Article

A Comparison of Optimized Sentinel-2 Super-Resolution Methods Using Wald’s Protocol and Bayesian Optimization

Sveinn E. Armannsson, Magnus O. Ulfarsson *, Jakob Sigurdsson, Han V. Nguyen and Johannes R. Sveinsson

Faculty of Electrical and Computer Engineering, University of Iceland, Hjardargata 2-6, 107 Reykjavik, Iceland; sea2@hi.is (S.E.A.); jakobs@hi.is (J.S.); hvn2@hi.is (H.V.N.); sveinsso@hi.is (J.R.S.)

* Correspondence: mou@hi.is

Abstract: In the context of earth observation and remote sensing, super-resolution aims to enhance the resolution of a captured image by upscaling and enhancing its details. In recent years, numerous methods for super-resolution of Sentinel-2 (S2) multispectral images have been suggested. Most of these methods depend on various tuning parameters that affect how effective they are. This paper’s aim is twofold. Firstly, we propose to use Bayesian optimization at a reduced scale to select tuning parameters. Secondly, we choose tuning parameters for eight S2 super-resolution methods and compare them using real and synthetic data. While all the methods give good quantitative results, Area-To-Point Regression Kriging (ATPRK), Sentinel-2 Sharpening (S2Sharp), and Sentinel-2 Symmetric Skip Connection convolutional neural network (S2 SSC) perform markedly better on several datasets than the other methods tested in this paper.

Keywords: data fusion; image sharpening; multispectral (MS) multiresolution images; Sentinel-2; sharpening of bands; super-resolution

1. Introduction

The European Space Agency’s Sentinel-2 (S2) satellite constellation comprises two satellites that carry arrays of multispectral sensors that cover 13 bands in the visible, near-infrared, and shortwave-infrared part of the spectrum. Spectral characteristics of Sentinel-2 A and B sensors can be assumed to be roughly identical for practical purposes and cover the same wavelengths and resolutions [1]. Four bands have 10 m spatial resolution, bands B2, B3, and B4 in the visible spectrum and B8 in the near-infrared spectrum. Bands B1, B9, and B10 have 60 m resolution. The remaining bands, B5, B6, B7, B8A, B11, and B12, have 20 m resolution. As Figure 1 illustrates, bands B9, B10, and B12 cover low transmittance wavelengths which makes them harder to process. B10, the cirrus band, is a water absorption band and is usually ignored outside of cloud cover detection applications, as it does not contain surface information.

Some multispectral satellites, for instance, Landsat-7, capture a high-resolution panchromatic band alongside lower resolution bands of varying spectra which can be used for pansharpening, i.e., transferring detail from the high resolution panchromatic band to the lower resolution bands. S2 does not have high-resolution panchromatic sensors, so pansharpening methods can not be directly applied to S2 data. It is necessary to use S2’s other high-resolution bands when generating super-resolution versions of low-resolution bands, which complicates the process somewhat. Popular pansharpening methods have still been applied to S2 data by selecting or generating high resolution bands, for example, in [2].
Figure 1. Sentinel-2 spectral response characteristics from [1] alongside earth surface solar irradiance from [3]. Spatial resolution and central spectral wavelength of each band appear in parentheses in the legend.

Methods have also been developed specifically for super-resolution of S2 images, and considerable prior work is available on the subject. Model based methods explicitly define a model of image degradation to be reversed based on prior knowledge of the image acquisition system, while machine learning based methods learn it directly from the data.

Numerous methods are based on a variational framework which is based on projections onto a subspace justified by a linear sensor model: Super-resolution for multispectral multiresolution estimation (SupReME) [4], Sentinel-2 Sharpening (S2Sharp) [5], Multiresolution Sharpening Approach (MuSA) [6], and Sentinel-2 Super-resolution via Scene-adapted Self-similarity (SSSS) [7].

Area-To-Point Regression Kriging (ATPRK) [8] and Sen2res [9] use somewhat different approaches. ATPRK combines regression and area-to-point kriging to predict higher resolution representations of the low-resolution bands. The point spread function (PSF) of the sensor, size of pixels, and spatial correlation are used for predictions based on the high-resolution bands [8]. The Sen2res method explicitly encodes geometric details from available high-resolution bands as pixel properties independent from their reflectance. It attempts to preserve the spectral content of each low-resolution band independently from the geometry and combine the two to solve the super-resolution problem [9].

A number of deep learning based methods have been proposed for S2 super-resolution. These include Deep Sentinel-2 (DSen2) [10] which was trained at reduced resolution on a random selection of 60 large scenes (∼110 × 110 km²). A Deep Residual Network that fuses the fine and coarse spatial resolution bands is suggested in [11]. Sentinel-2 Symmetric Skip Connection convolutional neural network (S2 SSC) [12] extends on this but uses an unsupervised single image learning (zero-shot) approach in which the training and testing are performed on the same image. The only training input is the image for which a super-resolved version is desired. The Remote Sensing Very Deep Super-Resolution [13] (RS-VDSR) network re-trains previously developed deep convolutional networks from [14,15] to work with S2 data with promising results. Sentinel-2 Parallel Residual Network [16] (SPRNet) trains two sets of parallel residual networks with lower scale data synthesized from the S2 data, obtaining good spatial fidelity and spectral preservation. The paper [17] was inspired by DSen2 and develops two large neural networks for deployment and training on high performance computing clusters, one for 20 m bands and one for 60 m bands. Furthermore, inspired by DSen2, ref. [18] suggests a single convolutional neural network for super-resolving the 20 and 60 m Sentinel-2 bands to 10 m simultaneously.
Variants of Generative Adversarial Networks [19] (GAN) have also been suggested for S2 sharpening. A deep convolutional design for a GAN that gives good results for the super-resolution of S2 images was developed in [20]. Sentinel-2 Generative Adversarial Network [21] (S2GAN) similarly uses the 10 m S2 bands as guidance for learning the super-resolution task on lower resolution bands. A GAN method inspired by [22] that focuses solely on the S2 10 m RGB bands was developed in [23]. It is trained using PeruSat-1 images with 2.8 m resolution as a reference to resolve S2 images.

When comparing different methods, finding the best parameters for each method can be a daunting task, particularly if a method has many parameters or if the user is not intimately familiar with all the nuances of the method. We present a framework that allows for fair comparisons between different methods by attempting to ensure optimal performance is achieved by each method. Having a framework that can be applied across different methods of super-resolution, or even different capture platforms makes an objective comparison of methods easier.

In this paper, we propose a tuning parameter selection method using Wald’s protocol [24] by first reducing the scale of the image being processed 2× and 6×. Bayesian optimization [25] with a Tree-structured Parzen Estimator [26,27] (TPE) is used to find the parameters that give the best results when resolving the 20 and 60 m bands back to their original resolution and applied to super-resolving the original full scale image. A preliminary version of this parameter selection approach was described in [28]. Our code is available on the first author’s github page (https://github.com/SveinnEirikur/s2-sr-parameter-optimization, accessed on 3 June 2021).

The proposed parameter selection approach is used to select tuning parameters for six S2 sharpening methods, i.e., SupReME, S2Sharp, MuSA, SSSS, ATPRK and S2 SSC. These methods along with DSen2, and Sen2res are critically compared using standard comparison criteria. The code used to generate these comparisons will be made available for download. Adding new methods through simple wrapper functions is easy and allows for both parameter optimization and direct comparisons between methods. While the focus in this paper is on Sentinel-2 images these same principles could also be applied to other types of multispectral multiresolution satellite data, e.g., MODIS, ASTER, Landsat-7, and Landsat-8 [29–31].

The paper is organized as follows. Section 2 describes the tuning parameter selection approach and the comparison methods. Section 3 compares the qualities of the methods and describes the quantitative metrics used to compare them. Quantitative comparisons use synthesized and reduced scale S2 data. For qualitative comparisons, real S2 images are used. This is followed by the final conclusions in Section 4.

**Notation**

In this paper, \( Y_{10m}, Y_{20m}, \) and \( Y_{60m} \) will be used to refer to the set of images captured by the S2 constellation at each resolution, respectively, \( Y_{10m} \in \mathbb{R}^{n_x \times n_y \times 4} \) is the scene as captured by bands B2, B3, B4, and B8 at 10 m resolution, \( Y_{20m} \in \mathbb{R}^{n_x/2 \times n_y/2 \times 6} \) is the scene as captured by bands B5, B6, B7, B8A, B10, and B11 at 20 m resolution and \( Y_{60m} \in \mathbb{R}^{n_x/6 \times n_y/6 \times 2} \) is the scene as captured by bands B1 and B9 at 60 m resolution. To indicate that a collection of bands has been resampled, either downsampled, upsampled or super-resolved, the new resolution is added as a superscript, i.e., \( Y_{10m}^{20m} \) would mean \( Y_{10m} \) at 20 m resolution and \( Y_{20m}^{10m} \) would mean \( Y_{20m} \) super-resolved to 10 m. The capital letter D is used in the same way to indicate downsampling in a more general manner.

**2. Methods**

**2.1. Model Based Methods**

**2.1.1. ATPRK**

Area-To-Point Regression Kriging (ATPRK) is a geostatistical approach to super-resolution that was first suggested for use with MODIS data in [32]. It was applied to pansharpening of WorldView-2 and Landsat data in [33] and adapted for S2 images in [8].
ATPRK uses regression-based overall trend estimation and area-to-point-based residual upsampling. Here, trend is the geostatistical term for the spatially varying mean of a spatial process, and residual refers to the variation remaining after removal of the trend. ATPRK accounts explicitly for size of support (pixel), spatial correlation, and Point Spread Function (PSF) of the sensor [8]. While the original code only performed 20 → 10 m super-resolution, extending it to perform 60 → 10 m super-resolving in exactly the same way is fairly simple.

ATPRK consists of two parts: regression and area-to-point kriging (ATPK). The final ATPRK estimate \( \hat{Y}_{20m}^{10m}(x) \), is given by

\[
\hat{Y}_{20m}^{10m}(x) = \hat{Y}_{20m}^{10m}(x) + \hat{Y}_{20m}^{10m/2}(x), \quad l \in \{1, \ldots, 6\},
\]

where \( \hat{Y}_{20m}^{10m}(x) \) is the regression prediction and \( \hat{Y}_{20m}^{10m/2}(x) \) is the ATPK prediction.

2.1.2. Sen2res

Sen2res [9] models the mix of elements within a pixel (e.g., 20% road and 80% vegetation) as a physical property of that pixel, exploiting the geometric consistency of subpixel elements across the MS bands in a similar manner to hyperspectral unmixing [34]. The reflectance of each element is modelled as dependent on the spectral band at which it is observed, relying on local consistency between neighbouring pixels. High resolution bands are used to separate band-independent information from band-dependent reflectance, and the geometric information is used to unmix low-resolution pixels while preserving their overall reflectance [9].

Shared values, \( S \), span over multiple pixels and represent the band dependent spectral reflectance of elements that are common to the spanned pixels. They are initialized as the average of the high-resolution pixels that partially cover them. Weights, \( W \), are internal to each pixel and represent the proportion of elements within the pixel and are initialized as \( \frac{1}{r} \), where \( r \) is the scale difference between the coarse and fine bands. Weights are independent of the spectral band, but \( S \) is not. The mixing equation for shared values \( S \) and weights \( W \) is given as

\[
\hat{Y}_{20m}^{10m}(x, y) = \hat{W}_0(x, y)\hat{S}_1(x, y) + \hat{W}_1(x, y)\hat{S}_1(x + 1, y) + \hat{W}_2(x, y)\hat{S}_1(x, y + 1) + \hat{W}_3(x, y)\hat{S}_1(x + 1, y + 1), \quad l \in \{1, \ldots, 6\}.
\]

The weights are non-negative and sum-to-one. \( \hat{W} \) and \( \hat{S} \) are estimated using an iterative method. The shared values are processed through various normalizations, and geometric means to produce a set of corrected shared values \( \hat{S} \) for each coarse band as detailed in [9] and they are then combined with the band independent weights, \( W \), to produce high-resolution estimates using (2). The general procedure for the \( Y_{60m} \) bands is similar.

2.1.3. SupReME

The SupReME (SUPer-REsolution for multispectral Multiresolution Estimation) method proposed in [4] relies strongly on the observation model of the imaging process that generates the low-resolution images (blurring and downsampling). It exploits the fact that spectral bands are correlated and can be represented in a lower-dimensional subspace, learned from the input data. The textural information from the high-resolution bands is utilized by encoding the discontinuities of the data and propagating the spatial information to the lower resolution bands.

The vectorized sharpened image at band \( i \) has a low-rank representation \( x_i = Zu_{(i)} \), where \( Z \) is \( n_i \times p \) and \( p = 7 \). The sharpened image is estimated by minimizing

\[
J(Z) = \sum_{i=1}^{L} \frac{1}{2} \| y_i - M_iB_iZu_{(i)} \|^2 + \lambda \sum_{j=1}^{p} q_j \phi_w(z_j).
\]
Here $y_i$ is the vectorized observed image at band $i$, $M_i$ is a downsampling operator, and $B_i$ is blurring. The regularization term is a weighted roughness penalty given by

$$\phi_w(z_j) = z_j^T \left( H_b^T W H_b + H_h^T W H_h \right) z_j.$$  \hspace{1cm} (4)

$W = [w_{ij}]$ is a diagonal spatial weighting matrix and $H_b$ and $H_h$ are $n \times n$ finite difference matrices that approximate vertical and horizontal derivatives of the images in the subspace, respectively [5].

An important part of the method is the estimation of $U = [u^T(i)]$. It is estimated using $XQ$, where $Q \in \mathbb{R}^{n \times n}$ is a blurring matrix representing a 2D cyclic convolution. As $X$, the $L \times n$ super-resolved image, is not known, $XQ$ is approximated by upsampling all the bands of $y$ to the same high-resolution using bicubic interpolation and blurring each band such that the blur of all the bands is equivalent to the strongest blur of the PSF. Singular value decomposition of this approximation of $XQ$ yields $U$.

SupReME fixes $q = [1 \ 1.5 \ 4 \ 8 \ 15 \ 15 \ 20]$ since the first few dimensions are expected to be less affected by noise. Per pixel weights, $w$, are for quadratic smoothing as a spatial regulariser. They are calculated from the Prewitt image gradient magnitude [35] and set to a value between 0.5 and 1. The minimization is then solved using the Constrained Split Augmented Lagrangian Shrinkage Algorithm (C-SALSA) [36–38] and the augmented Lagrangian as detailed in [4].

2.1.4. S2Sharp

S2Sharp [5] is closely related to SupReME. It uses a cyclic descent method that employs conjugate gradients and trust-region optimization to find the reduced-rank approximation and subspace $U$ and $Z$. The cost function minimized is

$$J(U, Z) = \sum_{i=1}^{L} \frac{1}{2} \| y_i - M_i B_i Z u(i) \|^2 + \sum_{j=1}^{p} \lambda_j \phi_w(z_j).$$  \hspace{1cm} (5)

This is essentially the same as SupReME apart from each dimension of the $p$-dimensional subspace having an independent regularization coefficient, $\lambda_j = \lambda q_j$, $j \in \{1, \ldots, p\}$, and S2Sharp optimizing $U$ as well. This is a non-convex optimization problem and is solved using an iterative algorithm. $U$ is initialized using the same method as SupReME and refined during cyclic descent. Then cyclic descent is performed using the linear conjugate gradient method and a trust-region method as described in [5].

2.1.5. MuSA

MuSA [6] is also related to SupReME and S2Sharp. SALSA is used to solve the same dimensionally reduced formulation of the problem

$$J(Z) = \sum_{i=1}^{L} \frac{1}{2} \| y_i - M_i B_i Z u(i) \|^2 + \lambda \phi(z).$$  \hspace{1cm} (6)

The subspace $U$ is estimated in an identical manner as in SupReME with the number of dimensions fixed at $p = 5$. Here, regularization is performed with the Color Block Matching 3D Denoiser (C-BM3D) [39], used as a plug-and-play patch-based regularizer [40]. C-BM3D is a variant of the block matching 3D denoiser developed in [41,42] meant for color images. It is used in place of a proximity operator of the kind described in [43,44]. The optimization is based on the SALSA algorithm as detailed in [6].
2.1.6. SSSS

Sentinel-2 Super-resolution via Scene-adapted Self-similarity (SSSS) [7], belongs to the same family of methods as SupReME, S2Sharp, and MuSA. It uses the same dimensionality reduction, but reformulates the regularization term as

\[ J(U, Z) = \sum_{i=1}^{L} \frac{1}{2} \| y_i - M_i B_i Z_{(i)} \|^2 + \lambda \Phi(z), \]

\[ \Phi(z) = \frac{1}{2} \sum_{j=1}^{p} \sum_{(k,l) \in K} a_{k,l} \left\| P_k z_j - P_l z_j \right\|^2, \]

where \( P_k \in \mathbb{R}^{q^2 \times n} \) serves as an operator that shifts the \( k \)th patch of size \( q^2 \) from an \( n \)-pixel image and \( K \) is the self-similarity graph collecting pairs of indexes of similar patches. The self-similarity measure \( a_{k,l} \geq 0 \) is generated using

\[ a_{k,l} := \left\| P_k \hat{Y} - P_l \hat{Y} \right\|^{-1} \]

from a reference image \( \hat{Y} \) generated by averaging the 10 m bands. \( K \) is learned by adding the most similar yet distinct patch \( l \) for each patch \( k \) to the graph. The optimization is a variant of the SALSA algorithm and is detailed in [7].

2.2. Deep-Learning Based Methods

2.2.1. DSen2

The DSen2 network was developed in [10] is a fully convolutional neural network, with ResNet inspired skip connections, and can handle images of arbitrary size. Two forms of the network were trained on 32 × 32 and 96 × 96 pixel patches, respectively, for 2 × and 6 × super-resolution. Inputs are upsampled to the target resolution using simple bi-linear interpolation before they are fed into the network. The first network is trained using the 10 and 20 m bands, while the second network is trained using all bands. They are otherwise the same. The network loss function is the mean absolute error between the network output and the training target. Training is performed using Wald’s protocol. Image bands are downsampled, 10 → 20 m and 20 → 40 m to train the 20 → 10 m network and 10 → 60 m, 20 → 120 m, and 60 → 360 m to train the 60 → 10 m network.

Bilinear interpolation is used, so all bands have the same pixel counts before inputting them as training data. The 20 → 10 m network is trained with \( Y_{20m} \) as the target and \( Y_{20m} \) and \( Y_{40m} \) as the input. The 60 → 10 m network with \( Y_{60m} \) as the target and \( Y_{60m} \), \( Y_{120m} \) and \( Y_{360m} \) as the input. Scale-invariance is assumed to assure the network’s usefulness for the actual 20 → 10 m and 60 → 10 m task.

The network has been trained on 60 varied scenes from across the globe. Actual S2 data tiles were gathered from the Copernicus Services Data Hub, picked randomly, aiming for a roughly even distribution over the globe and for variety in terms of climate zone, land-cover, and biome type. In regard to the generation of training data [10] directly references the applicability of the S2 PSF, but in practice, the code supplied by the authors does not utilize it. Instead training data is generated by simply filtering the bands with a Gaussian filter with \( \sigma = 1/s \) and an \( s \times s \) boxcar filter followed by downsampling by \( s \), where \( s \) is either 2 or 6.

2.2.2. S2 SSC

The S2 SSC [12] convolutional neural network is inspired by the U-net architecture [45] and the networks in [11,46]. It uses a single image learning approach that allows training and testing using the same image. Unlike the other methods studied in this paper, this network only super-resolves the 20 m bands of the S2 data, not the 60 m bands. The network is trained by using small overlapping training patches with a spatial size of 16 × 16 and a
training stride of 8 on the spatial axes. The loss function is given by the mean squared error between
the network output and the training target.

Training is performed using Wald’s protocol, the image bands are downsampled,
10 → 20 m and 20 → 40 m. Two different downsampling methods for training data are
tested. In [12], the network used a bilinear interpolation downsampling method to generate
the $Y_{20m}^{20m}$ and $Y_{20m}^{40m}$ training data, similar to DSen2. The tuned method uses each bands PSF
to generate the reduced scale training images.

As in the previous network, interpolation is used to ensure all bands are at the desired
pixel resolution before inputting them as training data. The network is trained with patches
from $Y_{20m}$ as the target and $Y_{10m}^{20m}$ and $Y_{20m}^{40m}$ as the input. The final result is then produced
by passing the original full resolution versions of the $Y_{10m}$ and $Y_{20m}$ bands as input through
the trained network, resulting in it outputting $Y_{20m}$ at the full 10 m resolution.

2.3. Performance Evaluation

Where ground truth is available, we use the root-mean-squared error (RMSE), relative
dimensionless global error (ERGAS), universal image quality index (UIQI), spectral
angle mapper (SAM) in degrees, signal-to-reconstruction error (SRE) in dB, and structural
similarity index measure (SSIM) metrics as illustrated in [24,47,48]. They are used for
quantitatively comparing the methods described in Sections 2.1 and 2.2. In order to give a
fair comparison, a set number of edge pixels are left out of metric calculations. The number
of pixels to ignore was chosen as six, as suggested for the SSSS method [7].

RMSE measures the size of the error between the estimated image and the ground
truth. Lower values are better. It is calculated for the 20 m and 60 m bands, separately, as
well as for the whole scene when ground truth data is available. RMSE is not useful when
comparing across different datasets, while the ERGAS and UIQI metrics were developed
for that purpose. ERGAS was specifically designed to compare how effectively an image
has been super-resolved. It not only normalizes the values with relation to the dynamic
range but also with regard to the scale difference between the low initial resolution and
high final resolution. ERGAS is therefore calculated separately for the 20 m and 60 m bands.
Lower values are better for ERGAS as well, and an ERGAS below 3 is considered good [24].
SAM maps the angular differences in the spectral signatures of each pixel and returns a
value in degrees. A value of zero implies that the spectra are identical, apart from possibly
a scaling factor applied identically to all bands. The mean SAM value is calculated over all
the pixels in the scene. When ground truth is available, it is calculated for all bands as well
as for only the 20 m bands.

UIQI takes a value between $-1$ and 1, with 1 signifying identical images. It measures
correlation, luminance distortion, and contrast distortion. SSIM compares local patterns of
pixel intensities. It includes normalization factors for the dynamic range of the image. Its
values are also in the range $[-1, 1]$ with 1 signifying very similar images. SRE measures
the ratio between the ground truth pixel values and the estimation error and is reported in
dB. Therefore, somewhat counter-intuitively, for a metric with error in its name, a higher
value is better for SRE. UIQI, SSIM, and SRE are calculated independently for each band,
and averages are reported for 20 m, 60 m, and all bands as applicable.

In short, for RMSE, ERGAS, and SAM values closer to 0 are better, while for UIQI and
SSIM values closer to 1 are better.

All experiments were run on an AMD Ryzen 9 3900X 12-Core Processor with 16 GB
of 2133 MHz DDR4 system memory and GeForce RTX 2070 SUPER graphics processing
unit with 8 GB of VRAM. Running times are the average of 10 runs of the same data on
this computer.

2.4. Tuning Parameter Selection

Since ground truth data is not available for real-world applications of S2 super-
resolution, tuning parameter selection is performed using Wald’s protocol [24]. The process
is outlined in the flow graph in Figure 2. Two reduced resolution images are generated, with
all bands at both $2 \times$ and $6 \times$ original pixel size for each image, respectively. Parameters are selected with Hyperopt and used for super-resolution of the full scale data to get the final image. For small images, padding the data may be necessary so that all methods can process the reduced results during the tuning parameter search. Padding is performed using reflective boundary conditions.

Figure 2. Flow graph showing the parameter tuning process. Scale reduced data is super-resolved back to its original resolution and evaluated against the unscaled data for parameter selection. The parameters are then used to super-resolve the unscaled data to 10 m resolution.

Appropriate filters are applied to their respective bands to simulate the PSF before downsampling by a factor of 2 and 6, respectively, resulting in $2 \times$ and $6 \times$ pixel scales. S2 sensor data from [49] is used to estimate the variance of the PSF of each S2 band based on the measured modulation transfer function (MTF) values for each band. The standard deviation is calculated as $\sigma = s \left( \frac{-2 \ln(MTF)}{\pi} \right)^{1/2}$, where $s$ is the scale difference.

Tuning parameters are selected with Bayesian optimization (BO) using the TPE algorithm of the Hyperopt Python package [25–27]. The parameters that result in the highest average SRE are found using 150 iterations of optimization. During each iteration, the method being tuned is run twice with the parameters suggested by the TPE algorithm, super-resolving the appropriate bands of both reduced images back to the resolution of the original data, $Y_{20m} \rightarrow Y_{20m}$ and $Y_{60m} \rightarrow Y_{60m}$. SRE is calculated between the original data and $Y_{20m}$ and $Y_{60m}$ and the average SRE is used as a loss for the optimization process. The tuning parameters that give the highest SRE, selected in this manner, at a reduced scale, can then be applied to super-resolve the original data to 10 m resolution. This parameter tuning method has previously been shown to improve the results obtained using S2Sharp [28].

BO attempts to build a probabilistic surrogate model from previous evaluations that allows it to predict which parameters might be ideal. It works best for search spaces with moderate dimensions [30]. For a specific function and a given set of parameters, BO treats the objective function as a black box and attempts to find the parameters that minimize (or maximize) the output. The aim is to find a good set of parameters as quickly as possible [25].

Tunable parameters and the objective or loss function can be defined to suit dissimilar tasks. For super-resolution using a quantitative image quality metric is appropriate. Tuning the parameters of a method for individual images can, in some cases, improve results dramatically. Different tuning parameters may also be needed for different image subjects. Remote sensing images can differ greatly depending on the subject, e.g., whether the area being captured is urban or rural, mountainous or oceanic, tropical or arctic, etc.

The TPE, or Tree-structured Parzen Estimator algorithm, is a hyperparameter optimization algorithm suggested in [26]. Standard BO using Gaussian Processes models the conditional probability distribution $p(y|x)$, the probability of output $y$ given input
TPE models $p(x|y)$, the probability of input $x$ given output $y$ and $p(y)$, the probability distribution of $y$. It fits one Gaussian Mixture Model (GMM), $l(x)$, to the set of hyperparameter values with the best loss function values, and another GMM, $g(x)$, to the remaining hyperparameter values. The TPE algorithm maintains sorted lists of observed variables and chooses the hyperparameter value $x$ that maximizes the ratio $l(x)/g(x)$ [27].

The different methods being tested have different tuning parameters. ATPRK has 6 tuning parameters, $\text{Range}_{\text{min}}, \text{Sill}_{\text{min}}, \text{L}_{\text{range}}, \text{L}_{\text{sill}}, \text{rate}$, and $H$. These parameters define a search space internal to the method used to select semivariogram functions for kriging. SupReME has a single tuning parameter, the regularization strength $\lambda$. For S2Sharp, $p$, the number of dimensions of the subspace $U$, can be tuned, as well as the regularization strength for each of the subspace dimensions, $\lambda_j, j \in \{1, \ldots, p\}$. MuSA and SSSS both have regularization tuning parameters, $\lambda$ and $\mu$, due to using similar implementations of SALSA. For S2 SSC, learning rate and batch size and scale reduction are tuned.

Parameter tuning was not performed for DSen2 and Sen2res. DSen2 is provided as a pre-trained neural network and can not be effectively tuned without extensive retraining. Sen2res does expose some tuning parameters, but they do not lend themselves well to optimization and offer minimal improvement possibilities without prohibitive growth in running times.

3. Results

3.1. Real Data Experiments

3.1.1. Qualitative Comparisons Using Real S2 Data

An unprocessed 324 × 324 pixel crop from real S2 data seen in Figure 3, showing parts of central Reykjavik in Iceland, was super-resolved using the methods and parameter tuning previously described. The image is part of an S2 tile from the Copernicus Open Access Hub that is 10,000 × 10,000 pixels in the 10 m bands captured by Sentinel-2B on a descending orbit at 13:03 on the 17th of October 2019, with Level-2A atmospheric corrections applied [51]. It was chosen for the favorable cloud conditions and presents a very clear view of the capital of Iceland and the surrounding area. Results for each method can be seen in Figures 4 and 5, as 100 × 100 pixel crops from bands B9 and B12, as seen in Figure 3. Very few outliers in the images, having intensity values over 10,000, the expected maximum value, have been clipped and the images normalized by dividing every value by 10,000. These bands are highlighted here because they are particularly hard to super-resolve. No ground truth is available, so quantitative image quality metrics can not be used, and the qualitative comparison is strictly visual.

Figure 3. An RGB image generated [52,53] using an S2 image of central Reykjavik, Iceland. A crop in the upper left corner of the image, shown on the right, is used in Figures 4 and 5.
Figure 4. Band B9 of real S2 image of Reykjavik super-resolved to 10 m resolution.

For the 60 m B9 in Figure 4, ATPRK, S2Sharp, and SSSS appear to give comparatively good results. Rooftop outlines and street layouts are easily discernable and the image contrast looks natural. SupReME, DSen2 and Sen2res, are each successively blurrier and Sen2res returns much lower reflectance values than all the other methods resulting in much lower contrast. MuSA appears oversharpened with overbright outlier pixels. The methods all give fairly similar results for most of the 20 m bands but struggle with band B12 in particular. Building shapes are harder to discern and appear irregular compared to the
other bands, as seen in Figure 5. Sen2res and DSen2, remain slightly blurry compared to the other methods. S2Sharp, SSSS, SupReME, and S2 SSC appear most natural and visually pleasing. MuSA again appears oversharpened. Making further qualitative judgements is hard due to how similar the results appear.

3.1.2. Quantitative Comparisons Using Reduced Scale S2 Data

To make a quantitative comparison of semi-real S2 data, two reduced scale versions of the same S2 image were generated. First a $1620 \times 2124$ crop from the data was chosen and reduced resolution versions of all bands were generated at both $2 \times$ pixel size and $6 \times$ pixel size. This was done by applying appropriate filters to simulate the PSF of each S2 band to their appropriate bands before downgrading by a factor of 2 and 6, respectively.

The blurring being applied is nearly identical to that modelled in the ATPRK, SupReME, S2Sharp, MuSA, and SSSS methods. These data are also scaled to the same scales as the data used for the training process of the DSen2 neural network, while S2 SSC will downgrade the data even further for training. If scale invariance holds, this should not affect the results greatly, but the model based methods and DSen2 may still have a slight advantage. A further crop to $216 \times 216$ pixels for $Y_{10m}^D$, $108 \times 108$ in $Y_{20m}^D$, and $36 \times 36$ in $Y_{60m}^D$ was then made for each version of the image, mainly to shorten the running times.

The results can be seen in Tables 1 and 2, the best results on each line are bold.

Although the reduction methods used are the same as for the synthetic data in the following subsection, the errors are much larger for semi-real data. One obvious difference is that here all bands are processed to even lower resolution scales. The scale reduction of the 10 m bands may also be adversely affecting the upscaling processes. The characteristics of real S2 data appear to be worse represented at reduced scales than by synthesized data like that used in the following subsection.

Based on average RMSE for the $40 \rightarrow 20$ m task, the methods ordered by performance are S2 SSC, S2Sharp, ATPRK, SupReME, MuSA, DSen2, Sen2res, and SSSS, as seen in Table 1. SSSS performs significantly worse than its relatives, S2Sharp, SupReME, and MuSA. All methods give fairly good UIQI and SSIM results. For both metrics, the best performances are in bands B6, B7 and B8A. The wavelengths captured in bands B7 and B8A are a subset of the wavelengths captured by the 10 m band B8, as can be seen in Figure 1, and B6 is spectrally close. For SRE, the metric being optimized during parameter tuning, ATPRK, S2 SSC, and S2Sharp outperform all other methods by some margin in every 20 m band. None of the methods manage to get an ERGAS below 3 for this task, which would be considered a good score.

Table 2 shows that the results are quite different for the $360 \rightarrow 60$ m task. MuSA and SSSS perform comparatively well, SSSS giving the best results, followed by MuSA, DSen2, SupReME, S2Sharp, ATPRK, and Sen2res. ATPRK does not perform as well for this data at a greatly reduced scale as for other data tested and in fact the magnitude of the RMSE is greatly increased at this scale for all methods apart from SSSS. ERGAS scores are also fairly high (lower is better) and SRE scores are low (higher is better) for all methods. UIQI and SSIM are also much worse than with the smaller scale reduction, particularly for DSen2 and Sen2res.

Table 1. Evaluation using reduced scale S2 data: Reduced by 2, super-resolved 40 $\rightarrow$ 20 m. The best results are given in bold typeface.

| Method  | 20 m  | 40 m  | 80 m  | 160 m | 320 m | 40 m  |
|---------|-------|-------|-------|-------|-------|-------|
| RMSE    | 132.80| 210.79| 253.24|       |       |       |
| SAM     | 5.23  | 6.36  | 4.72  | 4.38  |       |       |
| ERGAS   | 8.14  | 4.38  | 9.81  |       |       |       |
| Time    | 0.25  | 0.44  | 0.33  |       |       |       |

Table 2. Evaluation using reduced scale S2 data: Reduced by 2, super-resolved 360 $\rightarrow$ 60 m. The best results are given in bold typeface.

| Method  | 360 m | 60 m  | 120 m | 240 m | 360 m |
|---------|-------|-------|-------|-------|-------|
| RMSE    | 132.80| 210.79| 253.24|       |       |
| SAM     | 5.23  | 6.36  | 4.72  | 4.38  |       |
| ERGAS   | 8.14  | 4.38  | 9.81  |       |       |
| Time    | 0.25  | 0.44  | 0.33  |       |       |
Table 1. Cont.

|       | ATPRK | DSen2 | MuSA | S2Sharp | S2 SSC | SSSS | SupReME | Sen2res |
|-------|-------|-------|------|---------|--------|------|---------|---------|
| UIQI  |       |       |      |         |        |      |         |         |
| B5    | 0.94  | 0.90  | 0.91 | 0.94    | 0.93   | 0.90 | 0.92    | 0.84    |
| B6    | 0.96  | 0.91  | 0.93 | 0.94    | 0.94   | 0.91 | 0.93    | 0.85    |
| B7    | 0.95  | 0.91  | 0.94 | 0.94    | 0.94   | 0.91 | 0.93    | 0.85    |
| B8A   | 0.96  | 0.92  | 0.94 | 0.95    | 0.95   | 0.93 | 0.94    | 0.86    |
| B11   | 0.94  | 0.83  | 0.88 | 0.94    | 0.94   | 0.87 | 0.93    | 0.83    |
| B12   | 0.92  | 0.82  | 0.74 | 0.92    | 0.92   | 0.79 | 0.90    | 0.80    |
| 20 m  | 0.95  | 0.88  | 0.89 | 0.94    | 0.94   | 0.89 | 0.92    | 0.84    |
| SSIM  |       |       |      |         |        |      |         |         |
| B5    | 0.94  | 0.88  | 0.90 | 0.92    | 0.92   | 0.89 | 0.91    | 0.82    |
| B6    | 0.95  | 0.90  | 0.92 | 0.93    | 0.93   | 0.90 | 0.92    | 0.82    |
| B7    | 0.95  | 0.90  | 0.93 | 0.93    | 0.94   | 0.90 | 0.92    | 0.82    |
| B8A   | 0.96  | 0.91  | 0.93 | 0.94    | 0.94   | 0.92 | 0.93    | 0.84    |
| B11   | 0.93  | 0.80  | 0.86 | 0.93    | 0.93   | 0.86 | 0.91    | 0.80    |
| B12   | 0.90  | 0.78  | 0.72 | 0.91    | 0.91   | 0.77 | 0.88    | 0.76    |
| 20 m  | 0.94  | 0.86  | 0.88 | 0.93    | 0.93   | 0.87 | 0.91    | 0.81    |
| SRE [dB]  |       |       |      |         |        |      |         |         |
| B5    | 18.53 | 16.40 | 17.85| 19.22   | 19.17  | 17.31| 18.10   | 14.27   |
| B6    | 21.69 | 18.45 | 20.31| 21.30   | 21.54  | 18.43| 19.81   | 15.89   |
| B7    | 21.89 | 18.75 | 20.90| 21.68   | 21.99  | 18.70| 20.26   | 16.16   |
| B8A   | 22.64 | 19.25 | 21.16| 21.97   | 22.77  | 19.78| 20.83   | 16.78   |
| B11   | 21.16 | 16.59 | 17.53| 21.74   | 22.40  | 16.88| 19.90   | 16.09   |
| B12   | 16.63 | 14.12 | 12.30| 17.58   | 18.47  | 12.65| 16.10   | 13.41   |
| 20 m  | 20.43 | 17.26 | 18.34| 20.58   | 21.06  | 17.29| 19.16   | 15.43   |

Table 2. Evaluation using reduced scale S2 data: Reduced by 6, super-resolved 360 → 60 m. The best results are given in bold typeface.

|       | ATPRK | DSen2 | MuSA | S2Sharp | S2 SSC | SSSS | SupReME | Sen2res |
|-------|-------|-------|------|---------|--------|------|---------|---------|
| RMSE↓ | 60 m  | 337.50| 299.28| 277.25  | 317.93 | 200.08| 316.84  | 367.78  |
| ERGAS↓| 60 m  | 8.05  | 9.29  | 6.53   | 7.33   | 5.99 | 7.06   | 9.08    |
| Time ↓| Average| 11.44 | 0.45  | 1111.44| 8.53   | 40.24| 4.11   | 11.24   |
| STD   |       | 0.12  | 0.08  | 13.14  | 0.08   | 0.36 | 0.04   | 0.09    |
| UIQI↑ | B1    | 0.616 | 0.442 | 0.632  | 0.546  | 0.573| 0.598  | 0.46    |
| B9    | 0.579 | 0.615 | 0.655 | 0.593  | 0.712  | 0.589| 0.49   |
| 60 m  | 0.598 | 0.528 | 0.644 | 0.570  | 0.643  | 0.594| 0.48   |
| SSIM↑ | B1    | 0.533 | 0.372 | 0.573  | 0.470  | 0.507| 0.516  | 0.39    |
| B9    | 0.512 | 0.566 | 0.607 | 0.531  | 0.641  | 0.529| 0.44   |
| 60 m  | 0.523 | 0.469 | 0.590 | 0.500  | 0.574  | 0.522| 0.42   |
| SRE [dB]↑ | B1 | 8.56  | 6.60  | 10.44  | 9.53   | 10.48| 10.00  | 7.40    |
| B9    | 11.36 | 12.86 | 13.06 | 11.84  | 16.28  | 11.84| 10.66  |
| 60 m  | 9.96  | 9.73  | 11.75 | 10.68  | 13.38  | 10.92| 9.03   |

3.2. Synthesized Data Experiments

Using the synthetic datasets in Figure 6, quantitative comparisons can be performed. The images shown here as RGB color photos generated from the appropriate spectral bands, are all generated from hyperspectral images. Images a through f, which originate from [4–7], and the processes used to synthesize them are described in Appendix A. They present varied scenes with coastal, agricultural, rural, urban, and mixed geographies. Section 3.2.2 evaluates the methods using all of the datasets in Figure 6. Section 3.2.1 focuses on Figure 6g, which was generated from AVIRIS [54] data for this paper.
3.2.1. Quantitative Comparisons Using Synthesized S2 Data

The data for this experiment, seen in Figure 6g, were simulated using an AVIRIS image with 5 m resolution employing methods similar to [5,6]. The original AVIRIS image of an area near San Diego, CA, was acquired in November 2011. It is 790 × 6796 pixels with 5 m resolution in 224 bands that span a spectral range of 366–2496 nm that covers the range of the S2 sensors.

The S2 version was simulated by applying the S2 spectral response to a patch of the AVIRIS data depicting parts of the city of Escondido and its suburbs. Twelve bands of size 648 × 648 pixels at a spatial resolution of 5 m were obtained. These bands were filtered and subsampled and cropped to yield 252 × 252 pixel ground truth for all bands at 10 m resolution. The bands were then filtered using a 12 × 12 Gaussian approximation of the S2 sensor PSF from [49], and downgraded to yield four bands of size 252 × 252 pixels at 10 m resolution, six bands of size 126 × 126 pixels at 20 m resolution, and two bands of size 42 × 42 pixels at 60 m resolution. Because the PSF blur is applied to the synthesized image in a near identical manner to that modeled by the ATPRK, SupReME, S2Sharp, MuSA, and SSSS methods, those methods should be expected to perform particularly well on this data. The tuned version of S2 SSC also uses the same PSF blur during training. Sen2res and DSen2, on the other hand, do not explicitly model the S2 PSF in the same way, so they could be expected to perform worse.

Table 3 shows a comparison between the super-resolution methods. SupReME gets the best results for the 60 m bands while ATPRK does slightly better for the 20 m bands. Based on average RMSE for all bands, the methods ordered by performance are S2Sharp, ATPRK, SupReME, SSSS, MuSA, DSen2, and Sen2res. For the 20 m bands ATPRK is followed by S2 SSC, slightly outperforming S2Sharp. Otherwise, the order remains mostly the same apart from MuSA and SSSS trading places. For 60 m band performance, SupReME is followed by S2Sharp, SSSS, ATPRK, MuSA, Sen2res, and DSen2.
Table 3. Evaluation using S2 data simulated from AVIRIS image of Escondido. The best results are given in bold typeface.

|               | ATPRK | DSen2 | MuSA | S2Sharp | S2 SSC | SSSS | SupReME | Sen2res |
|---------------|-------|-------|------|---------|--------|------|---------|---------|
| **RMSE** ↓    | 20 m  | 40.18 | 85.38| 59.95   | 44.67  | 41.20| 69.91   | 53.04   | 116.95  |
|               | 60 m  | 50.33 | 113.28| 99.11   | 25.77  | -    | 37.75   | 25.13   | 100.54  |
|               | All   | 42.94 | 93.15 | 71.77   | 40.78  | -    | 63.42   | 47.62   | 113.07  |
| **SAM** ↓     | 20 m  | 1.65  | 2.60 | 2.58    | 1.97   | 1.68 | 2.73    | 2.14    | 3.78    |
|               | All   | 2.22  | 4.19 | 4.12    | 2.02   | -    | 2.74    | 2.17    | 4.30    |
| **ERGAS** ↓   | 20 m  | 2.57  | 4.96 | 3.83    | 2.86   | -    | 4.14    | 3.32    | 6.48    |
|               | 60 m  | 1.44  | 4.95 | 2.37    | 0.75   | -    | 0.93    | 0.37    | 2.69    |
| **Time** ↓    | Average| 14.01| 0.50 | 1461.99 | 13.24 | 15.53| 49.48   | 6.28    | 15.08   |
|               | STD   | 0.034 | 0.08 | 0.18    | 0.61   | 0.21 | 0.20    | 0.60    |
| **UIQI** ↑    | B1    | 0.88  | 0.49 | 0.95    | 0.96   | -    | 0.98    | 0.94    | 0.77    |
|               | B5    | 0.98  | 0.90 | 0.96    | 0.97   | 0.98 | 0.97    | 0.96    | 0.89    |
|               | B6    | 0.99  | 0.93 | 0.98    | 0.99   | 0.99 | 0.97    | 0.96    | 0.73    |
|               | B7    | 1.00  | 0.94 | 1.00    | 1.00   | 0.99 | 0.98    | 1.00    | 0.86    |
|               | B8A   | 1.00  | 0.94 | 1.00    | 1.00   | 0.99 | 0.98    | 1.00    | 0.86    |
|               | B9    | 0.92  | 0.86 | 0.95    | 0.98   | -    | 0.95    | 0.97    | 0.44    |
|               | B11   | 0.89  | 0.77 | 0.78    | 0.83   | 0.91 | 0.74    | 0.84    | 0.79    |
|               | B12   | 0.94  | 0.77 | 0.86    | 0.92   | 0.95 | 0.84    | 0.90    | 0.79    |
| **SSIM** ↑    | B1    | 0.88  | 0.48 | 0.94    | 0.96   | -    | 0.98    | 0.93    | 0.77    |
|               | B5    | 0.98  | 0.88 | 0.95    | 0.97   | 0.97 | 0.97    | 0.94    | 0.87    |
|               | B6    | 0.99  | 0.92 | 0.97    | 0.99   | 0.99 | 0.97    | 0.96    | 0.69    |
|               | B7    | 1.00  | 0.93 | 1.00    | 1.00   | 0.99 | 0.98    | 0.99    | 0.83    |
|               | B8A   | 1.00  | 0.94 | 1.00    | 1.00   | 0.99 | 0.98    | 1.00    | 0.84    |
|               | B9    | 0.93  | 0.86 | 0.95    | 0.98   | -    | 0.94    | 0.97    | 0.41    |
|               | B11   | 0.87  | 0.72 | 0.75    | 0.80   | 0.88 | 0.71    | 0.80    | 0.75    |
|               | B12   | 0.93  | 0.71 | 0.84    | 0.90   | 0.93 | 0.82    | 0.88    | 0.74    |
| **SRE [dB]** ↑| B1    | 22.42 | 8.35 | 27.06   | 27.84  | -    | 31.84   | 27.04   | 18.44   |
|               | B5    | 25.37 | 19.19| 23.03   | 25.36  | 24.71| 24.89   | 23.07   | 18.45   |
|               | B6    | 28.48 | 21.39| 25.12   | 27.36  | 29.21| 24.66   | 23.32   | 15.93   |
|               | B7    | 33.14 | 21.61| 34.17   | 34.96  | 30.00| 25.72   | 29.76   | 17.91   |
|               | B8A   | 35.26 | 22.08| 35.51   | 36.41  | 30.74| 26.22   | 33.03   | 18.40   |
|               | B9    | 20.85 | 17.07| 14.47   | 26.72  | -    | 22.92   | 27.11   | 14.63   |
|               | B11   | 22.23 | 19.76| 18.23   | 20.83  | 23.15| 17.97   | 21.10   | 20.16   |
|               | B12   | 24.79 | 20.16| 20.70   | 23.55  | 25.34| 19.99   | 22.96   | 20.45   |
| **20 m**      |       |       |       |         |       | -    | 27.11   | 27.11   | 14.63   |
| **60 m**      |       |       |       |         |       | -    | 27.38   | 27.07   | 16.54   |
| **All**       |       |       |       |         |       | -    | 27.28   | 25.92   | 18.05   |

SSSS shows particularly strong results in the 60 m bands, getting the best UIQI, SSIM, and SRE scores for B1. DSen2, in contrast, does very badly in band B1 and has the lowest average 60 m SRE of all the methods, while Sen2res does slightly worse in all other metrics. ATPRK, MuSA, S2Sharp, S2 SSC, SSSS, and SupReME all get very high 20 m UIQI and SSIM scores, with only bands B11 and B12 presenting a significant challenge. Depending on the application, any of these methods might be useful for the lower wavelength 20 m bands but based on the low RMSE and high SRE values, ATPRK, S2 SSC, and S2Sharp have a clear advantage.

As noted, due to the lack of PSF modeling, Sen2res and DSen2 were expected to perform slightly worse than the competing methods.

For all data, running times are dominated by DSen2. It processes this particular data in approximately half a second. The slowest method by far is MuSA, which takes almost
25 min to process the same image. The other methods are much quicker, taking about 15 s each, apart from SSSS which takes slightly under 1 min.

Error maps showing a graphical representation of the error between the estimates of each method and the reference ground truth as logarithmically scaled residual images, calculated using \( e = \log\left(1 + \frac{255 \hat{Y} - Y}{10,000}\right) \), can be seen in Figures 7 and 8. These residual images for bands B9 and B12, respectively, are shown on a scale of 0 to 4. These bands are again highlighted because they are hard to super-resolve.

MuSA appears to have a constant error in the 60 m band, B9, as can be seen in Figure 7c but does much better for B12. Strange geometrical patterns are also visible in the 60 m→10 m ATPRK image in subfigure a. The disadvantages of DSen2 and Sen2res are visible in subfigures b and f, respectively. Their residuals are larger than for the remaining methods. SSSS appears to give the best results in this case, with the least apparent structural textures visible in subfigure e, despite Table 3 showing that it has higher 60 m RMSE than S2Sharp and SupReME, whose localized errors are more eye-catching.

The results for B12 are more similar. The fact that building shapes are discernible in most error maps suggests structural image definition has been lost compared to the ground truths. In Figure 8 the errors appear to be more evenly distributed over the whole image for DSen2 and Sen2res. More localized errors reveal clear building shapes with Musa, S2Sharp, SSSS, and SupReME. ATPRK and S2 SSC give the visibly best result, which broadly matches the SRE results in Table 3.

![Error Maps](image1.png)

**Figure 7.** Absolute error map of Escondido crop, band B9, on log scale. Lower values are better.
3.2.2. Evaluating Effects of Parameter Tuning

The parameter tuning method was evaluated for each of the super-resolution methods by testing them on the 7 datasets seen in Figure 6. The 6 additional synthesized S2 datasets were originally used for testing some of the methods in their original papers and their generation is described in greater detail in Appendix A. These images all come with ground truths and were super-resolved using both default and tuned parameters.

Parameter tuning was performed as previously described and the boxplots of Figure 9 show a comparison of the SRE values for each method. While ATPRK, MuSa and SupReME show almost no improvements, S2Sharp, S2 SSC, and SSSS show significant gains from the use of parameter tuning method.

ATPRK performs a parameter search of its own to select the best values for the semivariograms used to generate the kriging matrix. The parameters being tuned here merely provides initialization for that search which might explain why ATPRK does not show much of an improvement. Parameter tuning does not appreciably improve MuSa performance and tweaking the single \( \lambda \) parameter of SupReME does not seem to affect the
results in any significant manner either. These methods may also have already been tuned to the ideally suited parameters.

Focusing on the mean SRE, represented by a line inside each box on the boxplot, S2Sharp, S2 SSC, and SSSS all benefit greatly from the use of tuned parameters. Unlike the related MuSA method the selection of different $\lambda$ and $\mu$ is very effective for SSSS. The end results still remain rather unimpressive compared to the other methods, even with tuning. SSSS appears ill suited for the type of synthetic data being tested.

For a fixed number of iterations, the tuning being performed on the S2 SSC batch size and learning rate is probably affecting how quickly the network finds a solution more than the quality of the solution. The improvement is mainly a result of the change in how training images are generated. Using PSF modeling for the synthesized data increases the methods performance significantly. It should still be reiterated that S2 SSC only processes the 20 m bands.

S2Sharp goes from middling, using default parameters, to leading performance when tuned. The performance boost is to be expected as the method was developed with an associated parameter tuning method presented in [5] and not really intended to be used with the default parameters for all data. As was shown in [28] the parameter tuning method used here can further improve its results.

Broadly, Figure 9 shows that with tuning S2Sharp gives the best results, on average, and S2 SSC does best for the 20 m bands. Without tuning ATPRK, MuSA, and SupReME can potentially perform better. The other methods, DSen2, Sen2res, and SSSS do not perform as well.

4. Conclusions

It is apparent that three methods, ATPRK, S2Sharp, and S2 SSC, outperform the other methods used in the experiments. ATPRK is easy to use and gives good results without the need for specifically tuning it to the data and may be good enough for many applications. Although improved results can be gained by using S2Sharp with tuned parameters, this requires parameter tuning for each dataset that is to be processed. It is less practical in general use even though on average it outperforms the other methods quantitatively for the data tested. For 20 m $\rightarrow$ 10 m super-resolution S2 SSC yields the best performance, highlighting the potential of deep learning methods. The tuning method proposed, while somewhat time-consuming, can be easily applied to all S2 super-resolution methods that benefit from parameter tuning. It was shown to greatly improve the results for S2Sharp and SSSS. The tuning parameters selected for the other methods had little effect on the end results, suggesting their default parameters were already a good fit for the data being tested and that the tuning method can reliably find similarly good parameters using reduced scale data.

Author Contributions: Investigation, data curation and writing—original draft preparation, S.E.A.; writing—review and editing, M.O.U., J.S. and J.R.S.; software, S.E.A., M.O.U. and H.V.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by The Icelandic Research Fund under Grant 207233-051.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://aviris.jpl.nasa.gov/, https://scihub.copernicus.eu/, https://github.com/lanha/SupReME.

Acknowledgments: We would like to thank the authors of Sen2res, Nicolas Brodu (GEOSTAT team at Inria), and of SupReME and DSen2, Charis Lanaras (ETH Zürich), for making their code publicly available, as well as Chia-Hsiang Lin (National Cheng Kung University), author of SSSS, Claudia Paris (University of Trento), author of MuSA, and Qunming Wang (Tongji University), author of ATPRK, for providing code and data upon request.
Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations
The following abbreviations are used in this manuscript:

- ATPRK Area-To-Point Regression Kriging
- C-BM3D Color Block Matching 3D Denoiser
- DSen2 Deep Sentinel-2
- ERGAS relative dimensionless global error
- GAN Generative Adversarial Networks
- GMM Gaussian Mixture Model
- MS Multi-Spectral
- MTF Modulation Transfer Function
- MuSA Multiresolution Sharpening Approach
- PSF Point Spread Function
- RMSE Root-Mean-Squared Error
- S2 Sentinel-2
- S2 SSC Sentinel-2 Symmetric Skip Connection
- S2Sharp Sentinel-2 Sharpening
- SALSA Split Augmented Lagrangian Shrinkage Algorithm
- SAM Spectral Angle Mapper
- SRE Signal-to-Reconstruction Error
- SRF Spectral Response Functions
- SSIM Structural Similarity Index Measure
- SSSS Sentinel-2 Super-resolution via Scene-adapted Self-similarity
- SupReME Super-resolution for multiresolution estimation
- TPE Tree-structured Parzen Estimator
- UIQI Universal Image Quality Index

Appendix A. Descriptions of Datasets
Descriptions of datasets used in Section 3.2.2 follow. Example data was provided with some of the published code for the methods in Section 2, and some authors were kind enough to share test data upon request.

Appendix A.1. S2Sharp
This dataset is simulated using an AVIRIS image of an area just outside Sacramento, CA, that was acquired in August 2013. The image is $408 \times 408$ pixels in 197 usable bands that span a similar spectral range as the S2 sensor. The S2 version was simulated by applying the S2 spectral response to the AVIRIS data. Twelve bands of size $408 \times 408$ pixels at a spatial resolution of 3.2 m were obtained.

Appropriate bands were filtered using an estimated point spread function of the S2 sensor and downgraded to yield four bands of size $408 \times 408$ pixels at 3.2 m resolution, six bands of size $204 \times 204$ pixels at 6.4 m resolution, and two bands of size $68 \times 68$ pixels at 19.2 m resolution. Zero-mean identically and independently distributed (i.i.d.) Gaussian noise is added for 40 dB SNR [5]. The scale ratio between bands is therefore the same, although the actual pixel sizes differ from S2’s 10 m, 20 m and 60 m.

Appendix A.2. SupReME
The SupReME method includes a sample S2 dataset simulated from part of the Open Science APEX dataset [55,56]. The method to generate this dataset is nearly identical to that used in the S2Sharp paper. The spatial resolution of this image is $198 \times 198$ pixels at 2 m resolution for $Y_{10m}$, $99 \times 99$ pixels at 4 m resolution for $Y_{20m}$ and $33 \times 33$ pixels at 12 m resolution for $Y_{60m}$. i.i.d. Gaussian noise is also added for 40 dB SNR [4].
Appendix A.3. MuSA and SSSS

The MuSA and SSSS methods both use the same four AVIRIS images of differing scenes showing coastal, mountainous, urban and agricultural geography. Ground-truth images were generated by low-pass filtering and subsampling the 5 m coastal and mountain data by 2 and the, respectively, 3.5 m and 3.2 m urban and crop datasets by 3.

After low-pass filtering, the S2 spectral responses were applied, followed by the S2 sensor PSF’s and further subsampling of $Y_{20m}$ and $Y_{60m}$ so that all bands have approximately the same pixel sizes as actual S2 data. Finally i.i.d. zero-mean Gaussian noise is added for 40 dB SNR [6].

References
1. European Space Agency. Sentinel-2 Spectral Response Functions (S2-SRF). Sentinel-2 MSI Document Library. 2017. Available online: https://sentinel.esa.int/web/sentinel/user-guides/sentinel-2-msi/document-library (accessed on 2 June 2021).
2. Kaplan, G. Sentinel-2 Pan Sharpening—Comparative Analysis. Proceedings 2018, 2, 345. [CrossRef]
3. G03 Committee. Tables for Reference Solar Spectral Irradiances: Direct Normal and Hemispherical on 37 Tilted Surface; Technical Report; ASTM International: West Conshohocken, PA, USA, 2012. [CrossRef]
4. Lanaras, C.; Bioucas-Dias, J.M.; Baltsavias, E.; Schindler, K. Super-Resolution of Multispectral Multiresolution Images from a Single Sensor. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 1505–1513. [CrossRef]
5. Ulfarsson, M.O.; Palsson, F.; Mura, M.D.; Sveinsson, J.R. Sentinel-2 Sharpening Using a Reduced-Rank Method. IEEE Trans. Geosci. Remote Sens. 2019, 57, 6408–6420. [CrossRef]
6. Paris, C.; Bioucas-Dias, J.M.; Bruzzone, L. A Novel Sharpening Approach for Superresolving Multiresolution Optical Images. IEEE Trans. Geosci. Remote Sens. 2019, 57, 1545–1560. [CrossRef]
7. Lin, C.H.; Bioucas-Dias, J.M. An Explicit and Scene-Adapted Definition of Convex Self-Similarity Prior With Application to Unsupervised Sentinel-2 Super-Resolution. IEEE Trans. Geosci. Remote Sens. 2020, 58, 3352–3365. [CrossRef]
8. Wang, Q.; Shi, W.; Li, Z.; Atkinson, P.M. Fusion of Sentinel-2 Images. Remote Sens. Environ. 2016, 187, 241–252. [CrossRef]
9. Brodú, N. Super-Resolving Multiresolution Images With Band-Independent Geometry of Multispectral Pixels. IEEE Trans. Geosci. Remote Sens. 2017, 55, 4610–4617. [CrossRef]
10. Lanaras, C.; Bioucas-Dias, J.M.; 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]
11. Palsson, F.; Sveinsson, J.R.; Ulfarsson, M.O. Sentinel-2 Image Fusion Using a Deep Residual Network. Remote Sens. 2018, 10, 1290. [CrossRef]
12. Nguyen, H.V.; Ulfarsson, M.O.; Sveinsson, J.R.; Sigurdsson, J. Zero-Shot Sentinel-2 Sharpening Using a Symmetric Skipped Connection Convolutional Neural Network. In Proceedings of the IGARSS 2020—2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 26 September–2 October 2020; pp. 613–616. [CrossRef]
13. Panagiotopoulou, A.; Grammatikopoulos, L.; Charou, E.; Bratsolis, E.; Madamopoulos, N.; Petrogonas, J. Very Deep Super-Resolution of Remotely Sensed Images with Mean Square Error and Var-Norm Estimators as Loss Functions. arXiv 2020, arXiv:2007.15417.
14. Kim, J.; Lee, J.K.; 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. [CrossRef]
15. Zioga, E.; Panagiotopoulou, A.; Stefouli, M.; Charou, E.; Grammatikopoulos, L.; Bratsolis, E.; Madamopoulos, N. Spatial Resolution Increase of Remotely Sensed Images from Mining Areas with Deep Learning Techniques. arXiv 2020, arXiv:2007.08791.
16. Wu, J.; He, Z.; Hu, J. Sentinel-2 Sharpening via Parallel Residual Network. Remote Sens. 2020, 12, 279. [CrossRef]
17. Zhang, R.; Cavallaro, G.; Jitsev, J. Super-Resolution of Large Volumes of Sentinel-2 Images with High Performance Distributed Deep Learning; Technical Report FZI-2020-04989; Jülich Supercomputing Center: Jülich, Germany, 2020.
18. Esquena Calm, O. Super-Resolution of Sentinel-2 Images a Single Network for All-Bands Upsampling. Bachelor’s Thesis, Universitat Politècnica de Catalunya, Barcelona, Spain, 2020.
19. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the 27th International Conference on Neural Information Processing Systems, NIPS’14; MIT Press: Cambridge, MA, USA, 2014; Volume 2, pp. 2672–2680.
20. Palsson, F.; Sveinsson, J.R.; Ulfarsson, M.O. Single Sensor Image Fusion Using A Deep Convolutional Generative Adversarial Network. In Proceedings of the 2018 9th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), Amsterdam, The Netherlands, 23–26 September 2018; pp. 1–5. [CrossRef]
21. Zhang, R.; Sumbul, G.; Demir, B. An Approach to Super-Resolution of Sentinel-2 Images Based on Generative Adversarial Networks. arXiv 2020, arXiv:1912.06013.
50. Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R.P.; de Freitas, N. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proc. IEEE* 2016, 104, 148–175. [CrossRef]

51. European Space Agency. S2B_MSIL2A_20191017T130259_N0213_R038_T27WVM_20191017T134229; Copernicus Open Access Hub. 2021. Available online: https://scihub.copernicus.eu/ (accessed on 2 June 2021).

52. Magnusson, M.; Sigurdsson, J.; Armansson, S.E.; Ulfarsson, M.O.; Sveinsson, J.R.; Deborah, H. Creating RGB Images from Hyperspectral Images Using a Color Matching Function. In Proceedings of the IGARSS 2020—2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 26 September–2 October 2020; pp. 2045–2048. [CrossRef]

53. Sovdat, B.; Kadunc, M.; Batić, M.; Milčinski, G. Natural Color Representation of Sentinel-2 Data. *Remote Sens. Environ.* 2019, 225, 392–402. [CrossRef]

54. NASA Jet Propulsion Laboratory. Flight f111116401p00r12; AVIRIS Data Portal. 2021. Available online: https://aviris.jpl.nasa.gov/ (accessed on 2 June 2021).

55. European Space Agency. APEX; APEX Open Science Data Set. 2021. Available online: https://apex-esa.org/en/data/free-data-cubes (accessed on 2 June 2021).

56. Schaepman, M.; Jehle, M.; Hueni, A.; D’Odorico, P.; Damm, A.; Weyermann, J.; Schneider, F.D.; Laurent, V.; Popp, C.; Seidel, F.C.; et al. Advanced radiometry measurements and Earth science applications with the Airborne Prism Experiment (APEX). *Remote Sens. Environ.* 2015, 158, 207–219. [CrossRef]