remote sensing

MDPI

Article

Fast Super-Resolution of 20 m Sentinel-2 Bands Using Convolutional Neural Networks

Massimiliano Gargiulo [1], Antonio Mazza [1], Raffaele Gaetano [2,3], Giuseppe Ruello [1] and Giuseppe Scarpa [1,*]

[1] Department of Electrical Engineering and Information Technology (DIETI), University Federico II, 80125 Naples, Italy; massimiliano.gargiulo@unina.it (M.G.); antonio.mazza@unina.it (A.M.); ruello@unina.it (G.R.)

[2] Centre International de Recherche Agronomique pour le Développement (CIRAD), Unité Mixte de Recherche Territoires, Environnement, Télédétéction et Information Spatiale (UMR TETIS), Maison de la Télédétéction, 34000 Montpellier, France; raffaele.gaetano@cirad.fr

[3] UMR TETIS, University of Montpellier, 34000 Montpellier, France

* Correspondence: giscarpa@unina.it; Tel.: +39-081-768-3768

Received: 17 October 2019; Accepted: 8 November 2019; Published: 11 November 2019

Abstract: Images provided by the ESA Sentinel-2 mission are rapidly becoming the main source of information for the entire remote sensing community, thanks to their unprecedented combination of spatial, spectral and temporal resolution, as well as their associated open access policy. Due to a sensor design trade-off, images are acquired (and delivered) at different spatial resolutions (10, 20 and 60 m) according to specific sets of wavelengths, with only the four visible and near infrared bands provided at the highest resolution (10 m). Although this is not a limiting factor in general, many applications seem to emerge in which the resolution enhancement of 20 m bands may be beneficial, motivating the development of specific super-resolution methods. In this work, we propose to leverage Convolutional Neural Networks (CNNs) to provide a fast, upscalable method for the single-sensor fusion of Sentinel-2 (S2) data, whose aim is to provide a 10 m super-resolution of the original 20 m bands. Experimental results demonstrate that the proposed solution can achieve better performance with respect to most of the state-of-the-art methods, including other deep learning based ones with a considerable saving of computational burden.

Keywords: pansharpening; data fusion; convolutional neural network; multi-resolution analysis; landcover classification

1. Introduction

The twin Sentinel-2 satellites ensure a global World coverage with a revisit time of five days at the equator, providing a multi-resolution stack composed of 13 spectral bands, between the visible and short-wave infrared (SWIR), distributed over three resolution levels. Four bands lying between visible and near-infrared (NIR) are given at the finer resolution of 10 m, while the remaining ones are provided at 20 (six bands) and 60 (three bands) m, as a result of a trade-off between storage and transmission bandwidth limitations. The 10 and 20 m bands are commonly employed for land-cover or water mapping, agriculture or forestry, estimation of biophysical variables, and risk management (floods, forest fires, subsidence, and landslide), while lower resolution 60 m bands can be used for monitoring of water vapor, aerosol corrections, pollution monitoring, cirrus clouds estimation and so forth [1,2]. Specifically, beyond land-cover classification, S2 images can be useful in such diverse applications as the prediction of growing stock volume in forest ecosystems [3], the estimation of the Leaf Area Index (LAI) [4,5], the retrieval of canopy chlorophyll content [6], the mapping of the extent

of glaciers [7], the water quality monitoring [8], the classification of crop or tree species [9], and the built-up areas detection [10].

In light of its free availability, world-wide coverage, revisit frequency and, not least, its above remarked wide applicability, several research teams have proposed solutions to super-resolve Sentinel-2 images, rising 20 m and/or 60 m bands up to 10 m resolution. Besides, several works testify the advantage of using super-resolved S2 images in several applications such as water mapping [11], fire detection [12], urban mapping [13], and vegetation monitoring [14].

According to the taxonomy suggested by Lanaras et al. [2] resolution enhancement techniques can be gathered in three main groups: (i) pansharpening and related adaptations; (ii) imaging model inversion; and (iii) machine learning. In addition to these category, it is also worth mentioning the matrix factorization approaches (e.g., [15,16]), which are more suited to the fusion of low resolution hyperspectral images with high resolution multispectral ones. In fact, the spectral variability becomes a serious concern to be handled carefully by means of unmixing oriented methodologies [17,18]. The first category refers to the classical pansharpening, where the super-resolution of low-resolution bands is achieved by injecting spatial information from a single spectrally-overlapping higher-resolution band. This is the case for many remote sensing systems such as Ikonos, QuickBird, GeoEye, WorldView, and so forth. The so-called component substitution methods [19,20], the multi-resolution analysis approaches [21,22], or other energy minimization methods [23–25] belong to this category. A recent survey on pansharpening can be found in [26]. Pansharpening methods can also be extended to Sentinel-2 images in different ways, although S2 bands at different resolutions present a weak or negligible spectral overlap, as shown by several works [27–31].

The second group refers to methods that face the super-resolution as an inverse problem under the hypothesis of known imaging model. The ill-posedness is therefore addressed by means of additional regularization constraints encoded in a Bayesian or a variational framework. Brodu's super-resolution method [32] separates band-dependent from cross-band spectral information, ensuring the consistency of the "geometry of scene elements" while preserving their overall reflectance. Lanaras et al. [33] adopted an observation model with per-band point spread functions that accounts for convolutional blur, downsampling, and noise. The regularization consists of two parts, a dimensionality reduction that implies correlation between the bands, and a spatially varying, contrast-dependent penalization of the (quadratic) gradients learned from the 10 m bands. In a similar approach, Paris et al. [34] employed a patch-based regularization that promotes self-similarity of the images. The method proceeds hierarchically by first sharpening the 20 m bands and then the coarser 60 m ones.

The last category casts machine learning approaches, and notably deep learning (DL) ones, which have recently gained great attention from the computer vision and signal processing communities and nearby fields, including remote sensing. In this case, contrarily to the previous categories, no explicit modeling (neither exact nor approximated) of the relationship between high and low resolution bands is required, since it is directly learned from data. Deep networks allow in principle to mimic very complex nonlinear relationships provided that enough training data are available. In this regard, it is also worth recalling that the pansharpening of multi-resolution images is somewhat related to the unmixing of multi-/hyper-spectral images [17,18], since in both cases the general aim is to derive the different spectral responses covered by a single, spatially coarse observation. However, more specifically, in these two problems, expectations are considerably different: spectral unmixing is a pertinent solution when the interest is focused on surface materials, hence requiring high precision on the retrieval of the corresponding spectral responses without the need to improve their spatial localization. In pansharpening, the focus is mainly on spatial resolution enhancement while preserving at most the spectral properties of the sources, and no specific information discovery about the radiometry of materials is typically expected. In fact, traditional pansharpening methods try to model spectral diversity, for example, by means of the modulation transfer function of the sensor [21,22], instead of using radiative transfer models associated to the possible land covers. In any case, from the deep learning perspective, it makes little difference once the goal is fixed and, more

importantly, a sufficiently rich training dataset is provided, as the knowledge (model parameters) will come from experience (data). To the best of our knowledge, the first notable example of DL applied to the super-resolution of remote sensing images is the pansharpening convolutional neural network (PNN) proposed by Masi et al. [35], which has been recently upgraded [36] with the introduction of a residual learning block and a fine-tuning stage for target adaptivity and cross-sensor usage. Another residual network for pansharpening (PanNet) is proposed in [37]. However, none of these methods can be applied to S2 images without some architectural network adaptation and retraining. Examples of convolutional networks conceived for Sentinel-2 are instead proposed in [2,11]. In [11], the super-resolution was limited to the SWIR 20 m band, as the actual goal was water mapping by means of the modified normalized difference water index (MNDWI), for which green and SWIR bands were requested. Lanaras et al. [2], instead, collected a very large training dataset which has been used to train two much deeper super-resolution networks, one for the 20 m subset of bands and the other for the remaining 60 m bands, achieving state-of-the-art results. In related problems, for example the single-image super-resolution of natural images or other more complex vision tasks such as object recognition or instance segmentation, thanks to the knowledge hidden in huge and shared training databases, deep learning has shown really impressive results compared to model-based approaches. Data sharing has represented a key enabling factor in these cases allowing researchers to compete with each other or reproduce others' models. In light of this consideration, we believe that Sentinel-2 is a very interesting case because of the free access to data that can serve as playground for a larger scale research activity on remote sensing super-resolution or other tasks. In the same spirit, Lanaras et al. [2] pushed on the power of the data by collecting a relatively large dataset to get good generalization properties. On the other hand, complexity is also an issue that end users care about. In this regard, the challenge of our contribute is to design and train a relatively small and flexible network capable of achieving competitive results at a reduced cost on the super-resolution of the 20 m S2 bands, exploiting spatial information from the higher-resolution 10 m S2/VNIR bands. Indeed, the proposed network being lightweight, apart from enabling the use of the method on cheaper hardware, allows quickly fine-tuning it when the target data are misaligned from the training data for some reason. The proposed method for Fast Upscaling of SEntinel-2 (FUSE) images is an evolution of the *proof-of-concept* work presented in [38]. In particular, the major improvements with respect to the method in [38] reside in the following changes:

a. Architectural improvements with the introduction of an additional convolutional layer.
b. The definition of a new loss function which accounts for both spectral and structural consistency.
c. An extensive experimental evaluation using diverse datasets for testing that confirms the generalization capabilities of the proposed approach.

The rest of the paper is organized as follows. In Section 2, we describe datasets and proposed method. Evaluation metrics, comparative solutions and experimental results are then gathered in Section 3. Insights about the performance of the proposed solution and related future perspectives are given in Section 4. Finally, Section 5 provides concluding remarks.

2. Materials and Methods

The development of a deep learning super-resolution method suited for a given remote sensing imagery involves at least three key steps, with some iterations among them:

a. Selection/generation of a suitable dataset for training, validation and test;
b. Design and implementation of one or more DL models;
c. Training and validation of the models (b) using the selected dataset (a).

By following this rationale, for ease of presentation, in this section, we first present the datasets and their preprocessing (a), then we describe design (b) and training (c) of the proposed model.

2.1. Datasets and Labels Generation

Regardless of its complexity and capacity, a target deep learning model remains a data-driven machinery whose ultimate behavior heavily depends on the training dataset, notably on its representativeness of real-world cases. Hence, we provide here detailed information about our datasets and their preprocessing.

For the sake of clarity, let us first recall the main characteristics of the 13 spectral bands of Sentinel-2, gathered in Table 1, and clarify symbols and notations that are used in the following with the help of Table 2.

Table 1. Sentinel-2 bands. The 10 m bands are highlighted in blue. In red are the six 20 m bands to be super-resolved. The remaining are 60 m bands.

Bands	**B1**	**B2**	**B3**	**B4**	**B5**	**B6**	**B7**	**B8**	**B8a**	**B9**	**B10**	**B11**	**B12**
Center wavelength [nm]	443	490	560	665	705	740	783	842	865	945	1380	1610	2190
Bandwidth [nm]	20	65	35	30	15	15	20	115	20	20	30	90	180
Spatial resolution [m]	60	10	10	10	20	20	20	10	20	60	60	20	20

Table 2. Notations and symbols.

Symbol	**Meaning**
$\mathbf{x}$	Stack of six S2 spectral bands (B5, B6, B7, Bba, B11, B12) to be super-resolved.
$\mathbf{z}$	Stack of four high-resolution S2 bands (B2, B3, B4, B8).
$\mathbf{x}^{\mathrm{hp}}$, $\mathbf{z}^{\mathrm{hp}}$	High-pass filtered versions of $\mathbf{x}$ and $\mathbf{z}$, respectively.
$\hat{\mathbf{x}}$	Super-resolved version of $\mathbf{x}$.
$\mathbf{r}$	Full-resolution *reference* (also referred to as *ground truth* or *label*), usually unavailable.
$x, \hat{x}, r$	generic band of $\mathbf{x}, \hat{\mathbf{x}}, \mathbf{r}$, respectively.
$\widetilde{\mathbf{x}}, \widetilde{x}, \widetilde{\mathbf{x}}^{\mathrm{hp}}$	Upsampled (via bicubic) versions of $\mathbf{x}, x, \mathbf{x}^{\mathrm{hp}}$, respectively.
$\overline{z}$	Single (average) band of $\mathbf{z}$.
$\mathbf{x}_{\downarrow}$, $\mathbf{z}_{\downarrow}$, $\mathbf{r}_{\downarrow}$,...	Reduced-resolution domain variables associated with $\mathbf{x}$, $\mathbf{z}$, $\mathbf{r}$,..., respectively. Whenever unambiguous subscript $_{\downarrow}$ will be dropped.

Except for some cases where unsupervised learning strategies can be applied, a sufficiently large dataset containing input–output examples is usually necessary to to train a deep learning model. This is also the case for super-resolution or pansharpening. In our case, as we decided to fuse 10 m bands ($\mathbf{z}$) with 20 m ($\mathbf{x}$) to enhance the resolution of $\mathbf{x}$ by a factor of 2 (resolution ratio), which means that we should have examples of the kind $((\mathbf{x}, \mathbf{z}); \mathbf{r})$, being $\mathbf{r}$ the desired (super-resolved) output corresponding to the composite input instance $(\mathbf{x}, \mathbf{z})$. In rare cases, one can rely on referenced data, for example thanks to ad hoc missions to collect full-resolution data to be used as reference, whereas in most cases referenced samples are unavailable.

Under the latter assumption, many deep learning solutions for super-resolution or pansharpening have been developed (e.g., [2,11,35,36,39–41]) by means of a proper schedule for generating referenced training samples from the same no-reference input dataset. It consists of a resolution downgrade process that each input band undergoes which involves two steps:

(i) band-wise low-pass filtering; and
(ii) uniform $R \times R$ spatial subsampling, being R the target super-resolution factor.

This is aimed to shift the problem from the original *full*-resolution domain to a *reduced*-resolution domain. In our case, $R = 2$ while the two original input components, $\mathbf{x}$ and $\mathbf{z}$, will be transformed in corresponding variables $\mathbf{x}_{\downarrow}$ and $\mathbf{z}_{\downarrow}$, respectively, lying in the reduced-resolution space, with associated reference $\mathbf{r}_{\downarrow}$ trivially given by $\mathbf{r}_{\downarrow} = \mathbf{x}$. How to filter the several bands before subsampling is an open question. Lanaras et al. [2] pointed out that with deep learning one does not need to specify sensor characteristics, for instance, spectral response functions, since sensor properties are implicit

in the training data. Contrarily, Masi et al. [35] asserted that the resolution scaling should be done accounting for the sensor Modulation Transfer Function (MTF), in order to generalize properly when applied at full resolution. Such a position follows the same rationale of the so-called Wald's protocol, a procedure commonly used for generating referenced data for objective comparison of pansharpening methods [26]. Actually, this controversial point cannot be resolved by looking at the performances in the reduced-resolution space, since a network learns from training data the due relationship whatever preprocessing has been performed on the input data. On the other hand, in full-resolution domain, no objective measures can be used because of the lacking referenced test data. In this work we follow the approach proposed in [35] making use of sensor MTF. The process for the generation of a training sample is summarized in Figure 1. Each band undergoes a different low-pass filtering, prior to being downsampled, whose cut-off frequency is related to the sensor MTF characteristics. Additional details can be found in [42].

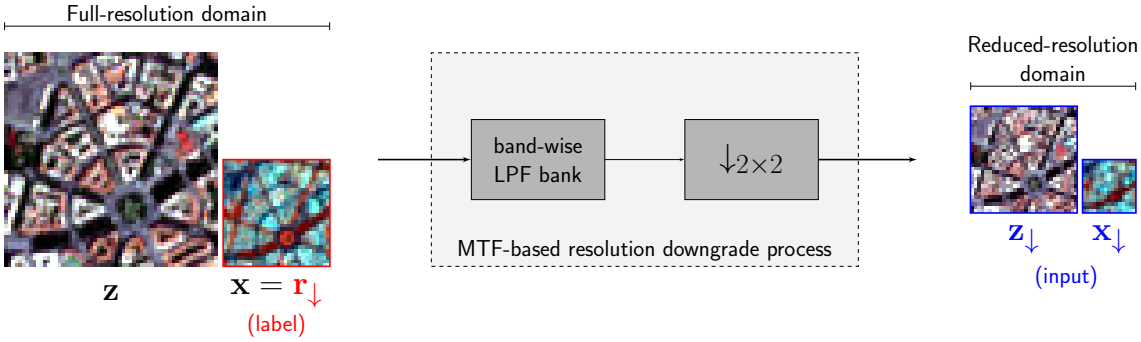
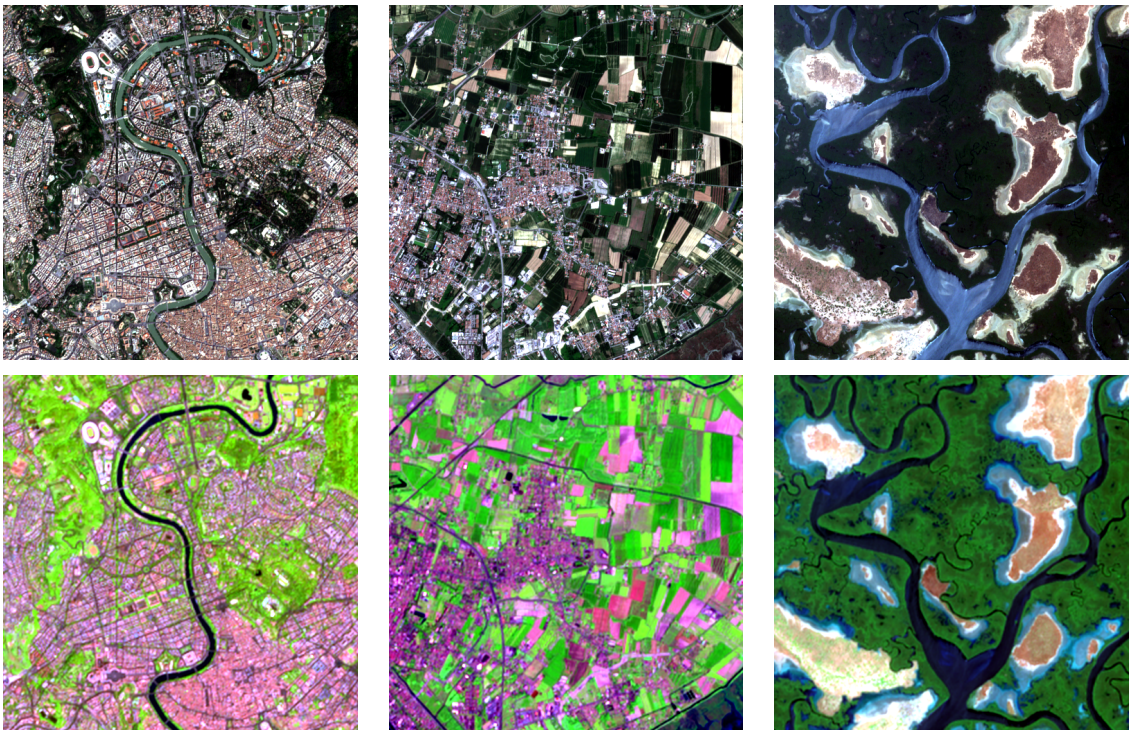


Figure 1. Generation of a training sample $((\mathbf{x}_{\downarrow}, \mathbf{z}_{\downarrow}); \mathbf{r}_{\downarrow})$ using Wald's protocol. All images are shown in false-color RGB using subsets of bands for ease of presentation. Each band is low-lass filtered with a different cut-off frequency according with the sensor MTF characteristics.

Another rather critical issue is the training dataset selection as it impacts the capability of the trained models to generalize well on unseen data. In the computer vision domain, a huge effort has been devoted to the collection of very large datasets in order to support the development of deep learning solutions for such diverse problems as classification, detection, semantic segmentation, tracking video and so forth (notable examples are ImageNet and Kitty datasets). Instead, within the remote sensing domain, there are no examples of datasets which are as large as ImageNet or Kitty. This is due to several obstacles, among which the cost of the data and the related labeling which requires domain experts, as well as the data sharing policy usually adopted in the past years by the remote sensing community. Luckily, for super-resolution, one can at least rely on the above-described fully-automated resolution downgrading strategy to avoid labeling costs. Due to the scarcity of data, most deep-learning models for resolution enhancement applied to remote sensing have been trained on a relatively small dataset, possibly taken from a few large images, from which non-overlapping sets for training, validation and testing are singled out [35,37,41]. The generalization limits of a pansharpening model trained on too few data have been stressed in [36], for both cross-image and cross-sensor scenarios, where a fine-tuning stage has been proposed to cope with the scarcity of data. In particular, it was shown that, for a relatively small CNN that integrates a residual learning module, a few training iterations (fine-tuning) on the reduced-resolution version of the target image allow quickly recovering the performance loss due to the misalignment between training and test sets. For Sentinel-2 imagery, thanks to the free access guaranteed by the Copernicus program, larger and more representative datasets can be collected, as done by Lanaras et al. [2], aiming for a roughly even distribution on the globe and for variety in terms of climate zone, land-cover and biome type. In this study, we opted for a lighter and flexible solution with a relatively small number of parameters to learn and a (pre-)training dataset of relatively limited size. This choice is motivated by the experimental observation that in

actual application the tuning of the parameters is still recommendable even if larger datasets have been used in training, making appealing lighter solutions that can be quickly tuned if needed.

To be aligned with the work of Lanaras et al. [2], we decided to keep their setting by using Sentinel-2 data without atmospheric correction (L1C product) for our experiments. For training and validation, we referred to three scenes (see Figure 2), corresponding to different environmental contexts: Venice, Rome, and Geba River. In particular, we randomly cropped 18,996 square tiles of size 33×33 (at 20 m resolution) from the three selected scenes to be used for training (15,198) and validation (3898). Besides, we have choosen four more scenes for the purpose of testing, namely Athens, Tokyo, Addis Abeba, and Sydney, which present different characteristics, hence allowing for a more robust validation of the proposed model. From such sites, we singled out three 512×512 crops at 10 m resolution, for a total of twelve test samples.



Rome Venice Geba River

Figure 2. Examples of images used for training. (top) RGB-composite images using 10 m bands B4(R), B3(G) and B2(B), subset of **z**; and (bottom) corresponding 20 m RGB subset of **x**, using B5(R), B8a(G) and B11(B).

2.2. Proposed Method

The proposed solution takes inspiration from two state-of-the-art CNN models for pansharpening, namely PanNet [37] and the target-adaptive version [36] of PNN [35], both conceived for very high resolution sensors such as Ikonos or WorldView-2/3. Both methods rely on a residual learning scheme, while main differences concern loss function, input preprocessing, and overall network backbone shape and size.

Figure 3 shows the top-level flowchart of the proposed method. As we deal with Sentinel-2 images, differently from Yang et al. [37] and Scarpa et al. [36], we have 10 input bands, six lower-resolution ones (**x**), to be super-resolved, plus four higher-resolution bands (**z**). Let us preliminarily point out that we train a single (relatively small) network for each band x to be super-resolved, as represented at the output in Figure 3. However, the deterministic preprocessing bounded by the dashed box is a shared part, while the core CNN, with fixed hyper-parameters, changes from one band to another to be super-resolved. This choice presents two main advantages. The first is that whenever users

need to super-resolve only a specific band, they can make use of a lighter solution with computational advantages. The second reason is related to the experimental observation that training separately the six networks allows reaching the desired loss levels more quickly than using a single wider network. This feature is particularly desirable if users need to fine-tune the network on their own dataset. Turning back to the workflow, observe that both input subsets, $\mathbf{x}$ and $\mathbf{z}$, are high-pass filtered (HPF) as also done by PanNet. This operation relies on the intuition that the missing details that the network is asked to recover lie in the high frequency range of the input image. Next, the HPF component $\mathbf{x}^{\mathrm{hp}}$ is upsampled ($R \times R$) using a standard bicubic interpolation, yielding $\widetilde{\mathbf{x}}^{\mathrm{hp}}$, in order to match the size of $\mathbf{z}^{\mathrm{hp}}$ with which to be concatenated prior to feed the actual CNN. The single-band CNN output $f_\Phi(\mathbf{x}, \mathbf{z})$ is therefore combined with the upsampled target band $\widetilde{x}$ to provide its super-resolved version $\hat{x} = \widetilde{x} + f_\Phi(\mathbf{x}, \mathbf{z})$. This last combination, obtained through a skip connection that retrieves the low-resolution content of $\hat{x}$ directly from the input, is known as residual learning strategy [43], and has soon became a standard option for deep learning based super-resolution and pansharpening [2,36,37], as it is proven to speed-up the learning process.

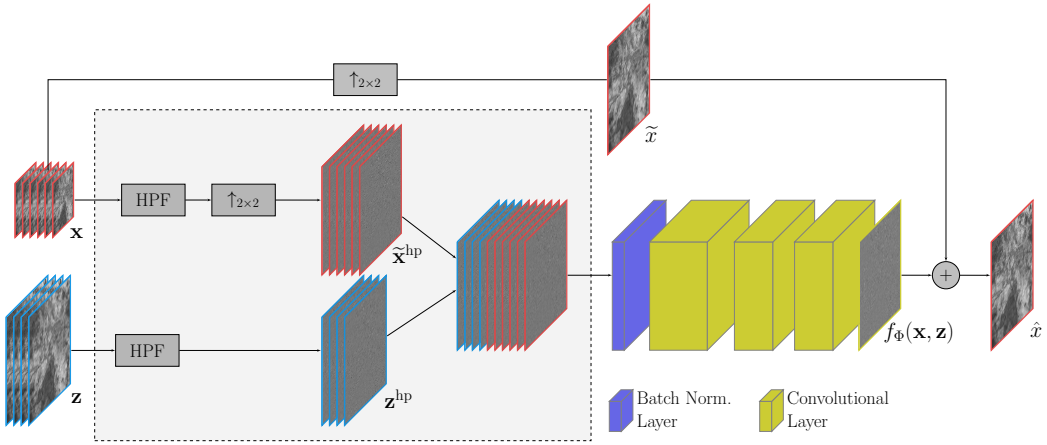


Figure 3. Top-level workflow for the super-resolution of any 20 m band of Sentinel-2. The dashed box gathers the shared processin,g which is the same for all predictors.

The CNN architecture is more similar to the pansharpening models [35,36] than to PanNet [37], making use of just four convolutional layers, whereas PanNet uses ten layers, each singling out 32 features (except for the output layer). Moreover, a batch normalization layer operating on the input stack precedes the convolutional ones. This has proven to make the learning process robust with respect to the statistical fluctuations of the training dataset [44]. In Table 3, the network hyper-parameters of the convolutional layers are summarized.

Table 3. Hyper-parameters of the convolutional layers for the proposed CNN model.

	ConvLayer 1	**ConvLayer 2**	**ConvLayer 3**	**ConvLayer 4**
Input Channels	10	48	32	32
Spatial Support	3×3	3×3	3×3	3×3
Output Channels	48	32	32	1
Activation	ReLU	ReLU	ReLU	tanh

Training

Once the training dataset and model are fixed, a suitable loss function to be minimized needs to be defined in order for the learning process to take place. L_2 or L_1 norms are typical choices [2,35,36,38,39] due to their simplicity and robustness, with the latter being probably more effective to speed-up the training, as observed in [2,36]. However, these measures do not account for structural consistency as they are computed on a pixel-wise basis and, therefore, assess only spectral dissimilarity. To cope

with this limitation, an option is to resort to a so-called *perceptual* loss [45], which is an indirect error measurement performed in a suitable feature space generated with a dedicated CNN. In [37], structural consistency is enforced by working directly on detail (HPF) bands. In the proposed solution, in addition to the use HPF components, we also define a combined loss that explicitly accounts for spectral and structural consistency. In particular, inspired by the variational approach [46], we make use of the following loss function:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{Spec}} + \lambda_2 \mathcal{L}_{\text{Struct}} + \lambda_3 \mathcal{L}_{\text{Reg}} \tag{1}$$

where three terms, corresponding to fidelity, or spectral consistency ($\mathcal{L}_{\text{Spec}}$), structural consistency ($\mathcal{L}_{\text{Struct}}$) and regularity ($\mathcal{L}_{\text{Reg}}$), are linearly combined. The weights were tuned experimentally using the validation set as $\lambda_1 = 1$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.01$.

By following the intuition proposed in [2,36], we decided to base the fidelity term on the L_1 norm, that is

$$\begin{aligned} \mathcal{L}_{\text{Spec}} &= \mathrm{E}\left\{ \left\| \hat{x}_{\downarrow} - r_{\downarrow} \right\|_1 \right\} \\ &= \mathrm{E}\left\{ \left\| f_{\Phi}(\mathbf{x}_{\downarrow}, \mathbf{z}_{\downarrow}) + \widetilde{x}_{\downarrow} - r_{\downarrow} \right\|_1 \right\} \end{aligned}$$

where the expectation $\mathrm{E}\{\cdot\}$ is estimated on the reduced-resolution training minibatches during the gradient descent procedure. $f_{\Phi}(\cdot)$ stands for the CNN function (including preprocessing) whose parameters to learn are collectively indicated with Φ. This loss term, as well as the other two, refers to a single band ($x_{\downarrow}$) super-resolution whose ground-truth is $r_{\downarrow} = x$. As the training is performed in the reduced-resolution domain, in the reminder on this section, we drop the subscript $\downarrow$ for the sake of simplicity.

The structural consistency term is given by

$$\mathcal{L}_{\text{Struct}} = \mathrm{E}\left\{ \sum_{i=1}^{4} \left\| G_i(\hat{x} - r) \right\|_{1/2} \right\},$$

where the operator $G = (G_1, \ldots, G_4)$ generalizes the gradient operator including derivatives in the diagonal directions that help to improve quality, as shown in [46]. It has been shown that the gradient distribution for real-world images is better fit with a heavy-tailed distribution such as a hyper-Laplacian ($p(x) \propto e^{-k|x|^p}$, $0 < p < 1$). Accordingly, we make use of a L_p-norm with $p = 1/2$, which we believe can be more effective [46]. This term penalizes discontinuities in the super-resolved band $\hat{x}$ if they do not occur, with the same orientation, in the panchromatic band. As the dynamics of these discontinuities are different, an additional prior regularization term that penalizes the total variation of $\hat{x}$ helps to avoid unstable behaviors:

$$\mathcal{L}_{\text{Reg}} = \mathrm{E}\left\{ \left\| \nabla \hat{x} \right\|_1 \right\} = \mathrm{E}\left\{ \left\| \nabla f_{\Phi}(\mathbf{x}, \mathbf{z}) + \nabla \widetilde{x} \right\|_1 \right\}.$$

Eventually, the network parameters were (pre-)trained by means of the Adaptive Moment Estimation (ADAM) optimization algorithm [47] applied to the above-defined overall loss (Equation (1)). In particular, we have set the ADAM default hyper-parameters, which are learning rate, $\eta = 0.002$, and decay rate of the first and second moments, $\beta_1 = 0.9$ and $\beta_2 = 0.999$, respectively [48]. The training was run for 200 epochs, being an epoch a single pass over all minibatches (118) in which the training set has been split, with each minibatch composed of 128 33×33 input–output samples.

3. Experimental Results

In this section, after a brief recall of the accuracy evaluation metrics (Section 3.1) and of the comparative methods (Section 3.2), we provide and discuss numerical and visual results (Section 3.3).

3.1. Accuracy Metrics

The quality assessment of pansharpening algorithms can be carried out in two frameworks, with or without ground-truth. Since normally the ground-truth is unavailable, the former context refers to the application of Wald's protocol [42], which is the same process used for the generation of training samples, as described in Section 2.1. Therefore, this evaluation frame, hereinafter referred to as *reference-based*, applies in the reduced-resolution domain and allows one to provide objective quality measurements. Because of the resolution shift (downgrade), the reference-based evaluation approach has a limited extent and it is therefore custom to complement it with a full-resolution assessment, referred to as the *no-reference* one, aimed to give qualitative measurements at full resolution.

In particular, we use the following reference-based metrics:

- Universal Image Quality Index (Q-Index) takes into account three different components: correlation coefficient, mean luminance distance and contrasts [49].
- *Erreur Relative Globale Adimensionnelle de Synthése* (ERGAS) measures the overall radiometric distortion between two images [50].
- Spectral Angle Mapper (SAM) measures the spectral divergence between images by averaging the pixel-wise angle between spectral signatures [51].
- High-pass Correlation Coefficient (HCC) is the correlation coefficient between the high-pass filtered components of two compared images [52].

On the other hand, as no-reference metrics, we use the following [26,53]:

- Spectral Distortion (D_λ) measures the spectral distance between the bicubic upscaling of the image component to be super-resolved, $\tilde{\mathbf{x}}$, and its super-resolution, $\hat{\mathbf{x}}$.
- Spatial Distortion (D_S) is a measurement of the spatial consistency between the super-resolved image $\hat{\mathbf{x}}$ and the high-resolution component $\mathbf{z}$.
- Quality No-Reference (QNR) index is a combination of the two above indexes that accounts for both spatial and spectral distortions.

For further details about the definition of the above metrics, the reader is referred to the associated articles.

3.2. Compared Methods

We compared our FUSE method with several state-of-the-art solutions. On the one side are classical approaches for pansharpening, properly generalized to the case of Sentinel-2, such as the following:

- Generalized Intensity Hue Saturation (GIHS) method [20].
- Brovey transform-based method [54].
- Indusion [55].
- Partial Replacement Adaptive Component Substitution (PRACS) [56].
- A Troús Wavelet Transform-based method (ATWT-M3) [22].
- The High-Pass Filtering (HPF) approach [21].
- Generalized Laplacian Pyramid with High Pass Modulation injection (MTF-GLP-HPM) [57].
- Gram-Schmidt algorithm with Generalized Laplacian Pyramid decomposition (GS2-GLP) [57].

Detailed information about these approaches can be found in the survey work of Vivone et al. [26].

Besides, we also compared with the following deep learning approaches native for Sentinel-2 images, including two ablations of our proposal:

- Our previous CNN-based method (M5) proposed in [11], extended (training from scratch) to all six 20 m bands.
- The CNN model (DSen2) proposed in [2], which is much deeper than ours and has been trained on a very large dataset.

- An enhancement of M5 where High-Pass filtering on the input and other minor changes have been introduced (HP-M5) [38], which represents a first insight on the improvements proposed in this work.
- FUSE with only three layers instead of four.
- FUSE trained using the L_1 norm without regularization and structural loss terms.

3.3. Numerical and Visual Results

To assess the performance of the proposed method, we collected twelve 512 × 512 images (at 10 m resolution) from four larger images taken over Athens, Adis Abeba, Sydney and Tokyo, respectively, from which no training or validation samples were extracted.

Numerical figures were computed for all compared methods on each test image. The average measures over the dataset are gathered in Table 4. Reference-based accuracy indicators shown on the left-hand side of the table are computed in the reduced-resolution space and provide objective measurements of the reconstruction error. Overall, we can see that the proposed FUSE method performs slightly better than DSen2 and outperforms all compared solution on three out of four indicators. On the other hand, M5 and ATWT-M3 show a slightly better spectral preservation compared to FUSE according to the Spectral Angle Mapper indicator.

As reduced-resolution data do not fully reproduce statistical fluctuations that may occur in the full resolution context, a common choice is to complement the low-resolution evaluation with a full-resolution assessment that, however, does not rely on objective error measurements. In particular, we resort to three well-established indicators that are usually employed in the pansharpening context: the spectral and spatial distortions, D_λ and D_S, respectively, and their combination, the QNR. According to these indicators, shown on the right-hand side of Table 4, the proposed method, again, outperforms the competitors. A slightly better spectral preservation is given by HP-M5, M5 and ATWT-M3.

Table 4. Accuracy assessment of several super-resolution methods. On top are model-based approaches and DL methods are on the bottom, including the proposed FUSE method.

	Reference-Based				**No-Reference**		
Method (**Ideal**)	**Q** (1)	**HCC** (1)	**ERGAS** (0)	**SAM** (0)	**QNR** (1)	$\mathbf{D_\lambda}$ (0)	$\mathbf{D_S}$ (0)
HPF	0.9674	0.6231	3.054	0.0641	0.8119	0.1348	0.0679
Brovey	0.9002	0.6738	4.581	0.0026	0.6717	0.2382	0.1241
MTF_GLP_HPM	0.8560	0.6077	19.82	0.2813	0.7802	0.1678	0.0643
GS2_GLP	0.9759	0.6821	2.613	0.0564	0.8129	0.1367	0.0647
ATWT-M3	0.9573	0.6965	3.009	**0.0019**	0.8627	0.0947	0.0473
PRACS	0.9767	0.7284	2.274	**0.0019**	0.8800	**0.0847**	0.0395
GIHS	0.8622	0.6601	5.336	0.0579	0.6112	0.2999	0.1444
Indusion	0.9582	0.6273	3.314	0.0425	0.8424	0.1311	0.0321
M5	0.9883	0.8432	1.830	**0.0019**	0.8715	0.0942	0.0389
HP-M5	0.9895	0.8492	1.720	0.0282	0.8779	0.0931	0.0329
DSen2	0.9916	0.8712	1.480	0.0194	0.8684	0.1028	0.0330
FUSE (3 layers)	0.9931	0.8602	1.631	0.0020	0.8521	0.1082	0.0474
FUSE (L_1 loss)	0.9930	0.8660	1.681	0.1963	0.8570	0.1081	0.0410
FUSE (full version)	**0.9934**	**0.8830**	**1.354**	0.0184	**0.8818**	0.1002	**0.0203**

Let us now look at some sample results starting from the full-resolution context. Figures 4 and 5 show some of the 512 × 512 images used for test, associated with urban and extra-urban contexts, respectively. For the sake of visualization, we use RGB false-color subsets of $\mathbf{z}$ and $\mathbf{x}$. In particular, we use three out of four bands of $\mathbf{z}$ (B2, B3 and B4), and three out of six bands of $\mathbf{x}$ (B5, B8a and B11—see Table 1). The input components $\mathbf{z}$ and $\widetilde{\mathbf{x}}$ are shown on the left and middle columns, while the super-resolution $\hat{\mathbf{x}}$ obtained with the proposed method is shown on the right.

Although at a first glance these results look pretty nice, a different observation scale would help to gain insight the behavior of the compared solutions. Therefore, in Figure 6, we show some zoomed details with the corresponding super-resolutions using different selected methods. In particular, for the sake of simplicity, we have restricted the visual inspection to the most representative DL and not DL approaches according to both reference-based and no-reference indicators reported in Table 4. A careful inspection reveals that some model-based approaches provide higher detail enhancement compared to DL methods. However, it remains difficult to appreciate the spectral preservation capability of the different methods due to the lack of objective references.

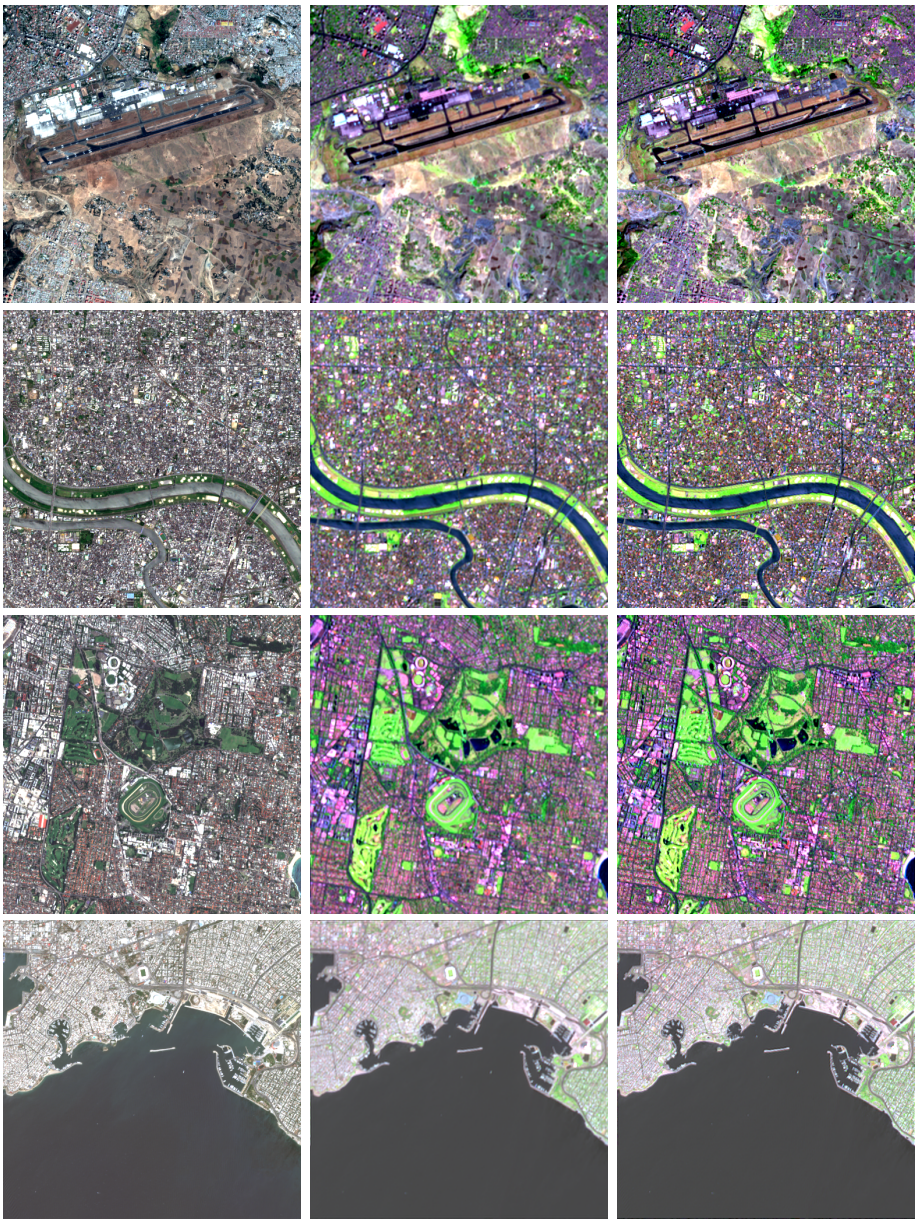


Figure 4. Super-resolution of the test images—Urban zones. From top to bottom: Adis Abeba, Tokyo, Sydney, and Athens. From left to right: High-resolution 10 m input component $\mathbf{z}$, low-resolution 20 m component $\tilde{\mathbf{x}}$ to be super-resolved, and super-resolution $\hat{\mathbf{x}}$ using the FUSE algorithm.

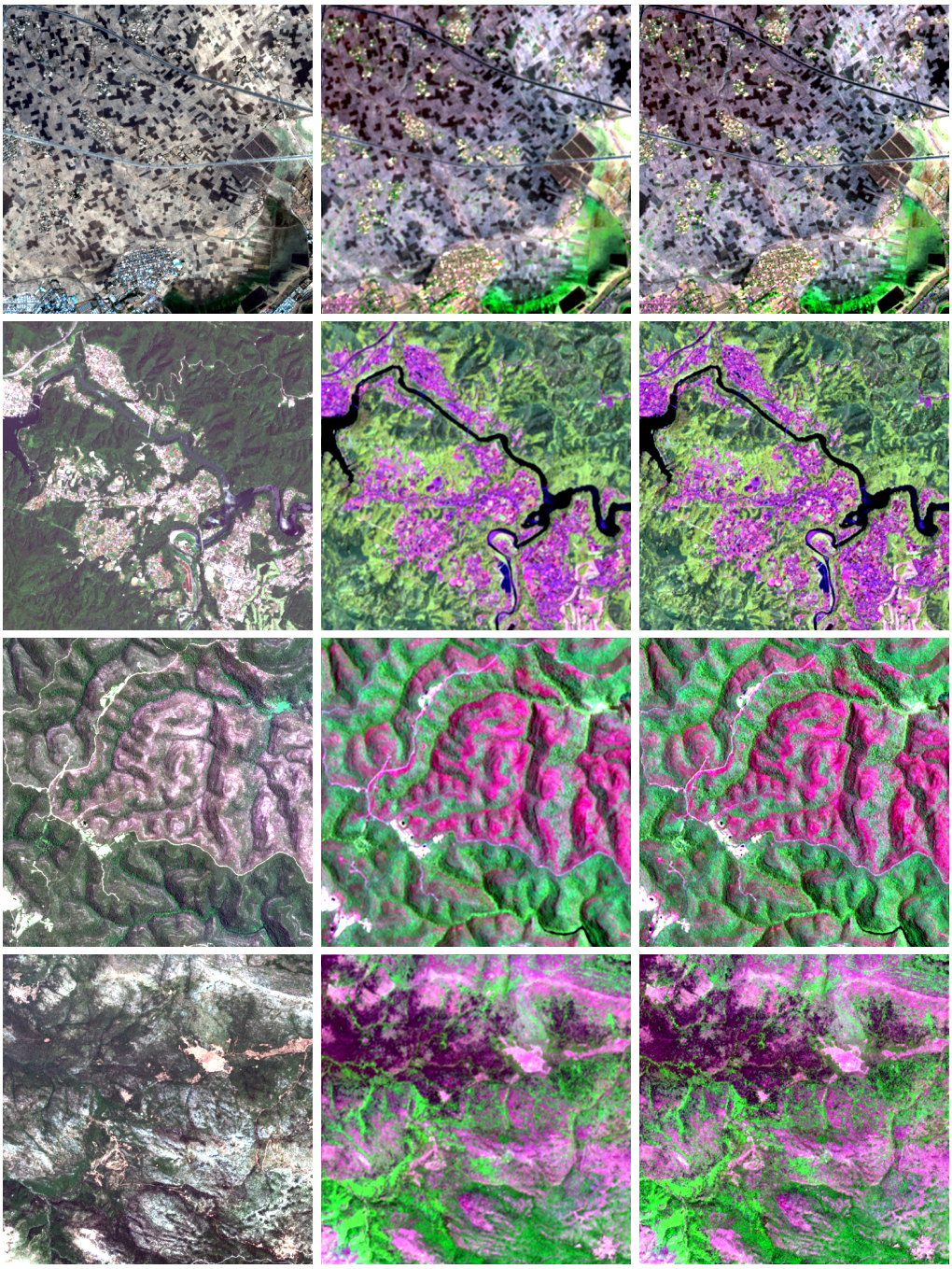


Figure 5. Super-resolution of the test images—Extra-urban zones. From top to bottom: Adis Abeba, Tokyo, Sydney, and Athens. From left to right: High-resolution 10 m input component $\mathbf{z}$, low-resolution 20 m component $\widetilde{\mathbf{x}}$ to be super-resolved, and super-resolution $\hat{\mathbf{x}}$ using the FUSE algorithm.

Actual errors can be visualized in the reduced-resolution domain instead. In Figure 7, we show in particular a few meaningful details processed in such a domain. For each sample, the composite input $(\widetilde{\mathbf{x}}_{\downarrow}, \mathbf{z}_{\downarrow})$ is shown in the leftmost column, followed by the reference ground-truth $\mathbf{r}_{\downarrow}$. Then, Columns 3–7 show a few selected solutions (odd rows) with the corresponding error maps (even rows) obtained as difference between the super-resolved image and the reference, $\hat{\mathbf{x}}_{\downarrow} - \mathbf{r}_{\downarrow}$. As it can be seen, the DL methods perform pretty well in comparison with model based approaches as the error map is nearly constant gray, whereas for PRACS and ATWT-M3 visible piece-wise color shifts are introduced. This observation does not contrast with the good values of SAM obtained by PRACS,

since this indicator accounts for the relative color/band proportions but not for their absolute intensity (some "colorful" error maps in Figure 7 are partially due to the band-wise histogram stretching used for the sake of visualization). Overall, by looking at both numerical accuracy indicators and visual results, in both reduced- and full-resolution contexts, the proposed method provides state-of-the-art results on our datasets, as does DSen2.

Figure 6. Full-resolution results for selected details. For each detail (row) from left to right are shown the two input components to be fused, followed by the corresponding fusions obtained by compared methods.

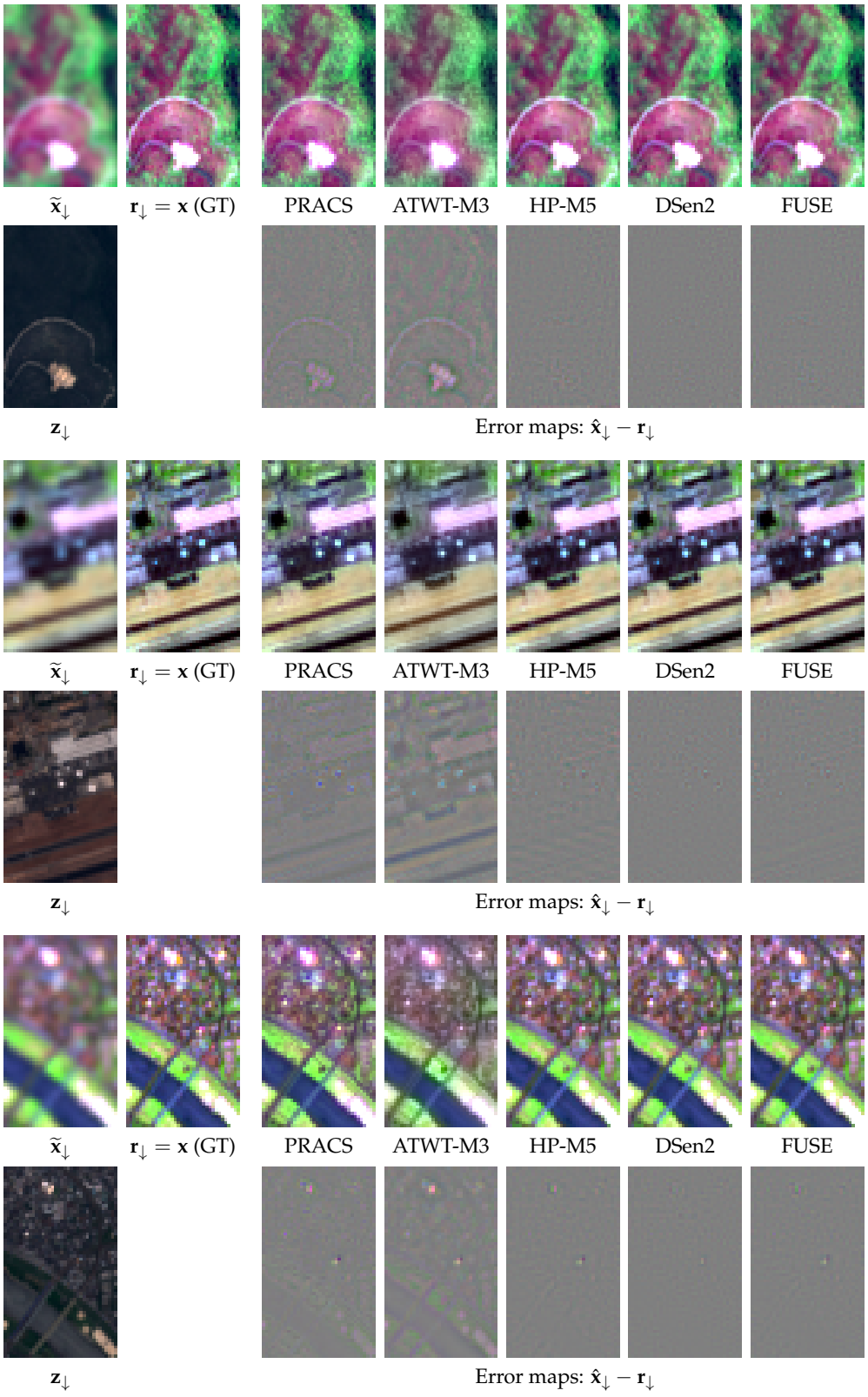


Figure 7. Reduced-resolution samples. Bottom images (Columns 3–7) show the difference with the ground-truth (GT).

4. Discussion

To assess the impact of the proposed changes with respect to the baseline HP-M5, an additional convolutional layer and a composite loss that adds a regularization term and a structural term to

the basic spectral loss (L_1-norm), we also carried out an ablation study. In particular, we have the three-layer scaled version of FUSE and the four-layer version trained without regularization and structural loss terms. These two solutions are also reported in Table 4. As can be seen, except for the SAM index, the full version of FUSE outperforms consistently both scaled versions, with remarkable gains on ERGAS, in the reference-based framework, and on the spatial distortion D_S, in the no-reference context. Focusing on the two ablations, it seems that the use of the composite loss has a relatively better impact compared to the network depth increase. This is particular evident looking at the SAM indicator.

The experimental evaluation presented above confirms the great potential of the DL approach in the context of the data fusion problem at hand, as already seen for pansharpening [35] and single-image super-resolution of natural images [39] a few years ago. The numerical gap between DL methods and the others is consistent and confirmed by visual inspection. In particular, we observe that the use of the additional structural loss term, the most relevant change with respect to our previous models M5 and HP-M5, allowed us to reach and slightly overcome the accuracy level of DSen2. Beside accuracy assessment, it is worth focusing on the related computational burden. DL methods, in fact, are known to be computationally demanding, hence potentially limited for large-scale applicability. Thus, we focused from the beginning on relatively small CNN models. Indeed, the proposed model involves about 28K parameters in contrast to DSen2 which has 2M parameters. In Table 5, we gather a few numbers obtained experimentally on a single GPU Quadro P6000 with 24 GB of memory. For both the proposed and DSen2, we show the GPU memory load and the computational time for the inference with respect to the image size.

Table 5. Computational burden of FUSE and DSen2 at test time for different image sizes.

	GPU Memory (Time)				
Im. Size	**512 × 512**	**512 × 1024**	**1024 × 1024**	**1024 × 2048**	**2048 × 2048**
DSen2	6.6 GB (3.4 s)	8.7 GB (4.3 s)	9.2 GB (7.4 s)	17.4 GB (9.8)	out of memory
FUSE	391 MB (6×0.45 s)	499 MB (6 × 0.47 s)	707 MB (6 × 0.50 s)	1.1 GB (6 × 0.55 s)	1.9 GB (6 × 0.60 s)

As the proposed model is replicated, with different parameters, for each of the six bands to be super-resolved, we assume either a sequential GPU usage (as done in the table) or a parallel implementation, therefore with 6× memory usage but also 6× faster processing. In any case, to have a rough idea of the different burden, it is sufficient to observe that, by using about one third of the memory necessary for DSen2 to super-resolve a 512 × 512 image, FUSE can super-resolve a 16× larger image (2048 × 2048) in the same time slot. In addition, it also has to be considered that, in many applications, the user may be interested in super-resolving a single band, hence saving additional computational and/or memory load. Finally, this picture does not consider the less critical training phase or an eventual fine-tuning stage, which would further highlight the advantage of using a smaller network. To have a rough idea of this, we recall that, according to Lanaras et al. [2], DSen2 was trained in about three days on a NVIDIA Titan Xp 12 GB GPU, whereas the training of our model took about 3 h using a Titan X 12 GB.

5. Conclusions

We presented and validated experimentally a new CNN-based super-resolution method for the 20 m bands of Sentinel-2 images, which blends high-resolution spatial information from the 10 m bands of the same sensor. The proposed network is relatively small compared to other state-of-the-art CNN-based models, such as DSen2, achieving comparable accuracy levels in both numerical and subjective visual terms. Overall, it is worth noticing that DL methods overcome model-based approaches especially in terms of spectral distortion (see Figure 7), which is rather interesting considering that the two band sets to be fused are only partially overlapped/correlated,

as can be seen in Table 1. In light of this, it will be interesting to explore in our future work the extension to 60 m bands of the proposed approach.

Author Contributions: Conceptualization, M.G. and G.S.; methodology, M.G., A.M., R.G, G.R. and G.S.; software, M.G. and A.M.; validation, M.G. and A.M.; investigation, M.G. and A.M.; data curation, M.G., A.M. and R.G.; writing—original draft preparation, G.S.; writing—review and editing, M.G., A.M., R.G., and G.R.; and supervision, G.S.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Drusch, M.; Del Bello, U.; Carlier, S.; Colin, O.; Fernandez, V.; Gascon, F.; Hoersch, B.; Isola, C.; Laberinti, P.; Martimort, P.; et al. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sens. Environ.* **2012**, *120*, 25–36. [CrossRef]
2. Lanaras, C.; Bioucas-Dias, J.; Galliani, S.; Baltsavias, E.; Schindler, K. Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network. *ISPRS J. Photogramm. Remote Sens.* **2018**, *146*, 305–319. [CrossRef]
3. Mura, M.; Bottalico, F.; Giannetti, F.; Bertani, R.; Giannini, R.; Mancini, M.; Orlandini, S.; Travaglini, D.; Chirici, G. Exploiting the capabilities of the Sentinel-2 multi spectral instrument for predicting growing stock volume in forest ecosystems. *Int. J. Appl. Earth Obs. Geoinf.* **2018**, *66*, 126–134. [CrossRef]
4. Castillo, J.A.A.; Apan, A.A.; Maraseni, T.N.; Salmo, S.G. Estimation and mapping of above-ground biomass of mangrove forests and their replacement land uses in the Philippines using Sentinel imagery. *ISPRS J. Photogramm. Remote Sens.* **2017**, *134*, 70–85. [CrossRef]
5. Clevers, J.G.P.W.; Kooistra, L.; Van den Brande, M.M.M. Using Sentinel-2 Data for Retrieving LAI and Leaf and Canopy Chlorophyll Content of a Potato Crop. *Remote Sens.* **2017**, *9*, 405. [CrossRef]
6. Delloye, C.; Weiss, M.; Defourny, P. Retrieval of the canopy chlorophyll content from Sentinel-2 spectral bands to estimate nitrogen uptake in intensive winter wheat cropping systems. *Remote Sens. Environ.* **2018**, *216*, 245–261. [CrossRef]
7. Paul, F.; Winsvold, S.H.; Kääb, A.; Nagler, T.; Schwaizer, G. Glacier Remote Sensing Using Sentinel-2. Part II: Mapping Glacier Extents and Surface Facies, and Comparison to Landsat 8. *Remote Sens.* **2016**, *8*, 575. [CrossRef]
8. Toming, K.; Kutser, T.; Laas, A.; Sepp, M.; Paavel, B.; Nõges, T. First Experiences in Mapping Lake Water Quality Parameters with Sentinel-2 MSI Imagery. *Remote Sens.* **2016**, *8*, 640. [CrossRef]
9. Immitzer, M.; Vuolo, F.; Atzberger, C. First Experience with Sentinel-2 Data for Crop and Tree Species Classifications in Central Europe. *Remote Sens.* **2016**, *8*, 166. [CrossRef]
10. Pesaresi, M.; Corbane, C.; Julea, A.; Florczyk, A.J.; Syrris, V.; Soille, P. Assessment of the Added-Value of Sentinel-2 for Detecting Built-up Areas. *Remote Sens.* **2016**, *8*, 299. [CrossRef]
11. Gargiulo, M.; Mazza, A.; Gaetano, R.; Ruello, G.; Scarpa, G. A CNN-Based Fusion Method for Super-Resolution of Sentinel-2 Data. In Proceedings of the 2018 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2018, Valencia, Spain, 22–27 July 2018; pp. 4713–4716.
12. Gargiulo, M.; Dell'Aglio, D.A.G.; Iodice, A.; Riccio, D.; Ruello, G. A CNN-Based Super-Resolution Technique for Active Fire Detection on Sentinel-2 Data. *arXiv* **2019**, arXiv:1906.10413.
13. Tzelidi, D.; Stagakis, S.; Mitraka, Z.; Chrysoulakis, N. Detailed urban surface characterization using spectra from enhanced spatial resolution Sentinel-2 imagery and a hierarchical multiple endmember spectral mixture analysis approach. *J. Appl. Remote Sens.* **2019**, *13*, 016514. [CrossRef]
14. Zhang, M.; Su, W.; Fu, Y.; Zhu, D.; Xue, J.H.; Huang, J.; Wang, W.; Wu, J.; Yao, C. Super-resolution enhancement of Sentinel-2 image for retrieving LAI and chlorophyll content of summer corn. *Eur. J. Agron.* **2019**, *111*, 125938. [CrossRef]
15. Yokoya, N.; Yairi, T.; Iwasaki, A. Coupled Nonnegative Matrix Factorization Unmixing for Hyperspectral and Multispectral Data Fusion. *IEEE Trans. Geosci. Remote. Sens.* **2012**, *50*, 528–537. [CrossRef]

16. Lanaras, C.; Baltsavias, E.; Schindler, K. Hyperspectral Super-Resolution by Coupled Spectral Unmixing. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 3586–3594.
17. Hong, D.; Yokoya, N.; Chanussot, J.; Zhu, X.X. An Augmented Linear Mixing Model to Address Spectral Variability for Hyperspectral Unmixing. *IEEE Trans. Image Process.* **2019**, *28*, 1923–1938. [CrossRef] [PubMed]
18. Ibarrola-Ulzurrun, E.; Drumetz, L.; Marcello, J.; Gonzalo-Martín, C.; Chanussot, J. Hyperspectral Classification Through Unmixing Abundance Maps Addressing Spectral Variability. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 4775–4788. [CrossRef]
19. Shah, V.P.; Younan, N.H.; King, R.L. An Efficient Pan-Sharpening Method via a Combined Adaptive PCA Approach and Contourlets. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 1323–1335. [CrossRef]
20. Tu, T.M.; Su, S.C.; Shyu, H.C.; Huang, P.S. A new look at IHS-like image fusion methods. *Inf. Fusion* **2001**, *2*, 177–186. [CrossRef]
21. Chavez, P.; Anderson, J. Comparison of three different methods to merge multiresolution and multispectral data: Landsat TM and SPOT panchromatic. *Photogramm. Eng. Remote Sens.* **1991**, *57*, 295–303.
22. Ranchin, T.; Wald, L. Fusion of high spatial and spectral resolution images: the ARSIS concept and its implementation. *Photogramm. Eng. Remote Sens.* **2000**, *66*, 49–61.
23. Fasbender, D.; Radoux, J.; Bogaert, P. Bayesian Data Fusion for Adaptable Image Pansharpening. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 1847–1857. [CrossRef]
24. Garzelli, A. Pansharpening of Multispectral Images Based on Nonlocal Parameter Optimization. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 2096–2107. [CrossRef]
25. Palsson, F.; Sveinsson, J.; Ulfarsson, M. A New Pansharpening Algorithm Based on Total Variation. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 318–322. [CrossRef]
26. Vivone, G.; Alparone, L.; Chanussot, J.; Mura, M.D.; Garzelli, A.; Licciardi, G.A.; Restaino, R.; Wald, L. A Critical Comparison Among Pansharpening Algorithms. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 2565–2586. [CrossRef]
27. Du, Y.; Zhang, Y.; Ling, F.; Wang, Q.; Li, W.; Li, X. Water bodies' mapping from Sentinel-2 imagery with modified normalized difference water index at 10-m spatial resolution produced by sharpening the SWIR band. *Remote Sens.* **2016**, *8*, 354. [CrossRef]
28. Wang, Q.; Shi, W.; Li, Z.; Atkinson, P.M. Fusion of Sentinel-2 images. *Remote Sens. Environ.* **2016**, *187*, 241–252. [CrossRef]
29. Vaiopoulos, A.D.; Karantzalos, K. PANSHARPENING ON THE NARROW VNIR AND SWIR SPECTRAL BANDS OF SENTINEL-2. *ISPRS Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci.* **2016**, *XLI-B7*, 723–730. [CrossRef]
30. Park, H.; Choi, J.; Park, N.; Choi, S. Sharpening the VNIR and SWIR Bands of Sentinel-2A Imagery through Modified Selected and Synthesized Band Schemes. *Remote Sens.* **2017**, *9*, 1080. [CrossRef]
31. Gašparović, M.; Jogun, T. The effect of fusing Sentinel-2 bands on land-cover classification. *Int. J. Remote Sens.* **2018**, *39*, 822–841. [CrossRef]
32. Brodu, N. Super-Resolving Multiresolution Images With Band-Independent Geometry of Multispectral Pixels. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 4610–4617. [CrossRef]
33. Lanaras, C.; Bioucas-Dias, J.; Baltsavias, E.; Schindler, K. Super-Resolution of Multispectral Multiresolution Images From a Single Sensor. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Honolulu, HI, USA, 21–26 July 2017.
34. Paris, C.; Bioucas-Dias, J.; Bruzzone, L. A hierarchical approach to superresolution of multispectral images with different spatial resolutions. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; pp. 2589–2592.
35. Masi, G.; Cozzolino, D.; Verdoliva, L.; Scarpa, G. Pansharpening by Convolutional Neural Networks. *Remote Sens.* **2016**, *8*, 594. [CrossRef]
36. Scarpa, G.; Vitale, S.; Cozzolino, D. Target-Adaptive CNN-Based Pansharpening. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 5443–5457. [CrossRef]
37. Yang, J.; Fu, X.; Hu, Y.; Huang, Y.; Ding, X.; Paisley, J. PanNet: A Deep Network Architecture for Pan-Sharpening. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017. [CrossRef]

38. Gargiulo, M. Advances on CNN-based super-resolution of Sentinel-2 images. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Saint Petersburg, Russia, 1–4 July 2019.
39. Dong, C.; Loy, C.; He, K.; Tang, X. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 295–307. [CrossRef] [PubMed]
40. Kim, J.K.L.; Lee, K.M. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1646–1654.
41. Yuan, Q.; Wei, Y.; Meng, X.; Shen, H.; Zhang, L. A Multiscale and Multidepth Convolutional Neural Network for Remote Sensing Imagery Pan-Sharpening. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2018**, *11*, 978–989. [CrossRef]
42. Wald, L.; Ranchin, T.; Mangolini, M. Fusion of satellite images of different spatial resolution: Assessing the quality of resulting images. *Photogramm. Eng. Remote Sens.* **1997**, *63*, 691–699.
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
44. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 448–456.
45. Johnson, J.; Alahi, A.; Li, F. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. *arXiv* **2016**, arXiv:1603.08155.
46. Jiang, Y.; Ding, X.; Zeng, D.; Huang, Y.; Paisley, J. Pan-sharpening with a Hyper-Laplacian Penalty. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
47. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
48. Ruder, S. An overview of gradient descent optimization algorithms. *arXiv* **2016**, arXiv:1609.04747.
49. Wang, Z.; Bovik, A. A universal image quality index. *IEEE Signal Process Lett.* **2002**, *9*, 81–84. [CrossRef]
50. Wald, L. Data Fusion: Definitions and Architectures–Fusion of Images of Different Spatial Resolutions. In *Les Presses de l'École des Mines*; Presses des Mines: Paris, France, 2002.
51. Chang, C.I. Spectral information divergence for hyperspectral image analysis. In Proceedings of the IEEE 1999 International Geoscience and Remote Sensing Symposium IGARSS'99 (Cat. No. 99CH36293), Hamburg, Germany, 28 June–2 July 1999; Volume 1, pp. 509–511.
52. Zhou, J.; Civco, D.; Silander, J. A wavelet transform method to merge Landsat TM and SPOT panchromatic data. *Int. J. Remote Sens.* **1998**, *19*, 743–757. [CrossRef]
53. Alparone, L.; Aiazzi, B.; Baronti, S.; Garzelli, A.; Nencini, F.; Selva, M. Multispectral and panchromatic data fusion assessment without reference. *Photogramm. Eng. Remote Sens.* **2008**, *74*, 193–200. [CrossRef]
54. Gillespie, A.R.; Kahle, A.B.; Walker, R.E. Color enhancement of highly correlated images. II. Channel ratio and "chromaticity" transformation techniques. *Remote Sens. Environ.* **1987**, *22*, 343–365. [CrossRef]
55. Khan, M.M.; Chanussot, J.; Condat, L.; Montanvert, A. Indusion: Fusion of multispectral and panchromatic images using the induction scaling technique. *IEEE Geosci. Remote Sens. Lett.* **2008**, *5*, 98–102. [CrossRef]
56. Choi, J.; Yu, K.; Kim, Y. A new adaptive component-substitution-based satellite image fusion by using partial replacement. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 295–309. [CrossRef]
57. Aiazzi, B.; Alparone, L.; Baronti, S.; Garzelli, A.; Selva, M. MTF-tailored multiscale fusion of high-resolution MS and Pan imagery. *Photogramm. Eng. Remote Sens.* **2006**, *72*, 591–596. [CrossRef]