Land Surface Temperature Super-Resolution with a Scale-Invariance-Free Neural Approach: Application to MODIS

Romuald Ait-Bachir, Carlos Granero-Belinchon, Aurélie Michel, Julien Michel, Xavier Briottet, Lucas Drumetz

Land Surface Temperature Super-Resolution with a Scale-Invariance-Free Neural Approach: Application to MODIS

Romuald Ait-Bachir*†, Carlos Granero-Belinchon*†, Aurélie Michel‡, Julien Michel§, Xavier Briottet‡ and Lucas Drumetz*†

*IMT Atlantique, Lab-STICC, UMR 6285, 29238, CNRS, Brest, France

†ODYSSEY Team-Project, INRIA Ifremer IMT-Atl., 35042, CNRS, Brest, France

‡ONERA-DOTA, University of Toulouse, F-31055 Toulouse, France

§ CESBIO, University de Toulouse, CNES, CNRS, INRAE, IRD, UT3, 31401 Toulouse, France.

***Abstract*—Due to the trade-off between the temporal and spatial resolution of thermal spaceborne sensors, super-resolution methods have been developed to provide fine-scale Land Surface Temperature (LST) maps. Most of them are trained at low resolution but applied at fine resolution, and so they require a scale-invariance hypothesis that is not always adapted. The main contribution of this work is the introduction of a Scale-Invariance-Free approach for training Neural Network (NN) models, and the implementation of two NN models, called Scale-Invariance-Free Convolutional Neural Network for Super-Resolution (SIF-CNN-SR) for the super-resolution of MODIS LST products. The Scale-Invariance-Free approach consists on training the models in order to provide LST maps at high spatial resolution that recover the initial LST when they are degraded at low resolution and that contain fine-scale textures informed by the high resolution NDVI. The second contribution of this work is the release of a test database with ASTER LST images concomitant with MODIS ones that can be used for evaluation of super-resolution algorithms. We compare the two proposed models, SIF-CNN-SR1 and SIF-CNN-SR2, with four state-of-the-art methods, Bicubic, DMS, ATPRK, Tsharp, and a CNN sharing the same architecture as SIF-CNN-SR but trained under the scale-invariance hypothesis. We show that SIF-CNN-SR1 outperforms the state-of-the-art methods and the other two CNN models as evaluated with LPIPS and Fourier space metrics focusing on the analysis of textures. These results and the available ASTER-MODIS database for evaluation are promising for future studies on super-resolution of LST.**

***Index Terms*—Super-Resolution, LST, Neural Networks, MODIS, ASTER**

I. Introduction

A. Overview

Land Surface Temperature (LST) is one of the Essential Climate Variables (ECVs) to describe the changing climate of the Earth [1]. It refers to the temperature of the Earth's surface and is a key variable in the physics of the land surface and its interactions with the atmosphere. It is used for a wide range of applications, such as the detection and characterization of urban heat islands (UHI) and heatwaves [2], [3], [4], [5], droughts [6], [7], [8], forest fires [9], [10], inland water bodies [11], or for the analysis of warming trends in different parts of the world [12], [13], [14] among others [15]. It also allows to estimate the surface energy balance [16] and the evapotranspiration of the vegetation [17]. LST is retrieved from two different kinds of sensors: those using Passive MicroWave (PMW) wavelengths and those using Thermal InfraRed (TIR) ones. On the one hand, PMW sensors acquire for all weather conditions, even through rain or clouds, but current PMW-based sensors present very low spatial resolutions of the order of 10-25 km [18]. On the other hand, commonly used TIR sensors require cloud-free conditions but present better spatial resolutions of 70 m to 1 km [19]. Both PMW and TIR sensors provide information at regional or global scales, but only TIR sensors provide LST maps at a sufficient spatial resolution for monitoring and studying local phenomena [20], [21]. Although TIR sensors provide finer spatial resolution maps, there is a trade-off between spatial resolution and temporal one: while high spatial resolution sensors with resolutions of around 100 m present very low acquisition frequencies of the order of one image per month or less (due to clouds), lower spatial resolution sensors such as MODIS, with resolutions of the order of 1 km acquire around two images per day. MODIS provides one of the most widely used LST products because it spans more than two decades [22], [23]. As LST displays a large heterogeneity in both space and time, monitoring the LST at both high temporal and spatial resolution is needed for a significant number of applications within the land and solid Earth, health and hazards and security and surveillance frameworks, as highlighted by [24]. For example, studies related to public health, volcanology, or urban climatology would benefit from LST series with a higher acquisition frequency [25], [26], [27]. Thus, with the objective of providing LST maps with both high spatial and temporal resolutions, there is a growing literature in super-resolution methods aimed at increasing the spatial resolution of the LST images, notably for MODIS [28], [29]. Although there are studies on improving the spatial resolution of PMW-based LST maps [30], [31], [32], this field is beyond the scope of this paper that aims to address the trade-off of TIR sensors.

corresponding author: carlos.granero-belinchon@imt-atlantique.fr

This research was funded by C.N.E.S in the A.P.R CES "Théia Température de Surface et Émissivité" (2022-2025) framework.

B. Related Works

Super-resolution of Land Surface Temperature images is commonly done with statistical models based on empirical relationships between LST and land features estimated from Visible and Near InfraRed (VNIR) satellite images. This approach was initially justified by the inverse relationship between Normalized Difference Vegetation Index (NDVI) and LST [33], [34]. Later, several studies showed that VNIR indices are adapted for LST super-resolution [35], [36], [37]. The most commonly used statistical models for LST super-resolution are based on simple linear regressions between the LST and VNIR features. Some of these models are Distrad, TsHARP, ATPRK or AATPRK [38], [39], [40], [41], [42]. Granero-Belinchon et al. 2019 illustrated that these models are only slightly dependent on the VNIR feature used in the regression and that ATPRK (Area-To-Point Regression Kriging) and AATPRK (Adaptive-Area-To-Point-Regression-Kriging) with a Kriging interpolation for the residuals' definition slightly outperform TsHARP and DisTrad [37]. Recently, more complex models such as Data Mining Sharpener (DMS) using regression trees appeared [43], [44], [45]. All these models use relationships between the LST and VNIR features that are optimized at the LST coarse resolution and applied at the high resolution of the VNIR feature. So, these models use a reduced scale training approach.

In the recent years, Deep Learning has been widely used for super-resolution applications of remote sensing images in the VNIR, where a lot of different approaches have been developed: Convolutional Neural Networks (CNN) [46], [47], Generative Adversarial Networks [48] or diffusion models [49] among others. However, Deep Learning has been only scarcely used for LST super-resolution [50], [51], [52]. A CNN has been used on images from MODIS [50] to increase the spatial resolution by a factor of four. Multilayer perceptrons have also been used to downscale MODIS LST to Landsat spatial resolution (by a factor of ten) [52] and more recently, [51] proposed a diffusion model to downscale Landsat LST by factors of four and eight. However, in the field of remote sensing, a frequent difficulty in super-resolution is the absence of a high resolution reference image to be used in the training step. This is particularly true for LST super-resolution. So, in order to apply supervised learning schemes, models are commonly trained at reduced scale, *i.e.* trained at degraded lower resolutions using the initial images as a reference to later apply these trained models at higher resolutions [50], [51], [52].

All the aforementioned statistical and deep learning models use a scale-invariance hypothesis: models optimized at coarse resolution can be used at high resolution. This hypothesis can be problematic depending on the heterogeneity of the studied surface or the range of scales studied.

C. Motivation and contributions

Several studies demonstrated that this scale-invariance hypothesis can lead to bad resolved small scale textures [50], [37], [53]. Changes in the statistical laws learned at different resolutions [37] lead to an important loss in performance of the models when applied to resolutions higher than the one of training [50]. To overcome this limitation, this article presents a new deep learning model that is trained at full scale, *i.e.* trained without scale-invariance hypothesis. The optimization problem statement of this model is based on a variational formulation [54], [55].

The main contributions of this research are:

- A Scale-Invariance-Free Convolutional Neural Network for super-resolution (SIF-CNN-SR). This model increases the spatial resolution of a given low resolution LST by imposing that 1) the initial low resolution LST image is recovered when the spatial resolution is degraded, and 2) the high resolution LST presents small scale textures from the NDVI. These two constraints correspond respectively to consistency and synthesis properties [56]. This approach aims at reconstructing the high frequencies with better fidelity.
- The demonstration that the SIF-CNN-SR approach presents a performance at the state of the art level while greatly outperforming some commonly used super-resolution approaches such as TsHARP and ATPRK.
- A super-resolution evaluation dataset associating concomitant coarse resolution MODIS LST images and high resolution ASTER LST ones which allows to evaluate LST super-resolution algorithms on non simulated data to quantify their performance.

II. Scale-Invariance-Free Convolutional Neural Network for super-resolution

This section presents the proposed model for LST super-resolution, as well as the NN architecture.

A. Model

Our model is inspired by classical variational schemes:

$$D(\theta) = \arg\min_{\theta}(||y - H\left(x(\theta)\right)|| + \lambda R(\theta)) \qquad (1)$$

where x is the reconstructed state of the system, y is an observation of the system state, H is an operator modeling the observation y from the state x, R is a regularization term used to impose desired properties to the reconstructed state x, λ is a constant weighting the significance of the regularization term and $||\,||$ is a given norm. Both x and R depend on θ, the parameters to be optimized. We can cast variational approaches as an optimization problem with two criteria: being able to obtain a state of the system x that is close to the observation y and matching the properties imposed by the regularization term.

We propose a Neural Network model (SIF-CNN-SR) to downscale a LST image observed at low spatial resolution $T_{obs}^{(l)}$ to a higher spatial resolution $T_{sr}^{(h)}$. The superscripts (l) and (h) indicates respectively low and high resolution products. The model is trained in a self-supervised way and takes as input a couple of (LST, NDVI) images of the same region acquired at the same time. The input LST correspond to the initial observation at low resolution $T_{obs}^{(l)}$ while the input NDVI is acquired at high resolution $V_{obs}^{(h)}$. The model is trained in order

to provide LST at high resolution that 1) when degraded to low spatial resolution recovers the initial LST used as input and 2) contains fine-scale textures informed by the high resolution NDVI used as input. Making the parallelism with variational approaches, these two objectives correspond respectively to provide a super-resolution LST close to the observed one, and close to a "background state" defined by the texture of the NDVI.

In order to ensure both conditions, the optimization problem is written as follows:

$$\hat{\theta} = \arg\min_{\theta} \alpha \left[J\left(\gamma G(V_{obs}^{(h)}), G(\Psi_\theta(V_{obs}^{(h)}, T_{obs}^{(l)}))\right)\right] + \\ + (1-\alpha)\left[J\left(T_{obs}^{(l)}, H(\Psi_\theta(V_{obs}^{(h)}, T_{obs}^{(l)}))\right)\right] \quad (2)$$

where Ψ_θ is the model providing high-resolution LST, θ are the parameters of the model to be learned, G is an operator defining small scale textures and H is an observation operator modeling the degradation of spatial resolution from high to low resolution. γ is a scaling coefficient to adapt the amplitude of the textures of NDVI and LST in the comparison. Finally, α is a weight coefficient used to compensate the significance of each term in the optimization and $J(a, b)$ is a given discrepancy measure comparing a and b. The LST at high spatial resolution provided by the model with optimized parameters $\hat{\theta}$ is:

$$T_{sr}^{(h)} = \Psi_{\hat{\theta}}(V_{obs}^{(h)}, T_{obs}^{(l)}) \quad (3)$$

We identify two terms in (2): a reconstruction term and a texture one.

$$\mathcal{L}_{rec} = \left[J\left(T_{obs}^{(l)}, H(\Psi_\theta(V_{obs}^{(h)}, T_{obs}^{(l)}))\right)\right] \quad (4)$$

$$\mathcal{L}_{texture} = \left[J\left(\gamma G(V_{obs}^{(h)}), G(\Psi_\theta(V_{obs}^{(h)}, T_{obs}^{(l)}))\right)\right] \quad (5)$$

The reconstruction term $\mathcal{L}_{rec}$ is used to obtain physical values of the LST since it seeks to recover the observed LST ($T_{obs}^{(l)}$) from the super-resolution LST ($T_{sr}^{(h)}$) by degrading the spatial resolution of the latter. Consequently, the observation operator H consists in a low pass filtering of $T_{sr}^{(h)}$ simulating the effect of the thermal sensor Modulation Transfer Function (MTF) [57]. This low pass filtering is done by convolving with a Gaussian kernel K. Then, a bicubic interpolation is used to reduce the size of the image to match the size of $T_{obs}^{(l)}$.

The texture term $\mathcal{L}_{texture}$ is inspired by the style-transfer defined by [58]. It is used to make the model add small scale textures that are present in $V_{obs}^{(h)}$ into $T_{sr}^{(h)}$ *i.e.* to transfer the textures from the NDVI into the LST. Two definitions of the small scale texture operator G are used:

- The first one combines the gradient with two additional diagonal derivatives. We write it $G(I) = \nabla_D I$ and is computed in practice by convolving any image I with four derivative Sobel filters of size 3×3 following respectively horizontal, vertical and diagonal directions.
- The second one consists in a high-pass filtering, done in practice by:

$$G(I) = I - K * I \quad (6)$$

with K the Gaussian kernel modeling the MTF of the thermal sensor. So, the high-frequency content of any image I is defined as the difference between the image itself and a low-pass filtered version $K * I$.

B. Architecture

The proposed SIF-CNN-SR model is not constrained by a specific architecture for Ψ_θ. In this work, Ψ_θ is a U-Net [59], which is a fully convolutional neural network with an encoder-decoder architecture and long-skip connections between the encoder and the decoder at each level. More particularly, our model grounds on the Multi-Residual U-Net used in Nguyen et al. [50] to downscale the LST of MODIS. U-Nets are commonly used for super-resolution applications [60], [61], [62].

Our multi-residual U-net takes as input the concatenation ($||$) along the channel dimension of $V_{obs}^{(h)}$ and the bicubic interpolation of $T_{obs}^{(l)}$ and is made of an encoder and its symmetrical decoder, each one with 3 levels, see figure 1. The model first present two consecutive convolutional blocks (Conv2d, Batchnorm2d, ReLU) and then the three levels of the encoder. Each level of the encoder consists of an average pooling reducing the size of the image by two on each dimension, followed by three convolutional blocks with a residual connection between the input of the first block and the output of the second one. The decoder has also three levels, each one with a non-trainable bilinear interpolation increasing the size of the image by two, followed by two convolutional blocks. Finally the output of the last convolutional block of the decoder is passed through a 2d convolutional layer. The corresponding stages of the encoder and the decoder are linked with concatenated long skip connections. All the 2d convolutional layers have kernel size 3, stride 1, padding 1 in replicate mode, and null bias. The total number of trainable parameters is 417 009 [1].

III. Application to MODIS LST super-resolution

In this work, we use SIF-CNN-SR to downscale MODIS LST from 1 km of spatial resolution to 250 m, that is the spatial resolution of the NDVI from MODIS. The model takes as input a couple of (LST, NDVI) MODIS images of the same region acquired at the same time. The model is trained in order to provide MODIS LST at 250 m that 1) when degraded into 1 km of spatial resolution recovers the initial MODIS LST and 2) contains fine-scale textures informed by the MODIS NDVI at 250 m used as input. For validation, we use ASTER data. Both sensors are described in the following.

A. Data

This section introduces the specifications of the Moderate-Resolution Imaging Spectroradiometer (MODIS) and the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) sensors and their products. Then, it presents the construction of the training, validation, and test datasets used in this study.

[1]The trained models are available at https://github.com/cgranerob/Land-Surface-Temperature-Super-Resolution-with-a-Scale-Invariance-Free-Neural-Approach.

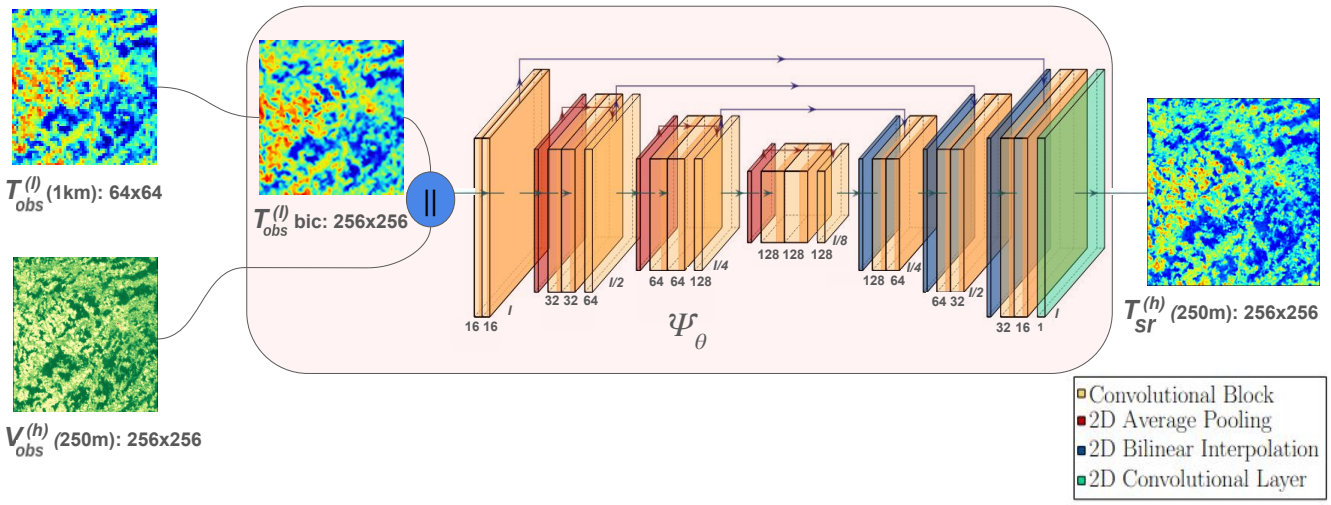

Fig. 1. Schematic representation of the U-net architecture of Ψ_θ. Orange blocks correspond to convolutional blocks with (Conv2d, Batchnorm2d, ReLU), red blocks are average pooling reducing the size of the image by a factor two along each dimension, blue blocks are bilinear interpolation increasing the size of the image by a factor two and the green block is a two dimensional convolutional layer. The number of channels of each two dimensional convolutional layer is indicated in the figure. The input is the concatenation (||) along the channel dimension of the high resolution NDVI $V_{obs}^{(h)}$ and the bicubic interpolation of $T_{obs}^{(l)}$. The output is the super-resolution $T_{sr}^{(h)}$ at the spatial resoluion of $V_{obs}^{(h)}$.

1) MODIS: MODIS is a sensor that acquires data in 36 spectral bands from the visible to the thermal infrared domains. MODIS is onboard the TERRA and AQUA platforms allowing the observation of the entire surface of the Earth with a revisit time of four times per day considering both satellites. Each MODIS image covers an area of 1200x1200 km^2.

In this work, we focus on two MODIS products: MOD21A1 which provides Land Surface Temperature at 1 km of spatial resolution and MOD09GQ which provides reflectances in the Red and NIR domains at 250 m of spatial resolution [63], [64]. Both products provide images of the same land surfaces and acquired at the same acquisition time. The LST MOD21A1 product results from the application of the Temperature and Emissivity Separation (TES) algorithm [65] to the radiance of the three MODIS bands in the TIR: bands 29, 31 and 32. MOD21A1D only contains LST images acquired during daytime from the MODIS sensor installed on the TERRA satellite. We combine the NIR and Red reflectances from MOD09GQ to generate NDVI images.

2) ASTER: ASTER is also a sensor onboard the TERRA platform. It acquires data over 14 spectral bands going from the VNIR to the TIR spectral domains. ASTER provides LST images at 90 m of spatial resolution, and a revisit time of 16 days. In this work, we use the ASTER AST_08 product which provides LST retrieved from the application of the TES algorithm to the 5 TIR bands of ASTER [66].

B. Training, validation and evaluation datasets

1) Training and validation dataset: The training and validation dataset contains LST and NDVI concomitant images coming from the MOD21A1D and MOD09GQ MODIS products, see table I. It spans over three years, from January 1st 2017 until December 31st 2019. The used images (tile h18v04) cover the center of Europe, mainly France, Germany, Italy, Austria and Switzerland, see figure 2. In total, 1095 pairs of MODIS (LST, NDVI) images covering areas of 1200×1200 km^2 are used. No nighttime images were considered to avoid any registration error between daytime NDVI and nighttime LST as well as to avoid different thermal dynamics between day and night that can complicate the learning process.

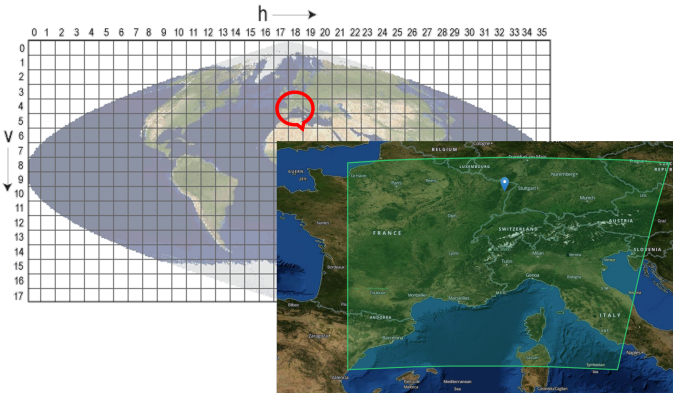

Fig. 2. Center Europe area (h18v04 MODIS tile) used in this study.

The data is further pre-processed by slicing the images into smaller patches covering a land surface of 64×64 km^2 (64×64 and 256×256 pixels for low resolution and high resolution products respectively) without clouds neither pixels evaluated by the NASA as low quality. In total, the training and validation dataset is made up of 8233 pairs of patches. It is then split following a 0.6/0.4 ratio leading to a training dataset containing 4900 pairs of images and a validation dataset containing 3333 pairs of images. During training, each LST and NDVI image is centered and standardized using the mean and variance of the dataset.

TABLE I
SUMMARY INFORMATION OF TRAINING, VALIDATION AND EVALUATION DATASETS.

Sensor	Product	Region	Time period	Spatial Resolution	Images for training	Images for validation	Images for evaluation
MODIS	MOD21A1D (LST)	Center Europe	2017-2019	1km	4900	3333	0
	MOD09GQ (NDVI)			250m			
MODIS	MOD21A1D (LST)		2020-2023	1km	0	0	79
	MOD09GQ (NDVI)			250m			
ASTER	AST_08 (LST)			90m			

2) Evaluation dataset: The dataset for evaluation contains concomitant MODIS and ASTER images covering most of the area of the MODIS tile h18v04 *i.e.* the same region used in the training. These images are acquired during three years starting from January 1st 2020 (so without overlapping with the training dataset). For each date, the corresponding MODIS data (MOD21A1D and MOD09GQ) is associated to the AST_08 data, see table I. Concomitant MODIS and ASTER LST acquisitions are needed because of the fast dynamics of the temperature that can lead to strong LST evolutions in times of the order of 1 hour [67].

In order to pair ASTER and MODIS images several pre-processing steps are done. First, the ASTER spatial resolution is degraded from 90 m to 250 m (the spatial resolution of the NDVI image of MODIS) by convolving it with a Gaussian kernel with half-kernel width of 250 m. Second, they are both projected into the same UTM (Universal Transverse Mercator) coordinate system and co-registered. Third, since ASTER images cover smaller areas, the MODIS image is cropped to cover the same region as the ASTER one. The evaluation dataset contains 79 pairs of concomitant LST images from ASTER and (LST, NDVI) couples from MODIS of 64×64 km^2.

Even if ASTER and MODIS are both onboard the TERRA platform, small errors in co-registration and small differences in acquisition times can still exist. In particular, for the couples defining our evaluation dataset the mean difference in acquisition times between ASTER and MODIS is 1.5 minutes with a variance of 4.2 minutes. The maximum and minimum time differences observed are respectively of 15.1 and 1.2 minutes.

C. Training procedure

The model Ψ_θ takes as inputs a couple of MODIS (LST, NDVI) images. In order for both images to have the same dimensions and observe the same area the MODIS LST is downscaled using bicubic interpolation. The model Ψ_θ provides a super-resolution LST image.

The loss function optimized during the training procedure is defined in (2) with J being the Huber loss for this specific application since it is specially robust to outliers [68].

The choice of the loss hyperparameters, α and γ, is critical. On the one hand, $\alpha \in [0, 1]$ weights the reconstruction and texture terms of the loss and should be chosen to equilibrate the significance of each term in the minimization. On the other hand, γ is a scale factor used to compensate the difference of the range of variations between the gradients of NDVI and LST. This difference is still present after standardization. γ is usually negative due to the inverse relationship between the LST and the NDVI [33], [34].

In this study, two models with slightly different definitions of the small scale texture G are trained: SIF-CNN-SR1 and SIF-CNN-SR2. SIF-CNN-SR1 uses Sobel filters to define the small scales, see section II-A. The choice of $\gamma = -0.5$ was inferred from a small grid search on the values $\{-0.1, -0.25, -0.5, -0.75, -1\}$. α was set to 0.99. We expected a much higher texture loss than the reconstruction one, explaining the choice of $\alpha = 0.99$ to compensate. SIF-CNN-SR2 uses high-pass filtering to define the small scales, see section II-A. The choice of $\gamma = -0.25$ was inferred from a grid search on $\{-0.1, -0.25, -0.5, -0.75, -1\}$. α was set to 0.1 due to the alternative representation of the texture term leading to much lower texture loss values.

The learning rate was set to 0.0001 and the batch size to 32. The model was trained for 200 epochs. The training was made using Python 3.9.19 and Pytorch 2.3.0. It used a Quadro RTX 8000 with 46 GO of VRAM. Table II synthesizes the main differences between SIF-CNN-SR1 and SIF-CNN-SR2.

TABLE II
SYNTHESIS OF THE PARAMETERS FOR MODELS SIF-CNN-SR1 AND SIF-CNN-SR2 BOTH TRAINED DURING 200 EPOCHS WITH A LEARNING RATE OF 0.0001 AND BATCH SIZE OF 32. THE LOSS FUNCTION USED IS DEFINED IN (2) WITH H DEFINED AS THE MTF OF MODIS.

Model	α	γ	G
SIF-CNN-SR1	0.99	-0.5	Sobel filter
SIF-CNN-SR2	0.10	-0.25	High pass filter

D. Benchmarks

1) State-of-the-art statistical methods for super-resolution: Four common methods from the state of the art are used for benchmarking SIF-CNN-SR1 and SIF-CNN-SR2: Bicubic interpolation, TsHARP, ATPRK and DMS. While bicubic interpolation is just a two dimensional interpolation of the image based on polynomials, TsHARP, ATPRK and DMS are sharpening techniques specifically developed for remote sensing applications and ground on a established statistical relationship between LST and variables obtained from the VNIR domain, such as the NDVI we use here.

TsHARP, ATPRK and DMS share a common strategy for sharpening:

- First, they consider a given relationship between LST and NDVI at low resolution (the initial resolution of the LST).

$$T_{obs}^{(l)} = f(V_{obs}^{(l)}) + \Delta^{(l)}T \tag{7}$$

with $\Delta^{(l)}T$ the residuals between the model and the ground-truth at low resolution. In the case of TsHARP and ATPRK, f is a linear function, and the slope and intercept of the linear relationship are obtained by least squares minimization. In the case of DMS, f is defined by m5 regression trees [69] with optimized parameters. DMS is much more complex than TsHARP and ATPRK by being a piece-wise linear model due to the leaves of the m5 tree being linear regressions. It is based on multiple locally and a globally fitted bagged regression trees which link non-linearly a NDVI to a LST. All these methods define the low resolution residuals $\Delta^{(l)}T$ for each pixel as $\Delta^{(l)}T = T^{(l)}_{obs} - f(V^{(l)}_{obs})$.
Straightforward generalizations of TsHARP and ATPRK, replacing the linear f by a bilinear one, and the current version of DMS are able to take as inputs the red and NIR bands of MODIS separately. However, this approach has not been tested in this study.

- The learned parameters (slope and intercept for TsHARP and ATPRK, and decision trees for DMS) at low resolution are considered scale-invariant and used at fine resolution to provide a first estimate of $T^{(h)}_{sr}$.

$$T'^{(h)}_{sr} = f(V^{(h)}_{obs}) \tag{8}$$

- Finally, a small scale residual correction is done:

$$T^{(h)}_{sr} = T'^{(h)}_{sr} + \Delta^{(h)}T \tag{9}$$

The definition of the small scale residuals, $\Delta^{(h)}T$, is different depending on the method. In the case of TsHARP and DMS $\Delta^{(h)}T$ is obtained by a nearest-neighbor interpolation of $\Delta^{(l)}T$. In the case of ATPRK, $\Delta^{(h)}T$ is obtained by kriging from $\Delta^{(l)}T$ [37].

A detailed description of TsHARP and ATPRK can be found in [37] and references therein, while a detailed description of DMS can be found in [43]. The TsHARP and ATPRK implementations used for benchmarking were implemented following exactly [37] and are available at https://github.com/cgranerob/ThUnmpy. The DMS approach used as a benchmark was developped by Guzinski and Nieto [45], and is freely accessible at https://github.com/radosuav/pyDMS.

2) U-Net trained under scale invariance hypothesis: We train a Neural Network (NN) model with the same architecture as SIF-CNN-SR, described in section II-B, and trained on the same MODIS dataset, described in section III-B1, but with a reduced scale training approach. We call it SC-Unet. This model is trained in a supervised way on MODIS data at degraded spatial resolutions [50], *i.e.* the LST of MODIS at 1 km of resolution is used as reference while the inputs are degraded versions of the LST and the NDVI at 4 km and 1 km of spatial resolution, respectively. Thus, SC-Unet is trained to provide LST at 1 km from LST and NDVI at 4 km and 1 km respectively. The loss function used to train the model is just the MSE (Mean-Squared Error) between the LST provided by the model and the MODIS LST at 1 km of resolution. Once the model is trained, it is used to provide LST at 250 m from LST at 1 km and NDVI at 250 m. Thus, SC-Unet grounds on the scale-invariance hypothesis. The model was trained during 100 epochs, with batch size 32 and learning rate 0.0001.

E. Evaluation Metrics

Seven complementary metrics were used to measure the performances of super-resolution models: Root Mean Squared Error (RMSE), Root Mean Square Error averaged only on the quartile of pixels with higher gradients (RMSE_{75-100}) [70], Structural Similarity Index Measure (SSIM) [71], Learned Perceptual Image Patch Similarity (LPIPS) [72], Frequency Restoration Rate (FRR) [73], Frequency Restoration Overshoot (FRO) [73] and the RMSE between attenuation spectra $\mathbf{F}(\nu)$, ($\text{RMSE}(\mathbf{F}(\nu))$. All of them are based on comparisons between super-resolution LST ($T^{(h)}_{sr}$) and the reference LST ($T^{(h)}_{ref}$). While the RMSEs and SSIM are comparisons performed directly in the LST space, LPIPS performs a comparison in a learned latent space and FRR, FRO and $\text{RMSE}(\mathbf{F}(\nu))$ are evaluations performed in the Fourier spectral domain.

1) Metrics on the space of LST: The RMSE measures the overall error between $T^{(h)}_{sr}$ and $T^{(h)}_{ref}$:

$$RMSE(T^{(h)}_{sr}, T^{(h)}_{ref}) = \sqrt{\frac{1}{N}||(T^{(h)}_{sr} - T^{(h)}_{ref}||^2_F} \tag{10}$$

where N is the number of pixels of the images and $||\,||^2_F$ is the Frobenius norm.

We also study the RMSE on specific areas of the images defined in function of the amplitude of the gradients of their pixels. The RMSE_{75-100} is estimated on pixels with an amplitude of the gradient in the fourth quartile, *i.e.* on areas with very heterogeneous textures, where super-resolution is more difficult.

The SSIM measures the perception-based similarity between two images x and y and between their variations [71]. We use an implementation based on the computation of the mean and variance within a sliding square window of size 7 pixels (1750 m). This provides a SSIM value per pixel of the image. Finally, the SSIMs are averaged to get an overall score per image. The SSIM between two windows, one for each image being compared, both centered on the pixel (x_1, x_2) is:

$$SSIM(x,y) = l(x,y)^{\beta_1} c(x,y)^{\beta_2} s(x,y)^{\beta_3} \tag{11}$$

where l, c and s are the radiance, contrast and structure terms and are defined as:

$$l(x,y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \tag{12}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \tag{13}$$

$$s(x,y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3} \tag{14}$$

where μ_x and μ_y are, respectively, the mean value of the pixels in the square window for each image, σ_x, σ_y and σ_{xy} are respectively the variance of the pixels in the window of each image and the cross covariance. $c_1 = (K_1 L)^2$, $c_2 = (K_2 L)^2$

and $c_3 = c_2/2$ are constants that avoid the denominator getting reduced to 0. $K_1 = 0.01$, $K_2 = 0.03$ and L is the data range corresponding to the difference between the maximum and the minimum LST in both images. β_1, β_2 and β_3 are set to 1, as usually. Images that are close in both intensity and variations present SSIM values close to 1.

2) Metrics on latent and Fourier spaces: LPIPS is a deep similarity metric defined by Zhang et al. [72]. LPIPS combines the Mean Squared Error between several latent representations of the super-resolution image and the reference. These representations are obtained using the intermediary features of a pretrained CNN such as VGG [74]. The greater the resemblance between the two images, the lower LPIPS will be.

Three metrics were used to compare directly the frequency content of $T_{sr}^{(h)}$ and $T_{ref}^{(h)}$ in the two-dimensional Fourier space as proposed in [73]. To facilitate the comparison, isotropy is considered and then an average is done on the angular dimension of the two-dimensional Fourier spectra, *i.e.* only the radial dimension of the spectra is considered.

The power spectra are computed following:

$$F(\nu) = \frac{1}{\#U_{f_m,f_M}} \sum_{(\nu_1,\nu_2)\in U_{f_m,f_M}} |F(\nu_1,\nu_2)| \qquad (15)$$

where $U_{f_m,f_M} = \left\{(\nu_1,\nu_2) : f_m \leq \sqrt{\nu_1^2+\nu_2^2} < f_M\right\}$ defines the discrete spatial frequencies lying in the ring defined by f_m and f_M and $\#U_{f_m,f_M}$ is the number of elements in U_{f_m,f_M}.

The attenuation spectra are:

$$\mathbf{F}(\nu) = 10(log_{10}(F(\nu)) - log_{10}(F(0))) \qquad (16)$$

The three metrics used are:

- the RMSE($\mathbf{F}(\nu)$) between the attenuation spectrum of ASTER LST $T_{ref}^{(h)}$ and each super-resolution product $T_{sr}^{(h)}$.
- the FRR, described by Michel et al. [73], that compares the attenuation spectra of: the bicubic interpolation of $T_{obs}^{(l)}$ considered here as the worst possible super-resolution methods, the reference LST $T_{ref}^{(h)}$, and the super-resolution products $T_{sr}^{(h)}$. It measures the improvement in frequency reconstruction of $T_{sr}^{(h)}$ compared to bicubic interpolation of $T_{obs}^{(l)}$.
- the FRO, described by Michel et al. [73], measuring the overshoot of the attenuation spectrum of $T_{sr}^{(h)}$ compared to $T_{ref}^{(h)}$.

IV. Results

A. Performance of the models in central Europe

Table III shows the performance of Bicubic interpolation, TsHARP, ATPRK, DMS, SC-Unet, SIF-CNN-SR1 and SIF-CNN-SR2 evaluated with the seven metrics presented in section III-E averaged over the full evaluation dataset of central Europe. The best RMSE is obtained with SC-Unet and SIF-CNN-SR2. Indeed, these two models are the only ones with a RMSE value smaller than 2 K. However, only a small difference of 0.2 K appears between these models and bicubic interpolation and only 0.4 K with the highest RMSE value for ATPRK with overlapping standard deviation values between all the methods. Nevertheless, these metrics are averaged on the whole dataset and we expect the standard deviation of the RMSE will decrease with the size of the dataset. In addition, the RMSE for each method can fluctuate from one image to another, meaning that the most efficient method can change according to the image. SC-Unet and SIF-CNN-SR2 show the best performances on $RMSE_{75-100}$, *i.e.* when only highly heterogeneous areas are considered in the RMSE estimation. SC-Unet, SIF-CNN-SR1 and SIF-CNN-SR2 present the highest SSIM values with very similar performances: SC-Unet has SSIM=0.64, SIF-CNN-SR1 has SSIM=0.61 and SIF-CNN-SR2 has SSIM=0.62. Concerning the metrics focusing on textures, the best LPIPs score is obtained with SIF-CNN-SR1 (LPIPS=0.28) closely followed by SC-Unet (LPIPS=0.29). In the Fourier domain, the model SIF-CNN-SR1 has the smallest RMSE($\mathbf{F}(\nu)$), indicating that the attenuation spectra obtained with this model is the closer to the ASTER spectra. In addition, it presents the best FRR and so the best improvement of frequency reconstruction compared to the bicubic interpolation. Finally, the best FRO is obtained with SC-Unet and bicubic. Very importantly FRO and FRR should be analyzed together since while FRO is blind to underestimation of the attenuation spectra, FRR is blind to overestimation. On the other hand the analysis of RMSE($\mathbf{F}(\nu)$) should be done with care since this metric does not discriminate between under and overestimation of the spectrum.

Table III allows us to conclude that SC-Unet and SIF-CNN-SR2 present the smaller overall errors on RMSE. However, SIF-CNN-SR1 seems to better recover the textures as evaluated with LPIPS. In the Fourier domain, which is indicative both of textures and overall errors, the best performances are obtained by the proposed model SIF-CNN-SR1. SIF-CNN-SR1 generates LSTs with more high-frequency content compared to the other models as evaluated by FRR and RMSE($\mathbf{F}(\nu)$) but it slightly overestimate textures as indicated by FRO. These results quantify the presence of more visible textures inside the super-resolution made by SIF-CNN-SR1. The three convolutional NN outperform state of the art approaches by a high margin in most metrics.

Figure 3 a) shows the attenuation spectra of ASTER LST, the different super-resolution LST products and the MODIS NDVI averaged over the 79 images of the evaluation dataset. We can observe that the bicubic interpolation and SC-Unet tend to underestimate the middle range and small scale textures of the LST while TsHARP, ATPRK, DMS and SIF-CNN-SR1 tend to overestimate them. SIF-CNN-SR2 underestimates the middle range scales while overestimating small scales. TsHARP, ATPRK, DMS and SIF-CNN-SR2 present an attenuation spectra following the shape of the attenuation spectra of the MODIS NDVI. This can be understood since the MODIS NDVI is used in these models to define the small scale textures. Very interestingly, the attenuation spectra of SIF-CNN-SR1 and SC-Unet present a dynamic across scales that follows the one of ASTER LST, even if they also use

TABLE III
AVERAGE RMSE, $RMSE_{75-100}$, SSIM, LPIPS, FRR, FRO AND $RMSE(\mathbf{F}(\nu))$ OVER THE EVALUATION DATASET CONTAINING 79 DAYTIME IMAGES OVER CENTRAL EUROPE FOR THE FIVE STUDIED SUPER-RESOLUTION APPROACHES. BOLD INDICATES THE BEST AVERAGE PERFORMANCE FOR EACH GIVEN METRIC AND BRACKETS () INDICATE THE STANDARD DEVIATION.

Methods	RMSE	$RMSE_{75-100}$	SSIM	LPIPS	FRR	FRO	$RMSE(\mathbf{F}(\nu))$
Bicubic	2.1 (0.7)	2.9 (0.8)	0.46 (0.09)	0.39 (0.03)	0.00 (0.00)	**0.00 (0.00)**	5.0 (0.9)
TsHARP	2.2 (0.7)	2.8 (0.9)	0.56 (0.08)	0.34 (0.05)	0.95 (0.05)	0.04 (0.01)	2.9 (0.6)
ATPRK	2.3 (0.7)	2.8 (0.9)	0.54 (0.08)	0.32 (0.04)	0.95 (0.10)	0.04 (0.01)	2.6 (0.7)
DMS	2.1 (0.7)	**2.5 (0.9)**	0.58 (0.09)	0.31 (0.03)	0.96 (0.06)	0.04 (0.01)	2.6 (0.6)
SC-Unet	**1.9 (0.7)**	**2.5 (0.9)**	**0.64 (0.07)**	0.29 (0.04)	0.62 (0.13)	**0.00 (0.00)**	1.8 (0.6)
SIF-CNN-SR1	2.2 (0.7)	2.6 (0.8)	0.62 (0.08)	**0.28 (0.04)**	**0.98 (0.03)**	0.03 (0.02)	**1.6 (0.7)**
SIF-CNN-SR2	**1.9 (0.7)**	**2.5 (0.8)**	0.61 (0.07)	0.32 (0.04)	0.79 (0.10)	0.01 (0.01)	1.8 (0.3)

MODIS NDVI to define the small scales. SIF-CNN-SR1 closely matches the ASTER LST spectra until scales ~ 500 m and overestimates the texture for smaller scales. On the other hand, SC-Unet underestimates the textures for scales smaller than 1 km. Figure 3 b) shows the error between the ASTER attenuation spectrum and the attenuation spectra of each super-resolution product. SIF-CNN-SR1 avoids significant errors at low and intermediary resolutions compared to SC-Unet and SIF-CNN-SR2 and it also avoids significant errors at high resolutions compared to DMS, TsHARP and ATPRK.

For a qualitative and deeper statistical analysis of the models' performances, we randomly choose one image from the evaluation dataset, then we study it visually as well as its attenuation spectrum and its distribution of the values of the high-pass filtered LST[2]. Figure 4 allows the visual comparison of the observed MODIS LST at 1 km and the LST from ASTER at 250 m together with the LSTs at 250 m obtained with the different methods explained in section III-D. While Bicubic is significantly blurred compared to ASTER LST and is visually similar to the MODIS at 1 km, TsHARP and ATPRK present similar patterns with more textures than ASTER LST. SC-Unet and SIF-CNN-SR2 also present blurrier LSTs than ASTER and the methods that seem to better recover the LST from ASTER are DMS and SIF-CNN-SR1.

Figure 5 a) represents the high-frequency filtered content of the super-resolved LSTs. ATPRK and TsHARP tend to overestimate the textures, while Bicubic, SC-Unet and SIF-CNN-SR2 tend to underestimate them. DMS and SIF-CNN-SR1 present the closest distributions of high-frequency filtered LSTs. Figure 5 b) illustrates the attenuation spectrum of ASTER LST in red, MODIS NDVI in dashed red and super-resolution LST from bicubic interpolation in magenta, TsHARP and ATPRK in blues and DMS in green, SC-Unet in black, SIF-CNN-SR2 in orange and SIF-CNN-SR1 in brown. As expected from previous results, bicubic interpolation SC-Unet and SIF-CNN-SR2 tends to underestimate mid-range and high frequencies of the LST. On the other hand, ATPRK, TsHARP, and DMS tend to overestimate high frequencies of the LST. SIF-CNN-SR1 presents the attenuation spectra closer to ASTER but with slightly overestimated high frequencies.

[2]The same analysis for each image of the validation dataset can be found in the supplementary materials and at https://github.com/cgranerob/Land-Surface-Temperature-Super-Resolution-with-a-Scale-Invariance-Free-Neural-Approach.

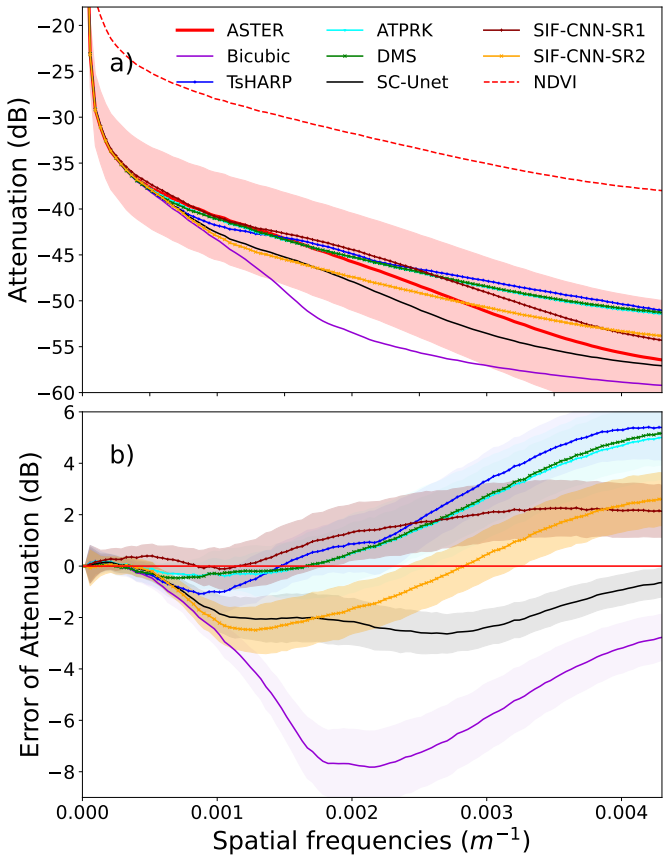

Fig. 3. a) Attenuation spectra of the ASTER LST (red), LST obtained with statistical super-resolution methods TsHARP and ATPRK (blue), DMS (green), SC-Unet (black), SIF-CNN-SR1 (brown) and SIF-CNN-SR2 (orange). The attenuation spectra of the MODIS NDVI is also shown in dashed red. The curves correspond to the mean spectra over the full test dataset. The shadow around the ASTER attenuation spectra correspond to $\pm$ a standard deviation around the mean value. b) Mean Error between ASTER attenuation spectrum and the attenuation spectra of TsHARP and ATPRK (blue), DMS (green), SC-Unet (black), SIF-CNN-SR1 (brown) and SIF-CNN-SR2 (orange). The curves correspond to the mean error over the full test dataset and the shadowed areas correspond to one standard deviation around the mean. The red horizontal line indicates the zero.

V. DISCUSSION

A. Comparison of the different models

All the super-resolution models studied in this work use as principal hypothesis to extract high resolution textures the relationship existing between the LST and the NDVI.

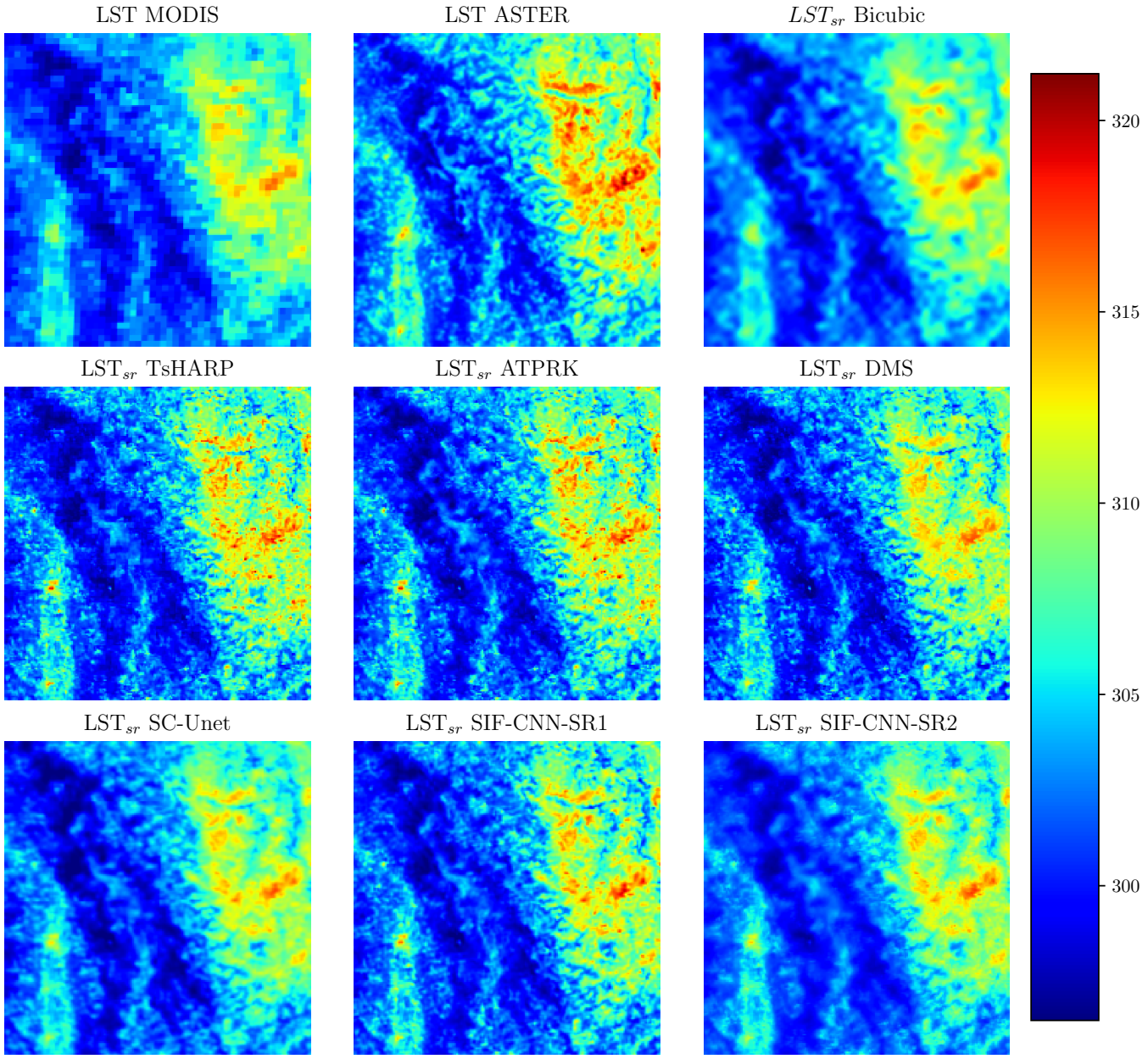

Fig. 4. For one random image from the validation dataset, visualization of the LST of MODIS and ASTER respectively at 1km and 250m of spatial resolution alongside the super-resolution LST obtained with the different approaches.

This relationship has been demonstrated by many [33], [34], [35]. However, this hypothesis is exploited differently by the different models.

First, TsHARP, ATPRK, DMS and SC-Unet use a scale-invariance hypothesis, *i.e.* the statistical models are learned at a spatial resolution that is lower than the spatial resolution of application. In the case of SIF-CNN-SR1 and SIF-CNN-SR2 there is not the scale-invariance hypothesis since the training is performed directly at the spatial resolution of application.

Second, TsHARP, ATPRK and DMS can be considered as models that predict the high resolution LST only from the high resolution NDVI: $T_{sr}^{(h)} = f_\theta(V_{obs}^{(h)})$ where f_θ has been trained using $(T_{obs}^{l}, V_{obs}^{l})$. In addition, they use a residual correction to overcome the use of an analytical relationship between NDVI and LST. This correction is computed at low resolution and interpolated at high resolution through a nearest neighbor approach for TsHarp and DMS or kriging for ATPRK, see III-D. On the other hand, SC-Unet, SIF-CNN-SR1 and SIF-CNN-SR2 predict the high resolution LST by directly taking as input both NDVI and LST. Our CNN models do not need any a posteriori correction.

Third, classical statistical methods require to learn a new f for each new data to process. These methods mostly learn to predict LST from the input NDVI through the function f. As NDVI does not contain important instantaneous drivers of the LST such as meteorological conditions, this relationship can only hold for a given instant, and can not generalize to different locations or times. This leads to a learned f

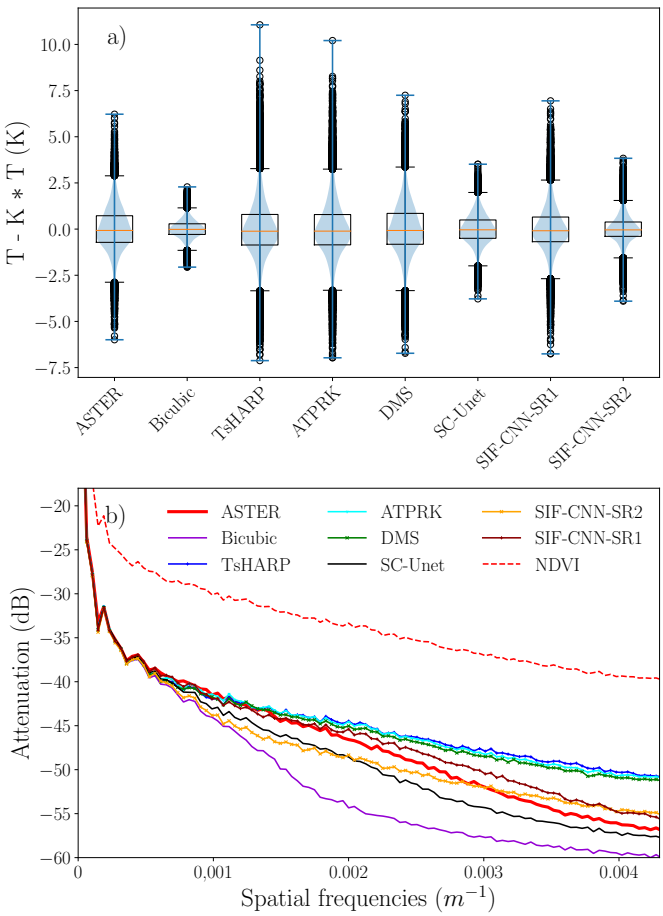

Fig. 5. Statistical analysis of the image visualized in figure 4. a) Boxplots and Violinplots representing the statistical distribution of the values of the high pass filtered LST (see equation 6) of ASTER and obtained with the different super-resolution approaches. b) Attenuation spectra of the ASTER LST (red), LST obtained with statistical super-resolution methods TsHARP and ATPRK (blue), DMS (green), SC-Unet (black), SIF-CNN-SR1 (brown) and SIF-CNN-SR2 (orange). The attenuation spectra of the MODIS NDVI is also shown in dashed red.

that specifically targets the observed area, but it is also a limitation that prevents to use more complex mapping f, because of the cost of re-training and the limited training data available to do it. On the other end, our proposed approach predict high resolution LST from both low resolution LST and high resolution NDVI, and is thus able to generalize to other locations and times, with comparable LST patterns, by using a more complex mapping f based on CNN, while saving inference time due to the absence of re-training.

Summarizing, the three main differences between the presented models are 1) the scale-invariance hypothesis, 2) training a function that explicitly depends on both NDVI and LST and 3) training a function able to generalize to different locations and times. These three differences can explain the better performance of our models with respect to the state of the art.

Another criteria to discriminate the super-resolution algorithm is their inference time since it strongly variates from one algorithm to another. The mean times for inference over the full test dataset are Bicubic 0.00007 s, TsHARP 0.01 s, ATPRK 19.56 s, DMS 0.31 s, and the three NNs 0.03 s. Our NNs present inference times of the order of TsHARP, ten times smaller than DMS and more than one hundred times smaller than ATPRK.

B. SIF-CNN-SR1 versus SIF-CNN-SR2

The main difference between SIF-CNN-SR1 and SIF-CNN-SR2 is the definition of the texture operator G. While in SIF-CNN-SR1, G corresponds to the gradients defined with Sobel filters, in SIF-CNN-SR2, G is a high-pass filter defined through a convolution with a Gaussian kernel. The gradient allows SIF-CNN-SR1 to generate LST products whose attenuation spectra better follows the evolution across spatial frequencies of the attenuation spectra of ASTER. The high-pass filter seems to not characterize scales between 1km and 300m leading to underestimated attenuation spectra of SIF-CNN-SR2 at these scales.

In particular, SIF-CNN-SR1 is promising for applications on LST from other sensors such as TRISHNA since it presents performance at the level of the state of the art, or even better when we look at the Fourier space metrics. Indeed, at higher spatial resolutions (< 100 m) the scale-invariance hypothesis seems to be more problematic [37]. The size of the pixel starts to be of the order of the observed objects and the heterogeneity of some landscapes make the scale-invariance hypothesis non-adapted. In addition, the scale-invariance hypothesis is specially not adapted for complex models such as Deep Learning models or DMS that fit better the NDVI-LST relationship at the scale of training. We expect models based on simple linear regression to be less impacted by the scale-invariance hypothesis [75], *i.e.* simple models define more generic LST-NDVI relationships, and so even if the learned relationship is not scale-invariant, the impact of this hypothesis on final performance is reduced.

C. On the evaluation metrics

In this study, we used seven complementary evaluation metrics. On the one hand, we studied metrics estimated directly on LST values: 1) RMSEs on the full images and in some parts of the images where the gradients are especially important and 2) SSIM. The RMSEs look at the average difference in Kelvin between super-resolution and reference images. Even if the information provided by these metrics is interesting, the RMSE, looking directly at LST values (not textures) and being averaged over whole areas, is not necessarily adapted to evaluate super-resolution performance. The SSIM looks at differences in the perceptual structure of the images. It depends on the LST values of the image and on their spatial distribution. SSIM is rarely discriminant in super-resolution. SSIM and RMSE may be prone to biases in the presence of radiometric or geometric distortions in cross-sensors dataset, therefore they will be jointly analyzed with complementary metrics [73]. On the other hand, we studied metrics focusing on the characterization of textures: 1) LPIPS working on latent learned representations and 2) RMSE($\mathbf{F}(\nu)$), FRR and FRO that are metrics on the Fourier domain. These metrics are less sensitive to specific LST values and are more concerned by

the scales resolved by the models. We consider that they are more representative for the characterization of the performance of super-resolution models. To the best of our knowledge, these latter metrics are not commonly used in studies of super-resolution of LST. This work highlights the interest of considering them for future studies.

D. On the relationship between the NDVI and the LST

The models considered in this study rely on an inverse relationship between the NDVI and the LST. In fact, the NDVI is associated with vegetation cover, and, at daytime, LST tends to be cooler for vegetated areas than for bare soils or impervious surfaces, so the LST and the NDVI are generally negatively correlated. Several works have studied the relationship between LST and NDVI in detail illustrating some variability depending on time, land cover and sensor. [76] showed a positive relationship between the LST and the NDVI during winter with GOES-8 data at 8km spatial resolution. [77] found varying negative correlations according to the season with low negative correlations during winter with Landsat 8 data. They also found that this relationship varies according to land cover, with the highest negative correlations for vegetated areas as expected and lower for built-up areas. For water bodies, the LST can be cooler than for vegetation but the NDVI generally has negative or close to zero values as pointed by [33].

The models SIF-CNN-SR1 and SIF-CNN-SR2 use a negative γ value due to the general assumption of the inverse relationship between the NDVI and the LST. However, these models are trained considering all seasons and different land cover types, and this can degrade their performance on some specific images with important water bodies. As a perspective, an adaptive γ value could be defined for each image to consider specific cases due to the variability of this relationship.

E. On the use of ASTER for validation

Another contribution of this work is the generation of a test database with pairs of concomitant ASTER LST and MODIS (LST,NDVI) images that will be useful for future studies on LST super-resolution[3]. The ASTER and MODIS LST products are obtained with the TES algorithm in order to reduce differences in the LST retrieval. However, we observed a bias between these LSTs. This bias can be explained since the number and spectral range of the TIR bands of ASTER and MODIS are different [78]. In addition, it may depend on the spatial heterogeneity of the observed landscape [79], [80]. Also, ASTER captures the LST variability at 90m which allows to better recover fine textures when MODIS has a smoother LST due to its spatial resolution of 1 km. These differences, together with other parameters such as the viewing angle, the intrinsic error of the TES algorithm or co-registration errors between ASTER and MODIS images, lead to a bias between the ASTER and MODIS LST that is different for the different couples of images of the validation dataset. Most of the images present a bias between 0 K and 2 K and only two couples present a bias higher than 2 K. The evaluation metrics based on the RMSE depend linearly on this bias: the higher the bias, the higher the RMSE. This is not the case for LPIPS, SSIM or Fourier space metrics which emphasize the interest of these metrics for the evaluation of super-resolution methods. Although an attempt to reduce the bias between ASTER and MODIS was found [81], this is an ongoing issue that should be treated in the future to improve the validation of super-resolution methods [29].

The ASTER LST images at 250 m of spatial resolution used for validation are obtained by low-pass filtering the initial ASTER LST at 90 m, see section III-B2. This low-pass filtering introduce a blurring effect that depends on the filter but its impact is supposed insignificant for spatial scales larger than 250 m. The energy content, as characterized by the attenuation spectra, of MODIS NDVI at small scales ~ 250m is significantly higher than the energy content of ASTER LST at these scales. This implies that the MODIS NDVI presents more small scale textures, *i.e.* high-frequency content, than the ASTER LST. This difference is in part explained by the fact that NDVI and LST are different physical variables, and so, the NDVI and ASTER textures would not be identical. However, the low-pass filter used to provide ASTER LST at 250 m could also attenuate small scales. Unfortunately, these two effects are completely mixed and cannot be discriminated. These remarks need to be considered when using the test database for future studies.

VI. Conclusions and perspectives

In order to overcome the trade-off between temporal and spatial resolution of the TIR sensors, current stat-of-the-art methods such as Tsharp, ATPRK or DMS rely on a scale-invariance hypothesis to perform the super-resolution of LST images: models optimized at coarse scales are applied at small ones. However, this hypothesis can lead to an important performance loss. In this work, we proposed a Scale-Invariance-Free approach to train two Convolutional Neural Networks, SIF-CNN-SR1 and SIF-CNN-SR2, that increase the resolution of MODIS LST images from 1 km to 250 m. We also propose a CNN sharing the same architecture as SIF-CNN-SR but being trained under the scale-invariance hypothesis, SC-Unet. This model allows to compare the proposed scale-invariance-free training approach with a classical reduced scale one, described in section III-D. All the models studied in this work use the MODIS NDVI at 250 m to provide information at high spatial resolution.

To evaluate the performance of each super-resolution method, we have generated a validation dataset by pairing concomitant LST images from ASTER with couples of (NDVI-LST) images from MODIS. Both ASTER LST and MODIS NDVI are provided at 250 m of spatial resolution while MODIS LST is provided at 1 km of resolution. Consequently, we generated an analysis-ready-dataset of MODIS and ASTER concomitant images that can be used for benchmarking super-resolution methods.

We compared the three proposed deep learning models with the aforementioned state-of-the-art methods. The results show

[3]This database is available at https://github.com/cgranerob/Land-Surface-Temperature-Super-Resolution-with-a-Scale-Invariance-Free-Neural-Approach.

that SIF-CNN-SR1 outperforms the other methods. Also, we showed that RMSE and SSIM are partially adapted for the evaluation of super-resolution methods since they do not focus on the characterization of textures. Indeed, metrics focusing on textures such as LPIPS and Fourier-space metrics appear as more appropriated for this case of studies.

Future research directions include: considering nighttime LST images without any concomitant NDVI image or testing the generalization of the model for different geographical regions. Also, the choice of an adaptive parameter γ to capture the variability of the relationship between the NDVI and the LST could lead to better performances. Finally, the ability of the models to operate at higher spatial resolutions should be investigated as future thermal missions are planned such as LSTM, TRISHNA or SBG that will provide LST images at around 50 m. Indeed, the monitoring of the LST for a wide range of applications would benefit from time series of LST images at both high temporal and spatial resolutions.

REFERENCES

[1] S. Bojinski, M. Verstraete, T. C. Peterson, C. Richter, A. Simmons, and M. Zemp, "The Concept of Essential Climate Variables in Support of Climate Research, Applications, and Policy," *Bulletin of the American Meteorological Society*, vol. 95, no. 9, pp. 1431–1443, Sep. 2014. [Online]. Available: https://journals.ametsoc.org/doi/10.1175/BAMS-D-13-00047.1

[2] C. R. d. Almeida, A. C. Teodoro, and A. Gonalves, "Study of the Urban Heat Island (UHI) Using Remote Sensing Data/Techniques: A Systematic Review," *Environments*, vol. 8, no. 10, p. 105, Oct. 2021. [Online]. Available: https://www.mdpi.com/2076-3298/8/10/105

[3] T. P. Albright, A. M. Pidgeon, C. D. Rittenhouse, M. K. Clayton, C. H. Flather, P. D. Culbert, and V. C. Radeloff, "Heat waves measured with MODIS land surface temperature data predict changes in avian community structure," *Remote Sensing of Environment*, vol. 115, no. 1, pp. 245–254, Jan. 2011. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0034425710002671

[4] B. Dousset, F. Gourmelon, K. Laaidi, A. Zeghnoun, E. Giraudet, P. Bretin, E. Mauri, and S. Vandentorren, "Satellite monitoring of summer heat waves in the Paris metropolitan area," *International Journal of Climatology*, vol. 31, no. 2, pp. 313–323, Feb. 2011. [Online]. Available: https://rmets.onlinelibrary.wiley.com/doi/10.1002/joc.2222

[5] Q. Yang, Y. Xu, T. C. Chakraborty, M. Du, T. Hu, L. Zhang, Y. Liu, R. Yao, J. Yang, S. Chen, C. Xiao, R. Liu, M. Zhang, and R. Chen, "A global urban heat island intensity dataset: Generation, comparison, and analysis," *Remote Sensing of Environment*, vol. 312, p. 114343, 2024.

[6] Z. Wan, P. Wang, and X. Li, "Using MODIS Land Surface Temperature and Normalized Difference Vegetation Index products for monitoring drought in the southern Great Plains, USA," *International Journal of Remote Sensing*, vol. 25, no. 1, pp. 61–72, Jan. 2004. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/0143116031000115328

[7] K. I. N. Rahmi and M. Dimyati, "Remote sensing and GIS application for monitoring drought vulnerability in Indonesia: a review," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 6, pp. 3507–3518, Dec. 2021. [Online]. Available: https://beei.org/index.php/EEI/article/view/3249

[8] I. R. Orimoloye, J. A. Belle, and O. O. Ololade, "Drought disaster monitoring using MODIS derived index for drought years: a space-based information for ecosystems and environmental conservation," *Journal of Environmental Management*, vol. 284, p. 112028, 2021.

[9] S. Kumar and A. Kumar, "Hotspot and trend analysis of forest fires and its relation to climatic factors in the western Himalayas," *Natural Hazards*, vol. 114, no. 3, pp. 3529–3544, Dec. 2022. [Online]. Available: https://link.springer.com/10.1007/s11069-022-05530-5

[10] J. Zhao, C. Yue, J. Wang, S. Hantson, X. Wang, B. He, G. Li, L. Wang, H. Zhao, and S. Luyssaert, "Forest fire size amplifies postfire land surface warming," *Nature*, vol. 633, no. 8031, pp. 828–834, Sep. 2024. [Online]. Available: https://www.nature.com/articles/s41586-024-07918-8

[11] J. A. Sobrino, "An analysis of the Lake Surface Water Temperature evolution of the worlds largest lakes during the years 2003-2020 using MODIS data," *Recent Advances in Remote Sensing*, Feb. 2024. [Online]. Available: https://www.recentadvancesin.com/remote-sensing/3020-6448-rars240001/

[12] J. Liu, D. F. T. Hagan, and Y. Liu, "Global land surface temperature change (20032017) and its relationship with climate drivers: Airs, modis, and era5-land based analysis," *Remote Sensing*, vol. 13, no. 1, 2021. [Online]. Available: https://www.mdpi.com/2072-4292/13/1/44

[13] Y.-R. Wang, D. O. Hessen, B. H. Samset, and F. Stordal, "Evaluating global and regional land warming trends in the past decades with both MODIS and ERA5-Land land surface temperature data," *Remote Sensing of Environment*, vol. 280, p. 113181, Oct. 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0034425722002930

[14] A. M. Waring, D. Ghent, M. Perry, J. S. Anand, K. L. Veal, and J. Remedios, "Regional climate trend analyses for Aqua MODIS land surface temperatures," *International Journal of Remote Sensing*, vol. 44, no. 16, pp. 4989–5032, Aug. 2023. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/01431161.2023.2240522

[15] P. Guillevic, F. Gttsche, J. Hulley, G. Ghent, J. Romn, M. Camacho, P. Guillevic, F.-M. Gttsche, J. Nickeson, G. Hulley, D. Ghent, Y. Yu, I. Trigo, S. Hook, J. Sobrino, J. Remedios, M. Romn, and F. Camacho de Coca, "Land surface temperature product validation best practice protocol. version 1.1," 01 2018.

[16] M. Anderson, J. Norman, W. Kustas, R. Houborg, P. Starks, and N. Agam, "A thermal-based remote sensing technique for routine mapping of land-surface carbon, water and energy fluxes from field to regional scales," *Remote Sensing of Environment*, vol. 112, no. 12, pp. 4227–4241, 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425708002289

[17] M. C. Anderson, R. G. Allen, A. Morse, and W. P. Kustas, "Use of Landsat thermal imagery in monitoring evapotranspiration and managing water resources," *Remote Sensing of Environment*, vol. 122, pp. 50–65, Jul. 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425712000326

[18] S.-B. Duan, X.-J. Han, C. Huang, Z.-L. Li, H. Wu, Y. Qian, M. Gao, and P. Leng, "Land Surface Temperature Retrieval from Passive Microwave Satellite Observations: State-of-the-Art and Future Directions," *Remote Sensing*, vol. 12, no. 16, p. 2573, Jan. 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/16/2573

[19] Z. Li, H. Wu, S. Duan, W. Zhao, H. Ren, X. Liu, P. Leng, R. Tang, X. Ye, J. Zhu, Y. Sun, M. Si, M. Liu, J. Li, X. Zhang, G. Shang, B. Tang, G. Yan, and C. Zhou, "Satellite Remote Sensing of Global Land Surface Temperature: Definition, Methods, Products, and Applications," *Reviews of Geophysics*, vol. 61, no. 1, p. e2022RG000777, Mar. 2023. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022RG000777

[20] M. C. Anderson, Y. Yang, J. Xue, K. R. Knipper, Y. Yang, F. Gao, C. R. Hain, W. P. Kustas, K. Cawse-Nicholson, G. Hulley, J. B. Fisher, J. G. Alfieri, T. P. Meyers, J. Prueger, D. D. Baldocchi, and C. Rey-Sanchez, "Interoperability of ECOSTRESS and Landsat for mapping evapotranspiration time series at sub-field scales," *Remote Sensing of Environment*, vol. 252, p. 112189, Jan. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0034425720305629

[21] J. Yang, J. Ren, D. Sun, X. Xiao, J. C. Xia, C. Jin, and X. Li, "Understanding land surface temperature impact factors based on local climate zones," *Sustainable Cities and Society*, vol. 69, p. 102818, Jun. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2210670721001098

[22] T. N. Phan and M. Kappas, "Application of MODIS land surface temperature data: a systematic literature review and analysis," *Journal of Applied Remote Sensing*, vol. 12, no. 04, p. 1, Oct. 2018. [Online]. Available: https://www.spiedigitallibrary.org/journals/journal-of-applied-remote-sensing/volume-12/issue-04/041501/Application-of-MODIS-land-surface-temperature-data--a-systematic/10.1117/1.JRS.12.041501.full

[23] P. Reiners, J. Sobrino, and C. Kuenzer, "Satellite-Derived Land Surface Temperature Dynamics in the Context of Global ChangeA Review," *Remote Sensing*, vol. 15, no. 7, p. 1857, Mar. 2023. [Online]. Available: https://www.mdpi.com/2072-4292/15/7/1857

[24] J. A. Sobrino, F. Del Frate, M. Drusch, J. C. Jimenez-Munoz, P. Manunta, and A. Regan, "Review of Thermal Infrared Applications and Requirements for Future High-Resolution Sensors," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 5, pp. 2963–2972, May 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7378483/

[25] H. Liu and Q. Weng, "Enhancing temporal resolution of satellite imagery for public health studies: A case study of West Nile Virus outbreak in Los Angeles in 2007," *Remote Sensing of Environment*, vol. 117, pp. 57–71, Feb. 2012. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0034425711002835

[26] Z. Mitraka, S. Stagakis, G. Lantzanakis, N. Chrysoulakis, C. Feigenwinter, and S. Grimmond, "High spatial and temporal resolution Land Surface Temperature for surface energy fluxes estimation," in *2019 Joint Urban Remote Sensing Event (JURSE)*. Vannes, France: IEEE, May 2019, pp. 1–4. [Online]. Available: https://ieeexplore.ieee.org/document/8808951/

[27] M. S. Ramsey and I. T. W. Flynn, "The Spatial and Spectral Resolution of ASTER Infrared Image Data: A Paradigm Shift in Volcanological Remote Sensing," *Remote Sensing*, vol. 12, no. 4, p. 738, Jan. 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/4/738

[28] W. Li, L. Ni, H. Wu, Z. Li, and S. Duan, "Evaluation of machine learning algorithms in spatial downscaling of modis land surface temperature," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 7, pp. 2299–2307, 2019.

[29] C. Yoo, J. Im, S. Park, and D. Cho, "Spatial donwscaling of MODIS Land Surface Temperature: recent research trends, challenges and future directions," *Korean Journal of Remote Sensing*, vol. 36, no. 4, pp. 609–626, 2020.

[30] S. Favrichon, C. Prigent, and C. Jimnez, "A Method to Downscale Satellite Microwave Land-Surface Temperature," *Remote Sensing*, vol. 13, no. 7, p. 1325, Jan. 2021. [Online]. Available: https://www.mdpi.com/2072-4292/13/7/1325

[31] Y. Zhong, L. Meng, Z. Wei, J. Yang, W. Song, and M. Basir, "Retrieval of All-Weather 1 km Land Surface Temperature from Combined MODIS and AMSR2 Data over the Tibetan Plateau," *Remote Sensing*, vol. 13, no. 22, p. 4574, Jan. 2021. [Online]. Available: https://www.mdpi.com/2072-4292/13/22/4574

[32] Y. Li, D. Sun, X. Zhan, P. Houser, C. Yang, and J. J. Qu, "Downscaling Land Surface Temperature Derived from Microwave Observations with the Super-Resolution Reconstruction Method: A Case Study in the CONUS," *Remote Sensing*, vol. 16, no. 5, p. 739, Jan. 2024. [Online]. Available: https://www.mdpi.com/2072-4292/16/5/739

[33] Z. Cai, G. Han, and M. Chen, "Do water bodies play an important role in the relationship between urban form and land surface temperature?" *Sustainable Cities and Society*, vol. 39, pp. 487–498, 2018.

[34] H. Govil, S. Guha, A. Dey, and N. Gill, "Seasonal evaluation of downscaled land surface temperature: A case study in a humid tropical city," *Heliyon*, vol. 5, no. 6, p. e01923, 2019.

[35] D. Kumar and S. Shekhar, "Statistical analysis of land surface temperature and vegetation indexes relationship through thermal remote sensing," *Ecotoxicology and Environmental Safety*, vol. 121, pp. 39–44, 2015.

[36] L. S. Ferreira and D. H. S. Duarte, "Exploring the relationship between urban form, land surface temperature and vegetation indices in a subtropical megacity," *Urban Climate*, vol. 27, pp. 105–123, 2019.

[37] C. Granero-Belinchon, A. Michel, J. Lagouarde, J. Sobrino, and X. Briottet, "Multi-resolution study of thermal unmixing techniques over madrid urban area: Case study of trishna mission," *Remote Sensing*, vol. 11, no. 10, p. 1251, 2019.

[38] W. Essa, B. Verbeiren, J. van der Kwast, T. Van de Voorde, and O. Batelaan, "Evaluation of the distrad thermal sharpening methodology for urban areas," *International Journal of Applied Earth Observation and Geoinformation*, vol. 19, pp. 163–172, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0303243412001183

[39] W. Kustas, J. Norman, M. Anderson, and A. French, "Estimating subpixel surface temperatures and energy fluxes from the vegetation indexradiometric temperature relationship," *Remote Sensing of Environment*, vol. 85, no. 4, pp. 429–440, 2003.

[40] N. Agam, W. Kustas, M. Anderson, F. Li, and C. Neale, "A vegetation index based technique for spatial sharpening of thermal imagery," *Remote Sensing of Environment*, vol. 107, no. 4, pp. 545–558, 2007.

[41] Q. Wang, W. Shi, P. Atkinson, and Y. Zhao, "Downscaling modis images with area-to-point regression kriging," *Remote Sensing of Environment*, vol. 166, pp. 191–204, 2015.

[42] Q. Wang, W. Shi, and P. M. Atkinson, "Area-to-point regression kriging for pan-sharpening," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 151–165, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271616000496

[43] F. Gao, W. Kustas, and M. Anderson, "A data mining approach for sharpening thermal satellite imagery over land," *Remote Sensing, vol. 4, issue 11, pp. 3287-3319*, vol. 4, pp. 3287–3319, 10 2012.

[44] J. Xue, M. C. Anderson, F. Gao, C. Hain, L. Sun, Y. Yang, K. R. Knipper, W. P. Kustas, A. Torres-Rua, and M. Schull, "Sharpening ECOSTRESS and VIIRS land surface temperature using harmonized Landsat-Sentinel surface reflectances," *Remote Sensing of Environment*, vol. 251, p. 112055, Dec. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0034425720304259

[45] R. Guzinski and H. Nieto, "Evaluating the feasibility of using sentinel-2 and sentinel-3 satellites for high-resolution evapotranspiration estimations," *Remote Sensing of Environment*, vol. 221, pp. 157–172, 2019.

[46] 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 Sensing*, vol. 11, p. 2635, 2019.

[47] 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 Journal of Photogrammetry and Remote Sensing*, vol. 146, pp. 305–319, 2018.

[48] N. Brodu, "Super-resolving multiresolution images with band-independent geometry of multispectral pixels," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 8, pp. 4610–4617, 2017.

[49] Y. Xiao, Q. Yuan, K. Jiang, J. He, X.Jin, and L. Zhang, "EDiffSR: an efficient diffusion probabilistic model for remote sensing image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, p. 5601514, 2023.

[50] B. Nguyen, G. Tian, M. Vo, A. Michel, T. Corpetti, and C. Granero-Belinchon, "Convolutional neural network modelling for modis land surface temperature super-resolution," *2022 30th European Signal Processing Conference (EUSIPCO)*, pp. 1806–1810, 2022.

[51] J. Chen, L. Jia, J. Zhang, Y. Feng, X. Zhao, and R. Tao, "Super-resolution for Land Surface Temperature retrieval images via cross-scale diffusion model using reference images," *Remote Sensing*, vol. 16, no. 8, p. 1356, 2024.

[52] Y.-J. Choe and J.-H. Yom, "Downscaling MODIS Land Surface Temperature to Landsat scale using multi-layer perceptron," *Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography*, vol. 35, pp. 313–318, 2017.

[53] O. Merlin, B. Duchemin, O. Hagolle, F. Jacob, B. Coudert, G. Chehbouni, G. Dedieu, J. Garatuza, and Y. Kerr, "Disaggregation of MODIS surface temperature over an agricultural area using a time series of Formosat-2 images," *Remote Sensing of Environment*, vol. 114, pp. 2500–2512, 2010.

[54] D. M. Barker, W. Huang, Y. Guo, A. J. Bourgeois, and Q. N. Xiao, "A three-dimensional variational data assimilation system for MM5: implementation and initial results," *Monthly Weather Review*, vol. 132, pp. 897–914, 2004.

[55] B. Melinc and Z. Zaplotnik, "3d-Var data assimilation using a variational autoencoder," *Quarterly Journal of the Royal Meteorological Society*, vol. 150, no. 761, pp. 2273–2295, 2024.

[56] F. Palsson, J. R. Sveinsson, M. O. Ulfarsson, and J. A. Benediktsson, "Quantitative quality evaluation of pansharpened imagery: consistency versus synthesis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 3, pp. 1247–1259, 2016.

[57] Y. Xiong, F. Shao, X. Meng, Q. Jiang, W. Sun, R. Fu, and Y.-S. Ho, "A large-scale remote sensing database for subjective and objective quality assessment of pansharpened images," *Journal of Visual Communication and Image Representation*, vol. 73, p. 102947, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1047320320301760

[58] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," *Computer Vision – ECCV 2016*, pp. 694–711, 2016.

[59] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, 2015.

[60] Y. Chen, R. Xia, K. Yang, and K. Zou, "MICU: Image super-resolution via multi-level information compensation and U-net," *Expert Systems with Applications*, vol. 245, p. 123111, 2024.

[61] A. Kalluvila, "Super-resolution of brain MRI via U-net architecture," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 5, pp. 26–31, 2023.

[62] X. Hu, M. A. Naiel, A. Wong, M. Lamm, and P. Fieguth, "RUNet: a robust U-net architecture for image super-resolution," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, CA, USA, 2019, pp. 505–507.

[63] G. Hulley and S. Hook, "MODIS/Terra Land Surface Temperature/3-Band Emissivity 5-Min L2 1km V061," 2021. [Online]. Available: https://lpdaac.usgs.gov/products/mod21v061/

[64] E. Vermote and R. Wolfe, "MODIS/Terra Surface Reflectance Daily L2G Global 250m SIN Grid V061," 2021. [Online]. Available: https://lpdaac.usgs.gov/products/mod09gqv061/

[65] A. Gillespie, S. Rokugawa, T. Matsunaga, J. Cothern, S. Hook, and A. Kahle, "A temperature and emissivity separation algorithm for advanced spaceborne thermal emission and reflection radiometer (aster) images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 36, no. 4, pp. 1113–1126, 1998.

[66] NASA/METI/AIST/Japan Spacesystems And U.S./Japan ASTER Science Team, "ASTER Level 2 Surface Temperature Product," 2001. [Online]. Available: https://lpdaac.usgs.gov/products/ast_08v003/

[67] L. Lu and X. M. Zhou, "A four-parameter model for estimating diurnal temperature cycle from modis land surface temperature product," *Journal of Geophysical Research: Atmospheres*, vol. 126, no. 8, p. e2020JD033855, 2021, e2020JD033855 2020JD033855. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020JD033855

[68] Y. Goshin and D. Arkhipova, "Noise compensation in super-resolution problem using the Huber loss function," in *2021 International Conference on Information Technology and Nanotechnology (ITNT)*, Samara, Russian Federation, 2021, pp. 1–4.

[69] J. Quinlan, "Learning with continuous classes," *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence : Hobart, Tasmania, 16-18 November 1992*, pp. 343–348, 1992.

[70] S. Rama, J. Michel, V. Rivalland, A. Michel, and C. Granero-Belinchon, "Assessing the usefulness of Land Surface Temperature spatial disaggregation for water stress mapping in the frame of the preparation of the Trishna mission," in *Recent Advances In Quantitative Remote Sensing*, Torrent (Valencia), Spain, Sep. 2022.

[71] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[72] R. Zhang, P. Isola, A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

[73] J. Michel, E. Kalinicheva, and J. Inglada, "Revisiting remote sensing cross-sensor single image super-resolution: the overlooked impact of geometric and radiometric distortion," *HAL*, p. 04723225, 2024.

[74] K. He, X. Zhang, S. Ren, and J. Sun, "Very deep convolutional networks for large-scale image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[75] R. Delair, H. Carfantan, A. Michel, X. Briottet, and V. Lonjou, "Multi-resolution analysis of urban morphological and spectral data and their relationships with the land surface temperature," Conference Poster, September 2024, Recent Advances in Quantitative Remote Sensing (RAQRS) VII Conference, 23-27 September 2024, Torrent (Valencia), Spain.

[76] D. Sun and M. Kafatos, "Note on the ndvi-lst relationship and the use of temperature-related drought indices over north america," *Geophysical Research Letters*, vol. 34, no. 24, 2007. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2007GL031485

[77] S. Guha and H. Govil, "Seasonal impact on the relationship between land surface temperature and normalized difference vegetation index in an urban landscape," *Geocarto International*, vol. 37, 08 2020.

[78] G. C. Hulley, C. G. Hughes, and S. J. Hook, "Quantifying uncertainties in land surface temperature and emissivity retrievals from ASTER and MODIS thermal infrared data," *Journal of Geophysical Research*, vol. 117, p. D23113, 2012.

[79] Y. Liu, T. Hiyama, and Y. Yamaguchi, "Scaling of land surface temperature using satellite data: A case examination on aster and modis products over a heterogeneous terrain area," *Remote Sensing of Environment*, vol. 105, no. 2, pp. 115–128, 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425706002379

[80] Y. Liu, Y. Noumi, and Y. Yamaguchi, "Discrepancy between aster- and modis- derived land surface temperatures: Terrain effects," *Sensors*, vol. 9, no. 2, pp. 1054–1066, 2009.

[81] Y. Liu, Y. Yamaguchi, and C. Ke, "Reducing the discrepancy between aster and modis land surface temperature products," *Sensors*, vol. 7, no. 12, pp. 3043–3057, 2007.