Revisiting remote sensing cross-sensor Single Image Super-Resolution: the overlooked impact of geometric and radiometric distortion

Julien Michel, Ekaterina Kalinicheva, Jordi Inglada

Revisiting remote sensing cross-sensor Single Image Super-Resolution: the overlooked impact of geometric and radiometric distortion

Julien Michel, Ekaterina Kalinicheva, Jordi Inglada

***Abstract*—In remote sensing, Single Image Super-Resolution can be learned from large cross-sensor datasets with matched High Resolution and Low Resolution satellite images, thus avoiding the domain gap issue that occurs when generating the Low Resolution image by degrading the High Resolution one. Yet cross-sensor datasets come with their own challenges, caused by the radiometric and geometric discrepancies that arise from using different sensors and viewing conditions. While those discrepancies can be prominent, their impact has been vastly overlooked in the literature, which often focuses on pursuing more complex models without questioning how they can be trained and fairly evaluated in a cross-sensor setting. This paper intends to fill this gap and provide insight on how to train and evaluate cross-sensor Single-Image Super-Resolution Deep Learning models. First, it investigates standard Image Quality metrics robustness to discrepancies and highlights which ones can actually be trusted in this context. Second, it proposes a complementary set of Frequency Domain Analysis based metrics that are tailored to measure spatial frequency restoration performances. Metrics tailored for measuring radiometric and geometric distortion are also proposed. Third, a robust training and evaluation strategy is proposed, with respect to discrepancies. The effectiveness of the proposed strategy is demonstrated by experiments using two widely used cross-sensor datasets: Sen2Venμs and Worldstrat. Those experiments also showcase how the proposed set of metrics can be used to achieve a fair comparison of different models in a cross-sensor setting. The code will be publicly available at https://github.com/Evoland-Land-Monitoring-Evolution/sisr4rs.git.**

***Index Terms*—Super-Resolution, Sentinel-2, Optical flows, Image Quality**

I. Introduction

SINGLE Image Super-Resolution (SISR), is the process of reconstructing a High Resolution (HR) image from a single low resolution (LR) observation by restoring or synthesizing HR details. In the computer vision field, after a first era of blind deconvolution or deblurring based on linear optimization and regularization priors [1], [2], Deep Learning based SISR has gained considerable attention in the last decade [3]–[5]. Meanwhile, in the remote sensing domain SISR has gradually gained an interest as an alternative or a complement to address the user need for ever higher resolution data [6]–[8]. In this context, SISR is envisioned either as a means to enhance lower resolution data from public archives such as Landsat-8 [9] or Sentinel-2 [10]–[12] in order to enable applications requiring higher resolutions for which Very High Resolution (VHR) commercial imagery would have to be purchased, or as a means to further increase the effective spatial resolution of the latter VHR imagery [13]–[15].

Julien Michel, Ekaterina Kalinicheva and Jordi Inglada are with CESBIO (Université de Toulouse, CNES, CNRS, INRAE, IRD, UT3), 18 avenue Edouard Belin, BPI 2801, TOULOUSE Cedex 9 31401, France.

A. The domain gap in SISR

As in most Deep Learning application fields, large datasets are required to achieve convergence of the optimization process and good generalization of the SISR models. In the general computer vision field, those images are harvested from the internet. In this context, corresponding LR images are derived from HR images by simulating image degradations [16]. However, using such simulated datasets leads to the so-called domain gap issue: the distribution of real images encountered at inference time differs from the distribution of simulated images generated for training, resulting in models that are not correctly adapted to the real data they will be used on. Efrat et al. [17] note that a critical concern is how well the synthetic forward model approximates real camera blur. In [18], Rad et al. show that applying these methods on real images, with unknown degradation from cameras, cell-phones, etc. often leads to poor results. More recently, Zhao et al. [19] analyzed how the down-sampling in the simulation process affects the training and performance evaluation, noting that super-resolution models are not correctly evaluated by using such a process. The domain gap is also well illustrated by Liu et al. [20] (fig. 2, p. 5463). In [21], the authors attempt to bridge the domain gap in reference-based Super-Resolution by integrating domain adaptation module in the training.

In remote sensing, the domain gap is even more critical. From a signal processing point of view, sensors are well characterized: optics and detector properties are carefully engineered and monitored during the lifetime of the instrument, motion is known with great accuracy and depth can be considered constant. The whole system acts as a relatively stable spatial low pass filter that is usually summarized by its cut-off spatial frequency for each spectral band [22] and its Ground Sampling Distance (GSD). On the other hand, there is a great diversity of designs from one sensor to another, including, but not limited to, different spectral bands, different Spectral Sensitivity Responses (SSR), or different designs of their focal planes [23], [24]. Simulating a SISR dataset for a given sensor implies using data from another sensor with a smaller GSD, which may have completely different characteristics. This makes the simulation impractical or unrealistic in many cases. However, remote sensing optical images are not used

only for visualization: they represent physical measurements of surface reflectances that are further processed by means of physical or mathematical modeling in order to derive value added indicators such as Essential Climate Variables [25] or Essential Biodiverisity Variables [26]. For instance, Sentinel-2 applications include vegetation monitoring [27], Land Cover and Land Use Mapping [28], European Common Agricultural Policy control [29] and monitoring of Water Bodies [30], . Current applications of SISR in remote sensing documented in the literature include object detection [31], [32], water bodies mapping [33], [34], and agriculture mapping [35], [36]. All those applications require meaningful and physically correct surface reflectance measurements as input. Therefore, remote sensing SISR bears far more correctness and accuracy expectations from downstream applications than the general computer vision SISR.

B. Cross-sensor SISR remote sensing datasets

Unlike the general computer vision field, where such process would be impractical [37], corresponding LR and HR remote-sensing images can be acquired by leveraging different sensors with different resolutions observing the same area, leading to cross-sensor datasets as opposed to simulated datasets. Cross-sensor datasets avoid the domain gap issue at inference time, but gathering such datasets is challenging [38], since same-day satellite overpasses depend on orbits of the different sensors.

As one of the most widely used sources of remote sensing imagery of the last decade, Sentinel-2 has naturally been matched with a selection of HR imagery from PlanetScope [39], PeruSat [40] and Worldview-3 [41], though none of those datasets have been released to the scientific community. More recently, several open datasets have been published to tackle either Multi-Image Super-Resolution (MISR) or SISR of Sentinel-2, and are listed in table I. Recently, Aybar et al. recently proposed OpenSR [42], which gathers data from Sen2Venμs, NAIP and Worldstrat into a dataset dedicated to SISR evaluation for remote sensing.

WorldStrat [43] , MuS2 [44], and BreizhSR [45] match several Sentinel-2 images with commercial VHR and do not guarantee that a same-day acquisition exists. Sen2NAIP [46] also may have temporal differences of up to 30 days. Those datasets are therefore prone to containing landscape changes. They also use different classes of sensors for LR and HR images, with a lot of critical differences: missing spectral bands, different levels of radiometric processing (Top of Canopy reflectances versus uncalibrated digital counts), geometric processing (ortho-rectification, pan-sharpening), and image encoding (8-bits aerial imagery in Sen2NAIP). For instance, achieving higher up-sampling factors with Worldstrat or BreizhSR requires to use pan-sharpening [47] in order to merge the 6 meter multispectral image with the 1.5 meter panchromatic image in order to produce a 1.5 meter multispectral image. Among those datasets, Worldstrat stands out for its fair sampling strategy which ensures high variability of patches and makes sure that their distribution is representative of end-users interests.

On the other hand, Sen2Venμs [48] is the cross-sensor dataset publicly available which better takes into account the above-mentioned limitations. Its super-resolution factor of 2 for the Sentinel-2 10 m bands and 4 for the 20 m bands may seem modest, but it is the only dataset that offers a same-day, 15-minute apart guarantee on the LR and HR pairs. Both the LR sensor (Sentinel-2) and the HR sensor (Venμs [49]) have similar spectral bands and similar SSR. Both have been processed to level 2A (TOC reflectance), using the same L2A processing software [50]. It is therefore tailored for remote sensing applications, where both radiometric and geometric accuracies matter. It must be stressed that if the up-sampling factor of Sen2Venμs looks small its HR resolution is native as opposed to Worldstrat.

TABLE I
MULTI-IMAGE SUPER-RESOLUTION AND SINGLE-IMAGE SUPER-RESOLUTION ORIENTED CROSS-SENSOR DATASETS USING SENTINEL-2 AS THE LR IMAGE (P IS FOR PANCHROMATIC BAND, NIR IS FOR NEAR INFRA RED BAND, P+XS STANDS FOR PANSHARPENED IMAGES).

Name	$1e^3$ km²	HR Sensor	HR bands	HR res.
Worldstrat [43]	9,8	SPOT6/7	P+RGB+NIR	6m (1.5 m)
MuS2 [44]	3,2	WV2	RGB+NIR	1.5 m (0.4 m)
BreizhSR [45]	35	SPOT6/7	P+RGB+NIR	6m (1.5 m)
Sen2Venμs [48]	216,7	Venμs	8 S2 bands	5 m
Sen2NAIP [46]	2, 3	NAIP	RGB+NIR	0.6 m

C. Geometric and radiometric discrepancies

Two kind of discrepancies arise in cross-sensor datasets: geometric discrepancies and radiometric discrepancies. The primary source of geometric discrepancies is the difference in viewing angles between the HR target image and the LR input image: different viewing angles will cause parallax effects resulting in local shifts that may be up to several pixels in areas with significant elevation variations. A secondary issue is the difference in accuracy of ground projection tools that are used to obtain map projected HR and LR images for superimposition. In MuS2 [44], the authors manually filtered out data with geometric discrepancies, while key-points based registration using SIFT [48] or SuperGlue points [42] have been used in Sen2Venμs, Worldstrat and OpenSR to improve the geometric consistency. Despite this pre-processing, Fig. 1 exhibits local shifts of several pixels in Sen2Venμs, probably caused by the large viewing angles of Venμs, and even larger and more systematic shifts in Worldstrat, the latter being likely caused by differences in location accuracy in addition to differences in viewing angles.

On the radiometric consistency side, in addition to slight differences in their SSR for corresponding HR and LR bands, the largest source of radiometric discrepancies is Bidirectional Reflectance Distribution Function (BRDF) effects, which cause variation of observed surface reflectances due to differences in observation angles. For non-synchronous datasets such as Worldstrat, another source of radiometric discrepancies is the difference in atmospheric conditions, that will incur changes in top-of-atmosphere surface reflectances and differences in the estimation of back-of-atmosphere surface reflectances.

Some datasets contain HR images processed by means of histogram matching [42], while in Sen2Venμs a simple linear transformation [48] is used in order to get closer to the LR radiometry. Fig. 2 shows scatter plots of sample patches from Worldstrat and Sen2venμs, for each spectral band. First, it can be observed that Worldstrat is significantly less radiometrically consistent that Sen2venμs. This is expected as Worldstrat does not offer same-day acquisitions as in Sen2Venμs, and it does not even provide cloud masks for Spot6/7 images, as the Level 2A of Venμs does. Regarding Sen2Venμs, it can be seen that despite of the radiometric pre-processing, slight radiometric discrepancies remain.

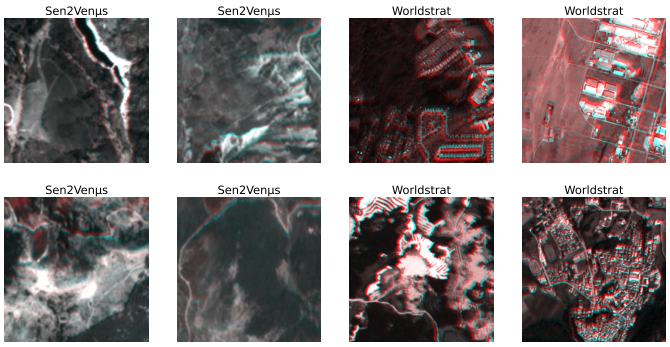

Fig. 1. Color compositions highlighting geometric discrepencies of images from the Sen2Venμs and Worldstrat datasets (R: B3 from Sentinel-2, B, G: B3 from Venμs or Spot6/7). Geometric discrepancies appear in red or blue fringes. The overall redish appearance of Worldstrat patches is caused by the radiometric bias in this dataset. Note that the overall redness of the Worldstrat patches in Fig. 1 is caused by radiometric bias.

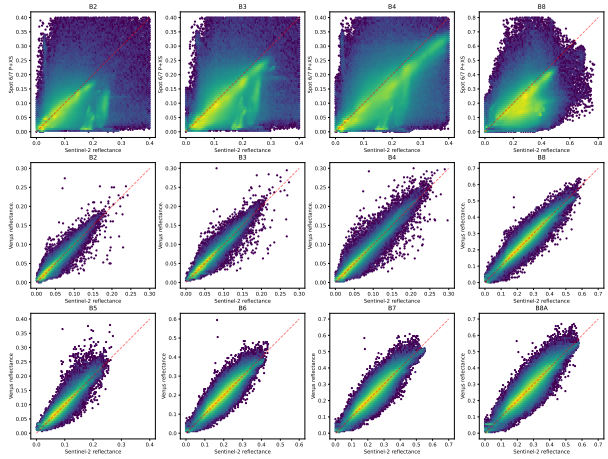

Fig. 2. Scatter plot of pan-sharpened Spot6/7 vs. Sentinel-2 reflectance from sample patches from the Worldstrat dataset (first row) and of Venμs vs. Sentinel-2 reflectance for sample patches from the Sen2Venμs dataset (second and third rows)

D. Contributions

Although discrepancies of cross-sensor datasets may impair the training of SISR Deep Learning models and the proper evaluation of their performances, their impact has been overlooked in recent remote sensing SISR cross-sensor works. In [51], the authors observe that Peak Signal to Noise Ratio (PSNR) is actually higher for SISR output with more blur. In [44], the authors find that bicubic interpolation consistently prevails over considered SISR techniques in some regions, when evaluating them with PSNR. In [52], the authors show that PSNR prefers blurred images over sharp but distorted images, and propose to turn to perceptual IQ metrics. In [53], the authors note that PSNR penalizes a bias in intensity much more than noise, and propose cPSNR, which consists in applying a set of pixel-wise shifts to one of the images and use the maximum PSNR, in order to make PSNR more robust to spatial mis-registration. We hypothesize that those observations are in fact related to the geometric and radiometric discrepancies of cross-sensor datasets.

In this paper, we intend to investigate the impact of those discrepancies, and propose a new strategy to train and assess the performances of SISR models in a cross-sensor setting. We intend to answer those questions by studying two major cross-sensor datasets: the Sen2Venμs dataset, with limited ×2 (×4 for the 20 m bands) up-sampling factor to 5 meter resolution and mild level of discrepancies, and the Worldstrat dataset with high level of discrepancies and a more challenging ×4 up-sampling factor to 2.5 m resolution. There are three main contributions in this work:

1) We analyze the limitations of Image Quality (IQ) metrics commonly used in remote sensing SISR with respect to radiometric and geometric discrepancies. This allows us to identify which metrics and losses should be used in a cross-sensor setting.
2) We propose a new set of Frequency Domain Analysis (FDA) metrics tailored for assessing image restoration in terms of spatial frequencies, complementing the identified set of metrics.
3) We propose a new strategy with respect to cross-sensor geometric and radiometric distortions for the training and evaluation of cross-sensor SISR models. We demonstrate that this strategy efficiently mitigates the impact of cross-sensor distortions and enables a fair evaluation of the performances.

The remainder of this paper is organized as follows. Section II focuses on the analysis of existing IQ metrics, describes the proposed FDA metrics and derives a set of robust metrics that allow to analyze every aspect of cross-sensor SISR. Section III proposes a strategy to mitigate cross-sensor discrepancies in SISR training, and precisely measures the benefit of each of its components in terms of radiometric and geometric faithfulness, as well as spatial frequency restoration and general Image Quality (IQ). Finally, section IV discusses the limitations of the proposed method, the impact of our findings, and potential extensions.

E. Notations

Throughout this paper, HR image patches will have width and height noted as $H \times W$, whereas LR patches will have width and height $H \times W$. $R_b(i,j)$ designates a $h \times w$ matrix of pixels (2D tensor in DL parlance) for band b of target HR patch, $P_b(i,j)$ designates a $W \times H$ tensor for band b of a predicted SISR patch, and $X_b(i,j)$ designates a $w \times h$ tensor

for band b of input LR patch, where $H = s \cdot h$ and $W = s \cdot w$ and $s \in \mathbb{N}^{+\star}$ is the integer scale factor. Indices (i, j) and subscript b may be dropped when unnecessary, for the sake of the reader.

II. Assessing cross-sensor SISR performances

This section first reviews common IQ metrics in remote sensing SISR. A benchmark of those metrics is then proposed, in order to evaluate how they react to geometric and radiometric discrepancies, but also to noise level or chessboard patterns. This benchmark allows to identify which metrics should be used to evaluate SISR models in a cross-sensor context. A new set of FDA based metrics is then introduced to complement this analysis. Last, the set of identified metrics is used in order to assess the SISR potential of the Sen2Venμs and Worldstrat datasets.

A. A short review of common IQ metrics used in SISR

IQ metrics can be divided into:

- Full Reference (FR) metrics, that require a reference HR image in the SISR context, further divided into:
  - traditional pixel-based metrics, which are derived from local statistics of the images difference,
  - perceptual metrics, which assess the discrepancies between predicted and reference images by projecting them both into a feature space, usually derived from pre-trained neural networks, and
- No Reference (NR) metrics, that attempt to evaluate intrisic IQ without any reference image.

Some metrics may also be used as objective functions (called loss functions in DL parlance) for model training. To qualify as loss functions, metrics require to be differentiable and numerically stable.

1) Full Reference local metrics: Surely the gold standard for pixel-based metrics in computer vision and SISR in particular has been the Peak Signal to Noise Ratio (PSNR), which has been used in almost all SISR published works. Starting from the Mean Squared Error (MSE) given by:

$$\text{MSE}(P_b, R_b) = \frac{1}{WH} \sum_{i=0}^{W-1} \sum_{j=0}^{H-1} (P_b(i,j) - R_b(i,j))^2, \quad (1)$$

and PSNR is given by:

$$\text{PSNR}(P_b, R_b) = 10 \cdot \log_{10} \Big(\frac{d^2}{\text{MSE}(P_b, R_b)} \Big), \quad (2)$$

where d denotes the expected data range.

Another widely used metric found in SISR papers is the Structural Similarity Index Measure (SSIM) [54], which is given by:

$$\text{SSIM}(P_b, R_b) = \frac{(2\mu_{P_b}\mu_{R_b} + c_1)(2\sigma_{P_bR_b} + c_2)}{(\mu_{P_b}^2 + \mu_{R_b}^2 + c_1)(\sigma_{P_b}^2 + \sigma_{R_b}^2 + c_2)}, \quad (3)$$

where μ_x is the mean of image x, σ_x is the standard deviation of image x, σ_{xy} is the covariance between images x and y, and c_1 and c_2 are user defined constants allowing to enhance the stability when images means or standard deviations are close to zero. SSIM is computed on small local windows, and averaged over the full image extent. As such, SSIM remains a local metric despite its formulation based on image statistics.

Although widely adopted, PSNR and SSIM are known to correlate poorly with visually assessed IQ [55]. Other pixel-based metrics that are less widely adopted include Image Fidelity Criterion [56] and the Gradient Magnitude Similarity Deviation [57]. Though advertised by the authors as perceptual IQ metrics, they still rely on local computations.

2) Full Reference Perceptual metrics: Obvious limitations of PSNR and SSIM have pushed researchers towards perceptual metrics, a class of metrics that tries to better match the human perception. Perceptual metrics consist in using a neural network that has been trained for a computer vision task, and compute the L_2 norm between the deep embeddings of reference and predicted patches, out of an intermediate layer of the network. The network itself can be a pre-trained VGG model [58] for instance. The most widely known perceptual metric is the Learned Perceptual Image Patch Similarity (LPIPS) [52]. LPIPS is used in SISR review papers to compare methods [3], [6]. Another lesser known perceptual metric is PieAPP [59], which directly predicts similarity scores based on a network trained using pair-wise human preference annotations. Because they are based on neural networks, perceptual metrics are fully differentiable and can thus also be used as a loss function [60], [61].

3) No Reference metrics: In order to be able to assess IQ in contexts where no reference image is available, many NR IQ metrics have been proposed in the literature. Mitall et al. proposed BRISQUE [62], which builds local features mapped to IQ score by a Support Vector Regressor. It is trained on human annotated scores, yielding a score between 0 and 100, 0 being the best IQ. NIQE [63] elaborates on the idea of collecting quality aware features from the image but differs from BRISQUE in that those features are fit with a multivariate Gaussian, which is then compared to a multivariate Gaussian fitted on a corpus of natural images. PIQE [64] differs from the previous approaches in that no human annotated supervision is required, the final score being derived from local blocks by a set a of expert rules. More recently, Ma et al. [65] used a human supervised study in order to derive SISR quality of natural images by means of a regression model working on features from Discrete Cosine Transform (DCT), Wavelet Decomposition and patch-level Principal Components Analysis. CLIP-IQA [66] uses Contrastive Language-Image Pre-training (CLIP), which is a model that is trained on pairs of image and text. The model uses the cosine similary between non-ambiguous antonym text prompts as the basis to provide a NR IQ metric. In [51] the authors propose to use CLIPA-v2 [67] with cosine similarity (agreement of 82.5% with visual inspection). In [15], the authors used BRISQUE and PIQE for the qualitative assesment of IQ of their SISR algorithm.

B. Benchmark of metrics for cross-sensor SISR

IQ metrics suited for cross-sensor SISR should be able to correctly assess the level of blur of SISR predictions in the

presence of geometric and radiometric discrepancies in the dataset. In particular, different SISR models yielding increasing levels of blur should be correctly ranked by the metric. In order to assess how cross-sensor dataset discrepancies affect the blur ranking capabilities of standard SISR metrics, the following experiments have been conducted.

Starting from a set of sample HR patches R_b for band b, geometric discrepancies are simulated by a simple translation in the diagonal direction, with a bicubic resampling, as follows:

$$R_t^{geom} = \omega\Big(R, (t/\sqrt{2}, t/\sqrt{2})\Big), \tag{4}$$

where ω is the bicubic resampling operator, and t is the magnitude of the diagonal translation.

Radiometric discrepancies are simulated by a simple linear transform of the radiometry of slope a around a fixed point c, as given in equation 5:

$$R_{a,c}^{rad} = b + a(R - c). \tag{5}$$

After radiometric or geometric distortion simulation, blur is simulated by convolving the data with a Gaussian kernel of width σ_0, where σ_0 is computed as the standard deviation of a Gaussian kernel having a *mtf* value m at Nyquist rate given by:

$$f_{mtf\to\sigma}(m) = \frac{1}{\pi}\sqrt{-2\ln(m)}. \tag{6}$$

The Gaussian kernel is given by:

$$\phi_\sigma(i,j) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{i^2+j^2}{2\sigma^2}}, \tag{7}$$

and its convolution with the input image, as well as optional addition of noise with standard deviation σ_{noise}, is formulated as:

$$R_{\sigma_0,\sigma_{noise}}^{smooth} = R * \phi_{\sigma_0} + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma_{noise}). \tag{8}$$

Note that the noise term ϵ can also be replaced by a 2D periodic pattern (also called chessboard pattern), of variable intensity, which will be used to assess the metrics robustness to chessboard artifacts that can be yielded by sub-pixel convolution layers.

This simulation process is applied to 128 HR patches from the Sen2Venμs dataset using band B4. The choice of this particular spectral band is arbitrary and does not impact the analysis. The simulation order is as follows: spatial distortion, radiometric distortion, blur, noise, and chessboard pattern. One parameter among those is varied at a time, for a fixed range of *mtf* values. Radiometric distortion uses $c = 0.1$ and $a \in [0, 0.1]$ for the slope values. Geometric distortion uses $t \in [0, 2]$ pixels. Noise level is given by $\sigma_{noise} \in [0, 0.01]$. For the pattern experiment, a fixed 4×4 random pattern has been generated. Each investigated metric is used to measure the similarity between the input images with no degradation and the simulated images with increasing levels of blur and studied distortion. Only a selection of the most interesting results are presented in the following sections, and more results can be found in the supplementary materials. Metric implementations from the PIQ library [68] have been used.

A perfect IQ metric for cross-sensor SISR would of course rank levels of blur correctly regardless of the level of distortion. If it were to be used as a loss function, this IQ metric should also not respond to distortion at all, regardless of the blur level: if less distortion yields better loss values, chances are that the model will learn the distortion along with the image sharpening.

1) Local metrics: Fig. 3 presents the result of the benchmark for PSNR. Since Sen2Venμs uses L2A surface reflectances, data range for PSNR is set as $d = 1$. Each chart in the Fig. shows how the metric reacts to one of the four investigated distortions, for different levels of blur ranging from None (no blur applied) or $m = 0.4$ (very sharp image) to $m = 0.001$ (very blurry image).

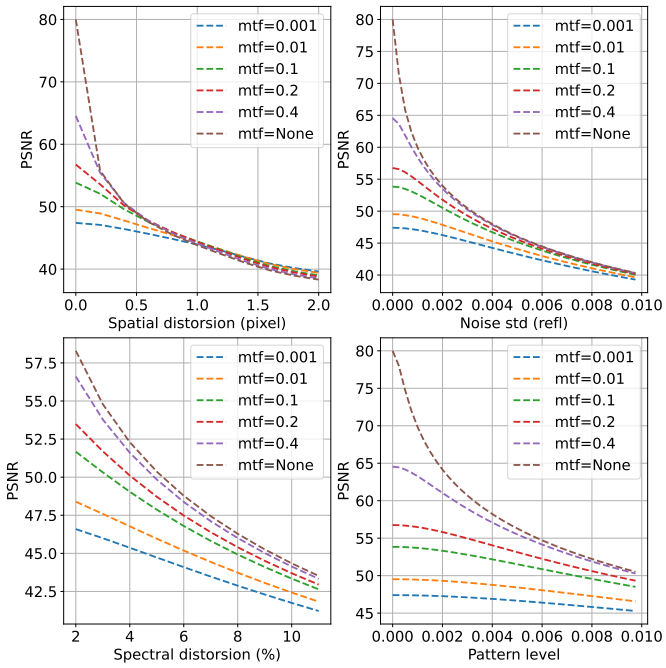

Fig. 3. Benchmarking of PSNR (higher is better) with respect to geometric distortion, radiometric distortion, noise level and chessboard pattern level, for different levels of blur, ranging from *mtf*=None (no blur applied) or *mtf*=0.4 (very sharp images) to *mtf*=0.001 (very blurry images).

Since PSNR is a metric that should be maximized (higher values are better), it would be expected that increasing levels of blur yield decreasing PSNR values. However, it can be observed that if geometric distortion is higher than 1 pixel, PNSR does not distinguish sharper images from smoother ones, and will even prefer the latter. For a given level of sharpness, PSNR favors less spatial distortion: a network learning some spatial distortion during training might exhibit higher PSNR, regardless of whether it actually has better super-resolution performances. PSNR is also sensitive to spectral distortion, and gradually looses ability to distinguish different levels of smoothness as spectral distortion gets higher. PSNR might favor networks that have learned the spectral distortion better over networks that have sharper, undistorted predictions. This experiment confirms observations on PSNR behavior with respect to visual inspection made in [44], [51], [52]: in a cross-sensor setup, PSNR will simply not measure SISR

performances and will favor smoothness, as well as learned radiometric and geometric distortions.

Despite being advertised as fixing the defects of PSNR and other MSE based-metrics [55], SSIM suffers from the same sensitivity to spatial and spectral distortion, as shown in Fig. 4. In fact, all pixel-based metrics that have been analyzed in these experiments show the same trends.

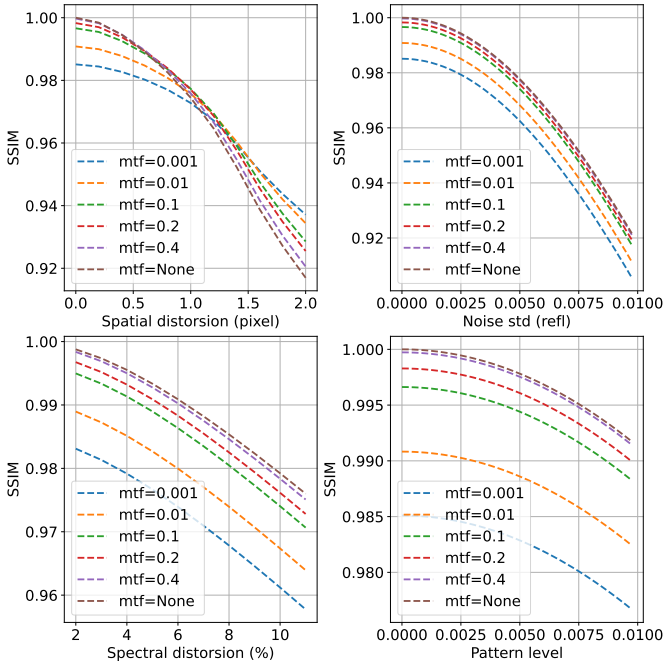

Fig. 4. Benchmarking of SSIM (higher is better) with respect to geometric distortion, radiometric distortion, noise level and chessboard pattern level, for different level of blur, ranging from *mtf*=None (no blur applied) or *mtf*=0.4 (very sharp images) to *mtf*=0.001 (very blurry images)

According to these findings, none of those metrics should be used to evaluate cross-sensor SISR, or as loss terms to train cross-sensor SISR models, because of their inability to measure or promote sharpness in presence of distortions. This of course does not hold for simulated datasets where such discrepancies do not exist.

2) Perceptual metrics: Defects of local metrics, though not characterized as in the present work, have been empirically observed by researchers and have justified the introduction of the perceptual metrics and losses, as presented in section II-A2. Since most of them are trained for RGB images, in this work a single band is used and duplicated to form a gray-scale RGB image, which is clipped to the range $[0, 1]$. Fig. 5 shows that the LPIPS metric indeed behaves consistently in the presence of geometric and radiometric distortion and consistently favors sharpness over blur. The same gentle slope toward less distortion can be observed regardless of the level of blur. This slope exists, which means that when used as a loss term, LPIPS will push the network toward learning to reduce those distortions. If LPIPS also behaves consistently with respect to the level of noise, it seems to favor a small amount of chessboard pattern, especially for the more blurry images. Though not shown in these experiments, LPIPS is also well conditioned to be used as a loss term, since it is

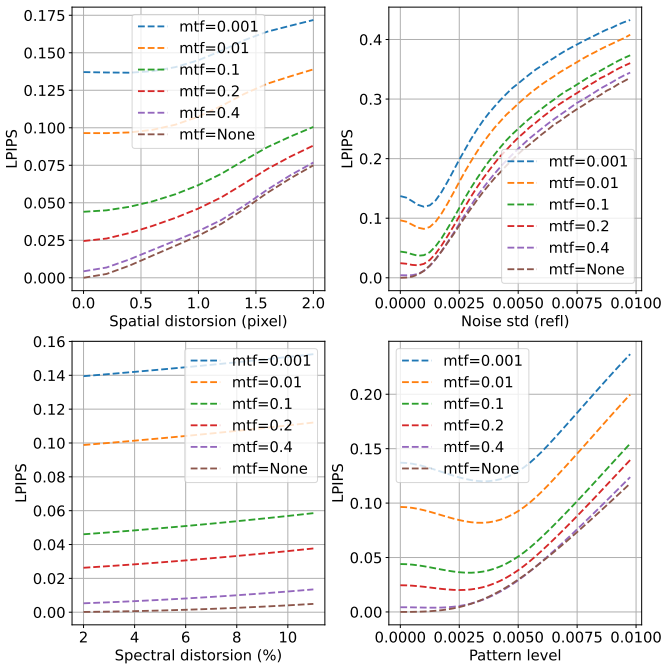

Fig. 5. Benchmarking of LPIPS (lower is better) with respect to geometric distortion, radiometric distortion, noise level and chessboard pattern level, for different levels of blur, ranging from *mtf*=None (no blur applied) or *mtf*=0.4 (very sharp images) to *mtf*=0.001 (very blurry images)

ultimately a L_2 loss applied to features extracted by a CNN, and is therefore fully differentiable.

3) No Reference Image Quality metrics: Fig. 6 shows how the BRISQUE score responds to the experiments. As a NR IQ metric, BRISQUE is blind to the radiometric and geometric distortions, the slight variations in response to geometric distortion being caused by the bicubic resampling. An interesting outcome of this experiment is the fact that BRISQUE favors a certain amount of noise or chessboard pattern. Being trained on human perception annotations, this surely reflects the fact that noise makes images look less synthetic. Nevertheless, in a SISR context, this trend might be problematic, especially when using BRISQUE for selecting the best model during training for instance.

C. Frequency Domain Analysis and derived metrics

As established by the experiments presented in the previous section, local metrics should be avoided in the context of cross-sensor SISR, while perceptual and NR IQ metrics are better choices for this task. However, they still have important limitations: perceptual metrics are not completely insensitive to discrepancies, while NR IQ metrics favor properties such as noise that may be related to the dataset they have been trained on.

This section proposes to leverage the conventional Discrete Fourier Transform (DFT), noted $\mathcal{F}$, in order to derive indicators focused on spatial frequency restoration. Analysing IQ through image decomposition has been proposed in [69] and [65]. DFT in particular has been largely used to derive Image Quality metrics [70]–[72]. More recently, FDA has been used [73] in order to study the effects of the simulation process

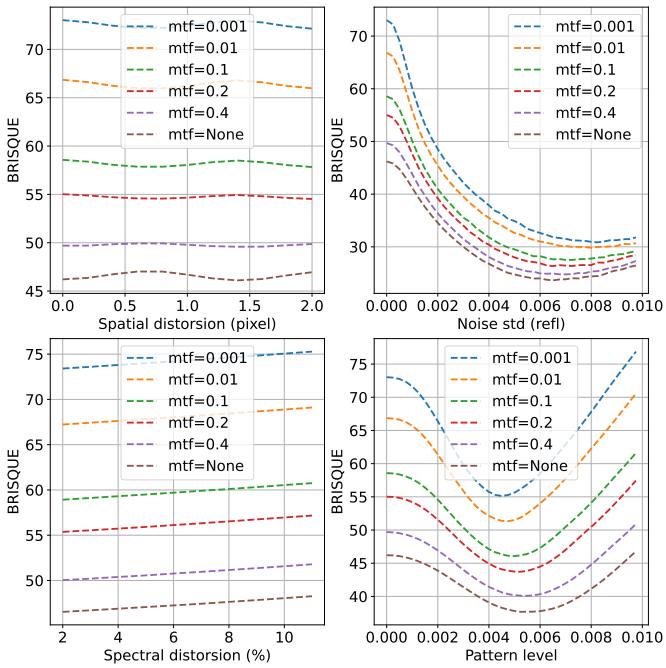

Fig. 6. Benchmarking of BRISQUE (lower is better) with respect to geometric distortion, radiometric distortion, noise level and chessboard pattern level, for different levels of blur, ranging from *mtf*=None (no blur applied) or *mtf*=0.4 (very sharp images) to *mtf*=0.001 (very blurry images)

in a remote sensing SISR context using simulated datasets. The proposed approach draws inspiration from [74], [75] but adapts it to the problem of SISR. It consists in analysing the Frequency Attenuation Profile ($\mathcal{F}_{\mathcal{AP}}$) for bandwidth $[f_m, f_M]$. Let

$$U_{f_m,f_M} = \left\{(u,v) : f_m \leq \sqrt{u^2+v^2} < f_M\right\} \tag{9}$$

denote the set of discrete spatial frequencies (u, v) that lies within a ring defined by f_m and f_M in Fourier plane, $\mathcal{F}_{\mathcal{AP}}$ is given by:

$$\mathcal{F}_{\mathcal{AP}}[P](f_m,f_M) = \frac{1}{\#U_{f_m,f_M}} \sum_{(u,v)\in U_{f_m,f_M}} |\mathcal{F}[P](u,v)|, \tag{10}$$

where $\#U_{f_m,f_M}$ is the number of elements in U_{f_m,f_M}, and $\mathcal{F}[P](u, v)$ is the DFT of image P.

$\mathcal{F}_{\mathcal{AP}}[P](f_m, f_M)$ is successively computed over a set of non-overlapping bandwidth intervals as given by the DFT quantization. For the sake of simplicity, the resulting set of values is denoted $\mathcal{F}_{\mathcal{AP}}[P](f_n)$, $n \in [0, N]$ in the following, with f_n the central frequency of each frequency intervals. It is then averaged across batches and dataset. Finally, the normalized logarithmic $\mathcal{F}_{\mathcal{AP}}$ is computed, which gives spatial frequency attenuation in dB:

$$\begin{aligned}\mathcal{F}_{\mathcal{AP}}{}^{\star}[P](f_n) = 10 \cdot \Big(&\log_{10}\left(\mathcal{F}_{\mathcal{AP}}[P](f_n)\right) \\ &- \log_{10}\left(\mathcal{F}_{\mathcal{AP}}[P](f_0)\right)\Big),\end{aligned} \tag{11}$$

Fig. 7 shows such average $\mathcal{F}_{\mathcal{AP}}{}^{\star}$ for original Venμs patches, bicubic up-sampled Sentinel-2 patches and blur levels in-between, using band B4 and the same 128 patches as used in section II-B. The increasing damping of higher frequencies can be clearly observed. The area highlighted in green, between the bicubic up-sampled Sentinel-2 curve and the Venμs curve, represents the maximum frequency restoration that can be expected from a SISR algorithm trained on this dataset.

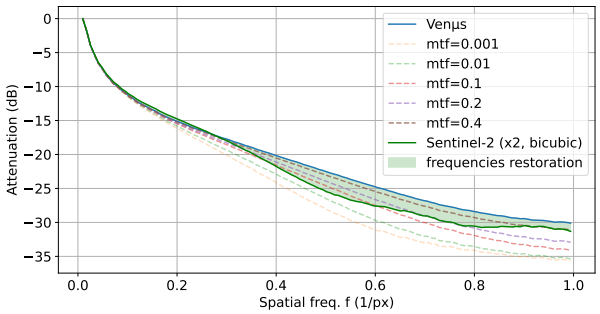

Fig. 7. Normalized log-Frequency Attenuation Profile $\mathcal{F}_{\mathcal{AP}}{}^{\star}$ for reference Venμs patches, Sentinel-2 2x bicubic up-sampled, and different levels of blur, for band B2. Spatial frequencies in x axis correspond to normalized spatial frequencies $f = \frac{f_n}{f_N}$.

Fig. 8 gives an example of the same graph as in Fig. 7 with the addition of a single SISR algorithm. The three curves allow to define different areas. In this work we define the Potential Frequency Restoration (PFR) w.r.t. bicubic spatial interpolation as the area between the $\mathcal{F}_{\mathcal{AP}}{}^{\star}$ of the original venus and the $\mathcal{F}_{\mathcal{AP}}{}^{\star}$ of the bicubic interpolation. It is estimated as :

$$\text{PFR}(R_b, X_b) = \sum \max\big((\mathcal{F}_{\mathcal{AP}}{}^{\star}[R_b] - \mathcal{F}_{\mathcal{AP}}{}^{\star}[X_b]), 0\big). \tag{12}$$

Likewise, we define the Actual Frequency Restoration (AFR), which measures the amount of spatial frequency content that has been restored with respect to bicubic up-sampling, as:

$$\begin{aligned}\text{AFR}(P_b, R_b, X_b) = \sum \max\Big(&\min\big(\mathcal{F}_{\mathcal{AP}}{}^{\star}[P_b], \mathcal{F}_{\mathcal{AP}}{}^{\star}[R_b]\big), \\ &\min\big(\mathcal{F}_{\mathcal{AP}}{}^{\star}[X_b], \mathcal{F}_{\mathcal{AP}}{}^{\star}[R_b]\big)\Big) \\ &-\min\big(\mathcal{F}_{\mathcal{AP}}{}^{\star}[R_b], \mathcal{F}_{\mathcal{AP}}{}^{\star}[X_b]\big).\end{aligned} \tag{13}$$

From PFR and AFR, we derive the Frequency Restoration Rate (FRR) as follows:

$$\text{FRR}(P_b, R_b, X_b) = \frac{\text{AFR}(P_b, R_b, X_b)}{\text{PFR}(R_b, X_b)}. \tag{14}$$

FRR measures how much of the PFR have actually been restored, as given by AFR. It ranges from 0 when no frequency have been restored to 1 when the full PFR has been restored by the algorithm.

There might be ranges of frequencies for which the SISR $\mathcal{F}_{\mathcal{AP}}{}^{\star}$ is actually higher than the reference $\mathcal{F}_{\mathcal{AP}}{}^{\star}$, meaning that the restoration of those frequencies is too strong. In order to measure this, we define the Frequency Restoration Overshoot (FRO) as the ratio between the FRO area and the

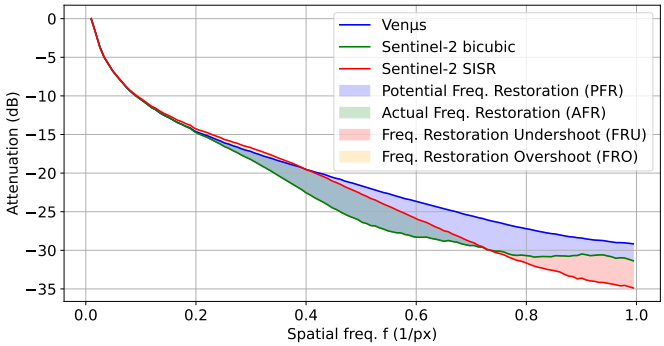

Fig. 8. Illustration of Potential Frequency Restoration (PFR), Actual Frequency Restoration (AFR), Frequency Restoration Overshoot (FRO) and Undershoot(FRU) based on logarithmic normalized Frequency Attenuation Profiles ($\mathcal{F}_{\mathcal{AP}}{}^{\star}$) from Venμs patches, bicubic up-sampled Sentinel-2 and a SISR prediction. Spatial frequencies in x axis correspond to normalized spatial frequencies $f = \frac{f_n}{f_N}$.

total area under the reference $\mathcal{F}_{\mathcal{AP}}{}^{\star}$, as illustrated in Fig. 8, and given by:

$$\begin{aligned}\text{FRO}(P_b, R_b, X_b) = &\frac{1}{\sum \mathcal{F}_{\mathcal{AP}}{}^{\star}[R_b]}\Big(\sum \mathcal{F}_{\mathcal{AP}}{}^{\star}[R_b] \\ &- \max(\mathcal{F}_{\mathcal{AP}}{}^{\star}[P_b], \mathcal{F}_{\mathcal{AP}}{}^{\star}[R_b])\Big).\end{aligned} \tag{15}$$

Similarly, there might be ranges of frequencies for which the SISR $\mathcal{F}_{\mathcal{AP}}{}^{\star}$ is actually lower than the bicubic upsampled LR $\mathcal{F}_{\mathcal{AP}}{}^{\star}$, meaning that the restoration of those frequencies is worse than a bicubic interpolation. In order to measure this, we define the Frequency Restoration Undershoot (FRU), as the ratio between the FRU area and total are under the bicubic up-sampled FRA, as illustrated in Fig. 8, and given by:

$$\begin{aligned}\text{FRU}(P_b, R_b, X_b) = &\frac{1}{\sum \mathcal{F}_{\mathcal{AP}}{}^{\star}[X_b]}\sum\Big(\mathcal{F}_{\mathcal{AP}}{}^{\star}[X_b] \\ &- \min(\mathcal{F}_{\mathcal{AP}}{}^{\star}[P_b], \mathcal{F}_{\mathcal{AP}}[X_b])\Big).\end{aligned} \tag{16}$$

According to these definitions, the SISR algorithm in Fig. 8 has a FRR of 43.6%, a FRO of 0.5% and a FRU of -2.49%. Note that by design, bicubic up-sampling always has 0% for AFR, FRR, FRU and FRO. Fig. 9 shows the results of applying the benchmark protocol of section II-B to FRR. The proposed metric exhibits a great sensitivity to blur and is almost insensitive to spectral distortion. It is also insensitive to spatial distortions, the oscillations of the sharper *mtf* being caused by the bicubic interpolation during the simulation process. As expected, noise increases the spatial frequency content, which can be mistakenly interpreted as frequency restoration. However, ordering of blur levels remain consistent even at higher noise levels. Last, it should be noted that the proposed FDA metrics make no assumption with respect to spatial resolution or even sensor modality, and can therefore be applied to the full range of SISR scenarios.

D. Robust metrics for cross-sensor SISR

The analysis presented in this section shows that PSNR and other local metrics should absolutely not be used for

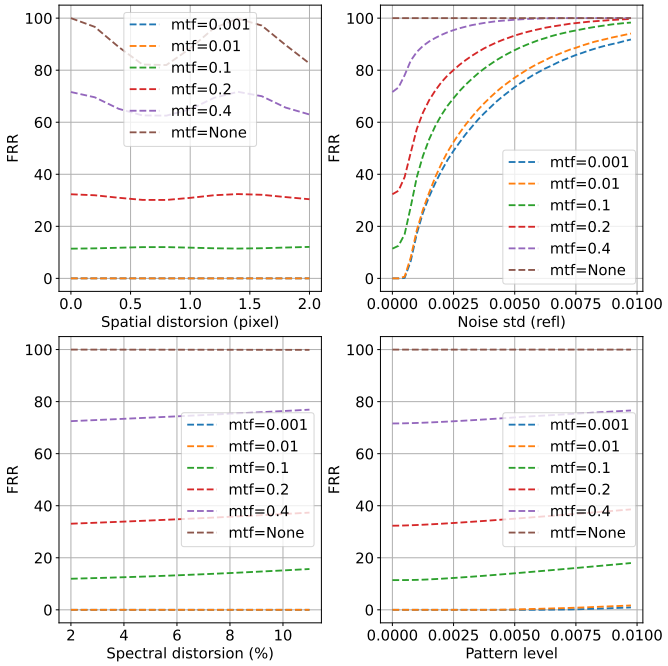

Fig. 9. Benchmarking of the proposed FRR metric (higher is better) with respect to geometric distortion, radiometric distortion, noise level and chessboard pattern level, for different level of blur, ranging from *mtf*=None (no blur applied) or *mtf*=0.4 (very sharp images) to *mtf*=0.001 (very blurry images)

the evaluation of cross-sensor SISR performances. It allows to define a set of robust metrics for this task:

- The LPIPS perceptual metric allows to measure proximity with target images, while being relatively insensitive to radiometric and geometric discrepancies, even if LPIPS has a tendency to favor chessboard patterns,
- The BRISQUE metric can be used to assess the general image quality and is insensitive to radiometric and geometric discrepancies. It should be noted that BRISQUE favors a bit of noise or chessboard patterns,
- The proposed AFR, FRR, FRO and FRU allows for fine grain characterization of the achieved spatial frequency restoration.

Finally, this set of metrics should be complemented with metrics that measure the spectral and spatial distortions induced by the SISR network. Measuring spectral distortion is obvious, and has been used for years in the pan-sharpening community [47]. The SISR output is Low Pass Filtered (LPF) and down-sampled back to the input patch size, after what traditional local metrics such as RMSE can be used, as given by:

$$\text{RMSE}_{\text{LR}}(P, X) = \sqrt{\frac{1}{WH}\sum_{w,h}\Big(X - \big(P * \phi_{\sigma_0}\big)\downarrow_s\Big)^2}, \tag{17}$$

where $\downarrow_s$ denotes the decimation operator by a stride of s and ϕ_{σ_0} is a Gaussian kernel as introduced in equation 7.

In order to measure geometric discrepancies, we propose to leverage an auxiliary pretrained UNet that estimates the optical flow between input LR patches X and predicted SISR

patches P, noted as $F_{P_b^\star \to X_b}$. Pretraining of this model is not specific to this work, and is detailed in appendix A. The proposed Geometric Distortion (GD) metric is the L_2 norm of the estimated optical flow:

$$\mathrm{GD}(P_b, X_b) = \sqrt{|F_{P_b^\star \to X_b}|^2}. \tag{18}$$

III. Robust strategy for cross-sensor SISR

In this section, we propose a strategy for training and evaluating models that is robust to radiometric and geometric distortions. This strategy is then applied to the 6 different training datasets as summarized in table II. An ablation study is performed in order to demonstrate the benefits of the proposed strategy in terms of learned radiometric and geometric discrepancies.

It should be noted that training sets **s2v1** and **s2v2** have both been sampled from the **Sen2Venµs** dataset. Training set **s2v1** has 20 sites for a total of 35 954 training patches, with a majority of sites with high viewing angles and low PFR, whereas training set **s2v2** has only 8 sites among those with PFR higher than 6%, for a total of 14 615 patches. This allows to investigate the relative importance of having a diverse dataset with respect to selecting good quality samples. The complete details on those datasets, as well as an analysis of their potential for the SISR task conducted with metrics identified in II-D can be found in appendix B.

TABLE II
Summary of the conducted experiments with training set, Sentinel-2 bands and associated up-sampling factor.

Training set	Source Dataset	Bands	Up-sampling
s2v1×2	Sen2Venµs	B2, B3, B4, B8	10 m → 5 m
s2v2×2	Sen2Venµs	B2, B3, B4, B8	10 m → 5 m
s2v1×4	Sen2Venµs	B5, B6, B7, B8A	20 m → 5 m
s2v2×4	Sen2Venµs	B5, B6, B7, B8A	20 m → 5 m
ws×2	Worldstrat	B2, B3, B4, B8	10 m → 5 m
ws×4	Worldstrat	B2, B3, B4, B8	10 m → 2.5 m

A. Proposed strategy

The proposed robust strategy for cross-sensor SISR is divided into two parts.

a) **During training**, increase dataset consistency by generating distortion-free HR patches, so that the model avoids learning discrepancies. This process is described in Fig. 10, and involves three main steps:

1) estimate $F_{X_b^\star \to X_b}$, the optical flow between the LR patch and the HR patch,
2) use the estimated flow in order to resample the HR patch onto the LR patch, thus reducing geometric distortion to the extent of what has been captured by the optical flow estimation,
3) inject low frequency radiometric residuals into the geometrically corrected HR patch in order to correct for radiometric discrepancies.

b) **During evaluation**, use the raw dataset, without applying the corrections used during training. In order to cope with radiometric and geometric distortions, use the set of metrics derived in section II, which have been selected for their robustness.

In the remaining of this section, the distortion correction process used in a) is further detailed. Step 1 is covered by the pre-trained optical flow estimation auxiliary UNet, as covered in appendix A. Steps 2 and 3 will be detailed in the following sections.

1) Geometric correction (step 2): The $F_{X_b^\star \to X_b}$ optical flow estimated by the pretrained UNet as described in section A-C is first low-pass filtered with a Gaussian kernel of width σ_1 in order to smooth out irregularities that might alter the quality of images that will be resampled using the flow. Value of σ_1 is empirically set to a large width of $\sigma_1 = f_{mtf \to \sigma}(1e^{-6})$ by means of eq. 6. The field is then up-sampled to R_b initial resolution using scaling factor s. The whole operation is summarized by:

$$F_{R_b^\star \to X_b} = s \times \left(F_{X_b^\star \to X_b} * \phi_{\sigma_1}\right) \uparrow_s, \tag{19}$$

where $\uparrow_s$ denotes a bicubic upsampling of factor s.

Then reference HR patches R can then be corrected from estimated optical flow (and thus geometric distortion), by means of:

$$\widetilde{R} = \omega(R, F_{R_b^\star \to X_b}). \tag{20}$$

This process forms $\widetilde{R}$, the geometrically corrected HR patch.

2) Radiometric correction (step 3): In order to correct for radiometric distortion, we propose to employ a strategy similar to the residual correction which is usual in the thermal sharpening literature [76]. $\widetilde{R}$ is down-sampled back to the LR resolution by mean of equation 25. A residual is formed by its difference with the input LR patches X. This difference is further low pass filtered with Gaussian kernel of width σ_2, and finally upsampled back to HR resolution. It is then added to $\widetilde{R}$. Value of σ_2 is experimentally set to a large width of $\sigma_2 = f_{mtf \to \sigma}(1e^{-5})$ by means of eq. 6, while value for σ_0 is the same as used in section A-C ($\sigma_0 = f_{mtf \to \sigma}(0.4)$). Intuitively, this ensures that down-sampling the radiometrically corrected HR patches would yield consistent radiometries with respect to LR patches. The low pass filtering of the LR residual avoids injecting back blurry details in the corrected HR patches. This correction requires a good geometric alignment between HR and LR patches, which is ensured by the geometric correction of equation 20. This process is summarized as follows:

$$\widetilde{R}^\star = \widetilde{R} + \left(\left((X - \widetilde{R} * \phi_{\sigma_0}) \downarrow_s\right) * \phi_{\sigma_2}\right) \uparrow_s . \tag{21}$$

The analysis of the impact of the proposed training strategy on the target image quality for each dataset can be found in appendix C.

B. Experimental setup

This section presents the experimental setup used to demonstrate the benefits of the proposed robust strategy. All experiments adopt the general framework of ESRGAN [60], which consists in training the SRResNet [61] architecture with

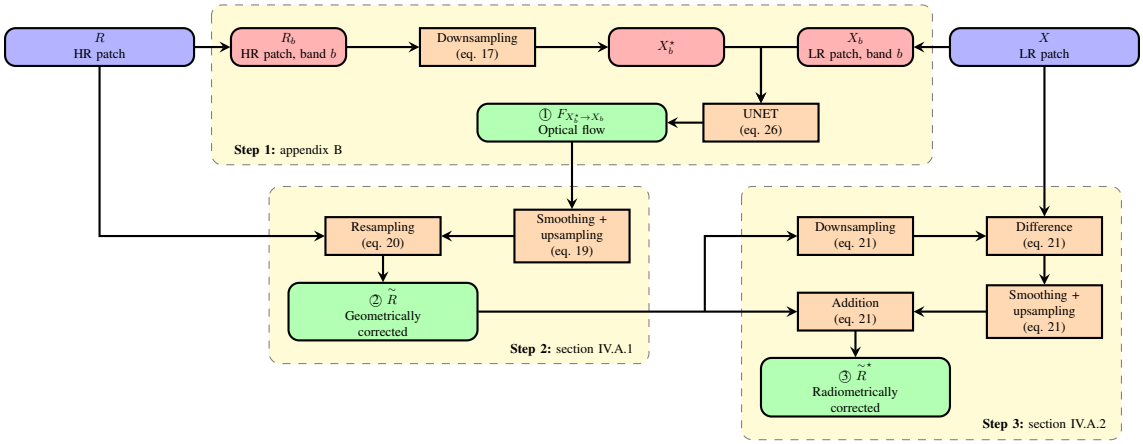

Fig. 10. Overview of robust strategy workflow. Outputs of the three main steps are highlighted in green.

Residual in Residual Dense Blocks (RRDB) in a generative adversarial training scheme, with the following modifications. We employ 6 residual blocks of 64 latent features for all experiments except for the Sen2Venμs 20 m $\rightarrow$ 5 m experiment, for which only 5 blocks can be used at most in order to get pixels with a full receptive field. While this architecture may seem relatively shallow, it can be trained in a reasonable amount of time and seems to perform well enough for the purpose of assessing proposed metrics and strategy in a cross-sensor remote sensing SISR setting. The discriminator is a standard UNet with 32 features at the highest level and four levels, using Spectral Normalization as in [61]. The GAN loss term L_G^{Ra} is given by the Relativistic Average Discriminator formulation introduced in [60] (eq. 1 and 2). Label smoothing of 0.1 is introduced on the real term in order to prevent the discriminator to become overconfident [77]. The total generator loss is given by:

$$L_G = L_{lpips}(P, R) + \lambda L_G^{Ra} + \eta L_2(P, R), \qquad (22)$$

where LPIPS replaces the perceptual loss used in ESRGAN. Parameters λ=0.005 and η=0.1 have been chosen experimentally. Experiments are named using the following convention:

- **baseline**: R (no corrections),
- **geom**: $\widetilde{R}$ (geometric corrections only, given by equation 20),
- **geom+rad**: $\widetilde{R}^\star$ (geometric and radiometric corrections, given by equation 21).

As already mentioned, corrected data are only used for loss computation during training and **not** during testing: IQ metrics are computed on raw data. This is important to avoid bringing back domain gap issue, were performances are evaluated on corrected data that are not representative of real world data. Details of the training process can be found in appendix D.

C. Ablation study

This section investigates the effects of geometric correction alone or geometric and radiometric corrections strategies with respect to the baseline. Table III shows the measured performances for all experiments. Note that for the sake of readability, only one band is analyzed, corresponding to the band used for optical flow estimation. Remaining results can be found in the supplementary materials. Fig. 11 shows the $\mathcal{F}_{\mathcal{AP}}{}^\star$ for each training set and each of the 3 strategies. Reconstructions for each strategy for the **wsx4** dataset are shown in Fig. 12. Reconstruction examples for all other datasets can be found in the supplementary materials.

1) Geometric corrections: In all cases, the geometric correction strategy (**geom**) allows to reduce the measured geometric distortion of the predicted image with respect to baseline. Gains range from 2 HR pixels to 0.5 HR pixels in the case of the **ws×4** training set, and from 0.247 HR pixels to 0.07 HR pixels for the **s2v1×2** training set. The smallest gain is obtained with **s2v1×4**, while still halving the error. Fig. 14 illustrates the effectiveness of the proposed geometric correction strategy for models trained on **s2v1×2** training set, using patches from the Sen2Venμs 10 m testing set. Distortion in predicted images is clearly visible in baseline images and vanishes for both geometric correction alone and geometric and radiometric correction.

The geometric correction also has a small impact on other IQ metrics. For Sen2Venμs based training sets, AFR and FRR are lowered by a small amount, and BRISQUE also increases slightly. This is caused by the bicubic resampling with the optical flows, which introduces a bit of blur and aliasing depending on the strength of the local geometric correction. In Fig. 11, it can be observed that this reduction of AFR is targeted on higher spatial frequencies, which is consistent with blur introduced by resampling. Interestingly, RMSE_{LR} decreases, which shows that with lower geometric distortion, the SISR predicted image is more coherent with the input image. For Worldstrat training sets, geometric correction improves FRR by almost 10%, which suggests that larger geometric distortion may impair proper training, even when using adversarial training and perceptual losses. A closer look at the $\mathcal{F}_{\mathcal{AP}}{}^\star$ in Fig. 11 shows that this improvement occurs on mid spatial frequencies rather than higher frequencies, which points at a better consistency between input and target images. FRU also improves for Worldstrat training set, which again

TABLE III
PERFORMANCE COMPARISON OF THE BASELINE STRATEGY WITH RESPECT TO THE GEOMETRIC CORRECTION ALONE AND GEOMETRIC AND RADIOMETRIC CORRECTION STRATEGIES, FOR ALL TRAINING SETS. PERFORMANCES ARE ESTIMATED ON SEN2VENμS AND WORLDSTRAT TESTING SETS RESPECTIVELY, AS DESCRIBED IN SECTION B. HERE, THE ↑ (RESP. ↓) INDICATES THAT THE METRIC SHOULD BE MAXIMIZED (RESP. MINIMIZED). BEST VALUES ARE IN BOLD.

TS	Strategy	AFR↑	FRR↑	FRU↑	FRO↓	BRISQUE↓	LPIPS↓	$RMSE_{LR}$↓	GD↓
s2v1×2 (B4)	baseline	**5.33**	**53.85**	-0.45	**0.00**	**46.90**	0.081	5.39e-03	0.247
	geom	5.28	53.37	-0.53	**0.00**	46.95	**0.077**	5.01e-03	0.088
	geom+rad	4.88	49.37	**-0.40**	0.04	48.60	0.079	**4.31e-03**	**0.070**
s2v1×4 (B7)	baseline	**16.18**	**81.85**	-0.08	**0.00**	**37.98**	0.296	7.90e-03	0.202
	geom	15.73	79.60	-0.08	**0.00**	38.53	0.295	7.85e-03	0.134
	geom+rad	14.94	75.61	**-0.05**	**0.00**	38.21	**0.294**	**6.33e-03**	**0.099**
ws×4 (B4)	baseline	16.63	78.69	-0.13	**0.00**	42.29	0.351	8.61e-02	1.923
	geom	**18.25**	**86.35**	**-0.00**	**0.00**	43.70	**0.348**	8.16e-02	0.493
	geom+rad	14.32	67.77	-0.11	**0.00**	**42.11**	0.355	**1.13e-02**	**0.288**
ws×2 (B4)	baseline	13.81	77.72	-0.29	**0.00**	**34.43**	0.245	8.64e-02	0.973
	geom	**15.47**	**87.09**	**-0.01**	0.02	34.82	**0.236**	7.56e-02	0.217
	geom+rad	12.95	72.88	-0.20	**0.00**	35.27	0.256	**1.33e-02**	**0.117**

points out at large spatial distortions effects on the ability to learn SISR.

2) Geometric and radiometric corrections: For all training sets, the addition of radiometric correction (**geom+rad**) allows to improve $RMSE_{LR}$ with respect to the geometric correction only. Interestingly, it also further reduces the geometric distortion, which can be explained by the fact that a part of the input data is injected into the reference data. For Sen2Venμs training sets, $RMSE_{LR}$ improves by around 1e-3 reflectance, which is minor. This is due to the good radiometric consistency of the Sen2Venμs dataset. For the less radiometrically consistent Worldstrat training sets, the radiometric correction allows to divide $RMSE_{LR}$ by 8, demonstrating its efficiency. Looking at Fig. 12, this dramatic improvement is clearly visible, and patches predicted from models trained with **geom+rad** configuration are very consistent with input patches in terms of radiometry. Fig. 13 shows a scatter plot similar to Fig. 2 showing values for band B2 in Sentinel-2 input images with respect to the target Worldview image as well as predicted images with all configurations. The benefit of using the radiometric correction is clearly visible, as it is the only configuration that shows an even distribution around the diagonal.

Regarding SISR frequency restoration performances, radiometric correction has a higher impact than geometric correction: AFR is lowered by 1 to 4% while FRR diminishes by 4 to 10% with respect to baseline depending on the case. Fig. 11 shows that for Sen2Venμs based training sets, this drop focuses on mid spatial frequencies, with a further drop in higher spatial frequencies with respect to the geometric correction alone. For the Worldstrat derived training sets, introducing radiometric corrections leads to a further drop in higher frequencies, which suggests that residual LPF standard deviation σ_2 in equation 21 could be adapted to prevent leakage of residuals higher spatial frequencies. Last, FRU and FRO are almost unaffected by radiometric correction.

With respect to baseline, the BRISQUE score is a bit higher (lower image quality) in **geom+rad** configuration, with the notable exception of the **ws×4** training set, which shows a slight improvement. LPIPS improves slightly with respect to baseline for Sen2Venμs training set, whereas it decreases slightly for Worldstrat dataset. LPIPS values are also one order of magnitude larger for Worldstrat when compared to Sen2Venμs, which may be caused by the lower consistency of Worldstrat dataset.

3) Summary of findings: The ablation study demonstrates the efficiency of the proposed strategy: geometric correction reduces learned geometric distortion, and the additional radiometric correction reduces learned radiometric distortion. When the geometric distortion level is high, such as in the Worldstrat datasets, introducing geometric corrections facilitates the learning of the SISR task, yielding higher AFR. However, the combination with the radiometric correction always yields slightly lower AFR than the baseline, underlining the effect of introducing a part of the LR signal in the target HR patches. The very limited impact on the BRISQUE score indicates that this lower AFR value is not noticeable in terms of general IQ. Besides, the significant gain in terms of geometric and radiometric consistency will be paramount in most downstream applications.

D. Real-world scenario: comparing SISR models

This section investigates how metrics proposed in section II-D allow to gain fine grain insight on the comparison of different models. To that aim, it compares models trained with geometric and radiometric robust strategy (**geom+rad**) on training sets **s2v1×2**, **s2v2×2** and **ws×2**. All 3 models provide the same up-sampling factor of ×2 (10 m → 5 m) for the same spectral bands and can thus be compared on both the Sen2Venμs (**sv**) and Worldstrat (**ws**) testing sets. For those three models and for both **sv** and **ws** testing sets, table IV shows the main performance metrics, while Fig. 15 shows the $\mathcal{F}_{\mathcal{AP}}{}^{\star}$. Fig. 16 (resp. 17) shows sample predicted patches on the **sv** (resp. **ws**) testing set.

1) Frequency Domain Analysis: For both **s2v** and **ws** testing sets, **ws×2** has higher AFR and FRR than **s2v2×2**, which has higher AFR and FRR than **s2v1×2**. Interestingly, those trends matches the prior datasets analysis presented in B. A closer look at the $\mathcal{F}_{\mathcal{AP}}{}^{\star}$ of Fig. 15 highlight that **ws×2** is almost 4dB higher than both **s2v1×2** and **s2v2×2** for higher spatial frequencies, which explain its FRR of almost 95% on the **sv** testing set. On this testing set, it can be observed that

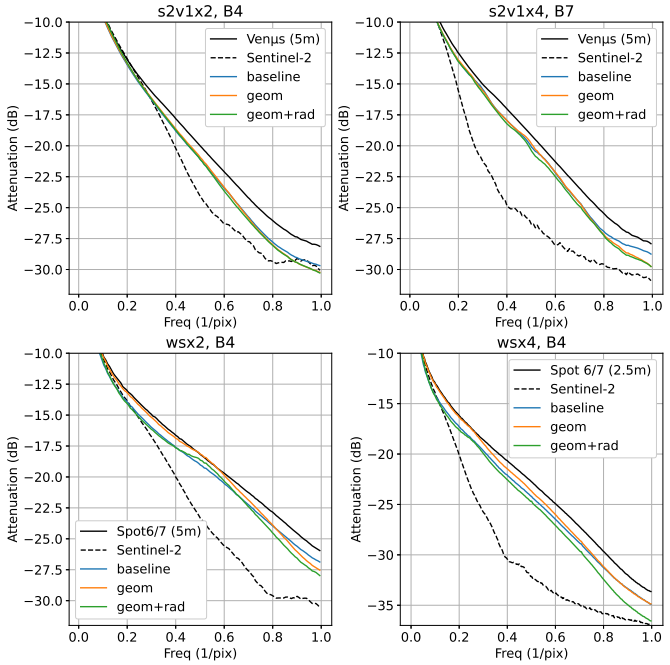

Fig. 11. Impact of proposed strategies on the $\mathcal{F}_{\mathcal{AP}}{}^{\star}$ for each of the training set, for band B4 (except for training set **s2v1×4**, where band B7 is used).

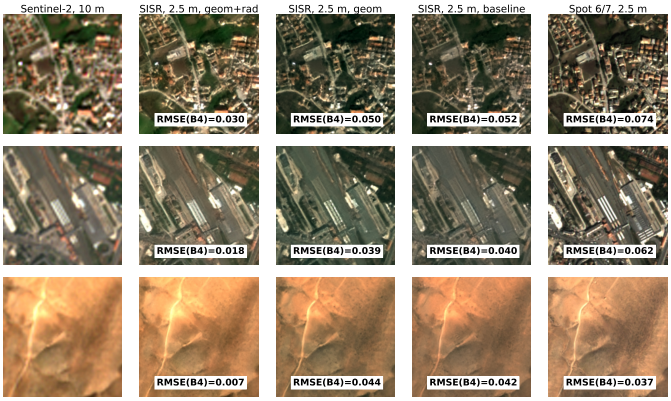

Fig. 12. Examples of predicted patches (Color composition: B4, B3, B2) from the testing set of Worldstrat, for SISR model trained on **ws×4** training set, with baseline, geom and geom+rad strategies. Note that the input image Sentinel-2 patches of first column have been up-sampled with bicubic interpolation (a larger version of this figure can be found in the supplementary materials).

the **ws×2** $\mathcal{F}_{\mathcal{AP}}{}^{\star}$ right end is actually higher than the reference HR profile, which explains its FRO of 2.77%. On the contrary, **s2v1×2** and **s2v2×2** have a slight restoration undershoot on the **ws** testing set, which is again visible on the right end of Fig. 15. This is confirmed by visual inspection of Fig. 16 and 17 where **ws×2** patches always appear sharper than the **s2v1×2** and **s2v2×2** patches. Patches from **s2v2×2** also appear sharper than those from **s2v1×2**, to a lesser extent.

2) Geometric and radiometric distortions: RMSE_{LR} and GD show that all models are free of geometric and radiometric distortions, though **ws×2** has higher RMSE_{LR} on the **sv** testing set and higher GD on the **ws** testing set. This is confirmed by visual inspection of Fig. 16 and 17.

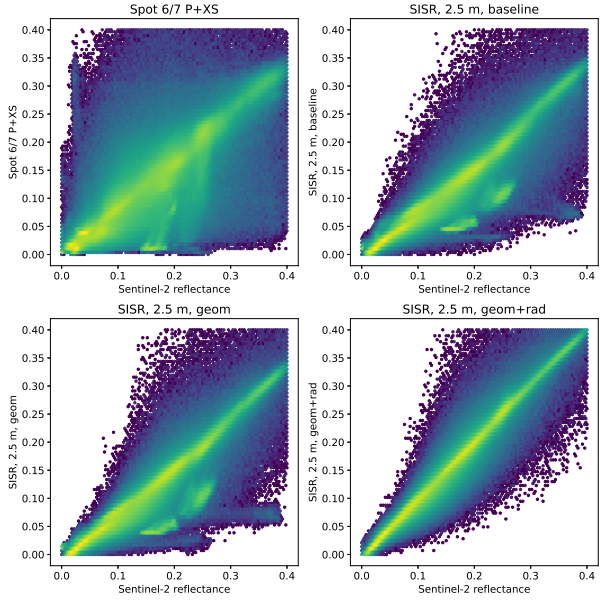

Fig. 13. Scatter plots of predicted patches from the testing set of Worldstrat, for SISR models trained on the **ws×4** training set, with baseline, geom and geom+rad strategies, with respect to the input Sentinel-2 reflectance, for band B4.

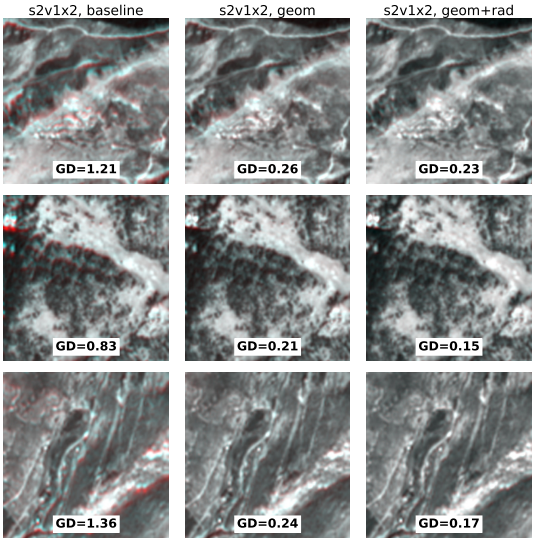

Fig. 14. Color composition highlighting geometric discrepancies for SISR models trained on the **s2v1×2** training set with the different strategies (R: B4 from Sentinel-2, B, G: B4 SISR predicted image). Geometric discrepancies appear in red or blue. Values for Geometric Distortion (GD) are computed for Sentinel-2 pixels, at 10 meter resolution.

3) General Image Quality: For any given training set, the BRISQUE score is lower on **ws** testing set than on **sv** testing set. A possible explanation is that urban patches with sharp edges, which BRISQUE may favor, are more prominent in patches from the **ws** testing set. For this testing set, the BRISQUE score models ordering follows FRR models ordering, which is the expected behavior. However, on the **sv** testing set, **ws×2** has the worst BRISQUE score and the higher FRR. One possible explanation is that due to the limited Worldstrat training set of **ws×2**, the model over-fits and fails

TABLE IV
PERFORMANCE COMPARISON OF MODELS TRAINED ON THE **s2v1×2**, **s2v2×2** AND **ws×2** TRAINING SETS, EVALUATED ON **s2v** TESTING SET (TOP) AND **ws** TESTING SET (BOTTOM) USING BAND B4. HERE, THE ↑ (RESP. ↓) INDICATES THAT THE METRIC SHOULD BE MAXIMIZED (RESP. MINIMIZED). BEST VALUES ARE IN BOLD.

Training	Test	AFR↑	FRR↑	FRU↑	FRO↓	BRISQUE↓	LPIPS↓	$RMSE_{LR}$↓	GD↓
s2v1×2	s2v	4.88	49.37	-0.40	**0.04**	48.60	0.079	4.31e-03	0.070
s2v2×2	s2v	6.58	66.51	**-0.11**	0.06	**45.40**	**0.073**	**4.02e-03**	**0.059**
ws×2	s2v	**9.34**	**94.43**	-0.45	2.77	50.38	0.103	5.55e-03	0.085
s2v1×2	ws	3.24	18.25	-0.91	**0.00**	43.69	0.271	**1.31e-02**	**0.074**
s2v2×2	ws	4.66	26.26	-0.50	**0.00**	40.94	0.267	1.32e-02	0.085
ws×2	ws	**12.95**	**72.88**	**-0.20**	**0.00**	**35.27**	**0.256**	1.33e-02	0.117

to generalize to the more natural landscapes of **sv** testing set. The same trend also affects LPIPS.

4) Summary of findings: This experiment shows that metrics proposed in sectionII-D allows to cross-compare the three models on two independent testing sets, using the raw HR images, and provide a good insight of the model relative strengths and weaknesses: the **ws×2** model provides higher spatial frequency restoration but seems to over-fit the Worldstrat dataset, while **s2v2×2** provides lower spatial frequency restoration but offers better generalization. Another important finding is that the *a priori* filtering of patches based on their PFR that yielded the **s2v1×2** and **s2v2×2** training sets allowed to consistently obtain higher FRR through SISR training. This opens interesting perspectives when dealing with large SISR datasets. Finally, the radiometric and geometric distortions robust strategy proposed in section III-A performs consistently well across all experiments.

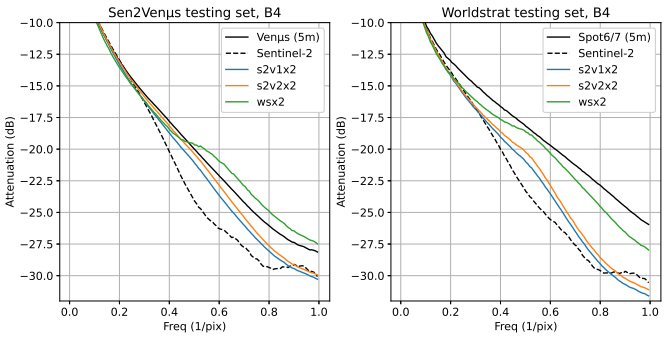

Fig. 15. $\mathcal{F}_{\mathcal{AP}}{}^{\star}$ of SISR model trained on **s2v1×2**, **s2v2×2**, and **ws×2** training sets, with the **geom+rad** strategy, computed on B4 of Sen2Venμs testing set (left) and Worldstrat testing set (right).

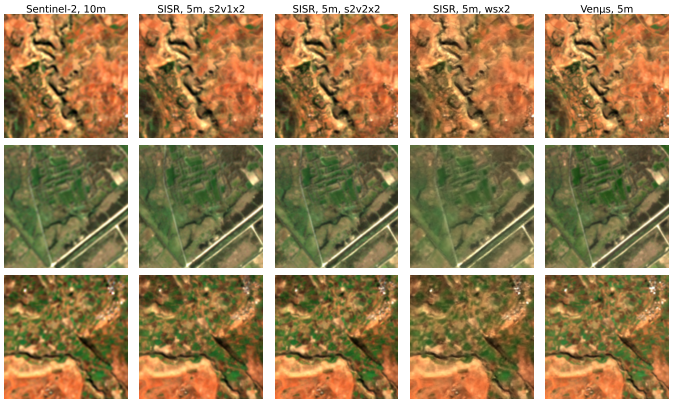

Fig. 16. Comparison of models **s2v1×2**, **s2v2×2**, and **ws×2** trained with the **geom+rad** strategy, on patches from the Sen2Venμs testing set (a larger version of this figure can be found in the supplementary materials).

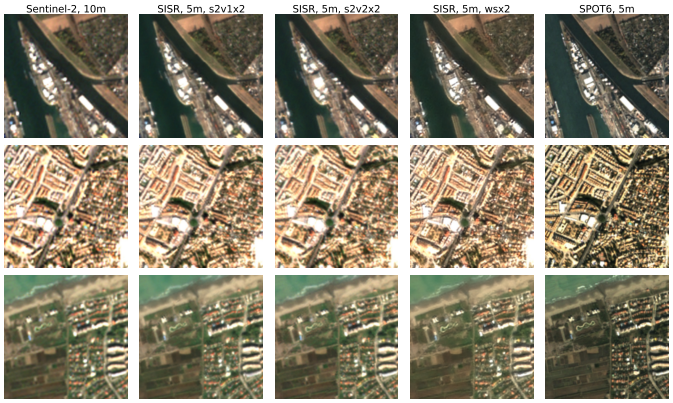

Fig. 17. Comparison of models **s2v1×2**, **s2v2×2**, and **ws×2** trained with the **geom+rad** strategy, on patches from the Worldstrat testing set (a larger version of this figure can be found in the supplementary materials).

E. Computational efficiency

It is important to note that the proposed method does not incur any additional computational cost at inference time. It induces a very small overhead during training time. Completing the training for the s2v2x2 dataset took 22.14 hours for the **baseline** configuration and 22.97 hours for the **spat+rad** configuration. Completing training of the wsx4 dataset took 8.8 hours for the **baseline** configuration and 9.15 hours for the **spat+rad** configuration. Most of the overhead comes from the use of the auxiliary UNet of the **spat** strategy, as the difference in training time between **spat** and **spat+rad** is insignificant. The training time overhead is estimated at around 4%. Of course, the training of the auxiliary UNet also needs to be accounted for: it converges in 5 hours for the Sen2Venμs dataset and 10.3 hours for the Worldstrat dataset. However, the same auxiliary UNet can be used for many training and also for computing the GD metric in many situations, it can not directly be imputed to the training overhead of the **spat** and **spat+rad** configurations.

IV. DISCUSSION AND CONCLUSION

A. Discussion

1) Alternative strategy: The robust strategy proposed in section III-A can be seen as a way of fixing the target patches for discrepancies, though this fix only occurs during training.

Experiments have demonstrated that, while this strategy is efficient in preventing the model to learn geometric and radiometric discrepancies, it comes at the price of slightly lowering AFR, and also impacts the BRISQUE score to a lesser extent. A possible alternative strategy could be to take advantage of the fully convolutional optical flow estimation network that is able to back-propagate gradients, and trade the spatial resampling of the target HR patches for an additional loss term aiming at minimizing GD of predicted patches directly. Additionally, the L_2 term of the loss given by equation 22 could be replaced by RMSE_{LR}, as it essentially plays the same role of constraining the network toward radiometric faithfulness. The full generator loss of equation 22 would then become:

$$\begin{aligned} L_G(P,R,X) =& L_{LPIPS}(P,R) + \lambda L_G^{Ra} + \eta RMSE_{LR}(P,X) \\ &+ \nu GD(P,X), \end{aligned} \tag{23}$$

where ν is an additional weighting parameter.

Though beyond the scope of this manuscript, this alternative loss completely avoids modifying the HR target patches and thus may yield better performances.

2) Limitations of FDA: Though the experiments of section III-D have demonstrate that FDA can be reliably used to assess and compare the frequency restoration performances of SISR models. Section B also demonstrated its dependency to the content of the patches used for its computation. Moreover, noise or periodic patterns may cause an artificial bump in higher spatial frequencies restoration which does not relate directly to sharper images. Evidence of the impact of noise can be observed for instance in Fig. 15 on the Sentinel-2 $\mathcal{F}_{\mathcal{AP}}{}^{\star}$ profile. The monotonic decreasing trend of frequency in the range [0, 0.8] suddenly changes for a plateau in the range [0.8,1.], which is better explain by noise in the B4 band rather than by an abrupt bump in sharpness. It is therefore very important to combine FDA with other metrics such as BRISQUE, LPIPS or PieAPP which allow to disambiguate such cases by providing different insights on image quality. Still, those perceptual metrics also exhibit a tropism toward noise or periodic patterns, as demonstrated in section II-B.

A final limitation of proposed FDA metrics is that their computation depends on a target reference image and therefore prevents a true No Reference assesment of spatial frequency restoration. However, equation 13 could easily be simplified in order to avoid clamping to and normalizing by the target $\mathcal{F}_{\mathcal{AP}}{}^{\star}$, thus creating a No Reference AFR that does not no longer depend on the target HR patches:

$$NRAFR(P_b, X_b) = \sum \mathcal{F}_{\mathcal{AP}}{}^{\star}[P_b] - \mathcal{F}_{\mathcal{AP}}{}^{\star}[X_b]. \tag{24}$$

3) About perceptual metrics: BRISQUE, LPIPS, PieAPP and other perceptual metrics are obtained by training machine learning algorithms on natural images, sometimes with human supervision. As such, they usually assume RGB bands, a limited data range and are tailored for images that are very far from the manifold of remote sensing images. Existing works as well as this paper show that these limitations can be solved by simple solutions such as data range rescaling and greyscale RGB composition with any given spectral band from remote sensing data. However, it would be interesting for the remote sensing community to build and maintain perceptual metrics tailored for remote sensing data. This may include for instance training for specific bands or range of sensors. Such *ad hoc* metrics could then avoid the bias toward noise and periodic patterns observed in sections II-B2 and II-B3 and drive models toward better SISR prediction when used as loss terms.

4) Is ×4 model better than ×2 ?: An everlasting question about remote sensing SISR is how far can we go in terms of up-sampling factor. Though many works have demonstrated impressive performances for factors far beyond ×4, it is also quite reasonable to think that the higher the up-sampling factor, the more SISR relies on generative capabilities of the network and learning probabilistic relationships between the input LR and the target HR manifolds, as opposed to restoring information that can actually be found in LR images. The proposed FDA gives a frequency domain perspective of this restoration, but can not determine if the restored higher frequencies correspond to any ground truth or are plain hallucination of the model. It is again of paramount importance to combine FDA with other metrics such as LPIPS, but this may not be enough to characterize the behavior of trained model on out of distribution samples.

Fig. 18 shows bicubic up-sampling, 5 m up-sampling with the model trained on **s2v2×2** and 2.5 m up-sampling with the model trained on **ws×4**, for a Sentinel-2 image which is not part of any of the datasets used in this paper, where words are painted on a track of the Le Bourget airport, France. From the analysis conducted in section III-D with the proposed metrics, we know that the **s2v2×2** model is better than bicubic up-sampling, since it has non null AFR and generalizes well. We also know that the **ws×4** model evaluated in section III-C has much higher AFR, and that it probably slightly overfits the training set as its sibling model **ws×2**. When fed with the painted words on the airport track, the **s2v2×2** model acts like a signal restoration model: letters become more visible and can almost be deciphered, contrary to the bicubic up-sampling. On the other hand, while the **ws×4** model provides a better overall sharpness of the image, its prediction of the painted words yields letters that are completely scrambled and are actually more difficult to decipher than the bicubic up-sampling. Of course, this poor performance may be caused by over-fitting the **ws×4** training set but it also highlights that larger up-sampling factors by lead to poorer generalization.

5) Impact on future cross-sensor SISR work: Many recent works on SISR have used PSNR, SSIM and other metrics on datasets that are probably affected by either radiometric distortion, geometric distortion, or both [21], [78]–[83]. Under those unreliable conditions, some of them advertise only marginal improvements of those metrics with respect to state-of-the-art methods. In this situation, the set of metrics proposed in section II-D would provide unbiased insights on the strengths and weaknesses of the compared methods, by providing a fair assessment of spatial frequency restoration, learned geometric and radiometric distortions as well as general IQ.

Fig. 18. Can you read it ? Sentinel-2 image of Le Bourget airport, France. From top to bottom, Sentinel-2 up-sampled to 5 m with bicubic, Sentinel-2 up-sampled to 5 m with SISR model trained on **s2v2×2**, and Sentinel-2 up-sampled to 2.5 m with SISR model trained on **ws×4** (a larger version of this figure can be found in the supplementary materials).

B. Conclusion

In this paper, we address the overlooked impact of geometric and radiometric discrepancies found in cross-sensor datasets for remote sensing SISR. Through a dedicated benchmark of common SISR IQ metrics, we demonstrate that widely used local metrics such as PSNR or SSIM are not adapted for cross-sensor SISR evaluation. Instead, we identify perceptual metrics such as LPIPS as robust to discrepancies, and complement them with new FDA metrics tailored to assess spatial frequency restoration performances. In addition, RMSE with respect to the input image can be used to measure the level of learned radiometric distortion, and we demonstrate that a dedicated UNet estimating optical flows can be used to assess the level of learned geometric distortion. We then propose a robust strategy for cross-sensor SISR DL model learning, divided into two parts. During training, we improve dataset consistency by means of spatial registration of HR targets with optical flows estimated by the pre-trained UNet to compensate for geometric distortion and LR residual injections to compensate for radiometric distortion. During evaluation, the proposed metric set is used on uncorrected reference data, which ensures robust performance evaluation with respect to distortions, and avoids potential biases from the correction data consistency enhancement process. Experiments demonstrate the effectiveness of the proposed robust strategy, but also the ability of the proposed metric set to provide a fair, in depth comparison of SISR models that is independent of the testing set and unaffected by its discrepancies.

A detailed comparison between the Worldstrat and the Sen2Venμs datasets is also proposed, which highlights that the former has higher potential for spatial frequency restoration while the latter has a strong geometric and radiometric consistency. Finally, this work highlights the paramount importance of dataset consistency for SISR, and we hope it paves the way to a better understanding and attention in future remote sensing SISR works and beyond. For instance, registration is probably a crucial but overlooked aspect of spatio-temporal fusion. Meanwhile, future work on cross-sensor SISR would benefit from the proposed strategy in order to take into account the effects of radiometric and geometric distortions in their training and bench-marking.

ACKNOWLEDGMENTS

- This work was partly performed using HPC resources from GENCI-IDRIS (Grant 2023-AD010114835)
- This work was partly performed using HPC resources from CNES Computing Center.
- The authors acknowledge funding from the EvoLand project (Evolution of the Copernicus Land Service portfolio, grant agreement No 101082130) funded from the European Union's Horizon Europe research and innovation programme.
- The authors would like to thank Julia Gottfriedsen and Christian Mollière for the fruitful discussion about Frequency Domain Analysis, and Jérémy Anger for the discussion about his work on [24], and especially on optical flow estimation, as well as for providing the testing location for Fig. 18, during ESA SUREDOS workshop (05.20224).
- The authors would to thank Mathieu Fauvel for the careful proof-reading of the manuscript.

REFERENCES

[1] T. Chan and C.-K. Wong, "Total variation blind deconvolution," *IEEE Transactions on Image Processing*, vol. 7, no. 3, pp. 370–375, 1998.

[2] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman, "Understanding blind deconvolution algorithms," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 12, pp. 2354–2367, 2011.

[3] S. Anwar, S. Khan, and N. Barnes, "A deep journey into super-resolution: A survey," *ACM Computing Surveys (CSUR)*, vol. 53, no. 3, pp. 1–34, 2020.

[4] K. Li, S. Yang, R. Dong, X. Wang, and J. Huang, "Survey of single image super-resolution reconstruction," *IET Image Processing*, vol. 14, no. 11, pp. 2273–2290, 2020.

[5] H. Chen, X. He, L. Qing, Y. Wu, C. Ren, R. E. Sheriff, and C. Zhu, "Real-world single image super-resolution: a brief review," *Information Fusion*, vol. 79, pp. 124–145, 2022.

[6] P. Wang, B. Bayram, and E. Sertel, "A comprehensive review on deep learning based remote sensing image super-resolution methods," *Earth-Science Reviews*, vol. 232, p. 104110, 2022.

[7] G. Rohith and L. S. Kumar, "Paradigm shifts in super-resolution techniques for remote sensing applications," *The Visual Computer*, vol. 37, no. 7, pp. 1965–2008, 2021.

[8] H. Liu, Y. Qian, X. Zhong, L. Chen, and G. Yang, "Research on super-resolution reconstruction of remote sensing images: a comprehensive review," *Optical Engineering*, vol. 60, no. 10, p. 100901, 2021.

[9] D. Pouliot, R. Latifovic, J. Pasher, and J. Duffe, "Landsat super-resolution enhancement using convolution neural networks and Sentinel-2 for training," *Remote Sensing*, vol. 10, no. 3, p. 394, 2018.

[10] C.-H. Lin and J. M. Bioucas-Dias, "An explicit and scene-adapted definition of convex self-similarity prior with application to unsupervised Sentinel-2 super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 5, pp. 3352–3365, 2019.

[11] C. Lanaras, J. Bioucas-Dias, S. Galliani, E. Baltsavias, and K. Schindler, "Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 146, pp. 305–319, 2018.

[12] M. Galar, R. Sesma, C. Ayala, and C. Aranda, "Super-resolution for Sentinel-2 images," *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 2019.

[13] F. Chouteau, L. Gabet, R. Fraisse, T. Bonfort, B. Harnoufi, V. Greiner, M. Le Goff, M. Ortner, and V. Paveau, "Joint super-resolution and image restoration for PLÉIADES NEO imagery," *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 43, pp. 9–15, 2022.

[14] X. Zhu, H. Talebi, X. Shi, F. Yang, and P. Milanfar, "Super-resolving commercial satellite imagery using realistic training data," in *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 498–502, 2020.

[15] Y. Tao and J.-P. Muller, "Super-resolution restoration of spaceborne ultra-high-resolution images using the UCL OpTiGAN system," *Remote Sensing*, vol. 13, no. 12, 2021.

[16] E. Agustsson and R. Timofte, "NTIRE 2017 challenge on single image super-resolution: Dataset and study," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[17] N. Efrat, D. Glasner, A. Apartsin, B. Nadler, and A. Levin, "Accurate blur models vs. image priors in single image super-resolution," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[18] M. S. Rad, T. Yu, C. Musat, H. K. Ekenel, B. Bozorgtabar, and J.-P. Thiran, "Benefiting from bicubically down-sampled images for learning real-world image super-resolution," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1590–1599, January 2021.

[19] N. Zhao, C. Zhang, H. Zhang, and Z. Jiang, "How down-sampling affects supervised-learning-based image super-resolutions," in *SPIE Future Sensing Technologies 2024*, 5 2024.

[20] A. Liu, Y. Liu, J. Gu, Y. Qiao, and C. Dong, "Blind image super-resolution: a survey and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–19, 2022.

[21] J. Min, Y. Lee, D. Kim, and J. Yoo, "Bridging the domain gap: a simple domain matching method for reference-based image super-resolution in remote sensing," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024.

[22] A. Valenzuela, K. Reinke, and S. Jones, "Assessing the spatial resolution distance of satellite images: Superdove versus Landsat 8," *International Journal of Remote Sensing*, vol. 45, no. 12, pp. 4120–4159, 2024.

[23] N. L. Nguyen, J. Anger, L. Raad, B. Galerne, and G. Facciolo, "On The Role of Alias and Band-Shift for Sentinel-2 Super-Resolution," in *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, pp. 4294–4297, 2023.

[24] N. L. Nguyen, J. Anger, A. Davy, P. Arias, and G. 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 (CVPR)*, 2023.

[25] S. Bojinski, M. Verstraete, T. C. Peterson, C. Richter, A. Simmons, and M. Zemp, "The concept of essential climate variables in support of climate research, applications, and policy," *Bulletin of the American Meteorological Society*, vol. 95, no. 9, pp. 1431 – 1443, 2014.

[26] W. Jetz, M. A. McGeoch, R. Guralnick, S. Ferrier, J. Beck, M. J. Costello, M. Fernandez, G. N. Geller, P. Keil, C. Merow, *et al.*, "Essential biodiversity variables for mapping and monitoring species populations," *Nature ecology & evolution*, vol. 3, no. 4, pp. 539–551, 2019.

[27] G. Misra, F. Cawkwell, and A. Wingler, "Status of phenological research using Sentinel-2 data: A review," *Remote Sensing*, vol. 12, no. 17, p. 2760, 2020.

[28] D. Phiri, M. Simwanda, S. Salekin, V. R. Nyirenda, Y. Murayama, and M. Ranagalage, "Sentinel-2 data for land cover/use mapping: A review," *Remote Sensing*, vol. 12, no. 14, p. 2291, 2020.

[29] B. Vajsová, D. Fasbender, C. Wirnhardt, S. Lemajic, and W. Devos, "Assessing spatial limits of Sentinel-2 data on arable crops in the context of checks by monitoring," *Remote Sensing*, vol. 12, no. 14, p. 2195, 2020.

[30] U. Bhangale, S. More, T. Shaikh, S. Patil, and N. More, "Analysis of surface water resources using Sentinel-2 imagery," *Procedia Computer Science*, vol. 171, pp. 2645–2654, 2020.

[31] F. Xiaolin, H. Fan, Y. Ming, Z. Tongxin, B. Ran, Z. Zenghui, and G. Zhiyuan, "Small object detection in remote sensing images based on super-resolution," *Pattern Recognition Letters*, vol. 153, pp. 107–112, 2022.

[32] Y. Liu, Z. Xiong, Y. Yuan, and Q. Wang, "Distilling knowledge from super-resolution for efficient remote sensing salient object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[33] A. Lac, J. Michel, V. Poulain, and N. Sfaksi, "Sentinel-2 Single Image Super-Resolution with the SEN2VENμS Dataset: architecture, training strategy, performances assessment and application to Water Bodies Detection." Submitted to IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Sept. 2023.

[34] M. Qin, L. Hu, Z. Du, Y. Gao, L. Qin, F. Zhang, and R. Liu, "Achieving higher resolution lake area from remote sensing images through an unsupervised deep learning super-resolution method," *Remote Sensing*, vol. 12, no. 12, p. 1937, 2020.

[35] H. Sun, Z. Wei, W. Yu, G. Yang, J. She, H. Zheng, C. Jiang, X. Yao, Y. Zhu, W. Cao, T. Cheng, and I. Ali, "Sidest: a sample-free framework for crop field boundary delineation by integrating super-resolution image reconstruction and dual edge-corrected segment anything model," *Computers and Electronics in Agriculture*, vol. 230, p. 109897, 2025.

[36] H. Han, Z. Feng, W. Du, S. Guo, P. Wang, and T. Xu, "Remote sensing image classification based on multi-spectral cross-sensor super-resolution combined with texture features: a case study in the liaohe planting area," *IEEE Access*, vol. 12, pp. 16830–16843, 2024.

[37] D. Chen, Z. Zhang, J. Liang, and L. Zhang, "SSL: A Self-similarity Loss for Improving Generative Image Super-resolution," 2024.

[38] A. Malczewska, J. Malczewski, and B. Hejmanowska, "Challenges in preparing datasets for super-resolution on the example of Sentinel-2 and Planet Scope images," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLVIII-1/W3-2023, pp. 91–98, 2023.

[39] M. Galar, R. Sesma, C. Ayala, L. Albizua, and C. Aranda, "Super-resolution of Sentinel-2 images using convolutional neural networks and real ground truth data," *Remote Sensing*, vol. 12, no. 18, p. 2941, 2020.

[40] F. Pineda, V. Ayma, and C. 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*, vol. 43, pp. 9–14, 2020.

[41] L. Salgueiro Romero, J. Marcello, and V. Vilaplana, "Super-resolution of Sentinel-2 imagery using generative adversarial networks," *Remote Sensing*, vol. 12, no. 15, p. 2424, 2020.

[42] C. Aybar, D. Montero, S. Donike, F. Kalaitzis, and L. Gómez-Chova, "A comprehensive benchmark for optical remote sensing image super-resolution," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024.

[43] J. Cornebise, I. Oršolić, and F. Kalaitzis, "Open High-Resolution Satellite Imagery: The WorldStrat Dataset –With Application to Super-Resolution," in *Advances in Neural Information Processing Systems* (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds.), vol. 35, pp. 25979–25991, Curran Associates, Inc., 2022.

[44] P. Kowaleczko, T. Tarasiewicz, M. Ziaja, D. Kostrzewa, J. Nalepa, P. Rokita, and M. Kawulok, "A real-world benchmark for Sentinel-2 multi-image super-resolution," *Scientific Data*, vol. 10, no. 1, p. 644, 2023.

[45] A. Okabayashi, N. Audebert, S. Donike, and C. Pelletier, "Cross-sensor super-resolution of irregularly sampled Sentinel-2 time series," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 502–511, June 2024.

[46] C. Aybar, D. Montero, J. Contreras, S. Donike, F. Kalaitzis, and L. Gómez-Chova, "SEN2NAIP: A large-scale dataset for Sentinel-2 Image Super-Resolution," *SEN2NAIP: A large-scale dataset for Sentinel-2 Image Super-Resolution*, 2024.

[47] G. Vivone, L. Alparone, J. Chanussot, M. Dalla Mura, A. Garzelli, G. A. Licciardi, R. Restaino, and L. Wald, "A critical comparison among pansharpening algorithms," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 5, pp. 2565–2586, 2014.

[48] J. Michel, J. Vinasco-Salinas, J. Inglada, and O. Hagolle, "Sen2venμs, a dataset for the training of Sentinel-2 super-resolution algorithms," *Data*, vol. 7, no. 7, p. 96, 2022.

[49] A. Dick, J.-L. Raynaud, A. Rolland, S. Pelou, S. Coustance, G. Dedieu, O. Hagolle, J.-P. Burochin, R. Binet, and A. Moreau, "Venμs: Mission characteristics, final evaluation of the first phase and data production," *Remote Sensing*, vol. 14, no. 14, 2022.

[50] V. Lonjou, C. Desjardins, O. Hagolle, B. Petrucci, T. Tremas, M. Dejus, A. Makarau, and S. Auer, "Maccs-atcor joint algorithm (MAJA)," in *Remote Sensing of Clouds and the Atmosphere XXI*, vol. 10001, p. 1000107, International Society for Optics and Photonics, 2016.

[51] P. Wolters, F. Bastani, and A. Kembhavi, "Zooming out on zooming in: Advancing super-resolution for remote sensing," 2023.

[52] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[53] M. Märtens, D. Izzo, A. Krzic, and D. Cox, "Super-resolution of PROBA-V images using convolutional neural networks," *Astrodynamics*, vol. 3, no. 4, pp. 387–402, 2019.

[54] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[55] Z. Wang and A. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[56] H. Sheikh, A. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on Image Processing*, vol. 14, no. 12, pp. 2117–2128, 2005.

[57] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: a highly efficient perceptual image quality index," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 684–695, 2014.

[58] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[59] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, "PieAPP: Perceptual Image-Error Assessment Through Pairwise Preference," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[60] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European conference on computer vision (ECCV) workshops*, pp. 0–0, 2018.

[61] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[62] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[63] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.

[64] V. N, P. D, M. C. Bh, S. S. Channappayya, and S. S. Medasani, "Blind image quality evaluation using perception based features," in *2015 Twenty First National Conference on Communications (NCC)*, 2 2015.

[65] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang, "Learning a no-reference quality metric for single-image super-resolution," *Computer Vision and Image Understanding*, vol. 158, pp. 1–16, 2017.

[66] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images,"

[67] X. Li, Z. Wang, and C. Xie, "CLIPA-v2: Scaling CLIP Training with 81.1% Zero-shot ImageNet Accuracy within a $10,000 Budget; An Extra $4,000 Unlocks 81.8% Accuracy," 2023.

[68] S. Kastryulin, J. Zakirov, D. Prokopenko, and D. V. Dylov, "Pytorch image quality: Metrics for image quality assessment," 2022.

[69] A. Balanov, A. Schwartz, Y. Moshe, and N. Peleg, "Image quality assessment based on DCT subband similarity," in *2015 IEEE International Conference on Image Processing (ICIP)*, 9 2015.

[70] M. Narwaria, W. Lin, I. V. McLoughlin, S. Emmanuel, and L.-T. Chia, "Fourier transform-based scalable image quality measure," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3364–3377, 2012.

[71] E. Mavridaki and V. Mezaris, "No-reference blur assessment in natural images using fourier transform and spatial pyramids," in *2014 IEEE International Conference on Image Processing (ICIP)*, 10 2014.

[72] K. De and V. Masilamani, "Image sharpness measure for blurred images in frequency domain," *Procedia Engineering*, vol. 64, pp. 149–158, 2013.

[73] M. Hou, Z. Huang, Z. Yu, Y. Yan, Y. Zhao, and X. Han, "Cswt-Sr: Conv-Swin Transformer for Blind Remote Sensing Image Super-Resolution With Amplitude-Phase Learning and Structural Detail Alternating Learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024.

[74] E. Cohen and Y. Yitzhaky, "No-reference assessment of blur and noise impacts on image quality," *Signal, Image and Video Processing*, vol. 4, no. 3, pp. 289–302, 2009.

[75] A. Chetouani, A. Beghdadi, and M. Deriche, "A new reference-free image quality index for blur estimation in the frequency domain," in *2009 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 12 2009.

[76] F. Gao, W. P. Kustas, and M. C. Anderson, "A data mining approach for sharpening thermal satellite imagery over land," *Remote Sensing*, vol. 4, no. 11, pp. 3287–3319, 2012.

[77] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?," in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.

[78] L. Rossi, V. Bernuzzi, T. Fontanini, M. Bertozzi, and A. Prati, "Swin2-MoSE: A New Single Image Super-Resolution Model for Remote Sensing," 2024.

[79] H. Zhang, P. Wang, and Z. Jiang, "Nonpairwise-trained cycle convolutional neural network for single remote sensing image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 4250–4261, 2021.

[80] H. Zhang, C. Zhang, F. Xie, and Z. Jiang, "A closed-loop network for single infrared remote sensing image super-resolution in real world," *Remote Sensing*, vol. 15, no. 4, p. 882, 2023.

[81] R. Dong, L. Zhang, and H. Fu, "Rrsgan: Reference-based super-resolution for remote sensing image," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022.

[82] Z. Zhang, K. Gao, J. Wang, L. Min, S. Ji, C. Ni, and D. Chen, "Gradient enhanced dual regression network: Perception-preserving super-resolution for multi-sensor remote sensing imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[83] Z. Tu, X. Yang, X. He, J. Yan, and T. Xu, "Rgtgan: Reference-based gradient-assisted texture-enhancement gan for remote sensing super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–21, 2024.

[84] D. Scheffler, A. Hollstein, H. Diedrich, K. Segl, and P. Hostert, "Arosics: an Automated and Robust Open-Source Image Co-Registration Software for Multi-Sensor Satellite Data," *Remote Sensing*, vol. 9, no. 7, p. 676, 2017.

[85] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, no. 1, pp. 185–203, 1981.

[86] B. D. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," in *IJCAI'81: 7th international joint conference on Artificial intelligence*, vol. 2, (Vancouver, Canada), pp. 674–679, Aug. 1981.

[87] M. Zhai, X. Xiang, N. Lv, and X. Kong, "Optical flow and scene flow estimation: a survey," *Pattern Recognition*, vol. 114, p. 107861, 2021.

[88] A. Ammar, A. Chebbah, H. B. Fredj, and C. Souani, "Comparative study of latest cnn based optical flow estimation," in *2022 International Conference on Intelligent Systems and Computer Vision (ISCV)*, 5 2022.

[89] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds.), (Cham), pp. 234–241, Springer International Publishing, 2015.

[90] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[91] D. P. Kingma and J. Ba, "Adam: a Method for Stochastic Optimization," 2014.

[92] I. Loshchilov and F. Hutter, "SGDR: Stochastic Gradient Descent With Warm Restarts," 2016.

[93] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), pp. 8024–8035, Curran Associates, Inc., 2019.

Julien Michel received the Telecommunications Engineer degree from the École Nationale Supérieure des Télécommunications de Bretagne, Brest, France, in 2006. He is currently with the Centre National d'Études Spatiales (French Space Agency), Toulouse, France, where he is working as a research engineer on remote sensing image processing at the Centre d'Études Spatiales de la Biosphère (CESBIO). He is also currently pursuing the Ph.D. degree with the Centre d'Études Spatiales de la Biosphère (CESBIO) Laboratory. His main research topic focuses on the fusion of heterogeneous Satellite Image Time Series (SITS) to enhance their spatial and temporal resolutions. His wider research interests include image processing and machine learning for remote sensing data.

Jordi Inglada received the master's degree in telecommunications engineering from the Universitat Politècnica de Catalunya, Barcelona, Spain, and the École Nationale Supérieure des Télécommunications de Bretagne, Brest, France, in 1997, and the Ph.D. degree in signal processing and telecommunications from the Université de Rennes 1, Rennes, France, in 2000. He is currently with the Centre National d'Études Spatiales (French Space Agency), Toulouse, France, where he is involved in the field of remote sensing image processing with the Centre d'Études Spatiales de la Biosphère (CESBIO) Laboratory. He is involved in the development of image processing algorithms for the operational exploitation of Earth observation images, mainly in the field of multitemporal image analysis for land use and cover change.

Ekaterina Kalinicheva received the engineering degree in Applied Geodesy from the Moscow State University of Geodesy and Cartography, Moscow, Russia, in 2012, and the M.Sc. degree in Geomatics from the University of Montpellier, Montpellier, France, in 2017. She obtained her Ph.D. degree in Informatics from Sorbonne University, Paris, France. Her Ph.D. research focused on unsupervised analysis of satellite image time series, particularly multitemporal change detection and multi-temporal clustering. She subsequently held a postdoctoral research position at the LASTIG Laboratory, where she worked on multi-modal bi-temporal analysis of forestry data, using airborne LiDAR scans paired with very high-resolution aerial imagery. She is currently a postdoctoral researcher at CESBIO Laboratory, Toulouse, France, working on the EVOLAND project. Her current research involves developing mono- and multi-date, mono- and multi-modal embedding algorithms for Sentinel-1 and Sentinel-2 imagery using self-supervised neural network algorithms. Her main research interests include machine learning, image analysis, 3D forestry data analysis, multi-temporal and multi-modal analysis, and neural networks applied to various remote sensing applications.

APPENDIX A
MEASURING CROSS-SENSOR GEOMETRIC DISTORTION

The idea of providing pixel wise registration in SISR datasets has been explored in [38], where they use the AROSICS algorithm [84] for the task. In this section, we propose to use an auxiliary optical flow estimation network similar to the Cross Spectral Registration network proposed in [24]. Optical flow is the 2-dimensional pixel-wise displacement field yielded by a moving imaging device over a 3-dimensional field, and its estimation from image pairs or sequences has been widely studied since the seminal work of Horn and Shunck [85] and Lucas and Kanade [86]. Since geometric discrepancies in cross-sensor SISR datasets are mainly caused by differences in viewing angles, it seems natural to address this issue with the tools from the stereo-vision field. Recent literature reviews [87], [88] show that UNet [89], which is also used in [24], is the most widely adopted solution for estimating optical flow, and will therefore be used in this work. It must be stressed that proposed usages of optical flow estimation in this paper are not specific of the UNet architecture and training procedure, and other solutions might be used while retaining their benefits. However, using a fully convolutional neural network comes with the benefit of being differentiable, and can thus be used in end-to-end training.

A. Optical flow estimation for cross-sensor datasets

Fig. 19 shows the overview of the proposed UNet based estimation of the optical flow in a cross-sensor SISR context. Though [24] suggest that the so-called Cross Registration Network can learn to estimate optical flow from different input bands. Here, corresponding spectral bands between LR and HR sensors will be used, since there are corresponding bands in all considered datasets.

The HR reference image R_b is first low-pass filtered and downsampled to LR resolution by means of:

$$X_b^\star = \left(R_b * \phi_{\sigma_0}\right) \downarrow_s, \tag{25}$$

where $\downarrow_s$ denotes the decimation operator by a stride of s and ϕ_{σ_0} is a Gaussian kernel as introduced in equation 7. LR input patches X_b are concatenated to $X_b^\star$ patches along the channel dimension and fed to the UNet with parameters Θ_{UNet}. The output of the UNet goes through a last 2D convolution layer forming a 2-channel output, which in turns goes through a hyperbolic tangent activation layer, bringing the channel data range to $[-1, 1]$. The resulting channels are scaled the with optical field range parameter r. It is worth noting that the actual maximum optical flow amplitude will be $r\sqrt{2}$. The whole process is described in the following equation, where $F_{X_b^\star} \in [-r, r]^{N\times W_{LR}\times H_{LR}\times 2}$ denotes the estimated optical flow:

$$F_{X_b^\star \to X_b} = r \cdot tanh\Big(conv2d\big(UNet([X_b^\star, X_b], \Theta_{UNet})\big)\Big). \tag{26}$$

Using the estimated flow $F_{X_b^\star \to X_b}$, $X_b^\star$ can be resampled by means of the grid-based bicubic interpolation ω:

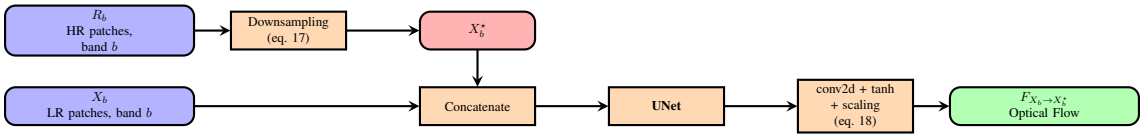

Fig. 19. Workflow of the proposed UNet based optical flow estimation, where b stands for a common spectral band between LR and HR sensors.

$$\widetilde{X}^\star = \omega(X^\star, F_{X_b^\star \to X_b}). \tag{27}$$

Flows can be composed by means of the resampling operator of equation 27. Let $F_{1\to2}$ and $F_{2\to3}$ denote two optical flows, the following equation gives the composition rule:

$$F_{2\to3} \circ F_{1\to2} = F_{1\to2} + \omega(F_{2\to3}, F_{1\to2}). \tag{28}$$

B. Training losses

The optical flow estimation network is pretrained with the same cross-sensor datasets used for the SISR task. Instead of using the Anchor Consistency Loss [24], the pretraining is achieved by optimizing the following loss function:

$$\begin{aligned} L_{flow}(X_b^\star, X_b, F_{rand}) = &\underbrace{L_{real}(X_b^\star, X_b)}_{\text{real term}} + \underbrace{L_{sim}(F_{rand}, X_b^\star, X_b)}_{\text{simulated term}} \\ &+ \underbrace{L_{sym}(X_b^\star, X_b) + L_{self}(X_b^\star)}_{\text{consistency terms}}, \end{aligned} \tag{29}$$

where F_{rand} is a simulated random optical flow. The real term, simulated term, and consistency terms as well as the generation of F_{rand} are described in details in the following subsections. The two first terms have similar order of magnitude, while the two last terms are meant to enforce consistency and thus will have values close to zero throughout the training: there is therefore no need to include weighting factors.

All loss terms make use of the Huber loss, also known as Huber loss, introduced in [90] given by:

$$L_1^{smooth}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise.} \end{cases} \tag{30}$$

The L_1^{smooth} loss behaves like the L_1 loss when differences between predicted and target values are large, which limits the impact of outliers, and like the L_2 when differences are small, which is better for optimization. In the experiments, it has proven to be beneficial over the standard MSE (L_2), especially in early stages of training.

It should also be noted that the remaining of this paper does not rely on this particular method for training, and other methods that reach convergence could be used as well.

1) Real loss term: The real loss term enforces that $F_{X_b^\star \to X_b}$, the flow estimated from $X_b^\star$ to X_b should allow to resample the former onto the latter:

$$L_{real}(X_b^\star, X_b) = L_1^{smooth}\big(X_b - \omega(X_b^\star, F_{X_b^\star \to X_b})\big). \tag{31}$$

2) Simulated loss term: Because the deformation between HR and LR includes the unobserved variation of elevation across the patch, the true optical flow may exhibit strong and abrupt variations that the real loss term is not likely to capture. Inspired by the Anchor Consistency Loss proposed in [24], this loss compares the simulated flow F_{rand} to an indirect estimate using X_b as a pivot image. This allows to use the simulated flow for supervision while still involving the real data.

$$\begin{aligned} L_{sim}(F_{rand}, X_b^\star, X_b) = L_1^{smooth}\big(&F_{X_b \to \omega(X_b^\star, F_{rand})} \\ &- F_{rand} \circ F_{X_b \to X_b^\star}\big). \end{aligned} \tag{32}$$

The random flow F_{rand} is obtained by combining random sinusoidal deformations and directional sigmoid deformations, as given by:

$$F_{rand}(x,y) = \begin{pmatrix} \alpha_x \cos\big(2\pi(\eta_x x + \phi_x)\big) + \beta_x \tanh\big(\frac{d(x,y,a,b)}{\gamma}\big) \\ \alpha_y \cos\big(2\pi(\eta_y y + \phi_y)\big) + \beta_y \tanh\big(\frac{d(x,y,a,b)}{\gamma}\big) \end{pmatrix}, \tag{33}$$

where a, b, γ, η_x, η_y, ϕ_x, ϕ_y, α_x, α_y, β_x and β_y are random simulation parameters drawn for each sample according to the distributions listed in table V, and $d(x, y, a, b)$ represents the signed distance between point (x, y) and sigmoid direction $y = ax + b$ as given by:

$$d(x, y, a, b) = \frac{y - ax - b}{\sqrt{1 + a^2}}. \tag{34}$$

The sinusoidal part of the random flow aims at simulating hills and smooth terrain changes, whereas the directional sigmoid term aims at simulating sharp pinches related to terrain abrupt changes. Those changes might be under-represented in the dataset, and using the simulated term also help ensuring that such changes are correctly learned by the model. Examples of simulated flows are displayed in Fig. 20.

TABLE V
SIMULATION PARAMETERS AND THEIR DISTRIBUTION. w IS THE WIDTH OF THE PATCH AND m IS THE MAXIMUM DEFORMATION VALUE

Parameter	Distribution	Role
a	$\mathcal{U}(-10, 10)$	Slope of sigmoid direction
b	$\mathcal{U}(-a * w, 0)$	Offset of sigmoid direction
γ	$\mathcal{U}(1, 10)$	Width of sigmoid
η_x, η_y	$\mathcal{U}(1/w, 1/(0.9w))$	Sinusoidal period
ϕ_x, ϕ_y	$\mathcal{U}(0, w)$	Sinusoidal phase
α_x, α_y	$\mathcal{U}(0, m)$	Strength of sinusoidal component
β_x, β_y	$\mathcal{U}(0, m)$	Strength of sigmoid component

While simulating such an optical flow brings back the domain gap issue, the experiments show its benefit for training, because it allows to express a loss term that is directly tied to the flow estimation instead of only indirectly measuring it

by comparing resampled images. Intuitively, this simulation process can be seen as a data augmentation strategy that regularize the training and thus leads to better generalization.

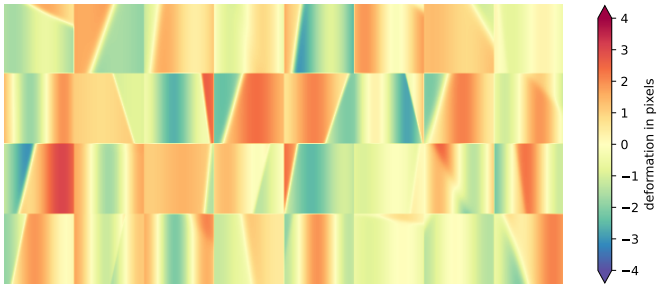

Fig. 20. Examples of simulated flows (x component)

3) Consistency loss terms: In addition, two consistency terms are introduced. The first term constrains F to be a symmetrical operator:

$$L_{sym}(X_b^\star, X_b) = L_1^{smooth}(F_{X_b \to X_b^\star} \circ F_{X_b^\star \to X_b}). \quad (35)$$

The second term forces the self flow to be null:

$$L_{self}(X_b^\star) = L_1^{smooth}(F_{X_b^\star \to X_b^\star}). \quad (36)$$

C. Results

The UNet used in these experiments has 4 levels and 64 features at the first level. The skip connections of the two top levels are removed to favor smoother optical flows. It is trained for 130 epochs with the Adam optimizer [91]. It uses cosine annealing with warm restarts [92], with an initial learning rate of 5e-5 and an initial restart period of 4000 steps. The maximum range parameter r in equation 26 is set to 10 pixels. Value of σ_0 is experimentally set to $\sigma_0 = \sigma(0.4)$ by means of eq. 6. The same datasets as in experiments from section III are used (see table II for datasets description). The red band is usually used in registration studies because it offers the best trade-off between signal to noise ratio and blur. Therefore, the Sentinel-2 red band (B4) is used as parameter b in equation 26, except for dataset **s2v1×4**, where B7 is used instead, since B4 is not available. A *mtf* value of 0.4 is used to generate values for σ_0 in equation 25.

TABLE VI
LOSS VALUES ACHIEVED BY BEST MODELS ON THE TESTING FROM EACH DATASET († THE **WS×2** RESULTS USE THE **WS×4** MODEL).

Dataset	L_{real}	L_{sim}	L_{sym}	L_{self}	**Total loss**
s2v1×2	0.0029	0.0025	0.0003	2.5553e-06	0.0057
s2v1×4	0.0028	0.0025	0.0001	4.9388e-06	0.0055
ws×4	0.1043	0.0462	0.0171	6.8638e-05	0.169
ws×2†	0.1046	0.0417	0.0126	7.4847e-05	0.159

Table VI shows the different loss terms evaluated on the testing set after convergence, for each model. The simulated flows are reconstructed with high accuracy, below 0.05 pixels in all cases. The real loss terms also exhibit very small values, the higher values for Worldstrat dataset being explained by the lower radiometric consistency. The consistency terms have very small values as expected, the higher values of the L_{sym} term on Worldstrat datasets being explained again by the low radiometric consistency, making it more difficult for the network to ensure the symmetric property. Those values demonstrates the ability of the trained model to generalize to unseen patches.

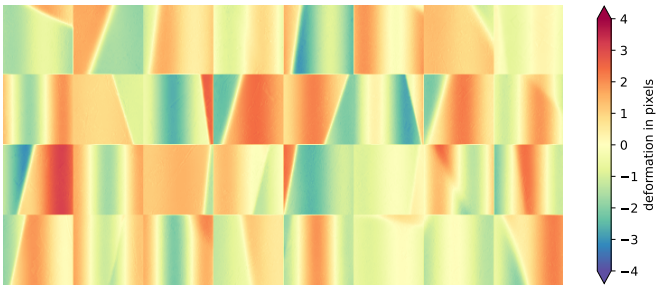

Fig. 21. Reconstruction of simulated flows presented in Fig. 20 using the trained UNet, on the testing set

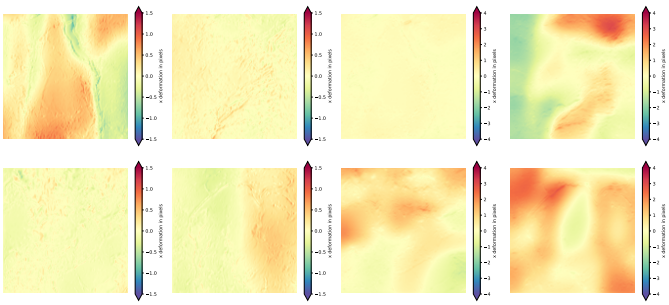

Fig. 22. Estimated flow on patches from Fig. 1, with x flow in the two leftmost columns and y flow in the two rightmost columns.

Fig. 22 shows the two components of the flow estimated for patches of Fig. 1, where strong geometric distortion could be observed. The estimated flow correlates well with the content of the images and underlying variation of relief.

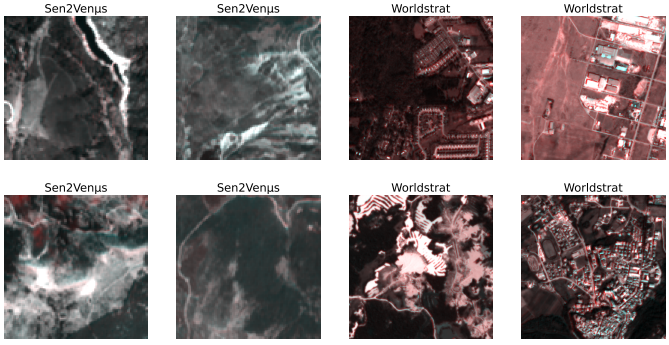

Fig. 23. Same patches and color composition as presented in Fig. 1, with Venμs and Worldstrat bands corrected with the optical flow estimated by UNet.

Fig. 23 shows the same patches as in Fig. 1, where the Venμs band (resp. Spot 6/7) has been resampled according to the estimated flow. One can observe that the geometric discrepancies have been greatly reduced. Residual smooth

discrepancies correspond to uncorrected radiometric discrepancies. This demonstrates that the proposed UNet trained with the proposed strategy is an efficient tool to estimate the optical flow on a SISR dataset, which can be used to measure radiometric discrepancies as well as to try to mitigate them. Note that the overall redness of the Worldstrat patches is caused by radiometric bias.

APPENDIX B
ANALYSING DATASETS WITH PROPOSED METRICS

In this section, we propose to use metrics identified in section II in roder to obtain a prior insight on the level of geometric and radiometric distortions, as well as on the general HR images IQ and the potential of frequency restoration of each dataset.

A. *Sen2Venμs*

The 130k patches of Sen2Venμs are spread across 29 sites, for each of which the Venμs zenith viewing angle is constant across time. Because of the optimization of the orbital resource, most of those 29 sites are acquired with a viewing angle higher than 20°. Since the actual spatial resolution follows the cosine of the zenith angle, sites with higher viewing angle might actually provide less interesting patches in terms of frequency restoration. Fortunately, the PFR FDA metric proposed in section II-C can be leveraged to gain insight on the potential of frequency restoration for each site. Fig. 24 presents a scatter plot comparing the Venμs zenith view angle and the PFR for each site averaged over 10 m bands. It can be observed that most of the highest viewing angle sites have indeed an average PFR below 6%. Most of sites with lower viewing angles have an average PFR between 6% and 12%, with the notable exception of sites K34-AMAZ and FGMANAUS, which exhibit PFR higher than 17%. Those sites are located in tropical forests. The number of patches is limited (less than 2000) with respect to other sites, and all patches are patches over forest areas, which have a specific texture that causes this boost in PFR. This points out that the PFR is not only related to the intrisic sensor parameters, but also to the observed landscape, as a very smooth landscape will exhibit low PFR even with good HR sensors.

In order to analyze the impact of mix of PFR in the training set, two different training sets have been selected, as shown in Fig. 24. The number of patches has been limited to 2000 patches per sites so as to avoid over-representation of some land cover types such as forest. Training set **s2v1** has 20 sites for a total of 35 954 training patches, with a majority of sites with high viewing angles and low PFR, whereas training set **s2v2** has only 8 sites among those with PFR higher than 6%, for a total of 14 615 patches. This allows to investigate the relative importance of having a diverse dataset with respect to selecting good quality samples. A separate testing set is formed with sites MAD-AMBO and ES-IC3XG, which have mid-range viewing angles and PFR, for a total of 3 392 patches.

Metrics proposed in section II-D can be used to analyze the quality of the two Sen2Venμs training sets **s2v1** and **s2v2**.

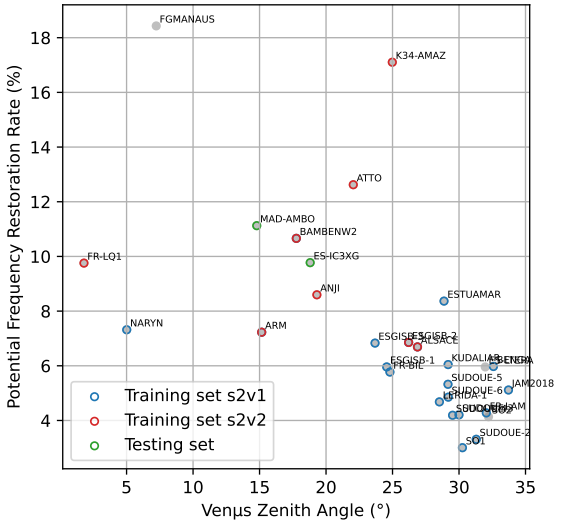

Fig. 24. Average PFR vs. Venμs zenith angle for each of the Sen2Venμs sites.

Table VII shows the PFR, radiometric bias and RMSE, and HR BRISQUE score for all 8 bands of both training sets. As expected, training set **s2v2** exhibits higher PFR values than training set **s2v1** for all bands, with increases of 2 to 3%. It also exhibits lower radiometric distortion on all bands, which can be explained by lower BRDF effects induced by lower viewing angles. The BRISQUE scores are rather similar between both training sets but surprisingly, are a bit better for training set **s2v1**. This may be explained by the ratio between blur and noise being slightly different in both sets.

TABLE VII
POTENTIAL FREQUENCY RESTORATION RATE (PFR), RADIOMETRIC BIAS AND RMSE, AND HR BRISQUE SCORE FOR RAW AND CORRECTED (BETWEEN PARENTS) ESTIMATED FROM 1600 PATCHES OF THE SEN2VENμS TRAINING SETS **S2V1** AND **S2V2**. HERE, THE ↑ (RESP. ↓) INDICATES THAT THE METRIC SHOULD BE MAXIMIZED (RESP. MINIMIZED).

TS	Band	PFR (%) ↑	Bias ± rmse ($1e^{-3}$)↓	BRISQUE ↓
s2v1×2	B2	5.60%	-0.012 ± 6.980	53.78
	B3	5.12%	-0.001 ± 8.018	49.06
	B4	5.88%	0.034 ± 9.755	44.60
	B8	6.79%	0.155 ± 19.572	35.74
s2v2×2	B2	7.51%	-0.060 ± 6.157	60.32
	B3	8.00%	-0.096 ± 6.925	47.75
	B4	8.70%	-0.050 ± 8.459	48.66
	B8	9.72%	-0.295 ± 17.800	37.10
s2v1×4	B5	15.11%	0.225 ± 8.632	40.85
	B6	15.23%	0.683 ± 13.632	37.12
	B7	14.95%	0.763 ± 15.565	36.36
	B8A	16.17%	0.757 ± 16.621	35.01
s2v2×4	B5	18.46%	-0.089 ± 7.502	40.16
	B6	18.73%	0.099 ± 11.585	38.40
	B7	18.64%	0.111 ± 13.441	37.96
	B8A	19.51%	0.082 ± 14.504	37.24

Table VIII shows the mean and standard deviation of the optical flow amplitude estimated with trained models on **s2v1** and **s2v2**, expressed in HR pixels. It can be observed that the level of geometric distortion is similar in all datasets,

with a mean amplitude around 0.6 HR (5 m) pixels and a standard deviation of the same amount. Geometric Distortion is similar in both training sets and coherent between 10 m and 20 m bands. Since geometric distortion is yielded by both high viewing angles and relief, and training set **s2v2** still has large viewing angles for most sites, this result is expected.

TABLE VIII
MEAN AND STANDARD DEVIATION OF SPATIAL DISTORTION AMPLITUDE MEASURED ON 1600 PATCHES OF THE SEN2VENµS TRAINING SETS **S2V1**, **S2V2**.

TS	Band	Mean (hr pix.)	Std. dev. (hr pix.)
s2v1×2	B4	0.578	0.667
s2v1×4	B7	0.603	0.603
s2v2×2	B4	0.685	0.655
s2v2×4	B7	0.621	0.599

B. Worldstrat

The Worldstrat dataset is oriented toward MISR, and offers several Sentinel-2 patches from different dates for a single HR date. In order to extract a SISR oriented dataset from it and minimize discrepencies related to temporal differences between HR and LR patches, only pairs of patches for which the acquisition date absolute difference is lower than 10 days have been selected. For POIs where more than one pair meets this criterion, only the pair with the closest dates is retained. Sentinel-2 10-meter bands B2, B3, B4 and B8 have been extracted from the L2A product. Since there are slight variations in patch size and HR patches have no geo-referencing, a 128x128 pixel patch has been extracted from the center of the Sentinel-2 L2A patch. The corresponding pan-sharpened HR patch, whose resolution is 1.6 meter, is first downsampled to 2.5 meters and 5 meters by means of equation 8, with *mtf* value set to 0.1. Then a 512×512 or 256×256 pixels patch is extracted. This process yields a ×4 dataset called **ws×4** (10 m $\rightarrow$ 2.5 m) and a ×2 dataset (10 m $\rightarrow$ 5 m) called **ws×2**. Training, validation and test split proposed in Worldstrat have been retained. Additionally, since no cloud screening has been performed on the dataset and HR patches have no cloud mask, the linear correlation between amplitude of HR and LR patches has been computed and pairs of patches with a correlation below 0.2 have been discarded. This yields a rather small dataset of 2 144 training patches, 260 validation patches and 274 testing patches.

Table IX shows the analysis of the Worldstrat dataset values for PFR, LR radiometric bias and RMSE, and HR BRISQUE score. With respect to the Sen2Venµs dataset, Worldstrat exhibits higher PFR with lower HR BRISQUE scores. The 2.5 m dataset yields a ×4 up-sampling factor that can explain higher PFR, but even the 5 m dataset, which has the same up-sampling factor as Sen2Venµs for 10 m bands, has PFR as large as twice those observed on Sen2Venµs. This points out that the Worldstrat HR images generated by the process described above are much sharper and clean than the Venµs images, and of general higher quality, as pointed out by the BRISQUE score. On the other hand, because of temporal changes and general consistency between sensors, radiometric distortion is also much larger on Worldstrat, with non null biases and larger RMSE values.

TABLE IX
POTENTIAL FREQUENCY RESTORATION RATE (PFR), RADIOMETRIC BIAS AND RMSE, AND HR BRISQUE SCORE ESTIMATED FROM 160 PATCHES OF THE WORLDSTRAT TRAINING SET. HERE, THE ↑ (RESP. ↓) INDICATES THAT THE METRIC SHOULD BE MAXIMIZED (RESP. MINIMIZED).

Dataset	Band	PFR (%) ↑	Bias ± rmse ($1e^{-3}$)↓	BRISQUE ↓
ws×4	B2	22.65%	-21.946 ± 87.776	43.10
	B3	22.60%	-32.887 ± 83.459	38.87
	B4	22.04%	-37.619 ± 83.835	39.96
	B8	26.01%	-81.538 ± 94.546	32.33
ws×2	B2	18.15%	-21.403 ± 88.224	33.13
	B3	18.40%	-32.457 ± 84.272	28.65
	B4	18.31%	-37.438 ± 84.511	30.02
	B8	22.42%	-80.350 ± 95.298	17.35

Table X shows the mean and standard deviation of the optical flow amplitude estimated with trained models on **ws×4** and **ws×2** training sets, expressed in HR pixels. It can be observed that Worldstrat based datasets exhibit higher geometric distortion than Sen2Venµs datasets, with more than 6 (2.5 m) pixels of mean amplitude and almost 5 (2.5 m) pixels of standard deviation.

TABLE X
MEAN AND STANDARD DEVIATION OF SPATIAL DISTORTION AMPLITUDE MEASURED ON 1600 PATCHES OF THE **WS×2** AND **WS×4** DATASETS.

TS	Band	Mean (hr pix.)	Std. dev. (hr pix.)
ws×4	B4	6.274	4.789
ws×2	B4	3.157	2.389

APPENDIX C
IMPACT OF PROPOSED STRATEGY ON DATASET QUALITY

TABLE XI
IMPACT OF THE PROPOSED STRATEGY ON THE QUALITY OF THE DATASET IN TERMS OF GEOMETRIC DISTORTION, MEASURED ON BAND B4 (EXCEPT FOR THE **S2V1×4** AND **S2V2×4** CASES, WHERE BAND B7 IS USED), ESTIMATED ON 1600 PATCHES FROM THE TRAINING SETS. SCORES FOR CORRECTED PATCHES ARE DISPLAYED FIRST, AND VARIATION IS DISPLAYED BETWEEN BRACKETS. VARIATION IS EXPRESSED AS PERCENT OF THE UNCORRECTED VALUE.

Dataset	Band	Mean (hr pix.)	Std. dev. (hr pix.)
ws×4	B4	0.197 (-96.86%)	0.426 (-91.10%)
ws×2	B4	0.102 (-96.77%)	0.226 (-90.54%)
s2v2×2	B4	0.073 (-89.34%)	0.077 (-88.46%)
s2v1×2	B4	0.107 (-82.77%)	0.071 (-89.16%)
s2v2×4	B7	0.075 (-87.02%)	0.092 (-84.64%)
s2v1×4	B7	0.123 (-79.60%)	0.107 (-82.26%)

Though only used during training, the proposed strategy has an impact on the target image quality. Table XII shows the PFR, LR radiometric bias and RMSE, and HR BRISQUE score of corrected sample training patches $\underset{\sim}{R}^{\star}$, for band B4 (and B7 for 20 m bands), whereas table XI shows Geometric Distortion of the same data. The radiometric correction is very effective especially for WorldStrat derived training sets, where there is almost no bias remaining and LR RMSE has been reduced respectively by 80%. Gains for Sen2Venµs training sets are less important, which is expected because of the

TABLE XII
IMPACT OF THE PROPOSED STRATEGY ON THE QUALITY OF THE DATASET IN TERMS OF PFR, LR RADIOMETRIC BIAS AND RMSE, AND BRISQUE SCORE, MEASURED ON BAND B4 (EXCEPT FOR THE **S2V1×4** AND **S2V2×4** 5 M CASES, WHERE BAND B7 IS USED), ESTIMATED ON 1600 PATCHES FROM THE TRAINING SETS. SCORES FOR CORRECTED PATCHES ARE DISPLAYED FIRST, AND VARIATION IS DISPLAYED BETWEEN BRACKETS. VARIATION IS EXPRESSED AS A PLAIN DIFFERENCE, EXCEPT FOR BIAS AND RMSE, WHERE VARIATION IS EXPRESSED IN PERCENT OF THE UNCORRECTED ABSOLUTE VALUE.HERE, THE ↑ (RESP. ↓) INDICATES THAT THE METRIC SHOULD BE MAXIMIZED (RESP. MINIMIZED).

Dataset	**Band**	**PFR (%) ↑**	**Bias ($1e^{-3}$)**	**rmse ($1e^{-3}$)↓**	**BRISQUE ↓**
ws×4	B4	17.09% (-4.95)	-37.619 (-98.21%)	16.644 (-80.15%)	40.49 (+0.53)
ws×2	B4	12.27% (-6.13)	-32.457 (-97.98%)	16.993 (-79.84%)	31.69 (+3.04)
s2v2×2	B4	7.85% (-0.85)	-0.050 (-100.00%)	5.264 (-37.77%)	49.80 (+1.14)
s2v1×2	B4	5.50% (-0.38)	0.034 (-91.18%)	6.089 (-37.58%)	45.60 (+1.00)
s2v2×4	B7	17.35% (-1.29)	0.111 (-95.50%)	7.910 (-41.15%)	38.26 (+0.30)
s2v1×4	B7	14.10% (-0.85)	0.763 (-95.81%)	8.801 (-43.46%)	36.64 (+0.28)

higher radiometric consistency of the dataset. Nevertheless, LR RMSE is consistently lower by around 40% in all cases. Geometric correction is also very effective, allowing to reduce the measured average distortion to approximately 0.1 pixel, with correspond to a decrease 80% for Sen2Venμs datasets and 96% for WorldStrat datasets. It should be reminded that raw Worldstrat data exhibited several pixels of average distortion. Standard deviation of measured distortion is reduced by at least 80% in all cases, which indicates that the network effectively correct for local geometric distortions. Those corrections come at the price of a decrease of the PFR of around 1 to 6% depending on the case, Worldstrat datasets being the worst case. However, this decrease seems to have very limited effect on the HR BRISQUE score, which suggests that corrected target images remain of high quality. A decrease of the PFR is expected and is mainly caused by the radiometric correction of equation 21, which consist of partially re-introducing the spatial frequency content of the input data into the reference data. The Low Pass Filtering performed by ϕ_{σ_2} limits this effect but cannot completely eliminate it.

APPENDIX D
TRAINING HYPER-PARAMETERS FOR EXPERIMENTS

Both the generator and the discriminator are trained with the Adam optimizer [91]. The initial learning rate is set to 2e-4 for the generator, and the cosine annealing with warm restarts [92] strategy is used, with a initial period of 1500 steps and a multiplicative factor of 2. The initial learning rate of discriminator is set to 1e-4, and the cosine annealing with warm restarts [92] strategy is used, with an initial period of 6000 steps and a multiplicative factor of 2. Batch size is set to 16 in all experiments. It should be noted that adversarial training only starts after one epoch. This allows for the generator to already yield plausible results when the discriminator enters the optimization process, and favor stability in the early stages of training. The training runs for 125 epochs of 1340 steps each for the Sen2Venμs datasets, and 500 epochs of 130 steps each for Worldstrat datasets. This lower number of steps is explained by the smaller size of the Worldstrat datasets, which starts to overfit earlier than Sen2Venμs. In all experiments, the model parameters of the last training step is used, since the use of GAN loss makes model selection based the total loss value hazardous: a weak discriminator may lead to local minima with poor generator quality. Moreover, section II demonstrated that a single metric can not account for all aspects of IQ, and thus selecting the best model according to a single metric may bias the study toward this metric. Nevertheless, both LPIPS and L_2 are monitored to ensure that training does not diverge. All codes use Pytorch [93] and runs on NDVIA GPU A100. Training times range between 12 and 20 hours depending on the experiment.