A CNN-Based Sentinel-2 Image Super-Resolution Method Using Multiobjective Training

Vlad Vasilescu, Mihai Datcu, *Fellow, IEEE*, and Daniela Faur, *Member, IEEE*

***Abstract*— Deep learning methods have become ubiquitous tools in many Earth observation applications, delivering state-of-the-art results while proving to generalize for a variety of scenarios. One such domain concerns the Sentinel-2 (S2) satellite mission, which provides multispectral images in the form of 13 spectral bands, captured at three different spatial resolutions: 10, 20, and 60 m. This research aims to provide a super-resolution mechanism based on fully convolutional neural networks (CNNs) for upsampling the low-resolution (LR) spectral bands of S2 up to 10-m spatial resolution. Our approach is centered on attaining good performance with respect to two main properties: consistency and synthesis. While the synthesis evaluation, also known as Wald's protocol, has spoken for the performance of almost all previously introduced methods, the consistency property has been overlooked as a viable evaluation procedure. Recently introduced techniques make use of sensor's modulation transfer function (MTF) to learn an approximate inverse mapping from LR to high-resolution images, which is on a direct path for achieving a good consistency value. To this end, we propose a multiobjective loss for training our architectures, including an MTF-based mechanism, a direct input–output mapping using synthetically degraded data, along with direct similarity measures between high-frequency details from already available 10-m bands, and super-resolved images. Experiments indicate that our method is able to achieve a good tradeoff between consistency and synthesis properties, along with competitive visual quality results.**

***Index Terms*— Consistency, convolutional neural networks (CNNs), Sentinel-2 (S2), super-resolution, synthesis.**

I. Introduction

THE Sentinel-2 (S2) satellite mission consists of a constellation of two identical satellites (Sentinel-2A and Sentinel-2B) designed to operate simultaneously for continuous environment monitoring with a high revisit frequency. Accurate monitoring is possible due to their multispectral instrument (MSI), which delivers multispectral images in the form of 13 bands, with different spectral profiles, acquired at three different spatial resolutions—10 m (B2, B3, B4, and B8), 20 m (B5, B7, B7, B8a, B11, and B12), and 60 m (B1, B9, and B10). S2 data come in the form of tiles spanning an approximate area of 100 × 100 km and can be downloaded for free from Copernicus Services Data Hub.[1] The various utilizations for such multispectral information include monitoring natural disasters, agricultural production, and deforestation oversight. However, the level of details is limited due to the lack of high-resolution (HR) profiles for 20- and 60-m bands, which narrows down the quality of some spectral indices used for evaluating and estimating application-specific characteristics. This motivates the development of methods for constructing HR bands that enclose the physical characteristics (light reflectance value distribution) of low-resolution (LR) bands while increasing the number of details in the spatial domain up to the maximum spatial resolution available in the current dataset.

Image fusion has maintained a popular position in Earth observation applications, serving a central role in monitoring the evolution of different environmental areas [1]. One such class of methods concerns multispectral image fusion [2], [3], which combines information from various multispectral images in order to increase the level of details in LR images. Approaches included in this category have shown applicability in super-resolving all LR S2 bands up to 10-m ground sampling distance (GSD). Applications that could benefit from a full set of HR images include time series evaluation of crop fields [4], [5], deforestation monitoring [6], ship detection and recognition [7], and so on. Recently proposed methods include model-based approaches, constructed as inverse problems that include a degradation process applied on the super-resolved bands, and machine learning-based methods, which directly combine the HR and LR bands to produce super-resolved images by exploiting interband information.

Recent deep learning architectures, mainly convolutional neural networks (CNNs), have imposed a great interest among different Earth observation applications [8], [9], through their ability of automatically learning complex spatial relationships. S2 image super-resolution is one particular domain that benefited from the power of representation of neural networks [10], [11], [12]. A representative method for this class was introduced in [13] as *DSen2*, in which the authors train two separate networks for upsampling the 20- and 60-m bands, both following a ResNet [14] architecture. The training process is designed in a supervised manner, constructing a synthetic

Manuscript received 19 July 2022; revised 24 October 2022 and 8 December 2022; accepted 23 January 2023. Date of publication 27 January 2023; date of current version 9 February 2023. This work was supported by CENTURION-H2020-SPACE-2018-2020–"Copernicus Datacube and AI Services for Society, Industry and New Market Generation", funded by EUROPEAN COMMISSION through European Health and Digital Executive Agency under Grant 101004311. *(Corresponding author: Vlad Vasilescu.)*

Vlad Vasilescu is with the Speech and Dialogue Laboratory and the Center for Spatial Information, University Politehnica of Bucharest, 060042 Bucharest, Romania (e-mail: vlad.vasilescu2111@upb.ro).

Mihai Datcu is with the Remote Sensing Technology Institute, EO Data Science, German Aerospace Center (DLR), 82234 Weßling, Germany, and also with the Center for Spatial Information, University Politehnica of Bucharest, 060042 Bucharest, Romania (e-mail: mihai.datcu@dlr.de).

Daniela Faur is with the Center for Spatial Information, University Politehnica of Bucharest, 060042 Bucharest, Romania (e-mail: daniela.faur@upb.ro).

Digital Object Identifier 10.1109/TGRS.2023.3240296

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

dataset by degrading the original bands, further considering as input–output pairs the degraded-observed images. This method has been shown to deliver competing results for a wide range of environments, suggesting good generalization capabilities. Other approaches rely on the use of GANs for generating HR patches from the original LR bands [15], [16], [17], training the generator network using a synthetic dataset, constructed in the same degradation-based manner.

A subset of model-based approaches incorporate a downsampling mechanism that mimics the sensor's modulation transfer function (MTF). Such transformation should restore a super-resolved image back to the originally observed one, produced by the sensor. Area-to-point regression kriging (ATPRK) [18] is a pansharpening technique originally tested on WorldView-2 and Landsat images, subsequently adapted for S2 super-resolution in [19]. In ATPRK, the super-resolved band is modeled as a linear combination of HR bands, with coefficients determined through least square minimization, followed by an adaptive averaging of the regression residuals for each pixel's neighborhood. SupReME [20] is a method that depends on the observation model of the imaging, i.e., the sensor's MTF, and formulates the super-resolution process as an inverse problem, exploiting the correlation between spectral bands by projecting them onto a lower dimensional space. S2Sharp [21] is an approach related to SupReME, optimizing the same objective function while accounting for interband correlations, but in this case, the low-dimensional space is estimated for each optimization step, while in the case of SupReME [20], it is defined beforehand for all steps. Sen2res [22] is a method based on extracting the local information from each LR band, i.e., reflectance values distribution, and exploiting the high-frequency characteristics of HR bands. The super-resolved patches are constructed as linear combinations of subpixel constituents from HR bands, using least square optimization for estimating the weighting coefficients. Another technique, introduced as SSSS [23], also formulates the super-resolution through the use of an MTF degradation model, additionally incorporating convex regularization terms intended for learning a self-similarity graph. This method has been shown to exhibit solid results, especially for upsampling the 60-m bands [24].

A more recent method introduced as S2SUCNN [25] combines these two classes of techniques, training a CNN in an unsupervised manner by introducing an MTF-based degradation layer as their final processing step. Their method is based on the idea of deep image prior [26], which provides a solving mechanism for inverse problems, such as denoising, inpainting, and super-resolution [27], [28], using neural network structures to model a reliable inverse mapping through standard gradient-based optimization, under a difference minimization objective.

He and Siu [29] applied Gaussian process regression (GPR) for single-image super-resolution, by fitting such a model on the local structures of each pixel and using a two-stage process to recover and refine the super-resolved image. Blix et al. [30], [31] utilized GPR for estimating quad-polarimetric parameters from dual-polarimetric synthetic aperture radar observations, providing insight by constructing uncertainty maps through the fit GPR model, thus establishing the level of trust in their predictions. S2 super-resolution could therefore benefit on these aspects, providing a new way to assess super-resolution results through measures indicated by such uncertainty maps. This is, however, yet to become a complete evaluation framework due to the difficulty of modeling multiple response variables, as discussed in [32], which for S2 super-resolution is a necessity when super-resolving multiple LR bands at once. Another obstacle is represented by the huge amount of input–output pairs a GPR model should be optimized on, in order to provide good generalization capabilities for different environmental areas.

In this article, we present a multiobjective training approach for fully CNNs, aimed for ensuring a good tradeoff between properties of consistency, synthesis, and enhanced visual details of super-resolved S2 bands. Our contributions can be summarized as follows.

1) We designed and implemented a super-resolution mechanism based on fully convolutional networks for upsampling the 20- and 60-m bands of S2, except for band B10 (we excluded band B10 from our study for two reasons: it exhibits poor radiometric quality[2] and it does not contain relevant surface information, as it is mainly used in applications dealing with cloud covered areas, e.g., cirrus cloud detection).
2) We formulated a multiobjective training for the proposed architectures, aimed at ensuring good super-resolution results with respect to two properties of importance: consistency and synthesis.
3) We included a direct similarity maximization term between high-frequency information from already available 10-m bands and upsampled results for 20- and 60-m bands, resulting in competitive visual results in terms of detail authenticity.

The remaining of this article is organized as follows. Section II presents the mathematical notations used throughout this article, along with an in-depth discussion about the elements implied by our method. Section III offers a view over super-resolution evaluation protocols, providing intuition for the adopted mechanisms in our method, and also their role in assuring a high level of reliability. Section IV begins with details regarding our method and a description for the multiobjective training process, followed by Section V, which includes discussions about various experimental results. The concluding remarks are given in Section VI.

II. Related Methods

A. Notations

Each S2 acquisition can be viewed as a 3-D tensor of spectral bands, each with its own spatial dimensions and spatial resolution. Let us denote the set of spectral bands for one acquisition as $\{X_{10}, X_{20}, X_{60}\}$, where X_r represents the channelwise concatenation of spectral bands with spatial resolution r. Considering each band from X_{10} to be of spatial dimension $H \times W$, the spatial dimensions of bands from

[2]Sentinel-2 L1C data quality report 2022.

X_{20} and X_{60} are $(H/2) \times (W/2)$ and $(H/6) \times (W/6)$, respectively. $X_r(i, j)$ denotes the vector of reflectance values from all spectral bands with spatial resolution r at location (i, j). The spatial degradation (blurring + downsampling) by a factor of s is denoted as $(\cdot) \downarrow_s$.

B. Deep Learning Methods

Deep learning methods have been extensively used as stand-alone solutions for S2 super-resolution, fusing bands at different spatial resolutions according to the process described by various neural network architectures. Such computational structures allow for obtaining the most representative abstract features for the task at hand, without any prior knowledge on the data acquisition process. However, these methods often require large training sets for achieving good generalization on unseen data, which is, by default, unachievable in certain domains. Since there is no ground truth for the observed 60- and 20-m bands of S2 data, the majority of methods based solely on neural network structures are trained on synthetically constructed LR–HR image pairs, in a supervised training paradigm. Given a neural network structure $T(\cdot)$, most approaches learn the following mapping (either for $\times 2$ or $\times 6$ upsampling):

$$T(X_{10} \downarrow_6, X_{20} \downarrow_6, X_{60} \downarrow_6) \rightarrow X_{60} \quad (1)$$

$$T(X_{10} \downarrow_2, X_{20} \downarrow_2) \rightarrow X_{20}. \quad (2)$$

The optimized structure $T(\cdot)$ is subsequently used on real, nonsynthetically degraded data.

One specific architecture recently used for super-resolving S2 bands is that of conditional GANs [17], [33], which trains the generator network by conditioning them on available information to produce super-resolved bands intended to fool a supposedly optimal discriminator network—optimal in that it distinguishes as good as possible between real S2 bands and bands produced by the generator. However, this type of neural architecture was found out to be very hard to train due to the necessity of finding an equilibrium state for the min–max optimization problem, often resulting in suboptimal final states, which leads to inconsistent super-resolved image patches. Since the generated images are significantly dependent on the ability of the discriminator to capture realistic HR patterns when distinguishing between real and generated images, the final image quality is tightly bounded by its performance.

One emerging network architecture in the field of computer vision, initially adopted by various natural language processing applications, is represented by transformer models [34]. Given the attention mechanism such models imply, they are able to capture wider spatial relationships, however, with the disadvantage that they require substantially more training than standard CNNs. One application in remote sensing that leverages the attributes of transformers was discussed in [35] for multi-image super-resolution of images provided by PROBA-V satellite. To reduce the transformer necessity for intensive training, the authors propose optimizing the model for each input acquisition, eliminating the reliance on a training dataset. Another super-resolution approach using transformers for remote sensing applications is described in [36], combining a CNN-based encoder with a transformer model, the latter acting on embedded image patches.

C. MTF-Based Methods

These methods are based on computations that simulate the sensor's transfer function (MTF), planning the super-resolution task as an inverse problem. Thus, for each LR image, a search in the HR image space is performed, imposing that the degraded HR solution should yield as close as possible to the LR image. Usually, the degradation process is described as follows, for an observed image X and its unknown super-resolved version $Y \in \mathbb{R}^{H \times W}$:

$$X = Y \downarrow_s = (Y * g_\sigma) * d_s \in \mathbb{R}^{\frac{H}{s} \times \frac{W}{s}} \quad (3)$$

where $*$ indicates a depthwise convolution operation, g_σ represents a 2-D Gaussian smoothing filter with variance σ, and d_s is an $s \times s$ averaging filter, applied with stride s. Thus, the problem of finding the unknown HR image Y can be formulated as follows:

$$Y = \arg\min_{\hat{Y}} ||X - (\hat{Y} * g_\sigma) * d_s||_p \quad (4)$$

for some distance metric $|| \cdot ||_p$ (usually, $p \in \{1, 2\}$). This problem is ill-posed, allowing for multiple possible solutions for the unknown image Y. Nguyen et al. [25] provided an iterative way of finding the above solution through the use of a CNN for constructing the possible image candidates $\hat{Y}$ that would minimize (4). However, their approach implies optimizing the network separately for each S2 acquisition, which could be bothersome given a computer with lower computing capabilities.

Applying such formulation in a super-resolution context aims at ensuring the property of consistency [37]—every super-resolved image once degraded to their initial spatial resolution should resemble as good as possible the originally observed image. This is a necessary property, but not sufficient to ensure good super-resolution results [37], given the ill-posedness characteristic of the problem at hand. The advantage of formulating the super-resolution problem in such manner is the possibility of training directly on observed data (as in [25]), eliminating the need of constructing synthetic input–output image pairs. Shocher et al. [38] raised the concern of testing a model trained solely on synthetically constructed data in real-world scenarios, where it may yield unsatisfactory results.

Finding a solution that would minimize (4) raises the following question: how much detail could $\hat{Y}$ encompass such that its degraded version does not significantly deviate from X? Since the super-resolution problem does not directly imply any term regarding the level of spatial detail, including an additional detail-related objective may force an optimization path to a visually better solution, while still preserving the consistency of results. We further test this assumption by combining a consistency-related objective with another that pushes the super-resolved image into maximizing the similarity between its level of detail with one of the other HR images.

(a) (b) (c) (d)

Fig. 1. S2 areas used for testing. Bands B2, B3, and B4 were used for constructing the RGB representation. (a) Bucharest, Romania. (b) Dilo, Ethiopia. (c) Tyrrhenian Sea, Italy. (d) NW Canada.

III. Evaluation Protocols

One of the mostly adopted evaluation protocols for S2 super-resolution methods is by verifying the synthesis property [37], [39] This protocol requires the method to be evaluated on synthetically degraded data, taking the observed images as the desired targets. This has been regarded as a sufficient property for ensuring good super-resolution results. However, the nature of the problem itself states that there exists no unique solution for the super-resolved image. This leads to the question of existence for a set of HR images, not necessarily similar to the observed one, that could still be considered viable candidate solutions.

A relatively unused evaluation protocol in the literature is the one based on assessing the consistency property, which states that degrading the obtained super-resolved HR band should yield a result close to the observed LR band. Thus, attaining a good consistency property does not impose only a specific solution for the super-resolved band but allows for a rather broad set of HR solutions. Advantages in using this property as the main objective for training a model are the removal for the need of constructing a synthetically degraded dataset to fit the model on, along with a wider range for possible output solutions. However, the latter does not account for the visual quality of each such possible solution, hence not necessarily implying good super-resolution results. As stated in [37], this property alone is not sufficient to always ensure reliable super-resolved images, but it is rather necessary.

While for consistency evaluation, one does not need an HR reference image to measure the performance, such an analysis does not consider all high-frequency details contained in the super-resolved band. An evaluation, which uses the full scale of predicted HR bands, is performed by measuring the quality no reference (QNR) metrics, which has received popularity among pan-sharpening applications [40], [41]. The QNR is usually described in terms of pan-sharpening evaluation, by considering a single HR image, which guides the fusion process of multiple LR bands. For multiple available HR bands (as in the S2 case), the QNR metric can be extended as follows:

$$\mathrm{QNR} = (1 - D_\lambda)^{t_1}(1 - D_S)^{t_2} \tag{5}$$

where

$$D_\lambda = \sqrt[p]{\frac{1}{L(L-1)} \sum_{\substack{l,r=1 \\ l \neq r}}^{L} \left| Q\left(\widehat{\mathbf{LR}}_l, \widehat{\mathbf{LR}}_r\right) - Q(\mathbf{LR}_l, \mathbf{LR}_r) \right|^p}$$

$$D_S = \sqrt[q]{\frac{1}{L} \sum_{l=1}^{L} \sum_{r=1}^{R} \left| Q\left(\widehat{\mathbf{LR}}_l, \mathbf{HR_r}\right) - Q(\mathbf{LR}_l, \mathbf{HR_r} \downarrow) \right|^q}.$$

Here, L and R are the numbers of LR and HR bands, respectively, Q is the universal image quality index [42], $\widehat{(\cdot)}$ denotes a super-resolved band, $\mathrm{HR}_i/\mathrm{LR}_i$ denotes the ith HR/LR observed band. In the S2 case, $\mathbf{HR} = X_{10}$ and $\mathbf{LR}$ is either X_{20} or X_{60}. All further results concerning this evaluation are obtained for $t_1 = t_2 = p = q = 1$, thus equal contribution from D_λ and D_S. A high QNR metric implies that the fusion process preserves the similarity between LR bands (measured through the spectral distortion D_λ) and the similarity between the degraded HR bands and original LR bands (spatial distortion D_S).

Another common evaluation is based on visual inspection, presenting a side-by-side comparison of observed LR bands and their super-resolved solutions. Displaying areas of interest analyzed through different super-resolution methods may provide meaningful insight for constructing a better comparison between these techniques. However, one should question the detail authenticity induced by each method, knowing the desired spatial resolution the images should be at. Each visual evaluation could therefore benefit from including an already available HR band next to the illustrations, providing authentic high-frequency details as guidance for the visual quality assessment.

IV. Proposed Method

For training/testing data, we selected Level-1C products, each tile spanning an area of $100 \times 100\ km^2$, from both Sentinel-2A and Sentinel-2B satellites. We used 14 such tiles for training and four for testing, capturing a diversity of environments. The RGB representation of test areas is shown in Fig. 1. Since S2 image data come in tiles with a spatial dimension of $10\,980 \times 10\,980$ pixels for 10-m bands, directly feeding them to a neural network would be impractical. Thus, we partitioned each training/testing tile into 192×192 patches to be directly processed by the network. Before any processing, the raw reflectance value is divided by 2000 (as in [13], for numerical stability). All networks were trained in typical settings: initializing the weights with Glorot uniform [43] and using Adam optimizer [44] with a learning rate of 10^{-4} with hyperparameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

The proposed CNN architecture is presented in Fig. 2, for $6\times$ super-resolution. We trained two such networks, one for

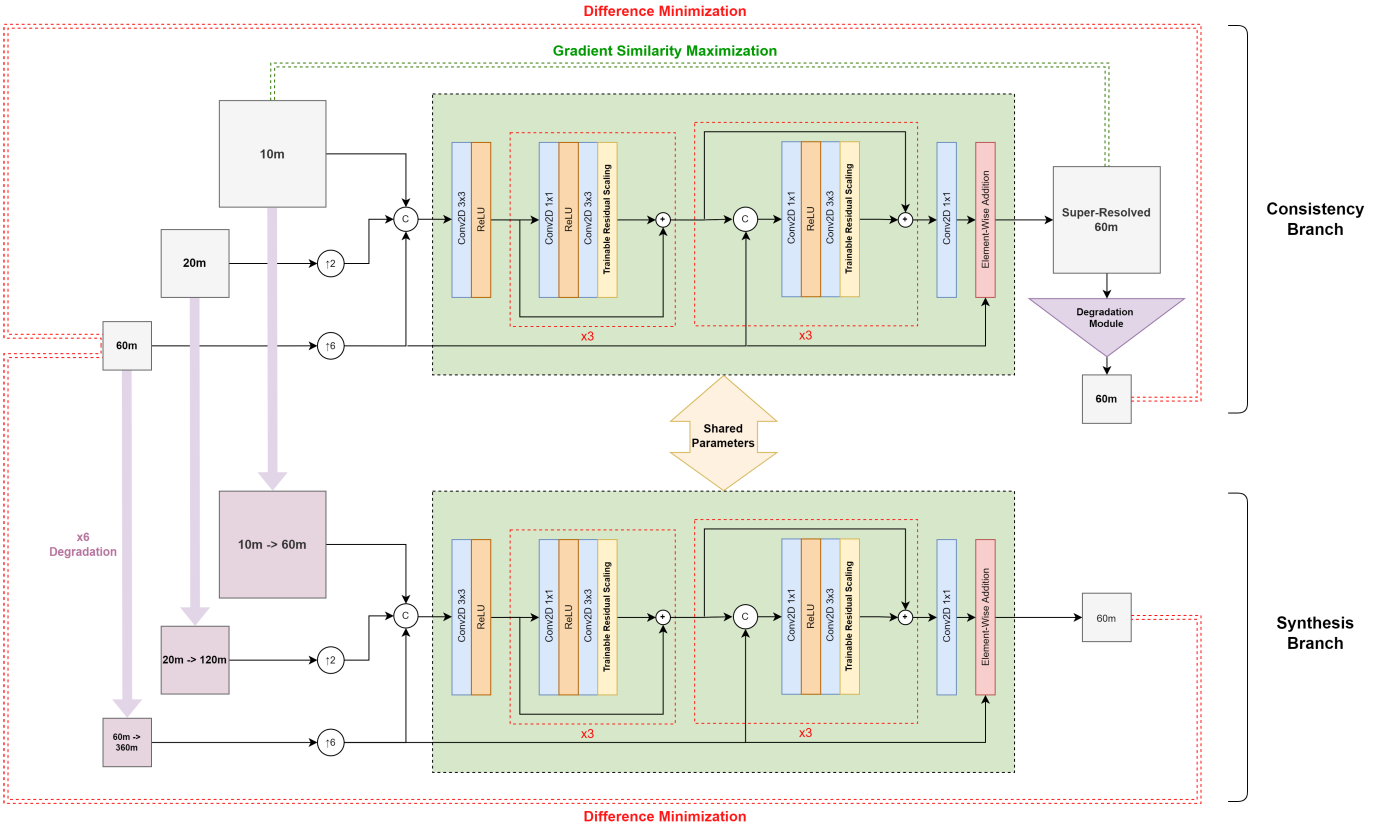


Fig. 2. Proposed neural network architecture (green boxes) and workflow diagram for 6× super-resolution. The two branches resemble the two consecutive forward propagations through the same network: one for the real S2 image patches and one for their degraded counterparts. Their outputs are further used for computing different error terms, from which a linear combination is formed to result in the final cost. The consistency branch aims at obtaining realistic high-frequency details while preserving the original physical characteristics in the super-resolved images. The synthesis branch, acting on degraded data, is designed to approximate a direct fusion to available ground truth, implying a supervised learning step in a reduced-resolution context.

super-resolving the 20-m bands and one for the 60-m bands. In the following, the network for super-resolving the 60-m bands is presented. All 20- and 60-m patches are upsampled using bilinear interpolation up to the spatial dimensions of 10-m patches, before any processing implied by the CNN. The network received as input the channelwise concatenation of these upsampled patches, which are first processed by a 2-D convolution layer to increase the number of channels. All 2-D convolution layers contain 64 filters of dimension 3×3 or 1×1. We used ReLU activation for all cases, due to its efficiency in terms of computation time and training stability. Following the first convolution layer, a series of residual blocks is utilized, each block implying two convolution operations. Based on the work in [45], the first convolution layer uses 1×1 filters, while the second one uses 3×3 filters, encapsulating a ReLU nonlinearity. The first three residual blocks are standard, combining the input from the previous residual block with the features computed by the current block. The last three residual blocks operate on the previous features channelwise concatenated with the upsampled 60-m patches, with the intent of retaining some of the radiometric properties from the original bands. Szegedy et al. [45] also introduced a mechanism called residual scaling by which the outputs of the final convolution layer in a residual block are scaled down by a subunitary constant value (usually chosen between 0.1 and 0.3). This additional step has been shown to stabilize training, which may present some benefit in cases where training data distribution spans over a wide domain, such as the multitude of environments captured by S2 images. The same mechanism is used in [13] by choosing a constant value of 0.1 for scaling the residual features. We have also included a similar step in our residual blocks, by introducing a trainable residual scaling layer (see Fig. 2) through which the scaling factor is learned during the optimization process. This avoids the need of arbitrarily setting each scaling factor to a constant value, thus allowing for an adaptive change for the importance of each residual block's computed features, by lowering/increasing their residual scaling factors. Each residual scaling factor is initialized with a fixed value of 0.05 and is constrained to always be positive during the optimization steps. Following the sequence of residual blocks, the computed feature maps are passed to a convolution layer with two 1×1 filters, fusing the previous feature volume in two channels. The final prediction is obtained through an elementwise addition between the original upsampled image patches and the previously computed two-channel features. For the ×2 super-resolution network, the only difference is the removal of 60-m input since it does not bring any helpful information to the process.

Both the proposed architectures, for 2× and 6× super-resolution, are trained using a weighted sum of three different loss terms, mainly aimed at preserving the two previously

discussed properties—consistency and synthesis—, along with a term penalizing the high-frequency details. This is achieved by implying two consecutive steps for each batch optimization: directly feeding the input patches to the network and forcing the degraded output to resemble as close as possible the LR input patches (consistency branch from Fig. 2), degrading the input patches and feeding them to the network in order to produce an output as close as possible to the original LR patches (synthesis branch from Fig. 2). In addition to preserving the consistency of results, the first branch also implies a direct similarity term between the high-frequency details contained in the super-resolved image and real details extracted from a subset of 10-m input patches. The motivation behind this optimization step resides in achieving a visual quality for the super-resolved image that closely matches the level of details contained by observed HR patches.

In the following, we present the training objective for the network tasked with super-resolving the 60-m bands up to 10-m spatial resolution. A similar mechanism can be easily derived for the loss implied in super-resolving the 20-m bands. Let $T(\cdot)$ denote the HR prediction of the proposed neural network model. The loss function $\mathcal{L}$ is described as follows:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{consistency}} + \beta \mathcal{L}_{\text{sym}} + \gamma \mathcal{L}_{\text{synthesis}} \tag{6}$$

$$\mathcal{L}_{\text{consistency}} = ||(T(X_{10}, X_{20}, X_{60})) \downarrow_6 - X_{60}||_1 \tag{7}$$

$$\mathcal{L}_{\text{sym}} = \sum_{j \in \mathcal{J}} (1 - \text{sym}(X_{10,j}, T(X_{10}, X_{20}, X_{60}))) \tag{8}$$

$$\text{sym}(X, Y) = \frac{\langle \nabla_X, \nabla_Y \rangle}{||\nabla_X||_2 ||\nabla_Y||_2} \tag{9}$$

$$\mathcal{L}_{\text{synthesis}} = ||T(X_{10} \downarrow_6, X_{20} \downarrow_6, X_{60} \downarrow_6) - X_{60}||_1 \tag{10}$$

where $\mathcal{J}$ represents the set of 10-m observed bands used for computing $\mathcal{L}_{\text{sym}}$, $X_{10,j}$ is the input patch from the jth 10 m band, $\text{sym}(x, y)$ denotes the similarity function (using cosine similarity), $\langle \cdot, \cdot \rangle$ is the dot product between vectorized images, ∇_X denotes the gradient of image X obtained by linear 2-D filtering with a Laplacian operator and $|| \cdot ||_1$ denotes the L1 norm. Note that before computing the 2-D gradient of an image we applied a 3×3 average smoothing filter to counter the sensitivity to noise of the Laplacian operator. In (6), α, β, and γ are hyperparameters meant for controlling the overall influence of each loss term over the final super-resolution properties. In Section V, we provide insight regarding the influence of the three loss terms included in (6), along with performance measures for our 6× and 2× super-resolution methods for S2 bands.

V. Experiments and Results

Given the three hyperparameters (α, β, and γ) from (6), we trained four separate networks using the architecture described in Section IV, with different loss configurations (further referred to by their corresponding (α, β, and γ) triplet) for 6× super-resolution, in order to examine their influence on full- and reduced-resolution evaluation.

1) $(\alpha, \beta, \gamma) = (1, 1, 1)$, equal contribution of all three loss terms.
2) $(\alpha, \beta, \gamma) = (1, 0.1, 1)$, less focus on minimizing the similarity component.
3) $(\alpha, \beta, \gamma) = (1, 0.1, 0)$, no contribution from the synthesis branch.
4) $(\alpha, \beta, \gamma) = (0, 0, 1)$, only the synthesis component is minimized.

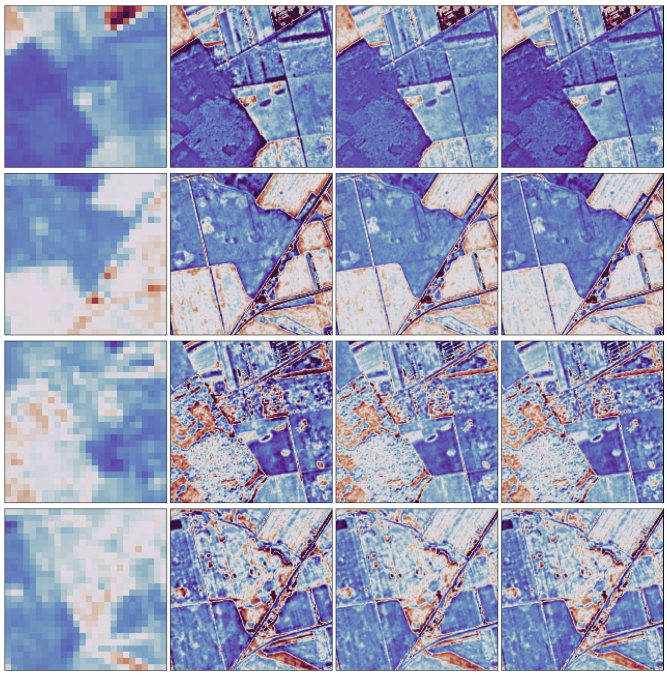


Fig. 3. Visual comparison between three network variants trained for × super-resolution. All results are obtained by feeding the network's real S2 images. (From left to right) Original, (1, 1, 1), (1, 0.1, 1), and (1, 0.1, 0). B1 on the first two rows and B9 on the last two rows.

TABLE I

Comparison Between Networks on 6× Super-Resolution for Reduced-Resolution Evaluation

Method – (α, β, γ)	QNR	B1		B9	
		RMSE	KL	RMSE	KL
(1, 1, 1)	0.9773	22.36	0.0465	24.01	0.0626
(1, 0.1, 1)	0.9774	18.23	0.0205	20.31	0.0286
(1, 0.1, 0)	0.9772	24.25	0.0468	21.80	0.0314
(0, 0, 1)	0.9782	23.33	0.0167	24.16	0.0250

TABLE II

Comparison Between Networks on 6× Super-Resolution for Reduced-Resolution Evaluation

Method – (α, β, γ)	B1		B9	
	RMSE	SSIM	RMSE	SSIM
(1, 0.1, 1)	66.71	0.8574	60.19	0.9032
(1, 0.1, 0)	70.02	0.8544	63.61	0.8966
(0, 0, 1)	78.41	0.8642	72.41	0.9046

The results on full-resolution evaluation for the four networks are shown in Table I and on reduced-resolution evaluation in Table II, as mean values over all test patches. Kullback–Leibler (KL) divergence is measured between the

TABLE III
PERFORMANCE COMPARISON USING WALD'S PROTOCOL FOR 2× SUPER-RESOLUTION ON SYNTHETIC DATA

Data	Method	B5↓	B6↓	B7↓	B8a↓	B11↓	B12↓	mRMSE↓	mSRE↑	mSSIM↑
Italy	Bicubic	35.18	51.38	59.84	63.55	42.66	38.78	48.57	37.79	0.9099
	DSen2	23.63	34.39	40.21	42.12	27.33	25.19	32.15	40.86	**0.9556**
	Proposed	**14.07**	**31.00**	**31.59**	**30.70**	**18.70**	**16.32**	**23.73**	**41.07**	0.9052
Romania	Bicubic	31.77	49.37	64.86	71.64	46.01	45.90	51.59	32.53	0.9454
	DSen2	21.33	33.16	43.53	47.59	29.29	29.63	34.09	36.23	0.9764
	Proposed	**13.31**	**31.25**	**36.00**	**37.68**	**20.03**	**19.04**	**26.22**	**38.58**	**0.9821**
Canada	Bicubic	164.12	171.89	178.40	179.30	50.86	38.23	130.47	23.23	0.9422
	DSen2	109.77	113.81	119.07	117.78	**33.07**	**25.03**	86.42	26.87	**0.9738**
	Proposed	**73.62**	**100.06**	**92.45**	**82.11**	44.79	33.66	**71.12**	**27.88**	0.9559
Ethiopia	Bicubic	33.05	31.79	33.70	34.19	29.47	31.88	32.35	35.82	0.9400
	DSen2	22.33	**21.42**	**22.79**	**22.98**	18.90	20.78	21.53	39.43	0.9740
	Proposed	**11.85**	28.24	30.18	30.65	**15.06**	**12.09**	**21.34**	**40.15**	**0.9781**

Columns B5, B6, B7, B8a, B11, B12 represent the mean RMSE of each band, computed on patches. mSRE is given in decibels. mRMSE, mSRE and mSSIM show the averaged values over all bands

super-resolved band and the original band, by quantizing the range 0–10 000 of reflectance values into 1000 equal width bins in order to construct a discrete probability distribution using the frequency of occurrence in each bin. In Table I, root-mean-squared error (RMSE) is measured between the original LR band and the degraded super-resolved band, while in Table II, RMSE is measured between super-resolved bands and ground truth—similarly for the structural similarity index (SSIM). Note that the mean QNR values are very close between the four networks, indicating little to no discrepancy between results at full-resolution evaluation, not aligning with the other metrics and the clear visual differences shown in Fig. 3. Thus, we decided to exclude the QNR evaluation from all future experiments. Results show that including the synthesis term $\mathcal{L}_{\text{synthesis}}$ during the training process is effective on reduced-resolution evaluation, as shown in Table II, network (1, 0.1, 1) delivering better results than (1, 0.1, 0). Training using only the synthesis term (network (0, 0, 1)) leads to the highest SSIM values for reduced-resolution evaluation but results in the highest RMSE. While network (1, 0.1, 1) obtains slightly lower SSIM values for the same evaluation, it leads to the best RMSE on both 60-m bands, indicating a more favorable solution with respect to radiometric accuracy, while also maintaining a similar visual aspect (Table II). The results on full-resolution evaluation indicate that configuration (0, 0, 1) achieves the highest mean RMSE on band B9 and the second highest on B1. Since both configurations (0, 0, 1) and (1, 0.1, 0) achieve relatively high RMSE values on full- and reduced-resolution evaluation, and given that configuration (1, 0.1, 1) results in the best performance on both evaluations, we concluded that using a mixture of all three components for optimization could lead to a reliable super-resolution algorithm. We found that assigning an equal contribution to all loss terms does not help in finding a good equilibrium state with both good visuals and good evaluation results. Reducing the value of β helps with achieving a good radiometric quality—according to Table I—while also not degrading the visual quality, as observed in Fig. 3. We hypothesize that this behavior is due to the opposing objective of $\mathcal{L}_{\text{sym}}$ and the other two loss terms, raising the need for finding a good tradeoff between visual quality and good consistency and synthesis metrics. Adding too much focus on correlating the super-resolved bands with observed 10-m bands clearly does not guarantee either good consistency or synthesis. This need of achieving a good tradeoff between visual quality and metrics translates into finding the appropriate weighting coefficients for the three loss terms.

In the following, we assessed the performance of both networks for 2× and 6× super-resolution with respect to the consistency and synthesis property. In the case of consistency assessment, the RMSE, signal-to-reconstruction error (SRE), and SSIM are computed between the original LR band and the degraded network output. Note that all results are measured given pairs of images with pixel values from the original S2 spectral bands. For reduced-resolution evaluation, we degrade all bands in order to reduce their spatial resolution s times, where $s \in \{2, 6\}$, feed them to the network, and compare the results with the original LR bands. Following the work in [25], the degradation process is designed according to (3), using a Gaussian kernel with standard deviation σ_s for depthwise filtering, followed by a downsampling operation using an $s \times s$ averaging filter applied with a stride of s. For 2× degradation, we use $\sigma_2 = 1$ with kernel size 7×7, and for 6× degradation, we use $\sigma_6 = 3$ with kernel size 15×15.

We focused on comparing the performance of the proposed model to another S2 super-resolution method based on

TABLE IV
PERFORMANCE COMPARISON USING WALD'S PROTOCOL FOR 2× SUPER-RESOLUTION ON SYNTHETIC DATA

Data	Method	B5↓	B6↓	B7↓	B8a↓	B11↓	B12↓	mRMSE↓	mSRE↑	mSSIM↑
Italy	Bicubic	75.61	109.84	127.58	136.66	102.14	88.86	106.78	31.01	0.7969
	DSen2	45.46	64.64	74.67	79.76	**65.36**	**57.30**	64.53	34.29	**0.8602**
	Proposed	**45.09**	**43.63**	**42.92**	**44.05**	67.32	59.65	**50.44**	**34.75**	0.8434
Romania	Bicubic	65.92	102.14	135.43	151.87	111.09	107.46	112.32	25.21	0.7426
	DSen2	**39.91**	61.75	81.23	91.16	71.51	69.13	69.11	29.40	0.9016
	Proposed	40.61	**49.29**	**46.14**	**47.30**	**69.37**	**68.53**	**53.54**	**31.53**	**0.9280**
Canada	Bicubic	371.67	387.52	398.52	406.83	126.13	93.74	297.40	17.62	0.7349
	DSen2	208.44	218.75	226.15	228.02	**84.16**	**62.17**	171.28	22.01	**0.8887**
	Proposed	**182.88**	**124.01**	**127.10**	**126.33**	141.35	111.26	**135.49**	**23.16**	0.8626
Ethiopia	Bicubic	68.21	66.63	70.67	72.49	70.21	72.53	70.12	28.59	0.7245
	DSen2	40.29	39.80	42.29	**43.94**	42.79	45.30	42.40	32.87	0.9030
	Proposed	**38.31**	**38.12**	**41.72**	44.97	**41.22**	**42.64**	**41.16**	**33.58**	**0.9360**

Columns B5, B6, B7, B8a, B11, B12 represent the mean RMSE of each band, computed on patches. mSRE is given in decibels. mRMSE, mSRE and mSSIM show the averaged values over all bands

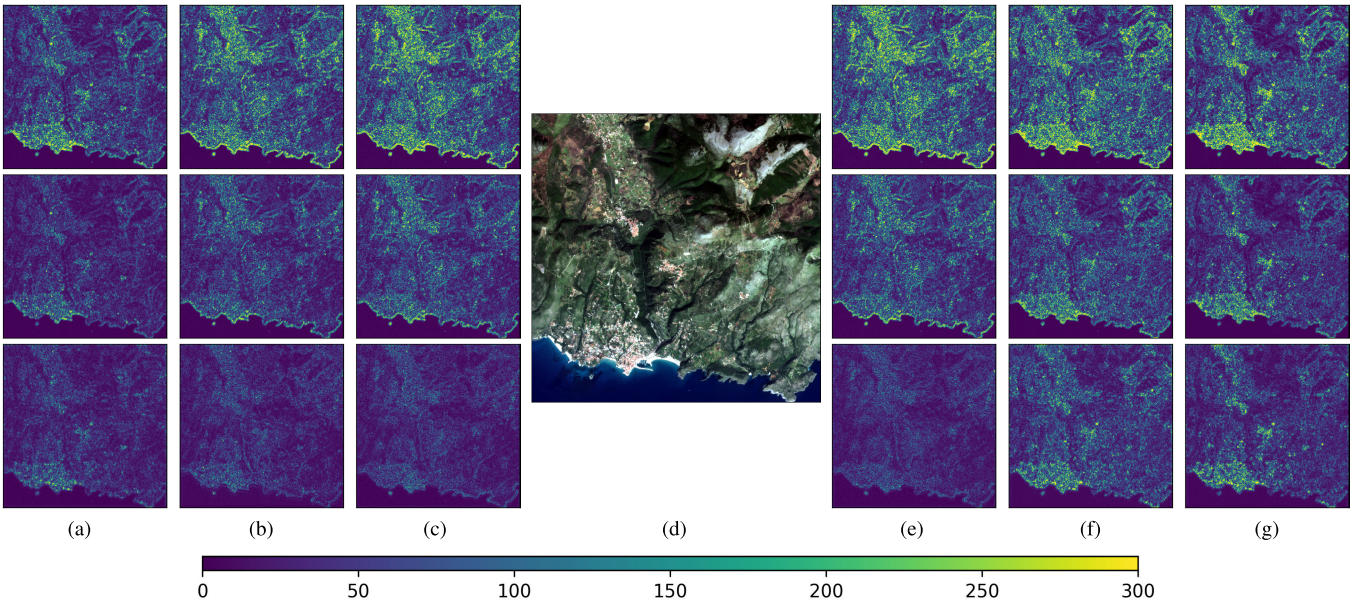


Fig. 4. Absolute differences between super-resolved 20-m bands and ground truth, using Wald's protocol (degraded S2 data). First row: bicubic interpolation. Second row: DSen2 [13]. Third row: proposed method. Region extracted from Italy tile. (a) B5. (b) B6. (c) B7. (d) RGB. (e) B8a. (f) B11. (g) B12.

deep learning techniques, namely, DSen2 [13]. We trained our ×2 super-resolution architecture using the configuration $(\alpha, \beta, \gamma) = (1, 0.1, 1)$. Through multiple experiments, we decided that for bands B5, B11, and B12, a direct similarity with any 10-m band resulted in worse visual and performance results; thus, the loss term from (8) is not considered during the optimization for these three bands. We decided for bands B6, B7, and B8a to include the similarity with band B8 (10 m) during training, after multiple experiments regarding the choice of β and the set of 10-m bands to extract details from. For bands B11 and B12, poor performance when increasing the similarity with a 10-m band could be explained by considering their spectral characteristics since their central wavelength is a lot more distanced from any 10-m band, compared with the other 20-m bands. In Table III, the results for consistency evaluation for 2× super-resolution are presented. Our architecture performed better with regard to averaged results over all 20-m bands, in terms of RMSE and SRE. For the Ethiopia tile, DSen2 [13] performed better on bands B6, B7, and B8a in terms of RMSE, allowing for a better recovery of the observed

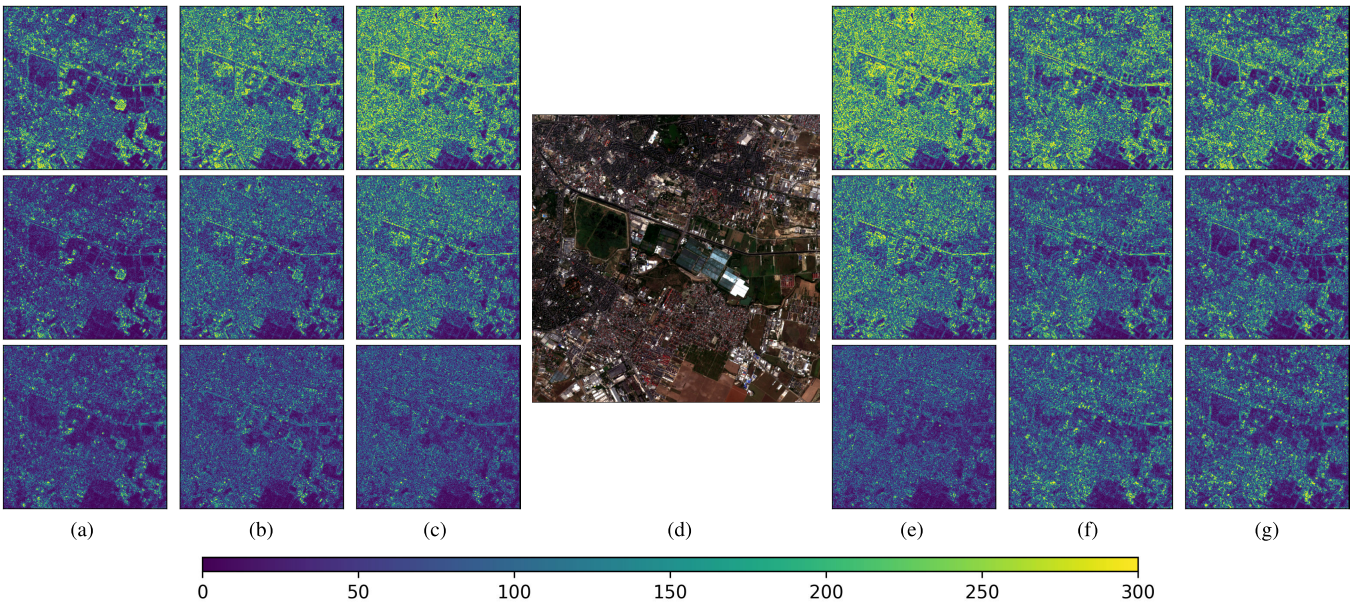


Fig. 5. Absolute differences between super-resolved 20-m bands and ground truth, using Wald's protocol (degraded S2 data). First row: bicubic interpolation. Second row: DSen2 [13]. Third row: proposed method. Region extracted from Bucharest area. (a) B5. (b) B6. (c) B7. (d) RGB. (e) B8a. (f) B11. (g) B12.

TABLE V
PERFORMANCE COMPARISON USING WALD'S PROTOCOL FOR 6× SUPER-RESOLUTION ON SYNTHETIC DATA

Data	Method	B1			B9			Average		
		RMSE	SRE	SSIM	RMSE	SRE	SSIM	mRMSE	mSRE	mSSIM
Italy	Bicubic	15.09	49.02	0.9101	35.75	39.84	0.9061	25.42	44.43	0.9081
	DSen2	14.73	45.19	0.8592	27.61	37.20	0.8391	21.17	41.19	0.8491
	Proposed	**8.11**	**52.13**	**0.9492**	**19.28**	**44.12**	**0.9449**	**13.69**	**48.12**	**0.9470**
Romania	Bicubic	10.1	44.47	0.9473	12.59	32.99	0.9470	22.69	38.73	0.9471
	DSen2	6.98	47.53	**0.9725**	9.36	35.55	0.9712	8.17	41.53	0.9718
	Proposed	**6.08**	**48.42**	0.9724	**5.22**	**40.55**	**0.9904**	**5.65**	**44.48**	**0.9814**
Canada	Bicubic	102.60	33.69	0.9444	103.18	24.11	0.9466	102.9	28.9	0.9455
	DSen2	66.31	37.18	0.9726	69.01	27.46	0.9741	67.66	32.32	0.9733
	Proposed	**50.10**	**38.73**	**0.9768**	**48.87**	**30.42**	**0.9851**	**49.48**	**34.57**	**0.9809**
Ethiopia	Bicubic	12.65	44.10	0.9442	9.77	35.80	0.9476	11.21	39.95	0.9459
	DSen2	9.82	46.47	**0.9708**	7.19	38.31	0.9718	8.50	42.39	0.9713
	Proposed	**8.71**	**47.02**	0.9671	**6.20**	**40.65**	**0.9857**	**7.45**	**43.83**	**0.9764**

image. Since $\mathcal{L}_{sym}$ was also used during training for these three bands, we hypothesized that the increased detail similarity with another 10-m band did not allow for a fairly good recovery of the observed image, for this particular test area. However, for the rest of the test images, our architecture performed better on these three bands, indicating that a majority of geographical areas could benefit from a direct transfer of details from 10-m bands while still being consistent with the observed LR bands. One thing to notice in Table III is the results for Canada tile, which are relatively high in magnitude compared to the other three areas, for both methods. This may be explained by a big difference in reflectance characteristics for snow-covered areas compared to other geographic zones, leading to poor results for deep learning methods that are usually exposed during the training process to areas with different radiometric properties. The results for reduced-resolution evaluation are presented in Table IV. Our method delivers good mean results over all 20-m bands in terms of RMSE and SRE, while for SSIM, it outperforms DSen2 on Romania and Ethiopia tile. On bands B11 and B12 for Italy and Canada tiles, applying DSen2

TABLE VI
PERFORMANCE COMPARISON USING WALD'S PROTOCOL FOR 6× SUPER-RESOLUTION ON SYNTHETIC DATA

Data	Method	B1			B9			Average		
		RMSE	SRE	SSIM	RMSE	SRE	SSIM	mRMSE	mSRE	mSSIM
Italy	Bicubic	69.01	36.24	0.5517	139.79	28.04	0.4853	104.40	32.14	0.5185
	DSen2	35.16	37.32	0.7994	59.92	31.91	0.7820	47.54	34.61	0.7907
	Proposed	**34.56**	**39.31**	**0.8339**	**42.11**	**36.34**	**0.8281**	**38.33**	**37.82**	**0.8310**
Romania	Bicubic	42.37	31.19	0.6237	52.51	20.33	0.5750	47.44	25.76	0.5993
	DSen2	**17.31**	**38.81**	**0.9447**	**27.78**	**25.86**	**0.9006**	**22.57**	**32.33**	**0.9226**
	Proposed	36.07	32.33	0.8478	45.78	21.51	0.8751	40.92	26.92	0.8614
Canada	Bicubic	515.09	18.42	0.6600	521.07	11.84	0.6408	518.08	15.13	0.6504
	DSen2	167.63	**28.18**	0.**9630**	211.03	19.58	0.9473	189.33	23.88	**0.9551**
	Proposed	**143.91**	27.89	0.9272	**124.02**	**24.18**	**0.9820**	**133.96**	**26.03**	0.9546
Ethiopia	Bicubic	81.84	28.03	0.6308	53.82	21.15	0.6516	67.83	24.59	0.6412
	DSen2	**30.45**	**36.26**	**0.9470**	**25.79**	**27.42**	0.9256	**28.12**	**31.84**	**0.9363**
	Proposed	52.40	30.13	0.8194	28.93	26.65	**0.9269**	40.66	28.39	0.8731

results in lower RMSE, which also leads to a higher average SSIM value. Similar to the previous evaluation, the results for Canada tile show a big difference in terms of magnitude, for both methods, indicating possible shortcomings in applying data-driven methods for snow-covered areas. DSen2 performs better on bands B11 and B12 on Italy and Canada tiles, and also on band B8a for Ethiopia and B5 for Romania, with minor differences for the last two. The remaining results are, however, considerably better for our method, which leads to an overall improved mean performance. Two visual examples are presented, in Fig. 4 for an area from Italy tile and in Fig. 5 for an area from Romania tile. The example in Fig. 4 illustrates a better performance for our method on bands B5, B6, B7, and B8a while achieving visually similar results for bands B11 and B12. This similarity, also verified through the results from Table IV (synthesis evaluation), may be explained given how the error for these two bands is computed during training: since only $\mathcal{L}_{\text{consistency}}$ and $\mathcal{L}_{\text{synthesis}}$ are employed for bands B11 and B12, and given that DSen2 is optimized through a cost function similar to $\mathcal{L}_{\text{synthesis}}$, it is natural to expect a similar performance on synthesis evaluation, along with a better performance on consistency evaluation for our method (as shown in Table III). Note here that our method performs especially well on areas containing land-to-water transitions, resulting in less artifacts than DSen2 for such regions. In Fig. 5, the example on a Bucharest region with an increased number of details also suggests that our method outperforms DSen2 on bands B5, B6, B7, and B8a.

For 6× super-resolution assessment, we conducted similar evaluation steps as in the 2× case, additionally including a visual comparison to accompany the numerical results. Table V contains the consistency evaluation results for super-resolving bands B1 and B9. For band B9, our method outperforms DSen2 on all test images while also achieving better mean results. On band B1, DSen2 achieves better SSIM scores on Romania and Ethiopia tiles, with very small differences compared to our method. Both methods also indicate deficiencies for the snow-covered image from Canada, in terms of magnitude for RMSE and SRE when compared to the other test tiles, achieving, however, a good SSIM score. One important element to consider in Tables III and V is the relatively poor performance of bicubic interpolation. Given that the process of bicubic upsampling does not introduce high-frequency details in the generated images, not aligning with the characteristics of an HR image, the degradation process applied during consistency evaluation is performed on an image with a lower resolution than expected. This results in degraded images having less details than the observed ones and, hence, low-performance measures between the two. The reduced-resolution evaluation results are presented in Table VI, along with one example from each test image shown in Fig. 6. Our method is shown to perform better on Italy and Canada tiles, while DSen2 performed overall better on the other two test images, given table results. Even though the numerical results indicate a poor performance on some test images, compared to DSen2, we determined that this is due to the increased number of details for the super-resolved images our method produces, not always aligning with the level of detail from the reference image. Visual results for full-resolution upsampling are presented in Fig. 7, with the intent of comparing the level of detail induced by each method with the level of detail from an observed 10-m band. Here, we include an additional visual comparison with another method, namely, SSSS [23], which was shown to produce among the best visual results

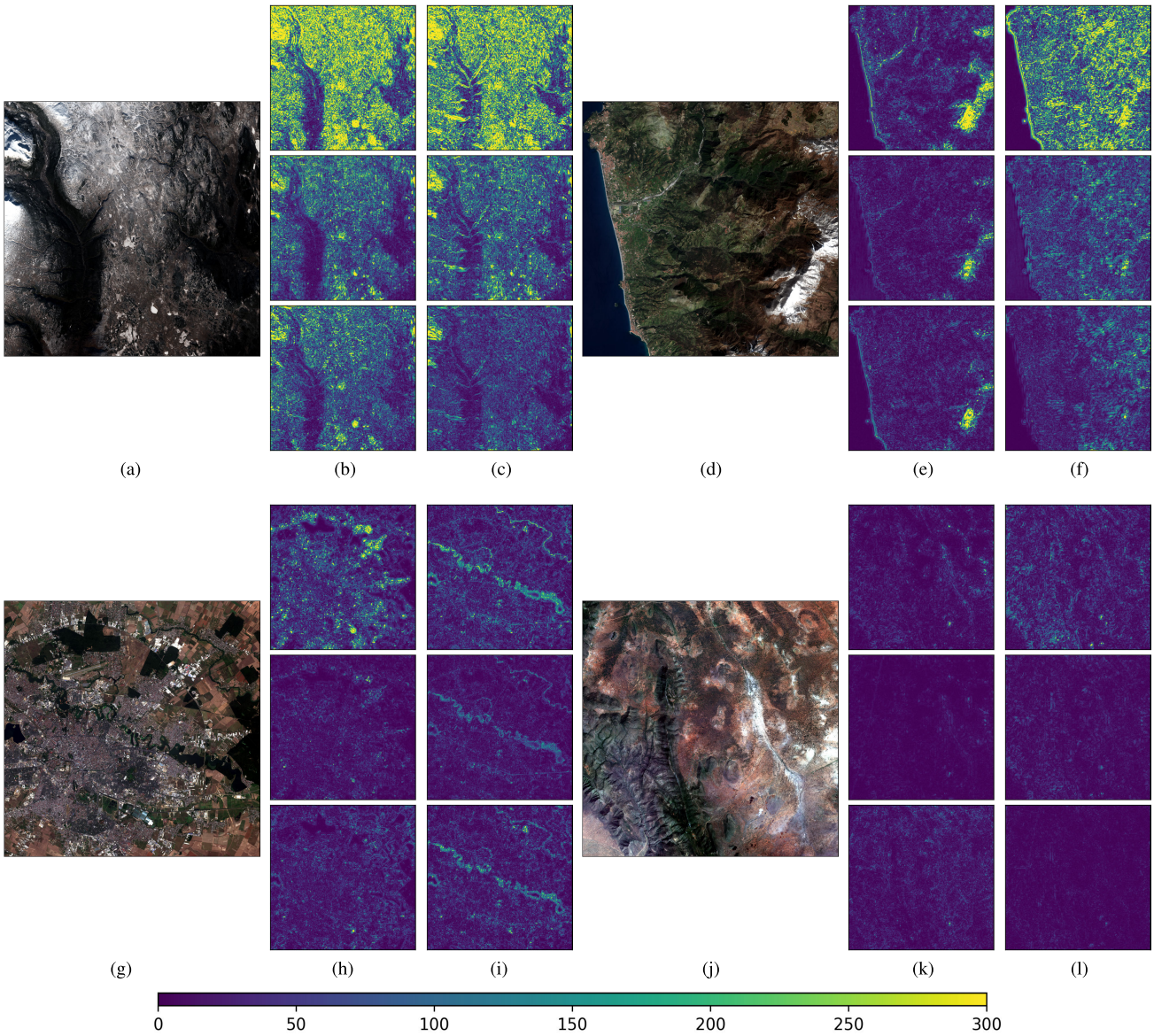


Fig. 6. Absolute differences between super-resolved 60-m bands and ground truth, using Wald's protocol (degraded S2 data). First row: bicubic interpolation. Second row: DSen2 [13]. Third row: proposed method. Region (a) extracted from Canada, region (d) from Italy, region (g) from Romania, and region (j) from Ethiopia. (a) RGB. (b) B1. (c) B9. (d) RGB. (e) B1. (f) B9. (g) RGB. (h) B1. (i) B9. (j) RGB. (k) B1. (l) B9.

on 6× super-resolution in an extensive comparative study conducted in [24]. Visually, our method produces super-resolved bands with an increased number of details, leveling up with the high-frequency components of existing 10-m bands. On band B1, SSSS results contain fairly different radiometric values, compared to the original band, while DSen2 and our method show little difference from the original reflectance distribution. On band B9, both SSSS and DSen2 are shown to obtain suboptimal results regarding the amount of detail, compared to our method. Note that along with an increased number of details, our method also performs considerably better on consistency evaluation, leading us to conclude that the method could be used as a reliable super-resolution mechanism for upsampling 60-m bands. In the light of the ill-posedness characteristic of the problem at hand, which was discussed in Section III, we acknowledge that our method leads to super-resolved images with good consistency properties and realistic high-frequency details, not always aligning with the reference solution considered in synthesis evaluation.

Choosing the right weighting parameters for the loss functions represents an important step, exerting a high influence over the numerical and visual results, as we have seen from Tables I and II and Fig. 3. However, training multiple models with slight differences in their loss configuration, in order to determine the right direction for modifying the weighting terms, is a highly time-consuming process. One solution would be to condition the model's output on these three weighting factors, eliminating the need of an intensive search process for the right combination. This approach has been recently discussed and presented in [46] as you only train once (YOTO), introduced as a new mechanism for loss-conditioned training of neural networks. Applying such method in applications with multiple training objectives, in particular to our solution for S2 super-resolution, could lead to a reliable searching algorithm

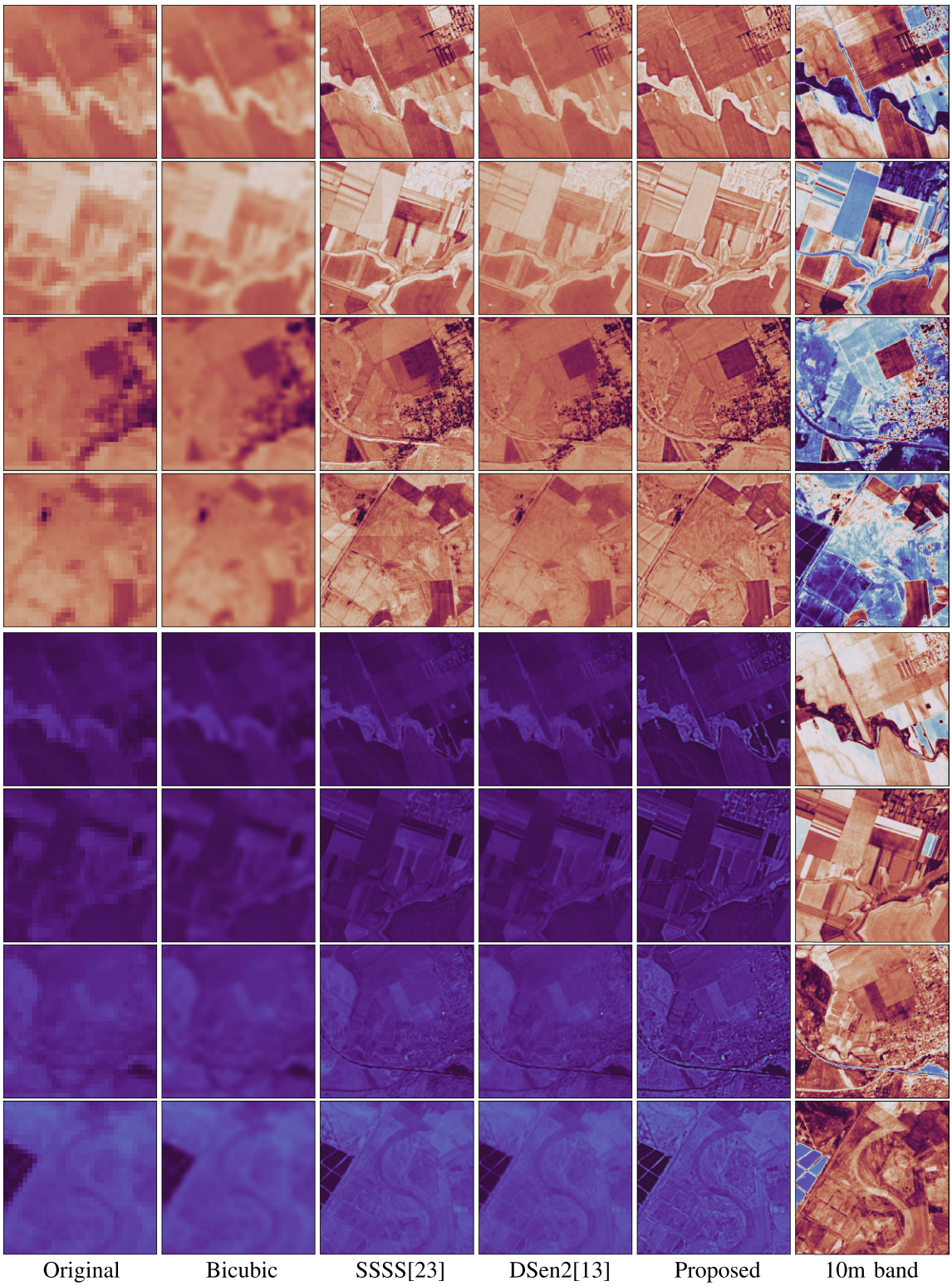


Fig. 7. Results on real S2 data for 6× super-resolution. The first four rows correspond to super-resolving band B1, while the last four are for band B9. (From left to right) Original patch, bicubic interpolation, SSSS [23], DSen2 [13], proposed method, band B2 (10 m) for the first four rows, and band B8 (10 m) for the last four rows.

for the right hyperparameter combination, ensuring a better final performance.

VI. Conclusion

In this article, we presented a mechanism for super-resolving the 20- and 60-m bands provided by S2 up to 10-m spatial resolution, based on fully CNNs. The architectures were trained using a multiobjective loss function, aimed at achieving a good tradeoff between three distinct features: good consistency properties, good synthesis properties, and visually realistic high-frequency components. The first objective was tackled by adopting previously

established MTF-based methods. The second one relied on an additional optimization step using degraded data, a common training strategy among many super-resolution methods. The third objective implied adding a direct similarity measure between details from generated super-resolved bands and real details extracted from already available 10-m bands. Our trained architectures delivered good results on both reduced-resolution evaluation (synthesis) and full-resolution evaluation (consistency) for 20-m bands, given a wide variety of environments, proving good generalization capabilities. For super-resolving the 60-m bands, our method was able to achieve a better consistency, along with enhanced realistic high-frequency components. While there exist a variety of evaluation protocols more or less adopted for super-resolution methods, we feel that the lack of an exact solution for any super-resolved real image should motivate the comparison between the results obtained using multiple evaluation schemes. Along with mechanisms designed for automatically finding a good tradeoff for multiobjective training, our future work will include building evaluation mechanisms for assessing the level of trust in super-resolution methods.

REFERENCES

[1] H. Ghassemian, "A review of remote sensing image fusion methods," *Inf. Fusion*, vol. 32, pp. 75–89, Nov. 2016.

[2] D. Jiang, D. Zhuang, Y. Huang, and J. Fu, "Survey of multispectral image fusion techniques in remote sensing applications," in *Image Fusion and Its Applications*. London, U.K.: IntechOpen, Jun. 2011, pp. 1–23.

[3] A. Azarang, H. E. Manoochehri, and N. Kehtarnavaz, "Convolutional autoencoder-based multispectral image fusion," *IEEE Access*, vol. 7, pp. 35673–35683, 2019.

[4] Y. T. Solano-Correa, F. Bovolo, L. Bruzzone, and D. Fernandez-Prieto, "A method for the analysis of small crop fields in Sentinel-2 dense time series," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 2150–2164, Mar. 2019.

[5] L. Tulczyjew, M. Kawulok, N. Longépé, B. Le Saux, and J. Nalepa, "Graph neural networks extract high-resolution cultivated land maps from Sentinel-2 image series," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[6] M. Finer et al., "Combating deforestation: From satellite to intervention," *Science*, vol. 360, no. 6395, pp. 1303–1305, Jun. 2018.

[7] H. Heiselberg, "A direct and fast methodology for ship recognition in Sentinel-2 multispectral imagery," *Remote Sens.*, vol. 8, no. 12, p. 1033, Dec. 2016.

[8] A. Lagrange et al., "Benchmarking classification of Earth-observation data: From learning explicit features to convolutional networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2015, pp. 4173–4176.

[9] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning Earth observation classification using ImageNet pretrained networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 105–109, Jan. 2016.

[10] M. Gargiulo, A. Mazza, R. Gaetano, G. Ruello, and G. Scarpa, "Fast super-resolution of 20 m Sentinel-2 bands using convolutional neural networks," *Remote Sens.*, vol. 11, no. 22, p. 2635, Nov. 2019.

[11] M. Gargiulo, D. A. G. Dell'Aglio, A. Iodice, D. Riccio, and G. Ruello, "A CNN-based super-resolution technique for active fire detection on Sentinel-2 data," in *Proc. Photon. Electromagn. Res. Symp. Spring (PIERS-Spring)*, Jun. 2019, pp. 418–426.

[12] M. Kawulok, T. Tarasiewicz, J. Nalepa, D. Tyrna, and D. Kostrzewa, "Deep learning for multiple-image super-resolution of Sentinel-2 data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2021, pp. 3885–3888.

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

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[15] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "Single sensor image fusion using a deep convolutional generative adversarial network," in *Proc. 9th Workshop Hyperspectral Image Signal Process., Evol. Remote Sens. (WHISPERS)*, Sep. 2018, pp. 1–5.

[16] K. Zhang, G. Sumbul, and B. Demir, "An approach to super-resolution of Sentinel-2 images based on generative adversarial networks," in *Proc. Medit. Middle-East Geosci. Remote Sens. Symp. (M2GARSS)*, Mar. 2020, pp. 69–72.

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

[18] Q. Wang, W. Shi, and P. M. Atkinson, "Area-to-point regression Kriging for pan-sharpening," *ISPRS J. Photogramm. Remote Sens.*, vol. 114, pp. 151–165, Apr. 2016.

[19] Q. M. Wang, W. Z. Shi, Z. B. Li, and P. M. Atkinson, "Fusion of Sentinel-2 images," *Remote Sens. Environ.*, vol. 187, pp. 241–252, Dec. 2016.

[20] C. Lanaras, J. Bioucas-Dias, E. Baltsavias, and K. Schindler, "Super-resolution of multispectral multiresolution images from a single sensor," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 20–28.

[21] M. O. Ulfarsson, F. Palsson, M. D. Mura, and J. R. Sveinsson, "Sentinel-2 sharpening using a reduced-rank method," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6408–6420, Sep. 2019.

[22] N. Brodu, "Super-resolving multiresolution images with band-independent geometry of multispectral pixels," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4610–4617, Aug. 2017.

[23] C.-H. Lin and J. M. Bioucas-Dias, "An explicit and scene-adapted definition of convex self-similarity prior with application to unsupervised Sentinel-2 super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3352–3365, May 2020.

[24] S. E. Armannsson, M. O. Ulfarsson, J. Sigurdsson, H. V. Nguyen, and J. R. Sveinsson, "A comparison of optimized Sentinel-2 super-resolution methods using Wald's protocol and Bayesian optimization," *Remote Sens.*, vol. 13, no. 11, p. 2192, Jun. 2021.

[25] H. V. Nguyen, M. O. Ulfarsson, J. R. Sveinsson, and M. D. Mura, "Sentinel-2 sharpening using a single unsupervised convolutional neural network with MTF-based degradation model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6882–6896, 2021.

[26] V. Lempitsky, A. Vedaldi, and D. Ulyanov, "Deep image prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9446–9454.

[27] X. Ma, Y. Hong, and Y. Song, "Super resolution land cover mapping of hyperspectral images using the deep image prior-based approach," *Int. J. Remote Sens.*, vol. 41, no. 7, pp. 2818–2834, Apr. 2020.

[28] S. Han, T. B. Lee, and Y. S. Heo, "Deep image prior for super resolution of noisy image," *Electronics*, vol. 10, no. 16, p. 2014, Aug. 2021.

[29] H. He and W.-C. Siu, "Single image super-resolution using Gaussian process regression," in *Proc. CVPR*, Jun. 2011, pp. 449–456.

[30] K. Blix, M. M. Espeseth, and T. Eltoft, "Machine learning for Arctic sea ice physical properties estimation using dual-polarimetric SAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 4618–4634, Jun. 2020.

[31] K. Blix, M. M. Espeseth, and T. Eltoft, "Up-scaling from quad-polarimetric to dual-polarimetric SAR data using machine learning Gaussian process regression," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 7332–7335.

[32] B. Wang and T. Chen, "Gaussian process regression with multiple response variables," *Chemometric Intell. Lab. Syst.*, vol. 142, pp. 159–165, Mar. 2015.

[33] J. Wang et al., "Multisensor remote sensing imagery super-resolution with conditional GAN," *J. Remote Sens.*, vol. 2021, pp. 1–11, Jan. 2021.

[34] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surveys*, vol. 54, no. 10s, pp. 1–41, Jan. 2022.

[35] T. An, X. Zhang, C. Huo, B. Xue, L. Wang, and C. Pan, "TR-MISR: Multiimage super-resolution based on feature fusion with transformers," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1373–1388, 2022.

[36] C. Ye, L. Yan, Y. Zhang, J. Zhan, J. Yang, and J. Wang, "A super-resolution method of remote sensing image using transformers," in *Proc. 11th IEEE Int. Conf. Intell. Data Acquisition Adv. Comput. Syst., Technol. Appl. (IDAACS)*, Sep. 2021, pp. 905–910.

[37] F. Palsson, J. R. Sveinsson, M. O. Ulfarsson, and J. A. Benediktsson, "Quantitative quality evaluation of pansharpened imagery: Consistency versus synthesis," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1247–1259, Mar. 2016.

[38] A. Shocher, N. Cohen, and M. Irani, "'Zero-shot' super-resolution using deep internal learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3118–3126.

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

[40] 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. 2013.

[41] X. Kang, S. Li, and J. A. Benediktsson, "Pansharpening with matting model," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 8, pp. 5088–5099, Aug. 2014.

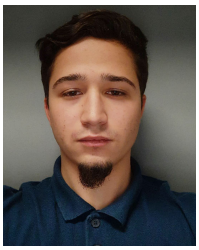
[42] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Aug. 2002.

[43] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[45] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1–7.

[46] A. Dosovitskiy and J. Djolonga, "You only train once: Loss-conditional training of deep networks," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–17.



Vlad Vasilescu received the B.S. degree in computer science and information technology from the University Politehnica of Bucharest (UPB), Bucharest, Romania, in 2021, where he is currently pursuing the M.S. degree in electronic engineering, telecommunications and information technologies.

He is currently a member of the Speech and Dialogue Research Laboratory (SpeeD), CAMPUS Research Center, and the Research Center for Spatial Information (CEOSpaceTech), UPB. His research focuses on adversarial robustness, super-resolution algorithms, Bayesian optimization, and deep reinforcement learning.

Mihai Datcu (Fellow, IEEE) received the M.S. and Ph.D. degrees in electronics and telecommunications from the University Politehnica of Bucharest (UPB), Bucharest, Romania, in 1978 and 1986, respectively, and the Habilitation a Diriger des Recherches degree in computer science from the University Louis Pasteur, Strasbourg, France, in 1999.

Since 1981, he has been with the Department of Applied Electronics and Information Engineering, Faculty of Electronics, Telecommunications and Information Technology, UPB, where he is currently a Full Professor and the Director of the Research Center for Spatial Information (CEOSapceTech). Since 1993, he has been with the German Aerospace Center (DLR), Weßling, Germany, where he is also a Senior Scientist with the Remote Sensing Technology Institute (IMF). From 1992 to 2002, he had a longer Invited Professor Assignment with the Swiss Federal Institute of Technology (ETH Zürich), Zürich, Switzerland. Since 2001, he had been initiating and leading the Competence Center on Information Extraction and Image Understanding for Earth Observation, ParisTech, Paris Institute of Technology, Paris, France, a collaboration of DLR with the French Space Agency (CNES). From 2005 to 2013, he was a Professor at the DLR-CNES Chair, ParisTech, Paris Institute of Technology. From 2011 to 2018, he was leading the Immersive Visual Information Mining Research Laboratory, Munich Aerospace Faculty, Munich, Germany. From 2018 to 2020, he was the holder of the Blaise Pascal International Chair of Excellence at Conservatoire national des arts et métiers (CNAM), Paris. From 2020 to 2022, he was involved in the DLR-French Aerospace Lab (ONERA) Joint Virtual Center for AI in Aerospace. He was a Visiting Professor with the University of Oviedo, Oviedo, Spain; University Louis Pasteur and International Space University, Strasbourg; the University of Siegen, Siegen, Germany; the University of Innsbruck, Innsbruck, Austria; the University of Alcala, Alcalá de Henares, Spain; University Tor Vergata, Rome, Italy; the Universidad Pontificia de Salamanca, Madrid, Spain; the University of Camerino, Camerino, Italy; the University of Trento, Trento, Italy, the China Academy of Sciences, Shenyang, China; the Universidade Estadual de Campinas (UNICAMP), Campinas, Brazil; the University of Wuhan, Wuhan, China, and the Swiss Center for Scientific Computing, Manno, Switzerland. He has initiated and implemented the European frame of projects for Earth Observation image information mining (IIM) and is involved in research programs for information extraction, data mining, big EO data knowledge discovery, and data understanding with the European Space Agency (ESA), NASA, and in a series of national and European projects. He is also a Visiting Professor with ESA's Φ-Lab. He and his team have developed the operational IIM processor in the Payload Ground Segment systems for the German mission TerraSAR-X, and data mining tools and systems for the Copernicus missions Sentinel-1 and Sentinel-2. He is developing algorithms for model-based information retrieval from high-complexity signals and methods for scene understanding from very-high-resolution synthetic aperture radar (SAR) and interferometric SAR data. His research interests include information theory, signal processing, explainable and physics-aware artificial intelligence, computational imaging, and quantum machine learning with applications in EO.

Dr. Datcu is a member of the ESA Working Group Big Data from Space. He was a recipient of the Romanian Academy Prize Traian Vuia for the development of the SAADI image analysis system and his activity in image processing in 1987, the Best Paper Award and the IEEE Geoscience and Remote Sensing Society Prize in 2006, the National Order of Merit with the rank of Knight, for outstanding international research results, awarded by the President of Romania in 2008, and the Chaire d'excellence internationale Blaise Pascal 2017 for international recognition in the field of data science in EO and the 2018 Ad Astra Award for Excellence in Science. He has served as a co-organizer for international conferences and workshops and as a guest editor and an associate editor for IEEE and other journals. In 2022, he received the IEEE GRSS David Landgrebe Award in recognition of outstanding contributions to Earth Observation analysis using innovative concepts for big data analysis, image mining, machine learning, smart sensors, and quantum resources.

Daniela Faur (Member, IEEE) received the M.S. and Ph.D. degrees in electrical engineering from the Politehnica University of Bucharest (UPB), Bucharest, Romania, in 2002 and 2006, respectively.

Since 2007, she has been holding a professorship in information theory with the Department of Applied Electronics and Information Engineering, UPB. She is involved in national and European research grants in the field of EO data processing and visualization targeting applications in agriculture, disaster and humanitarian crisis management, and biodiversity monitoring. In 2010, she co-founded CEOSpaceTech, Bucharest, Romania, the research center for spatial information that promotes technological and scientific areas of research in Earth observation and related fields. She acted as a Project Manager of the GEODIM-PNCDII-Platform for Geo-Information in Support of Disaster Management, LEOSITS-STAR-ROSA-Long Term Data Exploitation for Satellite Image Time Series-Extraction of Classes for Scene Dynamic, and VATEO-PNCDI III-Visual Analytics Tool for Earth Observation Images and as a Principal Investigator for eVADE-ESA ITT-Interactive Visual Analysis Tool for Earth Observation Data. As a result of winning Copernicus Incubator in 2019, she co-founded OGOR, an agri-tech start-up that builds a live journal based on Copernicus satellite images. She serves as a Europe Principal Investigator in the frame of ESA NASSC Dragon 5 Cooperation, coordinating the project on large-scale spatial–temporal analysis for dense satellite image series with deep learning for the term 2020–2024. From 2019 to 2021, she represented UPB in H2020 SPACE-END: ENDEAVOUR Space HUB. Currently, she leads, as UPB representative, the H2020 CENTURION: COPERNICUS datacube/AI data cube services for society, industry, and new market generation H2020-SPACE-2018-2021 for the term 2020–2024.

Dr. Faur is a member of the IEEE Geoscience and Remote Sensing Society.