Neurocomputing 530 (2023) 125–138

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

W-NetPan: Double-U network for inter-sensor self-supervised pan-sharpening

Ruben Fernandez-Beltran [a,*], Rafael Fernandez [b], Jian Kang [c], Filiberto Pla [b]

[a] Department of Computer Science and Systems, University of Murcia, 30100 Murcia, Spain
[b] Institute of New Imaging Technologies, University Jaume I, 12071 Castellón de la Plana, Spain
[c] School of Electronic and Information Engineering, Soochow University, Suzhou 215006, China

ARTICLE INFO

Article history:
Received 5 November 2021
Revised 1 February 2023
Accepted 3 February 2023
Available online 8 February 2023
Communicated by Zidong Wang

Keywords:
Pan-sharpening
Convolutional neural networks
Inter-sensor
Multi-modal
Remote sensing

ABSTRACT

The increasing availability of remote sensing data allows dealing with spatial-spectral limitations by means of pan-sharpening methods. However, fusing inter-sensor data poses important challenges, in terms of resolution differences, sensor-dependent deformations and ground-truth data availability, that demand more accurate pan-sharpening solutions. In response, this paper proposes a novel deep learning-based pan-sharpening model which is termed as the double-U network for self-supervised pan-sharpening (W-NetPan). In more details, the proposed architecture adopts an innovative W-shape that integrates two U-Net segments which sequentially work for spatially matching and fusing inter-sensor multi-modal data. In this way, a synergic effect is produced where the first segment resolves inter-sensor deviations while stimulating the second one to achieve a more accurate data fusion. Additionally, a joint loss formulation is proposed for effectively training the proposed model without external data supervision. The experimental comparison, conducted over four coupled Sentinel-2 and Sentinel-3 datasets, reveals the advantages of W-NetPan with respect to several of the most important state-of-the-art pan-sharpening methods available in the literature. The codes related to this paper will be available at https://github.com/rufernan/WNetPan.

1. Introduction

With the development of image acquisition technologies, spatial resolution plays a fundamental role in many important remote sensing (RS) applications, including land-cover mapping [1–3], environmental management [4–6], scenery recognition [7–9] and material analysis [10–12]. Nonetheless, designing multi-spectral (MS) instruments with a very high spatial resolution often becomes an infeasible task due to the diffraction effects of the incoming light as well as the high cost and complexity of this type of technology [13]. Consequently, many of the existing Earth Observation (EO) programmes, like Copernicus [14], try to relieve these limitations by including multiple specialized satellites that cover particular spatial-spectral needs. Within Copernicus, Sentinel-2 (S2) and Sentinel-3 (S3) constellations exemplify this trend. On the one hand, S2 satellites [15] carry the Multi-Spectral Instrument (MSI) which captures 13 spectral bands (B01-B12) in the wavelength range 443–2190 nm, using an spatial resolution between 10 m and 60 m. On the other hand, S3 counterparts [16] incorporate the Ocean and Land Colour Instrument (OLCI) which provides 21 bands (Oa01-Oa21) in the 390–1040 nm spectral range, with a fix spatial resolution of 300 m. In this scenario, MSI images generally become more suitable for land-cover characterization tasks, whereas OLCI products are more focused on the spectral features of oceans, inland waterways and coastal areas due to their limited spatial resolution [17]. Nevertheless, the unprecedented availability of inter-sensor Sentinel data brings us the opportunity of dealing with this kind of constraints from image processing and machine learning-based perspectives.

Over the past decade, extensive efforts have been done to spatially enhance RS images by means of the so-called pan-sharpening methods [18]. Specifically, pan-sharpening is a field, which could be considered a particular case of image super-resolution [19], where two given high-resolution (HR) panchromatic (PAN) and low-resolution (LR) MS images are combined to generate a HR version of the MS data. In this way, the final target of pan-sharpening consists in fusing two images that cover the same area over the Earth surface into a joint representation which gathers the finest spatial-spectral details of the input data.

From traditional algorithms [20], to more recent deep learning (DL) models [21], a wide variety of pan-sharpening methods have

* Corresponding author.

https://doi.org/10.1016/j.neucom.2023.02.002
0925-2312/

been proposed in the literature. Among the traditional group, one of the most popular trends is component substitution (CS). In CS, the spatial component of MS is replaced with its corresponding HR counterpart, which is extracted from PAN using a particular transformation model, such as, principal component analysis (PCA) [22] or intensity-hue-saturation (IHS) [23]. Another popular trend is multi-resolution analysis (MRA), which pursues to progressively inject the spatial information of PAN into the MS domain following a multi-resolution decomposition scheme, as in [24,25].

More recently, DL has attracted significant attention due to the excellent capabilities of convolutional neural networks (CNN) to extract highly relevant features from optical data. In more details, DL-based pan-sharpening methods are typically based on training CNN models to map the input data (i.e., HR PAN and LR MS) onto the target HR MS domain. In this way, different architectures, such as [26–28], have shown prominent results while setting the current state-of-the-art performance in the field.

Despite the positive results achieved by these and other relevant methods, there still are some essential open-ended problems when it comes to highly heterogeneous inter-sensor scenarios, such as S2/S3. In this case, fusing operational MSI and OLCI data poses additional challenges in terms of resolution differences, sensor-dependent deformations and missing ground-truth data, that are generally beyond pan-sharpening algorithms and need to be jointly addressed to provide more accurate solutions. Let us illustrate these problems by means of a visual example. Fig. 1 displays a sample S2 MSI image (a), its S3 OLCI counterpart (b) and their corresponding overlap (c) where OLCI pixels are colored in green and MSI pixels in purple. Although both S2/S3 image products are logically corrected to fit the same area over the Earth surface, the substantially bigger size of OLCI pixels does not allow accurately fitting MSI shapes since it has a much better resolution. In this situation, the deep spectral features uncovered from OLCI are always expected to contain some spatial distortions that can negatively affect the corresponding pan-sharpened result, specially with very deep networks. In response, this paper presents a new pan-sharpening network focused on three key aspects that take on special relevance with inter-sensor Sentinel data:

- First, the resolution differences between S2 and S3 may produce significant receptive field changes as the network depth increases. This effect may logically affect the final spatial quality of the pan-sharpened output.
- Second, the inter-platform nature of the data may generate the propagation of geolocation errors across sensors. This fact may eventually result in important spatial deviations at the output HR scale.
- Third, the use of operational S2/S3 imagery restricts the availability of ground-truth data for training. This situation may constrain pan-sharpening learning protocols to a reduced-reference strategy where inter-sensor resolution problems can be magnified.To cope with these challenges, we propose a novel DL-based pan-sharpening model, named the double-U network for self-supervised pan-sharpening (W-NetPan), which has been specifically designed for effectively managing data from different platforms, such as S2/S3. Unlike other methods available in the literature, the proposed architecture defines an innovative W-shape that integrates two sequential U-Net segments which simultaneously work for spatially matching and fusing inter-sensor data. Additionally, a new loss formulation is proposed to train the model from an end-to-end perspective without using neither external data supervision nor reduced-reference protocols. First, the considered U-Net backbone shapes allows our W-NetPan model to propagate receptive fields across layers with the objective of alleviating spatial resolution differences. Second, the two proposed U-Net segments aim at resolving inter-sensor spatial deviations while fusing the data, which generates a synergic effect where the first U-Net stimulates the second to find a more optimal solution. Third, the proposed loss has been formulated to only require operational input data by integrating three terms that work at different resolution levels: inter-sensor spatial matching, MS spectral consistency and PAN spatial consistency. In short, the main contributions of this work can be summarized as follows:

1. A new DL architecture (W-NetPan) is proposed for pan-sharpening inter-sensor data.
2. A novel joint loss formulation is defined for training the proposed network under a self-supervised scheme.
3. The performance of multiple state-of-the-art pan-sharpening methods is analyzed when fusing S3 OLCI and S2 MSI operational data.
4. The higher suitability of the proposed model is proven to resolve inter-sensor Sentinel data.The rest of this paper is organized as follows. section 2 introduces some related works while describing their main limitations with Sentinel data. section 3 defines the proposed pan-sharpening model, including its novel network topology and its joint loss formulation. section 4 pre-

Fig. 1. Spatial resolution differences between a sample S2 MSI product (a), its S3 OLCI counterpart (b) and their corresponding overlap (c). Note that in (c), OLCI pixels are colored in green and MSI pixels in purple. As it is possible to see, the substantially bigger size of OLCI pixels does not allow accurately fitting MSI shapes which can logically affect the extracted spectral features and distort the fused result.

sents the experiments, considering several datasets and state-of-the-art methods, and section 5 discusses the corresponding results. Finally, section 6 concludes the paper with some future research lines.

2. Related Work

Depending on their nature, pan-sharpening algorithms can be grossly divided into four major groups [20,29,21]: (1) component-substitution (CS), (2) multi-resolution analysis (MRA), (3) sparse factorization (SF) and (4) deep learning-based (DL). In this case, CS and MRA are typically identified as traditional methods, whereas SF and, above all, DL certainly represent more recent approaches. Let us provide a quick overview of these categories along the following lines.

In general, CS methods start from the assumption that LR MS can be separated into spatial and spectral components following a particular transformation model. Then, PAN can be a good substitution for the spatial component to generate HR MS via the corresponding inverse projection. Logically, each CS-based pan-sharpening method defines such transformation in a different way. For instance, Gillespie et al. [30] proposed using the Brovey transformation, which simply multiplies each re-sampled MS band by the intensity ratio between PAN and MS. In [22,31], the authors made use of the PCA for isolating the spatial information of MS into the first principal component. Analogously, Carper et al. [23] took advantage of the IHS transformation for extracting the spatial information as the intensity channel. Aiazzi et al. [32] used a modified version of the Gram-Schmidt (GS) orthogonalization for decorrelating MS bands using a simulated PAN, which is finally replaced by the actual PAN. In spite of their simplicity, the difficulties in completely isolating the spatial information from MS generally make CS-based methods prone to introduce spectral distortions.

To relieve these problems, MRA techniques opt to perform the spatial filtering into the HR domain. In this way, the high-frequency details are extracted from PAN, and then they can be injected into each interpolated MS band to produce the final HR MS result. For instance, Liu et al. [24] presented the smoothing filter-based intensity modulation (SFIM), which computes the difference between PAN and its low-pass filtered version for applying an additive injection of such spatial differences. In [33], the authors proposed using a discrete wavelet transform over PAN for obtaining the spatial details following a multi-resolution decomposition scheme. To further improve the filtering process, other authors exploited the modulation transfer function (MTF) of the instrument. As in [25], where Aiazzi et al. adopted the MTF of the sensor for building a generalized Laplacian pyramid (GLP), with the possibility of considering several injection models, such as, uniform weights (MTF-GLP) or high-pass modulation (MTF-GLP-HPM). Despite their advantages, MRA-based methods may still have important spatial limitations due to the own low-pass filtering process. In response, alternative factorization pan-sharpening mechanisms have also been developed in the literature. For example, it is the case of Yokoya et al. [34] who defined the coupled non-negative matrix factorization (CNMF) model. Specifically, CNMF factorizes the input data into their corresponding spectral signatures and fractional abundances. Then, MS signatures and PAN abundances are combined to obtain the target HR MS output.

Despite the remarkable performance achieved by these paradigms, DL is certainly one of the most emerging pan-sharpening trends due to its enormous success in many different related fields, e.g. [35–38]. In particular, the rationale behind DL-based pan-sharpening consists in learning a mapping function from the input MS/PAN data to the corresponding HR MS domain, in a similar fashion to super-resolution techniques [39]. For instance, it is the case of Masi et al. [26] who presented the pan-sharpening convolutional neural network (PNN). In PNN, the input MS data is first interpolated to the target resolution and concatenated to PAN. Then, these data are projected onto the target HR MS space using three convolutional layers. Extending this idea, Yang et al. [27] proposed the PanNet model which takes advantage of residual connections for propagating the input spectral information to the pan-sharpened output. In [40], Scarpa et al. further fine-tune the PNN by means of pre-training and a target-adaptive tuning phase. In addition to these architectures, other authors suggest alternative network designs. For example, it is the case of Xu et al. [28] who created the gradient projection pan-sharpening neural network (GPPNN). In details, GPPNN formulates two generative models, one for PAN and the other for MS, which are both regularized by deep image priors [41]. In this scenario, the gradient projection method [42] is used for obtaining the corresponding update rules. Then, two neural blocks are designed to embed these generative models and their updates into a CNN which is eventually optimized to produce the final pan-sharpened output. In [43], Uezato et al. also defined the guided deep decoder (GDD) which takes advantage of a two-stream network. On the one hand, an encoder-decoder CNN is used for uncovering multi-scale features from the input MS and PAN data. On the other hand, a generative decoder, guided by the previous network, is employed to produce the fused result. In another recent work, Ozcelik et al. [44] proposed the PanColorGAN model which exploits generative adversarial networks (GANs) for self-supervised pan-sharpening. In contrast to other approaches, PanColorGAN deals with the data fusion problem from a colorization perspective, where a U-Net [45] with color injection is used as generator and a conditional patch-GAN [46] as discriminator.

Without any doubt, DL models set a new path for learning pan-sharpening projections in a very effective way. However, the task of fusing highly heterogeneous inter-sensor data still raises some important challenges to pan-sharpening [18,21]. In general, DL-based methods start by concatenating LR MS and HR PAN features to project the input data onto the target HR MS domain. Although both LR MS and HR PAN images are logically captured to cover the same area over the Earth surface, their spatial resolution differences make not possible to perfectly match both sensors at a pixel-level since HR PAN have a better resolution. In this scenario, the spatial deviations inherent to the resolution change between LR and HR pixels can negatively affect the output result. Unlike other DL-based solutions that learn a direct pan-sharpening projection from the input data [26–28,44], this paper proposes a novel double-U topology to dynamically alleviate the spatial deviations of the input as the refined features are projected to the target HR MS domain. Specifically, this is the case of operational S2/S3 imaging data products where spatial differences, sensor-dependent deformations and the lack of actual ground-truth data can certainly affect the performance of the existing methods. The significant spatial resolution differences between S2 and S3 may produce a substantial widening of the convolutional receptive fields, which may eventually result in a blurring effect. Besides, S2/S3 inter-platform errors could also be propagated across sensors generating additional deviations in the output. What is more, the lack of actual ground-truth S2/S3 fused data may also limit the training protocol and the precision DL-based approaches. To address all these challenges, this article presents the W-NetPan model.

3. Methodology

This section describes the proposed W-NetPan model which has been specially designed for conducting self-supervised pan-sharpening from an inter-sensor perspective. In the following lines,

we formulate the pan-sharpening problem across different sensors based on the proposed CNN-based model while defining the considered loss functions and other implementation details. Nonetheless, let us first describe the notation considered in this work. Let $\mathbf{I}_{\text{MS}} \in \mathbb{R}^{(X_{\text{L}} \times Y_{\text{L}} \times B)}$ be a LR MS image with B bands and a $(X_{\text{L}} \times Y_{\text{L}})$ spatial size. Let $\mathbf{I}_{\text{PAN}} \in \mathbb{R}^{(X_{\text{H}} \times Y_{\text{H}} \times 1)}$ represent a HR panchromatic image with a spatial size of $(X_{\text{H}} \times Y_{\text{H}})$ that covers the same extent as $\mathbf{I}_{\text{MS}}$ over the Earth surface. Let R identify the scaling ratio between both sensors such that $X_{\text{H}} = X_{\text{L}}R$ and $Y_{\text{H}} = Y_{\text{L}}R$. In this sense, it is important to highlight that, with respect to the RS field, we assume in this work a Level-4 processing data nature, that is, both $\mathbf{I}_{\text{MS}}$ and $\mathbf{I}_{\text{PAN}}$ images are acquired by different instruments and platforms which logically introduce different error types and tolerances in each case. Additionally, let $\mathbf{I}_{\text{HR}} \in \mathbb{R}^{(X_{\text{H}} \times Y_{\text{H}} \times B)}$ be the corresponding HR ground-truth image which contains the spatial resolution of $\mathbf{I}_{\text{PAN}}$ and the spectral information of $\mathbf{I}_{\text{MS}}$.

In this context, the proposed network pursues to approximate a function of the form $\mathscr{F}(\mathbf{I}_{\text{MS}}, \mathbf{I}_{\text{PAN}}) = \mathbf{I}_{\text{HR}}$ following a self-supervised fashion, that is, without involving any ground-truth data. To achieve this goal, we define the W-NetPan architecture which is able to take advantage of the higher resolution of $\mathbf{I}_{\text{PAN}}$ for relieving inter-sensor deformations while projecting the input data onto the target HR MS domain. Table 1 provides a brief summary of the main notation considered in this section.

3.1. W-NetPan: Double-U Network for Pan-sharpening

Deeper CNNs can certainly extract higher level image features that may offer a better visual understanding for pan-sharpening. However, the deeper the network the higher the corresponding receptive fields which may eventually cause counterproductive effects in the spatial details of the output results [21]. In response, we adopt an U-shaped backbone architecture for propagating receptive fields across layers. Specifically, the U-Net architecture [45] is typically made of a symmetric encoder/decoder path where feature maps are subsequently down-sampled until a bottleneck layer (at the bottom of the U-Net) from which sequential up-samplings and concatenations are applied in order to propagate context information to higher resolution layers. In this way, feature maps corresponding to different scales can simultaneously be used to enhance spatial accuracy and abstraction ability in pan-sharpening [47]. Nonetheless, the standard U-Net architecture still has some important constraints when it comes to inter-sensor self-supervised pan-sharpening. On the one hand, the inter-sensor facet of the problem may introduce spatial deviations and uncertainties when fusing the information coming from two rather different instruments. Note that pan-sharpening techniques only make sense when there are important spatial differences between $\mathbf{I}_{\text{MS}}$ and $\mathbf{I}_{\text{PAN}}$ images, and in this circumstances, even small geolocation errors in $\mathbf{I}_{\text{MS}}$ may produce important deviations in the HR domain [48]. On the other hand, the lack of actual ground-truth HR data makes necessary to use a reduced-reference protocol for training CNN-based models, which may eventually exacerbate the inter-sensor resolution problem while substantially reducing the available training data.

Table 1
Summary of the considered notation.

Term	*Description*
$\mathbf{I}_{\text{MS}}$	Input LR MS image
$\mathbf{I}_{\text{PAN}}$	Input HR PAN image
$\mathbf{I}_{\text{HR}}$	Output HR MS image
$\widetilde{\mathbf{I}}_{\text{MS}}$	Up-sampled MS image (generated from $\mathbf{I}_{\text{MS}}$)
$\widetilde{\mathbf{I}}_{\text{PAN}}$	Simulated PAN image (generated from $\widetilde{\mathbf{I}}_{\text{MS}}$)
$\widehat{\mathbf{I}}_{\text{MS}}$	Spatially transformed MS image (generated from $\widetilde{\mathbf{I}}_{\text{MS}}$)
$\widehat{\mathbf{I}}_{\text{PAN}}$	Spatially transformed PAN image (generated from $\widetilde{\mathbf{I}}_{\text{PAN}}$)
$\mathbf{I}^{*}_{\text{MS}}$	Simulated output MS image (generated from $\mathbf{I}_{\text{HR}}$)
$\mathbf{I}^{*}_{\text{MS}}$	Simulated output PAN image (generated from $\mathbf{I}_{\text{HR}}$)
Φ	Deformation field between $\mathbf{I}_{\text{PAN}}$ and $\widetilde{\mathbf{I}}_{\text{PAN}}$
S_1	First U-Net segment of the proposed architecture
S_2	Second U-Net segment of the proposed architecture
$\mathscr{L}_1$	Spatial matching loss
$\mathscr{L}_2$	Spectral consistency loss
$\mathscr{L}_3$	Spatial consistency loss

To overcome these challenges, we propose the W-NetPan architecture which is depicted in Fig. 2. Note that, the proposed model adopts an innovative W-shape that integrates two sequential U-Net segments which jointly work for spatially matching and fusing inter-sensor data. Now, let us provide a detailed description of each one of the considered segments, which are identified by S_1 and S_2 in the figure. First, S_1 aims at matching the spatial information coming from the two input optical sources that operate at different resolutions. For this purpose, three different building blocks are used in this segment: S_1^{h} (head), S_1^{b} (body) and S_1^{t} (tail). The objective of S_1^{h} is based processing the input data to generate a uniform data cube focused on the spatial information. Hence, it contains the following layers: (1) up-sampling (Up), (2) pooling (Pool) and (3) concatenation (Cat). In (1), we employ a regular up-sampling layer with a bi-cubic filter for spatially up-scaling $\mathbf{I}_{\text{MS}}$ to the target resolution (ratio $R\times$) as $\widetilde{\mathbf{I}}_{\text{MS}}$. Then, an spectral average pooling is used to simulate its panchromatic counterpart ($\widetilde{\mathbf{I}}_{\text{PAN}}$) which is finally stacked onto the input $\mathbf{I}_{\text{PAN}}$ image. The body block (S_1^{b}) pursues to project these data on a two-dimensional deformation field which describes the vertical and horizontal displacements of each pixel in $\widetilde{\mathbf{I}}_{\text{PAN}}$ (simulated PAN generated from $\widetilde{\mathbf{I}}_{\text{MS}}$) with respect to $\mathbf{I}_{\text{PAN}}$ (input PAN image). Specifically, it is made of a standard U-Net with four encoding/decoding layers and two final convolutions.

Fig. 3 shows the considered U-Net backbone architecture, where K_{IN} represents the number of input bands, $K_{1,2}$ are the number of 3×3 filters, $S_{1,2}$ are the corresponding pixel-strides and K_{OUT} is the number of output bands. In the case of S_1^{b}, we set $K_{\text{IN}} = 2, K_1 = 16, K_2 = 32, S_1 = 2, S_2 = 1$ and $K_{\text{OUT}} = 2$ to generate the corresponding deformation field Φ. Regarding the tail block (S_1^{t}), it is directed to apply the estimated deformation over $\widetilde{\mathbf{I}}_{\text{MS}}$ and $\widetilde{\mathbf{I}}_{\text{PAN}}$ (as $\widehat{\mathbf{I}}_{\text{MS}}$ and $\widehat{\mathbf{I}}_{\text{PAN}}$, respectively) to be used in the subsequent stages. To achieve this goal, we build a differentiable spatial transformation block based on the so-called spatial transformer networks [49]. Specifically, for each i pixel in $\widetilde{\mathbf{I}}_{MS}$, we compute its corresponding sub-pixel location in $\widehat{\mathbf{I}}_{MS}$ as $j = i + \Phi(i)$. Since image locations are logically only defined at integer positions, we linearly interpolate each transformed location using its eight-pixel neighborhood as Eq. 1 shows. In this expression, $\mathscr{I}(j)$ represents the pixel neighbors of j and d iterates over width (X) and height (Y) spatial dimensions, being j_d and q_d the coordinates on each dimension for j and q_d, respectively.

$$\widehat{\mathbf{I}}_{\text{MS}}(i) = \sum_{q \in \mathscr{I}(j)} \widetilde{\mathbf{I}}_{\text{MS}}(q) \prod_{d \in \{X,Y\}} (1 - |j_d - q_d|) \tag{1}$$

As it is possible to see, the value of $\widehat{\mathbf{I}}_{\text{MS}}$ at a given pixel position i can be obtained by means of the following process. First, the transformed sub-pixel location j is calculated as $j = i + \Phi(i)$. Then, for each pixel q within the 8-pixel neighborhood of j, the value of $\widetilde{\mathbf{I}}_{\text{MS}}$ at q (i.e., $\widetilde{\mathbf{I}}_{\text{MS}}(q)$) is multiplied by its corresponding bi-linear re-sampling term, which is based on the distance between j and q, considering X and Y spatial dimensions. Finally, the weighted sum over the considered neighborhood produces the final re-sampled value

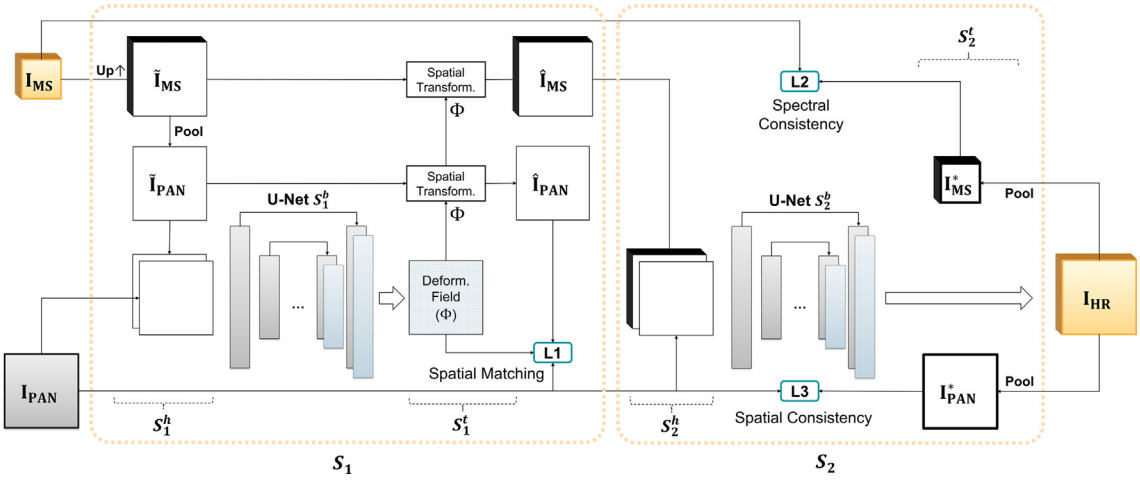

Fig. 2. W-NetPan: proposed double-U network for inter-sensor self-supervised pan-sharpening.

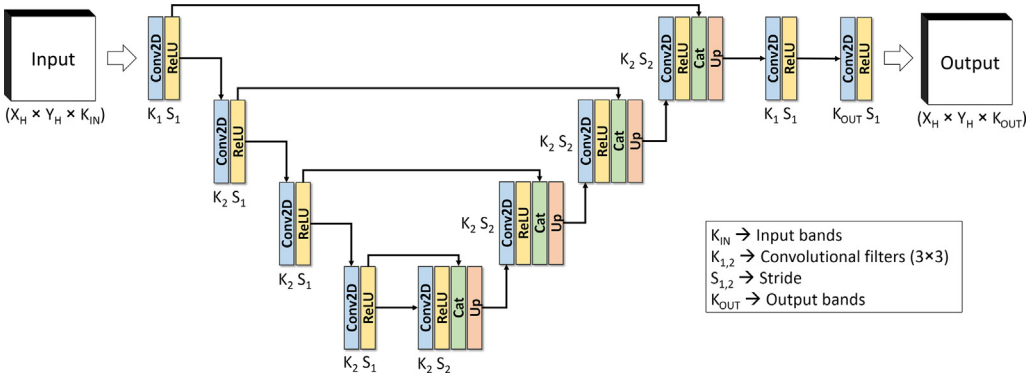

Fig. 3. U-Net backbone architecture considered within the proposed W-NetPan model.

$\widehat{\mathbf{I}}_{\text{MS}}(i)$. Note that this process corresponds to the bi-linear interpolation that can be found in [49]. In a similar fashion, $\widehat{\mathbf{I}}_{\text{PAN}}$ can be obtained from $\widetilde{\mathbf{I}}_{\text{PAN}}$ and Φ according to Eq. 2.

$$\widehat{\mathbf{I}}_{\text{PAN}}(i) = \sum_{q \in \mathscr{I}(j)} \widetilde{\mathbf{I}}_{\text{PAN}}(q) \prod_{d \in \{X,Y\}} (1 - |j_d - q_d|) \tag{2}$$

Once applied the uncovered deformation, the last segment of the proposed architecture is in charge of mapping the generated data onto the target HR space. In particular, three different blocks can be identified in $S_2 : S_2^{\text{h}}$ (head), S_2^{b} (body) and S_2^{t} (tail). First, the head block prepares the input data cube by concatenating $\widehat{\mathbf{I}}_{MS}$ and $\mathbf{I}_{PAN}$ using a single Cat layer. Second, the body block (S_2^{b}) projects these data to the final HR MS image $\mathbf{I}_{\text{HR}}$, which gathers the spatial details of $\mathbf{I}_{\text{HR}}$ and the spectral information of $\mathbf{I}_{\text{MS}}$. To this end, we follow the same U-Net backbone architecture used in S_1^{b} with the objective of designing the proposed model with two symmetric U-shaped segments. Note that these types of topological symmetries have shown to be effective for obtaining better CNN-based feature representations with limited data [50], which is precisely the case in the considered self-supervised pan-sharpening scenario. Hence, we set in S_2^{b} $K_{\text{IN}} = B + 1, K_1 = 16, K_2 = 32, S_1 = 2, S_2 = 1$ and $K_{\text{OUT}} = B$. Additionally, we also make use of a final skip connection to $\widehat{\mathbf{I}}_{\text{MS}}$ for driving the second U-Net towards the learning of spatial details not present in the low-spatial resolution domain. Finally, the tail block (S_2^{t}) is targeted at processing the generated HR output for allowing the use of an unsupervised training scheme based on the input $\mathbf{I}_{\text{MS}}$ and $\mathbf{I}_{\text{PAN}}$ images. Specifically, an spectral average Pool layer is used to simulate a panchromatic version of $\mathbf{I}_{\text{HR}}$ as $\mathbf{I}^*_{\text{PAN}}$, whereas an $R \times R$ spatial average Pool is applied to generate the simulated LR MS image (i.e., $\mathbf{I}^*_{\text{MS}}$).

3.2. Proposed Joint Loss Formulation

In this section, we describe the joint loss formulation proposed for training our W-NetPan architecture in a self-supervised manner. To that extent, it is important to highlight that the presented model only needs the input data volumes $\mathbf{I}_{\text{MS}}$ and $\mathbf{I}_{\text{PAN}}$ for training. As Fig. 2 shows, we consider a total three different loss functions: (a) spatial matching ($\mathscr{L}_1$), (b) spectral consistency ($\mathscr{L}_2$) and (c) spatial consistency ($\mathscr{L}_3$). Let us now describe them in more details:

(a) $\mathscr{L}_1$: The first loss is focused on the optimization of the initial segment of W-NetPan (i.e., S_1) in order to guarantee a good spatial matching between the two input optical sensors. For this purpose, $\mathscr{L}_1$ takes into account two different components $\mathscr{L}_{\text{LNCC}}$ and $\mathscr{L}_{\text{GRAD}}$. On the one hand, $\mathscr{L}_{\text{LNCC}}$ is a multi-modal reconstruction term between $\widehat{\mathbf{I}}_{\text{PAN}}$ and $\mathbf{I}_{\text{PAN}}$ to spatially match the transformed version of the simulated panchromatic and the original panchromatic image. Note that, at this point, inherent intensity variations are expected between both images due to the multi-modal nature of the data. Consequently, we make use of the Local Normalized Cross Correlation (LNCC) loss [51] as an efficient metric for quantifying the degree of alignment between two multi-

modal images. On the other hand, $\mathscr{L}_{\text{GRAD}}$ corresponds to a gradient-based regularization term for encouraging the generation of smooth deformation fields as well as spatially consistent local displacements. In this case, we employ a diffusion regularizer [52] on the spatial gradients of Φ. Eq. 3, 4,5 show the mathematical expressions for $\mathscr{L}_1$, being α a weighting hyper-parameter, Ω the 2D pixel grid of the image domain (given by *X*-*Y* spatial axis) and $\mathscr{I}(\cdot)$ the neighboring operator extracting an $(n \times n)$ output size.

$$\mathscr{L}_1\left(\mathbf{I}_{\text{PAN}}, \widehat{\mathbf{I}}_{\text{PAN}}, \Phi\right) = \mathscr{L}_{\text{LNCC}}\left(\mathbf{I}_{\text{PAN}}, \widehat{\mathbf{I}}_{\text{PAN}}\right) + \alpha \mathscr{L}_{\text{GRAD}}(\Phi) \tag{3}$$

$$\mathscr{L}_{\text{LNCC}}(I_1, I_2) = - \sum_{p \in \Omega} \frac{\sum_{q \in \mathscr{I}(p)} \left(\left(I_1(q) - \sum_{q_i \in \mathscr{I}(p)} \frac{I_1(q_i)}{n^2} \right) \left(I_2(q) - \sum_{q_i \in \mathscr{I}(p)} \frac{I_2(q_i)}{n^2} \right) \right)^2}{\sum_{q \in \mathscr{I}(p)} \left(I_1(q) - \sum_{q_i \in \mathscr{I}(p)} \frac{I_1(q_i)}{n^2} \right)^2 \sum_{q \in \mathscr{I}(p)} \left(I_2(q) - \sum_{q_i \in \mathscr{I}(p)} \frac{I_2(q_i)}{n^2} \right)^2} \tag{4}$$

$$\mathscr{L}_{\text{GRAD}}(\Phi) = \sum_{p \in \Omega} ||\nabla \Phi(p)|| = || \left(\frac{\partial \Phi(p)}{\partial X}, \frac{\partial \Phi(p)}{\partial Y} \right) || \tag{5}$$

(b) $\mathscr{L}_2$: The objective of the second loss consists in ensuring the spectral consistency between the output result (i.e., $\mathbf{I}_{\text{HR}}$) and the input MS image (i.e., $\mathbf{I}_{\text{MS}}$). To achieve this goal, $\mathscr{L}_2$ takes advantage of the simulated LR version of the output (i.e., $\mathbf{I}^*_{\text{MS}}$) in order to compute the mean squared error (MSE) with respect to $\mathbf{I}_{\text{MS}}$. Eq. 6 and 7 show the corresponding expressions, where $|\Omega|$ represents the total number of pixels of the image domain. It is important to note that, since our approach is a self-supervised model, this loss needs to work with the signal captured by the MS sensor. In this way, it is possible to fit the network output to the original MS data without using any ground truth information. This spectral consistency is computed at the low-level spatial resolution of $\mathbf{I}_{\text{MS}}$ to avoid the undesirable blurring effects generated when up-sampling the MS instrument to the target resolution. Otherwise, the considered MSE-based fit could compromise the sharpness of the solution.

$$\mathscr{L}_2\left(\mathbf{I}_{\text{MS}}, \mathbf{I}^*_{\text{MS}}\right) = \mathscr{L}_{\text{MSE}}\left(\mathbf{I}_{\text{MS}}, \mathbf{I}^*_{\text{MS}}\right) \tag{6}$$

$$\mathscr{L}_{\text{MSE}}(I_1, I_2) = \frac{1}{|\Omega|} \sum_{p \in \Omega} (I_1(p) - I_2(p))^2 \tag{7}$$

(c) $\mathscr{L}_3$: The third loss is aimed at guaranteeing the spatial consistency between $\mathbf{I}_{\text{HR}}$ and the input panchromatic image. For this purpose, $\mathscr{L}_3$ computes the similarity between the simulated panchromatic version of the network output (i.e., $\mathbf{I}^*_{\text{PAN}}$) and $\mathbf{I}_{\text{PAN}}$ by means of two different figures of merit: MSE and LNCC. Eq. 8 shows the considered loss expression, being β a weighting hyper-parameter. On the one hand, MSE quantifies the average squared differences between the simulated $\mathbf{I}^*_{\text{PAN}}$ and the original $\mathbf{I}_{\text{PAN}}$ with the objective of ensuring that the generated HR result does not have outlier predictions with huge spatial deviations. On the other hand, LNCC measures the relative local displacements between $\mathbf{I}^*_{\text{PAN}}$ and $\mathbf{I}_{\text{PAN}}$ in order to reduce the sensitivity to possible dynamic range changes between simulated and real panchromatic data. In this regard, it is important to highlight that we use an spectral pooling for simulating $\mathbf{I}^*_{\text{PAN}}$ and this process could introduce some artificial linear changes in the simulated panchromatic signal which may negatively affect MSE computations. In response, we add an LNCC-based term in $\mathscr{L}_3$ for making the spatial consistency loss more robust to some signal amplitude perturbations.

$$\mathscr{L}_3\left(\mathbf{I}_{\text{PAN}}, \mathbf{I}^*_{\text{PAN}}\right) = \mathscr{L}_{\text{MSE}}\left(\mathbf{I}_{\text{PAN}}, \mathbf{I}^*_{\text{PAN}}\right) + \beta \mathscr{L}_{\text{LNCC}}\left(\mathbf{I}_{\text{PAN}}, \mathbf{I}^*_{\text{PAN}}\right) \tag{8}$$

Finally, the proposed joint loss function for training W-NetPan can be formulated as follows,

$$\mathscr{L}_{\text{WNetPan}} = \mathscr{L}_1\left(\mathbf{I}_{\text{PAN}}, \widehat{\mathbf{I}}_{\text{PAN}}, \Phi\right) + \mathscr{L}_2\left(\mathbf{I}_{\text{MS}}, \mathbf{I}^*_{\text{MS}}\right) + \mathscr{L}_3\left(\mathbf{I}_{\text{PAN}}, \mathbf{I}^*_{\text{PAN}}\right) \tag{9}$$

4. Experiments

This section comprises the experimental part of the work, including the description of the considered datasets (section 4.1), the experimental settings (section 4.2) and the obtained results (section 4.3). Additional experiments are also included to provide a deeper understanding of the proposed model performance based on parameter sensitivity (section 4.4), ablation study (section 4.5) and trade-off analysis (section 4.6).

4.1. Datasets

This work includes four different datasets which are made of coupled S2 MSI and S3 OLCI products that cover several areas of interest across Europe. Table 2 summarizes the selected scenes as well as their acquisition dates and locations. Besides, Fig. 4 displays the corresponding images. The considered datasets are all cloud free Level-1C products that were downloaded from the Copernicus Open Access Hub[1] and processed via the Sentinel Application Platform (SNAP). In the case of S2, MSI images were atmospherically corrected using the Sen2Cor processor (with its default settings) and spatially resampled to 20 m for generating uniform data cubes. In the case of S3, OLCI images were corrected using the available Radiance to Reflectance processor. In addition, they were re-projected onto their associated S2 tiles to subset the overlapping areas between both sensors. With all these steps, we generated coupled images that represent the same extent over the Earth surface with an spatial-spectral size of $5490 \times 5490 \times 13$ in S2 and $366 \times 366 \times 21$ in S3. Now, let us define what are the input and output images of the considered inter-sensor pan-sharpening scheme, i.e. input LR MS ($\mathbf{I}_{\text{MS}}$), input HR panchromatic ($\mathbf{I}_{\text{PAN}}$) and output HR MS ($\mathbf{I}_{\text{HR}}$).

To relieve the lack of actual ground-truth data for a quantitative assessment, we make use of the following relaxations. First, we only consider those MSI and OLCI bands that are centered at the same wavelength, that is red (R), green (G), blue (B) and near infra-red (IR) bands, which are centered at $665, 560, 490$ and 865 nm, respectively. In this way, we define $\mathbf{I}_{\text{MS}} = \{\text{Oa04}, \text{Oa06}, \text{Oa08}, \text{Oa17}\}$. Second, we generate $\mathbf{I}_{\text{PAN}}$ by averaging MSI bands and resizing its spatial size to $R\times$ OLCI's resolution. Finally, we characterize $\mathbf{I}_{\text{HR}}$ by resizing MSI RGB-IR bands to $\mathbf{I}_{\text{PAN}}$ resolution and equalizing them to their corresponding OLCI's counterparts via an uniform mapping [53].

4.2. Experimental Settings

For validating the proposed inter-sensor self-supervised pan-sharpening model, we compare its performance to the one

[1] https://scihub.copernicus.eu/

Table 2
Description of the considered datasets.

Name	Scene	Location	Sensing dates		Tile (Ref. S2)
			S3	S2	
AN	Natural park	Andujar (Spain)	10/03/2017	10/03/2017	30SVH
MA	Southern Europe	Madrid (Spain)	10/04/2017	09/04/2017	30TVK
MI	Mountain range	Milan (Italy)	28/12/2016	07/01/2017	32TNR
UT	Northern Europe	Utrecht (Netherlands)	27/12/2016	27/12/2016	31UFT

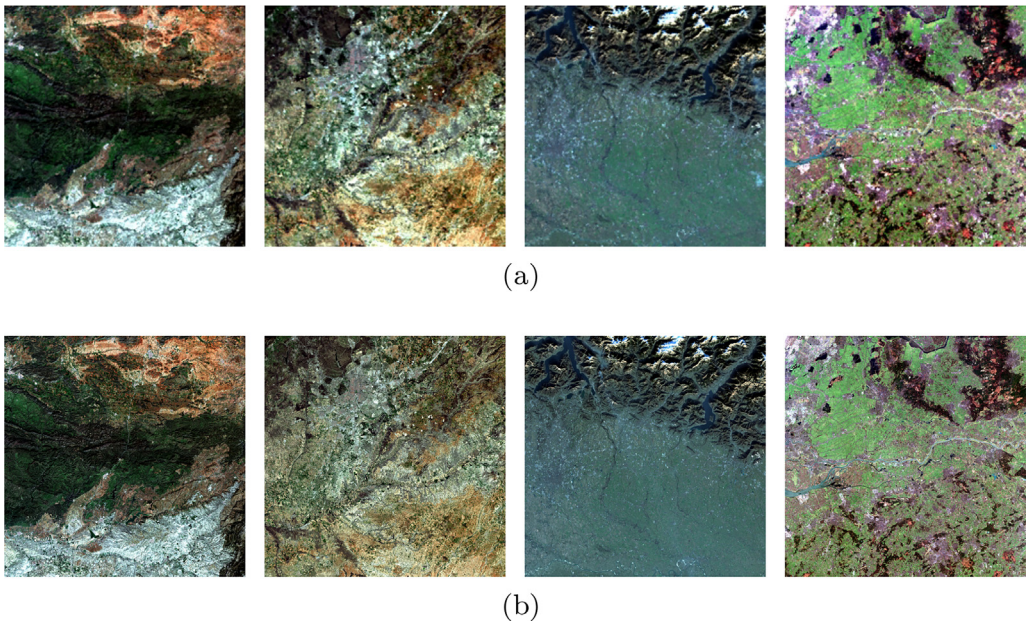

Fig. 4. Visualization of the considered datasets made of coupled S3 OLCI (a) and S2 MSI scenes (b).

obtained by several of the most popular pan-sharpening methods available in the literature, including Brovey [30], PCA [22], GFPCA [31], IHS [23], SFIM [24], GS [32], GSA [32], Wavelet [33], MTF-GLP [25], MTF-GLP-HPM [25], CNMF [34], PNN [26], PanNet [27], GPPNN [28] and PanColorGAN [44]. Additionally, we consider the bi-cubic kernel as up-scaling baseline. Regarding the inter-sensor spatial difference, we consider in this work a scaling ratio of $R = 4$ between $\mathbf{I}_{\text{MS}}$ and $\mathbf{I}_{\text{PAN}}$ images.

Under this scheme, we run all the considered pan-sharpening methods with their corresponding default settings a total of five Monte Carlo runs, reporting the corresponding average results. In the case of PNN, PanNet, GPPNN and PanColorGAN, it is important to note that we adopt a reduced-reference self-supervised training protocol to avoid requiring any HR ground-truth data for training (likewise the proposed network). Besides, all CNN models were trained via the ADAM optimizer for 200 epochs with (32×32) non-overlapping patches, $1e^{-3}$ learning rate and 16 batch size. Since the proposed architecture has been designed to process full-sized single images (not in a batch mode), we use in this case a total of 20000 iterations for the convergence of the model with $n = 9, \alpha = 0.5$ and $\beta = 1$ hyper-parameters. Note these hyper-parameters are set beforehand and the two segments of W-NetPan are simultaneously updated over the self-training process.

Regarding the assessment protocol, we use a total of six different metrics for quantitatively evaluating the obtained results [21]: mean squared error (MSE), peak signal to noise ratio (PSNR), structural similarity index measure (SSIM), spectral angle mapper (SAM), globale adimensionnelle de synthese (ERGAS) and spatial correlation coefficient (sCC). Note that MSE, PSNR, SSIM and ERGAS quantify global spatio-spectral deviations with respect to the ground-truth, whereas SAM and sCC are particularly focused on spectral and spatial variations, respectively. Additionally, different visual results are also considered for validating the methods from a qualitative perspective. In this work, we employed a computer with Intel(R) Core(TM) i7-6850 K, NVIDIA GeForce GTX 1080 Ti, 64 Gb of DDR4 RAM, Ubuntu 20.04 ×64 and Pytorch 1.6.0 with CUDA 10.1. The codes of this paper will be made available online[2].

4.3. Results

Table 3–6 provide the quantitative results obtained over the four considered datasets (i.e., Experiment 1: AN, Experiment 2: MA, Experiment 3: MI and Experiment 4: UT). Note that each table presents the assessment of a different image scene and Table 7 displays the overall average results. As it is possible to see, all the tables have been organized with the selected pan-sharpening methods in rows and the considered metrics in columns. Additionally, the two best results for each metric are highlighted in bold and the best result in gray shading font. In this case, the optimal values of the reported metrics are: MSE (0), PSNR ($+\infty$), SSIM (1), SAM (0), ERGAS (0) and sCC (1). For qualitative evaluation purposes, Fig. 5 also displays the pan-sharpened results obtained by the best performing methods (together with the bi-cubic baseline) when considering the AN dataset.

4.4. Parameter Sensitivity Analysis

Within the proposed model formulation, it is possible to find three main hyper-parameters: the window size $(n \times n)$ for the

[2] https://github.com/rufernan/WNetPan

Table 3
Experiment 1: Quantitative results for AN dataset.

Methods	MSE	PSNR	SSIM	SAM	ERGAS	sCC
Bicubic	0.0211	16.7625	0.1732	0.1784	10.3340	0.8081
Brovey [30]	0.0066	21.7718	0.8077	0.1836	5.1776	0.9386
PCA [22]	0.0111	19.5441	0.7783	0.3158	6.6860	0.9349
GFPCA [31]	0.0147	18.3289	0.3704	0.1663	8.5591	0.8551
IHS [23]	0.0062	22.0769	0.8216	0.1823	4.9716	0.9390
SFIM [24]	0.0152	18.1847	0.6531	0.2131	8.8011	0.8720
GS [32]	0.0063	22.0264	0.8087	0.1882	4.9966	0.9374
GSA [32]	0.0084	20.7560	0.7692	0.2259	6.0125	0.9391
Wavelet [33]	0.0130	18.8732	0.6724	0.1894	7.8306	0.8855
MTF-GLP [25]	0.0091	20.4329	0.7790	0.1956	6.2962	0.9227
MTF-GLP-HPM [25]	0.0110	19.5812	0.7328	0.2114	7.2994	0.9084
CNMF [34]	0.0081	20.9507	0.7393	0.2084	5.8798	0.9316
PNN [26]	0.0054	22.6657	0.8257	0.1666	4.7083	0.9450
PanNet [27]	0.0053	22.7447	0.8220	0.1701	4.7129	0.9466
GPPNN [28]	**0.0051**	**22.9654**	**0.8313**	**0.1615**	**4.5063**	**0.9494**
PanColorGAN [44]	0.0060	22.2660	0.8174	0.1720	5.0259	0.9439
W-NetPan (ours)	**0.0040**	**23.9985**	**0.8612**	**0.1491**	**4.0157**	**0.9625**

Table 4
Experiment 2: Quantitative results for MA dataset.

Methods	MSE	PSNR	SSIM	SAM	ERGAS	sCC
Bicubic	0.0260	15.8555	0.1760	0.2007	9.6313	0.7066
Brovey [30]	0.0114	19.4475	0.7504	0.2037	6.2663	0.8853
PCA [22]	0.0108	19.6717	0.7734	0.2432	6.2521	0.9068
GFPCA [31]	0.0183	17.3831	0.3940	0.1886	8.0489	0.7669
IHS [23]	0.0098	20.0801	0.7815	0.1944	5.8073	0.8989
SFIM [24]	0.0187	17.2833	0.6599	0.2236	8.1544	0.8295
GS [32]	0.0091	20.3947	**0.7845**	0.1887	5.6163	0.9046
GSA [32]	0.0148	18.3064	0.7271	0.2528	7.1827	0.9063
Wavelet [33]	0.0172	17.6449	0.6459	0.2088	7.7859	0.8278
MTF-GLP [25]	0.0149	18.2661	0.7355	0.2232	7.2190	0.8770
MTF-GLP-HPM [25]	0.0168	17.7487	0.7009	0.2258	7.6999	0.8606
CNMF [34]	0.0147	18.3856	0.7007	0.2199	7.1437	0.8837
PNN [26]	0.0082	20.8794	0.7552	0.1747	5.3066	0.9056
PanNet [27]	0.0082	20.8795	0.7675	0.1755	5.3131	0.9059
GPPNN [28]	**0.0082**	**20.8855**	0.7802	**0.1673**	**5.2980**	**0.9110**
PanColorGAN [44]	0.0083	20.8098	0.7752	0.1717	5.3472	0.9093
W-NetPan (ours)	**0.0077**	**21.1287**	**0.8064**	**0.1700**	**5.1459**	**0.9228**

Table 5
Experiment 3: Quantitative results for MI dataset.

Methods	MSE	PSNR	SSIM	SAM	ERGAS	sCC
Bicubic	0.0108	19.6586	0.4249	0.1214	8.0707	0.7298
Brovey [30]	0.0018	27.4786	**0.8371**	0.1212	3.1910	0.9590
PCA [22]	0.0036	24.4422	0.8178	0.2272	5.0267	0.9576
GFPCA [31]	0.0065	21.8861	0.5939	0.1242	6.2008	0.8311
IHS [23]	0.0018	27.5314	0.8341	0.1269	3.1738	0.9594
SFIM [24]	0.0067	21.7411	0.7532	0.1281	6.3917	0.8491
GS [32]	0.0018	27.4309	0.8335	0.1284	3.2228	0.9582
GSA [32]	0.0034	24.6565	0.7829	0.2024	4.4592	0.9482
Wavelet [33]	0.0066	21.8025	0.7299	0.1389	6.2764	0.8461
MTF-GLP [25]	0.0040	24.0095	0.8013	0.1461	4.8522	0.9158
MTF-GLP-HPM [25]	0.0044	23.5878	0.7956	0.1273	5.1113	0.9084
CNMF [34]	0.0039	24.1212	0.7618	0.1601	4.6986	0.9358
PNN [26]	0.0019	27.2581	0.8065	0.1276	3.2884	0.9566
PanNet [27]	0.0017	27.6669	0.8352	**0.1190**	**3.1413**	0.9607
GPPNN [28]	**0.0017**	**27.7110**	0.8297	**0.1198**	3.1933	**0.9614**
PanColorGAN [44]	0.0031	25.4021	0.8183	0.1212	4.1898	0.9334
W-NetPan (ours)	**0.0016**	**27.8355**	**0.8441**	0.1292	**3.0596**	**0.9643**

neighboring operator ($\mathscr{I}$) in $\mathscr{L}_{\text{LNCC}}$, the α weighting parameter in $\mathscr{L}_1$, and the β weighting parameter in $\mathscr{L}_3$.

Regarding the window size $(n \times n)$, this value indicates the amount of spatial context that is taken into account when computing the LNCC metric. Note that, in contrast to its global version, LNCC uses an sliding window to locally compute the cross correlation between both input images while accumulating such results. Considering the experimental configuration described in section 4.2, we test a single run of W-NetPan with the following window sizes $n = \{3, 5, 7, 9, 11, 13, 15\}$. The corresponding PSNR-based results are reported in Table 8. As it is possible to observe, window values below (7×7) are generally unable to provide satisfactory results since mean and variance local computations become rather biased estimations since too few pixel values are involved in such

Table 6
Experiment 4: Quantitative results for UT dataset.

Methods	MSE	PSNR	SSIM	SAM	ERGAS	sCC
Bicubic	0.0187	17.2902	0.2750	0.1704	7.7327	0.7580
Brovey [30]	0.0080	20.9633	0.7443	0.1712	5.1166	0.9052
PCA [22]	0.0085	20.7058	0.7458	0.1865	5.2848	0.9077
GFPCA [31]	0.0136	18.6540	0.4675	0.1643	6.6450	0.8022
IHS [23]	0.0075	21.2371	**0.7580**	0.1647	4.9602	0.9080
SFIM [24]	0.0142	18.4630	0.6890	0.2034	6.8040	0.8603
GS [32]	**0.0074**	**21.2996**	0.7549	0.1624	4.9282	0.9082
GSA [32]	0.0105	19.7884	0.7200	0.1932	5.8310	0.9132
Wavelet [33]	0.0124	19.0577	0.6604	0.1765	6.3439	0.8551
MTF-GLP [25]	0.0112	19.4952	0.7321	0.1849	6.0563	0.8975
MTF-GLP-HPM [25]	0.0135	18.7023	0.7157	0.2090	6.6358	0.8832
CNMF [34]	0.0126	19.0175	0.6813	0.2151	6.3801	0.8892
PNN [26]	0.0082	20.8584	0.7141	0.1555	5.1705	0.9092
PanNet [27]	0.0076	21.1722	0.7386	**0.1476**	5.0017	0.9181
GPPNN [28]	0.0075	21.2743	0.7473	**0.1388**	**4.9143**	**0.9232**
PanColorGAN [44]	0.0084	20.7471	0.7277	0.1536	5.2232	0.9128
W-NetPan (ours)	**0.0069**	**21.6326**	**0.7751**	0.1556	**4.7461**	**0.9195**

Table 7
Average quantitative results for all the considered datasets.

Methods	MSE	PSNR	SSIM	SAM	ERGAS	sCC
Bicubic	0.0191	17.3917	0.2622	0.1677	8.9421	0.7506
Brovey [30]	0.0070	22.4153	0.7849	0.1699	4.9379	0.9220
PCA [22]	0.0085	21.0909	0.7788	0.2432	5.8124	0.9267
GFPCA [31]	0.0133	19.0630	0.4565	0.1608	7.3635	0.8138
IHS [23]	0.0063	22.7314	**0.7988**	0.1671	4.7283	0.9263
SFIM [24]	0.0137	18.9180	0.6888	0.1920	7.5378	0.8527
GS [32]	0.0062	22.7879	0.7954	0.1669	4.6910	0.9271
GSA [32]	0.0093	20.8768	0.7498	0.2186	5.8713	0.9267
Wavelet [33]	0.0123	19.3446	0.6772	0.1784	7.0592	0.8536
MTF-GLP [25]	0.0098	20.5509	0.7620	0.1875	6.1059	0.9032
MTF-GLP-HPM [25]	0.0114	19.9050	0.7363	0.1934	6.6866	0.8901
CNMF [34]	0.0098	20.6188	0.7208	0.2009	6.0255	0.9101
PNN [26]	0.0059	22.9154	0.7754	0.1561	4.6185	0.9291
PanNet [27]	0.0057	23.1158	0.7908	0.1530	4.5423	0.9328
GPPNN [28]	**0.0056**	**23.2090**	0.7971	**0.1468**	**4.4780**	**0.9362**
PanColorGAN [44]	0.0064	22.3063	0.7846	0.1546	4.9466	0.9248
W-NetPan (ours)	**0.0051**	**23.6488**	**0.8217**	**0.1510**	**4.2418**	**0.9423**

operations. Additionally, using a window size beyond (11×11) leads to saturate the performance (or even slightly worsen it depending on the input image) because considering too large areas may logically affect the capability of detecting local image variations. For the sake of generality, we set a default window size of (9×9) as in [51].

In the case of α, this hyper-parameter weights the diffusion regularization of the deformation field uncovered by the first U-Net segment of the proposed architecture. More specifically, it modulates the gradient variations of the deformations that are internally used to correct the low spatial resolution image. Table 9 presents the corresponding PSNR-based evaluation when testing W-NetPan with $\alpha = \{0.0, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$ and fixing the other hyper-parameters to their default values. According to the reported results, the performance becomes quite stable for values higher than 0.1. In this way, we use an intermediate α of 0.5 as default value. (See Table 10).

Finally, β is another weighting hyper-parameter that aims at balancing the spatial consistency loss ($\mathscr{L}_3$) in the second U-Net segment of the proposed model. To analyze the impact of this hyper-parameter on the final performance, we test the following values $\beta = \{0.0, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$ likewise in the previous case. Table 9 contains the obtained results, where values larger than 0.5 provide similar performances. For the sake of simplicity, we set $\beta = 1$ as default value.

4.5. Ablation Study

Another important point that deserve to be analyzed is the contribution of each one of the U-Net segments that constitute the proposed architecture. To achieve this goal, this section includes an ablation study to compare W-NetPan to a simplified version that removes the first U-Net segment (we name this variant as single U-Net). In this way, it will be possible to see the contribution of the proposed W-shape with respect to the standard U-shape when using the same elemental configuration. Table 11 presents the results of this ablation study based on a single run and all the considered metrics. According to the results, it is possible to see that the proposed approach achieves consistent performance improvements with respect to its simplified version (single U-Net) over all the considered datasets. On average, W-NetPan provides gains of 0.45 in PSNR, 0.0112 in SSIM and -0.0072 in SAM, which reveals the spatial-spectral contribution of our newly designed W-shaped architecture.

4.6. Trade-off Analysis

The presented joint loss Eq. 9 is formulated according to three terms, i.e., spatial matching ($\mathscr{L}_1$), spectral consistency ($\mathscr{L}_2$) and spatial consistency ($\mathscr{L}_3$). This section provides a deeper analysis of the impact of each one of these terms into the final performance

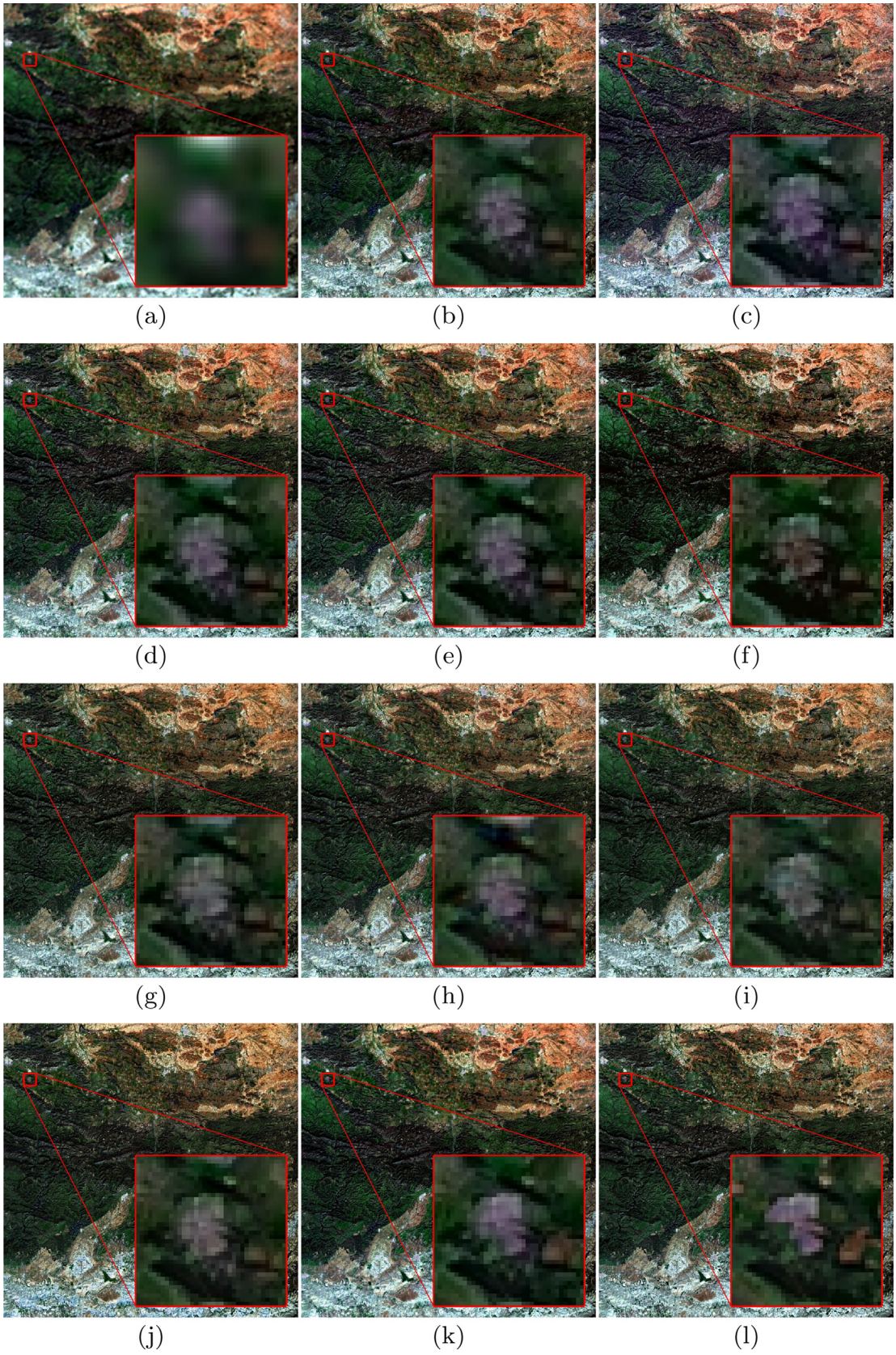

Fig. 5. Qualitative results for Experiment 1 (PSNR values in brackets): (a) Bicubic (16.76 dB), (b) Brovey (21.77 dB), (c) PCA (19.54 dB), (d) IHS (22.08 dB), (e) GS (22.03 dB), (f) CNMF (20.95 dB), (g) PNN (22.57 dB), (h) PanNet (22.71 dB), (i) GPPNN (23.07 dB), (j) PanColorGAN (22.42 dB), (k) W-NetPan (24.01 dB) and (l) ground-truth.

Table 8
Analysis of the window size ($n \times n$) for the neighboring operator ($\mathscr{X}$) in $\mathscr{L}_{\text{LNCC}}$ according to the PSNR (dB) metric.

Dataset	$n = 3$	$n = 5$	$n = 7$	$n = 9$	$n = 11$	$n = 13$	$n = 15$
AN	22.6254	23.9348	23.9258	23.9943	23.9829	23.9787	24.0027
MA	20.5611	21.0569	21.0719	21.0829	21.1317	21.1447	21.1646
MI	18.8067	24.8755	27.5464	28.0231	28.1468	28.1595	28.2241
UT	20.3981	21.5831	21.6127	21.6256	21.6637	21.6518	21.6423
Avg.	20.5978	22.8626	23.5392	23.6815	23.7313	23.7337	23.7585

Table 9
Analysis of the α hyper-parameter in $\mathscr{L}_1$ according to the PSNR (dB) metric.

Dataset	$\alpha = 0.0$	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 0.9$	$\alpha = 1.0$
AN	23.7690	23.9608	23.9857	23.9908	23.9715	23.9756	24.0071
MA	20.8683	21.0249	21.0905	21.0917	21.1074	21.0984	21.1099
MI	28.0059	28.0555	28.0250	28.0468	28.0485	28.0234	28.0411
UT	21.5226	21.6159	21.6396	21.6186	21.6228	21.6306	21.6007
Avg.	23.5414	23.6643	23.6852	23.6870	23.6875	23.6820	23.6897

Table 10
Analysis of the β hyper-parameter in $\mathscr{L}_3$ according to the PSNR (dB) metric.

Dataset	$\beta = 0.0$	$\beta = 0.1$	$\beta = 0.3$	$\beta = 0.5$	$\beta = 0.7$	$\beta = 0.9$	$\beta = 1.0$
AN	23.5114	23.8995	23.9603	23.9497	23.9649	23.9735	23.9760
MA	20.6989	20.9772	21.0659	21.0703	21.0664	21.1064	21.0873
MI	27.3351	27.9742	28.0293	28.0450	28.0407	28.0292	28.0040
UT	21.1960	21.5280	21.5832	21.6091	21.6161	21.6244	21.6166
Avg.	23.1853	23.5947	23.6597	23.6685	23.6721	23.6834	23.6710

Table 11
Ablation study to compare the performance of the proposed architecture (W-NetPan) with respect to its simplified version with a single U-Net block (single U-Net).

Dataset	Methods	MSE	PSNR	SSIM	SAM	ERGAS	sCC
AN	Ablation (single U-Net)	0.0047	23.3084	0.8509	0.1555	4.3705	0.9566
	Proposed (W-NetPan)	**0.0040**	**23.9816**	**0.8606**	**0.1497**	**4.0232**	**0.9623**
MA	Ablation (single U-Net)	0.0091	20.3861	0.7869	0.1824	5.6043	0.9115
	Proposed (W-NetPan)	**0.0077**	**21.1072**	**0.8061**	**0.1700**	**5.1581**	**0.9227**
MI	Ablation (single U-Net)	0.0016	27.9006	0.8311	0.1263	3.0317	0.9585
	Proposed (W-NetPan)	**0.0016**	**28.0522**	**0.8351**	**0.1216**	**2.9741**	**0.9594**
UT	Ablation (single U-Net)	0.0073	21.3872	0.7636	0.1609	4.8775	0.9149
	Proposed (W-NetPan)	**0.0069**	**21.6408**	**0.7754**	**0.1551**	**4.7411**	**0.9198**
Avg.	Ablation (single U-Net)	0.0057	23.2456	0.8081	0.1563	4.4710	0.9354
	Proposed (W-NetPan)	**0.0050**	**23.6955**	**0.8193**	**0.1491**	**4.2241**	**0.9410**

of the proposed pan-sharpening method. To this aim, we define three trade-off parameters (i.e., L1, L2 and L3) to weight the three considered loss terms (i.e., $\mathscr{L}_1, \mathscr{L}_2$ and $\mathscr{L}_3$, respectively). Then, we test the effect of setting each trade-off parameter to 0.0 (null-activation of the corresponding loss term), 0.5 (half-activation) and 1.0 (full-activation). Note that this configuration leads to 27 permutations per dataset and a total of 108 experiments. Fig. 6 presents the average results based on the PSNR metric. As it is possible to observe, the best results are always achieved when activating all three components which indicates the advantages of the proposed joint loss formulation with respect to other alternatives.

5. Discussion

When analyzing the obtained results, it is possible to note several important points that deserve to be mentioned. The first one is related to the global performance of the considered pan-sharpening techniques. In general, we can see that all the methods are able to outperform the bi-cubic baseline by a wide margin based on their average spatial-spectral metric assessments. This fact certainly indicates the high suitability of the pan-sharpening technology even when dealing with images coming from different platforms. In this sense, the exploitation of inter-platform data allows pan-sharpening algorithms to take better advantage of panchromatic images for alleviating some spatial deviations that do not need to be resolved in an intra-sensor scenario. Regarding the nature of the considered methods, we can also find remarkable performance differences between traditional and DL-based pan-sharpening techniques. In the case of traditional models, Brovey, IHS and GS have shown to be the three best traditional methods since they achieve the best average results from the perspective of MSE, PSNR, SSIM and ERGAS metrics. Nonetheless, their CS-based nature still makes them rather prone to generate spectral distortions as SAM reveals. In the case of DL methods, they are able to achieve a better spatial-spectral effectiveness, being PNN, PanNet and GPPNN the three best overall competitors and the proposed W-NetPan model certainly the best performing method.

Focusing on the nature of the considered metrics, a more detailed performance discussion can be made. In particular, four different analyses are possible based on the considered types of figures of merit: (1) error-based, (2) spatial, (3) spectral and (4) spatial-spectral. The first type (1), including MSE, PSNR and ERGAS, aims at computing average differences to globally measure the quality of the pan-sharpening results. Considering the PSNR metric as reference, we can see how W-NetPan clearly obtains the best

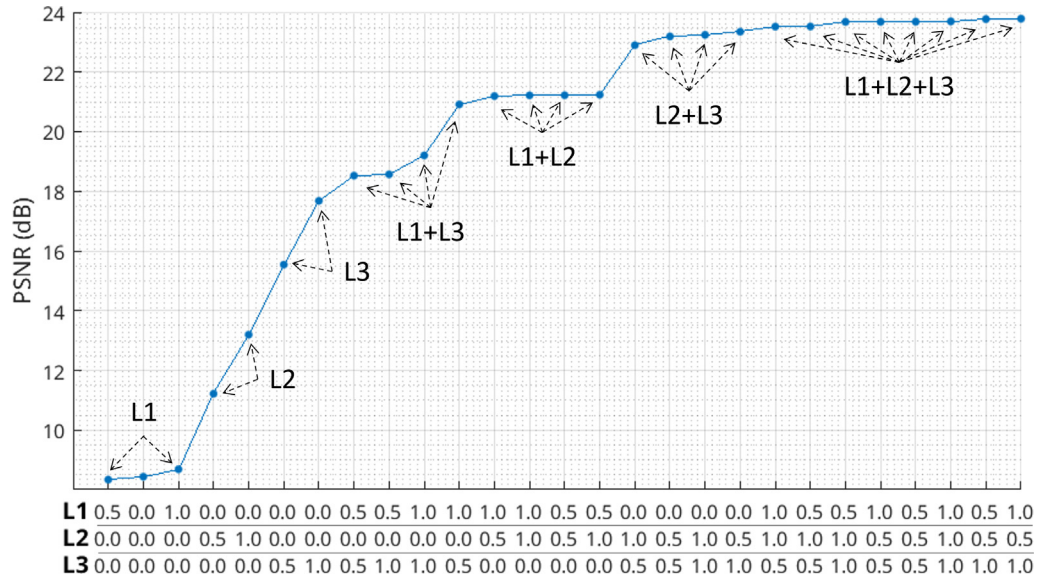

Fig. 6. Trade-off analysis of the considered loss terms. The horizontal axis displays the weighting values for $\mathscr{L}_1$ (L1), $\mathscr{L}_2$ (L2) and $\mathscr{L}_3$ (L3) terms, respectively. The vertical axis shows the PSNR (dB) results averaged over the considered datasets.

average value (23.65 dB), followed by GPPNN (23.21 dB), PanNet (23.12 dB), PNN (22.92 dB), GS (22.79 dB) and IHS (22.73 dB). Regarding MSE and ERGAS, analogous improvements can also be found in Table 7, which reveals the consistent error reductions provided by the proposed model. The second group (2), made of sCC, pursues to quantify the spatial correlations between pan-sharpened and ground-truth images. In this case, W-NetPan provides the best result followed by GPPNN, PanNet and PNN. Nonetheless, it is important to note that GS and IHS suffer a relevant sCC decrease, which indicates the lower spatial performance achieved by traditional methods. Regarding the third group (3), covering SAM, it is specially focused on the isolated computation of spectral differences. As it is possible to see in Table 7, the two best performing methods are GPPNN and W-NetPan followed by PanNet and PNN. Traditional pan-sharpening methods fail in this case since they are often unable to outperform the bi-cubic baseline. Although the proposed approach does not achieve the best average result, it is among the two best alternatives reaching state-of-the-art spectral performances in the conducted experiments. Finally, the forth group (4), composed of SSIM, makes a spatial-spectral assessment where correlation, luminance and contrast distortions are jointly taken into account for generating a complete image quality evaluation. According to Table 7, W-NetPan obtains the best quantitative result by a wide margin, followed by IHS, GPPNN, GS and PanNet. The displayed qualitative results also support these observations. As Fig. 4 shows, DL-based methods are generally able to produce better visual results than traditional ones, especially from an structural perspective. In more details, the proposed W-NetPan model certainly provides the most similar output with respect to the corresponding ground-truth. Considering both quantitative and qualitative performances, it is possible to see the significance of the achieved improvements with respect to those from other state-of-the-art models.

As an overall observation, the performed experimental comparison validates the higher suitability of W-NetPan for conducting inter-sensor pan-sharpening between S2 and S3 satellites. Unlike other instruments, OLCI has a particularly coarse spatial resolution (i.e., 300 m) that makes the straightforward up-scaling process rather uninformative from a data fusion perspective. As it is possible to observe in Fig. 4. (a), the use of a bi-cubic interpolation over S3 provides a very blurred result that makes difficult to identify even the most basic image regions and shapes. In this way, many of the existing DL-based pan-sharpening methods struggle at recognizing which spatial regions in S2 may correspond to which spectral data in S3 (see Fig. 4. (g)-(h) as an example). In contrast, the proposed network tries to relieve this effect by means of its W-shaped architecture. Specifically, the first U-Net segment pursues to identify the main spatial regions in S3 while adjusting and matching their boundaries to the corresponding S2 content. Note that these kinds of inter-sensor adjustments are highly convenient since MSI is logically more spatially reliable than OLCI. Then, the second segment takes advantage of these corrected data to generate the final pan-sharpened output. In this fashion, both U-Nets can provide feedback one another during training to achieve better fusion results from an end-to-end perspective.

6. Conclusions and Future Work

This paper has presented a new DL-based inter-sensor pan-sharpening model (W-NetPan), which has been specifically designed to deal with S2 and S3 data. In particular, the proposed architecture defines an innovative W-shape which jointly works for spatially matching and fusing inter-sensor data. Besides, the proposed loss formulation allows training the model without any external data supervision. The conducted experimental comparison, including several datasets and state-of-the-art pan-sharpening methods, reveals the competitive performance provided by the proposed approach.

One of the most important conclusions that can be extracted from this work is the high complexity of fusing rather heterogeneous data, such as in the case of S2 and S3 optical products. Specifically, the particularly low spatial resolution of OLCI together with the inherent inter-platform deviations often make state-of-the-art pan-sharpening models unable to obtain satisfactory results when projecting S3 spectral and S2 spatial information onto the corresponding fused space. In this sense, adopting a W-shaped network together with a self-supervised loss has shown to provide competitive advantages with respect several of the most important traditional and DL-based methods available in the literature. Although the obtained results are certainly promising, there is still room for further improvements. As future work, we plan to extend this research towards the following directions: 1) extending the proposed network to other inter-sensor platforms, and 2) expand-

ing this research to single-frame super-resolution. How to exploit the proposed model in other pixel-level vision tasks (e.g. [54–57]) certainly is another interesting research line that deserve to be mentioned.

CRediT authorship contribution statement

Ruben Fernandez-Beltran: Conceptualization, Methodology, Software, Writing - original draft. **Rafael Fernandez:** Conceptualization, Methodology, Software, Writing - original draft. **Jian Kang:** Investigation, Data curation, Writing - review & editing, Visualization. **Filiberto Pla:** Writing - review & editing, Supervision.

Data availability

Data will be made available on request.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Ministry of Science and Innovation of Spain (Grant No. PID2021-128794OB-I00) and the National Natural Science Foundation of China under Grant 62101371.

References

[1] M. ED Chaves, M. CA Picoli, I.D Sanches, Recent applications of landsat 8/oli and sentinel-2/msi for land use and land cover mapping: A systematic review, Remote Sensing 12 (18) (2020) 3062.
[2] Q. Bi, H. Zhang, K. Qin, Multi-scale stacking attention pooling for remote sensing scene classification, Neurocomputing 436 (2021) 147–161.
[3] R. Fernandez-Beltran, T. Baidar, J. Kang, F. Pla, Rice-yield prediction with multi-temporal sentinel-2 data and 3d cnn: A case study in nepal, Remote Sensing 13 (7) (2021) 1391.
[4] B. El Mahrad, A. Newton, J.D. Icely, I. Kacimi, S. Abalansa, M. Snoussi, Contribution of remote sensing technologies to a holistic coastal and marine environmental management framework: A review, Remote Sensing 12 (14) (2020) 2313.
[5] S. Chen, Y. Cao, X. Feng, X. Lu, Global2salient: Self-adaptive feature aggregation for remote sensing smoke detection, Neurocomputing (2021).
[6] R. Fernandez-Beltran, F. Pla, J. Kang, J. Moreno, A. Plaza, Sentinel-3/flex biophysical product confidence using sentinel-2 land-cover spatial distributions, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14 (2021) 3447–3461.
[7] J. Kang, R. Fernandez-Beltran, X. Sun, J. Ni, A. Plaza, Deep learning-based building footprint extraction with missing annotations, IEEE Geoscience and Remote Sensing Letters (2021).
[8] Y. Tan, S. Xiong, P. Yan, Multi-branch convolutional neural network for built-up area extraction from remote sensing image, Neurocomputing 396 (2020) 358–374.
[9] J. Kang, R. Fernandez-Beltran, D. Hong, J. Chanussot, A. Plaza, Graph relation network: Modeling relations between scenes for multilabel remote-sensing image classification and retrieval, IEEE Transactions on Geoscience and Remote Sensing 59 (5) (2020) 4355–4369.
[10] R. Fernandez-Beltran, A. Plaza, J. Plaza, F. Pla, Hyperspectral unmixing based on dual-depth sparse probabilistic latent semantic analysis, IEEE Transactions on Geoscience and Remote Sensing 56 (11) (2018) 6344–6360.
[11] X. Zhang, X. Li, Y. Dong, Robust hyperspectral unmixing based on dual views with adaptive weights, Neurocomputing 461 (2021) 204–216.
[12] R. Fernandez-Beltran, F. Pla, A. Plaza, Endmember extraction from hyperspectral imagery based on probabilistic tensor moments, IEEE Geoscience and Remote Sensing Letters 17 (12) (2020) 2120–2124.
[13] N.K. Keppy, M. Allen, Understanding spectral bandwidth and resolution in the regulated laboratory, Thermo Fisher Scientific Technical Note 51721 (2008).
[14] J. Aschbacher, M.P. Milagro-Pérez, The European Earth monitoring (GMES) programme: Status and perspectives, Remote Sensing of Environment 120 (2012) 3–8.
[15] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, et al., Sentinel-2: Esa's optical high-resolution mission for gmes operational services, Remote sensing of Environment 120 (2012) 25–36.
[16] C. Donlon, B. Berruti, A. Buongiorno, M.-H. Ferreira, P. Féménias, J. Frerick, P. Goryl, U. Klein, H. Laur, C. Mavrocordatos, et al., The global monitoring for environment and security (GMES) sentinel-3 mission, Remote Sensing of Environment 120 (2012) 37–57.
[17] Z. Malenovský, H. Rott, J. Cihlar, M.E. Schaepman, G. García-Santos, R. Fernandes, M. Berger, Sentinels for science: Potential of Sentinel-1,-2, and-3 missions for scientific observations of ocean, cryosphere, and land, Remote Sensing of Environment 120 (2012) 91–101.
[18] X. Meng, H. Shen, H. Li, L. Zhang, R. Fu, Review of the pansharpening methods for remote sensing images based on the idea of meta-analysis: Practical discussion and challenges, Information Fusion 46 (2019) 102–113.
[19] R. Fernandez-Beltran, P. Latorre-Carmona, F. Pla, Single-frame super-resolution in remote sensing: a practical overview, International Journal of Remote Sensing 38 (1) (2017) 314–354.
[20] G. Vivone, L. Alparone, J. Chanussot, M. Dalla Mura, A. Garzelli, G.A. Licciardi, R. Restaino, L. Wald, A critical comparison among pansharpening algorithms, IEEE Transactions on Geoscience and Remote Sensing 53 (5) (2014) 2565–2586.
[21] F.D. Javan, F. Samadzadegan, S. Mehravar, A. Toosi, R. Khatami, A. Stein, A review of image fusion techniques for pan-sharpening of high-resolution satellite imagery, ISPRS Journal of Photogrammetry and Remote Sensing 171 (2021) 101–117.
[22] P. Kwarteng, A. Chavez, Extracting spectral contrast in landsat thematic mapper image data using selective principal component analysis, Photogramm. Eng. Remote Sens 55 (1) (1989) 339–348.
[23] W. Carper, T. Lillesand, R. Kiefer, The use of intensity-hue-saturation transformations for merging spot panchromatic and multispectral image data, Photogrammetric Engineering and remote sensing 56 (4) (1990) 459–467.
[24] J. Liu, Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details, International Journal of Remote Sensing 21 (18) (2000) 3461–3472.
[25] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, M. Selva, Mtf-tailored multiscale fusion of high-resolution ms and pan imagery, Photogrammetric Engineering & Remote Sensing 72 (5) (2006) 591–596.
[26] G. Masi, D. Cozzolino, L. Verdoliva, G. Scarpa, Pansharpening by convolutional neural networks, Remote Sensing 8 (7) (2016) 594.
[27] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, J. Paisley, Pannet: A deep network architecture for pan-sharpening, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 5449–5457.
[28] S. Xu, J. Zhang, Z. Zhao, K. Sun, J. Liu, C. Zhang, Deep gradient projection networks for pan-sharpening, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1366–1375.
[29] L. Alparone, A. Garzelli, G. Vivone, Intersensor statistical matching for pansharpening: Theoretical issues and practical solutions, IEEE Transactions on Geoscience and Remote Sensing 55 (8) (2017) 4682–4695.
[30] A.R. Gillespie, A.B. Kahle, R.E. Walker, Color enhancement of highly correlated images. ii. channel ratio and?chromaticity? transformation techniques, Remote Sensing of Environment 22 (3) (1987) 343–365.
[31] W. Liao, X. Huang, F. Van Coillie, G. Thoonen, A. Pižurica, P. Scheunders, W. Philips, Two-stage fusion of thermal hyperspectral and visible rgb image by pca and guided filter, in: 2015 7th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), Ieee, 2015, pp. 1–4.
[32] B. Aiazzi, S. Baronti, M. Selva, Improving component substitution pansharpening through multivariate regression of ms + pan data, IEEE Transactions on Geoscience and Remote Sensing 45 (10) (2007) 3230–3239.
[33] R.L. King, J. Wang, A wavelet based algorithm for pan sharpening landsat 7 imagery, in: IGARSS 2001. Scanning the Present and Resolving the Future. Proceedings. IEEE 2001 International Geoscience and Remote Sensing Symposium (Cat. No. 01CH37217), Vol. 2, IEEE, 2001, pp. 849–851.
[34] N. Yokoya, T. Yairi, A. Iwasaki, Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion, IEEE Transactions on Geoscience and Remote Sensing 50 (2) (2011) 528–537.
[35] D. Zhang, D. Meng, J. Han, Co-saliency detection via a self-paced multiple-instance learning framework, IEEE transactions on pattern analysis and machine intelligence 39 (5) (2017) 865–878.
[36] D. Cheng, J. Zhou, N. Wang, X. Gao, Hybrid dynamic contrast and probability distillation for unsupervised person re-id, IEEE Transactions on Image Processing 31 (2022) 3334–3346.
[37] R. Fernandez, R. Fernandez-Beltran, J. Kang, F. Pla, Sentinel-3 super-resolution based on dense multireceptive channel attention, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14 (2021) 7359–7372.
[38] R. Fernandez-Beltran, D. Ibañez, J. Kang, F. Pla, Time-resolved sentinel-3 vegetation indices via inter-sensor 3-d convolutional regression networks, IEEE Geoscience and Remote Sensing Letters (2021).
[39] Z. Wang, J. Chen, S.C.H. Hoi, Deep learning for image super-resolution: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (10) (2021) 3365–3387, https://doi.org/10.1109/TPAMI.2020.2982166.
[40] G. Scarpa, S. Vitale, D. Cozzolino, Target-adaptive cnn-based pansharpening, IEEE Transactions on Geoscience and Remote Sensing 56 (9) (2018) 5443–5457.
[41] D. Ulyanov, A. Vedaldi, V. Lempitsky, Deep image prior, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 9446–9454.

[42] R.T. Rockafellar, Monotone operators and the proximal point algorithm, SIAM journal on control and optimization 14 (5) (1976) 877–898.
[43] T. Uezato, D. Hong, N. Yokoya, W. He, Guided deep decoder: Unsupervised image pair fusion, European Conference on Computer Vision, Springer (2020) 87–102.
[44] F. Ozcelik, U. Alganci, E. Sertel, G. Unal, Rethinking cnn-based pansharpening: Guided colorization of panchromatic images via gans, IEEE Transactions on Geoscience and Remote Sensing 59 (4) (2020) 3486–3501.
[45] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.
[46] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1125–1134.
[47] W. Yao, Z. Zeng, C. Lian, H. Tang, Pixel-wise regression using u-net and its application on pansharpening, Neurocomputing 312 (2018) 364–371.
[48] R. Fernandez-Beltran, F. Pla, A. Plaza, Intersensor remote sensing image registration using multispectral semantic embeddings, IEEE Geoscience and Remote Sensing Letters 16 (10) (2019) 1545–1549.
[49] M. Jaderberg, K. Simonyan, A. Zisserman, et al., Spatial transformer networks, Advances in neural information processing systems 28 (2015) 2017–2025.
[50] M. Zhang, M. Gong, H. He, S. Zhu, Symmetric all convolutional neural-network-based unsupervised feature extraction for hyperspectral images classification, IEEE Transactions on Cybernetics (2020).
[51] X. Cao, J. Yang, L. Wang, Z. Xue, Q. Wang, D. Shen, Deep learning based inter-modality image registration supervised by intra-modality similarity, in: International workshop on machine learning in medical imaging, Springer, 2018, pp. 55–63.
[52] J. Modersitzki, FAIR: flexible algorithms for image registration, SIAM, 2009.
[53] M. Kaur, J. Kaur, J. Kaur, Survey of contrast enhancement techniques based on histogram equalization, International Journal of Advanced Computer Science and Applications 2 (7) (2011).
[54] P. Huang, J. Han, N. Liu, J. Ren, D. Zhang, Scribble-supervised video object segmentation, IEEE/CAA Journal of Automatica Sinica 9 (2) (2021) 339–353.
[55] N. Liu, J. Han, Dhsnet: Deep hierarchical saliency network for salient object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 678–686.
[56] D. Zhang, J. Han, Y. Zhang, D. Xu, Synthesizing supervision for learning deep saliency network without human annotation, IEEE transactions on pattern analysis and machine intelligence 42 (7) (2019) 1755–1769.
[57] D. Zhang, G. Huang, Q. Zhang, J. Han, J. Han, Y. Yu, Cross-modality deep feature learning for brain tumor segmentation, Pattern Recognition 110 (2021).

Ruben Fernandez-Beltran earned a B.Sc. degree in Computer Science, a M.Sc. in Intelligent Systems and a Ph.D. degree in Computer Science, from the University Jaume I (Castellon de la Plana, Spain) in 2007, 2011 and 2016, respectively. He is currently an Associate Professor within the Department of Computer Science and Systems at the University of Murcia, Spain. He has been visiting researcher at the University of Bristol (UK), the University of Caceres (Spain), the Technical University of Berlin (Germany) and the Autonomous University of Mexico State (Mexico). His research interests lie in multimedia retrieval, spatio-spectral image analysis, pattern recognition techniques applied to image processing and remote sensing. Dr. Fernandez-Beltran was awarded with the Outstanding Ph.D. Dissertation Award at Universitat Jaume I in 2017.

Rafael Fernandez received a B.S. degree in Computer Science from Universitat Jaume I, Castellon de la Plana, Spain, in 2005. In 2019, he certified his M.Sc. level in Computer Science under the European Qualification Framework. Since then, he is pursuing a Ph.D. degree in Computer Science from Universitat Jaume I. His research interests include computer vision and machine learning, with special interest in remote sensing applications.

Jian Kang received the B.S. and M.E. degrees in electronic engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2013 and 2015, respectively, and the Dr.Ing. degree from the Signal Processing in Earth Observation (SiPEO) Group, Technical University of Munich (TUM), Munich, Germany, in 2019.,In August 2018, he was a Guest Researcher with the Institute of Computer Graphics and Vision (ICG), TU Graz, Graz, Austria. From 2019 to 2020, he was with the Faculty of Electrical Engineering and Computer Science, Technische Universit?t Berlin (TU Berlin), Berlin, Germany. He is currently with the School of Electronic and Information Engineering, Soochow University, Suzhou, China. His research focuses on signal processing and machine learning techniques, and their applications in remote sensing. In particular, he is interested in intelligent synthetic aperture radar (SAR)/interferometric SAR (InSAR) data processing and deep learning-based techniques for remote sensing image analysis. Dr. Kang received the First Place of the Best Student Paper Award from the European Conference on Synthetic Aperture Radar (EUSAR) 2018, Aachen, Germany. His joint work was selected as one of the Ten Student Paper Competition Finalists at the International Geoscience and Remote Sensing Symposium (IGARSS) 2020. He has served as a Guest Editor for IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.

Filiberto Pla received the B.Sc. and Ph.D. degrees in physics from the Universitat de Valencia, Spain, in 1989 and 1993, respectively. He is currently a Full Professor with the Departament de Llenguatges i Sistemes Informatics, University Jaume I, Castellon de la Plana, Spain. He has been a Visiting Scientist with the Silsoe Research Institute, the University of Surrey, the University of Bristol, U.K., CEMAGREF, France, the University of Genoa, Italy, the Instituto Superior Tecnico, Lisbon, Portugal, the Swiss Federal Institute of Technology, ETH-Zurich, the Idiap Research Institute, Switzerland, and the Technical University of Delft, The Netherlands. He is a faculty member of the Institute of New Imaging Technologies, University Jaume I. His current research interests are color and spectral image analysis, visual motion analysis, 3-D image capture and visualization, and pattern recognition techniques applied to image processing. He is a member of the Spanish Association for Pattern Recognition and Image Analysis, which is a partner of the International Association for Pattern Recognition.