Pansharpening by Convolutional Neural Networks in the Full Resolution Framework

Matteo Ciotola, *Graduate Student Member, IEEE*, Sergio Vitale, *Student Member, IEEE*, Antonio Mazza, *Student Member, IEEE*, Giovanni Poggi, *Member, IEEE*, and Giuseppe Scarpa, *Senior Member, IEEE*

***Abstract*— In recent years, there has been a growing interest in deep learning-based pansharpening. Thus far, research has mainly focused on architectures. Nonetheless, model training is an equally important issue. A first problem is the absence of ground truths, unavoidable in pansharpening. This is often addressed by training networks in a reduced-resolution domain and using the original data as ground truth, relying on an implicit scale invariance assumption. However, on full-resolution images, results are often disappointing, suggesting such invariance not to hold. A further problem is the scarcity of training data, which causes a limited generalization ability and a poor performance on off-training-test images. In this article, we propose a full-resolution training framework for deep learning-based pansharpening. The framework is fully general and can be used for any deep learning-based pansharpening model. Training takes place in the high-resolution domain, relying only on the original data, thus avoiding any loss of information. To ensure spectral and spatial fidelity, a suitable two-component loss is defined. The spectral component enforces consistency between the pansharpened output and the low-resolution multispectral input. The spatial component, computed at high resolution, maximizes the local correlation between each pansharpened band and the panchromatic input. At testing time, the target-adaptive operating modality is adopted, achieving good generalization with a limited computational overhead. Experiments carried out on WorldView-3, WorldView-2, and GeoEye-1 images show that methods trained with the proposed framework guarantee a pretty good performance in terms of both full-resolution numerical indexes and visual quality.**

***Index Terms*— Convolutional neural network (CNN), data fusion, deep learning, image enhancement, multiresolution analysis (MRA), spectral distortion, structural consistency, super resolution, unsupervised learning.**

I. Introduction

GIVEN the ever-increasing number of satellites acquiring images of the Earth, data fusion is becoming a key asset in remote sensing, enabling cross-sensor [1], [2], cross-resolution [3], or cross-temporal [4] analysis and information extraction. Due to technological constraints, many Earth observation systems, such as GeoEye, Plaiades, or WorldView, acquire a single full-resolution panchromatic band (PAN), responsible for the preservation of geometric information, along with a multispectral (MS) image at lower spatial resolution, with rich spectral information. A multiresolution fusion process, called pansharpening, is then employed to estimate a full-resolution MS image from the original PAN and MS components [3], [5].

Pansharpening is a challenging task, object of intense research for three decades but still far from being solved, also because of the continuously increasing resolutions at which new generation satellites operate. Several approaches and a large number of methods have been proposed over the years.

In the component substitution (CS) approach [6], the MS image is transformed in a suitable domain, one of its components is replaced by the spatially rich PAN, and the image is transformed back in the original domain. If only three bands are concerned, the intensity–hue–saturation (IHS) transform can be used, with the intensity component replaced by the panchromatic band [7]. The method is generalized in [8] (GIHS) to handle a larger number of bands. Many other transforms have been considered for CS, including principal component analysis [9], Brovey transform [10], and Gram–Schmidt (GS) decomposition [11]. More recently, adaptive CS methods have also been introduced, such as the advanced versions of GIHS and GS [12], the partial replacement CS (PRACS) method [13], or the band-dependent spatial detail (BDSD) injection method and its variants [14]–[16].

With the multiresolution analysis (MRA) approach [17], instead, pansharpening is addressed from the spatial perspective. These methods extract the high-frequency spatial details through a multiresolution decomposition, such as decimated or undecimated wavelet transforms [17]–[20], Laplacian pyramids [5], [21]–[24], or other nonseparable transforms, e.g., contourlet [25]. Extracted details are then properly injected into the resized MS component.

A further set of methods address the pansharpening problem through the variational optimization (VO) of suitable acquisition or representation models. In [26], the optimization functional involves the degradation filters mapping high-resolution to low-resolution images, whereas work [27] focuses on the sparse representations of injected details. Palsson *et al.* [28]–[30] proposed several methods of

Manuscript received September 21, 2021; revised November 9, 2021, January 22, 2022, and February 23, 2022; accepted March 22, 2022. Date of publication March 31, 2022; date of current version April 19, 2022. *(Corresponding author: Giuseppe Scarpa.)*

Matteo Ciotola, Antonio Mazza, Giovanni Poggi, and Giuseppe Scarpa are with the Department of Electrical Engineering and Information Technology, University of Naples Federico II, 80125 Naples, Italy (e-mail: matteo.ciotola@unina.it; antonio.mazza@unina.it; poggi@unina.it; giscarpa@unina.it).

Sergio Vitale is with the Department of Science and Technology, University of Naples Parthenope, 80143 Naples, Italy (e-mail: sergio.vitale@uniparthenope.it).

Digital Object Identifier 10.1109/TGRS.2022.3163887

this class, using a total variation regularized least square formulation, defining a maximum *a posteriori* problem, and, very recently, looking for low-rank representations of the joint PAN–MS pair organized in a suitable matrix, respectively. Other methods do not fit the above categories and can be roughly classified as statistical [31]–[35], dictionary-based [36]–[41], or matrix factorization approaches [42]–[44]. The reader is referred to [3] for a more comprehensive review.

In recent years, a paradigm shift from model-based to data-driven approaches has revolutionized all fields of image processing, from computer vision [45]–[49] to remote sensing [50]–[53]. In pansharpening, the first method based on convolutional neural networks (CNN) was proposed by Masi *et al.* [54], after which many more followed in a few years' span [50], [55]–[66]. It seems safe to say that deep learning is currently the most popular approach for pansharpening. Nonetheless, it suffers from a major problem: the lack of ground-truth data for supervised training. In fact, multiresolution sensors can only provide the original MS-PAN data, downgraded in space or spectrum, never their high-resolution versions, which remain to be estimated.

A widespread solution to this problem is to perform a resolution shift. The PAN–MS data undergo a downsampling process, after which they are used as input samples to train a network where the original MS data play the role of ground truth. By doing so, the network is trained in a fully supervised manner, although in a low-resolution domain. Eventually, it will be used for pansharpening the original data. Therefore, this solution relies on a sort of scale-invariance assumption: a network trained at low resolution is expected to work equally well at high resolution that this hypothesis holds up, however, is by no means obvious.

In the literature, this problem is well known [67] and, in fact, great attention is devoted to mimic the sensor modulation transfer functions (MTFs) to ensure correct downgrading of data. Even with an ideal scaling process, however, an inherent information gap exists between scales. For example, objects whose typical size amounts to a few pixels at the original resolution will necessarily lose their shape when brought at low resolution. There is no hope that a network trained at reduced resolution will "experience" such tiny geometries. Not surprisingly, networks trained with this approach work very well on reduced-resolution data but show significant quality losses on full-resolution target data [50], [54], [55], [68]. Interestingly, these problems have often been overlooked precisely because, in the absence of ground truth, it is not possible to objectively measure the performance at target resolution.

A further limit of deep learning-based pansharpening is the endemic scarcity of remote sensing training data. Due to the high cost of multiresolution data, networks are usually trained on just a few images, which, however large, cannot ensure an adequate diversity in terms of geographical position, territorial conformation, atmospheric conditions, acquisition geometry, direction and intensity of light, and so on. As a consequence, such networks will hardly generalize to images acquired by sensors not seen in training or even just to different-looking images.

Motivated by these considerations, in this article, we propose a new framework for training pansharpening models in the high-resolution domain. Networks are trained using the original PAN–MS pairs as input, at their native resolution, with no downgrading and hence no loss of information. To obviate the lack of a ground truth, a new *ad hoc* loss is defined, which weights suitably defined indicators of spatial and spectral consistency. These indicators are computed by comparing the pansharpened output with the original PAN and MS components in their respective domains. In addition, to ensure correct operations on images with the most diverse characteristics, notwithstanding the limited datasets available for training, we use the target-adaptive modality proposed originally in [68], which fine-tunes the network on the fly to the target image. Finally, it is worth underlining that the proposed learning framework is fully general and can be used for any deep learning-based pansharpening model. Experiments with three state-of-the-art CNN-based pansharpening models on images acquired by different multiresolution sensors demonstrate the broad and seamless applicability of this framework, as well as the significant quality improvements ensured by high-resolution training.

In summary, the main innovative contribution of this work is the proposal of a new fully unsupervised framework, which allows training deep learning-based pansharpening models at high resolution. To validate the proposal, we retrain several state-of-the-art methods in the new framework and carry out a wide range of experiments on images acquired by several sensors. Moreover, to ensure research reproducibility, we publish online a user-friendly software package for high-resolution training and testing of pansharpening networks, together with several trained models.[1]

The rest of this article is organized as follows. In Section II, we account for related work. Section III describes the proposed full-resolution training framework. Section IV presents the experimental result, and finally, Section V draws conclusions.

II. Related Work

In recent years, there has been a growing awareness that the resolution-shift approach to training pansharpening networks has inherent weaknesses and may cause a performance cap. Starting in 2020, several papers have begun to address this issue and to propose new solutions that carry out training, at least partially, in the high-resolution domain.

The first of these papers [69], to the best of our knowledge, has been proposed by some of the authors of the present work. Training is carried out in fully supervised modality with a loss that includes both reduced- and full-resolution terms. At low resolution, the resolution-shift approach is used, with the original MS acting as ground truth. At high resolution, instead, the output of the MTF-GLP-HPM model-based algorithm [22] takes the role of ground truth. Indeed, this algorithm is known to ensure a very good preservation of high-resolution details, which justifies using it as a proxy of the unknown ground truth for the only purpose of improving spatial quality. Needless to say, spatial accuracy cannot be better than that of the auxiliary algorithm, certainly nonoptimal. An enhanced

[1] GitHub repository: https://github.com/matciotola/Z-PNN

version of the method was later proposed in [70], with spatial loss terms relying also on the preservation of spatial gradients. Eventually, both versions provide only minor improvements with respect to methods relying on reduced-resolution learning schemes. A further development [71] concerns the fusion of high- and low-resolution spectral bands in Sentinel-2 images, a closely related task.

In [72], a rather complex residual network is proposed, trained at high resolution. Features extracted from the PAN are used in a sequence of fusion units to refine the high-pass details extracted from the upsampled MS. The loss includes spatial and spectral terms, compounding both Euclidean norm and structural similarity (SSIM), together with a term depending on a no-reference quality index. Despite the stated goal of overcoming the resolution-shift approach, these loss terms depend heavily on cross-scale consistency indexes, thereby reintroducing a sort of scale invariance assumption. In addition, an MS-to-PAN operator is used (called it $\mathcal{G}$, in Section III) which combines linearly the MS bands through coefficients estimated, again, at low resolution. Experimental results seem promising, but training and test data come from the same scene and do not allow to test generalization ability.

A deep CNN, called UPSNet, comprising 28 residual blocks plus two adaptation blocks, is proposed in [73]. Loss terms are computed exclusively on high-resolution data, with spatial accuracy pursued by working on the PAN gradients. However, they depend again on some ill-defined pieces of information, such as "grayed" or upsampled versions of the MS. To make up for errors originated by such grayed MS, a further loss is introduced, which, however, involves also nondifferentiable functions. Despite these shortcomings, good quality pansharpened images are obtained, although a bit oversmoothed.

A group of recent papers on this topic rely on generative adversarial networks (GANs). Indeed, GANs seem to fit very well the pansharpening task. The generator may be charged with the task of producing the high-resolution output starting from the available PAN and MS, while two dedicated discriminators validate the quality of results by comparing the panchromatic and low-resolution projections of the output with the original counterparts. None of these processes require a resolution shift. PanGAN [74], PercepPAN [75], and PGMAN [76] all follow this approach, with minor variations. However, despite the elegant formulation, results turn out to be much below expectations, with visible spectral aberrations (PanGAN and PercepPAN) or spatial blurring (PGMAN). Arguably, such poor results may be due to seemingly minor inaccuracies that disrupt the delicate training process of GANs. Such inaccuracies include the use of arbitrary MS-to-PAN linear projections with coefficients estimated on unrepresentative data and imperfect MS interpolation.

Despite their obvious value, these contributions present some common limits and flaws.

1) They concern individual pansharpening methods trained at high resolution, not a general training framework.
2) They rely heavily on potentially detrimental cross-scale processing steps, such as arbitrary forms of interpolation or decimation, or MS-to-PAN conversions.

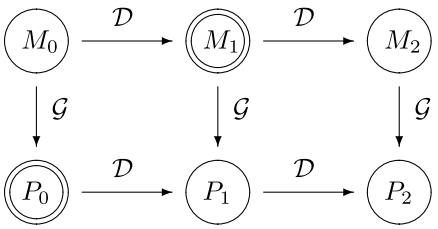


Fig. 1. Images and scales involved in the pansharpening process. The only available pieces of information are the full-resolution panchromatic image, P_0, and the low-resolution MS image, M_1, from which the target high-resolution MS image, M_0, is estimated. Deterministic (only partially known) operators, $\mathcal{D}$ and $\mathcal{G}$, relate images with their spatially or spectrally downgraded versions.

3) They generalize poorly to images with characteristic not seen in training, especially if acquired by sensors not represented in the dataset.
4) Methods and results are hardly reproducible due to the lack of software code online.

On the contrary, we propose a high-resolution training *framework*, applicable to any deep learning-based network, even if designed originally for reduced-resolution training. We minimize cross-scale processing, limited to a single downsizing step for loss computation. Correct operations on the most diverse images are ensured by the target-adaptive modality. Finally, we make all our software available online to allow easy reproduction of results and easy development of further improvements.

III. Proposed Full-Resolution Training Framework

In the following, we will use M and P, respectively, to denote MS and panchromatic images. A subscript will indicate their spatial scale, with 0 associated with the highest resolution, and a fixed resolution ratio R between scales n and $n+1$, for each n. The relationship between these images is shown in Fig. 1 where it is also assumed that low-resolution images can be obtained from their higher resolution versions through a deterministic operator, $\mathcal{D}$, and panchromatic images from the corresponding MS ones through another operator, $\mathcal{G}$. This assumption holds with good approximation for the downscaling operator, $\mathcal{D}$, while MS-to-PAN operators, though often used in applications, are necessarily far from ideal because of the sensors' physical characteristics. Of course, such operators imply a loss of information and hence are not invertible.

In multiresolution remote sensing, M_1 and P_0 are the only available pieces of information (the MS-PAN pair), and in fact, the goal of pansharpening is to estimate the unknown high-resolution MS image M_0 from these spatially and spectrally degraded images

$$\widehat{M}_0 = \phi_0(M_1, P_0). \tag{1}$$

In deep learning-based pansharpening, the estimator ϕ_0 is learned from a suitable collection of training data. This would be a standard task if complete training data were available, that is, for each training input pair (M_1^i, P_0^i), the corresponding desired output M_0^i was also provided. However, this is not the case that no full-resolution MS images are available to be used as ground truth.

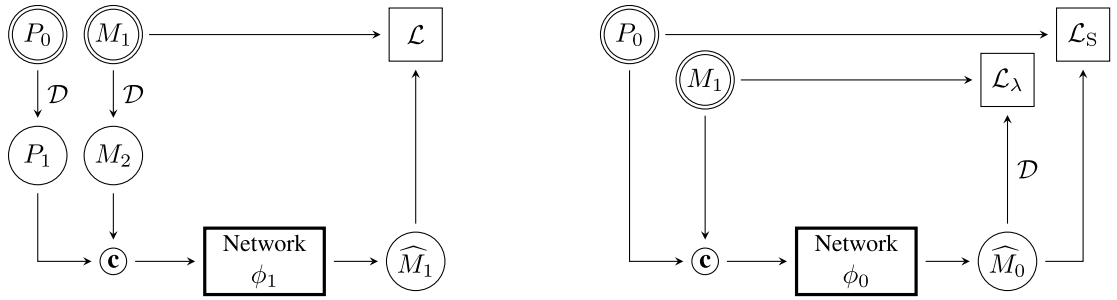


Fig. 2. (Left) Wald-like and (Right) proposed training frameworks. In the first case, training takes place in the reduced-resolution domain, MS and PAN are immediately downgraded, and the latter is not used to compute the loss. In the proposed framework, only the original high-resolution PAN and MS are used for training, and they are both used to compute the loss.

Most deep learning-based pansharpening methods proposed thus far [50], [54], [56], [68] have circumvented this problem by means of a domain shift approach, known as Wald's protocol [67], which allows to assess their synthesis ability. All available images in the dataset are downscaled to the next lower resolution level

$$M_2^i = \mathcal{D}\big(M_1^i\big), \quad P_1^i = \mathcal{D}\big(P_0^i\big). \tag{2}$$

For these pairs, the original MS images, M_1^i, represent the perfectly known ground truth. Therefore, a conventional training procedure can be used to estimate the network weights, that is, the pansharpening function ϕ_1. This function is eventually used to perform pansharpening at the original scale. A block diagram of this training procedure is shown in the left of Fig. 2.

Of course, underlying this approach is the assumption that the same network can operate equally well at low resolution and high resolution, that is, $\phi_1 \simeq \phi_0$. This is a convenient approximation, but experimental evidence accumulated over the years proves it to be largely inaccurate. Networks trained under Wald's resolution downgrading protocol work very well on the low-resolution images they have been trained for, but only fairly well [77] on the full-resolution images. In practice, there is a significant domain mismatch between low- and high-resolution pansharpening.

Before proposing our alternative training framework, let us justify intuitively the unsatisfactory behavior of the resolution-shift solution. The fundamental observation is that the network, during the entire training process, never sees the full-resolution data. In particular, the panchromatic images, the only data available at full resolution, are immediately resized, causing an irrecoverable loss of information. To fully realize the importance of this loss, one should also keep in mind that these images are acquired at a fixed resolution. For example, all panchromatic images provided by the WorldView-3 sensor have a spatial resolution of 0.31 m. At this resolution, a number of small urban objects, such as cars and traffic signs, are fully characterized with well-defined geometric shapes. With the help of low-resolution spectral information, they can be accurately recovered. However, with the resolution-shift approach, the network sees only images of much lower resolution, 1.24 m (with 4.96-m MS) where these tiny objects loose completely their shape, reducing to a very few pixels or even subpixels. Contrary to what happens in super-resolution, there is no 8-cm resolution WorldView-3 image available to make up for this loss of information.

An additional problem is that resized images are much smaller than the original ones, providing much less data for training. Sticking to the WorldView-3 example, at low resolution, there are 16 times less pixels than at full resolution. Considering the scarcity of training data, due to the restrictive policies of most data providers, this turns out to be a nonnegligible drawback.

These considerations, together with experimental results much below expectations, motivate our proposal of a full-resolution training framework. We will train pansharpening networks using the original data, thereby including full-resolution panchromatic images. Clearly, we must do without the ground-truth images, which do not exist. Therefore, the cornerstone of our proposal is the definition of a new loss function that takes the role of the conventional full-reference loss.

Since we lack the full-resolution reference, M_0, we use the next most valuable pieces of information, that is, its projections on the low-resolution and panchromatic domains, M_1 and P_0. The network output $\widehat{M}_0$ is compared with these two references, in their respective domains, to ensure spectral and spatial consistency. Accordingly, the proposed loss becomes

$$\mathcal{L}\Big(M_1, P_0; \widehat{M}_0\Big) = \mathcal{L}_\lambda\Big(M_1; \mathcal{D}\Big(\widehat{M}_0\Big)\Big) + \beta\mathcal{L}_S\Big(P_0; \mathcal{G}\Big(\widehat{M}_0\Big)\Big) \tag{3}$$

with β a suitable parameter that weighs the spectral and spatial loss terms.

Fig. 3 shows our approach geometrically. The target image M_0 (red dot) is regarded as the combination of its M_1 and P_0 projections plus a third unknown component (call it U), which cannot be explained by neither of the former two. By minimizing the loss of (3), we are pushing the estimate $\widehat{M}_0$ (black dot) toward the projections of M_0 on the (M_1, P_0) plane. The origin of the third component has been critically explored in [78], comparing alternative perspectives and assumptions. Our working hypothesis is that this unpredictable part is indeed small and, therefore, our final estimate will be very close to the actual image. It is left to the experimental results to say the final word in favor or against this hypothesis. At the very least, with our approach, we are not discarding any relevant data in the training process.

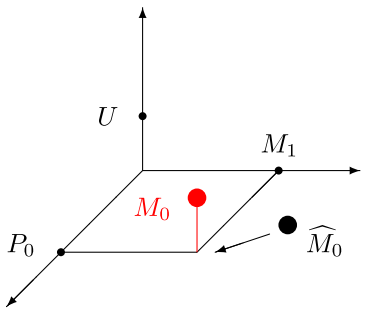


Fig. 3. Graphical sketch of the ideal proposed approach. Lacking ground truth, the pansharpening process aims at generating an image, $\widehat{M}_0$, whose projections coincide with the known original data, P_0 and M_1. An unknown residual, U, orthogonal to this plane, remains unaccounted for. Our conjecture is that this latter component is small.

In the practical implementation, we depart slightly from the elegant symmetric formulation of (3). Indeed, while the $\mathcal{D}$ operator can be reasonably assumed to be known, such that $M_1 = \mathcal{D}(M_0)$, there is no consensus in the literature on the exact form and even on the conceptual correctness of the $\mathcal{G}$ operator. To circumvent this problem, this operator is bypassed here, and the spatial loss term is computed as the sum of B individual contributions, one for each spectral band of $\widehat{M}_0$. Synthetically, the proposed loss reads as

$$\mathcal{L}\left(M_1, P_0; \widehat{M}_0\right) = \mathcal{L}_\lambda\left(M_1; \mathcal{D}\left(\widehat{M}_0\right)\right) + \beta\mathcal{L}_S\left(P_0; \widehat{M}_0\right). \quad (4)$$

A block diagram of the proposed training procedure is shown in Fig. 2, next to the Wald-like training scheme with resolution downgrading, for easy comparison. Visual inspection provides an immediate appreciation of the fundamental changes.

1) In the Wald-like framework, P_0 is immediately downgraded and never used further; therefore, high-resolution information is lost forever.
2) In the proposed framework, instead, an additional spatial loss term $\mathcal{L}_S$ is introduced to take advantage of the information conveyed by the PAN.
3) In the proposed framework, the only resolution downgrade takes place after pansharpening and only for the purpose of comparison with the original MS.

In the following, we describe in detail the spectral and spatial loss terms.

A. Spatial Loss

The role of the spatial loss is to inject in the pansharpened image the high-resolution structures observed in the PAN. Accordingly, the PAN can be used to perform a prediction, necessarily imperfect, of the output image bands, and preferably a linear prediction, lacking any reasons to prefer more complex solutions. Following this point of view, here, we define the spatial loss term as a function of the correlation coefficient between the PAN and the spectral bands of the output image.

Let X and Y be two equal-size single-band images, and let σ_X^2, σ_Y^2, and σ_{XY} indicate their sample variances and covariance. Then, the correlation coefficient between X and Y is defined as

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}, \quad -1 \le \rho_{XY} \le 1. \quad (5)$$

The correlation coefficient indicates to what extent one image can be linearly predicted from the other, with $|\rho| = 1$ implying perfect predictability and $\rho = 0$ total incorrelation.

Now, we expect to find in the pansharpened bands mostly the same spatial layout of the PAN and, therefore, a strong correlation with it. However, to preserve the spectral information, such a correlation cannot be unitary. Actually, it can be expected to vary spatially and from band to band, as a function of the observed scene. For example, in vegetated areas, we expect the PAN to have a strong correlation with the "green" band of the output and a weaker correlation with other bands, while the opposite will happen in other regions. In rare cases, even negative correlations are observed, due to local contrast inversions between the PAN and some MS bands [78]. This leads us to consider a 3-D spatial–spectral correlation field rather than a single coefficient. Thus, in (5), let X be a square patch of size $\sigma \times \sigma$ extracted from the PAN at spatial location (i, j) and Y be the corresponding patch extracted from band b of $\widehat{M}_0$; then, we obtain the 3-D correlation field

$$\rho^\sigma(i, j, b) \triangleq \rho^\sigma_{P, \widehat{M}_0(b)}(i, j) \quad (6)$$

which depends on spatial coordinates (i, j), spectral coordinate b, and size parameter σ.

Now, we could think of defining a local spatial loss as

$$\ell^\sigma(i, j, b) = 1 - \rho^\sigma(i, j, b), \quad 0 \le \ell \le 2 \quad (7)$$

and the global spatial loss term as its average. However, by doing so, we would neglect the inherent spatial–spectral variability of the correlation mentioned above and push it uniformly toward 1. Therefore, to address this problem, we define an auxiliary reference correlation field, $\rho^{\sigma,\text{ref}}(i, j, b)$, computed between a low-pass filtered version of the PAN and an expanded version (plain interpolation) of the MS and define the local loss as

$$\ell^\sigma(i, j, b) = \begin{cases} 1 - \rho^\sigma(i, j, b), & \rho^\sigma < \rho^{\sigma,\text{ref}} \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

The reference correlation field can be computed exactly from the available data and provides a rough approximation of the target correlation field. A positive loss $\ell^\sigma = 1 - \rho^\sigma$ is incurred at site (i, j, b) whenever the local correlation is too small, forcing the output band to follow the spatial layout of the PAN. When ρ^σ exceeds the reference value $\rho^{\sigma,\text{ref}}$, however, there is no further contribution to the global loss, and the network is free to optimize the output based on other inputs.

Although the use of correlation is certainly not new in pansharpening, we point out that our approach is very different from what encountered in conventional methods. CS, for example, relies on the strong assumption of a perfect global correlation between the pansharpened MS bands and the PAN [78]. When this assumption is violated, especially in the presence of occultation or contrast inversion phenomena, strong spectral aberrations are observed. In some traditional injection-based methods [79], instead, local correlation is used

just to exert a consistency check. Injection of PAN details takes place only when the local correlation is high, switching to a plain upsampling of the MS bands otherwise. We assume a generally large local correlation between PAN and MS, but verify our hypothesis on the reference field, $\rho^{\sigma,\text{ref}}$, and leverage deep learning with a suitable loss to exploit this dependence.

B. Spectral Loss

The spectral loss is computed in a straightforward manner by comparing the low-resolution projection of the pansharpened image, $\mathcal{D}(\widehat{M}_0)$, with its natural reference M_1

$$\mathcal{L}_\lambda = \left\| \mathcal{D}\left(\widehat{M}_0\right) - M_1 \right\|_1 \tag{9}$$

where $\|\cdot\|_1$ indicates the ℓ_1-norm.

As already said, the low-resolution projection operator $\mathcal{D}$ has been widely studied in the literature and can be assumed to be known. It consists of band-dependent low-pass filtering followed by spatial decimation at pace R

$$\mathcal{D}(M_n(\cdot,\cdot,b)) = [M_n(\cdot,\cdot,b) * h_b] \downarrow R. \tag{10}$$

Under this assumption, $\mathcal{L}_\lambda$ can be expected to completely vanish in the presence of correct pansharpening, $\widehat{M}_0 = M_0$, a property not always satisfied by other quality indicators [80].

However, this is really the case only if the original spectral bands are correctly aligned; otherwise, a coregistration step is required. Indeed, in multiresolution imagery, the MS bands are often misaligned. This is due to technological constraints of the sensing systems and may also depend on the specific product released. As a result, spectral aberrations appear in the image, easily spotted in false-color representations as thin lines with weird colors near object boundaries. Therefore, it is good practice to coregister the MS spectral bands beforehand, a step often neglected by researcher and practitioners alike. Interestingly, in the proposed framework, bands are automatically coregistering. In fact, to maximize their spatial correlation with the PAN, they are eventually aligned with it and hence among themselves. This good thing, however, has a perverse effect. After decimation, in fact, the well-aligned low-resolution projection will be compared with a misaligned reference, generating a nonzero loss even in the presence of a perfect output. However, this problem is readily solved. The band-to-PAN shifts resulting after the fine-tuning are used in the decimation step to realign $\mathcal{D}(\widehat{M}_0)$ with M_1.

In the proposed loss function of (4), two critical hyperparameters must be set: the patch size σ used in the spatial loss term and the weight β that balances spatial and spectral losses. In Section IV-E, we describe and discuss the preliminary experiments carried out to select the values of σ and β used in our implementation.

C. Target-Adaptive Operating Modality

Remote sensing images present a large variability, due to the portrayed scene, the sensor characteristics, the acquisition conditions, and so on. Even a large and well-designed dataset could hardly capture this wide variety, but the training sets used in practical applications consist often of just one or a few (large) images, often acquired by the same sensor. This is mostly due to the high cost of multiresolution images and the scarcity of data freely available for the research community. Understandably, models trained in these conditions work poorly on new off-training images. To address this problem, target-adaptive pansharpening was proposed in [68]. This operating modality (see Fig. 4) consists in unfreezing the network weights, $\phi^{(0)}$, and performing a few cycles of fine-tuning to the target image, using some selected samples extracted by the target image itself. With a sensible choice of parameters, only a limited increase in complexity is incurred. On the positive side, the generalization ability improves sharply, with performance gains that may be also very significant, depending on training-test mismatch. We, therefore, regard target adaptation as an essential ingredient for real-world pansharpening methods and an integral part of the proposed framework. At test time, the user is only asked to provide/select the pretrained network, and then, the algorithm runs a few iterations of fine-tuning to optimize the weights for the target image, before carrying out the actual pansharpening using the updated parameters, $\phi^{(\infty)}$. The default number of tuning iterations was set to 50 in [68]. Here, we raise it to 100, based on the experimental results discussed in Section IV-C.

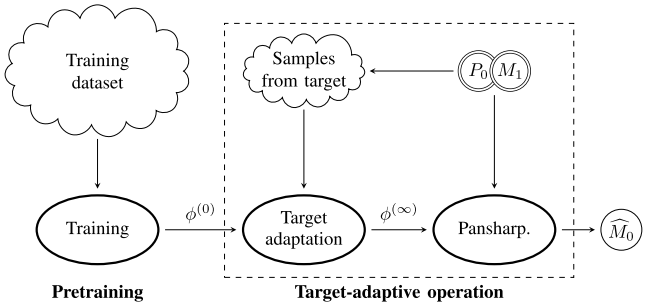


Fig. 4. High-level flowchart of target-adaptive pansharpening.

IV. Experimental Analysis

A. Reference Methods, Datasets, and Performance Measures

1) Comparative Methods: For all comparative analyses, we rely on the benchmark toolbox [77], which implements a large number of methods belonging to the four main categories recalled in Section I: CS, MRA, VO, and ML. All methods available in the toolbox are used in the experiments, except for a few VO solutions that suffer software compatibility issues. In addition, we consider two more state-of-the-art ML methods, PanNet [50] and DRPNN [56], retrained on our datasets to ensure a fair comparison.

2) Datasets: Table I lists the datasets used for training, validation, and fine-tuning of the deep learning-based models and for testing of all methods. In particular, three satellites have been considered, WorldView-3 (WV3), WorldView-2 (WV2), and GeoEye-1 (GE1). In some cases, we use baseline models pretrained on other datasets detailed in the reference papers.

3) Performance Measures: Assessing the performance of pansharpening methods is an open issue, given the lack of full-resolution ground truths. A widespread approach is to measure performance objectively in a reduced-resolution setting. Popular indexes used to this end are spectral angle

TABLE I

DATASETS. GSD: PAN GROUND SAMPLE DISTANCE AT NADIR (m). PAN/MS RESOLUTION RATIO, $R = 4$. ADELAIDE AND WASHINGTON, COURTESY OF DIGITALGLOBE. MEXICO CITY, NAPLES, WATERFORD, AND GENOA (DIGITALGLOBE) PROVIDED BY ESA

Sensor-site	# tiles	PAN size	GSD	Usage
WV3-Mexico City	1	2048×2048	0.31	Training
WV3-Mexico City	2	2048×2048	0.31	Validation
WV3-Adelaide	10	2048×2048	0.31	Testing
WV2-Napoli	1	2048×2048	0.46	Training
WV2-Napoli	2	2048×2048	0.46	Validation
WV2-Washington	13	2048×2048	0.46	Testing
GE1-Waterford	1	2048×2048	0.41	Training
GE1-Waterford	2	2048×2048	0.41	Validation
GE1-Genova	10	2048×2048	0.41	Testing

mapper (SAM), Erreur Relative Globale Adimensionnelle de Synthése (ERGAS), and $Q2^n$ [multiband extension of the universal image quality index (UIQI)] [81]–[83], also provided in the benchmark toolbox [77]. However, this approach is at odds with our goals and is not followed here.

Instead, we consider full-resolution no-reference indexes, which assess separately spectral and spatial fidelity. Many such indexes have been proposed in recent years toward this end, for example, [84]–[86]. For spectral fidelity, we consider here the spectral distortion index, $D_\lambda^{(\mathrm{K})}$, proposed by Khan *et al.* [87], in the slightly modified implementation of the assessment toolbox [77], together with the reprojection indexes, R-SAM, R-ERGAS, and R-$Q2^n$, proposed in [80]. Note that R-$Q2^n$ equals 1-$D_\lambda^{(\mathrm{K})}$ if the latter is implemented as originally proposed. For spatial fidelity, instead, we consider the spatial distortion index, D_S, proposed in [88], and the correlation distortion index, D_ρ, also proposed in [80]. Unlike for the spectral case, these two indexes have a deeply different rationale and sometimes provide contrasting results. In particular, experiments carried out in [80] show that D_ρ correlates better than D_S with experts' visual assessment, especially for high-quality pansharpening.

B. Does Full-Resolution Training Improve Performance?

The aim of this section is to prove that the proposed full-resolution training framework does indeed improve the performance of deep learning-based pansharpening, as measured by full-resolution quality indexes and especially visual inspection. Toward this end, we consider three state-of-the-art networks, PanNet [50], DRPNN [56], and A-PNN [68], a variant of PNN [54] with a skip connection for residual learning. For each network, we consider three versions. First of all, the basic model originally trained by the authors using losses based on L_1-norm (A-PNN) or L_2-norm (the others). By doing so, we have a solid starting point, the network optimized by the authors on their own data and available online. Then, we add two target-adaptive versions, with adaptation carried out at reduced resolution, with the Wald-like approach (model-TA), or at full resolution, with the proposed framework (model-TA-FR). Unlike in normal operations, where only a few iterations are used to save time, we use a large number of iterations here, 2000, to ensure a very good adaptation to the target image.

TABLE II

FOR EACH MODEL ∈ {A-PNN, PANNET, DRPNN}, WE CONSIDER SEVERAL VERSIONS, DIFFERING IN PRETRAINING AND TARGET ADAPTATION. Z-PNN IS A PROPOSED PNN VARIANT LATER DETAILED (SECTION IV-C)

full acronym	pretraining		target-adaptation		
	dataset	resolution	applied	resolution	# iter.
model	authors'	reduced	no	–	–
*model**	ours	reduced	no	–	–
model-TA	authors'	reduced	yes	reduced	2000
model-TA-FR	authors'	reduced	yes	full	2000
Z-PNN (0 iter.)	ours	full	no	–	–
Z-PNN	ours	full	yes	full	100

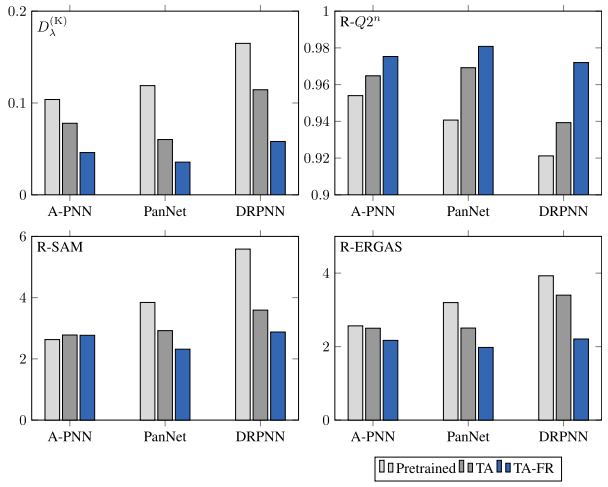


Fig. 5. Full-resolution spectral accuracy indexes for Adelaide.

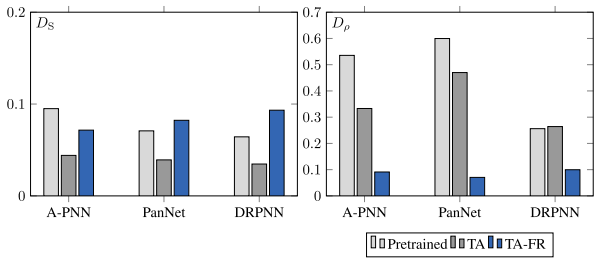


Fig. 6. Full-resolution spatial accuracy indexes for Adelaide.

This allows the network to "forget" the initial parameters, removing possible biases due to the different pretraining conditions. At this point, differences in performance will depend only on the architecture and, for each architecture, on the use of the low- or high-resolution training framework. Table II summarizes these models and variants.[2]

Figs. 5 and 6 report spectral and spatial quality indexes, respectively, obtained for the WorldView-3 Adelaide test image. Similar results are obtained with different test images.

A first observation concerns the significant performance gaps observed between different pretrained models (light gray bins). For example, A-PNN has an R-SAM index about half

[2]In the table, we also include models used in subsequent experiments, that is, the versions retrained on our datasets (marked by an asterisk), and Z-PNN, a PNN variant proposed here (Section IV-C). We warn the reader that the toolbox [77] uses a slightly different acronym [advanced PNN with fine-tuning (A-PNN-FT)] to indicate the reduced-resolution target-adaptive A-PNN [68] that we name here A-PNN-TA.

that of DRPNN. However, such differences depend more on the limited generalization ability of the methods than on their intrinsic effectiveness. A-PNN was originally trained on data well-aligned with our WorldView-3 test image, something that probably did not happen with DRPNN. This interpretation is strongly supported by results obtained with the TA models (dark gray bins). In fact, with target adaptation, the performance improves almost always significantly, and the quality indexes become much more uniform across the three methods. Overall, target adaptation mitigates the mismatch between training and test set and the resulting indexes can be regarded as more reliable indicators of the actual potential of the various pansharpening tools.

We now turn to the real objective of our analysis, the performance obtained with target adaptation at high resolution (blue) to be compared with that obtained at low resolution (dark gray). Regarding spectral quality, a significant gain is observed for all methods over all indexes (again, with minor exceptions), and the performance appears to be even more uniform than before. For spatial quality, instead, results are more controversial. While the D_ρ index drops, suggesting a large quality improvement, the D_S index grows again, indicating a spatial accuracy comparable to that of pretrained models. Two facts motivate this strong mismatch. On one hand, we argue that D_S is not really a reliable indicator when quality is very high. Indeed, as also noted in [80], D_S does not really measure spatial quality, but rather a sort of cross-scale spatial quality consistency. Thus, it may be small even in the presence of strong spatial distortion, provided that the same distortion occurs across the various scales of interest, and it may be large even with perfect pansharpening, $\widehat{M}_0 = M_0$. On the other hand, since the spatial loss used in our training framework follows closely the definition of D_ρ, this indicator may be biased in favor of TA-FR methods. Since such contradictions cannot be reconciled, we will keep using both indicators, leaving the final say to visual inspection.

In Fig. 7, for some crops of the Adelaide test image, we show the original MS and PAN data together with the output pansharpened images obtained with the pretrained, TA, and TA-FR versions of the three CNN-based methods. Since we are interested in comparing training schemes against one another, not architectures, we show different crops for different architectures so as to offer a richer yet compact picture. First of all, visual inspection fully confirms the improvements in terms of spectral accuracy ensured by target adaptation. With respect to pretrained models, colors are better preserved and some evident errors are avoided. In addition, the TA-FR solutions seem to ensure clear improvements also in terms of spatial accuracy. Some strange patterns created by pretrained or TA networks disappear. Small objects (e.g., cars) are reconstructed with higher accuracy and, in general, all contours are sharper. High-frequency textures observed in the PAN are preserved (sometimes, even oversharpened). Overall, we see a huge improvement with respect to the pretrained models, as predicted by D_ρ and also a consistent improvement with respect to the TA versions. While further work is certainly necessary to obtain fully satisfactory pansharpening, we believe that these results represent convincing indications that high-resolution training is the right path to follow.

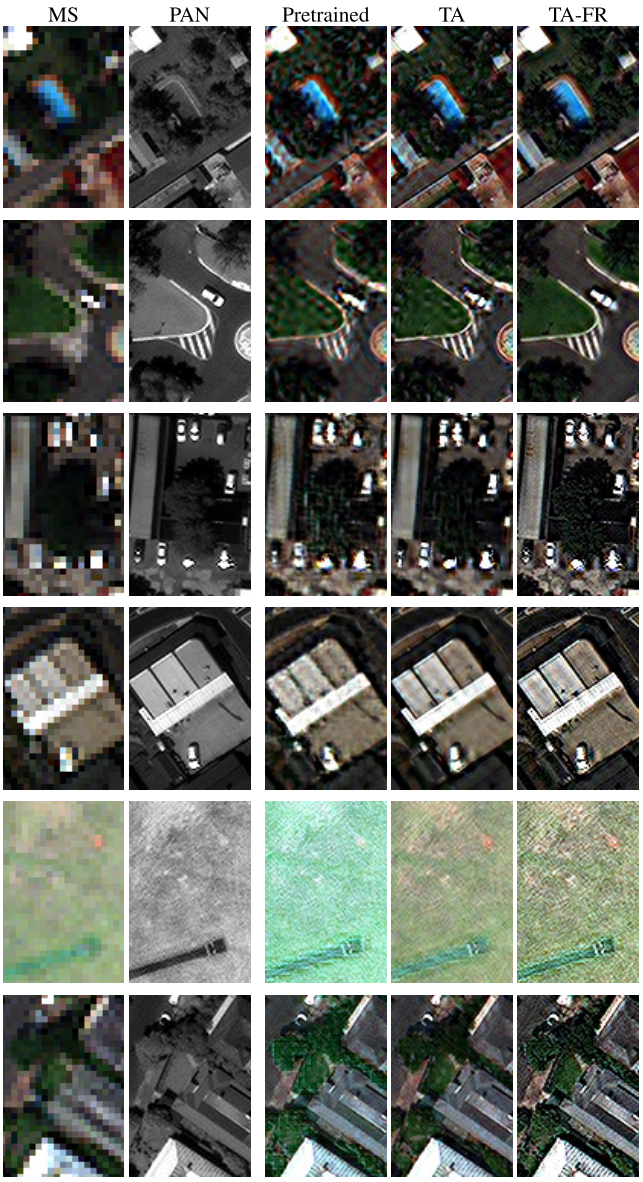


Fig. 7. Results on crops from the Adelaide image. (From Left to Right) MS, PAN, Pretrained, TA, and TA-FR. (From Top to Bottom) A-PNN (rows 1 and 2), PanNet (rows 3 and 4), and DRPNN (rows 5 and 6). Red, green, and blue bands are used for RGB composition.

C. Z-PNN: A CNN-Based Pansharpening Method Pretrained at Full Resolution

The analysis of Section IV-B sheds light on the potential of high-resolution training. However, it relies on intensive target adaptation, which has nonnegligible costs in terms of both memory and time. Such costs are summarized in Table III for the three models analyzed thus far, considering a 2048 × 2048-pixel multiresolution image and an NVIDIA Quadro P6000 GPU. In practice, 2000 iterations require 1-h processing time or more. This was not the case with the target-adaptive method proposed in [68], as it worked on much smaller (16×) low-resolution images and used only 50 iterations.

TABLE III

COMPUTATION TIME (PER ITERATION) AND MEMORY REQUIREMENTS TO PERFORM TARGET ADAPTATION ON A 2048 × 2048 WV3 IMAGE

	Time [seconds]			GPU Memory [GB]		
	A-PNN	PanNet	DRPNN	A-PNN	PanNet	DRPNN
TA	0.862	0.346	0.615	0.38	0.74	2.03
TA-FR	1.885	1.885	11.172	8.96	12.84	24.09

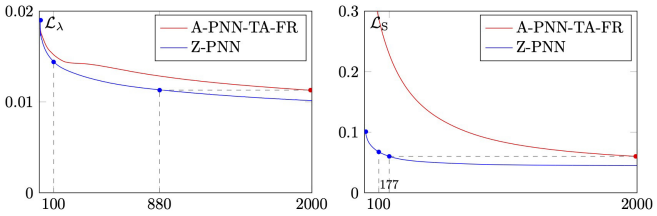


Fig. 8. (Left) Spectral and (Right) spatial losses versus number of iterations for adapting A-PNN-TA-FR (red lines) and Z-PNN (blue lines) to the target image.

To obtain fast high-quality pansharpening, we refined the original model weights through a further pretraining phase carried out at full resolution, using a dedicated training image for each sensor (see again Table I). By doing so, we expect that much fewer iterations will be necessary for target adaptation. Since all three architectures appear to perform equally well, from now on, we focus only on the simplest one, A-PNN. The resulting network will be referred to as Z-PNN, short for Zoom-PNN. We test the impact of this modification on an off-training WorldView-3 image. Fig. 8 shows the progress of spectral and spatial loss terms, as adaptation proceeds, for the versions without (A-PNN-TA-FR) and with (Z-PNN) this further pretraining phase. The right part, concerning the spatial loss, is especially telling. While the A-PNN-TA-FR curve lowers very slowly, reaching eventually the value $\mathcal{L}_S \simeq 0.06$ after 2000 iterations, the Z-PNN curve reaches the same value after less than 200 iterations. Actually, the spatial loss is quite low from the beginning, $\mathcal{L}_S \simeq 0.10$, ensuring a good performance even in the absence of any adaptation. The left figure, instead, shows that the spectral loss benefits from fine-tuning also when starting from the Z-PNN weights. In any case, a small number of iterations seem to be sufficient to observe a significant improvement. Fig. 9 shows, for two crops of the test image, the evolution of the pansharpened output as adaptation goes on. The images fully confirm all previous observations. In summary, it appears that Z-PNN could be safely used even without adaptation or with just a few iterations. In the following experiments, we set conservatively the number of iterations to 100. However, the user is free to change this value depending on both available resources and quality target.

D. Comparative Analysis

We can now move to a full-fledged comparative analysis. Experiments will be carried out on test images acquired by three different sensors (WV2, WV3, and GE1), listed in Table I, and the results will be compared with those of the reference methods summarized in Table IV.

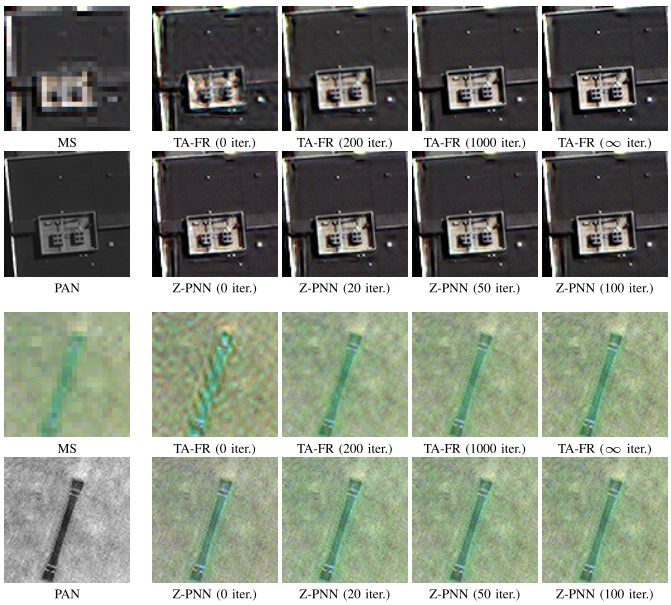


Fig. 9. Impact of target adaptation with increasing iterations on image quality for (A-PNN-)TA-FR (odd rows) and Z-PNN (even rows) for two WV3 crops. (Left) MS and PAN. Z-PNN reaches a satisfactory quality long before A-PNN-TA-FR.

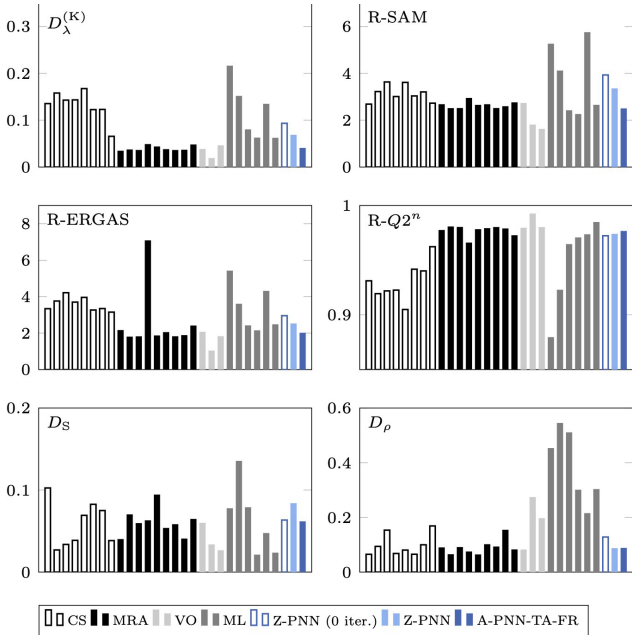


Fig. 10. Numerical results on the WV3-Adelaide dataset.

Numerical results are shown in the bar graphs of Figs. 10–12, for the WV3, WV2, and GE1 datasets, respectively. Each graph refers to a different full-resolution measure, and each bar refers to a different method. The reference methods are those listed in Table IV, grouped according to their approach (CS, MRA, VO, and ML), and shown with a different bar style for each group. Newly developed methods, Z-PNN without (0) and with (100 iterations) target adaptation and A-PNN-TA-FR, are shown in shades of blue at the end. Note that PanNet and DRPNN have been retrained on our dataset to ensure a fairer comparison, an asterisk marks this version.

TABLE IV
DETAILED LIST OF ALL REFERENCE METHODS

Component Substitution (CS)
BT-H [89], BDSD [14], C-BDSD [15], BDSD-PC [16], GS [11], GSA [12], C-GSA [24], PRACS [13]
Multiresolution Analysis (MRA)
AWLP [90], MTF-GLP [90], MTF-GLP-FS [91], MTF-GLP-HPM [90], MTF-GLP-HPM-H [89], MTF-GLP-HPM-R [92], MTF-GLP-CBD [93], C-MTF-GLP-CBD [24], MF [94]
Variational Optimization (VO)
FE-HPM [26], SR-D [27], TV [28]
Machine Learning (ML)
PNN [54], PNN-IDX [54], A-PNN [68], A-PNN-TA [68], DRPNN* [56], PanNet* [50]

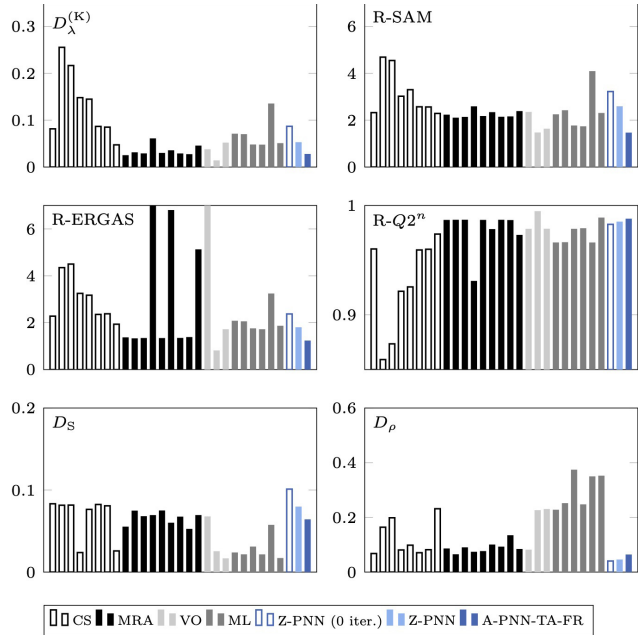


Fig. 11. Numerical results on the WV2-Washington dataset.

To begin, let us focus on Fig. 10, also because similar considerations, with minor differences, hold for the other cases. The most notable outcome is that, contrary to the widespread belief, ML methods do not outperform conventional methods (smaller is best for all measures but R-$Q2^n$). As an example, the MRA methods (black) are generally[3] superior to ML methods (dark gray) in terms of both spectral quality and spatial quality indicators. This surprising result is due, in our opinion, to the low-resolution versus high-resolution mismatch. ML methods are usually trained at low resolution, with the Wald-like protocol of Fig. 2 (left), and then tested again with the Wald protocol. Therefore, it is not surprising that numerical results speak largely in their favor. Visual analyses on full-resolution data, however, have always casted some shadows on the superiority of ML methods. Such doubts are confirmed here, where results are computed only in terms of high-resolution indexes. These provide a more unbiased assessment of performance and are better predictors of the quality of pansharpened images the end users can expect.

[3]Note that individual methods show occasional failures on some images, we neglect these special cases in this high-level analysis.

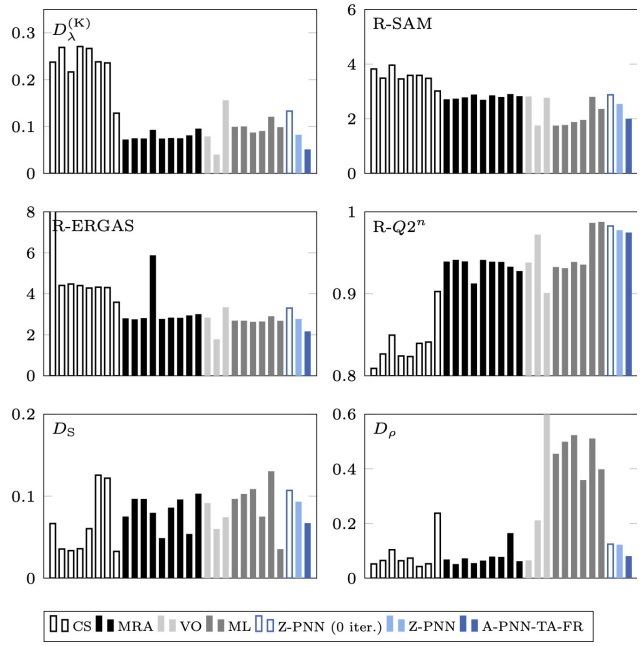


Fig. 12. Numerical results on the GE1-Amsterdam dataset.

The performance of ML methods improves significantly only when training takes place at high resolution, as proved by the last three (blue) bars. This behavior is observed, more or less pronounced, with all sensors, see Figs. 11 and 12, and the improvement is especially significant in terms of spatial quality, according to the D_ρ indicator. In particular, the fully (2000 iterations) adapted method, A-PNN-TA-FR (last bar), has one of the smallest D_ρ values consistently on all datasets. Moreover, it has also very good spectral quality indicators, suggesting an excellent overall performance. On the other hand, it is fair to underline that D_S results depict a very different situation, almost opposite to D_ρ. Again, this calls for accurate visual inspection of pansharpened images, which is the next step of our analysis.

Figs. 13–15 show the visual results for some crops acquired by the WV2, WV3 and GE1, respectively. For each crop, next to the original MS and PAN, we show the output of two methods trained at high resolution (A-PNN-TA-FR and Z-PNN), together with six reference methods. The latter are chosen as the best and second best ranking methods in terms of D_ρ, D_S, and $D_\lambda^{(K)}$.

Let us consider Fig. 13, first, and compare the A-PNN-TA-FR with the original PAN–MS pair. By suitably enlarging the figure, one can fully appreciate the impressive spatial quality of the result. All details are faithfully preserved with their original shapes and textures, and no alien pattern is introduced by the pansharpening process. Spectral quality is also very good, but this property is shared with several other methods. Z-PNN also provides very good results, and we only observe a minor loss of spectral accuracy. Continuing along the row, MTF-GLP-HPM-H and MTF-GLP are the best reference methods in terms of D_ρ, and in fact, we observe a very good spatial fidelity also for them. Things are very different, instead, for A-PNN-TA and PanNet*, the best methods according to D_S. Besides a reduced precision on object shapes and

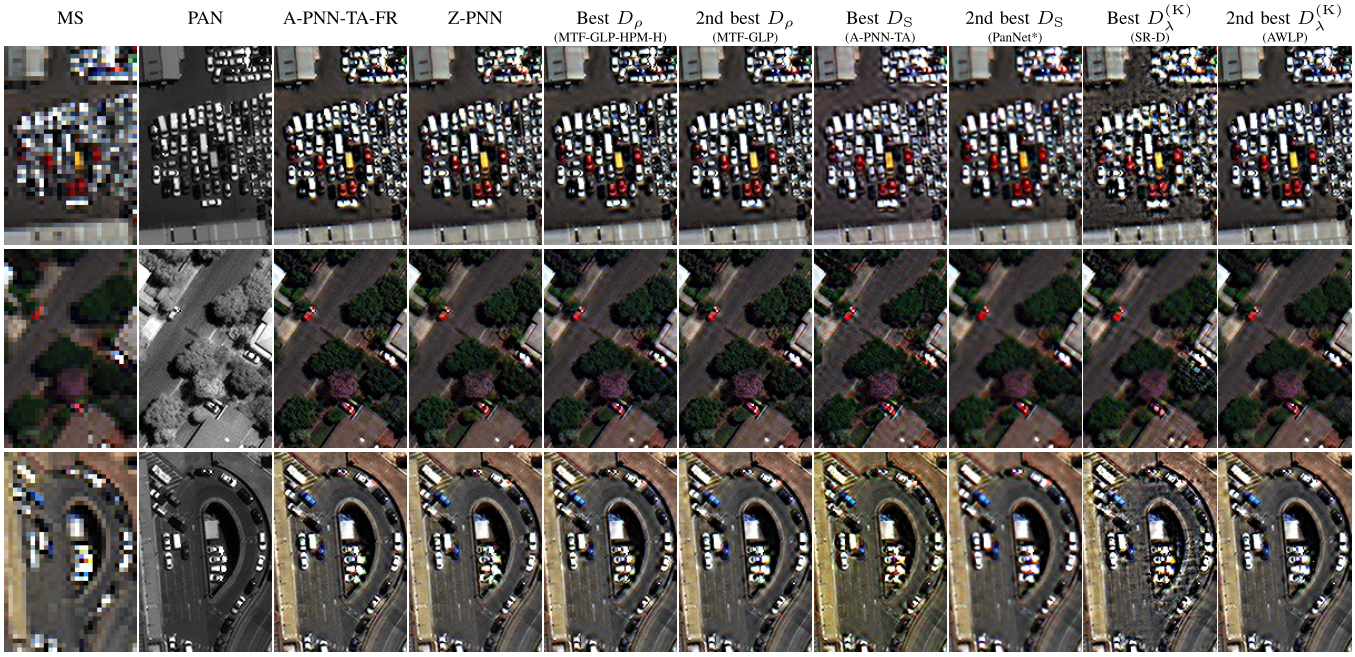


Fig. 13. Results for some WV3 crops. (From Left to Right) MS, PAN, A-PNN-TA-FR, and Z-PNN, best references for D_ρ, D_S, and $D_\lambda^{(K)}$.

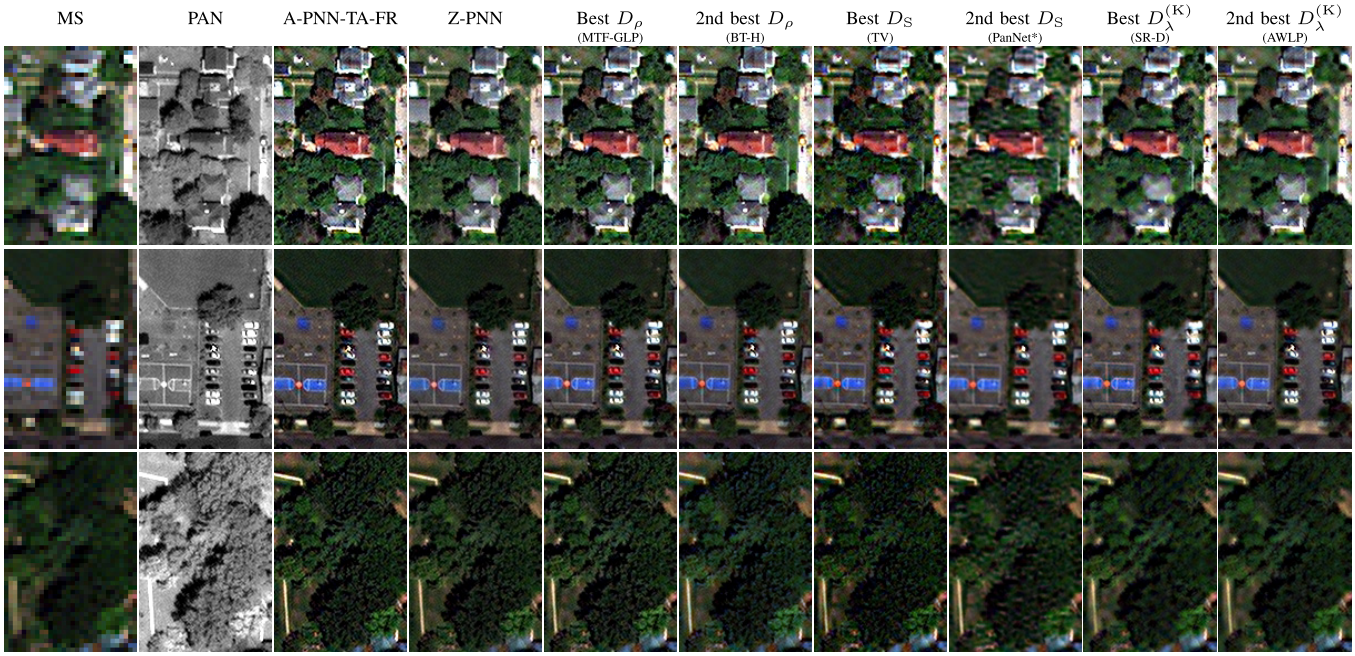


Fig. 14. Results for some WV2 crops. (From Left to right) MS, PAN, A-PNN-TA-FR, and Z-PNN, best references for D_ρ, D_S, and $D_\lambda^{(K)}$.

some loss of resolution, especially for PanNet*, we observe annoying periodic patterns over the whole scene, confirming that D_S cannot be considered a fully reliable predictor of spatial fidelity. Finally, the best methods in terms of $D_\lambda^{(K)}$, SR-D and AWLP, ensure indeed a good spectral quality, although comparable to that of other methods, but exhibit some problems in terms of spatial fidelity.

Figs. 14 and 15 show similar results for the WV2 and GE1 images, respectively. Beyond minor differences, the same phenomena described before are observed in all cases. A-PNN-TA-FR and Z-PNN keep providing very good results, especially in terms of spatial quality, only rarely matched by other methods, typically those performing best in terms of D_ρ.

E. Setting Loss Hyperparameters: Testing Alternative Losses

In all previous experiments, we used the loss of (4) with hyperparameters σ and β optimized experimentally. Here, we discuss their impact on the performance and motivate experimentally the values selected in the implementation. In addition, we test an alternative loss function proposed in the literature for use in our framework.

1) Setting σ: The patch size σ is the only critical parameter of the proposed spatial loss. We already motivated the need

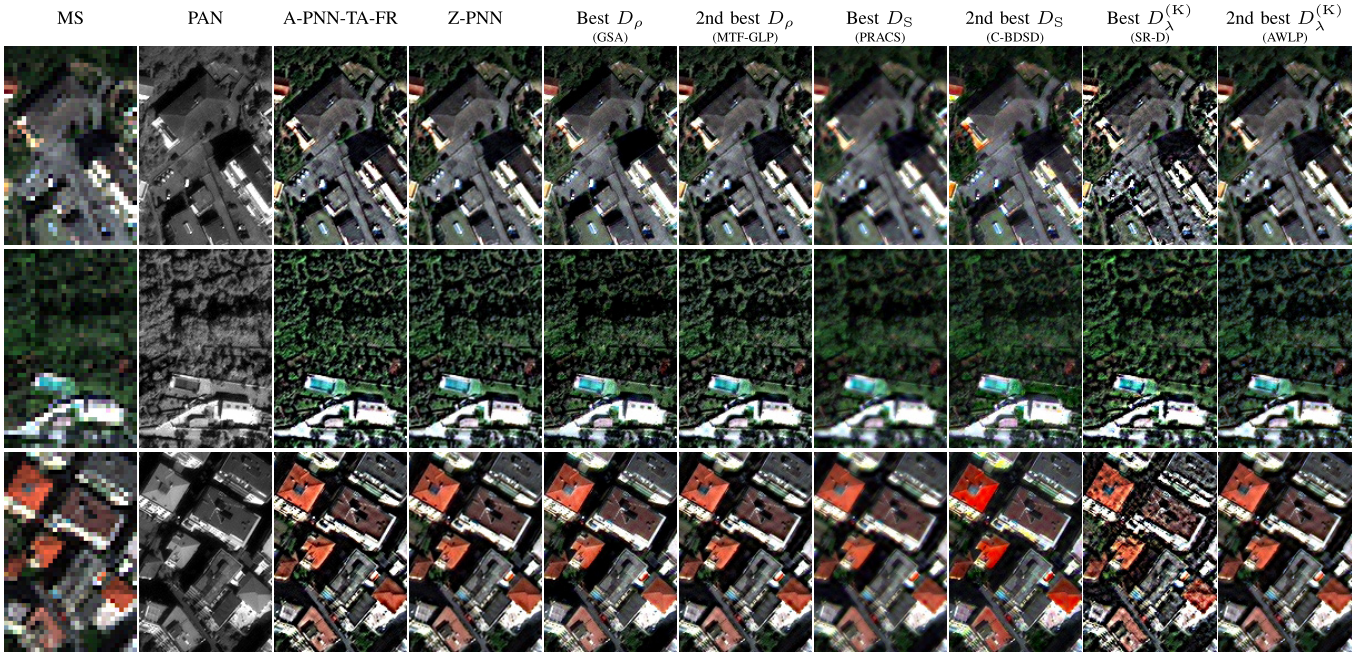


Fig. 15. Results for some GE1 crops. (From Left to right) MS, PAN, A-PNN-TA-FR, and Z-PNN, best references for D_ρ, D_S, and $D_\lambda^{(K)}$.

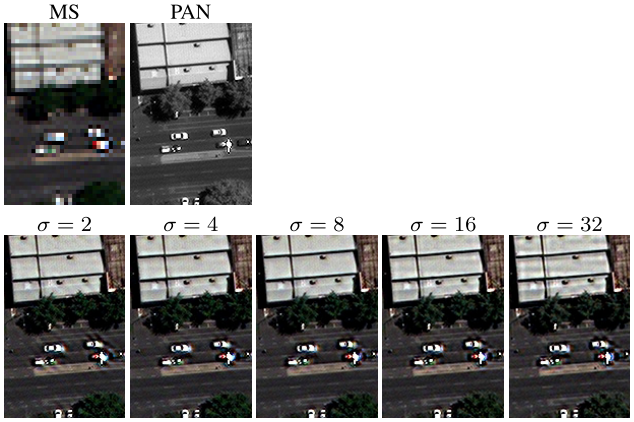


Fig. 16. Impact of patch size (σ) on pansharpening quality.

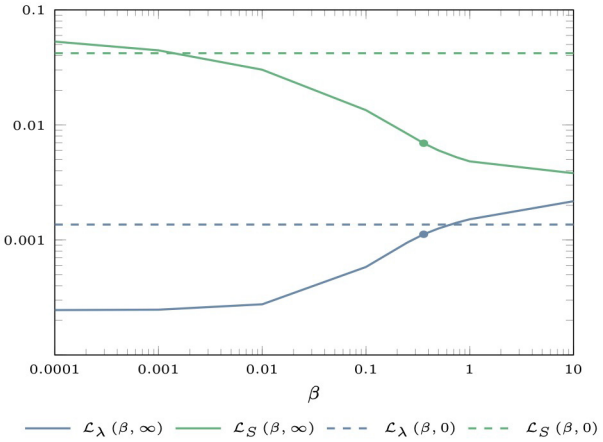


Fig. 17. Spatial and spectral losses as a function of β.

to estimate the MS-PAN correlation on a local as opposed to global scale (small σ), thereby limiting long-range spatial dependencies and preserving spectral fidelity. On the other hand, with a very small value for σ, the correlation ends up being estimated on just a few points. Lacking any more precise theoretical guidance, we carried out a number of experiments on test images with σ doubling progressively from 2 to 32. A sample result is shown in Fig. 16. With $\sigma = 2$, obvious artifacts are visible on the roads, in the form of diagonal patterns. These disappear already with $\sigma = 4$ and then $\sigma = 8$, and however, for larger values, other spectral checkerboard aberrations appear on the building rooftops, such as echoes of the existing black separation lines. We observed a similar behavior on many more test images, which suggests choosing small values for σ, between 4 and 8. Eventually, we set σ equal to the resolution ratio, R, which is always 4 for our images.

2) Setting β: The parameter β balances the relative importance of the spatial and spectral loss terms. When $\beta = 0$, only the spectral loss is considered, which negatively affects the spatial quality, and the opposite happens when $\beta \to \infty$. To quantify this behavior, Fig. 17 reports the values of the spatial and spectral loss terms observed for A-PNN-TA-FR when β grows from 0.0001 to 10. There is a large range of values where both losses (solid lines) decrease with respect to the case without adaptation (dashed lines). Thus, to gain a better insight, we resort again to visual inspection for a sample test image. In Fig. 18, we show the original MS (enlarged) and PAN, in the first row, the pansharpened outputs for various values of β, in the second row, and the difference between the former and an interpolated version of the MS, in the third row. For $\beta = 0.01$ and even 0.1, the output images appear blurred, with an insufficient spatial quality. For $\beta = 10$, instead, and to a lesser extent also for $\beta = 1$, there are color distortions on the vegetation and other details, especially visible in the difference images. A good compromise is obtained with values between

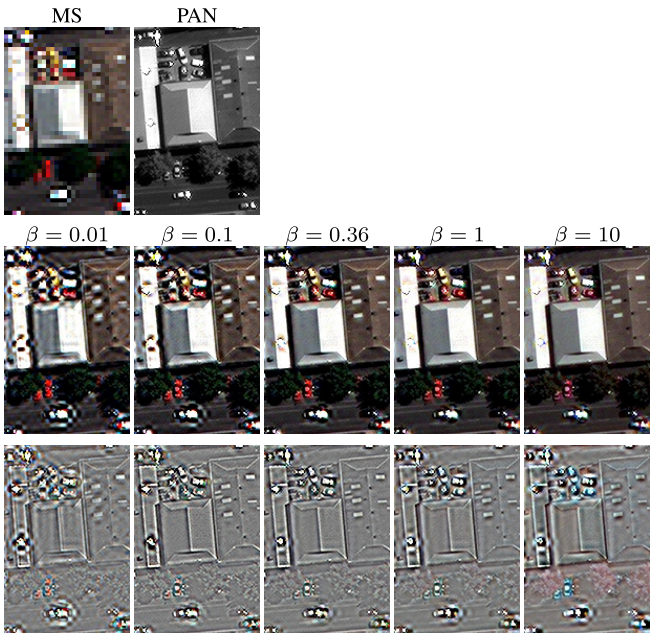


Fig. 18. Impact of loss balance (β) on pansharpening quality.

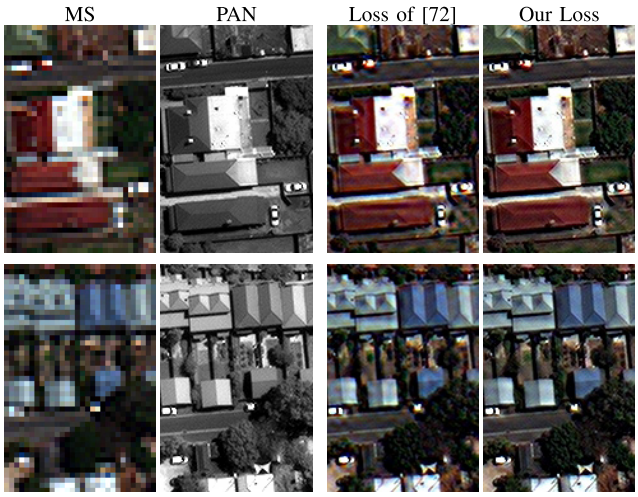


Fig. 19. Comparing the loss of [72] with the proposed loss.

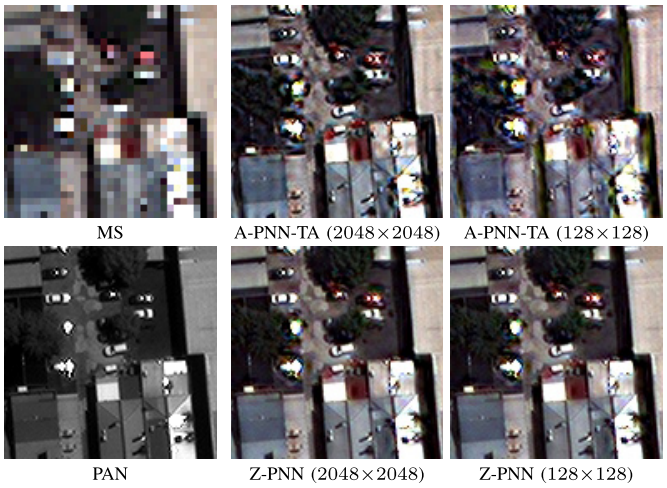


Fig. 20. For Z-PNN, the quality of fine-tuning does not appreciably depend on the size of the target scene, (Middle) 2048 × 2048 or (right) 128 × 128. Therefore, it can be used to "zoom" in real time on any detail of interest.

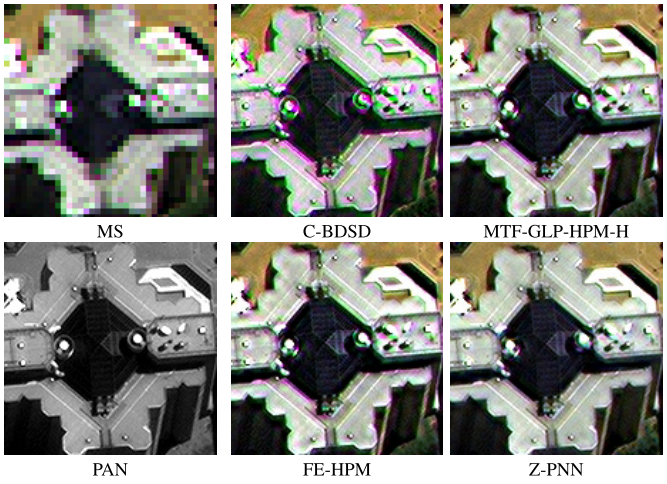


Fig. 21. In the red–yellow–blue display, the effects of MS bands misalignment are highlighted. Spurious green or magenta lines appear along object borders in all pansharpened images except Z-PNNs, where this issue is automatically addressed.

0.1 and 1, and in fact, we selected eventually $\beta = 0.25$ for GeoEye and $\beta = 0.36$ for WorldView.

3) Testing an Alternative Loss: Our system has been conceived based on a clear rationale, discussed in Section III, and our training loss was designed to fulfill it. Nonetheless, one can legitimately wonder what happens if different losses are used in the same framework. Thus, we fine-tuned the A-PNN-TA-FR model replacing our loss with a very different one, recently proposed [72] for high-resolution training.

This latter, call it $\mathcal{L}'$, comprises four terms

$$\mathcal{L}' = \mathcal{L}'_\lambda + \mathcal{L}'_S + \mathcal{L}'_{\text{QNR}} + \lambda \|\Theta\|_2^2. \tag{11}$$

The first two aim at improving spectral and spatial quality by minimizing combinations of MSE and SSIM indexes in the upscaled MS and panchromatic domains. Instead, the third term directly targets the QNR [3] a well-known full-resolution quality measure, while the last one, the weighted norm of the parameters, serves only for regularization. The reader is referred to [72] for all details.

Sample pansharpened crops are shown in Fig. 19, next to the original MS and PAN, for two samples for the WorldView3 Adelaide image. The spectral fidelity is quite good in both cases, although slightly better indexes are obtained with our loss, $D_\lambda^{(\text{K})} = 0.03$ as opposed to 0.05. When considering spatial fidelity, however, an obvious performance gap appears. Images pansharpened with our loss are much sharper, fine textures and small details are much better preserved, as obvious from the comparison with the PAN, and no spurious pattern is generated in the process.

F. Strengths and Weaknesses of the Proposed Framework

The above discussed results make clear what the main strengths of the proposed framework are. By using the original PAN–MS pairs to train a deep learning model, we make sure that the most informative data are considered and lay the basis for obtaining high spectral and spatial fidelity in

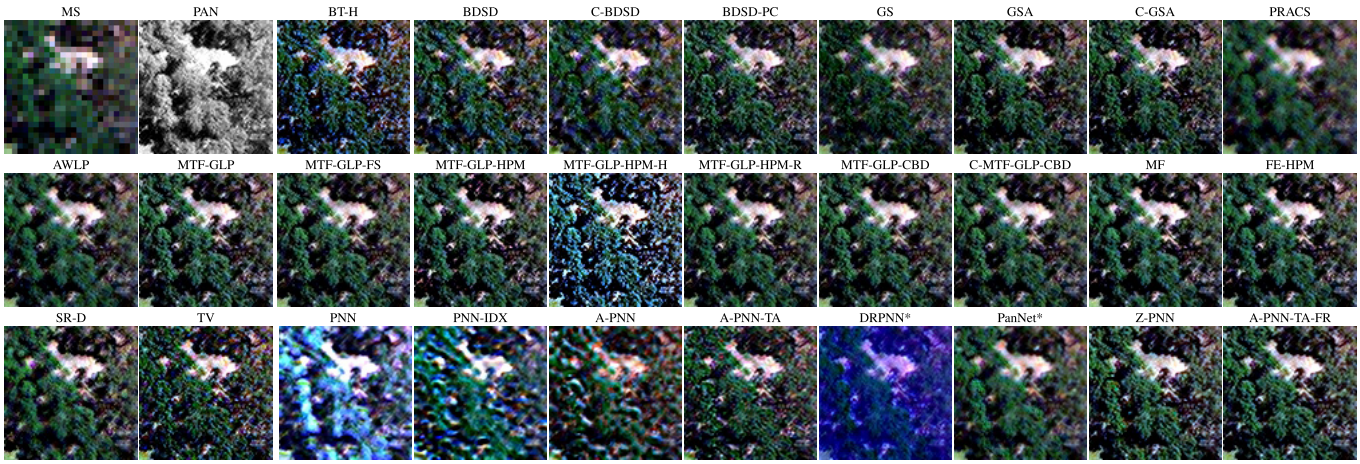


Fig. 22. Results of all methods on a small WV3 vegetation crop. Most methods, including ML methods trained at low resolution, show chromatic aberrations and resolution loss. ML methods trained at high resolution ensure high spatial and spectral fidelity.

pansharpening. Results obtained with A-PNN, PanNet, and DRPNN are just examples of the potential of this approach. On the downside, working at high resolution incurs costs. Using the original data, without subsampling, causes pretraining to become much slower and memory intensive, a nuisance, but not a major problem, considering that pretraining takes place off-line. On the other hand, target adaptation is important to ensure the best performance, and this process takes place online. For Z-PNN and 100 iterations, it requires about 3 min, as shown in Table III. Depending on application and mode of use, this may be overly annoying. With the following experiment, however, we show that this cost may be significantly reduced.

Fig. 20 shows, on the left column, the original PAN–MS pair for a 128 $\times$ 128 pixel WV3 crop. In the middle column, we see the output of A-PNN-TA on the top and Z-PNN on the bottom, both adapted on the 2048 $\times$ 2048-pixel target image including our crop, displaying the by-now usual quality gap. Our focus, though, is on the right column. Here, adaptation is carried out only on the very same 128 $\times$ 128 pixel target crop, not the whole image, hence, using much less data and computing time. While the quality of the A-PNN-TA image further degrades, likely for the lack of sufficient data, this is not the case for the Z-PNN image, which is almost indistinguishable from the previous case. As this behavior is observed consistently in our experiments, we conclude that Z-PNN can be safely fine-tuned on the very same scene of interest, even very small, providing stable and high-quality results. Needless to say, this comes at a fraction of the original computational cost, just about 1 s in our example. Therefore, one can use Z-PNN in this modality to "zoom" on the details of interest, each time upgrading the original Z-PNN output (already good) in a matter of seconds.

Another critical point regards the different speed of adaptation of the spectral and spatial loss terms (see Fig. 8). Since the latter improves much faster than the former, Z-PNN and A-PNN-TA-FR turn out to have a very similar spatial score (D_ρ) but, in some cases, a nonnegligible gap in terms of spectral score ($D_\lambda^{(\mathrm{K})}$, R-SAM). This may call for a longer adaptation phase in the presence of very strict spectral accuracy requirements.

A valuable strength of the proposed framework is the automatic coregistration of pansharpened spectral bands. To better appreciate this feature, in Fig. 21, we show, for another WV3 150 $\times$ 150 pixel crop, the input PAN–MS pair and the output images generated by Z-PNN and some reference methods where the coregistration problem is not addressed. This time, however, we use an unusual red–yellow–blue false-color representation. In fact, while the red, green, and blue bands are usually well aligned, other bands, such as the yellow one, may be slightly shifted, due to the imaging system that acquires subsets of bands in slightly different time intervals. As expected, severe color distortions are visible in all the output images except for Z-PNN, where spectral fidelity remains high also near sharp boundaries.

To conclude this section, in Fig. 22, we show the output of all reference methods for a single small vegetation crop. Vegetation is extremely common in multiresolution imagery, but its correct pansharpening is often prohibitive due to the presence of fine textures at multiple scales. This is confirmed by the results in the figure. Apart from some methods that present a clear failure (e.g., DRPNN*), many more provide disappointing results, with large chromatic aberrations and/or a significant loss of detail. In general, MRA methods perform quite well on this image, much better than CS and VO. Also, ML methods trained at low resolution are among the worst in this task. Instead, due to the spatial loss based on local correlation, A-PNN-TA-FR and Z-PNN, trained with our high-resolution framework, show again a very good performance, preserving faithfully even the most subtle vegetation textures.

G. Implementation Details

All experiments were run on a server equipped with Nvidia Quadro P6000 GPU with 24-GB memory, and all networks were implemented in PyTorch. Some of the tested CNN models, i.e., Z-PNN, PanNet*, and DRPNN*, needed a pretraining phase. For Z-PNN, as stated in Section IV-C, the model weights have been produced as refinement of the original A-PNN model parameters [68], using a dedicated training image for each sensor, as indicated in Table I. The whole

image is used as a one-sample batch, running 2000 iterations that involve all layers, with a learning rate of 10^{-5} on WV2/3 and 5×10^{-5} on GeoEye-1, and using the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$.

The models for PanNet* and DRPNN*, instead, have been reimplemented in PyTorch and trained from scratch on our training datasets for all three sensors, using the same hyperparameters (learning rate, optimizer, loss, epochs, and so on) of the original versions. For these models, however, since the training occurs in the reduced-resolution domain, we used a 4×4 times larger tile, hence 8192×8192 pixels, to compensate for data volume reduction.

V. Conclusion

We have proposed a framework for full-resolution training of pansharpening models, with the aim of exploiting all the information carried by the original data, with no resolution downgrading. Lacking a ground truth, we defined a suitable compound loss, with two components accounting separately for spectral and spatial fidelity. We used the proposed framework to train several state-of-the-art pansharpening models. Experimental results are extremely encouraging. Besides numerical indicators, visual inspection confirms that the quality of the pansharpened images is largely improved due to high-resolution training. Beyond the framework itself and the trained pansharpening methods, though, the main contribution of this work is to prove the potential of this training approach. Many improvements are certainly possible, and we hope to stimulate research on this topic. We are currently working on a refined spatial loss component.

References

[1] M. Gargiulo, A. Mazza, R. Gaetano, G. Ruello, and G. Scarpa, "A CNN-based fusion method for super-resolution of Sentinel-2 data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 4713–4716.

[2] A. Errico *et al.*, "SAR/multispectral image fusion for the detection of environmental hazards with a GIS," in *Proc. SPIE*, vol. 9245, Oct. 2014, Art. no. 924503.

[3] G. Vivone *et al.*, "A critical comparison among pansharpening algorithms," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2565–2586, May 2015.

[4] R. Gaetano *et al.*, "Exploration of multitemporal COSMO-SkyMed data via interactive tree-structured MRF segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 7, pp. 2763–2775, Jul. 2014.

[5] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, "MTF-tailored multiscale fusion of high-resolution MS and PAN imagery," *Photogramm. Eng. Remote Sens.*, vol. 72, no. 5, pp. 591–596, May 2006.

[6] V. K. Shettigara, "A generalized component substitution technique for spatial enhancement of multispectral images using a higher resolution data set," *Photogram. Eng. Remote Sens.*, vol. 58, no. 5, pp. 561–567, 1992.

[7] T. M. Tu, S. C. Su, H. C. Shyu, and P. S. Huang, "A new look at IHS-like image fusion methods," *Inf. Fusion*, vol. 2, pp. 177–186, May 2001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1566253501000367

[8] T. Tu, "A fast intensity hue-saturation fusion technique with spectral adjustment for IKONOS imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 1, pp. 309–312, 2004.

[9] P. S. Chavez and A. Y. Kwarteng, "Extracting spectral contrast in Landsat thematic mapper image data using selective principal component analysis," *Photogramm. Eng. Remote Sens.*, vol. 55, no. 3, pp. 339–348, 1989.

[10] A. R. Gillspie, A. B. Kahle, and R. E. Walker, "Color enhancement of highly correlated images. II. Channel ratio and 'chromaticity' transformation techniques," *Remote Sens. Environ.*, vol. 22, no. 3, pp. 343–365, 1987.

[11] C. Laben and B. Brower, "Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening," U.S. Patent 6 011 875, Jan. 4, 2000.

[12] B. Aiazzi, S. Baronti, and M. Selva, "Improving component substitution pansharpening through multivariate regression of MS+Pan data," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3230–3239, Oct. 2007.

[13] J. Choi, K. Yu, and Y. Kim, "A new adaptive component-substitution-based satellite image fusion by using partial replacement," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 1, pp. 295–309, Jan. 2011.

[14] A. Garzelli, F. Nencini, and L. Capobianco, "Optimal MMSE pan sharpening of very high resolution multispectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 1, pp. 228–236, Jan. 2008.

[15] A. Garzelli, "Pansharpening of multispectral images based on nonlocal parameter optimization," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2096–2107, Apr. 2015.

[16] G. Vivone, "Robust band-dependent spatial-detail approaches for panchromatic sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6421–6433, Sep. 2019.

[17] T. Ranchin and L. Wald, "Fusion of high spatial and spectral resolution images: The ARSIS concept and its implementation," *Photogramm. Eng. Remote Sens.*, vol. 66, no. 1, pp. 49–61, Jan. 2000.

[18] J. Nunez, X. Otazu, O. Fors, A. Prades, V. Pala, and R. Arbiol, "Multiresolution-based image fusion with additive wavelet decomposition," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 3, pp. 1204–1211, May 1999.

[19] X. Otazu, M. González-Audicana, O. Fors, and J. Núñez, "Introduction of sensor spectral response into image fusion methods. Application to wavelet-based methods," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 10, pp. 2376–2385, Oct. 2005.

[20] M. M. Khan, J. Chanussot, L. Condat, and A. Montanvert, "Indusion: Fusion of multispectral and panchromatic images using the induction scaling technique," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 1, pp. 98–102, Jan. 2008.

[21] B. Aiazzi, L. Alparone, S. Baronti, and A. Garzelli, "Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 10, pp. 2300–2312, Oct. 2002.

[22] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, "An MTF-based spectral distortion minimizing model for pan-sharpening of very high resolution multispectral images of urban areas," in *Proc. GRSS/ISPRS Joint Workshop Remote Sens. Data Fusion Urban Areas*, May 2003, pp. 90–94.

[23] J. Lee and C. Lee, "Fast and efficient panchromatic sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 1, pp. 155–163, Jan. 2010.

[24] R. Restaino, M. D. Mura, G. Vivone, and J. Chanussot, "Context-adaptive pansharpening based on image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 753–766, Feb. 2017.

[25] V. P. Shah, N. H. Younan, and R. L. King, "An efficient pan-sharpening method via a combined adaptive PCA approach and contourlets," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1323–1335, May 2008.

[26] G. Vivone *et al.*, "Pansharpening based on semiblind deconvolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 1997–2010, Apr. 2015.

[27] M. R. Vicinanza, R. Restaino, G. Vivone, M. D. Mura, and J. Chanussot, "A pansharpening method based on the sparse representation of injected details," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 1, pp. 180–184, Jan. 2015.

[28] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "A new pansharpening algorithm based on total variation," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 318–322, Jan. 2014.

[29] F. Palsson, J. R. Sveinsson, M. O. Ulfarsson, and J. A. Benediktsson, "Model-based fusion of multi- and hyperspectral images using PCA and wavelets," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2652–2663, May 2015.

[30] F. Palsson, M. O. Ulfarsson, and J. R. Sveinsson, "Model-based reduced-rank pansharpening," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 4, pp. 656–660, Apr. 2020.

[31] D. Fasbender, J. Radoux, and P. Bogaert, "Bayesian data fusion for adaptable image pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 6, pp. 1847–1857, Jun. 2008.

[32] L. Zhang, H. Shen, W. Gong, and H. Zhang, "Adjustable model-based fusion method for multispectral and panchromatic images," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 6, pp. 1693–1704, Dec. 2012.

[33] X. Meng, H. Shen, H. Li, Q. Yuan, H. Zhang, and L. Zhang, "Improving the spatial resolution of hyperspectral image using panchromatic and multispectral images: An integrated method," in *Proc. WHISPERS*, Jun. 2015, pp. 1–4.

[34] H. Shen, X. Meng, and L. Zhang, "An integrated framework for the spatio–temporal–spectral fusion of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7135–7148, Dec. 2016.

[35] S. Zhong, Y. Zhang, Y. Chen, and D. Wu, "Combining component substitution and multiresolution analysis: A novel generalized BDSD pansharpening algorithm," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 6, pp. 2867–2875, Jun. 2017.

[36] S. Li and B. Yang, "A new pan-sharpening method using a compressed sensing technique," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 2, pp. 738–746, Feb. 2011.

[37] S. Li, H. Yin, and L. Fang, "Remote sensing image fusion via sparse representations over learned dictionaries," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 9, pp. 4779–4789, Sep. 2013.

[38] X. X. Zhu and R. Bamler, "A sparse image fusion algorithm with application to pan-sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2827–2836, May 2013.

[39] M. Cheng, C. Wang, and J. Li, "Sparse representation based pansharpening using trained dictionary," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 293–297, Jan. 2014.

[40] X. X. Zhu, C. Grohnfeld, and R. Bamler, "Exploiting joint sparsity for pansharpening: The J-SparseFI algorithm," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 5, pp. 2664–2681, May 2016.

[41] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

[42] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 2, pp. 528–537, Feb. 2012.

[43] C. Lanaras, E. Baltsavias, and K. Schindler, "Hyperspectral super-resolution by coupled spectral unmixing," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3586–3594.

[44] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "CoSpace: Common subspace learning from hyperspectral-multispectral correspondences," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4349–4359, Jul. 2019.

[45] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1106–1114.

[46] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[47] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Mar. 2017, pp. 2980–2988.

[48] F. Lateef and Y. Ruichek, "Survey on semantic segmentation using deep learning techniques," *Neurocomputing*, vol. 338, pp. 321–348, Apr. 2019.

[49] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.

[50] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, "PanNet: A deep network architecture for pan-sharpening," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5449–5457.

[51] G. Scarpa, M. Gargiulo, A. Mazza, and R. Gaetano, "A CNN-based fusion method for feature extraction from sentinel data," *Remote Sens.*, vol. 10, no. 2, p. 236, Feb. 2018. [Online]. Available: http://www.mdpi.com/2072-4292/10/2/236

[52] P. Benedetti, D. Ienco, R. Gaetano, K. Ose, R. G. Pensa, and S. Dupuy, "M^3Fusion: A deep learning architecture for multiscale multimodal multitemporal satellite data fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 12, pp. 4939–4949, Dec. 2018.

[53] A. Mazza, F. Sica, P. Rizzoli, and G. Scarpa, "TanDEM-X forest mapping using convolutional neural networks," *Remote Sens.*, vol. 11, no. 24, p. 2980, Dec. 2019.

[54] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sens.*, vol. 8, no. 7, p. 594, Jul. 2016. [Online]. Available: http://www.mdpi.com/2072-4292/8/7/594

[55] Y. Wei and Q. Yuan, "Deep residual learning for remote sensed imagery pansharpening," in *Proc. RSIP*, May 2017, pp. 1–4.

[56] Y. Wei, Q. Yuan, H. Shen, and L. Zhang, "Boosting the accuracy of multispectral image pansharpening by learning a deep residual network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1795–1799, Oct. 2017.

[57] Y. Rao, L. He, and J. Zhu, "A residual convolutional neural network for pan-shaprening," in *Proc. IEEE Int. Workshop Remote Sens. Intell. Process. (RSIP)*, Sep. 2017, pp. 1–4.

[58] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "CNN-based pansharpening of multi-resolution remote-sensing images," in *Proc. Joint Urban Remote Sens. Event (JURSE)*, Mar. 2017, pp. 1–4.

[59] A. Azarang and H. Ghassemian, "A new pansharpening method using multi resolution analysis framework and deep neural networks," in *Proc. 3rd Int. Conf. Pattern Recognit. Image Anal. (IPRIA)*, Apr. 2017, pp. 1–6.

[60] Q. Yuan, Y. Wei, X. Meng, H. Shen, and L. Zhang, "A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 3, pp. 978–989, Mar. 2018.

[61] X. Liu, Y. Wang, and Q. Liu, "PSGAN: A generative adversarial network for remote sensing image pan-sharpening," in *Proc. ICIP*, Oct. 2018, pp. 873–877.

[62] Z. Shao and J. Cai, "Remote sensing image fusion with deep convolutional neural network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 5, pp. 1656–1669, May 2018.

[63] S. Vitale, G. Ferraioli, and G. Scarpa, "A CNN-based model for pansharpening of WorldView-3 images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 5108–5111.

[64] Y. Zhang, C. Liu, M. Sun, and Y. Ou, "Pan-sharpening using an efficient bidirectional pyramid network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5549–5563, Aug. 2019.

[65] W. Dong, T. Zhang, J. Qu, S. Xiao, J. Liang, and Y. Li, "Laplacian pyramid dense network for hyperspectral pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.

[66] W. Dong, S. Hou, S. Xiao, J. Qu, Q. Du, and Y. Li, "Generative dual-adversarial network with spectral fidelity and spatial enhancement for hyperspectral pansharpening," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 10, 2021, doi: 10.1109/TNNLS.2021.3084745.

[67] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolution: Assessing the quality of resulting images," *Photogramm. Eng. Remote Sens.*, vol. 63, no. 6, pp. 691–699, 1997.

[68] G. Scarpa, S. Vitale, and D. Cozzolino, "Target-adaptive CNN-based pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5443–5457, Sep. 2018.

[69] S. Vitale and G. Scarpa, "A detail-preserving cross-scale learning strategy for CNN-based pansharpening," *Remote Sens.*, vol. 12, no. 3, p. 348, Jan. 2020.

[70] S. Vitale and G. Scarpa, "A cross-scale loss for CNN-based pansharpening," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Sep. 2020, pp. 645–648.

[71] M. Ciotola, M. Ragosta, G. Poggi, and G. Scarpa, "A full-resolution training framework for Sentinel-2 image fusion," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2021, pp. 1260–1263.

[72] S. Luo, S. Zhou, Y. Feng, and J. Xie, "Pansharpening via unsupervised convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4295–4310, 2020.

[73] S. Seo *et al.*, "UPSNet: Unsupervised pan-sharpening network with registration learning between panchromatic and multi-spectral images," *IEEE Access*, vol. 8, pp. 201199–201217, 2020.

[74] J. Ma, W. Yu, C. Chen, P. Liang, X. Guo, and J. Jiang, "Pan-GAN: An unsupervised pan-sharpening method for remote sensing image fusion," *Inf. Fusion*, vol. 62, pp. 110–120, Oct. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253520302591

[75] C. Zhou, J. Zhang, J. Liu, C. Zhang, R. Fei, and S. Xu, "PercepPan: Towards unsupervised pan-sharpening based on perceptual loss," *Remote Sens.*, vol. 12, no. 14, p. 2318, Jul. 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/14/2318

[76] H. Zhou, Q. Liu, and Y. Wang, "PGMAN: An unsupervised generative multiadversarial network for pansharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6316–6327, 2021.

[77] G. Vivone *et al.*, "A new benchmark based on recent advances in multispectral pansharpening: Revisiting pansharpening with classical and emerging pansharpening methods," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 1, pp. 53–81, Mar. 2021.

[78] C. Thomas, T. Ranchin, L. Wald, and J. Chanussot, "Synthesis of multispectral images to high spatial resolution: A critical review of fusion methods based on remote sensing physics," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1301–1312, May 2008.
[79] T. Ranchin, B. Aiazzi, L. Alparone, S. Baronti, and L. Wald, "Image fusion—The ARSIS concept and some successful implementation schemes," *ISPRS J. Photogramm. Remote Sens.*, vol. 58, nos. 1–2, pp. 4–18, Jun. 2003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271603000133
[80] G. Scarpa and M. Ciotola, "Full-resolution quality assessment for pansharpening," 2021, *arXiv:2108.06144*.
[81] L. Wald, *Data Fusion: Definitions and Architectures–Fusion of Images of Different Spatial Resolutions*. Paris, France: Les Presses de lÉcole des Mines, 2002.
[82] L. Alparone, S. Baronti, A. Garzelli, and F. Nencini, "A global quality measurement of pan-sharpened multispectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 1, no. 4, pp. 313–317, Oct. 2004.
[83] A. Garzelli and F. Nencini, "Hypercomplex quality assessment of multi/hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 662–665, Oct. 2009.
[84] P. S. Pradhan, R. L. King, N. H. Younan, and D. W. Holcomb, "Estimation of the number of decomposition levels for a wavelet-based multiresolution multisensor image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 12, pp. 3674–3686, Dec. 2006.
[85] J. Zhou, D. L. Civco, and J. A. Silander, "A wavelet transform method to merge Landsat TM and SPOT panchromatic data," *Int. J. Remote Sens.*, vol. 19, no. 4, pp. 743–757, 1998.
[86] X. Meng *et al.*, "A blind full-resolution quality evaluation method for pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.
[87] M. M. Khan, L. Alparone, and J. Chanussot, "Pansharpening quality assessment using the modulation transfer functions of instruments," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 11, pp. 3880–3891, Nov. 2009.
[88] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, and M. Selva, "Multispectral and panchromatic data fusion assessment without reference," *Photogramm. Eng. Remote Sens.*, vol. 74, no. 2, pp. 193–200, Feb. 2008.
[89] S. Lolli, L. Alparone, A. Garzelli, and G. Vivone, "Haze correction for contrast-based multispectral pansharpening," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2255–2259, Dec. 2017.
[90] L. Alparone, A. Garzelli, and G. Vivone, "Intersensor statistical matching for pansharpening: Theoretical Issues and practical solutions," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4682–4695, Aug. 2017.
[91] G. Vivone, R. Restaino, and J. Chanussot, "Full scale regression-based injection coefficients for panchromatic sharpening," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3418–3431, Jul. 2018.
[92] G. Vivone, R. Restaino, and J. Chanussot, "A regression-based high-pass modulation pansharpening approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 984–996, Feb. 2018.
[93] L. Alparone, L. Wald, J. Chanussot, C. Thomas, P. Gamba, and L. M. Bruce, "Comparison of pansharpening algorithms: Outcome of the 2006 GRS-S data-fusion contest," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3012–3021, Oct. 2007.
[94] R. Restaino, G. Vivone, M. D. Mura, and J. Chanussot, "Fusion of multispectral and panchromatic images based on morphological operators," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2882–2895, Jun. 2016.

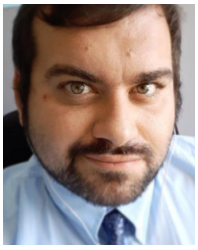
Matteo Ciotola (Graduate Student Member, IEEE) received the B.Sc. and M.Sc. *(cum laude)* degrees in automation engineering from the University of Naples Federico II, Naples, Italy, in 2018 and 2020, respectively. He is currently pursuing the Ph.D. degree with the University of Naples Federico II.

He is also a member of the GRIP Research Team. In 2020, he held a traineeship with the Centre de coopération internationale en recherche agronomique pour le développement (Cirad), Montpellier, France, where he conducted his research at the Maison de la télédétection. His main research interests focus on data fusion of remotely sensed images, in particular super-resolution and pansharpening, through deep learning algorithms.

Sergio Vitale (Student Member, IEEE) received the Laurea degree *(summa cum laude)* in telecommunication engineering from the University of Naples Federico II, Naples, Italy, in May 2017, and the Ph.D. degree from the Department of Engineering, University of Naples Parthenope, Naples, in September 2021.

He is currently a Post-Doctoral Researcher with the Department of Science and Technology, University of Naples Parthenope. His research deals with deep learning methods and their application to remote sensing image processing in the fields of image enhancement, pansharpening, synthetic aperture radar (SAR) despeckling, and interferometry.

Dr. Vitale is also a Guest Editor of the MDPI *Remote Sensing*—Special Issue on Advances in Multiresolution Fusion in Remote Sensing. In 2018, he received the 2017 Best Italian Remote Sensing Thesis Prize for the IEEE Geoscience Remote Sensing South Italy Chapter with a master's degree thesis proposing a deep learning-based method for pansharpening.



Antonio Mazza (Student Member, IEEE) received the master's degree in telecommunication engineering and the Ph.D. degree in information technology and electrical engineering from the University of Naples Federico II, Naples, Italy, in 2017 and 2021, respectively.

He is currently a Post-Doctoral Researcher with the University of Naples Federico II. His research interests include image enhancement, super-resolution, pansharpening, data fusion, multitemporal interpolation, image classification and segmentation, coregistration, synthetic aperture radar (SAR) despeckling, and deep learning.

Dr. Mazza is also Guest Editor of the MDPI *Remote Sensing*—Special Issue on Advances in Multiresolution Fusion in Remote Sensing.

Giovanni Poggi (Member, IEEE) is currently a Full Professor of telecommunications with the University of Naples Federico II, Naples, Italy. His research interests are in statistical image processing, including compression, restoration, segmentation, and classification, with application to remote-sensing (both optical and SAR images) digital forensics, and biometry.

Prof. Poggi is an Associate Editor of *Remote Sensing* (MDPI) and has been an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING and *Signal Processing*.

Giuseppe Scarpa (Senior Member, IEEE) is currently an Associate Professor of telecommunications with the University of Naples Federico II, Naples, Italy. His research activity concerns image segmentation, texture modeling and classification, object detection, pansharpening, feature extraction, data fusion, synthetic aperture radar (SAR) despeckling, image coregistration, and deep learning, with applications in the remote sensing domain.

Prof. Scarpa has been a Senior Area Editor of IEEE SIGNAL PROCESSING LETTERS. He has also served as a Guest Editor for several special issues of the *Remote Sensing* journal (MDPI).