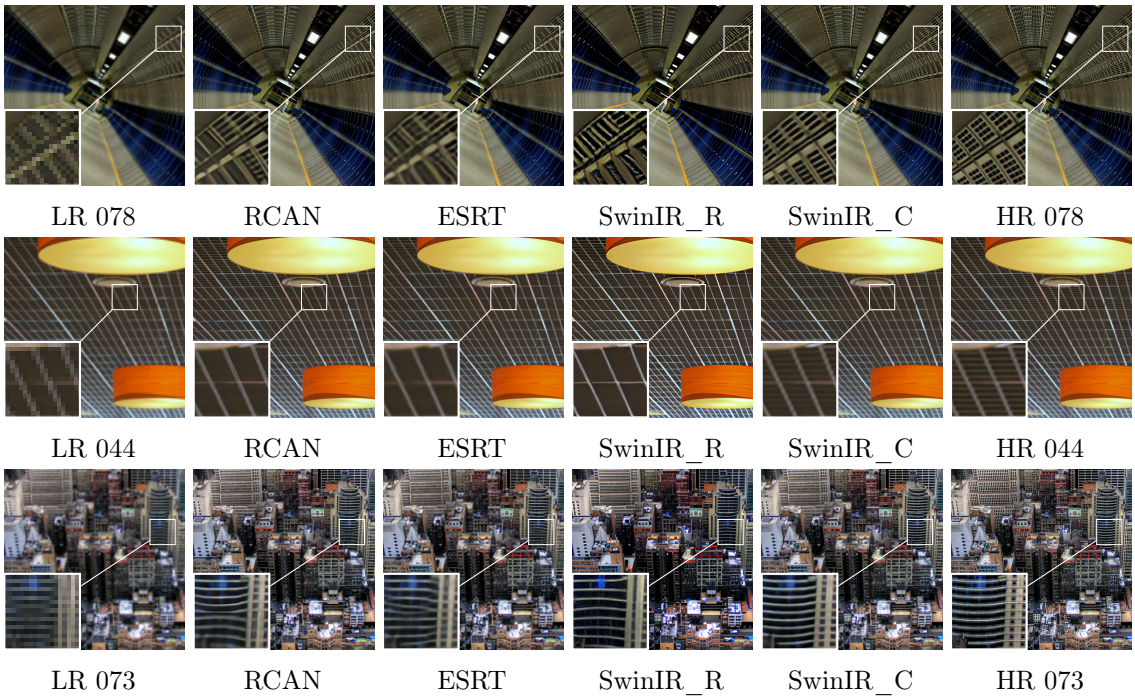


Figure 2.6: Comparaison qualitative entre les deux modèles SwinIR, RCAN, et ESRT sur le dataset Urban100. Super-résolution par un facteur de 4.

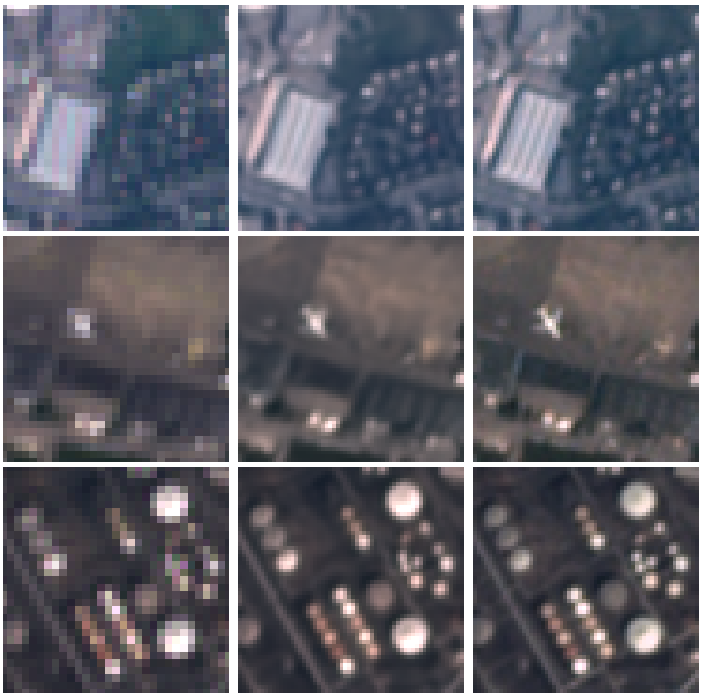
En utilisant le dataset Urban100, abondant en structures auto-similaires, nous mettons en lumière le mérite du mécanisme d'auto-attention de SwinIR à exploiter ces similarités pour améliorer la reconstruction de super-résolution, en particulier dans les zones à faible contraste ou aliasées (Fig. 2.6). Cependant, lorsque le modèle est appliqué à l'imagerie satellitaire, ses performances diminuent en raison de défis spécifiques au domaine.

Chapitre 8: Sur le rôle de l'alias et du décalage de bande pour la super-résolution des produits L1C

Contrairement aux méthodes SISR conventionnelles qui dépendent de la restauration perceptive ou des GANs pour faire face à l'ill-posedness de SISR, ce chapitre dévoile le potentiel de récupération réelle de détails à haute fréquence dans l'imagerie Sentinel-2. Ce succès est dû aux caractéristiques distinctes de Sentinel-2, spécifiquement l'aliasing et le décalage inter-bandes.

L'aliasing résulte d'un échantillonnage spatial faible par rapport à la fonction de transfert de modulation de l'instrument, tandis que les décalages inter-bandes sont attribués aux retards temporels dans les lignes d'acquisition de différentes bandes spectrales. L'effet synergique de ces caractéristiques transforme le problème SISR en un scénario mieux posé (comme MISR), ouvrant la voie à une véritable récupération d'informations à haute fréquence.

Notre étude est corroborée par des expériences impliquant à la fois des données synthétiques et réelles, ces dernières utilisant PlanetScope pour la supervision de la super-résolution de Sentinel-2. Les performances sur les vraies données L1C de Sentinel-2 sont présentées à la Fig. 2.7.



Sentinel-2 L1C Super-resolution x2 PlanetScope

Figure 2.7: Résultats SISR obtenus avec la perte L_1. Nous soutenons que l'alias caractéristique et le décalage de bande sont clés pour la SR x2 de l'imagerie Sentinel-2.

Chapitre 9: Exploiter le chevauchement des détecteurs pour une super-résolution auto-supervisée des produits L1B

Le chapitre précédent a révélé le potentiel de SISR pour les images de niveau 1C (L1C) de Sentinel-2. Ce chapitre présente une approche unique pour l'alignement conjoint auto-supervisé SISR et de bande des produits de niveau 1B (L1B) de Sentinel-2. L'accent passe des produits L1C aux produits L1B bruts, de stade précoce, principalement en raison d'une caractéristique inhérente à la conception du capteur de Sentinel-2 - le chevauchement des détecteurs.

Ce chevauchement de détecteurs entraîne des images L1B qui se chevauchent, ce qui nous permet d'entraîner notre réseau par apprentissage auto-supervisé. La stratégie est simple mais efficace: une image chevauchante est utilisée comme entrée et une autre comme cible ; le réseau a pour tâche de produire une image à haute résolution (HR) alignée en bande de telle sorte que, après déformation et sous-échantillonnage, elle corresponde à la cible cachée. Mathématiquement, cette stratégie équivaut à minimiser la perte auto-supervisée telle qu'exprimée dans l'équation:

$$\ell^{\text{L1BSR}}(\mathbf{Net}(I_0^{LR}), I_1^{LR}) = \left\| \mathbf{Warp}^{\downarrow}(\mathbf{Net}(I_0^{LR}), F_{1\to 0}) - I_1^{LR} \right\|_1, \tag{2.13}$$

Néanmoins, les produits L1B présentent leur propre ensemble de défis: ils contiennent un désalignement significatif entre leurs bandes spectrales, ce qui complique le calcul du mouvement entre les deux images qui se chevauchent (c'est-à-dire $F_{1\to 0}$ dans (2.13)). Pour y remédier, nous introduisons un nouveau module d'enregistrement spectral croisé (CSR), qui est également formé avec auto-supervision. Il est important de noter que le module CSR est essentiel à notre cadre, mais n'est utilisé que pendant l'entraînement pour guider la reconstruction. Au stade de l'inférence, notre réseau produit directement

une sortie HR alignée en bande à partir d'une image L1B mal alignée en bande. Dans la

Figure 2.8: L1BSR produit une sortie à haute résolution (HR) de 5m avec toutes les bandes correctement enregistrées à partir d'une seule image L1B Sentinel-2 à basse résolution (LR) de 10m avec des bandes mal alignées. Notez que notre méthode est formée sur des données réelles avec auto-supervision, c'est-à-dire sans aucune cible HR.

section expérimentale, nous menons des expériences complètes pour démontrer l'efficacité de notre cadre auto-supervisé. Nous montrons également que les performances de divers modules au sein de notre cadre sont à égalité avec celles des modules supervisés, à la fois quantitativement et qualitativement. La Fig. 2.8 présente la reconstruction HR par notre méthode sur de vraies images L1B.

2.6 Liste des publications

Présentées dans cette thèse

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. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 3881–3884. IEEE, jul 2021

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1121–1131, 2021

Ngoc Long Nguyen, Jérémy Anger, Axel Davy, Pablo Arias, and Gabriele Facciolo. Self-supervised push-frame super-resolution with detail-preserving control and outlier detection. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 131–134. IEEE, 2022

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1858–1868, 2022

Ngoc-Long Nguyen. A brief analysis of the swinir image super-resolution. *Image Processing On Line*, 12:582–589, 2022

Ngoc Long Nguyen, Jérémy Anger, Lara Raad, Bruno Galerne, and Gabriele Facciolo. On the role of alias and band-shift for sentinel-2 super-resolution. *arXiv preprint arXiv:2302.11494*,

2023

Ngoc Long Nguyen, Jérémy Anger, Axel Davy, Pablo Arias, and Gabriele Facciolo. L1bsr: Exploiting detector overlap for self-supervised single-image super-resolution of sentinel-2 l1b imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2012–2022, 2023

Autres

Jamy Lafenetre, Ngoc Long Nguyen, Gabriele Facciolo, and Thomas Eboli. Handheld burst super-resolution meets multi-exposure satellite imagery. *arXiv preprint arXiv:2303.05879*, 2023

Part I

Multi-image super-resolution in satellite imagery

3 Repurposing the Proba-V challenge for reference-aware super-resolution

The PROBA-V Super-Resolution challenge aims to advance research on Multi-Image Super Resolution (MISR) for satellite images by providing real low-resolution image series and corresponding high-resolution targets. However, a crucial piece of information is missing in the PROBA-V dataset—the identification of the low-resolution image that corresponds to the high-resolution target. This absence leads to a ranking of proposed methods that heavily relies on the heuristics used to determine which image in the series is most similar to the high-resolution target. In this chapter, we demonstrate the significance of this issue by achieving performance improvements for the two challenge winners through the use of a different reference image determined by a simple heuristic. To address this limitation, we propose PROBA-V-REF, a variant of the PROBA-V dataset where the reference image in the low-resolution series is explicitly identified. By providing the reference image, we observe a notable change in the ranking between different methods, highlighting the importance of reference-aware MISR. This variant better aligns with practical use cases of MISR, where the objective is to super-resolve a specific image from the series with a known reference. PROBA-V-REF serves as a valuable resource for evaluating the performance of various methods in tackling the reference-aware MISR problem in real-world scenarios.

3.1 Introduction

In 2019, the Advanced Concepts Team of the European Space Agency (ESA) organized a challenge [MIKC19] to super-resolve multi-temporal images from the PROBA-V satellite. This novel challenge was instrumental in enabling the training of deep learning MISR methods on real-world data, leveraging the unique capabilities of the PROBA-V satellite. This satellite is equipped with two types of cameras that capture images with different resolutions and revisit times, a setup that presents a special opportunity for supervising MISR methods on real-world data.

The challenge dataset consists of sets of LR images captured within a one-month time window over various sites. For each site, a high-resolution target image (HR) is provided. In each sequence, one of the LR images corresponds to the HR image, which we refer to as the true reference. However, the identity of the true reference images is not provided, a key limitation that hampers the full potential of the dataset. Knowledge of the LR reference can aid in producing results that better match the HR image, as significant changes may occur between images taken at different dates.

Traditional MISR methods, such as shift-and-add, kernel regression [TFM07], and poly-

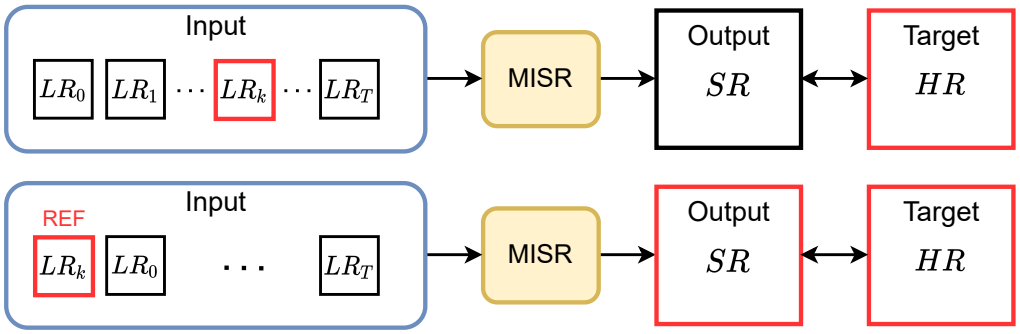


Figure 3.1: The PROBA-V dataset (top) does not make any distinction between the LR images. One of them was acquired at the same time as the target HR image which is used for training and evaluation. The MISR methods need to determine a reference without knowing which is the one corresponding to the target. We propose PROBA-V-REF (bottom), a version of PROBA-V where the identity of true reference is known.

nomial fitting [AEdFF20] all need a LR reference image. They start by registering all LR images to a common domain, often chosen to be the domain of a specific LR image in the series, typically the one targeted for super-resolution. The two top-performing methods in the Proba-V challenge, DeepSUM [MVFM19] and HighRes-net [DKG+20], also select a specific LR image as a reference for reconstruction. DeepSUM chooses the LR image with the highest clarity as the reference for the registration step, while HighRes-net selects the median of the nine clearest LR images as the reference in the fusion step.

Numerous teams have participated in the challenge, and a "post-mortem" contest continues to benchmark new MISR methods. All these works attempt to solve the problem without knowledge of the reference images. We believe that the problem of MISR without a reference image is interesting and holds several applications. However, in such cases, the reference image should be completely random, which is not the case in the PROBA-V challenge. For example, a cloud-free LR image has a higher chance of being the reference than a cloudy LR image. This bias introduces noise into the resulting benchmark, and a method's performance may be influenced not by a more suitable architecture or training but by a better heuristic for guessing the reference image.

On the other hand, reference-aware MISR is a relevant problem in its own right. In many practical scenarios, the goal is to super-resolve a specific image from the sequence, such as one corresponding to a specific date. While this problem is considerably easier, it is far from being solved. In other domains, such as video super-resolution or burst image super-resolution, the standard definition of MISR includes the reference image. Therefore, we believe that a variant of the PROBA-V dataset with the true reference images would be valuable for the computer vision community.

In this chapter, we first demonstrate the impact of the heuristic used to select the reference LR image in the PROBA-V challenge. By simply changing the reference images of the two winning methods with a different one chosen through a simple heuristic, we observe improved performance. We also highlight that the true reference image can be obtained in the training and validation splits of the dataset by comparing with the HR target. Consequently, we propose PROBA-V-REF, a version of the PROBA-V dataset that includes the true LR references. Finally, we retrain the top two methods from the challenge on the PROBA-V-REF dataset and show that the ranking between them becomes inverted.

3.2 Recovering the true LR reference image

The PROBA-V dataset comprises 566 scenes from the NIR spectral band and 594 scenes from the RED band. Each scene consists of a single HR image of size 384 x 384 pixels and multiple LR images (ranging from 9 to 35) captured over a period of one month, with dimensions of 128 x 128 pixels. The LR images within a set can exhibit variations due to changes in illumination, presence of clouds, shadows, or ice/snow covering.

A status map is provided for each LR image to indicate the reliability of the pixels for fusion. The "clearance score" measures the percentage of clear pixels in the status map. The dataset is carefully curated to ensure that the LR images have a clearance score of at least 60% and the HR image has a clearance score of at least 75%. If multiple HR images meet this clearance criterion within a 30-day period, only the one with the highest clearance score is selected as the target HR image. Since the PROBA-V dataset does not distinguish between the LR images, MISR methods are required to produce an average SR image. However, to accurately recover the true details in the SR image, knowledge of the true LR reference image is essential (see Figure 3.1).

To determine the true LR reference for each element in the training set, a comparison is made between the LR images and the HR image. A filtered and subsampled version of the HR image is computed, and the LR frames are aligned with the downsampled HR image using the inverse compositional algorithm [BM01]. Pixel-wise root-mean-square errors are then calculated between the LR images and the downsampled HR image. The true reference image is identified as the LR image with the minimum error. The computed indexes of the true references for the PROBA-V dataset can be found at the following link: https://github.com/centreborelli/PROBAVref.

3.3 Experiments

In this section, we emphasize the significance of the reference image in the super-resolution process and highlight its impact on the overall technique performance. Additionally, we illustrate and discuss the advantages of utilizing the PROBA-V-REF dataset for real-world applications.

To evaluate the quality of the reconstructions, we employ the "corrected clear" PSNR (cPSNR) metric introduced during the PROBA-V challenge [MIKC19]. This metric is particularly relevant as it incorporates the status map of the ground truth HR image, enabling the consideration of intensity biases and small pixel translations between the super-resolved image and the target. By utilizing the cPSNR metric, we can assess the reconstruction quality more accurately and account for potential discrepancies caused by variations in pixel intensity and minor pixel-misalignments.

3.3.1 Experimental Settings

As mentioned earlier, the top two competitors in the PROBA-V challenge, DeepSUM and HighRes-net, rely on specific LR images as anchor references.

DeepSUM [MVFM19] — The winner of the challenge, DeepSUM uses the LR image with the highest clearance as the reference. A registration step aligns all other images to this reference.

HighRes-net [DKG+20] — HighRes-net achieved the second place in the challenge. It

Table 3.1: Average cPSNR (dB) over the validation dataset for DeepSUM and HighRes-net. The original performance is highlighted in orange and the best performances are highlighted in blue

Methods	**Training ref.**	**Evaluation ref.**			
		Simil.	*Clearance*	*Median*	*Heuristic*
DeepSUM	*Clearance*	**47.99**	**47.75**	47.62	47.87
HighRes-net	*Median*	**47.77**	47.26	**47.48**	47.57

selects the median of the nine images with the highest clearance as a shared representation for multiple LR images. Each LR image is jointly embedded with this reference image before being recursively fused.

To demonstrate that the choices of reference images by DeepSUM and HighRes-net may not be optimal, we retrain these models from scratch using the true LR references (as described in Section 3.2). We refer to these adjusted methods as **DeepSUM-ref** and **HighRes-net-ref**, respectively. Furthermore, we compare the performance of these MISR methods with a single-image super-resolution (SISR) algorithm trained on the true LR references. We introduce **DeepSUM-SI**, which is a modified version of DeepSUM designed for SISR by replacing all input images with the true references.

Tables 3.1 and 3.2 present the performance of these methods on the validation set for the NIR spectral band, which consists of 170 scenes.

We explore different approaches for selecting the reference image on the validation set:

Similarity — This approach uses the true reference image computed in Section 3.2.

Highest clearance — The LR image with the highest clearance score is chosen as the reference, following the method used in [MVFM19].

Median — The reference image is determined as the median of the nine clearest LR observations, following the approach in [DKG+20].

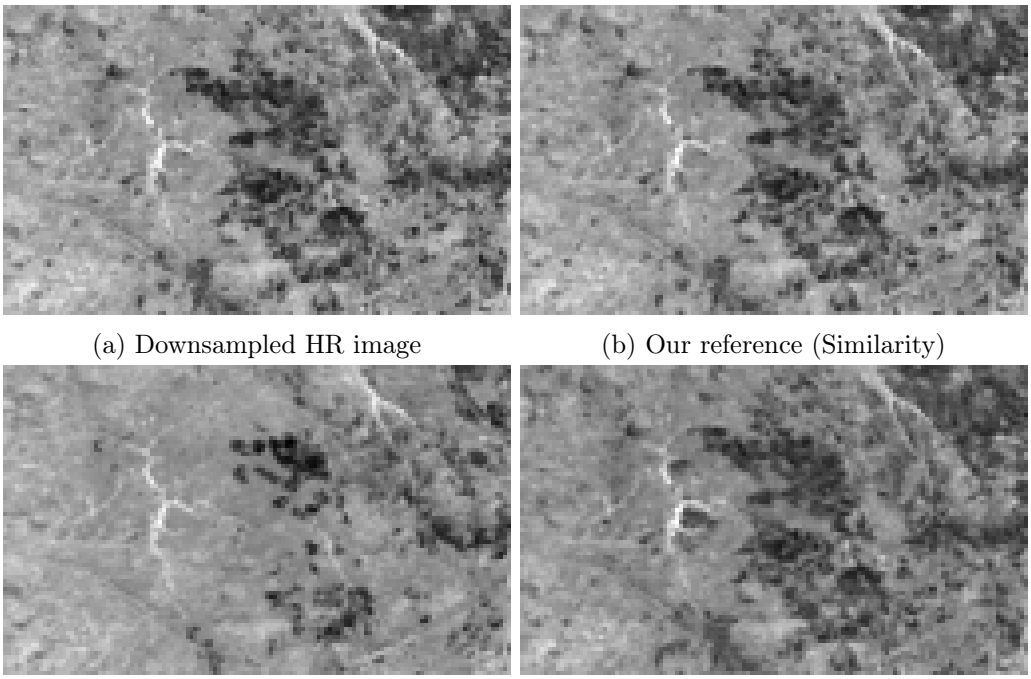
Heuristic — In the absence of ground truth HR images in the test set, a heuristic is employed to predict the reference images. This is accomplished by minimizing the objective function:

$$\begin{aligned} i_{\text{heur}} = \operatorname{argmin}_i \Big\{ & \|\text{Mask}_i^{\text{LR}} - \text{Downscale}(\text{Mask}^{\text{HR}})\|_1 \\ & + \alpha \, |\text{median}(\text{LR}_i) - \text{median}(\text{LRset})| \\ & + \beta \, \text{clearance}(\text{LR}_i) \Big\}, \end{aligned} \tag{3.1}$$

where Mask designates the status map of an image, LRset is the set of input LR images, clearance is the sum of all clear pixels of a LR, we manage to guess the true references in more than 50% of scenes in the training set. We set $\alpha = 0.1, \beta = 0.3$ in our experiments. Fig. 3.2 provides a visual comparison between different heuristics.

3.3.2 Discussion

Upon analyzing the results presented in Table 3.1, we observe that the performance of the top competitors in the PROBA-V challenge is influenced by the choice of reference



(a) Downsampled HR image (b) Our reference (Similarity)

(c) DeepSUM reference (Clearance) (d) HighRes-net reference (Median)

Figure 3.2: Visual comparision between different heuristics.

Table 3.2: Average cPSNR (dB) over the validation dataset for DeepSUM-SI, DeepSUM-ref and HighRes-net-ref. For each methods, the best performance is highlighted in blue.

Methods	**Training ref.**	**Evaluation ref.**			
		Simil.	*Clearance*	*Median*	*Heuristic*
DeepSUM-ref	*Similarity*	**50.24**	46.38	46.69	**49.10**
HighRes-net-ref	*Similarity*	**50.49**	46.35	46.47	**49.29**
DeepSUM-SI	*Similarity*	**49.05**	45.57	45.85	**47.96**

images. Without retraining, utilizing the true references or even the “heuristic references” consistently improves the results, with DeepSUM outperforming HighRes-net in this scenario.

After retraining the models with the true references (as shown in Table 3.2), DeepSUM-ref and HighRes-net-ref exhibit significantly superior performance compared to the original DeepSUM and HighRes-net, achieving a large margin of improvement (2.49 and 3.01 dB, respectively). Even with the “heuristic references”, the adjusted methods still outperform the original approaches, with improvements of 1.35 dB for DeepSUM-ref and 1.81 dB for HighRes-net-ref. It should be noted that using Mask^{HR} in this context deviates from the contest rules. Nevertheless, as a proof of concept, we submitted the results of HighRes-net-ref with the "heuristic references" to the official post-mortem PROBA-V challenge[1], where it currently holds the second-place position in the leaderboard, surpassing the performance of the original DeepSUM and HighRes-net methods. Although this heuristic approach relies on the mask of the HR image, it demonstrates the substantial impact that the choice of reference image can have on the results.

Furthermore, when the true references are provided, HighRes-net-ref performs better than DeepSUM-ref. This indicates that the design of the challenge strongly influences its outcome and highlights the importance of the reference image selection.

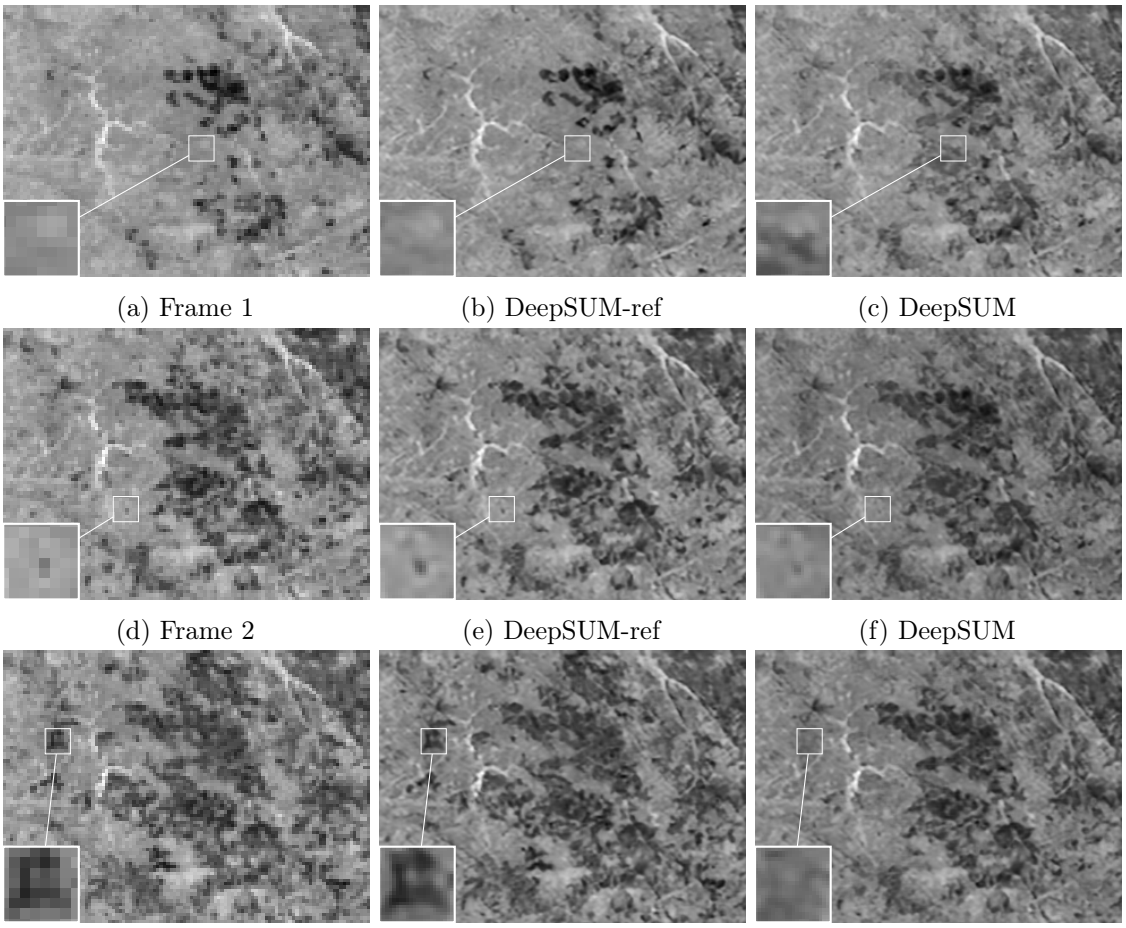
On the other hand, the SISR algorithm DeepSUM-SI achieves significantly better results compared to the MISR algorithm DeepSUM. This can be attributed to the temporal variability among the LR observations. In the absence of knowledge about the reference image, networks trained on MISR tasks are required to simultaneously guess the reference image and super-resolve that specific image using information from other images in the set. As a result, the network tends to predict an average SR image. Incorporating information about the reference image allows the networks to focus solely on the SR problem, leading to improved performance.

Overall, these findings emphasize the crucial role of reference image selection in SR tasks and highlight the potential for significant performance gains by incorporating true references or employing effective heuristics for their estimation.

In order to assess the impact of the reference image on the results of DeepSUM and DeepSUM-ref, we carefully select three LR images captured on different days as the reference images (refer to Figure 3.3). Notably, DeepSUM-ref consistently produces SR images that accurately recover fine details corresponding to the references. In contrast, the vegetation representation in the outputs of DeepSUM does not align well with that of the references. This indicates that the reconstructions of DeepSUM lack the desired correlation with the reference images. As a result, DeepSUM-ref proves to be more suitable for practical super-resolution applications where the objective is to enhance a specific image within a time series.

In summary, this chapter underscores that the evaluation of the PROBA-V challenge is not solely based on the MISR performance of the methods, but also heavily influenced by the selection of LR reference images. However, in many practical applications, the choice of the reference image is predetermined by the specific requirements of the task at hand. To address this use case, we introduced PROBA-V-REF, a modified version of the dataset that includes the true LR reference images in the training and validation splits. These reference images were determined by comparing the LR images with a downscaled

[1] https://kelvins.esa.int/PROBA-v-super-resolution-postmortem



(a) Frame 1 (b) DeepSUM-ref (c) DeepSUM

(d) Frame 2 (e) DeepSUM-ref (f) DeepSUM

(g) Frame 3 (h) DeepSUM-ref (i) DeepSUM

Figure 3.3: Examples of reconstruction by DeepSUM-ref and DeepSUM with different references. The first column corresponds to crops of three different LR images in a set. The second and the third column show the reconstruction by DeepSUM-ref and DeepSUM respectively when using each of these three LR as the reference image.

version of the ground truth HR image. By providing the true LR references, we enable future methods to leverage this unique real-world dataset and focus on the fundamental challenge of MISR, which is effectively utilizing the complementary information in the LR images.

This concept of reference-aware super-resolution emerges as a cornerstone of this thesis. All MISR techniques developed and discussed in this thesis underline the importance of a reference frame. From the methods developed within this work to those we benchmark against in our evaluations, the reference frame takes precedence. This initial foray into reference-aware super-resolution sets the stage for the ensuing chapters, establishing a foundation for our exploration of this crucial aspect of satellite image super-resolution.

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

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 chapter, 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.

4.1 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 [MSS+14, AEdFF20].

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.

In this chapter, we focus on MISR from push-frame satellite sensors such as the SkySat constellation from Planet. The SkySat satellites [MSS+14] contain a full-frame sensor capable of capturing bursts of overlapping frames. So the same point on the ground is

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

Figure 4.1: 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).

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 [AEdFF20] for the formation of the low-resolution images I_t^{LR}

$$I_t^{LR} = \Pi A_t(\mathcal{I} * k) + n_t, \tag{4.1}$$

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 [AEdFF20]), Π is the bi-dimensional sampling operator due to the sensor array, 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× zoom factor. Our goal in this work is 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 [CZY+19], where it was shown that models trained on synthetic data [AT17] 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 [MDS]). An exception is the PROBA-V dataset, proposed in [MIKC19], which allows to train supervised deep-learning MISR

[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.

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 [WGDE+19, MBH20]. Good results in denoising of real images have been obtained using synthetic datasets [BMX+19, ZAK+20]. 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 [EDM+19, EDAF19, DAD+21, YPPJ20] 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 chapter, we proposed a framework for self-supervised training of MISR networks without requiring high resolution ground truth images.

Our framework (Section 4.3) 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 (Sections 4.4 and 4.5) 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.

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 Figure 4.1).

4.2 Related works

Video and burst super-resolution. There is a long history of MISR techniques from bursts of images and videos (see [NM14, YSL+16] 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 [WGDE+19, TFM07], variational formulations [TK95, MO08, FAAC09, RF18], and fusion in transformed domains [KBV90, LR00, NM00, AEdFF20].

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 [KPB88]. 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 [SVB18, XCW+19, DZL20]. 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 [TGL+17], 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 [JWOKJK18], deformable convolutions [WCY+19], progressive fusion residual blocks [YWJ+19]. 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 [FGT19, IJG+20a].

Super-resolution for satellite images. Most MISR methods for satellite images are still based on classic model-based techniques [LR00, MN07, MSS+14, AEdFF20, AEF21]. 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.

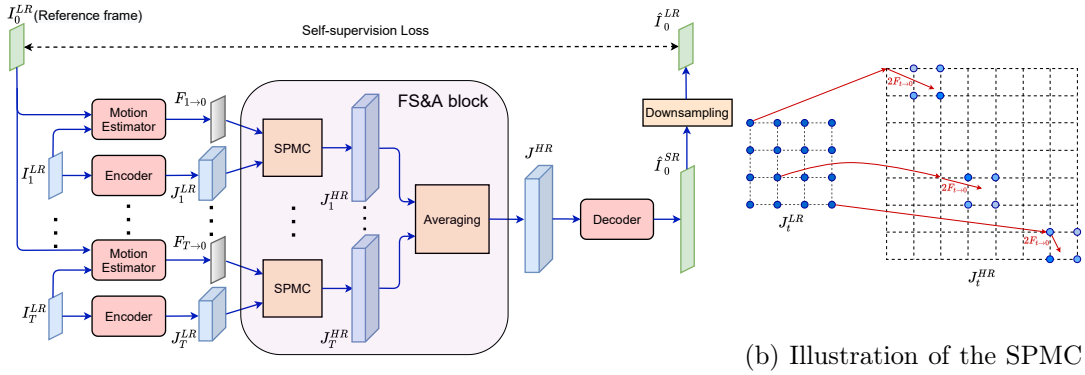
In the case of SISR, some methods resort to simulating realistic data [ZTS+20] or to combine images acquired from different satellites with different resolutions [PLPD18, SRMV20] so as to avoid the synthetic downsampling.

To the best of our knowledge, the only dataset with real LR and HR satellite images is the PROBA-V dataset [MIKC19]. This dataset and the associated challenge have triggered research in MISR of satellite imagery [DKG+20, SMKC20, MVFM19, MVFM20]. In the PROBA-V dataset, the LR reference (the LR image associated to the HR target) is unknown. This last point was analyzed in Chapter 3, where PROBA-V-ref was proposed, 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. [LMH+18] 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:

$$\mathcal{R}_{\text{N2N}}(\mathbf{Net}) = \sum_j \ell(\mathbf{Net}(N_j), N'_j). \tag{4.2}$$

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 estima-



(a) Overview of our architecture and framework

(b) Illustration of the SPMC layer

Figure 4.2: (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 $\times 2$ layer [TGL+17]: Splatting LR features onto the HR domain using the flow $F_{t\to0}$.

tors 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, [KBJ19, BR19] 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 [EDM+19] applied this idea to fine-tune a single frame denoising (and/or demosaicking [EDAF19]) network requiring only a single noisy video or burst. Extensions were proposed in [DAD+21, YPPJ20] 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 [EDM+19, EDAF19, DAD+21].

Other self-supervision strategies were also explored for SISR. In [SCI18], an algorithm that exploits the internal recurrence of information across scales inside a single image is proposed. The authors of [YLZ+18, KPL+20] propose to use cycle-consistency and adversarial losses to train a SISR neural network without supervision using unpaired LR and HR images. In [LLTMK19b], an extension of the Deep Image Prior [UVL18] is applied to fine-tune a SISR network on a single image.

4.3 Self-supervised multi-image SR

We first present an overview of our proposed *Deep Shift-and-Add* in Section 4.3.1. Then we describe our framework for self-supervised MISR training in Section 4.3.2, and in Section 4.3.3 we provide details about the training.

4.3.1 Architecture

Our neural network (illustrated in Figure 4.2a) 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 [FH02, MN07, GCLK08, ABHY00, Jia12].

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 [TGL+17] 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 [BDVGT21, DKG+20, XNC+20]. 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 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 Section 4.3.2.

The trainable modules of the proposed architecture (shown in red in Figure 4.2a) include the Motion Estimator, the Encoder and the Decoder.

Motion Estimator. We follow the work of [SVB18] 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}

$$F_{t\to 0} = \mathbf{ME}(I_t^{LR}, I_0^{LR}; \Theta_{\mathbf{ME}}) \in [-R, R]^{H\times W\times 2}. \tag{4.3}$$

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 [VSVV07]. This network will be trained with a maximum range of motions $[-R, R]^2$ (in this chapter, $R = 5$ pixels).

The **ME** network follows a simple hourglass style architecture (4 scales with 32, 64, 128 and 256 features, 2 convolutions blocks per scale [SVB18]). 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

$$J_t^{LR} = \mathbf{Encoder}(I_t^{LR}; \Theta_{\mathbf{E}}) \in \mathbb{R}^{H\times W\times N}, \tag{4.4}$$

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 [TGL+17], then a weighted average is computed.

The SPMC module uses the flow $F_{t\to 0}$ to compute the positions of the samples from J_t^{LR} in the HR grid

$$J_t^{HR} = \text{SPMC}(J_t^{LR}, \{F_{t\to 0}\}) \in \mathbb{R}^{rH\times rW\times N}, \tag{4.5}$$

where r is the upscaling factor ($r = 2$ in our case). As in [TGL+17], every LR pixel is "splatted" on a neighborhood of the computed HR position using bilinear weights (see Figure 4.2b). 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}

$$J^{HR} = (\textstyle\sum_t J_t^{HR})(\sum_t W_t^{HR})^{-1}, \tag{4.6}$$

where $W_t^{HR} = \text{SPMC}(1, \{F_{t\to 0}\})$ 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

$$\hat{I}_0^{SR} = \textbf{Decoder}(J^{HR}; \Theta_{\mathbf{D}}) \in \mathbb{R}^{rH\times rW}, \tag{4.7}$$

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.

4.3.2 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 (4.1), 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

$$\ell_{self}(\hat{I}_0^{SR}, I_0^{LR}) = \|D_2(\hat{I}_0^{SR}) - I_0^{LR}\|_1, \tag{4.8}$$

where $\hat{I}_0^{SR} = \textbf{Net}(\{I_t^{LR},\}_{t=1}^T, I_0^{LR})$ 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 [LMH+18], 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 [LMH+18, EDM+19].

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 [BR19, DAD+21]. 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 [YHD16]. 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:

$$\ell_{me}(\{F_{t\to 0}\}_{t=1}^{T}) = \sum_t \|I_t^{LR} - \mathbf{Pullback}(I_0^{LR}, F_{t\to 0})\|_1 + \lambda_1 TV(F_{t\to 0}), \tag{4.9}$$

where **Pullback** computes a bicubic warping of I_0^{LR} according to a flow, TV is the finite difference discretization classic Total Variation [ROF92] 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.

4.3.3 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 [TGL+17]. As a result, we separate the training into two phases.

To pre-train the motion estimation network we use the motion estimation loss (4.9). We initialize the weights of the motion estimator with Xavier's initialization [GB10]. In our experiments, we set λ_1 to 0.01 and batch size to 64, then use Adam [KB14] 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:

$$\text{loss} = \ell_{self} + \lambda_2 \ell_{me}. \tag{4.10}$$

We set $\lambda_2 = 10$ in our experiments. The initial weights are set using He initialization [HZRS15], 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 Section 4.4.2). In that case, we replace ℓ_{self} in (4.10) by

$$\ell_{supervised}(\hat{I}_0^{SR}, I^{HR}) = \|\hat{I}_0^{SR} - I^{HR}\|_1, \tag{4.11}$$

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.

4.4 Experiments

In our experiments, we use real push-frame images acquired by satellites from the SkySat constellation [MSS+14]. These images are also used to create a simulated dataset used for a quantitative evaluation.

4.4.1 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 $\times 1.25$ 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.

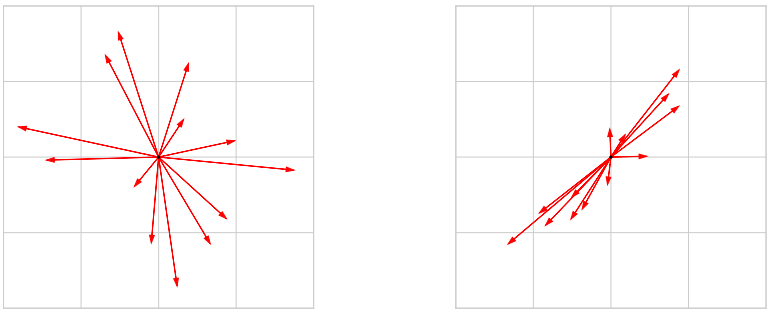
Simulated dataset. A part of our experiments will be conducted on a simulated dataset generated from a set of crops of L1B products. For a given crop B, the ground truth HR image I^{HR} is computed by filtering B with a small Gaussian kernel with $\sigma = 0.3$ so as to simulate a small optical blur. 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

$$\begin{aligned} I_0^{LR} &= D_2(I^{HR}) + n_0, \\ I_t^{LR} &= D_2(\mathrm{Shift}_{\Delta_t}(I^{HR})) + n_t, \quad t = 1, \dots, T, \end{aligned} \tag{4.12}$$

where D_2 is the subsampling operator, $\mathrm{Shift}_{\Delta_t}$ 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 Figure 4.3.



(a) Uniform sampling (b) Degenerate sampling

Figure 4.3: 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.

Table 4.1: 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**

We also generate datasets with 100% uniform and 100% degenerate samplings. We refer to them as *mixed*, *uniform*, and *degenerate*.

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.

4.4.2 Super-resolution on simulated data

We evaluate the performance of our super-resolution network on the simulated dataset described in Section 4.4.1 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 [MN07, GCLK08, ABHY00, Jia12]. *ACT-Spline* is a state-of-the-art method based on spline fitting [AEdFF20]. In Shift-and-Add and ACT-Spline, the LR images are aligned using the inverse compositional method [BM01, BFS18]. *HighRes-net* is a MISR CNN with implicit motion estimation trained originally for the PROBA-V challenge [DKG+20]. Here we use a variant that was shown in Chapter 3 to have a better performance on the PROBA-V-ref dataset.

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

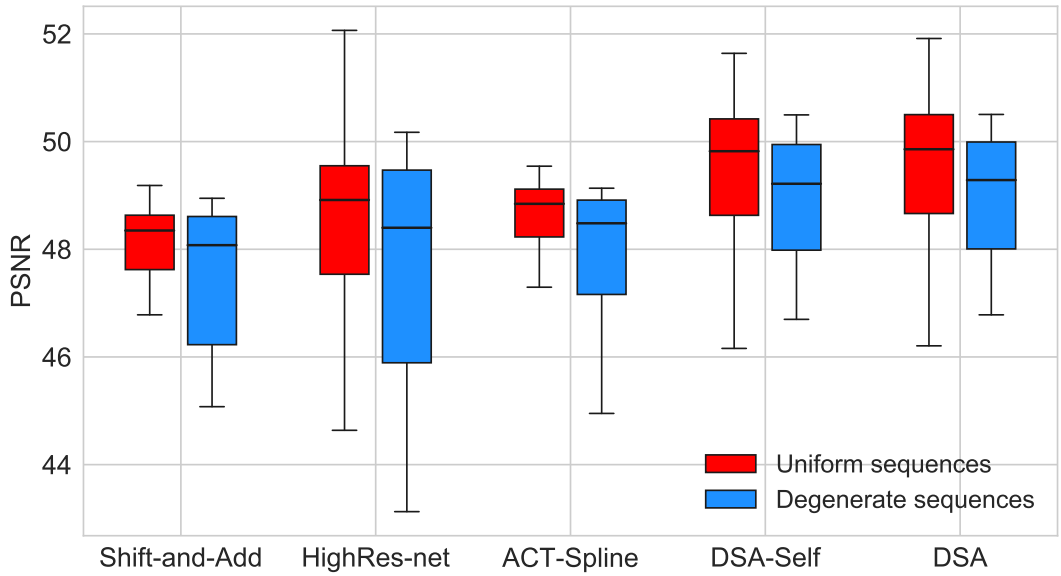


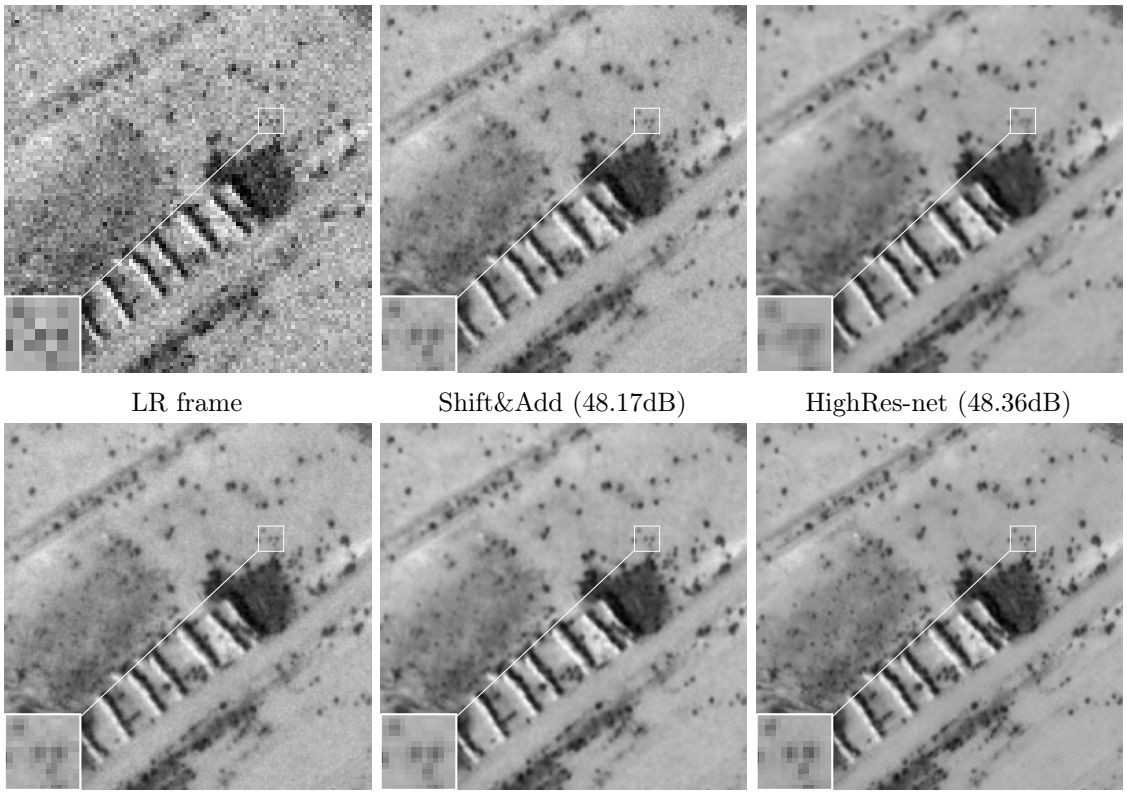
Figure 4.4: PSNR of different methods over our main validation set with 16 input frames per sequence.

Table 4.2: 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).

Train		**Test/mixed dataset**			**Test/uniform dataset**			**Test/degenerate dataset**		
Number of images	**Train dataset**	**Number of images** 5	16	30	**Number of images** 5	16	30	**Number of images** 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

Table 4.1 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. Figure 4.4 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 [AEdFF20] 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}.

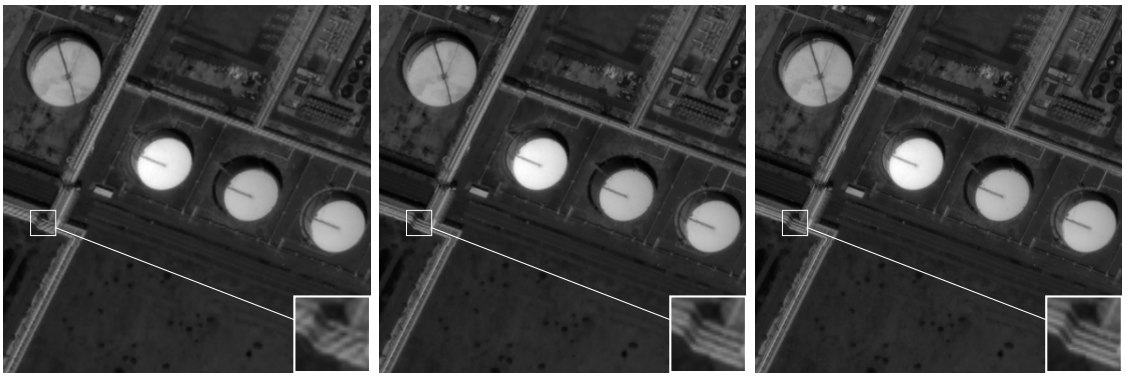
Figure 4.5 presents a qualitative comparison between Shift-and-Add, HighRes-net, ACT-Spline and the proposed DSA-Self on a sequence of 16 frames from the validation set with uniform sampling of the shifts. The output of our supervised DSA (49.93dB) was not included as it was indistinguishable to the DSA-Self result. In this example, DSA-Self



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

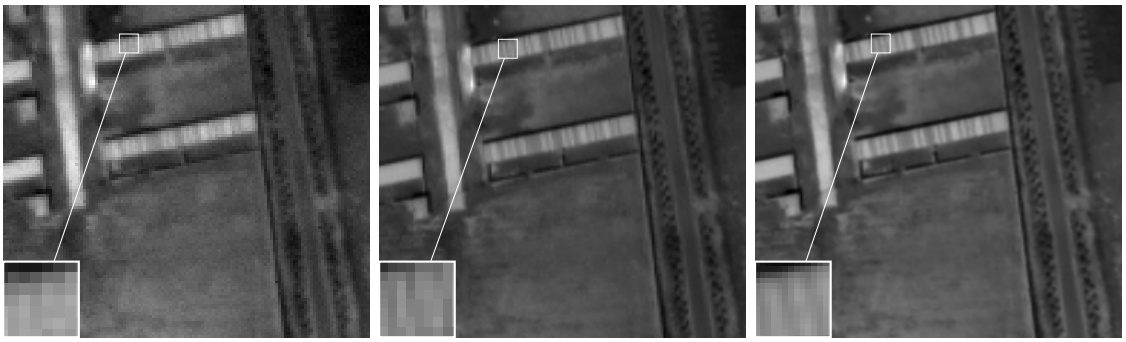
ACT-Spline (48.81dB) | DSA-Self (**49.86dB**) | Ground truth

Figure 4.5: 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 4.6: 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 4.7: Super-resolution from a sequence of 15 SkySat L1A frames. (c) was obtained using Eq. (4.13) as reconstruction loss with deconvolution.

Table 4.3: 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

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.

Reproducing the same experiment but with the degenerate samplings, we observe that HighRes-net fails to remove aliasing artifacts of the LR frames (see Figure 4.6), 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.

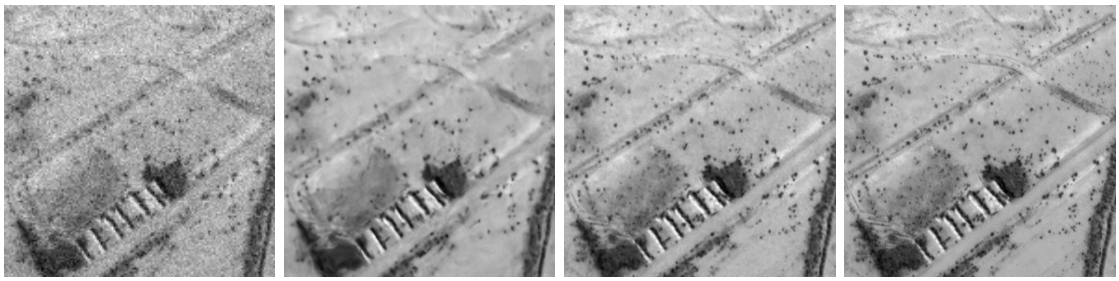
4.4.3 Super-resolution trained on real data

We applied our framework to train our DSA-Self network on the dataset of real SkySat L1A bursts. Since there 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 [AEdFF20] 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 [ADF19, KF09] as in [AEdFF20]. Instead, we opt for incorporating the deconvo-

Table 4.4: 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(dB)	43.92	45.82	49.33	50.50



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

Figure 4.8: 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.

lution in the self-supervision loss

$$\ell_{self}(\hat{I}_0^{SR}, I_0^{LR}) = \|D_2(\hat{I}_0^{SR} * k') - I_0^{LR}\|_1. \tag{4.13}$$

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).

Figure 4.1 and 4.7 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 Figure 4.1 and the vertical bars in the Figure 4.7 are much sharper in the proposed method.

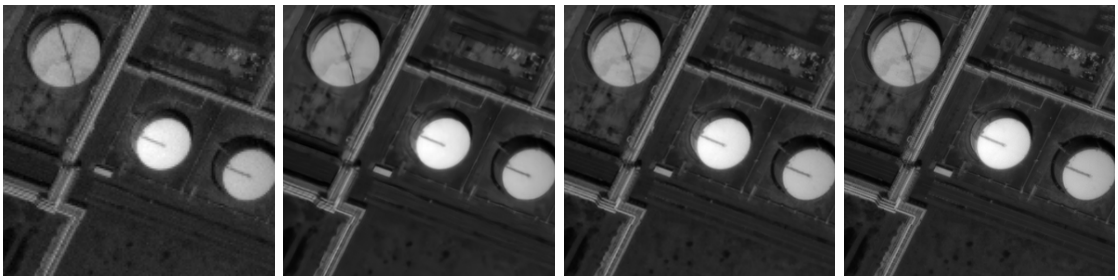
At inference time, our proposed method takes 0.6 seconds to produce a $\times 2$ super-resolved image from a sequence of 15 L1A images (256×256 pixels).

4.4.4 Comparison with single-image super-resolution approaches

In this section, we conduct a comprehensive comparison between our MISR method and a popular SISR method, namely SRGAN [LTH+17]. To ensure a fair evaluation, we retrained the SRGAN model using our simulated dataset.

It is important to note the Nyquist-Shannon sampling theorem, which states that high-frequency details cannot be fully recovered from a single aliased LR image. Consequently, SISR techniques often introduce "hallucinations" or spurious details during the reconstruction process, which are inappropriate for accurate remote sensing analysis. On the other hand, the fundamental goal of MISR methods is to genuinely increase the optical resolution of the SR images, enabling more reliable and precise analysis.

Quantitative results in Table 4.4 shows that our MISR method consistently outperforms the SISR approach by a large margin. The improved performance can be attributed to the utilization of multiple LR input images, which provide complementary information for



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

Figure 4.9: Visual comparison with the SISR method SRGAN. In this example, SRGAN fails to remove the alias present in the LR image.

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

Figure 4.10: Loss-based result contains less noise and no ringing artifacts (on the top of the building).

more accurate and faithful SR image reconstruction. Furthermore, the qualitative visual comparisons in Figures 4.8 and 4.9 vividly illustrate the superiority of our MISR method in preserving important details, minimizing artifacts, and delivering more visually appealing SR images.

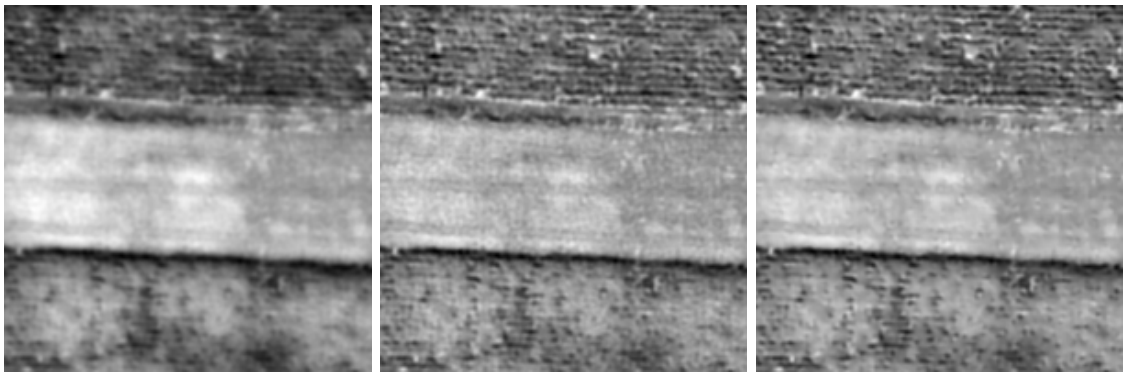
These compelling results reinforce the suitability of our MISR method for remote sensing applications, as it effectively increases the true optical resolution of SR images without introducing misleading or artificial features.

4.4.5 Image sharpening: comparison with a variational method

In this section, we delve into a detailed examination of two formulas for the self-supervision loss, namely Equation (4.8) and Equation (4.13). These formulas play a crucial role in our approach and enable a comparison with a variational method for image sharpening.

Equation (4.8) represents the self-supervision loss used to train the network. The objective is to produce an SR image, $\hat{I}_0^{SR}$, that, after subsampling, matches the corresponding LR reference image, I_0^{LR}. Following our image formation model, the output of the network is a high-resolution image that is inherently blurry due to the image restoration process.

Using the network trained with the loss (4.8), we can subsequently restore a sharp image, denoted as I, from the blurry SR output, $\hat{I}_0^{SR}$. This involves solving a non-blind deconvolution problem, expressed as:



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

Figure 4.11: 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.

$$\arg\min_{I} |I * k - \hat{I}_0^{SR}|_2^2 + \lambda |\nabla I|_1, \tag{4.14}$$

where the blur kernel k is defined in the Fourier domain as $\hat{k}(\boldsymbol{\omega}) = (5|\boldsymbol{\omega}|+1)^{-1}$ [AEdFF20]. The regularization weight, λ, can be set to a very low value due to the low noise level in the SR results. An efficient solution to this inverse problem can be achieved using a half-quadratic splitting method, similar to prior works [KF09, ADF19].

Alternatively, instead of relying on the variational method described above, we can integrate the deconvolution step (with the same blur kernel k) directly into the self-supervision loss, as indicated in Equation (4.13). In this scenario, the network is trained to produce a sharp SR image directly, such that when blurred, it aligns with the observed blurry samples.

To provide a visual comparison between our loss-based method and the variational method, we present Figures 4.10 and 4.11. Figures 4.10a and 4.11a showcase the outputs of the network trained with the self-supervision loss in Equation (4.8), which are inherently blurry. Meanwhile, Figures 4.10b and 4.11b depict the deconvolved results obtained from these blurry images using the variational method. Lastly, Figures 4.10c and 4.11c represent the outputs of the network trained with the self-supervision loss (4.13). Notably, the loss-based results exhibit sharpness comparable to the variational approach while avoiding unwanted high-frequency artifacts and demonstrating reduced noise. Moreover, our loss-based method is simple, efficient, and does not necessitate additional regularization techniques.

These findings highlight the effectiveness of our loss-based approach for image sharpening, offering a viable alternative to the more complex variational methods.

4.5 Ablation Study

To thoroughly investigate the significance of different elements within the proposed architecture, we conducted a series of experiments in a supervised setting. Our analysis focused on three key aspects: the role of the encoder, the impact of the number of features gener-

ated by the encoder, and the influence of training with a variable number of input images and considering degenerate sampling configurations.

Role of the encoder Firstly, we examined the role of the encoder in the DSA architecture. By comparing the performance of two methods – one with the encoder and one without – we sought to evaluate the effect of feature fusion versus pixel fusion. Our findings indicate that the network utilizing feature fusion demonstrates superior noise reduction capabilities in the SR image. This advantage becomes particularly pronounced when working with a limited number of LR input images. Notably, when using only five input images, the network with the encoder outperformed the one without by 0.22dB (refer to Table 4.3).

Impact of number of features Continuing our investigation, we delved into the influence of the number of features generated by the encoder. In a supervised setting, we re-trained our DSA architecture using three different numbers of features: 4, 16, and 64. Subsequently, we evaluated these models on 50 validation sequences, each comprising 16 images. The results revealed that employing 64 features yielded the best outcomes with the best high-frequency reconstruction, while an increase in the number of features beyond this threshold led to diminishing returns. This observation underscores the importance of selecting an appropriate number of features for optimal performance in terms of noise reduction and detail recovery in the SR image.

Variable number of input images and degenerate sampling Lastly, we investigated the impact of training with a variable number of input images and considering degenerate sampling configurations. By analyzing the networks trained under different conditions – including *variable* or *fixed* (16) number of input images and the use of *mixed* or *uniform* datasets – we sought to understand the network's resilience to fewer input frames and its ability to handle degenerate sampling. Our evaluations were conducted on the uniform, degenerate, and mixed test datasets, using different numbers of input images (5, 16, 30). The results in Table 4.2 demonstrated that training with a variable number of input images and a mixed dataset led to a network that exhibited greater resilience when faced with fewer input frames and could effectively handle degenerate sampling configurations.

Overall, our experimental findings highlight the significance of feature extraction within the DSA architecture. The presence of the encoder, along with the appropriate selection of the number of features, contributes to improved performance in terms of noise reduction and detail recovery in the SR image. Furthermore, training with a variable number of input images and considering degenerate sampling can enhance the network's robustness and adaptability in practical scenarios.

4.6 Chapter summary

In this chapter, 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.

In Chapter 5, we will enhance the detail preservation and introduce outlier management mechanisms, enriching the functionality of DSA. Chapter 6 will further explore extensions of DSA to multi-exposure sequences and improved temporal fusion, highlighting our ongoing endeavour to optimise the performance and utility of our self-supervised MISR architecture.

5 Adding detail-preserving control and outlier detection

Utilizing self-supervised training, we continue to advance the application of deep-learning methods for multi-image super-resolution in satellite imagery. In this chapter, we introduce two substantial enhancements to our previously proposed Deep Shift-and-Add (DSA) method. The first improvement extends the self-supervised loss of DSA, adding a spatially varying parameter that empowers users to strike a balance between detail preservation and noise reduction during testing. In the second improvement, we equip the DSA architecture with a module that handles outliers, such as those caused by dead pixels, reflections, or registration errors. These developments further strengthen the flexibility and adaptability of the DSA method in handling challenges in satellite multi-image super-resolution.

5.1 Introduction

As we have established earlier in this thesis, multi-image super-resolution (MISR) has recently emerged as an essential technique for enhancing the resolution of push-frame satellites [MSS+14, AEdFF20]. This approach leverages high framerate, low-resolution acquisitions to enable low-cost satellite constellations to compete effectively against traditional high-cost satellites.

Inherent noise in satellite imagery frequently presents complications [MSS+14]. This noise, a result of a myriad of factors such as system calibration errors, defective sensors, faulty channels, and various types of photonic and thermal noise, is generally viewed as an unwelcome interference and is consequently targeted for removal. However, the act of eliminating this noise invariably brings forth the possibility of also discarding critical detail, a risk we must attentively manage. As we have discussed in earlier chapters, our central objective within this research is to execute a robust joint super-resolution and denoising process from a series of satellite images, thereby carefully balancing the elimination of noise and preservation of vital detail.

The focus of our study is on push-frame satellites, particularly the SkySat constellation from Planet, which is capable of capturing burst of images for each scene. Despite the burgeoning attention that MISR for push-frame satellite images has received, the majority of these methods still rely heavily on classical model-based techniques [MSS+14, AEdFF20, AEF21]. For instance, Anger et al. [AEdFF20] propose solving a least-squares problem fitting spline polynomials to the observed samples, which they termed as ACT. Furthermore, Farsiu et al. [FREM04c] proposed a variational method that extends the

classic shift-and-add MISR to be robust to outliers, which amounts to applying pixel-wise medians.

Recently, there has been a surge of significant advancements in the field of deep learning-based methods for image and video processing. However, training a supervised MISR for remote sensing application is challenging due to the lack of ground truth high-resolution (HR) data. In light of these challenges, in the previous chapter (Chapter 4), we proposed a novel self-supervised deep learning approach named Deep Shift-and-Add (DSA) (Chapter 4). The innovative feature of DSA is its ability to train on satellite image bursts without the need for ground truth HR data. DSA peforms joint denoising and super-resolution and it can handle variable number of frames.

Notwithstanding its advantages, DSA also carries certain limitations. First, like many deep-learning-based image restoration algorithms, DSA tends to smooth out textures or details that have a lower contrast relative to the noise level [BM18]. This is a known problem for restoration methods based on minimizing a distortion measure (such as the MSE or L_1 loss) [BM18]. 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 [LTH+17, BM18]. However, this requires the network 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.

The second drawback of DSA is that it does not handle outliers, which are particularly prevalent and problematic in the realm of satellite imagery. Among these outliers, moving object misregistrations can create double imaging or ghosting artifacts due to discrepancies in the expected and actual positions of moving objects across different frames. Reflections, caused by sunlight bouncing off bodies of water or metallic surfaces, can generate unexpected intensities that may skew the SR results. Additionally, sensor issues such as dead pixels consistently generate incorrect readings, introducing additional noise into the dataset. These anomalies, if not accurately handled, can negatively affect the quality of the resulting SR images, underscoring an area of potential improvement for the DSA framework.

To address the limitations inherent in DSA, in this chapter we offer two significant contributions:

- A data-fitting term that allows for controlling the amount of detail in the solution via a spatially varying map provided to the network as input during testing. By doing so, users can manage the trade-off between noise reduction and detail preservation based on the specific application. This concept was inspired by a common photography trick to recover low contrast details lost to image denoising, which involves adding back to the output a fraction of the noisy input image.
- A more robust version of DSA achieved through the introduction of outlier masks computed by a neural network (see Figure 5.1). 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.

The first part of this chapter elucidates our proposed enhancements to DSA. We will delve into the mechanisms that enable adaptive noise reduction without compromising intricate

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

Figure 5.1: Super-resolution from a sequence of 15 real low-resolution SkySat L1A frames. (a) L1B from Planet, (b) DSA (Chapter 4), (c) Our improvement with an additional CNN to detect the outliers.

detail preservation and explain how we bolster the model's resilience against outliers in the satellite imagery.

The second part of the chapter is dedicated to empirical validation. Through rigorous tests and comparisons against existing methods, we will showcase the marked improvements our enhancements bring to the DSA framework.

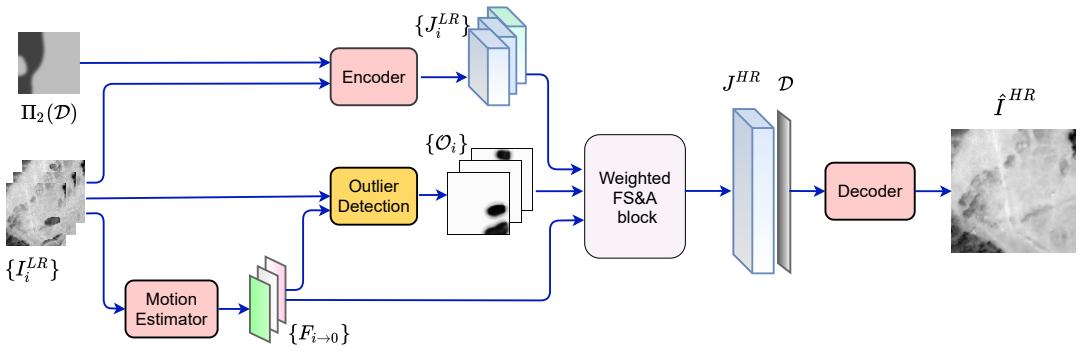
5.2 Proposed method

Our network (Figure 5.2) is built upon the Deep Shift-and-Add (DSA) architecture (Chapter 4) with four major modules: Motion Estimator, Encoder, Feature Shift-and-Add block (FS&A) and Decoder. The self-supervised DSA loss (Chapter 4) 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}

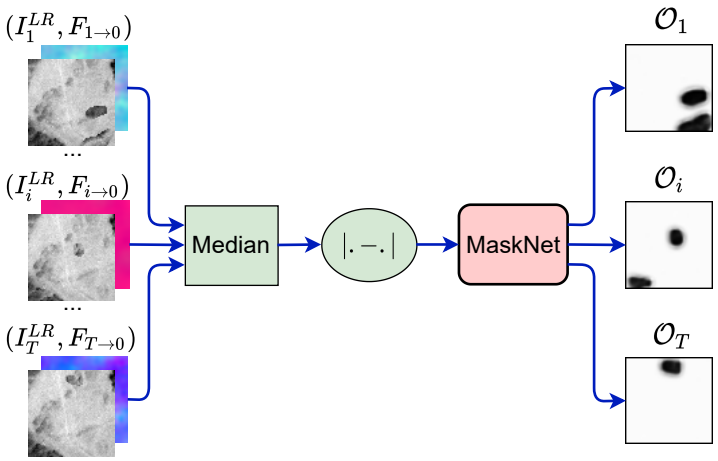
$$\text{DSA loss} = \left\| \Pi_2(\widehat{I}^{HR}) - I_0^{LR} \right\|_1, \tag{5.1}$$

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 [BM18].

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 $\mathcal{D}$ (Section 5.2.1). In addition we introduce in the DSA network architecture MaskNet, a new trainable module (Figure 5.2), whose purpose is to produce outlier masks $\mathcal{O}$ that indicate the presence of outliers (Section 5.2.2).



(a) Our architecture



(b) Outlier detection block

Figure 5.2: Overview of our method, which builds upon the Deep Shift-and-Add (DSA) architecture (Chapter 4).

5.2.1 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 [AEdFF20, FGS+95] when $p = 2$:

$$\text{LL}_p \text{ loss} = \frac{1}{T}\sum_{i=1}^{T} \left\| \Pi_2 \left(\text{Warp}(\hat{I}^{HR}, 2F_{i\to 0}) \right) - I_i^{LR} \right\|_p^p, \tag{5.2}$$

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 (Chapter 4) for details). This loss corresponds to the likelihood of the data under a generalized Gaussian noise model [FREM04c].

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 [FREM04c], 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 (5.1) and (5.2). 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

$$\text{loss} = \left\| \left(\Pi_2(\widehat{I}^{HR}) - I_0^{LR} \right) \cdot (1 - \Pi_2(\mathcal{D})) \right\|_1 + \frac{1}{T} \sum_{i=1}^{T} \left\| \left(\Pi_2(\text{Warp}(\widehat{I}^{HR}, 2F_{i\to 0})) - I_i^{LR} \right) \cdot \Pi_2(\mathcal{D}) \right\|_1, \quad (5.3)$$

where $\widehat{I}^{HR} = \mathbf{Net}(I^{LR}_{i=1,\dots,T}, \mathcal{D})$ 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 pre-aligned so that the $\mathcal{D}$ does not have to be warped in the loss.

5.2.2 Outlier handling

In DSA (Chapter 4), 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 [MHL+21] 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 [MHL+21] between the warped median image and each image is used as input for MaskNet

$$\mathcal{O}_i = \text{MaskNet}\left(|\,\text{Warp}(M^{LR}, F_{i\to 0}) - I_i^{LR}| \right). \quad (5.4)$$

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

$$J^{HR} = \frac{\sum_{i=1}^{T} \text{SPMC}(J_i^{LR} \cdot \mathcal{O}_i, F_{i\to 0})}{\sum_{i=1}^{T} \text{Max}(\text{SPMC}(\mathcal{O}_i, F_{i\to 0}), \epsilon)}, \quad (5.5)$$

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

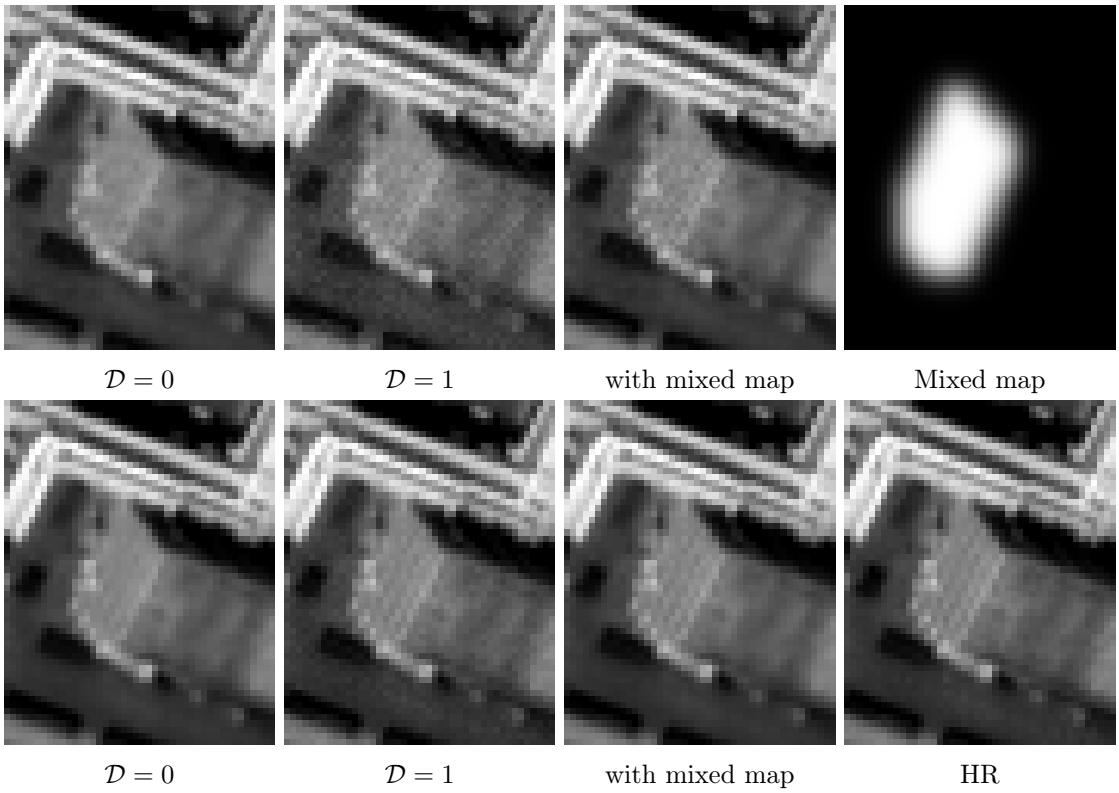


Figure 5.3: Super-resolution from a stack of 4 (first row) and 14 (second row) noisy synthetic images. From left to right: Reconstruction with $\mathcal{D} = 0$ (without detail preservation), with $\mathcal{D} = 1$, and with a mixed $\mathcal{D}$ map.

module [TGL+17, NAD+21b] 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.

5.3 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.

5.3.1 Examining detail preservation map $\mathcal{D}$

In order to train the network, we prepare sets of training input data $\{I^{LR}_{i=0,\dots,T}; \mathcal{D}\}$. 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 ($\sigma = 3$).

Figure 5.3 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

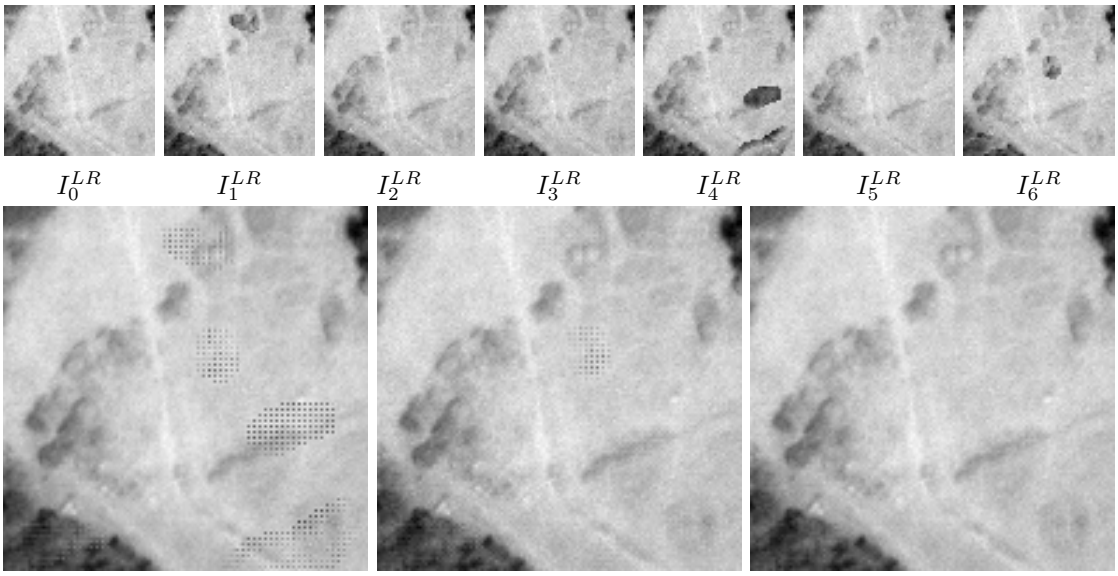


Figure 5.4: Effect of the architecture and the loss in robustness to outliers. First line: 7 LR images with outliers. Second line: Reconstruction with the LS or LAV loss, with and without MaskNet.

Table 5.1: 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

noise in uniform regions and preserve details in textured regions as shown in Figure 5.3. 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.

5.3.2 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 [FREM04c] 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 (Figure 5.4) 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 5.1 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. Figure 5.1 illustrates how our architecture with and without MaskNet handles moving objects. As expected, the motion estimator of L1B [MSS+14] 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.

5.4 Chapter summary

In this chapter, we presented an extension to the self-supervised DSA method (Chapter 4) 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.

We posited that outliers are elements inconsistently visible in the majority of frames. Nevertheless, there may be a need to preserve the content of the reference image, which might require additional modifications.

In the next chapter, we'll discuss extensions to multi-exposure sequences, highlighting a parallel evolution and optimization of our MISR system.

6 Extension to multi-exposure sequences and improved feature fusion

Modern Earth observation satellites capture multi-exposure bursts of push-frame images, offering a new frontier for computational super-resolution. This chapter builds on our previous research with the Deep Shift-and-Add (DSA) method and introduces a super-resolution approach designed explicitly for multi-exposure sequences—a problem relatively unexplored in the existing literature. Our proposed method not only handles signal-dependent noise effectively but also accommodates sequences of any length and compensates for inaccuracies in exposure times. Most significantly, it can be trained end-to-end in a self-supervised manner, negating the need for ground truth high-resolution frames, and is thus well-suited for real data applications. Key to our method are three critical contributions: i) a base-detail decomposition to handle exposure time errors, ii) a noise-level-aware feature encoding to enhance the fusion of frames with varying signal-to-noise ratio, and iii) a permutation invariant fusion strategy via temporal pooling operators. Evaluations on both synthetic and real data reveal that our method substantially outperforms existing single-exposure approaches when adapted to the multi-exposure scenario. This advancement marks a significant step forward in the realm of multi-exposure super-resolution processing of satellite imagery.

6.1 Introduction

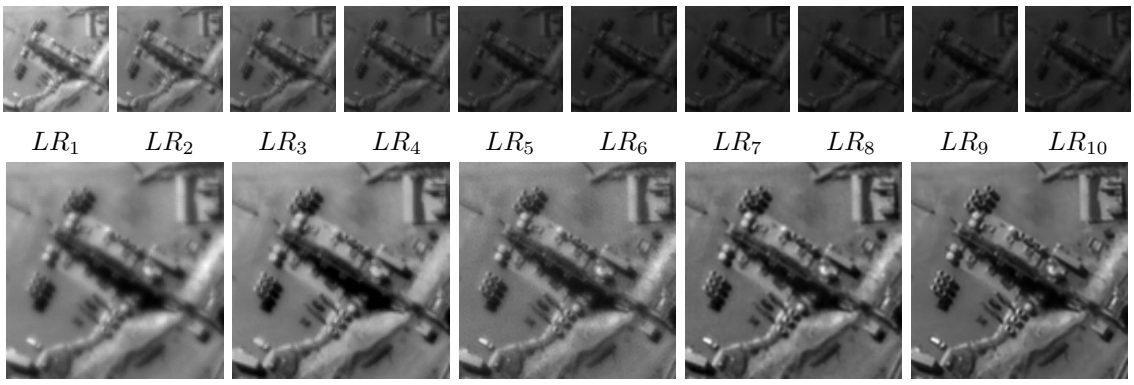


Figure 6.1: Super-resolution from a real multi-exposure sequence of 10 SkySat images. Top row: Original low resolution images with different exposures. Bottom row: Reconstructions from five methods, including ours trained with self-supervision (right).

High-resolution (HR) satellite imagery is vital for a plethora of applications, such as human activity monitoring and disaster relief. Recent trends in the remote sensing industry (Planet SkySat, Satellogic Aleph-1) point towards the adoption of computational super-resolution techniques [MSS+14, AEdFF20], enabling low-cost satellite constellations to compete effectively with their high-cost counterparts.

In order to capture the full dynamic range of the scene, some satellites use exposure bracketing, resulting in sequences with varying exposures. While several works have addressed multi-image super-resolution (MISR) of single-exposure sequences, almost no previous work considers the multi-exposure case.

MISR techniques utilize the aliasing present across multiple low-resolution (LR) captures to reconstruct a HR image, though the final resolution is inherently limited by the spectral decay of the system's blur kernel. On the other hand, frame aggregation offers significant noise reduction and, when working with bracketed exposures, the potential to create super-resolved high dynamic range (HDR) images. Long exposures have higher signal-to-noise ratio (SNR) which helps reduce the noise in dark regions, whereas short exposures provide information in bright regions which can cause saturation with longer exposure times.

In this chapter, we build on our previous work (Chapter 4), which presented the Deep Shift-and-Add (DSA) method—a self-supervised approach for MISR of single-exposure satellite image bursts. Our focus now turns to a more complex scenario: performing joint super-resolution and denoising from a time series of bracketed satellite images. While we concentrate on push-frame satellite sensors like the SkySat constellation from Planet, our technique is versatile and applies to a broad range of cameras capable of multi-exposure burst or video acquisition. We increase the resolution by a factor of two, which is the frequency cutoff of the combined optical and sensor's imaging system.

Several methods have addressed either MISR or HDR imaging from multiple exposures, but their combination has received little attention. Existing works consider an ideal setup in which frames can be aligned with an affinity [TA14, AEdFF20] or a homography [VSR18], and the number of acquisitions is large enough to render the problem an overdetermined system of equations. Such motion models are good approximations for satellite bursts, but ignore parallax [AEF21], which can be particularly prominent for mountains and tall buildings.

In the case of satellite imaging, push-frame cameras capable of capturing multi-exposure bursts are relatively recent, which explains why all previous works on MISR focus on the single-exposure case [MVFM20, DKG+20, AEdFF20, NAD+21b], except for SkySat's proprietary method [MSS+14] producing the L1B product, whose details are not public. While deep learning methods generally outperform traditional model-based approaches, they are often challenged by the need for large, realistic datasets with ground truth for training, since methods trained on synthetic data often fail to generalize to real images. Such remote sensing data are usually not available.

A promising direction is to use self-supervised learning techniques, which have been applied to video restoration tasks such as denoising and demosaicing [EDM+19, EDAF19, DAD+21, YPPJ20, SMV+21], and recently to MISR (Chapter 4). These techniques benefit from the temporal redundancy in videos. Instead of using ground truth labels, one of the degraded frames in the input sequence is withheld from the network and used as label.

In this chapter, we explore this path further, extending our DSA framework to handle multi-exposure bursts. By exploiting frame redundancy, we aim to perform joint MISR

and HDR processing of multi-exposure bursts, pushing the boundaries of what is achievable in the realm of satellite imagery processing.

Contributions. In this chapter, we propose *High Dynamic Range Deep Shift-and-Pool*, HDR-DSP a self-supervised method for joint super-resolution and denoising of bracketed satellite imagery. The method is able to handle time-series with a variable number of frames and is robust to errors in the exposure times, as the ones provided in the metadata are often inaccurate. This makes our method directly applicable to real image data (see Figure 6.1). This is, to the best of our knowledge, the first multi-exposure MISR method for satellite imaging, and beyond satellite imagery, it is the first approach based on deep-learning.

Our contributions are the following:

Feature Shift-and-Pool. We propose a *shift-and-pool* module that merges features (computed by an encoder network on each input LR frame) into a HR feature map by temporal pooling using permutation invariant statistics: average, maximum, and standard deviation. This gives a rich fused representation which yields a substantial improvement over the average [NAD+21b], in both single and multiple exposure cases.

Robustness to inaccurate exposure times via base-detail decomposition. We propose normalizing the input frames and decomposing them into base and detail. The errors caused by the inaccurate exposure times affect mainly the base, whereas the detail containing the aliasing required for super-resolution can be safely processed by the network. Note that vignetting and stray light can also cause exposure issues that affect single and multi-exposure MISR alike.

Noise-level-aware detail encodings. The noise present in the LR images is signal-dependent, its variance being an affine function of the intensity. To deal with such noise, we provide the un-normalized LR images to the encoder in addition to the normalized detail components. This gives the encoder information about the noise level of each pixel, necessary for an optimal fusion.

Self-supervised loss with grid shifting. We validate our contributions with an ablation study on a synthetic dataset (Section 6.5.3), designed to model the main characteristics of real bracketed SkySat sequences. Since there are no previous works on multi-exposure MISR, we compare against state-of-the-art single-exposure MISR methods which we adapt and retrain to (Section 6.5.4).

We also introduce a dataset of 2500 multi-exposure real SkySat bursts (Section 6.5.5). The dataset only consists of noisy LR images, but we can nevertheless train our network on it, since it is self-supervised. Both on synthetic and real data, the proposed HDR-DSP method attains the best results by a significant margin *even though it is trained without high resolution ground truth data.* The dataset is available on the project website.

6.2 Related work

Most works on video and burst super-resolution focus on the single-exposure case [LPM21, BDVGT21, SVB18, TGL+17, AEdFF20, DKG+20, MVFM19, NAD+21b]. The problem of super-resolution from multi-exposure sequences has received much less attention. In [TA14] it is modeled as an overdetermined system and solved via a non regularized least-squares approach. An affine motion model and exact knowledge of the exposure times are assumed.

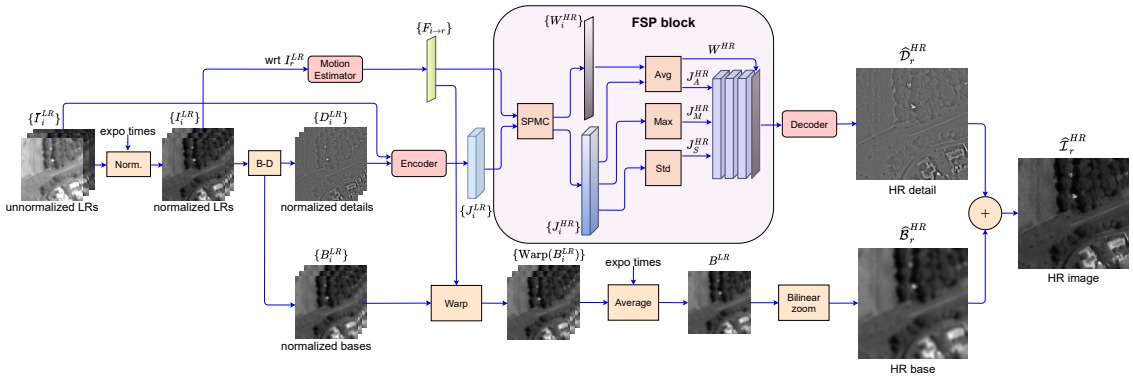


Figure 6.2: Overview of our proposed multi-exposure super-resolution network architecture HDR-DSP at inference time.

The authors in [VSR18] address the case in which the images have motion blur due to the camera shake. They also consider a static scene and do not consider noise. A related method for HDR imaging uses dual exposure sensors, which interlace two exposures in even and odd columns of the image [HST+14, ÇBM+20]. This can be seen as horizontally super-resolving the video.

Other works perform a related task: joint super-resolution and reverse tone-mapping [KK18, KOK19, KOK20]. The difference with our problem is that the input video is a single-exposure LR video, and the goal is to artificially increase its dynamic range to adapt it to HDR screens.

Methods for HDR imaging from multiple exposures need to deal with the noise. Granados et al. [GAW+10] address the case of signal-dependent noise and propose a fixed point iteration of the MLE estimator which is close to the Cramer-Rao bound [ADGM14]. In these works, the denoising comes only from the temporal fusion. In [ADGM13, AAD+17], this is incorporated in into spatio-temporal patch-based denoisers.

Our work can also be related to burst and video joint denoising and demosaicing [HSG+16, WGDE+19, EDAF19], as demosaicing can be regarded as a super-resolution problem.

6.3 Observation model

We denote by $\mathfrak{I}_t$ a dynamic infinite-resolution ideal scene. The camera on the satellite captures a sequence of m low resolution images $\bar{I}_i^{LR}$ with different exposures. For the $i-$th acquisition, the dynamic scene $\mathfrak{I}_t$ is integrated during an exposure time e_i centered at t_i. Even if satellites travel at a very high speed relative to the ground, precise electro-optical image stabilization systems (with piezo-electric actuators [KRV, KSD+20] or steering mirrors [RDL+]) assure that the observed scene $\mathfrak{I}_t$ is mostly constant during the exposure time (∼2ms), which allows us to approximate the temporal integration with a product in our observation model

$$\bar{I}_i^{LR} = e_i \Pi_1 \left(\mathfrak{I}_{t_i} * k \right) + n_i = e_i \mathcal{I}_i^{LR} + n_i. \tag{6.1}$$

Here k is the Point Spread Function (PSF) modeling jointly optical blur and pixel integration, Π_1 is the bi-dimensional sampling operator due to the sensor array, $\mathcal{I}_i^{LR}$ is the clean low-resolution image corresponding to an exposure of 1 unit of time and n_i denotes the noise. Throughout the text, calligraphic fonts $\mathcal{I}_i$ denote noise-free images and regular

fonts I_i noisy ones. A bar $\bar{I}_i = e_i I_i$ indicates that the image is multiplied by its exposure time (i.e. as it is acquired by the sensor), while its absence denotes images *normalized* to an exposure time of 1. We consider the r-th image $\bar{I}_r^{LR}$ in the time series as the *reference*, and without loss of generality we assume its exposure time to be one, $e_r = 1$.

We model the noise as spatially independent, additive Gaussian noise with zero mean and signal-dependent variance $n_i(x) \sim \mathcal{N}(0, \sigma^2(\bar{\mathcal{I}}_i^{LR}(x)))$, where

$$\sigma^2(\bar{\mathcal{I}}_i^{LR}(x)) = a e_i \mathcal{I}_i^{LR}(x) + b, \tag{6.2}$$

is an approximation of the Poisson shot noise plus Gaussian readout noise [PLZ+07, FTKE08], with parameters a and b.

Because of the spectral decay imposed by the pixel integration and optical blur (k), the images $\mathfrak{I}_{t_i} * k$ are band limited with a cutoff at about twice the sampling rate of the LR images for SkySat. *Our goal is to increase the resolution by a factor* 2 *by estimating* $\widehat{\mathcal{I}}_r^{HR}$*, a non-aliased sampling of* $\mathfrak{I}_{t_r} * k$ *from several LR observations* $\{\bar{I}_i^{LR}\}_{i=1}^m$ *with varying exposures* $\{e_i\}_{i=1}^m$. A sharp super-resolved image can then be recovered by partially deconvolving k.

In order for the method to be applicable in practice, it needs to handle time series with a variable number of frames m, and to be robust to inaccuracies in the exposure times e_i, as the exposure times in the image metadata are only a coarse approximation of the real ones.

6.4 Proposed method

Our method builds upon the DSA method for MISR introduced in Chapter 4, which can be regarded as a trainable generalization of the traditional shift-and-add (S&A) algorithms [FH02, MN07, GCLK08, ABHY00, Jia12]. A *feature S&A* is used to fuse feature representations produced from the LR images by an encoder network. A motion estimation network computes the optical flows between each input LR frame and the reference frame. The output of the feature S&A is a high-resolution aggregated feature map, which is then decoded by another network to produce the output image.

The DSA method could be extended to multi-exposure sequences by applying it to the normalized images $I_i^{LR} = \bar{I}_i^{LR}/e_i$. This approach however is sub-optimal because it neglects the fact that the normalization alters the noise variance model, and fails if the reported exposure times are inaccurate, which is the case in practice.

To better exploit multiple exposures, we propose two modifications: (1) A base-detail decomposition, which provides robustness to errors in the exposure times; (2) An encoding of the images that is made dependent on the noise variance, which allows the encoder to weight different contributions according to their signal-to-noise ratio. In addition, we also propose a new feature pooling fusion intended to capture a richer picture of the encoded features, leading to a substantial improvement in reconstruction quality, both for single and multiple exposure cases. The resulting network can be trained end-to-end with self-supervision, i.e. without requiring ground truth.

6.4.1 Architecture

Figure 6.2 shows a diagram of our proposed architecture which takes as input a sequence of multi-exposed LR images $\{\bar{I}_i^{LR}\}_{i=1}^m$ along with the corresponding exposure times e_i and

produces one super-resolved image $\widehat{\mathcal{I}}_r^{HR}$. The input LR images are first normalized to unit exposure time. The normalized LR images $\{I_i^{LR}\}_{i=1}^m$ are then decomposed into base $\{B_i^{LR}\}$ and detail $\{D_i^{LR}\}$ components. The bases contain the low frequencies. We align and average them to reduce the low frequency noise and upsample the result using bilinear zooming to produce the HR base component. The LR detail images are fed to a shared convolutional *Encoder* network that outputs a feature representation of each LR image. The features are then merged into a HR feature map by our *shift-and-pool* block (FSP), which aligns the LR features into the HR grid of the reference frame, and applies different pooling operations. The pooled features are then concatenated and fed to a *Decoder* CNN module that produces the HR detail image. The final HR image is obtained by adding the HR base and detail $\widehat{\mathcal{I}}_r^{HR} = \widehat{\mathcal{B}}_r^{HR} + \widehat{\mathcal{D}}_r^{HR}$.

The trainable modules of the proposed architecture (shown in red in Figure 6.2) include the Motion Estimator, the Encoder, and the Decoder.

Base-Detail decomposition. As mentioned above, normalizing a sequence of the frames $\bar{I}_i^{LR}$ by their reported exposures e_i does not result in stable intensity levels across the sequence. This can be due to small errors in e_i. However, uncorrected vignetting or stray light also contribute the same effect, even in single-exposure imagery.

The nature of the super-resolution task makes it very sensitive to these exposure fluctuations. The shift-and-add operation would merge the LR features into an incoherent high-resolution feature map, making the task of the decoder more difficult, resulting in loss of details or high-frequency artifacts (see Figure 6.3). Refining the initial e_i could limit this problem. But this entails its own challenges, especially if one also considers vignetting and stray light sources.

Instead, in this paper we propose a more robust and simple alternative, which is based on a base-detail decomposition [OABB85] of the normalized LR images defined as follows

$$B_i^{LR} = I_i^{LR} * G, \qquad D_i^{LR} = I_i^{LR} - B_i^{LR}, \tag{6.3}$$

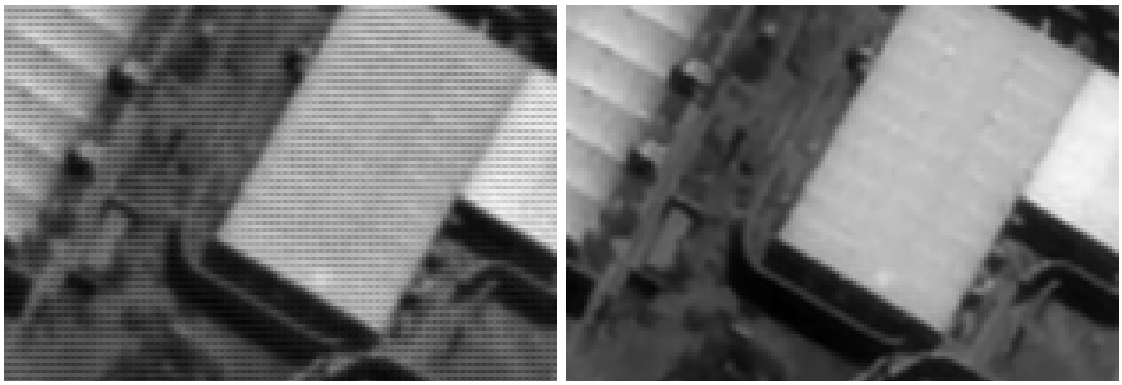
for $i = 1, \ldots, m$. Here G is a Gaussian kernel of standard deviation 1. We then process independently the details $\{D_i^{LR}\}$ and the bases $\{B_i^{LR}\}$ to produce the corresponding high resolution estimates $\widehat{\mathcal{D}}_r^{HR}$ and $\widehat{\mathcal{B}}_r^{HR}$. This decomposition is linear and does not affect the super-resolution since the alias is preserved in the detail components $\{D_i^{LR}\}$.

As the detail images span a smaller intensity range than the complete image I_i^{LR}, an error δ in the exposure time results in a small deviation in the detail and a large one in the base: $\delta\, B_i^{LR} + \delta\, D_i^{LR} = \delta\, I_i^{LR}$. The small error in the detail can be handled by a super-resolution method.

On the other hand, the base images do not need to be super-resolved, but still need to be denoised. In this chapter we propose a simple processing that aligns and averages the bases and upsamples the result. To fully exploit the high signal-to-noise ratio of longer exposures, the average is weighted by the exposure times e_i

$$B^{HR} = \text{Zoom}\left(\frac{\sum_i e_i \,\text{Warp}(B_i^{LR})}{\sum_i e_i}\right). \tag{6.4}$$

This weighting is an approximation of the ML estimator of Granados et al. [GAW+10] (details in the supplementary material).



DSA [NAD+21b] (without BD) Our HDR-DSP (with BD)

Figure 6.3: High frequency artifacts in a reconstruction from a real SkySat sequence (using DSA [NAD+21b]) with exposure time errors (left). HDR-DSP with the proposed base-detail (BD) decomposition does not present artifacts (right).

Base and detail decompositions have been used in super-resolution networks [KOK19, IJG+20b] to focus the network capacity on the details. In our case, the decomposition also provides robustness to errors in the radiometric normalization.

Motion Estimator. We follow the works of [SVB18, NAD+21b] to build a network (with the same hourglass architecture) that estimates the optical flows between the normalized LR frames $\{I_i^{LR}\}_{i=1}^m$ and the normalized reference frame I_r^{LR}

$$F_{i\to r} = \mathbf{MotionEst}(I_i^{LR}, I_r^{LR}; \Theta_{\mathbf{M}}) \in [-R, R]^{H\times W\times 2}, \tag{6.5}$$

where $\Theta_{\mathbf{M}}$ denotes the network parameters. A small Gaussian filter ($\sigma = 1$) is applied to the input images to reduce the alias [VSVV07, NAD+21b]. The network is trained with a maximum motion range of $[-R, R]^2$ (with $R = 5$ pixels). The training was adapted to better handle the noise difference due to the multi-exposure setting (see Section 6.4.2).

Noise-level-aware detail encodings. The Encoder module generates relevant features J_i^{LR} for each normalized LR detail image D_i^{LR} in the sequence

$$J_i^{LR} = \mathbf{Encoder}(D_i^{LR}, \bar{I}_i^{LR}; \Theta_{\mathbf{E}}) \in \mathbb{R}^{H\times W\times N}, \tag{6.6}$$

where $\Theta_{\mathbf{E}}$ is the set of parameters of the encoder and $N = 64$ is the number of produced features. The network architecture is detailed in the supplementary material.

The un-normalized low resolution frames $\bar{I}_i^{LR}$ are also fed to the encoder. This is motivated by the fact that the maximum likelihood fusion of noisy acquisitions into a (HDR) image is a weighted average, where the weights are the inverse of the noise variances [GAW+10, ADGM14]. In the proposed architecture, the normalized details D_i^{LR} are fused to produce a high resolution detail $\widehat{\mathcal{D}}_r^{HR}$. The noisy un-normalized images are unbiased estimators of an affine function of the noise variances $\sigma^2(\tilde{I}_i^{LR})/a - b/a$, thus they provide to the encoder the information required to compute the optimal fusion weights. The resulting features J_i^{LR} are then aggregated via a set of pooling operations, without any particular handling related to different source exposures.

Feature Pooling. We propose the Feature Shift-and-Pool block (FSP) which maps the LR features into their positions on the reference HR grid and pools them. First the features

are "splatted" bilinearly onto the HR grid by the SPMC module [TGL+17]. Each LR frame is upscaled by introducing zeros between samples and motion compensated following the flows $F_{i\to r}$. This is differentiable with respect to the intensities and the optical flows. Each splatted pixel is assigned a bilinear weight depending on the fractional part of its position in the HR grid. See [TGL+17, NAD+21b] for details.

This results in a set of aligned sparse HR feature maps

$$J_i^{HR} = \text{SPMC}(J_i^{LR}, \{F_{i\to r}\}) \in \mathbb{R}^{sH\times sW\times N}, \tag{6.7}$$

and the corresponding bilinear splatting weights $W_i^{HR} = \text{SPMC}(1, \{F_{i\to r}\})$. The upscaling factor s is set to 2.

As in Chapter 4, we use a weighted average pooling in the temporal direction (6.8). In addition, we propose computing the standard deviation and the max (6.9):

$$J_A^{HR} = (\sum_i J_i^{HR})(\sum_i W_i^{HR})^{-1}, \tag{6.8}$$

$$J_M^{HR} = \max_i J_i^{HR}, \qquad J_S^{HR} = \operatorname*{std}_i J_i^{HR}. \tag{6.9}$$

Note that this block does not have any trainable parameters, a trainable layer may attain a similar performance at a much higher computational cost (see the supplementary material).

These feature pooling operations render the architecture invariant to permutations of the input frames [AD18]. The key idea is that through end-to-end training, the encoder network will learn to output features for which the pooling is meaningful. Therefore, it is essential that the pooling operation is capable of passing all the necessary information to the decoder. Indeed, average pooling captures a consensus of the features, which amounts to a temporal denoising. But in aliased image sequences, it is common to come across features that are only visible in a single frame. Thus, the idea of the max-pooling operation is to preserve these unique features that would otherwise be lost in the average. The standard deviation pooling completes the picture by measuring the point-wise variability of the features.

The pooled features are independent of the number of processed frames. But this information is important as the decoder may interpret features resulting from aggregating many images differently than those resulting from just a few. For this reason, the aggregation weights $W^{HR} = \sum_i W_i^{HR}$ are also concatenated with the pooled features. As we will see in Section 6.5.3, incorporating W^{HR} improves the network ability to handle a variable number of input frames.

Decoder. The Decoder network reconstructs the HR detail image $\widehat{\mathcal{D}}_r^{HR}$ from the pooled features

$$\widehat{\mathcal{D}}_r^{HR} = \textbf{Decoder}(J_A^{HR}, J_M^{HR}, J_S^{HR}, W^{HR}; \Theta_{\mathbf{D}}) \in \mathbb{R}^{sH\times sW}, \tag{6.10}$$

where $\Theta_{\mathbf{D}}$ denotes the set of parameters of the decoder. The architecture is detailed in the supplementary material.

6.4.2 Self-supervised learning

To train the HDR-DSP detail fusion network, we adapt the fully self-supervised framework of DSA, which requires no ground truth HR images. During training, the LR frames

are randomly selected and for every sequence, one frame is set apart as the reference I_r^{LR}. Then, all the other LR images in each sequence are registered against the reference using the **MotionEst** network yielding the flows $F_{i\to r}$. The reference frame serves as the target for the self-supervised training similarly to noise-to-noise [LMH+18, EDM+19]. The procedure relies on the minimization of a reconstruction loss in the LR domain plus a motion estimation loss to ensure accurate alignment of the frames. The losses and the proposed adaptations are detailed in the following paragraphs.

Self-supervised SR loss. The self-supervised loss forces the network to produce an HR detail $\widehat{\mathcal{D}}_r^{HR}$ such that when subsampled, it coincides (modulo the noise) with the withheld target detail D_r^{LR}

$$\ell_{self}(\widehat{\mathcal{D}}_r^{HR}, D_r^{LR}) = \|\Pi_2(\widehat{\mathcal{D}}_r^{HR} * k) - D_r^{LR}\|_1, \tag{6.11}$$

where $\widehat{\mathcal{D}}_r^{HR} = \mathbf{Net}(\{D_i^{LR}\}_{i\neq r}, \{\bar{I}_i^{LR}\}_{i=1}^m)$ is the SR output, and Π_2 is the subsampling operator that takes one pixel over two in each direction. As in Chapter 4 we include the convolution kernel k in the loss. This forces the network to produce a deconvolved HR image that once convolved with k and subsampled matches the optical blur present in D_r^{LR}.

During training, the LR reference is only used in the motion estimator to compute the optical flows, but it is not fused into the HR result to avoid unwanted trivial solutions [BR19, DAD+21, NAD+21b]. At inference time we use the reference as this leads to improved results [NAD+21b].

Grid shifting. The self-supervised loss (6.11) downsamples the super-resolved detail to compare it with the reference LR detail. But since the downsampling is fixed, only the sampled positions intervene in the loss, which breaks the translation equivariance of the method. To avoid this issue, during training we augment the data by adding to the estimated optical flows a random shift of 0.5ϵ in each dimension ($\epsilon \in \{0, 1\}$). As a result, the super-resolved image is shifted by ϵ, which is easily compensated before computing the loss. This yields an improvement in PSNR of 0.2dB.

Motion estimation loss. The motion estimator is trained with unsupervised learning as in [YHD16]. The loss consists of a warping term and a regularization term. We observed that the optical flow is very sensitive to the intensity fluctuations between frames (as in our normalized LR frames I_i^{LR}), which result in imprecise alignments. To prevent this issue we compute the warping loss on the details rather than on the images, which is common in traditional optical flow [SVB18, LLKX19] . The loss is computed for each flow $F_{i\to r}$ estimated by the **MotionEst** module

$$\ell_{me}(\{F_{i\to r}\}_{i=1}^m) = \lambda_1 TV(F_{i\to r}) + \textstyle\sum_i \|\mathbf{Detail}\left(I_i^{LR} - \mathbf{Pullback}(I_r^{LR}, F_{i\to r})\right)\|_1, \tag{6.12}$$

where **Pullback** computes a bicubic warping of I_r^{LR} according to a flow, **Detail** applies a high-pass filter, TV is the finite difference discretization of the classic Total Variation regularizer [ROF92], and $\lambda_1 = 0.003$ is a hyperparameter controlling the regularization strength.

Training details. The self-supervised training of HDR-DSP is done in two stages. We first pretrain the motion estimator, then we train the entire system end-to-end.

Training the motion estimator for multi-exposure images poses a significant challenge. To ensure accurate flow estimations, we first pretrain the motion estimator using a simulated

dataset. The quality of the estimations is monitored by comparing them with ground truth flows until an average error of 0.05 pixels is achieved.

Initially, our approach was to employ the L_1 distance between the reference image and the radiometrically corrected warped image as the training objective for the motion estimator. However, this resulted in unacceptable flow estimations with errors exceeding 0.1 pixels. The imprecise alignments were primarily due to the sensitivity of motion estimation to intensity fluctuations between the normalized LR frames I_i^{LR}.

To address this issue, we compute the warping loss on the details rather than on the images, following the approach commonly employed in traditional optical flow methods [SVB18, LLKX19]. The loss is computed for each flow $F_{i\to r}$ estimated by the **MotionEst** module:

$$\ell_{me}(\{F_{i\to r}\}_{i=1}^{m}) = \sum_i \|\mathbf{Detail}(I_i^{LR}) - \mathbf{Detail}\left(\mathbf{Pullback}(I_r^{LR}, F_{i\to r})\right)\|_1 + \lambda_1 TV(F_{i\to r}), \tag{6.13}$$

where $\lambda_1 = 0.003$ is a hyperparameter controlling the regularization strength.

We set the batch size to 32 and use Adam [KB14] with the default Pytorch parameters and a initialized learning rate of 10^{-4} to optimize the loss. The pre-training converges after 50k iterations and takes about 3 hours on one NVIDIA V100 GPU.

In the next phase, we utilize the pretrained motion estimator and train the entire system end-to-end using the total loss defined in Eq (6.14):

$$\text{loss} = \ell_{self} + \lambda_2 \ell_{me}. \tag{6.14}$$

In our experiments, we set $\lambda_2 = 3$. Additionally, to mitigate boundary issues, the loss computation excludes values within a 2-pixel distance from the frame borders.

During training, LR crops of size 64×64 pixels are used, while validation is performed on LR images of size 256×256 pixels. The network receives a random number of LR input images ranging from 4 to 14 in each sequence. We employ a batch size of 16 and optimize the loss using the Adam optimizer with default parameters. The learning rates are initialized at 10^{-4} and scaled by 0.3 every 400 epochs. The training process completes in approximately 20 hours (1200 epochs) using a single NVIDIA V100 GPU.

6.5 Experiments

For our experiments, we use real multi-exposure push-frame images (L1A) acquired by SkySat satellites [MSS+14]. For the quantitative evaluations we also simulated a multi-exposure and a single-exposure datasets from L1B products (super-resolved products by Planet with a factor of 1.25).

6.5.1 Exposure error analysis

We observed a discrepancy between the reported exposure times provided by Planet and the correct normalization ratios. This discrepancy can be attributed to measurement imprecision, as the quantities involved are in the sub-millisecond range, or local illumination effects like vignetting. In order to estimate the accurate exposure ratio for a given pair of images, we employed phase correlation for image registration, masked saturated pixels, and computed the spatial median of the ratio between the two frames. Visual validation confirmed that these estimated exposure ratios were more precise than the reported exposure times, as they exhibited reduced flickering.

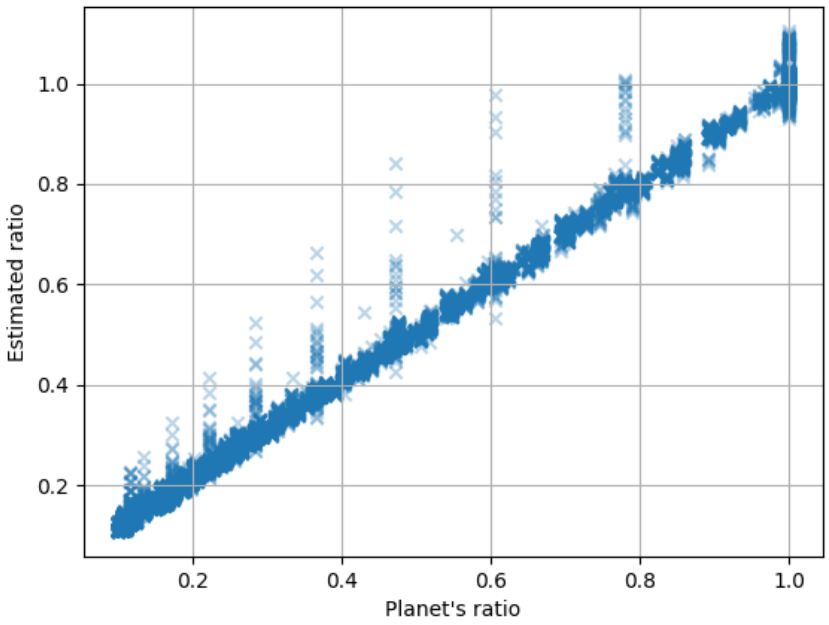


Figure 6.4: Normalized estimated exposure ratio with respect to provided exposure time.

Figure 6.4 illustrates the relationship between the reported ratio and the estimated ratio. We observed that errors typically ranged from a few percent, but occasionally larger errors were present. The nominal exposure times spanned from 0.4ms to 4.5ms. It is important to note that the absolute error in exposure time measurements is likely constant regardless of the exposure time itself. However, when computing the ratio of two exposures with errors, this can result in a substantial divergence of the ratio, especially if the exposure in the denominator is relatively short.

For the proposed super-resolution method, it is worth mentioning that we utilized the imprecise reported exposure times instead of the estimated ones. This decision was made due to the potential failure of the estimation method itself.

Additionally, it is important to clarify that the inaccuracy mentioned, such as a 20% discrepancy, indicates that the error measurements are approximately 20 with respect to the exposure time of the reference image (e_r is always set to 1). In our measurements, we have already determined the actual values of the exposure time and observed an error of up to ± 0.26 after normalization.

6.5.2 Simulated multi-exposure dataset

The two simulated datasets were generated from 1371 crops of L1B products (1096 train, 200 test, 75 val). First, we generate the noise-free LR images normalized to an exposure time of 1. Random subpixel translations of $\{\Delta_i\}_{i=1}^{m}$ are applied to the ground truth followed by $\times 2$ subsampling

$$\begin{aligned} \mathcal{I}_r^{LR} &= \Pi_2(\mathcal{I}^{HR}), \\ \mathcal{I}_i^{LR} &= \Pi_2(\mathrm{Shift}_{\Delta_i}(\mathcal{I}^{HR})), \qquad i \neq r \end{aligned} \tag{6.15}$$

where Π_2 is the subsampling operator. The exposure times are simulated as $e_i = \alpha^{c_i}$, where $c_i \in \{-5, .., 5\}$, and $\alpha = \mathrm{uniform}(1.2, 1.4)$. The noises $n_i = \sqrt{ae_i\mathcal{I}_i^{LR} + b\mathcal{N}(0,1)}$ are

Table 6.1: Handling of multi-exposure sequences with base-detail decomposition (BD) and using the un-normalized LR frames I_i^{LR} as an additional encoder input.

Methods (all HDR-DSP)	full	w/o BD	w/o BD (trained SE)	w/o LR
PSNR(dB) ME	**54.70**	53.76	52.91	53.94
PSNR(dB) SE	**54.72**	54.16	54.54	54.16

Table 6.2: Feature pooling choice. Using average (A), maximum (M), and standard deviation (S) pooling improves the results.

Features	AMS (HDR-DSP)	AS	AM	A
PSNR(dB) ME	**54.70**	54.46	54.44	54.17
PSNR(dB) SE	**54.72**	54.47	54.48	54.20

then added to all the un-normalized frames to produce the noisy multi-exposure sequence $\bar{I}_i^{LR} = e_i \mathcal{I}_i^{LR} + n_i$. The constants $a = 0.119, b = 12.050$ were estimated from real SkySat images with the Ponomarenko noise curve estimation method [CB13,PLZ+07]. The single-exposure dataset is generated in the same manner but with all $e_i = 1$. To simulate the exposure inaccuracies, during training and testing the e_i values are contaminated with noise within a range of 5%.

We use a PSNR score in our evaluation. The SkySat L1A images have a dynamic range of 12 bits, but we observed that the peak signal is at about 3400 DN. Therefore, our PSNR is normalized with a peak of 3400. We denote PSNR ME (resp. PSNR SE) as the average PSNRs computed on all the multi-exposure (resp. single-exposure) test sequences.

6.5.3 Ablation study

We study in Table 6.1 the importance of the base-detail decomposition. We consider simulated multi-exposure (ME) and single-exposure (SE) sequences presenting small exposure errors that match the ones observed in real sequences. If we train HDR-DSP without the proposed base-detail (w/o BD), the performance drops noticeably, which is also visible on real sequences (Figure 6.3). Even when training specifically for a single-exposure setting, as with DSA, the performance with base-detail is superior. In addition, we can see that removing the un-normalized LR frame from the encoder inputs (w/o LR) leads to a large performance drop for both single- and multi-exposure.

The experiment shown in Table 6.2 studies the impact of using multiple feature pooling strategies: average, maximum, and standard deviation. It shows that using the three greatly improves the results: about 0.5dB with respect to just using average. We observed that not including the average among the pooling strategies yields much worse results.

The aggregation weight feature W^{HR} was added to improve the handling by the decoder of sequences with variable number of input frames. The results in Table 6.3 confirm the importance of providing these weights. We also compare with networks trained for a fixed number of frames (HDR-DSP 4 and 14) and observe that in this case the performance drops even when testing for those specific configurations. We conclude that the weights become useless if the training does not consider a variable number of frames.

Lastly, removing the grid shifting (Section 6.4.2) from the training also reduces the PSNR

Table 6.3: Handling variable number of frames (PSNR ME (dB)).

Methods (all HDR-DSP)	full	w/o W^{HR}	HDR-DSP 4	HDR-DSP 14
4 frames	**52.81**	52.60	52.69	51.31
14 frames	**55.85**	55.59	54.26	55.53
variable n frames	**54.70**	54.45	53.85	54.07

Table 6.4: PSNR ME (dB) over the synthetic test set with 15 images in the case of 0%, 5% and 20% exposure time errors.

Methods	RAMS	ME S&A	HR-net	BD-ACT	DSA	HDR-DSP
0% exp. error	52.05	53.33	54.30	54.24	55.55	**56.00**
5% exp. error	51.84	52.43	54.22	54.23	54.99	**55.99**
20% exp. error	49.95	49.19	53.82	54.20	54.30	**55.90**

ME: from 54.70 to 54.49dB.

6.5.4 Comparison with the state-of-the-art

We compare our self-supervised network on the simulated dataset against state-of-the-art MISR methods for satellite images: *DSA* [NAD+21b], *HighRes-net* (HR-net) [DKG+20], *RAMS* [SMKC20], and *ACT* [AEdFF20]. A weighted *Shift-and-add* [MN07] with bicubic splatting adapted to multi-exposure sequences (ME S&A) serves as the baseline. HR-net and RAMS are two supervised networks designed to perform super-resolution of multi-temporal PROBA-V satellite images. In the context of push-frame satellites, we use the reference-aware version (Chapter 3) of HR-net and RAMS rather than the original approaches, as they achieve higher quality results. DSA and ACT are two state-of-the-art super-resolution methods for SkySat imagery. ACT also serves as a proxy for comparison with other interpolation-based methods from the literature [WGDE+19].

We adapt these methods to multi-exposure sequences. The deep learning approaches are fed with the normalized input images, whereas for ACT method we apply the same base-detail decomposition described in Section 6.4.1 and use ACT to restore the details (denoted BD-ACT). The registration step of ME S&A, BD-ACT, and RAMS are done with the inverse compositional algorithm [BM01, BFS18], which is robust to noise and brightness changes. The motion estimator of DSA is also trained with the loss on the details (Section 6.4.2).

Table 6.4 shows a quantitative comparison of the methods over the test set in the case of adding exposure time errors of 5% (as during training) and 20%. These errors are estimated from SkySat data (exposures ranging from 0.5 to 4.5 ms); see the supplementary material for details. Note that even with exact exposure times (row 0%), vignetting or stray light effects still justify the use of the proposed base-detail decomposition. Our self-supervised network ranks first in all cases with a significant gain of more than 1dB over all others (see Figure 6.5). Interestingly, the performance of most methods degrades quickly for large inaccuracy in exposure times. Only the methods using the base-detail decomposition (BD-ACT and ours) are robust to these inaccuracies. Note that HDR-DSP has never seen errors of 20% during training.

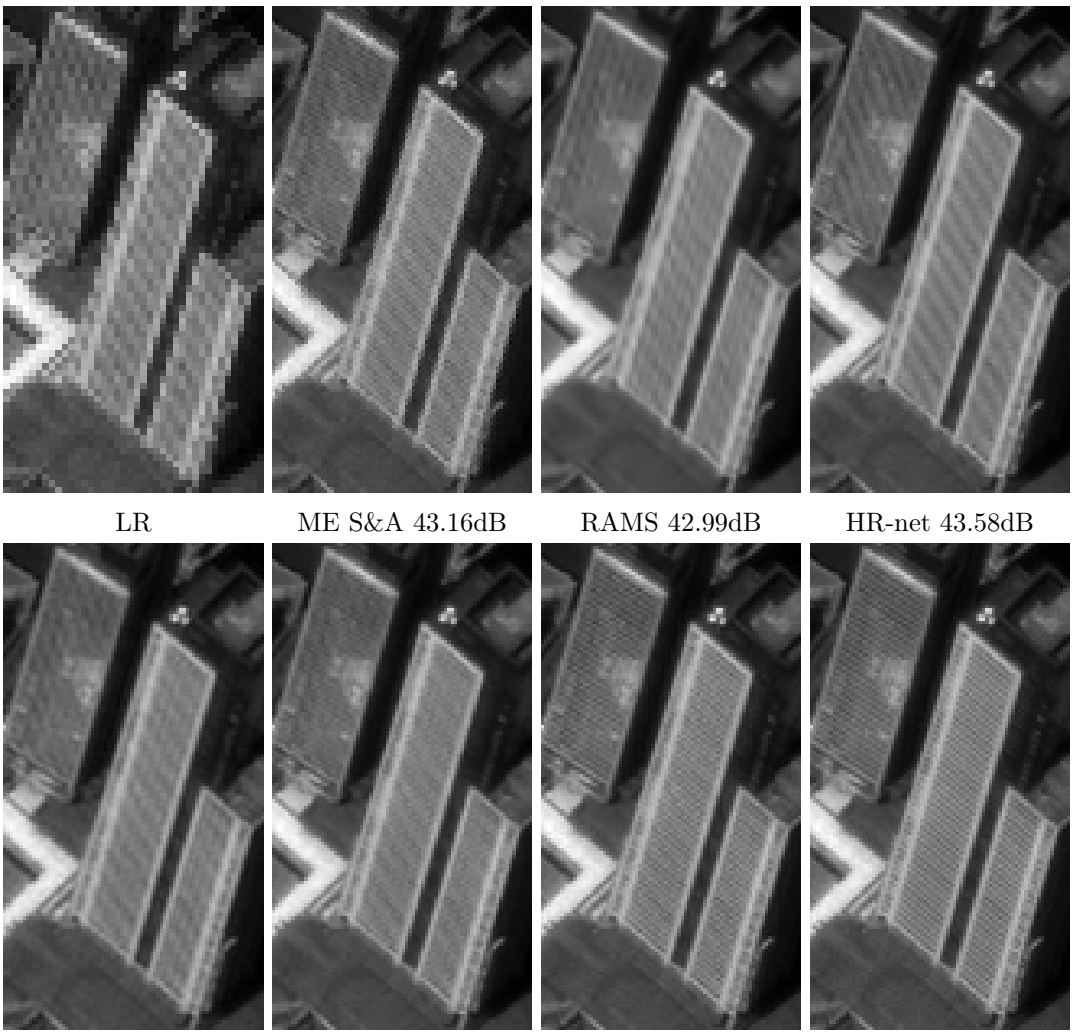


Figure 6.5: Super-resolution from a synthetic multi-exposure sequence (5% exp. error) of 15 aliased LR images. Methods are trained on a synthetic dataset and receive as inputs the normalized ME images except BD-ACT and HDR-DSP, which use the base-detail decomposition.

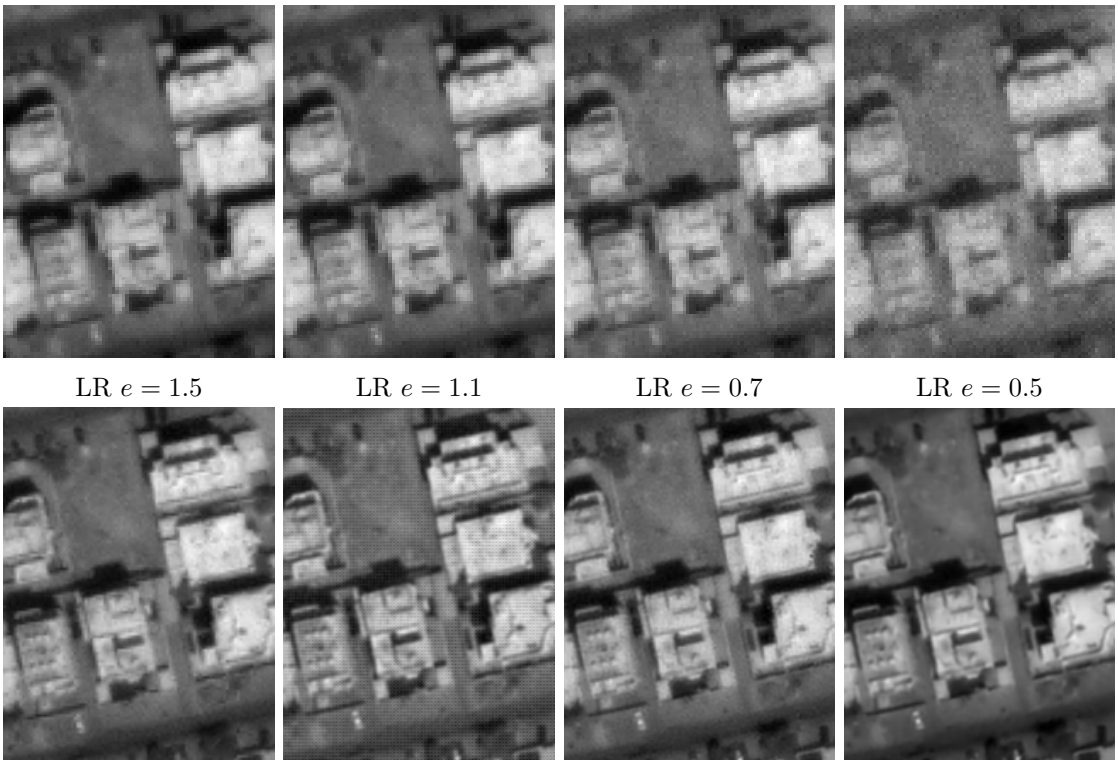


Figure 6.6: Super-resolution from a real multi-exposure sequence of 9 SkySat images. The first line corresponds to 4 normalized LR images in that sequence with different exposure times. The second line shows the reconstructions by Planet (L1B), DSA, BD-ACT and our method HDR-DSP .

Table 6.5: Execution time (s) on 200 sequences of size $15 \times 256 \times 256$ pixels.

Methods	RAMS	ME S&A	HR-net	BD-ACT	DSA	HDR-DSP
Time (s)	93	276	22	555	82	97

6.5.5 Results on real data

The proposed self-supervised training allows to train HDR-DSP on real multi-exposure sequences taken from SkySat satellites. From the L1A product of Planet SkySat, we extracted 2500 sequences (128×128 pixels) pre-registered up-to an integer translation. Out of 2500 sequences, 300 are used for testing. Each sequence contains from 4 to 15 frames. In about 75% of the sequences the exposure time varies within each sequence and we used the exposure time information provided in the metadata.

Figure 6.6 compares HDR-DSP against Planet L1B, DSA, and BD-ACT. The top row shows four normalized frames of the sequence, where we can notice the dependence of the noise level on the exposure time. The method used in the Planet L1B product is unknown. It super-resolves by a factor of 1.25 but contains noticeable artifacts and lacks fine details. The result from DSA exhibits a high-frequency pattern due to the imprecise exposure times. BD-ACT is able to cope with the exposure changes thanks to the base-detail decomposition, but the result is still very noisy. In contrast, HDR-DSP shows a clean and detailed reconstruction.

Figure 6.1 also shows a multi-exposure LR sequence along with the results from ME S&A, Planet L1B, ACT, DSA and HDR-DSP. Comparing HDR-DSP with DSA, we see that the former provides a cleaner result thanks to the base-detail decomposition and the proposed improvements over the DSA architecture and training procedure, which is also observed in the synthetic experiments.

6.5.6 Execution time

Table 6.5 presents the execution times of the methods evaluated on the synthetic multi-exposure dataset. Among the methods, HighRes-net demonstrates the fastest performance owing to its convolutional architecture. Comparatively, HDR-DSP is slightly more computationally expensive than DSA due to its feature pooling process and the requirement to fuse the bases together. ME S&A and BD-ACT are both executed on the CPU, with the latter incurring higher computational costs due to the linear spline system inversion it performs.

6.6 Chapter summary

The proposed HDR-DSP method is able to reconstruct high-quality results from multi-exposure bursts, providing fine details, low-noise, and high dynamic range. The proposed base-detail processing allows robustness to errors in the exposure time that are common in practice. In addition, a significant performance improvement is obtained by making the image encoding dependent on the noise variance, and using a new feature pooling designed to capture richer representations. Thanks to its fully self-supervised training, the method requires no ground truth and can thus be applied on real data. We show its effectiveness by training a model that super-resolves multi-exposure SkySat L1A acquisitions, leading to a substantial resolution gain with respect to the state-of-the-art.

However, it's important to acknowledge limitations within this remote sensing context. For instance, our assumptions do not cater to the challenging photon-limited noise regime, nor do they fully handle complexities related to motion and occlusions. Additionally, the method does not currently address saturation. These points will require further investigation.

The journey we've embarked on through MISR has prepared us for the subsequent exploration into SISR, a domain that exhibits a distinct, yet equally intricate, set of challenges. It's important to note that satellite multi-spectral SISR shares similarities with MISR, considering that each spectral band views the scene from a unique angle, hence capturing additional information. As we progress to the next part of the thesis, these commonalities will become increasingly evident, particularly in Chapter 9. There, we develop a self-supervised SISR method specifically for Sentinel-2 L1B data. The inspiration for this method comes from the DSA MISR framework discussed in Chapter 4, demonstrating the continuity in our super-resolution research journey and the interconnectedness of the concepts within.

6.7 Appendix

6.7.1 Weights for the base fusion

Since the base component only contains low frequencies and cannot be super-resolved, we propose a simple pipeline consisting of *i)* alignment of the LR base components B_i to the reference, *ii)* temporal fusion via weighted average to attenuate noise, *iii)* upscaling using bilinear interpolation. For the temporal fusion the weights in the weighted average are simply the exposure times:

$$B^{LR}(x) = \frac{\sum_i e_i \text{Warp}(B_i^{LR}(x))}{\sum_i e_i} \tag{6.16}$$

In this section we will provide a justification for this choice, which is based on two approximations.

Approximate noise model for the base. The base results from the convolution with a Gaussian kernel G. At pixel x we have

$$B_i^{LR}(x) = \sum_h G(h) I_i^{LR}(x+h).$$

Assuming the signal-dependent Gaussian noise model of Eq (6.2), we have that $B_i^{LR}(x)$ also follows a Gaussian distribution with the following mean and variance:

$$\begin{aligned}
\mathbb{E}\{B_i^{LR}(x)\} &= \sum_h G(h) \mathcal{I}_i^{LR}(x+h) \\
\mathbb{V}\{B_i^{LR}(x)\} &= \frac{a}{e_i} \sum_h G^2(h) \mathcal{I}_i^{LR}(x+h) + \frac{b}{e_i^2} \sum_h G^2(h).
\end{aligned}$$

We are going to assume that the clean LR image $\mathcal{I}_i^{LR}$ varies smoothly in the filter support, and thus

$$\mathbb{E}\{B_i^{LR}(x)\} \approx \mathcal{I}_i^{LR}(x), \quad \mathbb{V}\{B_i^{LR}(x)\} \approx \frac{\alpha e_i \mathcal{I}_i^{LR}(x) + \beta}{e_i^2}. \tag{6.17}$$

where $\alpha = a\sum_h G^2(h)$ and $\beta = b\sum_h G^2(h)$. This rough approximation allows us to use a signal-dependent Gaussian noise model like (6.2). The approximation is only valid in regions where the image is smooth (away from edges, textures, etc.). However, these are the regions in which we are mainly interested, since it is where the low frequency noise present in the base becomes more noticeable.

Approximate MLE estimator for the weights. After alignment, for a given pixel x we have different values acquired with varying exposure times, which we are going to denote as $z_i = \text{Warp}(B_i^{LR}(x))$ to simplify notation. We also have the corresponding clean LR base images $\mathcal{B}_i^{LR}$, and we are going to assume that they coincide after alignment, i.e. $y = \text{Warp}(\mathcal{B}_i^{LR})(x)$ for $i = 1, ..., m$. We would like to estimate y from the series of observations

$$z_i \sim \mathcal{N}\left(y, \sigma_i^2(y)\right), \quad \sigma_i^2(y) = \frac{\alpha e_i y + \beta}{e_i^2}.$$

This problem occurs in HDR imaging, when estimating the unknown irradiance given noisy acquisitions with varying exposure times [GAW+10, ADGM14]. Each z_i is an unbiased estimator of y. Therefore, if the variances were known, we can minimize the MSE with the following weighted average, where the weights are the inverse of the variances:

$$\hat{y} = \frac{\sum_i w_i z_i}{\sum_i w_i}, \quad w_i = \frac{e_i^2}{\alpha e_i y + \beta}. \tag{6.18}$$

The problem is that the weights depend on the unknown y. In [GAW+10] Granados et al. solve this problem with an iterative weighted average:

$$w_i^0 = \frac{e_i^2}{\alpha e_i z_i + \beta}.$$
$$w_i^k = \frac{e_i^2}{\alpha e_i \hat{y}^k + \beta}, \quad \hat{y}^{k+1} = \frac{\sum_i w_i^k z_i}{\sum_i w_i^k}, \quad k = 1, 2, ...$$

It can be shown that this converges to the maximum likelihood estimate.

In our case, we are going to simplify expression (6.18) by assuming that $\alpha e_i y \gg \beta$, and therefore $w_i \approx \frac{e_i}{\alpha y}$. Under this assumption, we obtain

$$\hat{y} = \frac{\sum_i e_i z_i}{\sum_i e_i}. \tag{6.19}$$

This assumption holds for brighter pixels and well exposed images [ADGM14].

6.7.2 Alternative exposure weighting strategies

As discussed in the main paper, the LRs with longer exposure time should contribute more to the reconstruction because of their high signal-to-noise ratio. In our proposed method, we use the un-normalized LR images as additional input to the Encoder so as that the Encoder perceives the noise level in each LR image. Subsequently, the Encoder can decide which features are more important.

We also evaluated an alternative strategy to weight the features (WF) based on the exposure times. This simply consists in weighting the features J_i^{LR} by the corresponding exposure time in the SPMC module. Actually, this was inspired from the ME S&A method.

This strategy leads to slightly worse yet adequate feature encodings (-0.08dB) as shown in Table 6.6. Moreover, using both feature weighting and LRs encoding (third column) leads to the same performance as only using LRs encoding. This implies that the Encoder already encodes the necessary information about the signal-dependent noise on the features.

Table 6.6: Handling of the signal-dependent noise

Methods	HDR-DSP	DSP (+WF - LR)	DSP (+WF)
PSNR(dB) ME	**54.70**	54.62	54.70

6.7.3 Adaptation of existing methods to multi-exposure sequences

We detail here the adaptations to the algorithms we used in the comparisons.

ME S&A. *Multi-exposure Shift-and-add* is a weighted version of the classic shift-and-add method [FH02,MN07,GCLK08,Jia12] designed for multi-exposure sequences. Usually, S&A produces the HR image by registering the LR images onto the HR grid using the corresponding optical flows. After the registration step, the intensities of the LR images are splatted to the neighborhood integer-coordinate pixels using some kernel interpolation. Finally, pixel-wised aggregation is done to obtain the HR output image. Therefore a naive method consists of using the classic S&A method on the normalized LR images. However this ignores the different signal-to-noise ratios in the normalized images and fails to greatly reduce the noise. Using the same arguments as in the Section 6.7.1, we propose the weighted S&A for multi-exposure sequence

$$\widehat{I}^{HR} = \frac{\sum_{i=1}^{m} \mathbf{Register}(\bar{I}_i^{LR})}{\sum_{i=1}^{m} e_i}, \tag{6.20}$$

where **Register** maps and splats the un-normalized images $\bar{I}_i^{LR}$ onto the HR grid.

Base-detail ACT (BD ACT). ACT [AEdFF20] is a traditional multi-image super-resolution method developed for Planet SkySat single-exposure sequences. It formulates the reconstruction as an inverse problem and solves it by an iterative optimization method. BD ACT extends ACT to support multi-exposure images by adopting the same base-detail strategy as proposed in HDR-DSP : the details of the images are fused by ACT, and the base is reconstructed by the upsampled average of the bases of the input images.

HighRes-net (HR-net) and RAMS. HighRes-net [DKG+20] and RAMS [SMKC20] are two super-resolution methods for multi-temporal PROBA-V satellite images. However in the PROBA-V dataset, the identity of the LR reference image is unavailable. This hinders the true potential of the methods trained on this dataset. As a result we use the reference-aware super-resolution (Chapter 3) of HighRes-net and RAMS. In HighRes-net, the reference image is used as a shared representation for all LR images. Each LR image is embedded jointly with this reference before being recursively fused. In RAMS, each LR image is aligned to the reference image before being input to the residual attention block. The registration step of RAMS is done with inverse compositional algorithm [BFS18], which is robust to noise and brightness change. As HighRes-net and RAMS are supervised methods, we also use a radiometric correction on the output before computing the loss [DKG+20].

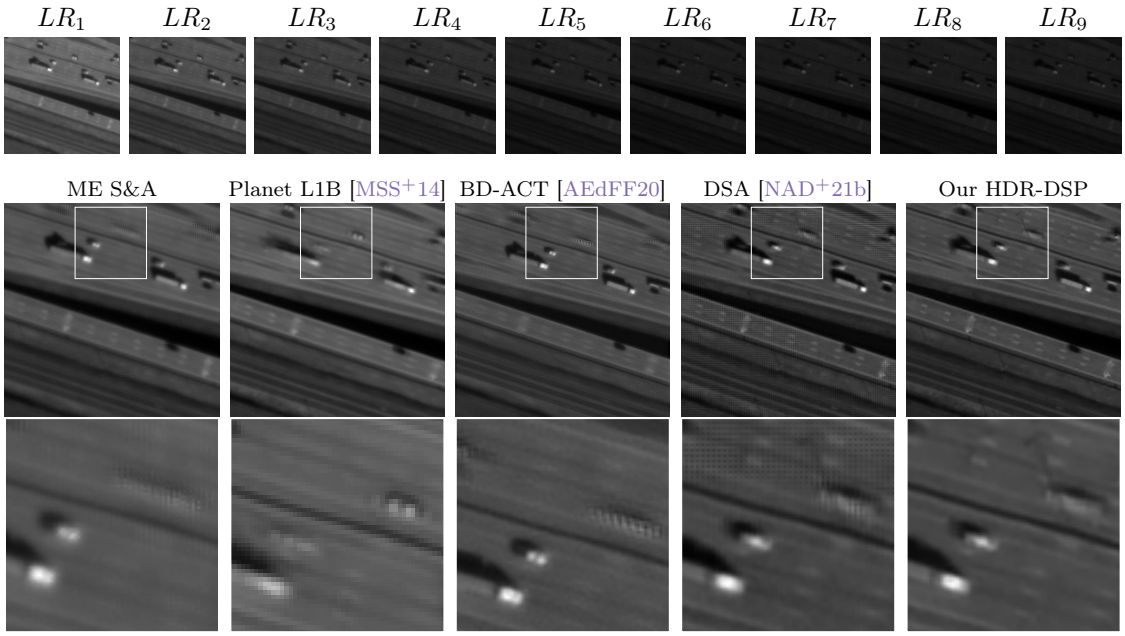


Figure 6.7: Super-resolution from a real multi-exposure sequence of 9 SkySat images. Top row: Original low resolution images with different exposures. Middle row: Reconstructions from five methods, including ours trained with self-supervision (right). Bottom row: Zoom on a detail of the results.

DSA. *Deep shift-and-add* [NAD+21b] DSA is a self-supervised method for super-resolution of push-frame single-exposure satellite images. We adapt DSA to multi-exposure case by using the normalized LR images as input. We also use the loss on the details to train the motion estimator in DSA.

6.7.4 Additional comparisons using real SkySat sequences

Figure 6.7 presents results obtained on real multi-exposure SkySat images using 9 frames. This is a challenging sequence as it contains moving vehicles. Note how the road markings are better seen in the HDR-DSP result. However, since HDR-DSP does not account for moving objects (the motion estimator only predicts smooth motion within a range of 5 pixels) the cars are blurry.

Figure 6.8 shows another example of reconstruction on a real sequence of 7 SkySat images. Even though there are only 7 images in this sequence and most of them are very noisy, HDR-DSP is able to produce a clean image. The fine details are well restored.

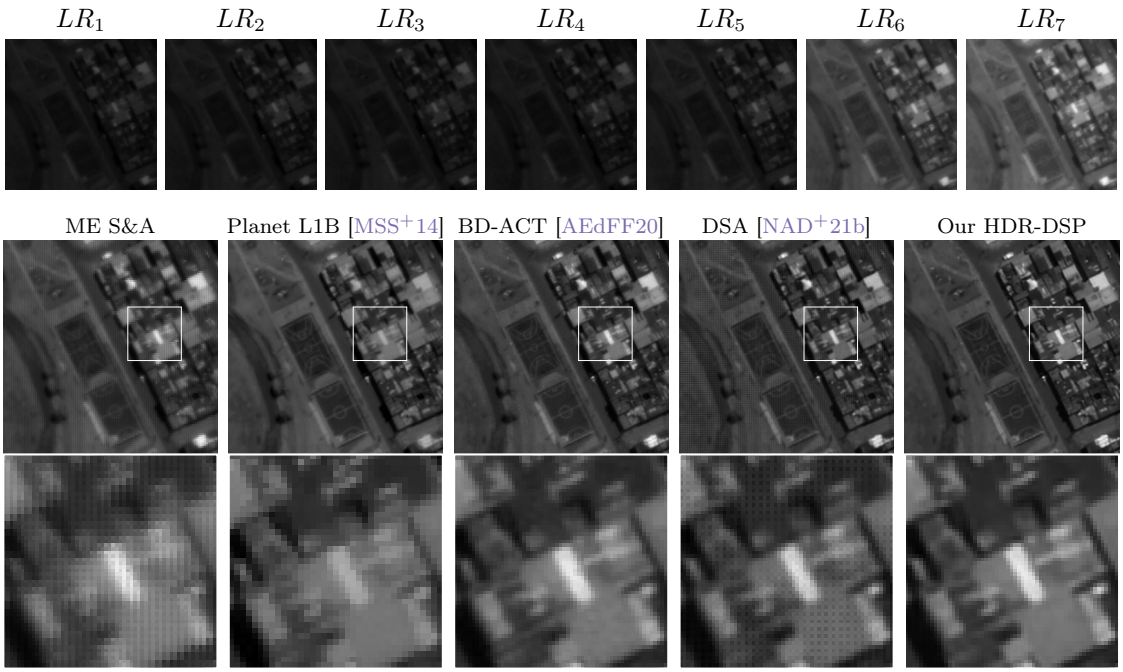


Figure 6.8: Super-resolution from a real multi-exposure sequence of 7 SkySat images. Top row: Original low resolution images with different exposures. Middle row: Reconstructions from five methods, including ours trained with self-supervision (right). Bottom row: Zoom on a detail of the results.

Part II

Single-image super-resolution in satellite imagery

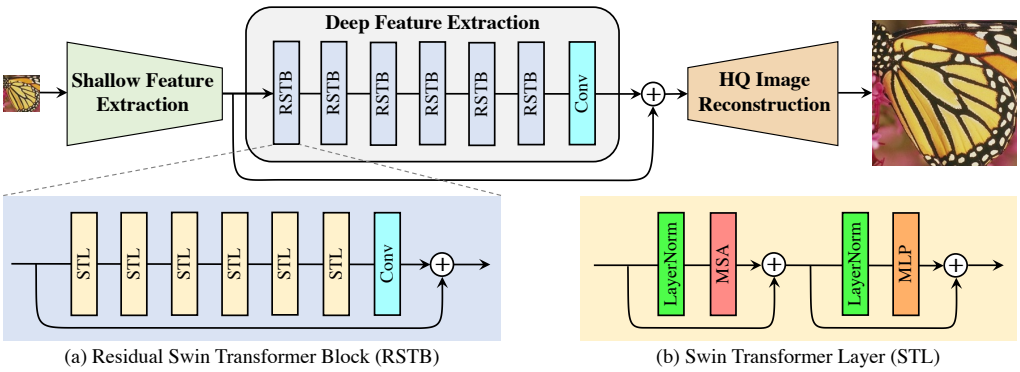
7 A brief analysis of the SwinIR super-resolution method

SwinIR, utilizing the innovative Swin Transformer architecture, presents a significant advancement in the domain of image restoration. Unlike traditional convolutional neural networks, SwinIR's capacity to capture intricate relationships between image patches results in exceptional outcomes. This chapter studies the application of SwinIR for single-image super-resolution, scrutinizing the distinct characteristics of its architecture. Furthermore, our discussion extends beyond theoretical exposition. To put SwinIR's efficacy into perspective, we conduct rigorous experiments on satellite images. We contrast its performance against other prevalent deep learning methodologies, offering an insightful comparison that highlights the potential and the limitation of SwinIR in the realm of super-resolution in satellite imagery.

7.1 Introduction

Single image super-resolution (SISR) is a fundamental problem in computer vision that aims to obtain a high resolution (HR) output from its degraded low-resolution (LR) counterpart. Recently, deep-learning methods have outperformed traditional SISR algorithms by a huge margin in both quantitative and qualitative results. As a matter of fact, SISR can be seen as an interpolation problem since SISR tries to recover pixels in the HR from their neighboring pixels in the LR image. Being a local problem, SISR has been dominated by convolutional neural networks (CNN). In particular, [ZZGZ17] uses dilated convolution to increase the receptive field over twice and get better result. RDN [ZTK+18] proposes a residual dense network to exploit the hierarchical features from the convolutional layers. RCAN [ZLL+18] adds an attention mechanism inside the CNN framework to exploit better feature representation produced by the channels.

On the other side of deep-learning, Transformer is the backbone of natural language processing (NLP). Since its invention in 2017, Transformer with its powerful self-attention mechanism has refreshed and dominated all modern architectures in NLP. The question we all wanted to ask is whether Transformer could be applicable to computer vision. One naive approach is to consider image pixels as tokens and put all of them into the self-attention mechanism. However, this approach is intractable due to enormous amount of pixels in natural images. To this aim, Dosovitskiy et al. [DBK+20] introduce the Vision Transformer (ViT) which applies Transformer directly on non-overlapping image patches. Achieving state-of-the-art performance in image classification, ViT is very promising in computer vision. Notwithstanding its great potential, the limitation of the ViT resides in its quadratic computational complexity on image size, which makes it unscalable to higher-



(a) Residual Swin Transformer Block (RSTB) (b) Swin Transformer Layer (STL)

Figure 7.1: SwinIR architecture

resolution images. Another related work IPT [CWG+21] uses a pretrained Transformer model to perform image processing tasks. Like other Vision Transformer-based models, IPT is computationally intensive and requires a large training dataset. Liu et al. [LLC+21] propose the Swin Transformer to overcome the main drawbacks of the Vision Transformer and achieve state-of-the-art results in image classification, object detection, and semantic segmentation. The Swin Transformer alleviates the computational burden of the ViT by computing self-attention only locally, but also models long-range dependency by using the shifted window scheme. The Swin Transformer is used in many state-of-the-art super-resolution methods, including stereo image super-resolution [JWY+22] and burst raw super-resolution [LLC+22].

Recently, Liang et al. [LCS+21] proposed SwinIR, an excellent baseline for image restoration based on the Swin Transformer. SwinIR is actually a hybrid model with two CNN modules (shallow feature extraction and high-quality image reconstruction) at the two ends, and specially a Swin Transformer-based module (deep feature extraction) as the crucial component of the method. SwinIR is proven to achieve state-of-the-art performance on single image super-resolution, image denoising, and JPEG artifact removal with a reasonable number of parameters.

In this study, we not only analyze the performance of SwinIR on SISR, investigating the role of long-range information captured by the Transformer in addressing local problems, but also extend our exploration to the application of SwinIR for satellite imagery. We scrutinize its potential and limitations within this unique context, illuminating its viability and areas for potential enhancement when deployed for earth observation data.

7.2 Method

7.2.1 SwinIR architecture

As shown in Figure 7.1, SwinIR has a hybrid architecture consisting of three modules: shallow feature extraction (CNN), deep feature extraction (Swin Transformer), and high quality image reconstruction (CNN). In this project, we focus on two SwinIR networks: *classical SR* and *realistic SR*. The classical model is a medium-sized network designed and trained for quantitative measurement. The realistic model is larger and is trained to perform real-world super-resolution.

Shallow feature extraction The shallow feature extraction can be considered a preprocessing step, which serves to map the LR image $I_{LR} \in \mathbb{R}^{H\times W\times 3}$ to a richer dimensional feature space with C feature channels. The shallow feature extraction is a convolutional layer H_{SF} with kernel size 3×3

$$F_0 = H_{SF}(I_{LR}), \tag{7.1}$$

where $F_0 \in \mathbb{R}^{H\times W\times C}$ is the shallow extracted feature. Applying an early small convolutional layer at the beginning of the Vision Transformer was reported to help the training to stabilize and converge faster [XSM+21]. The embedded dimension C is set to 180 for the classical model and 240 for the realistic model.

Deep feature extraction The deep feature extraction, composed of K residual Swin Transformer blocks (RSTB) and a CNN, comes after the shallow layer H_{SF}. K is set to 6 in the classical model and 9 in the realistic model. Concretely, first these blocks RSTB compute the transitional features $F_1, F_2, \ldots, F_K$ sequentially

$$F_i = H_{RSTB_i}(F_{i-1}), i = 1, 2, \ldots, K, \tag{7.2}$$

where H_{RSTB_i} denotes the $i-$th RSTB. And then a small CNN H_{CONV} at the end extracts the output deep feature F_{DF}

$$F_{DF} = H_{CONV}(F_K). \tag{7.3}$$

This CNN is presumed to introduce the image domain-specific inductive biases into the Transformer. The CNN in the classical model is just a simple convolutional layer that keeps the embedded dimension $C = 180$. For the realistic model, it is an hourglass-shaped CNN with 3 convolutional layers and hidden dimension 60 in order to save parameters and memory.

High resolution image reconstruction Finally, the reconstruction module H_{REC} produces the high resolution output from the computed shallow and deep features,

$$I_{HR} = H_{REC}(F_0 + F_{DF}). \tag{7.4}$$

The shallow features and the deep features contain mainly the low-frequency and the high-frequency information, respectively. While the former is pretty simple to extract with a convolutional layer, the latter is much more sophisticated to reconstruct. Hence, a long skip connection from F_0 up to F_{DF} is used to help the deep feature extraction focus on recovering the high frequency details. The reconstruction module H_{REC} is built from an upsample operator (pixel shuffle [SCH+16] in the classical model, and nearest-neighbor interpolation in the realistic model) and several convolution layers.

7.2.2 Residual Swin Transformer block

Each residual Swin Transformer block (RSTB) is composed of L ($L = 6$ for the two models) Swin Transformer Layers (STL) followed by a CNN (Figure 7.1a). More specifically, given the input feature $F_{i,0}$ of the $i-$th RSTB, the intermediate features $F_{i,1}, \ldots, F_{i,L}$ by L STL are computed as

$$F_{i,j} = H_{STL_{i,j}}(F_{i,j-1}), j = 1, 2, \ldots, L, \tag{7.5}$$

where $H_{STL_{i,j}}$ is the $j-$th STL of the $i-$th RSTB. Then a CNN is applied to enhance the translation equivariance of the Swin Transformer just before the residual connection

$$F_{i,out} = H_{CONV_i}(F_{i,L}) + F_{i,0}, \tag{7.6}$$

where H_{CONV_i} is the CNN in the $i-$th RSTB. Note that this CNN has the same architecture as the H_{CONV} in (7.3). The residual connection stabilizes the training and allows the accumulation of the features at different depths. It is worth noticing that unlike the original Swin Transformer architecture, in the RSTB there is no patch-merging operation (i.e., combine 2×2 image patches into a larger patch) between STL. Moreover, the embedded dimension is kept constant through the layers.

Swin Transformer layer The Swin Transformer Layer (Figure 7.1b) has the same structure as in [LLC+21]. Basically, first the input features of a STL are partitioned into non-overlapping $M \times M$ local windows ($M = 8$ pixels). Then the standard multi-head self-attention (MSA) is computed for the patches in each window. The number of heads h is fixed to 6 in the classical model and 8 in the realistic model. Next a multi-layer perceptron (MLP) with 2 connected layers (the hidden dimension is double the embedded dimension C) and GELU non-linarity is used for further feature transformation. LayerNorm (LN) is applied before both MSA and MLP, and the residual connection is applied after both modules. The whole process is then repeated but with the shifted window mechanism (that is, by cyclic shifting the windows by $\frac{M}{2}$ in each direction) to enable cross-window connections.

7.3 Training details

7.3.1 Training set

For the classical model, the authors use two datasets DIV2K (800 images), and DIV2K + Flickr2K (2650 images), with bicubic downsampling to create training sets. They observe that the model trained with more data has better PSNR performance (+0.3dB) when tested on the dataset Manga109. On the other hand, a large collection of diverse datasets (DIV2K +Flickr2K + OST (10324 images, nature) + WED(4744 images) + FFHQ (first 2000 images, face) + Manga109 (manga) + SCUT-CTW1500 (first 100 images, texts)) are used to train the realistic model. Furthermore, a sophisticated degradation model from [ZLVGT21] is adopted to simulate real-world scenarios.

7.3.2 Training loss and optimization

The classical model is trained with a simple L_1 loss, while the realistic model is trained with a combination of L_1 loss, GAN loss, and perceptual loss to obtain better visual quality. The two models are both trained 10^6 epochs on 8 GPUs. They are optimized using Adam solver (initial learning rate $= 1e-4$) and MultiStepLR learning rate scheduler with 5 steps and $\gamma = 0.5$. The batch size is set to 32 and the LR image size is 64×64 pixels.

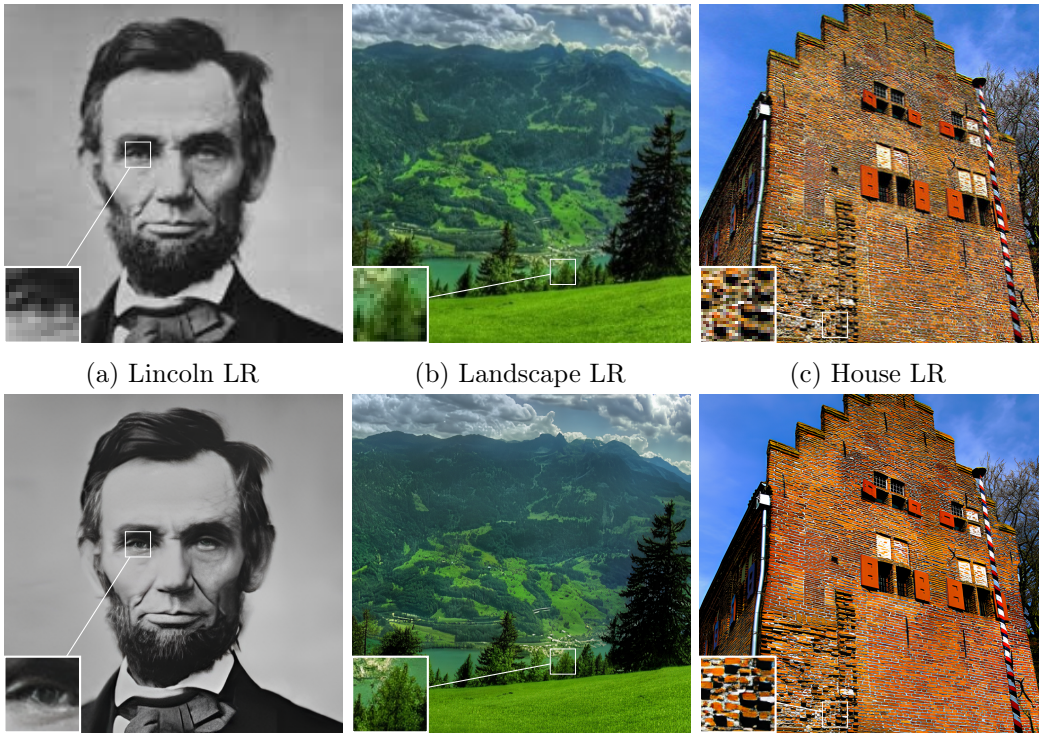
7.4 Experiments

In the demo, we fix the super-resolution factor to 4 since the authors only provide the x4 pretrained model for the realistic SwinIR.

7.4.1 Real-world super-resolution

This section presents the qualitative performance of the realistic model on real-world images. Note that the realistic model is trained to perform not only super-resolution but

also image denoising and JPEG artifacts removal, which makes it particularly suitable to restore old pictures or to enhance the quality of natural images. Figure 7.2 shows the super-resolution reconstruction of the realistic model on real images. Generally, SwinIR excelled at removing noise, JPEG artifacts, and producing plausible high-frequency details. But we also notice that when dealing with highly compressed or very noisy images, SwinIR may present unwanted artifacts such as residual noise or cartooned textures, respectively.



(a) Lincoln LR (b) Landscape LR (c) House LR

(d) Lincoln SR (e) Landscape SR (f) House SR

Figure 7.2: Visual quality of SwinIR super-resolution on real-world images. Top line corresponds to the LR input. Bottom line shows the x4 SR reconstruction of SwinIR.

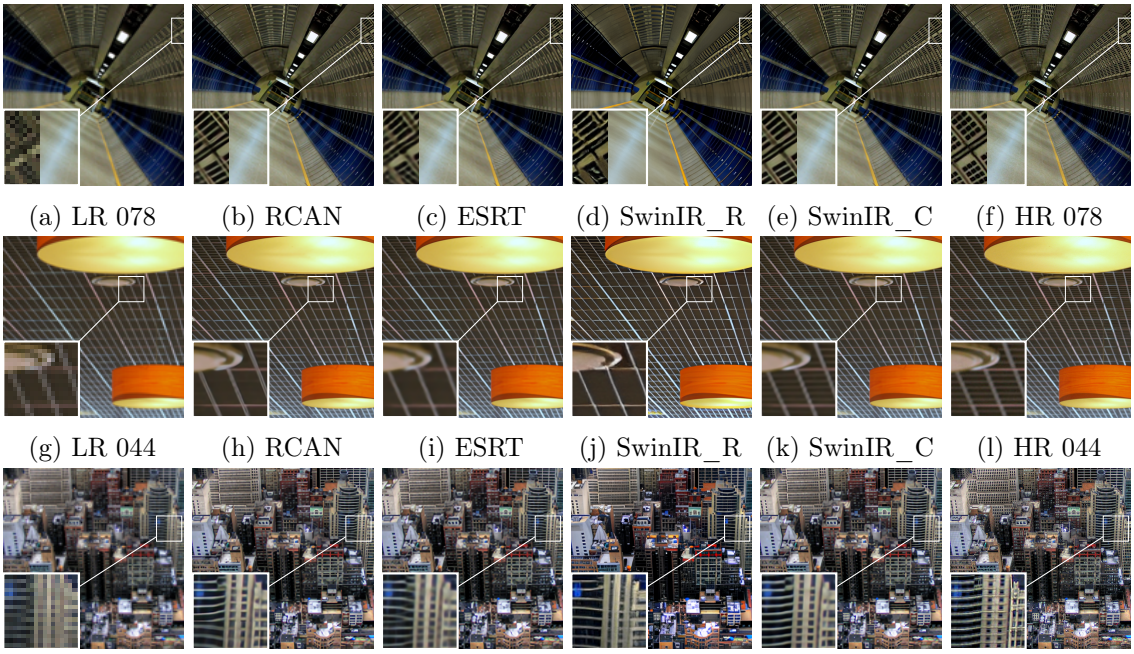
7.4.2 Auto-similarity and single image super-resolution

We know that SISR is a local problem per se. In this experiment, we want to study the impact of self-attention in the Swin Transformer in the Urban100 dataset. We choose this dataset because it contains a lot of auto-similar structures. The importance of auto-similarity in image restoration was first exploited in the Non-Local Means denoising method [BCM11]. The authors of [BCM11] demonstrate that we can reduce the noise of an image patch by aggregating its similar patches, which are not necessarily spatially close to the patch of interest. Since this chapter, auto-similarity has become more and more popular in image processing. Recently, ESRT [LLL+22] exploits auto-similarity to train a Transformer network for single-image super-resolution. It is arguable that Transformer (and Swin Transformer) can make use of attention in similar patches to get a better SR reconstruction, especially in the low-contrast or aliased regions. We also compare these Transformer networks with RCAN [ZLL+18], a classic CNN for SISR.

We use bicubic interpolation to create low-resolution images from the Urban100 dataset. Note that we do not include recent GAN-based state-of-the-art methods in this study since they will hallucinate low-contrast details. Both ESRT and RCAN are trained with

L_1 loss on the DIV2K dataset with bicubic degradation. The classic Swin Transformer model is trained with L_1 loss but on the DIV2K + Flickr2K dataset.

Figure 7.3 shows the comparison between the two SwinIR models, RCAN and ESRT on the Urban100 dataset. First, we observe that the SwinIR realistic model is not reliable for recovering the true details due to its generative nature (Figure 7.3d). Second, we expected ESRT to perform better on this particular test set using global attention (compare for example, Figure 7.3b and Figure 7.3c). Maybe the performance of ESRT is restricted by its capacity (ESRT is a lightweight network). Finally, the classical SwinIR network recovers genuinely the low-contrast and aliased textures and achieves the best results. Unfortunately, we could not claim whether this boost of performance comes from the long-range dependency mechanism. First, the window size of SwinIR is really small (8 pixels), which makes SwinIR a rather local network. Second, the classical SwinIR is trained on a larger dataset. Finally, this gain in performance may be due to the advance in network design (i.e., large kernel size, GELU activation, Layer norm, etc) rather than the superiority of Transformer over traditional CNN [LMW+22]. In conclusion, SwinIR is a promising and powerful method for SISR, but we still need to carry out more experiments to fully understand its competence.



(a) LR 078 (b) RCAN (c) ESRT (d) SwinIR_R (e) SwinIR_C (f) HR 078

(g) LR 044 (h) RCAN (i) ESRT (j) SwinIR_R (k) SwinIR_C (l) HR 044

(m) LR 073 (n) RCAN (o) ESRT (p) SwinIR_R (q) SwinIR_C (r) HR 073

Figure 7.3: Qualitative comparison between the two SwinIR models, RCAN, and ESRT on the Urban100 dataset. Super-resolution by factor of 4.

7.4.3 Application to satellite imagery

We use the realistic SwinIR SISR model to super-resolve actual satellite images from this thesis, including those from PROBA-V, SkySat, and Sentinel-2 (L1C and L1B). A visual representation of the results can be found in Figure 7.4, providing a comparative insight into the performance of the methodologies proposed throughout this chapter. These methodologies include DeepSUM-ref for multi-date SR of PROBA-V (Chapter 3); DSA for SkySat burst SR (Chapter 4); L1CSR for SISR of Sentinel-2 L1C (Chapter 8); and

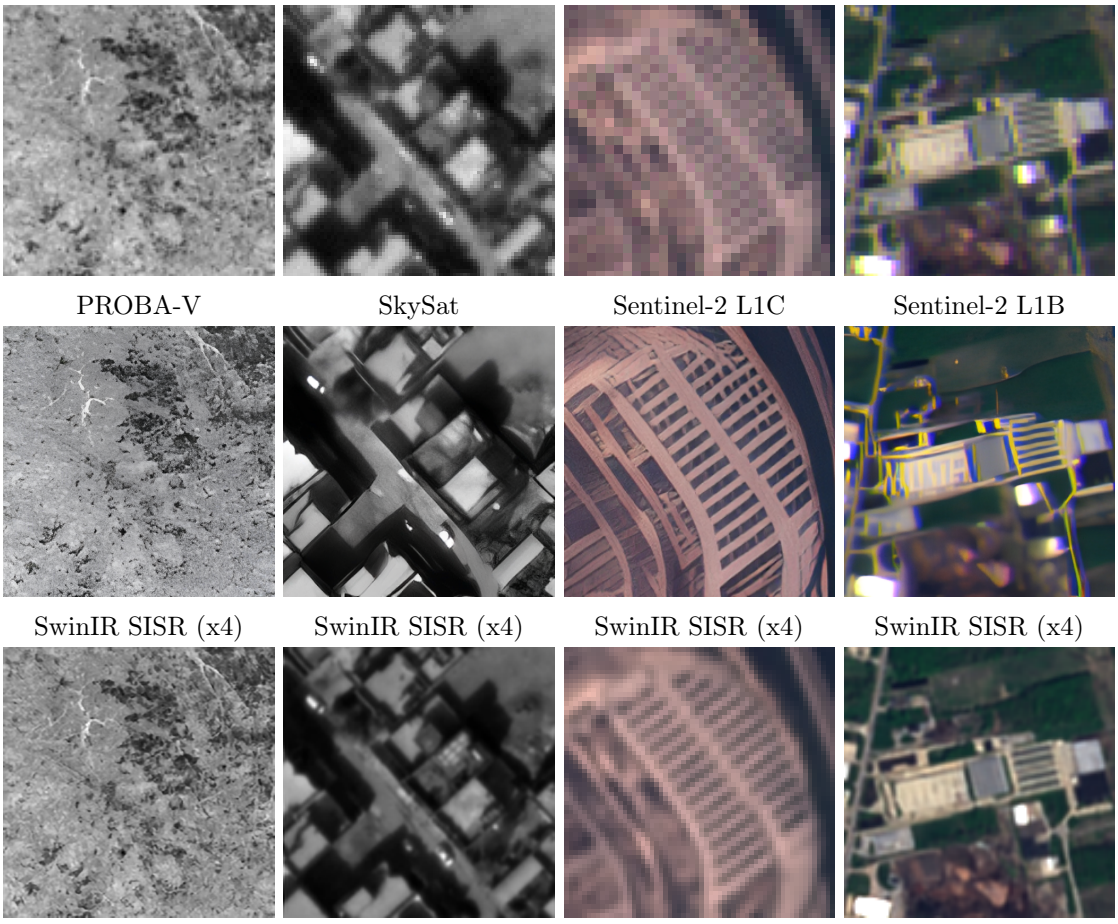


Figure 7.4: Top row: Real LR satellite images used in this thesis. Middle row: SISR reconstruction (x4) by the realistic SwinIR model. Bottom row: SR reconstruction (x2) obtained using the methodologies proposed in this thesis.

L1BSR for SISR of Sentinel-2 L1B (Chapter 8).

As expected, SwinIR doesn't excel on these satellite images. This is likely due to the model being trained on different types of data, thus leading to a domain gap problem. While no definitive judgments can be made about SwinIR's overall performance given these constraints, this experimentation yields important insights into the complexities and limitations of applying GANs for satellite image super-resolution. Notably, severe hallucination artifacts are visible in the SwinIR reconstruction of SkySat and Sentinel-2 images.

We argue that the use of GANs for SISR in satellite imagery deserves careful scrutiny, and we should avoid it when retrieving true high-frequency details is feasible (Chapter 8). We also propose that self-supervision may offer a solution to the domain-gap problem (Chapter 9).

7.5 Chapter Summary

In this chapter, we examined SwinIR, a Swin Transformer network applied to image super-resolution. We explored how Transformers can leverage auto-similarity in natural images to enhance SISR performance. While SwinIR achieves outstanding results with both synthetic and real-world imagery, becoming an increasingly popular backbone in computer vision, its performance with satellite data is not without complications. Domain-gap problems and hallucination artifacts induced by GANs limit SwinIR's effectiveness for satellite imagery super-resolution. However, as we will reveal in the next chapter, the application of GANs may not be indispensable for Sentinel-2's SISR problem. The sufficient information available makes it plausible to recover true high-frequency details without resorting to GANs, marking a promising avenue for our continuing investigation.

8 On the role of alias and band-shift for super-resolution of L1C products

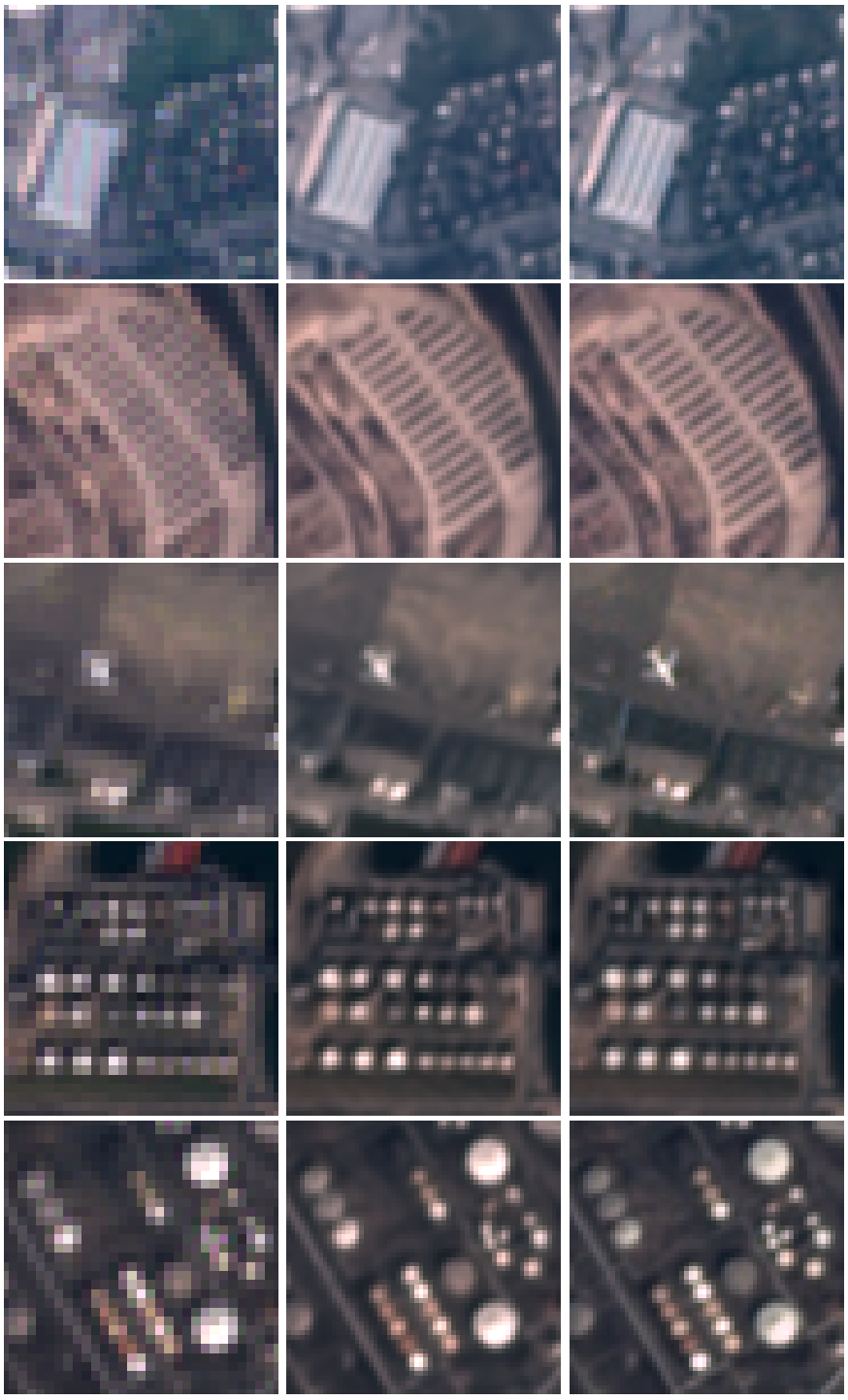
This chapter takes a focused look at the single-image super-resolution (SISR) of Sentinel-2 imagery, an integral part of our overall exploration on improving satellite image resolution. We discover that due to Sentinel-2's distinct sensor specifications, specifically the inter-band shift and alias, deep learning methods can effectively recover fine details. By employing a simple L_1 loss for training a model, we can obtain results that are devoid of hallucinated details—a crucial element in our ongoing investigation. We construct a dataset of paired images from Sentinel-2 and PlanetScope to train and assess our super-resolution model, underlining our commitment to grounded, real-world applicability in our approach.

8.1 Introduction

The use of satellite imagery has become increasingly prevalent in a variety of fields, from environmental monitoring to urban planning. One such satellite is the Sentinel-2 constellation, which provides recurrent 10m/pixel resolution optical imagery. The high frequency revisit of Sentinel-2 makes it useful for monitoring temporal changes, such as the growth of crops or the spread of urban development. However, the relatively low spatial resolution can be a limitation for certain applications, such as identifying small objects.

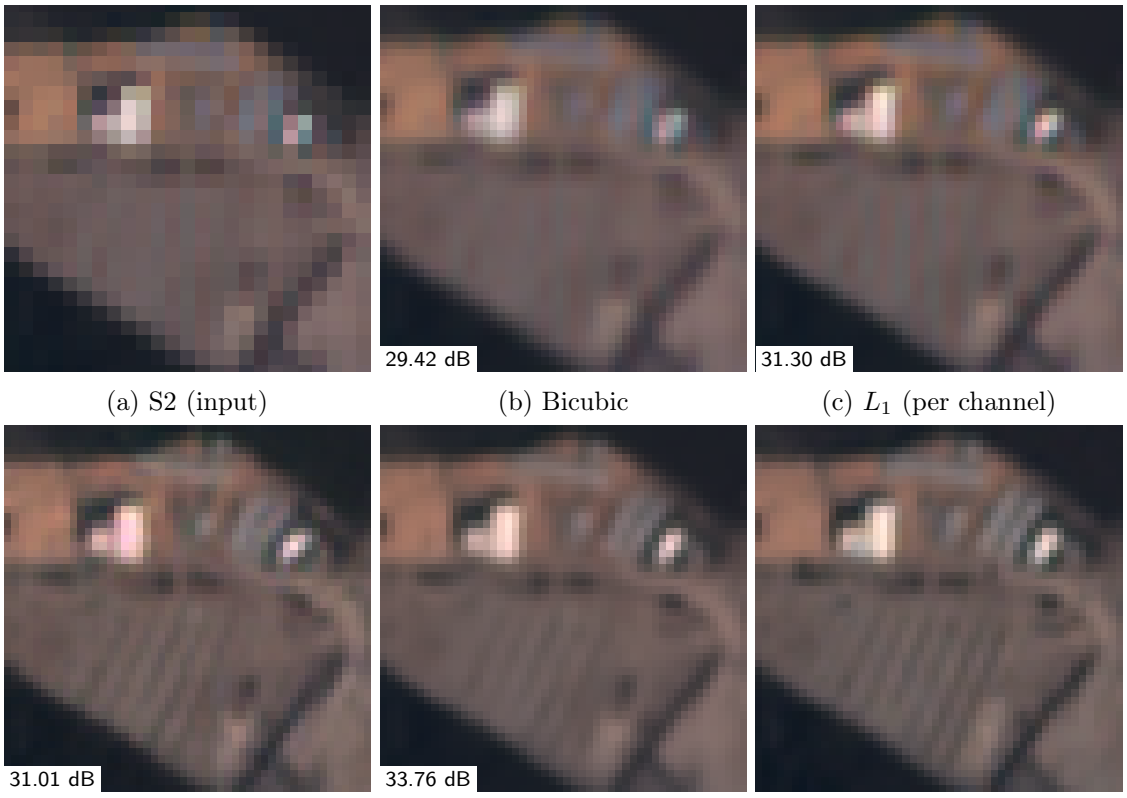
In this chapter, we propose a deep learning approach for SISR of Sentinel-2 imagery. Unlike previous methods that aim for a x4 increase in spatial resolution, our work focuses on a x2 increase, which we argue is a more reasonable and practical choice. Additionally, we avoid using generative adversarial networks (GANs) in our method, as they have been known to introduce hallucinations and artifacts that can be undesirable for sensitive applications. Instead, we use an L_1 cost function, which has been shown to effectively preserve image details while minimizing distortion [BM18] (see Figure 8.1).

This study focuses on understanding what makes SISR of Sentinel-2 imagery possible. To this aim, we explore two unique characteristics of Sentinel-2: the alias and the inter-band shift and find that they enable the reconstruction of fine structures. It is worth noting that super-resolving the 10m bands of Sentinel-2 is a relatively new problem, and while we do not aim to achieve the best possible results, we analyze the specific features of Sentinel-2 imagery relevant for SR.



Sentinel-2 L1C Super-resolution x2 PlanetScope

Figure 8.1: SISR results obtained with the L_1 loss. We argue that the characteristic alias and band-shift are key for x2 SR of Sentinel-2 imagery.



(a) S2 (input) (b) Bicubic (c) L_1 (per channel)

(d) GAN loss (e) L_1 loss (f) PS (ground-truth)

Figure 8.2: SR results from different models. The per-channel result corresponds to one model per channel trained independently. The GAN and the L_1 loss are able to restore similar details.

8.2 Related work

Early research on SISR of Sentinel-2 images focuses on pan-sharpening the lower-resolution (20m and 60m) bands to complete a uniform 10m GSD data cube [LBDG+18, GMG+19]. Recent trends are super-resolving the 10m bands of Sentinel-2 using other relevant very high-resolution satellites. For example, [PAB20] generates low-resolution/high-resolution (LR-HR) image pairs from the PeruSat-1 satellite (2.8m GSD) to train a x4 SR model and use it to reconstruct fine textures in the Sentinel-2 10m bands. However, these techniques use a pre-determined degradation model, like bicubic downsampling, to create LR from HR. So when the input deviates from the pre-defined degradation model, the performance may drop substantially. To fill the gap between simulated and real-world remote sensing images, real HR satellites such as PlanetScope [GSA+20, ZB22], VENµS [MVSIH22], and WorldView [SRMV20] are used directly to supervise the SR of Sentinel-2. Perceptual losses, such as GAN or high-level feature matching, are used in these works to produce sharp outputs.

Besides focusing on perceptual restoration, most past studies do not justify why the reconstruction of actual high-frequency details is feasible from a single multi-band image. Results reported in Chapter 4 suggest that alias and displacement between frames are crucial for exploiting complementary information in different frames (or different spectral bands in our case) and obtaining up-to-par SR performance.

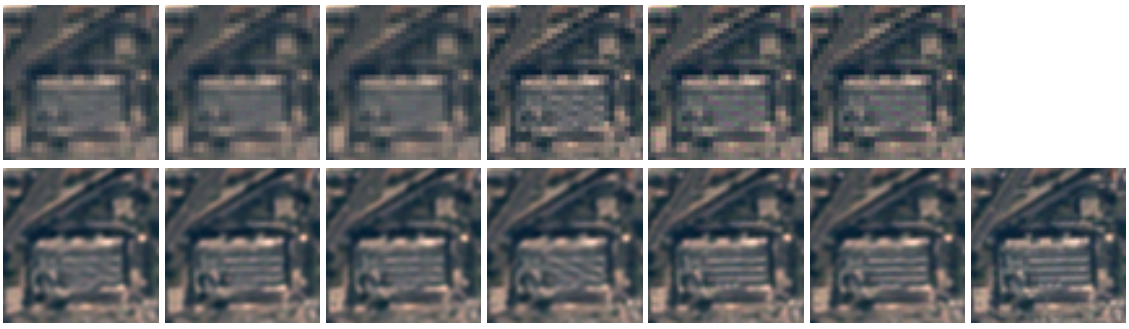


Figure 8.3: Synthetic dataset. The top row shows simulated LR images corresponding to the six acquisition configurations, and the bottom row shows the SR results obtained with networks trained on these specific synthetic datasets. The bottom-right image is the ground-truth HR.

8.3 Method

Our method is specifically designed for x2 SR of Sentinel-2 images. It is based on the ESRGAN architecture [WYW+19], adapted for a smaller network and a single-term loss. The model is trained on pairs of Sentinel-2 and PlanetScope images, suitable for x2 SR, as described in Section 8.4.

Architecture. Given that ESRGAN was developed for a x4 SR factor, we propose some adjustments for a factor of 2. We found that using only 8 RRDB blocks instead of 23 RRDB blocks [WYW+19] was enough to obtain satisfactory results while significantly reducing the training and inference time.

Cost function. The ESRGAN model [WYW+19], was initially trained on a set of HR natural images from the DIV2K dataset [AT17]. It uses a base model trained with a loss L_1, followed by a second training phase using a cost function that includes a relativistic discriminator loss [JM18], a perceptual loss [JAFF16], and the L_1 loss. However, when adapting the model for Sentinel-2 SR, we found that training the model on Sentinel-2/PlanetScope image pairs using only the L_1 loss instead of the complete loss function with perceptual terms resulted in a similar detail reconstruction (second row in Figure 8.2). This suggests that, thanks to the alias and inter-band shift present in Sentinel-2 imagery, the problem is better posed. Hence the L_1 loss is sufficient for successful x2 SR. This is further explored in Section 8.5.

8.4 S2/PS dataset

For this study, we built a dataset of Sentinel-2 L1C and PlanetScope image pairs, referred to as the S2/PS dataset. The Sentinel-2 L1C images have a spatial resolution of 10m/px, while the PlanetScope images have a resolution of 3m/px. The latter were resampled to 5m/px and registered with the S2 coordinate system with bicubic interpolation. Since PlanetScope images tend to be well-sampled [AdFF19], the resampling step does not introduce a significant loss of information. The pairs are from images taken on the same day, a few hours apart, to minimize changes due to vegetation or human activity. The remaining changes in the images include the presence of clouds, differences in satellite

Table 8.1: Shift and alias influence. Best PSNR in **bold** and second best underlined.

		PSNR (dB)		
		Test set	Train set	Val set
no alias	no shift	46.69	47.13	47.35
	fixed shift	47.20	47.38	47.57
	random shift	46.87	47.15	47.37
alias	no shift	46.67	47.29	47.51
	fixed shift	**49.30**	**49.25**	**49.54**
	random shift	48.12	48.44	48.73

perspective, and shadows. In this work, we use the PlanetScope images acquired with the PS2 instrument (Dove Classic), so we restrict our study to the three visible bands (blue, green, and red).

To prepare this dataset, we performed equalization of the mean and standard deviation for each band from each PlanetScope image to the corresponding Sentinel-2 image. The residual spatial shift between the downsampled PlanetScope image and the Sentinel-2 one is estimated using the phase correlation algorithm. Then, the PlanetScope image is resampled using a third-order spline interpolation and introducing zeros where information was missing. Pairs with a phase correlation score below 0.55 were removed from the dataset.

We used 380 full-scene images and extracted up to 20 LR crops of size 200×200 from each image. The test set consists of 65 of these scenes (or 693 crops) and is geospatially disjoint from the train set (3680 crops). The validation set (406 crops) is selected from different dates. Figure 8.1 shows zoom-in crops from the dataset, where alias and inter-band shift are clearly visible in the S2 images.

8.5 Experiments

This section presents experiments that empirically show that Sentinel-2 imagery is well-suited for the problem of SISR.

A surprisingly performant L_1 loss. We compare two models: one trained with the L_1 loss and one trained with the original ESRGAN loss, with the relativistic discriminator and feature similarity terms. Quantitatively, the average PSNR (over 12 bits) computed over the test set yields 42.21dB for the L_1 loss and 37.29dB for the ESRGAN loss. Visually, we observe that the results obtained with the L_1 loss are slightly smoother than those obtained with the ESRGAN loss, but do not contain any color artifacts. This can be seen in the second row of Figure 8.2. In addition, the details in the images generated by the L_1 model are much better than those obtained through bicubic interpolation. Overall, these experiments suggest that, in the case of Sentinel-2, the L_1 loss is an effective solution to increase the resolution by a factor 2 without risking the introduction hallucinated details [BM18].

Additional results using the L_1 model on the test set are shown in Figure 8.1. We find that the network is able to resolve aliased patterns into high-frequency details, with strong fidelity to the ground-truth PlanetScope images. Given that the L_1 loss minimizes distor-

tion [BM18], one can be confident that there are few hallucinated details. In ambiguous cases, the network will likely favor a blurry result instead of sharp, but potentially wrong details.

Cross-spectral information. We claim that our network exploits cross-spectral information to increase spatial resolution. To validate this hypothesis, we perform the following experiment: from the S2/PS dataset, we train three networks, each dedicated to super-resolving one specific spectral band, and only this band is given as input. On the test set, we observe a drop of 0.88dB in the PSNR, and we observe visually that the network is no longer able to resolve fine structures such as very high-frequency patterns. Even though the LR signal is aliased in each spectral band, the network no longer has the ability to perform a consistent, joint reconstruction of the signal. This can be observed in the top-right image of Figure 8.2. A related observation was reported in [GSA+20] in which a network trained with both RGB and NIR bands performed better than just with RGB bands.

Aliasing and band-shift influence. We argue that the model described in Section 8.3 is able to exploit specific characteristics of the Sentinel-2 sensor, namely the presence of alias in each band and the inter-band shifts. The alias is due to a low spatial sampling with respect to the modulation transfer function of the instrument [GBT+17], and the inter-band shifts originate from time delays between the acquisition of the lines of the different spectral bands [GBT+17]. Combined, these two aspects yield a configuration that is better-posed than standard SISR, and real information can be recovered under these acquisition specificities.

Next, we provide experimental evidence that the acquisition configuration of Sentinel-2 is indeed favorable to SISR. To this aim, we construct synthetic datasets using six different acquisition configurations: with and without alias, and with and without fix/random inter-band shifts. In each configuration, we use the PlanetScope images as ground-truth and we synthesize LR images according to each configuration. The presence of alias is controlled by the amount of blur introduced before downsampling. The shifts are +/-1 offsets applied to the bands before downsampling and then compensated by 0.5 offsets on the LR images. In each configuration, 0.1% Gaussian noise was added to match Sentinel-2 noise level. The first row of Figure 8.3 shows the effects of these configurations over the generated LR images. The configuration *with alias* and *random inter-band shift* is the most faithful simulation of Sentinel-2 imagery.

We train one network per scenario according to the same training details as in Section 8.5. Table 8.1 shows the PSNR (over 12 bits) over the train, validation and test sets for the different settings, and the bottom row of Figure 8.3 shows SR results. These results highlight that the combined presence of the inter-band shift and the alias allows the network to retrieve significant information from the signal. In contrast to the usual SISR scenario where little alias and no inter-band shift are present, our experiments assert that Sentinel-2 imagery is well-suited for SISR.

8.6 Chapter Summary

In this chapter, our focus was on the contributing factors that enable the Single Image Super-Resolution (SISR) of Sentinel-2 L1C imagery. We found that the network's capacity to leverage Sentinel-2's specific characteristics—alias and inter-band shift—underpins the gains in resolution. We substantiated these insights through extensive experiments on

meticulously assembled synthetic datasets. By using a straightforward L_1 loss in training our model, we've demonstrated successful enhancement of Sentinel-2's spatial resolution from 10m to 5m GSD, whilst curbing distortion and preventing the formation of false details. As we progress, we will carry forward these insights into the next chapter, where our goal will be to conduct a novel self-supervised SISR study, but this time on Sentinel-2's L1B data.

9 Exploiting detector overlap for self-supervised super-resolution of L1B products

Building upon our investigation into Single Image Super-Resolution (SISR) of Sentinel-2 imagery in previous chapters, this chapter addresses the lack of reliable high-resolution (HR) ground truth target. We introduce L1BSR, a self-supervised deep learning method designed to super-resolve and align bands of Sentinel-2 L1B 10m images. Training directly on real L1B data, our method leverages overlapping areas in images produced by adjacent CMOS detectors, circumventing the need for HR ground truth. The method utilizes a novel Cross-Spectral Registration network (CSR) to enforce correct band alignment in the super-resolved output. The CSR network is also trained with self-supervision using an innovative Anchor-Consistency loss, which we also introduce in this chapter. Our method's performance, evaluated on both synthetic and real L1B data, proves to be on par with supervised approaches, demonstrating the potency of self-supervised learning for satellite imagery super-resolution.

9.1 Introduction

Earth observation (EO) satellites play a crucial role in our understanding of the Earth systems including climate, natural resources, ecosystems, and natural and human-induced disasters. The Sentinel-2 mission, which is a part of the Copernicus Programme by the European Space Agency (ESA), is considered a significant EO effort alongside other missions such as Landsat. Sentinel-2 provides optical images of Earth's surface in 13 spectral bands, 4 bands at 10m resolution, 6 bands at 20m, and 3 bands at 60m. The blue (B), green (G), red (R), and near-infrared (N) bands at a ground sample distance (GSD) of 10m/pixel are particularly useful for a variety of applications, including land cover classification, vegetation monitoring, and urban mapping [DDBC+12]. However, for certain tasks, such as identifying small objects or analyzing fine-scale features, this spatial resolution is still inadequate. To address this limitation, super-resolution (SR) techniques can be used to achieve a GSD better than 10m for the RGBN bands.

SR approaches can be broadly classified into multi-image super-resolution (MISR) and single-image super-resolution (SISR). MISR aims at reconstructing a high-resolution (HR) image from a set of low-resolution (LR) images, typically captured with different viewpoints [AEdFF20, NAD+21b, NAD+22b, LNFE23] or at different satellite passes [MIKC19, AMSC+20]. If the LR images contain alias and sub-pixel misalignment, they present a perfect scenario for MISR to leverage complementary information in different frames and to recover the true details in the HR output [NAD+21b]. SISR, on the other hand, is often considered an ill-posed problem due to the potential loss or corruption of high-

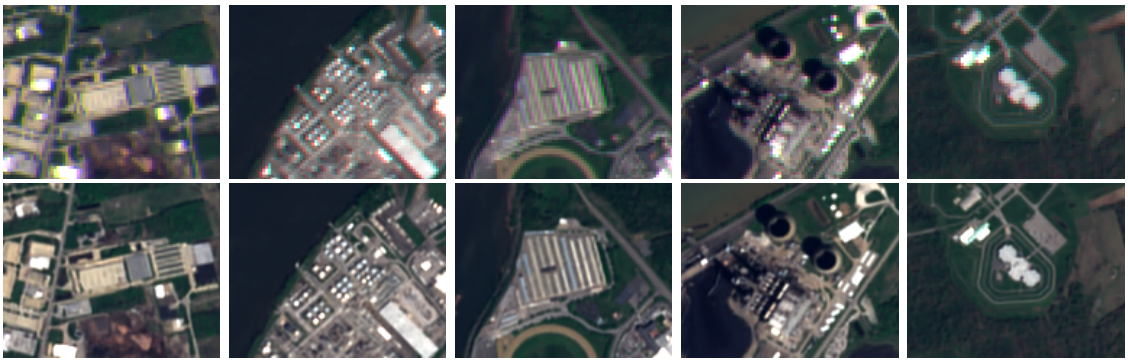


Figure 9.1: L1BSR produces a 5m high-resolution (HR) output with all bands correctly registered from a single 10m low-resolution (LR) Sentinel-2 L1B image with misaligned bands. Note that our method is trained on real data with self-supervision, i.e. without any ground truth HR targets.

frequency information caused by factors such as noise, blur, or compression. Nonetheless, our study in Chapter 8 demonstrates the possibility of SISR for Sentinel-2 10m bands thanks to its unique sensor specifications, namely the inter-band shift and aliasing. The misaligned bands sample the ground at different positions. Since they are correlated each band obtains complementary information from the other bands, in a situation similar to a demosaicing problem.

Deep learning (DL) SISR methods currently outperform traditional model-based approaches by a large margin [WCH20]. To date, all learning-based methods for SISR of Sentinel-2 10m bands have used supervised training, which penalizes a loss between the HR image predicted by the network and a ground truth HR image. Some studies attempt to train a SISR model on a simulated dataset where LR images are generated using a predefined degradation model [PAB20]. However, the performance may drop substantially if the real low-resolution input deviates from the simulated degradation model. Other works use real HR images acquired by other satellites to directly supervise the SR of Sentinel-2 [GSA+20, NAR+23]. However, obtaining the HR ground truth images can be costly. In addition, the use of HR images from different satellites introduces challenges such as spectral response discrepancies, and acquisition viewpoint and time differences, which complicate the process of dataset creation and negatively impact performance.

A promising direction is to use self-supervised learning techniques, which have been applied to multi-image restoration tasks such as video/burst denoising and demosaicing [EDM+19, EDAF19, DAD+21, YPPJ20, SMV+21], and recently to MISR in the context of push-frame satellites [NAD+21b, NAD+22b] that acquire bursts of images at high frame rate. These techniques exploit redundant information from multiple observations: one of the degraded frames in the input sequence is withheld from the network and used as label instead of the ground truth. As a consequence they require at least two degraded observations from the same HR signal.

In this chapter we leverage a unique feature of the Sentinel-2 hardware design that enables self-supervised training of single-image super-resolution (Figure 9.1). Sentinel-2 is equipped with a MultiSpectral Instrument (MSI) that has 12 detectors capturing information in the in the visible and near infrared (VNIR) wavelength range.These detectors operate in a push-broom fashion, scanning the image line-by-line as the satellite moves over the ground (as illustrated in Figure 9.2). Of note, adjacent detectors share a 2 km overlap

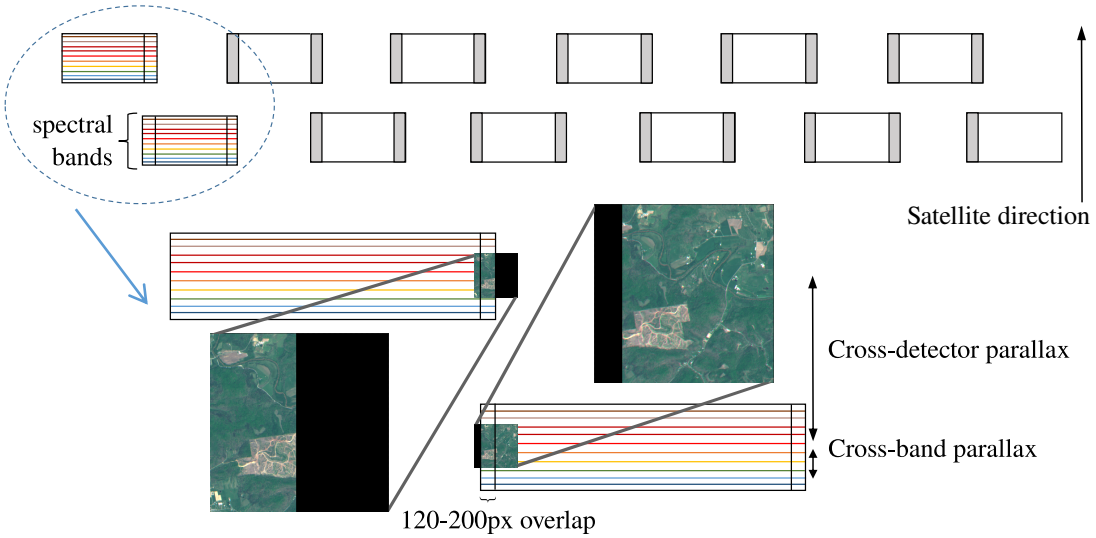


Figure 9.2: Sensor layout of the Sentinel-2 MSI (figure adapted from [GBT+17]). The push-broom acquisition is done in the vertical direction. The overlap between detectors provides two near-simultaneous observations of the scene.

area across the track (120 to 200 pixels), which offers opportunities for self-supervised image restoration techniques.

These overlapping regions are only available in early products in the Sentinel-2 processing pipeline, such as the level-1B (L1B) products, which present a significant inter-band parallax due to the hardware design of the detectors. This parallax, while beneficial for the super-resolution task, is undesirable for human interpreters, which is why it is removed later in the pipeline (e.g. L1C products) using camera calibration information to align the different bands.

Contributions. In this chapter we propose L1BSR, a novel self-supervised method for SISR and band alignment of Sentinel-2 L1B RGBN bands. The method is trained directly on real L1B data using image crops contained in the detector overlap regions. One of the overlapping L1B crops is given as input to our network, and the other is used as target in the loss. The network is tasked to generate a super-resolved image such that, when properly aligned and downsampled, matches the target 10m LR crop (Chapter 4). It should be noted that once trained, as a SISR method, our network has the capability to produce 5m HR images throughout the image domain, rather than just at the overlapping regions.

Our self-supervised loss is designed to enforce the super-resolution network to output an HR image with the bands correctly aligned. This is achieved by aligning all bands of the target image with the green channel of the super-resolved output.

To that aim, as a second contribution of our work, we present a novel cross-spectral registration (CSR) method which allows to compute an optical flow between images of different spectral bands. To train the CSR network we propose Anchor-Consistency, a simple yet effective self-supervised loss for cross-spectral registration. Our self-supervised cross-spectral registration simultaneously learns to handle all possible band combinations. We use our CSR network only during training of the SR network as part of the self-supervised loss. Once the reconstruction network is trained, the cross-spectral registration network is

no longer required. The reconstruction network can directly generate high-quality HR images with correctly registered bands (as shown in Figure 9.1) without requiring an explicit alignment step, nor calibration information.

We validate our contributions with an empirical study on a synthetic dataset obtained from L1C products (Section 9.4), designed to model the main characteristics of Sentinel-2 L1B data. We show that our L1BSR network as well as our cross-spectral registration module trained with the proposed self-supervision strategy attain performance on par with those obtained with supervised training.

We train our self-supervised method on a dataset of 3740 pairs of L1B RGBN overlapping crops (Section 9.4.4) and compare with our supervised method designed for Sentinel-2 L1C (Chapter 8) in Section 9.4.5. It is worth noting that our training dataset, which can find applications in various image restoration and cross-spectral registration tasks, will be soon available on our project website.

9.2 Related work

SISR for Sentinel-2 Early research on SISR of Sentinel-2 images primarily focused on pan-sharpening the lower-resolution (20m and 60m) bands to create a uniform 10m GSD data cube [LBDG+18, GMG+19]. Recent years have seen an increased interest in enhancing the resolution of the Sentinel-2 10m bands. Some studies [PAB20] generated synthetic LR-HR pairs to train SISR models, but these models tend to suffer from generalization issues [CZY+19]. Another trending approach is to directly supervise the SISR of Sentinel-2 using another high-resolution satellite such as PlanetScope [NAR+23, GSA+20, ZB22], VENµS [MVSIH22], or WorldView [SRMV20]. However, creating the training dataset for these approaches requires significant engineering work due to the radiometric and geometric differences between the two constellations.

Self-supervised SR Self-supervised and unsupervised learning are promising approaches to avoid the need for large labeled datasets in SR tasks. ZSSR [SCI18] and MZSR [SCC20] have been proposed to model image-specific LR-HR relations during the testing phase using example pairs generated from the LR test image and its degraded version. Although the idea is interesting, it may not be practical to train on each test image. Another approach is to use cycle-consistency and adversarial losses [YLZ+18, KPL+20, LDT19, WZYW21] to train a neural network without requiring pairs of LR-HR images. However, these GAN-based models are prone to producing hallucinations, which may not be acceptable for certain applications.

Our SISR method is both fully self-supervised and free from hallucinations. We drew inspiration from the Noise2Noise framework [LMH+18], which introduced a pioneering approach to train a neural network for image denoising task using pairs of noisy images instead of pairs of noisy-clean images. The key idea behind Noise2Noise is that when comparing pairs of noisy images with independent noise realizations, the network learns to identify the underlying patterns in the noise and removes them accordingly. Similarly by comparing pairs of overlapping L1B images, our network can learn to recognize the aliasing patterns in each LR image and leverage them to recover the high-frequency details in the HR. The closest work to ours is the DSA framework (Chapter 4), which addresses MISR for SkySat imagery. During training, DSA hides the LR reference image and asks the network to produce a HR image from the other $n-1$ images such that after downsampling,

it coincides with the reference image. Our work can be seen as the SISR version of DSA. However, our network also learns to perform implicit cross-spectral registration at inference time, which is a challenging and compelling task by itself.

Cross-spectral registration Cross-spectral registration refers to the process of aligning two or more images that are captured using different sensors or imaging modalities. Cross-spectral registration has become increasingly important in various fields such as remote sensing [YSBS17,PWÖ+20], medical imaging [LHS+09], and computer vision [AGD+20] as it allows for the integration of information from different spectral bands, thereby yielding richer scene representations. While increasing efforts have been made in the past few years to improve the performance of cross-spectral registration, this still remains an open problem [JMX+21]. Feature-based methods [YSBS17, LAZW18] involve identifying distinctive features, such as edges or corners, in both images and then matching them to establish correspondences. Intensity-based methods [CSCL20, ZZH+18] rely on the similarity of the pixel values in both images. Examples of intensity-based methods include normalized cross-correlation, mutual information, and phase correlation. In recent years, deep learning-based methods have also been explored and achieved state-of-the-art (SOTA) performance. Most DL studies including [PWÖ+20, AGD+20, WMJB21, XMY+22] have employed image-to-image translation [IZZE17] techniques to map two images to the same image space and then register them accordingly. However, these methods require extensive work in designing models and sophisticated training losses. In contrast, we propose Anchor-Consistency a novel and straightforward loss for training a cross-modal registration network. Notwithstanding its simplicity, our method provides a strong baseline for the task. In addition, our loss can also be easily integrated into existing frameworks to improve consistency or used as a quantitative metric for evaluating cross-modal registration quality.

9.3 Proposed Method

Our primary aim is to leverage detector overlaps in Sentinel-2 L1B images to learn to recover high-frequency details hidden in its misaligned bands. Note that the maximum attainable resolution is capped by the spectral decay of the blur kernel resulting from the sensor's pixel integration and the camera optics, which imposes a frequency cutoff beyond which there is no usable high frequency information [BK02]. For this reason, our aim in this chapter is to increase the resolution by a factor 2. In this section, we first present an overview of our proposed L1BSR framework (Section 9.3.1). Then, we describe our self-supervised losses in Section 9.3.2 and provide details about the training in Section 9.3.3.

Throughout the text, we denote by $I_t, t \in \{0, 1\}$ the two 4-channel overlapping images. We refer to I_0 as the input (or reference image) for the SR task and I_1 as the target for our self-supervised losses, which are explained in more detail in Section 9.3.2. $I_{t,i}$ is the grayscale image extracted from the channel i of I_t, where $i \in \{b, g, r, n\}$ and b, g, r, and n stand for the blue, green, red, and near-infrared channels, respectively.

9.3.1 Architecture

Our proposed L1BSR framework (Figure 9.3) consists of two main components: a cross-spectral registration network (**CSR**) and a reconstruction network (**REC**). The **CSR** module computes dense correspondences between $I_{0,g}$ and all the bands of I_1. By utilizing

these motion fields during training, the **REC** network learns to produce a HR output $\widehat{I}_0^{HR}$ with all four channels aligned with $I_{0,g}$. Of note, the **CSR** module is not used at test time. Instead, the **REC** network performs cross-channel registration implicitly.

Reconstruction Network Our **REC** network is built on the Residual Channel Attention Networks (RCAN) architecture [ZLL+18], which has been shown to achieve state-of-the-art performance on many image restoration tasks. We chose this architecture mainly due to its channel attention mechanism, which can be viewed as a weighting function that enables **REC** to selectively focus on informative channels in the feature space and disregard irrelevant ones.

The reconstruction network takes a LR image $I_0 \in \mathbb{R}^{H\times W\times 4}$ with misaligned bands as input and produces a super-resolved output by a factor of two $\widehat{I}_0^{HR}$ with all four bands aligned with the green channel of the input image

$$\widehat{I}_0^{HR} = \mathbf{REC}(I_0; \Theta_{\mathbf{REC}}) \in \mathbb{R}^{2H\times 2W\times 4}, \tag{9.1}$$

where $\Theta_{\mathbf{REC}}$ denotes the network parameters. We opt for the default RCAN configuration to strike a balance between computational efficiency and performance. Overall, the **REC** network contains 10 residual groups, each with 20 residual channel attention blocks (RCAB), and a global skip connection. Each RCAB is a combination of a residual block and a channel attention layer implemented using a "squeeze-and-excitation" technique [HSS18]. The number of feature channels is fixed to 64 across all layers.

It is important to highlight that the task our **REC** network must accomplish is particularly challenging, as it involves both super-resolution and cross-spectral registration at the same time. To tackle this problem, we incorporate a dedicated **CSR** module into the training process, which enables the **REC** network to learn efficiently the task in a self-supervised way.

Cross-Spectral Registration Network The **CSR** module is instrumental during training. We use it to train the **REC** network to produce an HR output where all bands are aligned to the green one (as justified in Section 9.4.2). By having to align the channels, it becomes easier for the **REC** network to learn inter-band correlations, and thereby to leverage the complementary information in each band.

The cross-spectral registration network takes any two spectral bands of Sentinel-2 L1B images $I_{\cdot,i}$ and $I_{\cdot,j}$, with $i, j \in \{b, r, g, n\}$ as input and produces a dense correspondence between them

$$F_{I_{\cdot,j}\to I_{\cdot,i}} = \mathbf{CSR}(\bar{I}_{\cdot,i}, \bar{I}_{\cdot,j}; \Theta_{\mathbf{CSR}}) \in [-R, R]^{H\times W\times 2}, \tag{9.2}$$

where $\Theta_{\mathbf{CSR}}$ denotes the parameters of **CSR**, and $\bar{I}$ is the normalization of image I according to its mean and standard deviation. The network is trained with a maximum motion range of $[-R, R]^2$ (with $R = 10$ pixels). Note that $I_{\cdot,i}$ and $I_{\cdot,j}$ should represent the same scene and be extracted either from the same image or from two overlapping images. The **CSR** follows a simple U-Net architecture [RFB15] with 4 scales to increase the receptive field of the convolutions.

9.3.2 Self-supervised learning

Our framework is trained in a fully self-supervised manner, i.e. without requiring ground truth. In this section, we describe our self-supervised losses that are used to train the **REC** and the **CSR** modules.

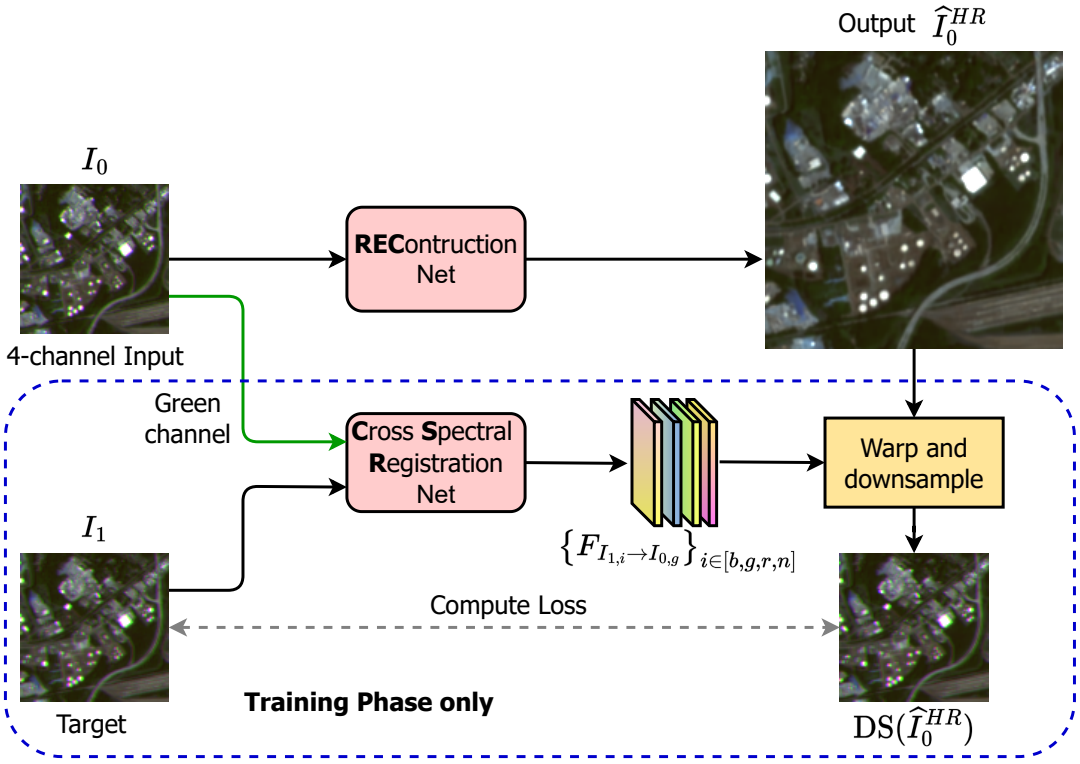


Figure 9.3: Overview of our proposed self-supervised L1BSR framework for Sentinel-2 L1B at training time. The depicted loss represents the self-supervision term $\ell_{\text{Self-SR}}$ (9.3). Note that at inference time, only one input and the reconstruction module are required.

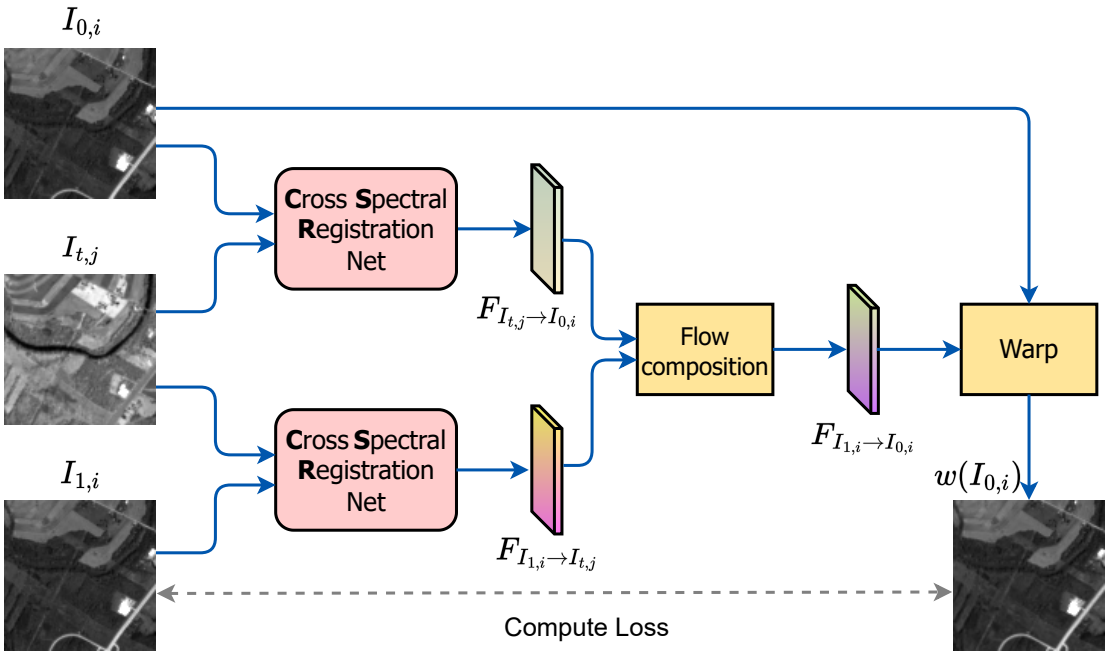


Figure 9.4: Training setup of our proposed cross-spectral registration (**CSR**) module. The motion from $I_{0,i}$ to $I_{1,i}$ via the anchor $I_{t,j}$ should represent the direct motion between them. The depicted loss represents the Anchor-Consistency loss (9.7).

Self-SR Loss By utilizing the motion fields between the green band of I_0 and all bands of I_1 computed by **CSR**, **REC** can produce a HR output $\widehat{I}_0^{HR}$ with all bands correctly registered to $I_{0,g}$. We achieve this by minimizing the Self-SR loss:

$$\ell_{\text{Self-SR}} = \|\Pi_2 \omega_2(\widehat{I}_0^{HR}, F_{I_1 \to I_{0,g}}) - I_1\|_1, \tag{9.3}$$

where Π_2 denotes the subsampling operator and $\omega_2(-, F)$ computes a bicubic sampling (Pullback) in the HR domain using the LR flow F. F denotes actually 4 optical flows, one for each band in I_1. The operator ω_2 is equivalent to a backward warping with an upscaled version of the flow $2F$.

The Self-SR loss forces $\omega_2(\widehat{I}_0^{HR}, F_{I_1 \to I_{0,g}})$ to be aligned to I_1, resulting in the requirement that all bands of the output $\widehat{I}_0^{HR}$ be registered with $I_{0,g}$. Following the work on DSA in Chapter 4, we can also incorporate a blur kernel k into the Self-SR loss to directly produce a sharp HR image

$$\ell^*_{\text{Self-SR}} = \|\Pi_2 \omega_2(\widehat{I}_0^{HR} * k, F_{I_1 \to I_{0,g}}) - I_1\|_1. \tag{9.4}$$

During training, we randomly choose the reference and the target between the two overlapping images. At inference time, we only need one single LR input and the **REC** network to obtain a high-quality HR output.

Anchor-Consistency Loss The **CSR** network is trained with self-supervision. Figure 9.4 illustrates our training setup for the **CSR**. For that we need 3 images $I_{0,i}$, $I_{t,j}$, and $I_{1,i}$. The image $I_{t,j}$ serves as an anchor image extracted from either the I_0 or I_1 (i.e. $t \in \{0, 1\}$) but may come from different spectral band than the two other images (i.e. $j \neq i$). We compute the motion fields between these 3 images in two steps:

$$\begin{aligned} F_{I_{t,j} \to I_{0,i}} &= \mathbf{CSR}(\bar{I}_{0,i}, \bar{I}_{t,j}), \\ F_{I_{1,i} \to I_{t,j}} &= \mathbf{CSR}(\bar{I}_{t,j}, \bar{I}_{1,i}). \end{aligned} \tag{9.5}$$

These two motion fields should be consistent in such a way that their composition enables alignment between $I_{0,i}$ and $I_{1,i}$:

$$\widehat{F}_{I_{1,i} \to I_{0,i}} = F_{I_{1,i} \to I_{t,j}} \circ F_{I_{t,j} \to I_{0,i}}. \tag{9.6}$$

The Anchor-Consistency loss constrains that the motion from $I_{0,i}$ to $I_{1,i}$ via the anchor should represent the direct motion between them.

$$\ell_{\text{Anchor-Consistency}} = \|\omega(I_{0,i}, \widehat{F}_{I_{1,i} \to I_{0,i}}) - I_{1,i}\|_1, \tag{9.7}$$

where ω is the classic pullback operator.

During training, we randomly choose the reference and the target between the two overlapping images. The spectra i and j are also picked arbitrarily in $\{b, r, g, n\}$. It is important to note that the case where i and j are identical is also considered for **CSR** to learn to register images of the same band.

9.3.3 Training details

We train first the **CSR** network, as it is important to ensure that the **CSR** output can be effectively utilized by the **REC** network. We employ the Anchor-Consistency loss (9.7) to train the network, with weights initialized using Xavier's initialization [GB10]. We set

the batch size to 64 and used Adam [KB14] with the default PyTorch parameters and a learning rate of $5e-5$ to optimize the loss. The training converged after 200k iterations and took approximately 24 hours on a single NVIDIA V100 GPU.

The second phase consists of training the **REC** network using the Self-SR loss (9.3) and the trained **CSR**. We train **REC** on LR crops of size $96 \times 96 \times 4$ pixels and validate on LR images of size $256 \times 256 \times 4$ pixels. We set the batch size to 16 and optimize the loss using the Adam optimizer with default parameters. Our learning rate is initialized to $5e-5$ and decayed by a factor of 0.6 every 12k iterations. The training converges after 60k iterations and takes about 20 hours to complete on a single NVIDIA V100 GPU. We apply data augmentation (DA) techniques such as flips and rotations. The **CSR** network is fixed during the training of the **REC** network.

9.4 Experiments

In this section, we present experimental results that demonstrate the effectiveness of our fully self-supervised approach for Sentinel-2 SISR. To this aim, we conduct experiments on both real Sentinel-2 L1B data and a simulated dataset that we generated from Sentinel-2 L1C products. Through extensive ablation studies and quantitative analyses, we aim to demonstrate the efficacy of our method in addressing the challenges posed by the cross-spectral registration and SISR tasks in Sentinel-2 imagery. Additionally, we compare our self-supervised approach to a state-of-the-art supervised SISR method.

9.4.1 Simulated dataset

The simulated dataset used in our experiments was generated from 20 Sentinel-2 L1C products, with 18 used for training and 2 for testing. The products were extracted from 5 different continents in both summer and winter seasons to ensure geographic and radiometric diversity. For the training set, we selected 6,998 crops, each with a size of $512 \times 512 \times 4$ pixels. For the testing set, we selected 184 crops of the same size. Provided that Sentinel-2 imagery contains significant alias (Chapter 8) which is unsuitable to use as ground truth HR, we first applied a Gaussian kernel with $\sigma = 0.7$ to each crop to remove some aliasing, approximating the effect of an optical blur. The ground truth images I_0^{HR} and I_1^{HR} were then generated by applying 2 random homography transformations $\mathcal{H}_0$ and $\mathcal{H}_1$ to the blurred HR image (denoted as B^{HR}). These ground truth HR should be aligned to $I_{0,g}$ and $I_{1,g}$, respectively. We also simulated band-misalignment in the LR by applying a small homography transformation, where the translation component is dominant. Additionally, a little Gaussian noise (0.1%) was added to the LR to match the Sentinel-2 noise level. Overall, the simulation process can be summarized as follows:

$$\begin{aligned} I_t^{HR} &= \mathcal{H}_t(B^{HR}), \qquad t \in \{0,1\} \\ I_{t,g} &= \Pi_2(I_{t,g}^{HR}) + n_g, \\ I_{t,i} &= \Pi_2\left((\mathcal{H}_{t,i} \circ \mathcal{H}_t)(B_i^{HR})\right) + n_i, \quad i \neq g, \end{aligned} \tag{9.8}$$

where $\mathcal{H}_{t,i}$ $(i \in \{b,r,n\})$ is a translation-dominant homography modeling the band-misalignment between $I_{t,g}$ and $I_{t,i}$. n_i models the noise in the Sentinel-2 L1B. The largest distortion between 2 bands of 2 images can be up to 10 pixels. To enable diverse ablation studies for both the **CSR** and the **REC** networks, the homographies are stored as ground truth flows.

Table 9.1: Cross-spectral registration error (in pixel) of our self-supervised and supervised **CSR** networks over the synthetic test set ($184 \times 256 \times 256 \times 4$ pixels). The score of same-band registration is highlighted in **bold**.

		Target bands			
	Ref. bands	B	G	R	N
Self-supervised	B	**0.026**	0.035	0.039	0.106
	G	0.034	**0.026**	0.038	0.092
	R	0.035	0.037	**0.026**	0.104
	N	0.100	0.086	0.098	**0.027**
Supervised	B	**0.016**	0.028	0.029	0.088
	G	0.027	**0.014**	0.027	0.076
	R	0.029	0.027	**0.017**	0.090
	N	0.083	0.072	0.081	**0.017**

The Sentinel-2 L1C products are derived from the L1B products by the Ground Segment. During this process, the bands are aligned and resampled using camera altitude and geometric models. However, due to imperfect parameter estimation, there may be residual shifts between the bands. These shifts are typically less than 0.3 pixels [GBT+17].

9.4.2 Cross-spectral registration

Table 9.1 reports the mean absolute pixel error of the self-supervised and supervised **CSR** networks on our test set when aligning the reference bands $I_{0,i}$ to the target bands $I_{1,j}$ with $i, j \in \{b, g, r, n\}$. The first half of the table shows the performance of the self-supervised **CSR** network, where the diagonal entries correspond to the same-band registration. The self-supervised **CSR** network performs exceptionally well for the RGB bands, in particular for the green band, exhibiting low mis-registration (less than 0.04 pixel), indicating a high correlation between the RGB bands. However, the registration between NIR and RGB is much more challenging, with an error around 0.1 pixel, which is twice as large as that of the RGB bands, suggesting a much lower correlation between NIR and RGB bands.

To validate the effectiveness of our self-supervised approach, we also performed supervised training of **CSR** on the same training set, where we penalized the error between the output flows and the ground truth flows obtained from the homographies. The second half of Table 9.1 presents the results of the supervised model over the test set. The table shows a small gap between the performance of the two models, with a maximum mean error of 0.03 pixels for RGB and 0.09 pixels for NIR in the supervised setting, compared to 0.04 pixels and 0.11 pixels for self-supervision. Overall, the self-supervised **CSR** approach performs well for many applications, without requiring any ground truth flows, knowledge of the optical instrument or scene modeling.

9.4.3 Multi-band super-resolution

We conducted ablation experiments using the proposed synthetic L1C dataset to evaluate the performance of our self-supervised reconstruction network. We found that the residual misalignment in the L1C product affects PSNR measurements: a super-resolved result with well-aligned bands will be slightly misaligned with respect to the ground truth. To

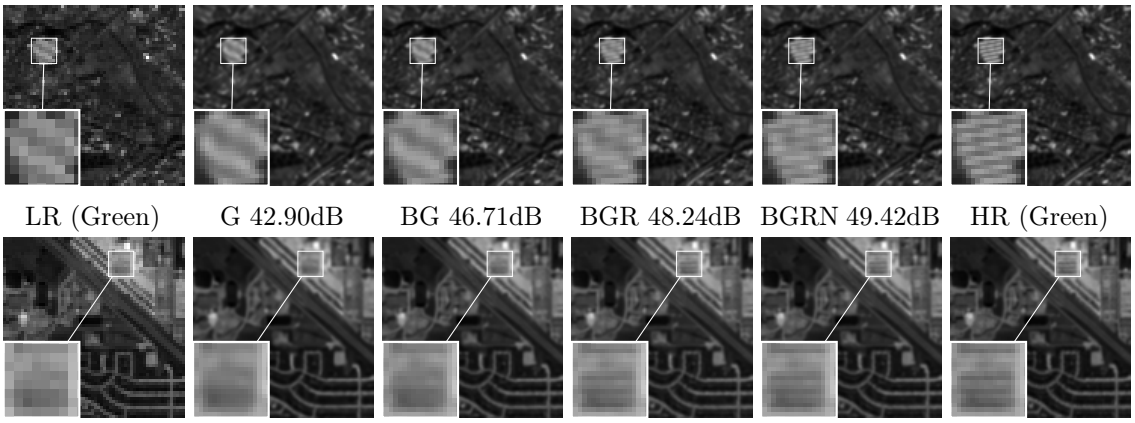


Figure 9.5: SR reconstruction of the green band when trained jointly with different input spectral bands. We observe that the other bands (B, R, N) provide valuable information for improving the reconstruction of the green band.

avoid this, we align each band of the ground truth to the corresponding band of the super-resolved output before computing the PSNR (note that this alignment is between images of the same band). We use a classical TV-L1 optical flow [PMLF13], setting the weight of the data attachment term to 0.3. The PSNRs shown in Tables 9.2 and 9.3 were computed in this way.

First, we studied the impact of the number of input bands on the reconstruction quality. Table 9.2 shows the PSNR results for four different networks trained with different input bands. We observed a significant improvement in performance as we increased the number of bands provided to the **REC** network, consistent with previous findings (Chapter 8).

Furthermore, Figure 9.5 illustrates the performance improvements in restoring the G band by providing different bands as input. As more bands are used, the network is able to restore aliased patterns into the true pattern and reach higher signal-to-noise ratio.

Secondly, we compared the performance of our self-supervised framework against a supervised training approach using the synthetic L1C dataset. The supervised training minimized an L_1 loss between the restoration and the ground truth HR image. Table 9.3 shows the per-band mean PSNRs over the test set. We observe a significant PSNR gap in favour of the proposed self-supervised L1BSR method (0.7dB for the visible bands and 0.3 for the near-infrared). This is rather unexpected. Self-supervised training is at best equivalent to supervised training [LMH+18, EDM+19, DAD+21, NAD+21b]. The worse performance of the supervised method can be explained by the residual misalignment of the L1C images (see Section 9.4.1) used as ground truth during training, resulting in blurry super-resolved images. The self-supervised method does not suffer from this problem: during training each band of the target image is aligned to the green band of the super-resolved image by the **CSR** module. It should be possible to compensate the misalignment of the ground truth in a supervised setting by aligning the ground truth bands to the super-resolved prediction (see for example [DKG+20]). We did not explore this approach here as our interest lies mainly in the self-supervised training.

Table 9.2: Multi-band super-resolution with the Green band being the reference. PSNRs obtained after aligning GT bands to SR bands. Best PSNR in **bold**.

	Testing bands			
Training bands	B	G	R	N
G		46.39		
BG	50.87	49.77		
BGR	51.62	51.09	48.51	
BGRN	**51.80**	**51.67**	**48.82**	41.34

Table 9.3: Comparison with supervised training. PSNRs obtained after aligning GT bands to SR bands. Best PSNR in **bold**.

	Testing bands			
Methods	B	G	R	N
Supervised	50.90	51.04	48.14	40.98
L1BSR	**51.80**	**51.67**	**48.82**	**41.34**

9.4.4 Real L1B dataset

The Sentinel-2 MSI includes 12 detectors for the VNIR bands arranged on a focal plane. The L1B product is composed of individual rasters, one per detector and per band. By design, two successive detectors acquire the scene with significant overlap, i.e. 120-200 pixels for the RGBN bands, which allows us to train our model on real L1B data.

However, due to the sensor layout, there is a noticeable vertical offset between consecutive detectors. To prepare the training and testing datasets, we pre-register the bands from different detectors using an integer translation. This process involves estimating a coarse translation between detectors using a SIFT-based matching method, refining the offset for each crop using an optical flow method, and regressing an integer translation. Since the bands of a given detector are acquired almost simultaneously, it is possible to ensure a registration precision up to a few pixels. For overlaps, registration is less precise due to the time delay, induced parallax and viewpoint changes, but typically remains less than 10 pixels. Overall, a refined registration is not required for our framework since the **CSR** network accurately captures residual shifts.

We used only two Sentinel-2 L1B products in this work due to their current scarcity, and extracted training data around overlaps, consisting of 3740 pairs of height 400 pixels and width depending on the overlap width between detectors for RGBN bands.[1] For testing we extracted crops outside of the overlapping regions, typically near the center of the detectors.

9.4.5 Qualitative analysis

We train the **CSR** and **REC** networks sequentially as described in Section 9.3 on the L1B dataset. For **REC**, the $\ell^*_{\text{Self-SR}}$ loss is used. The reconstruction results over the images of

[1] Note that L1B products will be systematically available through the Copernicus Data Space Ecosystem starting October 2023.

(a) L1B (b) L1BSR

(c) L1C (d) L1C-based SR (Chapter 8)

Figure 9.6: SR (x2) reconstruction by our proposed method (b) using L1B and the method in Chapter 8 (d) using L1C imagery. (a) and (c) are from the same acquisition.

the test set are shown in Figure 9.1. The **REC** network is able to successfully restore a high-quality HR image with aligned bands and fine details recovered.

To evaluate our self-supervised method, we compare it against our $\times 2$ SISR method in Chapter 8. The method in Chapter 8 relies on ground truth data obtained from PlanetScope imagery. It is trained with a L_1 loss, and is designed to work with Sentinel-2 L1C products. For comparison, we identified the L1C products from the same acquisition as the L1B samples and extracted the corresponding crops. Figure 9.6 shows the SR results of the proposed methods on L1B, and ours on L1C (Chapter 8). We found that the recovered details are very similar, yet our approach does not suffer from the gap between Sentinel-2 and PlanetScope images, which results in radiometric and geometric degradation by the network.

9.5 Chapter summary

This chapter presents a novel self-supervised method, L1BSR, for single-image super-resolution of Sentinel-2 L1B 10m bands. Leveraging the unique hardware design of Sentinel-2, the proposed method achieves remarkable performance without the need for

HR ground truth images. By training a cross-spectral registration module using an innovative self-supervised loss, Anchor-Consistency, L1BSR is capable of reconstructing high-resolution outputs with all bands correctly registered. To the best of our knowledge, our work is the first to explore the overlapping regions between detectors of Sentinel-2 and their potential for self-supervised SISR. Through an ablation study on a synthetic dataset and comparison with other supervised SISR methods for Sentinel-2 L1C images, we have demonstrated that L1BSR achieves performance on par with that of supervised training. The availability of our training dataset on the project website makes it a valuable resource for various image restoration and cross-spectral registration tasks.[2] As we move into the concluding discussions of this thesis, the advancements made in this chapter set a strong precedent for the further development and refinement of self-supervised methods in satellite image super-resolution.

[2]https://centreborelli.github.io/L1BSR/

10 Conclusion

Synopsis of thesis and key contributions

This thesis represents the chronicle of my exploration of multi-image and single-image super-resolution in satellite imagery. Each chapter uncovers a different facet, unravels a unique challenge, and showcases a novel solution.

Our models in this thesis, though rooted in deep-learning, lean heavily on principles of signal processing. This delicate balance between advanced neural networks and fundamental mathematical principles forms the crux of our models. Theory guides us, but the data teaches us, fostering a dynamic, symbiotic relationship.

In Part I, "Multi-image super-resolution in satellite imagery", we commenced by examining the multi-date PROBA-V super-resolution dataset. The inherent challenges posed by this dataset led us to propose a more practical variant known as PROBA-V-REF. This improved dataset explicitly identifies the reference frame that corresponds to the HR target image, thereby more meaningful for most real-world applications.

Our exploration continued with the development of the Deep Shift-and-Add (DSA) framework, specifically designed for push-frame satellite sensors like the SkySat constellation by Planet. DSA stands as the pioneering self-supervised MISR methodology. The unique amalgamation of shift-and-add fusion and self-supervised learning in DSA enables effective training in the absence of high-resolution ground truth. Especially, the shift-and-add fusion allows DSA to retain permutation-invariance and efficiently handle a variable number of frames - qualities not often found in other deep learning techniques. Achieving the state-of-the-art in MISR for satellite imagery, DSA forms part of an extensive super-resolution toolbox developed in this thesis.

We improved DSA with detail-preserving control, which refines super-resolved images by retaining vital details often lost in traditional super-resolution methods. An outlier detection feature was integrated, bolstering DSA's ability to process input images with substantial discrepancies and enhancing the robustness of the super-resolution process.

We further extended the potential of the DSA framework with HDR-DSP, to tackle multi-exposure sequences in remote sensing. HDR-DSP is the first integrated solution for joint super-resolution and High Dynamic Range (HDR) imaging. Uniquely, it is capable of handling signal-dependent noise and significant inaccuracies in reported exposure time. HDR-DSP not only improves the performance of DSA, but also extends its applicability to a broader array of real-world scenarios, enhancing its practical utility.

As we shifted our focus to "Single-image super-resolution in satellite imagery" in the

second part of this thesis, we first engaged with the state-of-the-art image restoration model, SwinIR. With perceptual loss, SwinIR was effective on most real-world images. However we found it to have limitations with satellite imagery. This opened our eyes to the importance of crafting domain-specific strategies for the genuine super-resolution of satellite data.

In our pursuit of these strategies, we discovered the hidden potential within the distinct characteristics of Sentinel-2. Rather than perceiving aliasing and inter-band shift as hurdles, we leveraged them for true high-frequency detail recovery. This redefining approach transitioned SISR problem towards a more favourable scenario, akin to MISR, with promising experimental results.

Further deepening our exploration, we proposed an innovative approach for self-supervised joint SISR and band-alignment of Sentinel-2 Level 1B (L1B) products. This novel method took advantage of the unique detector overlap feature inherent in Sentinel-2's sensor design. Despite the complex band misalignment challenges presented by L1B products, our self-supervised framework demonstrated a performance comparable to supervised methods. The introduction of an innovative cross-spectral registration (CSR) module proved instrumental in overcoming these challenges and offered a unique contribution to our super-resolution toolbox.

Potential impact

The potential implications of this thesis span both academic and practical dimensions, contributing to the progress of remote sensing research and enhancing the capabilities of satellite imagery utilization.

For the academic landscape, the techniques and methodologies proposed in this thesis offer a new perspective on super-resolution applications for satellite imagery, particularly with the introduction of the DSA, HDR-DSP, and L1BSR frameworks. These self-supervised learning mechanisms can serve as robust foundations for future research, facilitating additional developments and innovations in the field.

From a practical standpoint, the improved quality of satellite imagery achievable through the techniques discussed in this thesis can significantly enhance the level of detail and accuracy in data obtained for a variety of applications.

In climate change studies, more precise satellite data can provide an increased level of detail, facilitating more accurate monitoring and modelling of environmental changes, from polar ice melt to deforestation rates.

For agricultural applications, the implications are similarly profound. Precision farming, for instance, could benefit immensely from enhanced satellite imagery. Detailed information regarding crop health, soil conditions, and irrigation requirements could all be made more readily available, aiding in optimizing farming operations.

Urban planning and development could also see significant advantages. High-resolution satellite imagery can assist planners in accurately tracking city growth, monitoring land use changes, and making informed infrastructure development decisions.

Beyond these fields, the potential impact extends to disaster management, military surveillance, archaeology, and many more areas where improved satellite imagery could provide considerable benefits. While the advancements presented in this thesis are significant,

they represent potential pathways for continued progress and innovation in the world of satellite imagery and super-resolution.

Future research directions

This thesis has sought to extend the boundaries of super-resolution techniques in satellite imagery. However, the journey doesn't end here. The principles, methodologies, and algorithms developed can act as stepping stones for an array of intriguing future research directions. Here are a few potential pathways to be explored.

- MISR for Satellogic's Aleph-1: Aleph-1 is another push-frame satellite. Each Aleph-1 frame is multi-spectral with inherent band misalignment. Therefore, the L1BSR method (Chapter 9) could be initially applied for band-registration, followed by the use of the DSA framework (Chapter 4) for MISR.
- Complete super-resolution of Sentinel-2: The potential for L1BSR to super-resolve all 13 bands of Sentinel-2 Level 1B is discussed. This process involves a recursive fusion approach: first, 60m bands are enhanced to 20m GSD. These are then incorporated with existing 20m bands and upscaled to 10m. The final step is to uniformly super-resolve all 10m bands to 5m resolution. In each phase, higher-resolution bands also guide the lower-resolution bands' super-resolution, akin to pan-sharpening.
- Transfer learning from L1BSR to L1C: The orthorectified nature of Sentinel-2 L1C products makes them better suited for global analysis when compared to L1B products. Given this advantage, it's crucial to consider extending the super-resolution methodologies of L1BSR, initially developed for L1B images, to Sentinel-2's L1C products.
- Extension of CSR applications: The Cross-Spectral Registration (CSR) module, introduced within the L1BSR framework in Chapter 9, showcases an ability to refine band-alignment in various commercial satellites, even without fine-tuning. This suggests a promising avenue for the expansion of CSR for broader cross-modal registration applications. This includes, but is not limited to, SAR/optical registration and medical image registration.
- Optical flow improvement: Satellite imagery benefits from simple motion, with minimal occlusion and parallax. However, extending the methodologies presented in this thesis to real-world data such as video super-resolution or raw burst super-resolution requires dealing with more complex motion. Future research should focus on the development or adoption of more sophisticated optical flow models to handle these scenarios.

Final Remarks

The journey undertaken in this thesis has offered a fresh perspective on satellite imagery super-resolution, interlinking deep-learning and signal processing principles to push the boundaries of the field. We have dissected the challenges and potentials of satellite imagery, providing a robust platform for future investigation.

Yet, the work here is just the beginning. As satellite technology progresses and new data streams continually emerge, so too will the landscape of super-resolution evolve. The

hope is that this thesis will serve as a launchpad for more research, fostering continuous innovation in this dynamic field.

To the readers, my colleagues, and future researchers, I extend my deepest gratitude for your interest in this work. I hope this work proves insightful, inspires innovation, and lays the groundwork for revolutionary advancements in super-resolution methodologies.

Bibliography

[AAD+17] Cecilia Aguerrebere, Andrés Almansa, Julie Delon, Yann Gousseau, and Pablo Musé. A bayesian hyperprior approach for joint image denoising and interpolation, with an application to hdr imaging. *IEEE Transactions on Computational Imaging*, 3(4):633–646, 2017. 86

[ABHY00] 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. 62, 66, 87

[ACHR06] Andrés Almansa, Vicent Caselles, Gloria Haro, and Bernard Rougé. Restoration and zoom of irregularly sampled, blurred, and noisy images by accurate total variation minimization with local constraints. *Multiscale Modeling & Simulation*, 5(1):235–272, 2006. 12, 30

[AD18] Miika Aittala and Frédo Durand. Burst image deblurring using permutation invariant convolutional neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 731–747, 2018. 90

[ADF19] Jérémy Anger, Mauricio Delbracio, and Gabriele Facciolo. Efficient blind deblurring under high noise levels. In *IEEE ISPA*, pages 123–128, 2019. 69, 72

[AdFF19] Jérémy Anger, Carlo de Franchis, and Gabriele Facciolo. Assessing the sharpness of satellite images: Study of the planetscope constellation. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 389–392. IEEE, 2019. 9, 27, 118

[ADGM13] Cecilia Aguerrebere, Julie Delon, Yann Gousseau, and Pablo Muse. Simultaneous hdr image reconstruction and denoising for dynamic scenes. In *IEEE International Conference on Computational Photography (ICCP)*, pages 1–11. IEEE, 2013. 86

[ADGM14] Cecilia Aguerrebere, Julie Delon, Yann Gousseau, and Pablo Musé. Best algorithms for hdr image generation. a study of performance bounds. *SIAM Journal on Imaging Sciences*, 7(1):1–34, 2014. 86, 89, 100

[AEdFF20] 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 & Spatial Information Sciences*, 5(1), 2020. 10, 12, 22, 28, 31, 41, 50, 57, 58, 60, 66, 67, 69, 72, 75, 78, 83, 84, 85, 95, 101, 102, 103, 123

[AEF21] 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. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 2679–2682. IEEE, 2021. 60, 75, 84

[AGD+20] Moab Arar, Yiftach Ginger, Dov Danon, Amit H Bermano, and Daniel Cohen-Or. Unsupervised multi-modal image registration via geometry preserving image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13410–13419, 2020. 127

[AMSC+20] Md Rifat Arefin, Vincent Michalski, Pierre-Luc St-Charles, Alfredo Kalaitzis, Sookyung Kim, Samira E. Kahou, and Yoshua Bengio. Multi-image super-resolution for remote sensing using deep recurrent networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 206–207, 2020. 13, 32, 123

[AT17] 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. 16, 34, 58, 118

[BCM05] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 2, pages 60–65. Ieee, 2005. 16, 35

[BCM11] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. Non-local means denoising. *Image Processing On Line*, 1:208–212, 2011. 111

[BDVGT21] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deep burst super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9209–9218, 2021. 16, 34, 62, 85

[BFS18] Thibaud Briand, Gabriele Facciolo, and Javier Sánchez. Improvements of the Inverse Compositional Algorithm for Parametric Motion Estimation. *IPOL*, 8:435–464, 2018. 66, 95, 101

[BK02] 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. 127

[BKSI19] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. *Advances in Neural Information Processing Systems*, 32, 2019. 16, 35

[BM01] Simon Baker and Iain Matthews. Equivalence and efficiency of image alignment algorithms. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001. 51, 66, 95

[BM18] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 14, 33, 76, 77, 115, 119, 120

[BMX+19] 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. 59

[BP22] Haoran Bai and Jinshan Pan. Self-supervised deep blind video super-resolution. *arXiv preprint arXiv:2201.07422*, 2022. 16, 35

[BR19] Joshua Batson and Loic Royer. Noise2self: Blind denoising by self-supervision. In *International Conference on Machine Learning*, pages 524–533. PMLR, 2019. 16, 35, 61, 64, 91

[BRE19] Alon Brifman, Yaniv Romano, and Michael Elad. Unified single-image and video super-resolution via denoising algorithms. *IEEE Transactions on Image Processing*, 28(12):6063–6076, 2019. 13, 32

[CB13] Miguel Colom and Antoni Buades. Analysis and extension of the ponomarenko et al. method, estimating a noise curve from a single image. *Image Processing On Line*, 3:173–197, 2013. 94

[CBFAB97] Pierre Charbonnier, Laure Blanc-Féraud, Gilles Aubert, and Michel Barlaud. Deterministic edge-preserving regularization in computed imaging. *IEEE Transactions on image processing*, 6(2):298–311, 1997. 14, 32

[ÇBM+20] Uğur Çoğalan, Mojtaba Bemana, Karol Myszkowski, Hans-Peter Seidel, and Tobias Ritschel. Hdr denoising and deblurring by learning spatio-temporal distortion models. *arXiv preprint arXiv:2012.12009*, 2020. 86

[CDL18] Kan Chang, Pak Lun Kevin Ding, and Baoxin Li. Single image super resolution using joint regularization. *IEEE Signal Processing Letters*, 25(4):596–600, 2018. 14, 33

[CLK21] Young-Ju Choi, Young-Woon Lee, and Byung-Gyu Kim. Wavelet attention embedding networks for video super-resolution. In *2020 25th International conference on pattern recognition (ICPR)*, pages 7314–7320. IEEE, 2021. 13, 32

[COK22] Julien Cornebise, Ivan Oršolić, and Freddie Kalaitzis. Open high-resolution satellite imagery: The worldstrat dataset–with application to super-resolution. *Advances in Neural Information Processing Systems*, 35:25979–25991, 2022. 16, 34

[CSCL20] Si-Yuan Cao, Hui-Liang Shen, Shu-Jie Chen, and Chunguang Li. Boosting structure consistency for multispectral and multimodal image registration. *IEEE Transactions on Image Processing*, 29:5147–5162, 2020. 127

[CVTGC+11] Gustavo Camps-Valls, Devis Tuia, Luis Gómez-Chova, Sandra Jiménez, and Jesús Malo. Remote sensing image processing. 2011. 9, 27

[CWG+21] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained

image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021. 108

[CWY+21] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4947–4956, 2021. 13, 14, 32

[CXLT18] Mengyu Chu, You Xie, Laura Leal-Taixé, and Nils Thuerey. Temporally coherent gans for video super-resolution (tecogan). *CoRR*, abs/1811.09393, 2018. 14, 32

[CZY+19] 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. 15, 34, 58, 126

[DAD+21] 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. 17, 36, 59, 61, 64, 84, 91, 124, 133

[DBK+20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 107

[DDBC+12] Matthias Drusch, Umberto Del Bello, Sébastien Carlier, Olivier Colin, Veronica Fernandez, Ferran Gascon, Bianca Hoersch, Claudia Isola, Paolo Laberinti, Philippe Martimort, et al. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote sensing of Environment*, 120:25–36, 2012. 123

[DFKE07] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8):2080–2095, 2007. 16, 35

[DKG+20] 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. 13, 14, 32, 50, 51, 52, 60, 62, 66, 84, 85, 95, 101, 133

[DLHT14] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pages 184–199. Springer, 2014. 15, 33

[DZL20] 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. 60

[EDAF19] 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. 17, 36, 59, 61, 84, 86, 124

[EDM+19] 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. 17, 36, 59, 61, 63, 84, 91, 124, 133

[EPC21] Mohammad Emad, Maurice Peemen, and Henk Corporaal. Dualsr: Zero-shot dual learning for real-world super-resolution. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1630–1639, 2021. 16, 35

[FAAC09] 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. 12, 30, 60

[FGS+95] Hans G Feichtinger, Karlheinz Gr, Thomas Strohmer, et al. Efficient numerical methods in non-uniform sampling theory. *Numerische Mathematik*, 69(4):423–440, 1995. 78

[FGT19] 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. 60

[FH02] 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. 62, 87, 101

[FJP02] William T Freeman, Thouis R Jones, and Egon C Pasztor. Example-based super-resolution. *IEEE Computer graphics and Applications*, 22(2):56–65, 2002. 14, 16, 33, 35

[FREM04a] Sina Farsiu, Dirk Robinson, Michael Elad, and Peyman Milanfar. Advances and challenges in super-resolution. *International Journal of Imaging Systems and Technology*, 14(2):47–57, 2004. 12, 13, 31

[FREM04b] Sina Farsiu, Dirk Robinson, Michael Elad, and Peyman Milanfar. Fast and Robust Multiframe Super Resolution. *IEEE Transactions on Image Processing*, 13(10):1327–1344, oct 2004. 11, 30

[FREM04c] Sina Farsiu, Dirk Robinson, Michael Elad, and Peyman Milanfar. Robust shift and add approach to superresolution. In *Applications of Digital Image Processing XXVI*, 2004. 13, 14, 31, 32, 75, 78, 79, 81

[FTKE08] Alessandro Foi, Mejdi Trimeche, Vladimir Katkovnik, and Karen Egiazarian. Practical poissonian-gaussian noise modeling and fitting for single-

image raw-data. *IEEE Trans. Image Process.*, 17(10):1737–1754, 2008. 87

[GAW+10] Miguel Granados, Boris Ajdin, Michael Wand, Christian Theobalt, Hans-Peter Seidel, and Hendrik PA Lensch. Optimal hdr reconstruction with linear digital cameras. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 215–222. IEEE, 2010. 86, 88, 89, 100

[GB10] 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. 64, 130

[GBI09] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *2009 IEEE 12th international conference on computer vision*, pages 349–356. IEEE, 2009. 14, 33

[GBT+17] Ferran Gascon, Catherine Bouzinac, Olivier Thépaut, Mathieu Jung, Benjamin Francesconi, Jérôme Louis, Vincent Lonjou, Bruno Lafrance, Stéphane Massera, Angélique Gaudel-Vacaresse, et al. Copernicus sentinel-2a calibration and products validation status. *Remote Sensing*, 9(6):584, 2017. 120, 125, 132

[GCLK08] 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. 62, 66, 87, 101

[GMG+19] Massimiliano Gargiulo, Antonio Mazza, Raffaele Gaetano, Giuseppe Ruello, and Giuseppe Scarpa. Fast super-resolution of 20 m sentinel-2 bands using convolutional neural networks. *Remote Sensing*, 11(22):2635, 2019. 117, 126

[Gre09] Hayit Greenspan. Super-resolution in medical imaging. *The computer journal*, 52(1):43–63, 2009. 14, 33

[GSA+20] Mikel Galar, Rubén Sesma, Christian Ayala, Lourdes Albizua, and Carlos Aranda. Super-resolution of sentinel-2 images using convolutional neural networks and real ground truth data. *Remote Sensing*, 12(18):2941, 2020. 117, 120, 124, 126

[HFBP+18] Juan Mario Haut, Ruben Fernandez-Beltran, Mercedes E Paoletti, Javier Plaza, Antonio Plaza, and Filiberto Pla. A new deep generative network for unsupervised remote sensing single-image super-resolution. *IEEE Transactions on Geoscience and Remote sensing*, 56(11):6792–6810, 2018. 14, 33

[HLVDMW17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 15, 34

[HSA15] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE*

conference on computer vision and pattern recognition, pages 5197–5206, 2015. 14, 33

[HSG+16] Samuel W. Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T. Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (ToG)*, 35(6):1–12, 2016. 86

[HSS18] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 128

[HST+14] Felix Heide, Markus Steinberger, Yun-Ta Tsai, Mushfiqur Rouf, Dawid Pajak, Dikpal Reddy, Orazio Gallo, Jing Liu, Wolfgang Heidrich, Karen Egiazarian, et al. Flexisp: A flexible camera image processing framework. *ACM Transactions on Graphics (ToG)*, 33(6):1–13, 2014. 86

[HSU19] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3897–3906, 2019. 13, 32

[HZ03] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 11, 30

[HZRS15] 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. 64

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 15, 34

[IJG+20a] 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. 13, 32, 60

[IJG+20b] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. Video Super-Resolution with Recurrent Structure-Detail Network. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2020. 89

[IZZE17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 127

[JAFF16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 694–711, Cham, 2016. Springer International Publishing. 118

[Jia12] Yunwei Jia. Method and apparatus for super-resolution of images, November 6 2012. US Patent 8,306,121. 62, 66, 87, 101

[JM18] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018. 118

[JMX+21] Xingyu Jiang, Jiayi Ma, Guobao Xiao, Zhenfeng Shao, and Xiaojie Guo. A review of multimodal image matching: Methods and applications. *Information Fusion*, 73:22–71, 2021. 127

[JWOKJK18] 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. 13, 32, 60

[JWY+22] Kai Jin, Zeqiang Wei, Angulia Yang, Sha Guo, Mingzhi Gao, Xiuzhuang Zhou, and Guodong Guo. Swinipassr: Swin transformer based parallax attention network for stereo image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 920–929, 2022. 108

[KB14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 64, 92, 131

[KBJ19] 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. 61

[KBP+19] Michal Kawulok, Pawel Benecki, Szymon Piechaczek, Krzysztof Hrynczenko, Daniel Kostrzewa, and Jakub Nalepa. Deep learning for multiple-image super-resolution. *IEEE Geoscience and Remote Sensing Letters*, 17(6):1062–1066, 2019. 13, 32

[KBV90] 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. 12, 31, 60

[KF09] Dilip Krishnan and Rob Fergus. Fast image deconvolution using hyper-laplacian priors. In *NIPS*, pages 1033–1041, 2009. 69, 72

[KJK20] Jonghee Kim, Chanho Jung, and Changick Kim. Dual back-projection-based internal learning for blind super-resolution. *IEEE Signal Processing Letters*, 27:1190–1194, 2020. 16, 35

[KK18] Soo Ye Kim and Munchurl Kim. A multi-purpose convolutional neural network for simultaneous super-resolution and high dynamic range image reconstruction. In *Asian Conference on Computer Vision*, pages 379–394. Springer, 2018. 86

[KKLML16] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the*

IEEE conference on computer vision and pattern recognition, pages 1646–1654, 2016. 15, 34

[KLNK18] Soo Ye Kim, Jeongyeon Lim, Taeyoung Na, and Munchurl Kim. 3dsr-net: Video super-resolution using 3d convolutional neural networks. *arXiv preprint arXiv:1812.09079*, 2018. 13, 32

[KOK19] Soo Ye Kim, Jihyong Oh, and Munchurl Kim. Deep SR-ITM: Joint learning of super-resolution and inverse tone-mapping for 4K UHD HDR applications. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3116–3125, 2019. 86, 89

[KOK20] Soo Ye Kim, Jihyong Oh, and Munchurl Kim. Jsi-gan: Gan-based joint super-resolution and inverse tone-mapping with pixel-wise task-specific filters for uhd hdr video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11287–11295, 2020. 86

[KPB88] Danny Keren, Shmuel Peleg, and Rafi Brada. Image sequence enhancement using sub-pixel displacements. In *CVPR 88*, pages 742–743, 1988. 12, 13, 31, 60

[KPL+20] 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. 61, 126

[KRV] Emiliano Kargieman, Gerado Gabriel Richarte, and Juan Manuel Vuletich. Imaging device for scenes in apparent motion. U.S. Patent 9813601B2, issued November 7, 2017. 86

[KSD+20] Mary Knapp, Sara Seager, Brice-Olivier Demory, Akshata Krishnamurthy, Matthew W. Smith, Christopher M. Pong, Vanessa P. Bailey, Amanda Donner, Peter Di Pasquale, Brian Campuzano, Colin Smith, Jason Luu, Alessandra Babuscia, Robert L. Bocchino, Jr., Jessica Loveland, Cody Colley, Tobias Gedenk, Tejas Kulkarni, Kyle Hughes, Mary White, Joel Krajewski, and Lorraine Fesq. Demonstrating High-precision Photometry with a CubeSat: ASTERIA Observations of 55 Cancri e. *The Astronomical Journal*, 160(1):23, jun 2020. 86

[KYDK16] Armin Kappeler, Seunghwan Yoo, Qiqin Dai, and Aggelos K Katsaggelos. Video super-resolution with convolutional neural networks. *IEEE Transactions on Computational Imaging*, 2(2):109–122, 2016. 14, 32

[LAZW18] Xiangzeng Liu, Yunfeng Ai, Juli Zhang, and Zhuping Wang. A novel affine and contrast invariant descriptor for infrared and visible image registration. *Remote Sensing*, 10(4):658, 2018. 127

[LBDG+18] Charis Lanaras, José Bioucas-Dias, Silvano Galliani, Emmanuel Baltsavias, and Konrad Schindler. Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 146:305–319, 2018. 117, 126

[LCF+22] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *arXiv preprint arXiv:2201.12288*, 2022. 14, 32

[LCS+21] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021. 15, 34, 108

[LDT19] Andreas Lugmayr, Martin Danelljan, and Radu Timofte. Unsupervised learning for real-world super-resolution. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3408–3416. IEEE, 2019. 126

[LGZ13] Jing Liu, Zongliang Gan, and Xiuchang Zhu. Directional bicubic interpolation—a new method of image super-resolution. In *3rd International Conference on Multimedia Technology (ICMT-13)*, pages 463–470. Atlantis Press, 2013. 14, 33

[LHD+19] Sheng Li, Fengxiang He, Bo Du, Lefei Zhang, Yonghao Xu, and Dacheng Tao. Fast spatio-temporal residual network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10522–10531, 2019. 13, 32

[LHS+09] Daewon Lee, Matthias Hofmann, Florian Steinke, Yasemin Altun, Nathan D Cahill, and Bernhard Scholkopf. Learning similarity measure for multi-modal 3d image registration. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 186–193. IEEE, 2009. 127

[LKC15] Thomas Lillesand, Ralph W Kiefer, and Jonathan Chipman. *Remote sensing and image interpretation*. John Wiley & Sons, 2015. 9, 27

[LLC+21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 108, 110

[LLC+22] Ziwei Luo, Youwei Li, Shen Cheng, Lei Yu, Qi Wu, Zhihong Wen, Haoqiang Fan, Jian Sun, and Shuaicheng Liu. Bsrt: Improving burst superresolution with swin transformer and flow-guided deformable alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 998–1008, 2022. 108

[LLKX19] Pengpeng Liu, Michael Lyu, Irwin King, and Jia Xu. Selflow: Self-supervised learning of optical flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4571–4580, 2019. 91, 92

[LLL+22] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tieyong Zeng. Transformer for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 457–466, 2022. 111

[LLTMK19a] Alice Lucas, Santiago Lopez-Tapia, Rafael Molina, and Aggelos K Katsaggelos. Generative adversarial networks and perceptual losses for video

super-resolution. *IEEE Transactions on Image Processing*, 28(7):3312–3327, 2019. 13, 14, 32

[LLTMK19b] 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. 61

[LMH+18] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2Noise: Learning Image Restoration without Clean Data. In *35th International Conference on Machine Learning, ICML 2018*, 2018. 16, 17, 35, 36, 60, 63, 91, 126, 133

[LMW+22] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 112

[LNFE23] Jamy Lafenetre, Ngoc Long Nguyen, Gabriele Facciolo, and Thomas Eboli. Handheld burst super-resolution meets multi-exposure satellite imagery. *arXiv preprint arXiv:2303.05879*, 2023. 25, 45, 123

[LPM21] Bruno Lecouat, Jean Ponce, and Julien Mairal. Lucas-kanade reloaded: End-to-end super-resolution from raw image bursts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2370–2379, 2021. 13, 32, 85

[LPME22] Bruno Lecouat, Jean Ponce, Julien Mairal, and Thomas Eboli. High Dynamic Range and Super-Resolution from Raw Image Bursts. *ACM Transactions on Graphics (TOG)*, 2022. To appear. 13, 32

[LR00] Christophe Latry and Bernard Rougé. Optimized sampling for CCD instruments: the supermode scheme. In *IGARSS*, volume 5, pages 2322–2324. IEEE, 2000. 60

[LSK+17] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPR Workshops*, pages 136–144, 2017. 15, 34

[LTH+17] 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. 15, 34, 70, 76

[MBC+18] Ben Mildenhall, Jonathan T Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll. Burst denoising with kernel prediction networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2502–2510, 2018. 13, 32

[MBH20] 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. 59

[MBP+09] Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Non-local sparse models for image restoration. In *2009 IEEE 12th international conference on computer vision*, pages 2272–2279. IEEE, 2009. 14, 33

[MDS] MDSP super-resolution and demosaicing datasets. https://users.soe.ucsc.edu/~milanfar/software/sr-datasets.html. Accessed: 2021-03-15. 58

[MHL+21] Matteo Maggioni, Yibin Huang, Cheng Li, Shuai Xiao, Zhongqian Fu, and Fenglong Song. Efficient multi-stage video denoising with recurrent spatio-temporal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3466–3475, 2021. 79

[MIKC19] Marcus Märtens, Dario Izzo, Andrej Krzic, and Daniël Cox. Super-resolution of proba-v images using convolutional neural networks. *Astrodynamics*, 3:387–402, 2019. 16, 35, 49, 51, 58, 60, 123

[Mil17] Peyman Milanfar. *Super-resolution imaging*. CRC press, 2017. 11, 30

[MN07] Maria Teresa Merino and Jorge Nunez. Super-resolution of remotely sensed images with variable-pixel linear reconstruction. *IEEE TGRS*, 45(5):1446–1457, 2007. 12, 31, 60, 62, 66, 87, 95, 101

[MO08] Antonio Marquina and Stanley J Osher. Image super-resolution by tv-regularization and bregman iteration. *Journal of Scientific Computing*, 37(3):367–382, 2008. 12, 14, 30, 33, 60

[MSS+14] Kiran Murthy, Michael Shearn, Byron D. Smiley, Alexandra H. Chau, Josh Levine, and 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. 10, 12, 13, 22, 28, 31, 41, 57, 60, 65, 75, 82, 83, 84, 92, 102, 103

[MVFM19] 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. 13, 14, 32, 50, 51, 52, 60, 85

[MVFM20] 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. 60, 84

[MVSIH22] Julien Michel, Juan Vinasco-Salinas, Jordi Inglada, and Olivier Hagolle. Sen2venμs, a dataset for the training of sentinel-2 super-resolution algorithms. *Data*, 7(7):96, 2022. 117, 126

[NAD+21a] 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. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 3881–3884. IEEE, jul 2021. 25, 44

[NAD+21b] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1121–1131, 2021. 22, 25, 41, 44, 80, 83, 84, 85, 89, 90, 91, 95, 102, 103, 123, 124, 133

[NAD+22a] Ngoc Long Nguyen, Jérémy Anger, Axel Davy, Pablo Arias, and Gabriele Facciolo. Self-supervised push-frame super-resolution with detail-preserving control

and outlier detection. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 131–134. IEEE, 2022. 25, 44

[NAD+22b] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1858–1868, 2022. 25, 44, 123, 124

[NAD+23] Ngoc Long Nguyen, Jérémy Anger, Axel Davy, Pablo Arias, and Gabriele Facciolo. L1bsr: Exploiting detector overlap for self-supervised single-image super-resolution of sentinel-2 l1b imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2012–2022, 2023. 25, 45

[NAR+23] Ngoc Long Nguyen, Jérémy Anger, Lara Raad, Bruno Galerne, and Gabriele Facciolo. On the role of alias and band-shift for sentinel-2 super-resolution. *arXiv preprint arXiv:2302.11494*, 2023. 25, 44, 124, 126

[Ngu22] Ngoc-Long Nguyen. A brief analysis of the swinir image super-resolution. *Image Processing On Line*, 12:582–589, 2022. 25, 44

[NM00] 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. 12, 31, 60

[NM14] Kamal Nasrollahi and Thomas B Moeslund. Super-resolution: a comprehensive survey. *Machine vision and applications*, 25(6):1423–1468, 2014. 59

[OABB85] Joan M. Ogden, Edward H. Adelson, James R. Bergen, and Peter J. Burt. Pyramid-based computer graphics. *RCA Engineer*, 30(5):4–15, 1985. 88

[PAB20] Ferdinand Pineda, Victor Ayma, and César Beltran. A generative adversarial network approach for super-resolution of sentinel-2 satellite images. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43:9–14, 2020. 117, 124, 126

[PC+19] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. 13, 32

[PJ07] Vorapoj Patanavijit and Somchai Jitapunkul. A lorentzian stochastic estimation for a robust iterative multiframe super-resolution reconstruction with lorentzian-tikhonov regularization. *EURASIP Journal on Advances in Signal Processing*, 2007:1–21, 2007. 12, 30

[PLPD18] 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. 60

[PLZ+07] Nikolay N. Ponomarenko, Vladimir V. Lukin, M.S. Zriakhov, Arto Kaarna, and Jaakko Astola. An automatic approach to lossy compression of aviris images. In *2007 IEEE International Geoscience and Remote Sensing Symposium*, pages 472–475. IEEE, 2007. 87, 94

[PMLF13] Javier Sánchez Pérez, Enric Meinhardt-Llopis, and Gabriele Facciolo. Tv-l1 optical flow estimation. *Image Processing On Line*, 2013:137–150, 2013. 133

[PWÖ+20] Nicolas Pielawski, Elisabeth Wetzer, Johan Öfverstedt, Jiahao Lu, Carolina Wählby, Joakim Lindblad, and Natasa Sladoje. Comir: Contrastive multimodal image representation for registration. *Advances in neural information processing systems*, 33:18433–18444, 2020. 127

[RDL+] Dirk Robinson, Jonathan Dyer, Joshua Levine, Brendan Hermalyn, Ronny Votel, and Matt William Messana. Controlling a line of sight angle of an imaging platform. U.S. Patent 10432866B2, issued October 1, 2019. 86

[RF18] 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. 60

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 128

[RK99] Seunghyeon Rhee and Moon Gi Kang. Discrete cosine transform based regularized high-resolution image reconstruction algorithm. *Optical Engineering*, 38(8):1348–1356, 1999. 12, 31

[ROF92] 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. 64, 91

[SCC20] Jae Woong Soh, Sunwoo Cho, and Nam Ik Cho. Meta-transfer learning for zero-shot super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3516–3525, 2020. 126

[SCH+16] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, pages 1874–1883, 2016. 13, 32, 109

[SCI18] 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. 16, 35, 61, 126

[SDW+17] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1279–1288, 2017. 13, 32

[SJC+22] Huihui Song, Yutong Jin, Yong Cheng, Bo Liu, Dong Liu, and Qingshan Liu. Learning interlaced sparse sinkhorn matching network for video super-resolution. *Pattern Recognition*, 124:108475, 2022. 13, 32

[SMKC20] 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. 13, 32, 60, 95, 101

[SMV+21] Dev Yashpal Sheth, Sreyas Mohan, Joshua L. Vincent, Ramon Manzorro, Peter A. Crozier, Mitesh M. Khapra, Eero P. Simoncelli, and Carlos Fernandez-Granda. Unsupervised deep video denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1759–1768, 2021. 84, 124

[SRMV20] Luis Salgueiro Romero, Javier Marcello, and Verónica Vilaplana. Super-resolution of Sentinel-2 imagery using generative adversarial networks. *Remote Sensing*, 12(15):2424, 2020. 60, 117, 126

[SVB18] 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. 13, 14, 32, 60, 62, 85, 89, 91, 92

[SVE19] Jacob Shermeyer and Adam Van Etten. The effects of super-resolution on object detection performance in satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 9, 27

[TA14] Yann Traonmilin and Cecilia Aguerrebere. Simultaneous high dynamic range and superresolution imaging without regularization. *SIAM Journal on Imaging Sciences*, 7(3):1624–1644, 2014. 84, 85

[TAH06] Matt W Thornton, Peter M Atkinson, and DA Holland. Sub-pixel mapping of rural land cover objects from fine spatial resolution satellite sensor imagery using super-resolution pixel-swapping. *International Journal of Remote Sensing*, 27(3):473–491, 2006. 9, 27

[TDV20] Matias Tassano, Julie Delon, and Thomas Veit. Fastdvdnet: Towards real-time deep video denoising without flow estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1354–1363, 2020. 13, 32

[TFM07] Hiroyuki Takeda, Sina Farsiu, and Peyman Milanfar. Kernel regression for image processing and reconstruction. *IEEE Transactions on image processing*, 16(2):349–366, 2007. 49, 60

[TGL+17] 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. 13, 14, 32, 60, 61, 62, 63, 64, 80, 85, 90

[TK95] 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. 60

[TLLG17] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image super-resolution using dense skip connections. In *Proceedings of the IEEE international conference on computer vision*, pages 4799–4807, 2017. 15, 34

[TMPE09] Hiroyuki Takeda, Peyman Milanfar, Matan Protter, and Michael Elad. Super-resolution without explicit subpixel motion estimation. *IEEE Transactions on Image Processing*, 2009. 12, 31

[TOS92] A Murat Tekalp, Mehmet K Ozkan, and M Ibrahim Sezan. High-resolution image reconstruction from lower-resolution image sequences and space-varying image restoration. In *ICASSP*, volume 3, pages 169–172, 1992. 12, 31

[Tsa84] Roger Tsai. Multiframe image restoration and registration. *Advance Computer Visual and Image Processing*, 1:317–339, 1984. 12, 30

[UPWB10] Markus Unger, Thomas Pock, Manuel Werlberger, and Horst Bischof. A convex approach for variational super-resolution. In *Pattern Recognition: 32nd DAGM Symposium, Darmstadt, Germany, September 22-24, 2010. Proceedings 32*, pages 313–322. Springer, 2010. 14, 33

[UVL18] 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. 61

[VBW13] Singanallur V Venkatakrishnan, Charles A Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *2013 IEEE global conference on signal and information processing*, pages 945–948. IEEE, 2013. 13, 32

[VSR18] Subeesh Vasu, Abhijeet Shenoi, and A.N. Rajagopazan. Joint hdr and super-resolution imaging in motion blur. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2885–2889. IEEE, 2018. 84, 86

[VSVV07] 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. 62, 89

[WCH20] Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep learning for image super-resolution: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3365–3387, 2020. 124

[WCY+19] 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. 60

[WGDE+19] 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. 12, 31, 59, 60, 86, 95

[WMJB21] Celyn Walters, Oscar Mendez, Mark Johnson, and Richard Bowden. There and back again: Self-supervised multispectral correspondence estimation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5147–5154. IEEE, 2021. 127

[WYW+19] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In Laura Leal-Taixé and Stefan Roth, editors, *Computer Vision – ECCV 2018 Workshops*, pages 63–79, Cham, 2019. Springer International Publishing. 15, 34, 118

[WZYW21] Wei Wang, Haochen Zhang, Zehuan Yuan, and Changhu Wang. Unsupervised real-world super-resolution: A domain adaptation perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4318–4327, 2021. 126

[XCW+19] 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. 60

[XMY+22] Han Xu, Jiayi Ma, Jiteng Yuan, Zhuliang Le, and Wei Liu. Rfnet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19679–19688, 2022. 127

[XNC+20] 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. 62

[XSM+21] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better. *Advances in Neural Information Processing Systems*, 34:30392–30400, 2021. 109

[YHD16] Jason J. Yu, 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. 64, 91

[YLXC15] Daiqin Yang, Zimeng Li, Yatong Xia, and Zhenzhong Chen. Remote sensing image super-resolution: Challenges and approaches. In *2015 IEEE international conference on digital signal processing (DSP)*, pages 196–200. IEEE, 2015. 9, 27

[YLZ+18] 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. 16, 35, 61, 126

[YPPJ20] 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. 17, 36, 59, 61, 84, 124

[YSBS17] Yuanxin Ye, Jie Shan, Lorenzo Bruzzone, and Li Shen. Robust registration of multimodal remote sensing images based on structural similarity. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5):2941–2958, 2017. 127

[YSL+16] 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. 59

[YWHM10] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010. 14, 33

[YWJ+19] 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. 60

[YYF+20] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5791–5800, 2020. 15, 34

[ZAK+20] 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. 59

[ZB22] Maialen Zabalza and Angela Bernardini. Super-resolution of sentinel-2 images using a spectral attention mechanism. *Remote Sensing*, 14(12):2890, 2022. 117, 126

[ZEP12] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7*, pages 711–730. Springer, 2012. 14, 33

[ZGFK16] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging*, 3(1):47–57, 2016. 14, 32

[ZLL+18] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. 107, 111, 128

[ZLVGT21] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings*

of the IEEE/CVF International Conference on Computer Vision, pages 4791–4800, 2021. 110

[ZTK+18] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018. 15, 34, 107

[ZTS+20] 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. 60

[ZZGZ17] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3929–3938, 2017. 107

[ZZH+18] Junhao Zhang, Masoumeh Zareapoor, Xiangjian He, Donghao Shen, Deying Feng, and Jie Yang. Mutual information based multi-modal remote sensing image registration using adaptive feature weight. *Remote Sensing Letters*, 9(7):646–655, 2018. 127

[ZZSL10] Liangpei Zhang, Hongyan Zhang, Huanfeng Shen, and Pingxiang Li. A super-resolution reconstruction algorithm for surveillance images. *Signal Processing*, 90(3):848–859, 2010. 14, 33