Deep Learning for Remote Sensing Image Downscaling: Fusion and Super-Resolution

Maria Sdraka, Ioannis Papoutsis, *Member, IEEE*, Bill Psomas, Konstantinos Vlachos, Konstantinos Ioannidis, Konstantinos Karantzalos, *Member, IEEE*, Ilias Gialampoukidis, Stefanos Vrochidis, *Member, IEEE*

***Abstract*—The past few years have seen an accelerating integration of Deep Learning (DL) techniques into various Remote Sensing applications, highlighting their adaptability power and achieving unprecedented advancements. In the present review, we provide an exhaustive exploration of the DL approaches proposed specifically for the spatial downscaling of Remote Sensing imagery. A key contribution of our work is the presentation of the major architectural components and models, metrics and datasets available for this task, and the construction of a compact taxonomy for navigating through the various methods. Furthermore, we analyze the limitations of the current modeling approaches and provide a brief discussion on promising directions for image enhancement, following the paradigm of general Computer Vision practitioners and researchers as a source of inspiration and constructive insight.**

***Index Terms*—deep learning, remote sensing, downscaling, super-resolution, satellite, enhancement, spatial improvement, spatiotemporal, spatiospectral, pansharpening, image fusion**

I. Motivation

RECENT technological advances have significantly increased the volume and distribution rate of Remote Sensing (RS) data, reaching the level of tens of Terabytes on a daily basis. For that reason, such data have become a ubiquitous source of information for the monitoring of Earth's physical, chemical and biological systems, assisting atmospheric, geological and oceanic research, as well as hazard assessment and resource management applications, to name a few.

Satellite RS currently drives Earth Observation (EO) research and applications. There is a large number of operational satellites orbiting the Earth mounted with active and passive RS sensors, providing a continuous stream of information on various aspects of the planet's physical processes. Satellite imagery from these sensors is characterized by its spatial, spectral, temporal and radiometric resolution [1]. *Spatial resolution* (or *ground sample distance*) represents the size of a single satellite image pixel on the ground and corresponds to the level of spatial detail that can be acquired with this particular sensor. *Spectral resolution* represents the range of the electromagnetic (e/m) spectrum (wavebands) that the sensor acquires observations in, while *temporal resolution* (or *revisit time*) represents the time interval between two consecutive image acquisitions of the same location. Finally, *radiometric resolution* refers to the numerical precision or bit depth of a single pixel. Unfortunately, due to technical and financial constraints, there is usually a trade-off between these factors and no available sensor can capture information in the highest possible spatial and temporal resolution across all wavebands.

For that matter, one of the hottest topics in RS is the fusion of multi-source data with the aim to combine their strengths and enhance the resolution along the spatial, spectral or temporal dimension. In this particular study we focus on the spatial downscaling problem which can be greatly aided by the integration of Deep Learning (DL) methods and comprises an essential part in the pipeline of various RS research fields such as land use and land cover classification [2] [3], deforestation monitoring [4] [5], crop yield forecasting, precipitation forecasting [6], disaster monitoring [7] [8], stream flow monitoring [9], and many more.

Several review papers were published recently, which to a certain extent address the problem of image downscaling with deep neural networks. The present study aims to differ and ultimately add a methodological framework and a valuable condensation of the most recent literature on enhancing the spatial resolution of satellite imagery data specifically, using advanced Deep Learning architectures. These Deep Learning models are tailored to Earth Observation data with their unique and heterogeneous spatial, temporal and spectral characteristics, which differ significantly from the imagery traditionally used by the Computer Vision community. In fact, research on Computer Vision applications has motivated the production of valuable review papers, mainly for (non-satellite) image super-resolution, like [10], [11], [12], [13], [14], [15] and [16]. Our work targets exclusively the RS field and provides a broader overview of methods and applications than [17], [18] and [19] that focus solely on pansharpening approaches or [20] that only examines Single Image Super-resolution, non-DL methods. Additionally, a number of noteworthy studies [21]–[23] provide a thorough analysis of the use of DL techniques in RS but they are not limited to the spatial downscaling problem and address the entire spectrum of applications. Other review works ([1], [24]) focus on multi-modal data fusion state-of-the-art, partially addressing image resolution enhancement without focusing on Deep Learning techniques. Finally, a study similar

M. Sdraka and I. Papoutsis are with the Institute of Astronomy, Astrophysics, Space Applications & Remote Sensing, National Observatory of Athens, Greece.

B. Psomas and K. Karantzalos are with the School of Rural & Surveying Engineering, National Technical University of Athens, Greece.

K. Vlachos, K. Ioannidis, I. Gialampoukidis and S. Vrochidis are with the Information Technologies Institute, Center for Research & Technology Hellas.

This work has received funding from the European Union's Horizon2020 research and innovation projects i) DeepCube, under grant agreement No 101004188 (M. Sdraka and I. Papoutsis), ii) NEANIAS, under grant agreement No 863448 (B. Psomas and K. Karantzalos), and iii) CALLISTO, under grant agreement No 101004152 (K. Vlachos, K. Ioannidis, I. Gialampoukidis and S. Vrochidis).

to ours [25] reviews the literature up to mid 2019 therefore missing the most recent state-of-the-art approaches.

Indeed, the last three years have been productive for scientific works on image downscaling with DL. For example, while RS image super-resolution publications have been steadily increasing, the ratio of the studies that use DL has blown-up from 5% in 2017, to almost 40% in 2020 (Fig. 1). Similarly in Computer Vision, DL for image super-resolution publications [26] exhibit a steady increase.

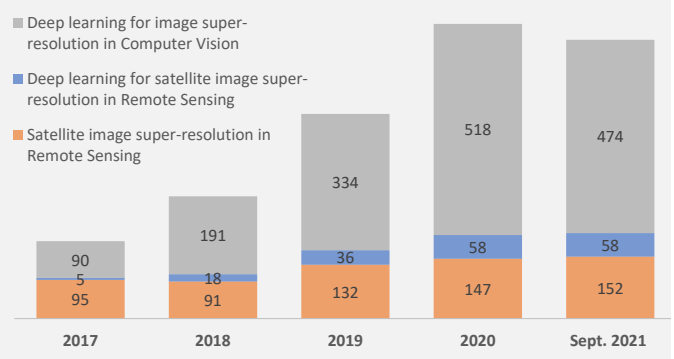


Fig. 1: Number of published papers related to image super-resolution for traditional and Deep Learning based techniques, for satellite Remote Sensing and computer vision fields [26].

In this review paper we are presenting the recent advancements (up to July 2021) of spatial downscaling on satellite imaging through Deep Learning approaches and analyse their strengths and shortcomings. We are only interested in the enhancement of surface reflection products and do not address geophysical variables, such as land surface temperature (LST), vegetation indices, etc. First, in Section III we formalise the problem definition, while in Section IV we provide a list of relevant metrics that have been proposed for evaluating and benchmarking different DL models. Section V condenses the building blocks, frameworks and key models used extensively in literature to solve the downscaling problem given the selected metrics, and is a key contribution of our work. These components are the "bricks" used by researchers as stand-alone or in combination depending on the nature of the downscaling problem. In Section VI we provide an overall taxonomy for the different families of RS downscaling methods, and we analyse these methods in Sections VII, VIII & IX. We extend our review in Section IX-D with some special interest downscaling cases tailored to Synthetic Aperture Radar (SAR) and Unmanned Aerial Vehicle (UAV) data, while we exclude from our analysis 3D point cloud data obtained by LiDAR sensors. Section X lists the publicly available datasets targeting this problem. Finally, we provide an outlook for promising state-of-the-art methods researched in general Computer Vision field, but not yet applied in RS, where the interested reader could seek further ideas and approaches (Section XI).

Terminology

Before moving forward we need to clarify which terminology is used in this paper as far as spatial resolution increase/decrease is concerned. In climate and meteorological (e.g. [27]), as well as Remote Sensing [28] studies the term "downscale" refers to the transition from low to high resolution, i.e. less to more detail representation. However, in the Computer Vision field it is the term "upscale" that refers to the increase of (spatial) resolution and "downscale" to the decrease of it (e.g. [29]) and are synonymous to upsample and downsample, respectively. Indicatively, Zhan et al. [30] conducted a research on land surface temperature downscaling terminology, among others, and found that terms such as "enhancement", "sharpening", "fusion", "super-resolution", "unmixing", "subpixel" and "disaggregation" are also relevant to spatial resolution increase. In this paper we use the term "downscale".

II. Deep Learning for Remote Sensing

The governing principle of DL is the construction of artificial neural networks with a large number of layers (indicated by the adjective "deep" in the term) which mostly comprise convolutional, pooling and fully connected units. Although several architectures with these building blocks have been proposed, some of which have been carefully handcrafted for a specific task, the main idea is the construction of a hierarchy of features extracted from raw input data. This hierarchy is computed through representation learning approaches that can be supervised, semi-supervised or unsupervised. Overall, the strongest advantage of DL is its ability to process raw data, thus mitigating the need for manual feature extraction, and unravel complex non-linear dependencies in the input.

One critical factor for the success of any DL method is the existence of a big and diverse dataset to train on. The abundance and availability of data in EO provide therefore a fertile ground for the application of advanced Machine Learning algorithms and notable progress has been made over the last decade ([21]–[23]). For example, a number of works which exploit deeper architectures have recently been published and achieve impressive results in problems such as land use and land cover classification [31], scene classification [32], object detection [33], image fusion [1] and image registration [34], [35], highlighting the great potential of DL in RS applications and research.

However EO poses a unique challenge for DL since it involves the manipulation of multimodal and multitemporal data. Remote sensors acquire information from multiple segments of the electromagnetic spectrum, differentiating themselves from typical computer vision data which lie in the RGB range mostly. In addition, time is a quite important variable in EO applications. When studying dynamic systems, information is captured at regular time intervals and successive observations must be assessed and compared. Finally, RS images often suffer from information loss either due to hardware failure or atmospheric conditions that are difficult to penetrate by certain sensor types (e.g. cloud coverage, haze, etc. are a common obstacle for optical sensors). Therefore, any researcher willing to design and implement novel DL algorithms for EO must take all above points into consideration.

III. Problem Definition

Given a set of *n* low resolution (LR) images $(x_1, x_2, ..., x_n)$, where $x_i \in X^{H \times W}$, and their corresponding high resolution (HR) images $(y_1, y_2, ..., y_n)$, where $y_i \in Y^{kH \times kW}$, the goal is to estimate a downscaling function: $f : X \rightarrow Y$. Note that *H* is the image height, *W* is the image width and *k* is the scaling factor. This survey presents the approaches that have been proposed for the estimation of this non-linear downscaling function f through deep neural networks.

Imaging Model

The process of obtaining the LR x image from its HR y equivalent is commonly represented in the literature by the imaging model:

$$x = (y \otimes b) \downarrow_k + n \tag{1}$$

where $\otimes b$ is the convolution with a blurring kernel b, $\downarrow_k$ is the downsampling operation by a scaling factor k and n is a noise term. This formula is a simple model of the image degradation taking place during the capture of the scene and attempts to simulate the physics inside the imaging sensor. Some researchers have proposed modifications of this model which account for parameters like motion blur, quantization error of the compression process, zooming effects, exposure time, white balancing, etc. For a thorough investigation of the imaging model and its many extensions, please refer to [14].

Wald's Protocol

Due to the lack of paired LR-HR images in most cases, an alternative approach described by the *Wald's protocol* [36] is employed. This protocol assumes that the performance of data fusion models is independent of the scale, provided that certain conditions hold. In their seminal work, Wald et al. suggest first degrading the input image according to a factor *k*, thus creating LR-HR image pairs, and proceed to design a model tasked to downscale it to the original resolution. Then the developed method can be transferred to downscale the original image into one of much higher resolution according to the same downscaling factor *k*. Effectively, this is a self-supervised modeling approach.

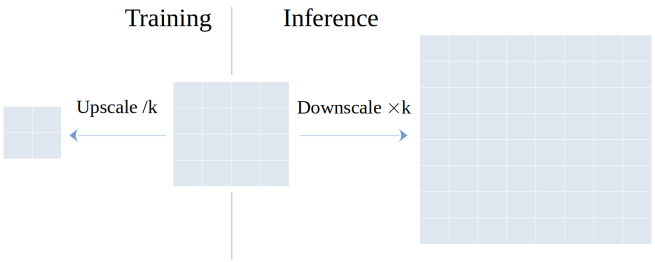


Fig. 2: Illustration of the the Wald's Protocol pipeline. The original image (middle one) is upscaled by a $/k$ factor and the resulting pair is used for model training. The trained model is then transferred to downscale the original image by a $\times k$ factor.

Note that throughout this document we will refer to the low-resolution images as C (coarse) and the high-resolution images as F (fine) respectively.

IV. Metrics

Several quality metrics have been proposed in order to assess the output of image restoration algorithms. Depending on the availability of a reference HR image these metrics can be divided into three broad categories [37]: (i) *full-reference*, where a complete HR reference image is required for comparison with the reconstructed image, (ii) *no-reference*, where only the reconstructed image is required, and (iii) *reduced-reference*, where only a set of features extracted from a HR image are available and used for comparison. Table I presents some of the most popular quality metrics found in literature for the task of spatial enhancement.

Perception-Distortion Trade-off

Full-reference metrics are also referred to as *distortion* metrics and typically measure the similarity/dissimilarity between the reconstructed image and the corresponding HR image. The goal of such metrics is to assess the reconstruction algorithm's ability to respect the structure and semantic content of the target image and can be generally formulated as

$$\Delta(I_{HR}, \hat{I}_{HR}) \tag{2}$$

where Δ is a similarity metric, I_{HR} is the HR image and $\hat{I}_{HR}$ the reconstructed one.

Accordingly, no-reference metrics are also known as *perceptual quality* metrics and they aim to quantify the "natural look" of a reconstructed image, i.e. how close it looks to a valid natural image, regardless of its similarity to the corresponding I_{HR}. Such metrics tend to approximate the perceptual quality of the human visual system (HVS) and can be formulated as

$$d(p_{I_{HR}}, p_{\hat{I}_{HR}}) \tag{3}$$

where d is a distribution similarity metric, $p_{I_{HR}}$ is the distribution of the natural HR images and $p_{\hat{I}_{HR}}$ the distribution of the reconstructed images.

Reduced-reference metrics provide an intermediate approach to full- and no-reference metrics, and can be either regarded as distortion or perceptual depending on the extracted features. Such metrics are primarily used for quality of service (QoS) monitoring of image/video broadcasting systems, where only a selected number of features are transmitted along with the compressed image in order to assess the transmission quality. In the image enhancement domain, no such metrics have been noted to be in wide use.

It has been empirically observed and then mathematically proven [57] that distortion and perceptual quality metrics act in a complementary, yet competitive manner. The *Perception-Distortion Trade-off* theorem dictates that as the distortion error of an algorithm decreases, the visual quality must also decrease, and vice versa. In practice, pursuing low distortion rate results in more blurry and over-smoothed images because the produced output approximates the statistical average of

Metric	Range	Description	Category
Mean Squared Error (MSE)	$[0, \infty)$	Pixel-based mean squared error	FR
Root Mean Squared Error (RMSE)	$[0, \infty)$	Pixel-based root mean squared error	FR
Mean Absolute Error (MAE)	$[0, \infty)$	Pixel-based mean absolute error	FR
Correlation Coefficient (CC)	[-1, 1]	Pixel-based correlation	FR
Coefficient of Determination (R^2)	[0, 1]	Per-pixel proportion of total variation	FR
Signal to Reconstruction Error Ratio (SRE)	[0, ∞)	Error relative to the mean image intensity	FR
Peak Signal-to-Noise Ratio (PSNR)	$(-\infty, \infty)$	Peak SNR based on MSE and expressed in dB	FR
Weighted Peak Signal-to-Noise Ratio (WPSNR) [38]	$(-\infty, \infty)$	Weighted PSNR to evaluate differently specific regions of the image	FR
Universal Image Quality Index (UIQI or UQI) [39]	[-1, 1]	Local differences in correlation, luminance and contrast	FR
Structural Similarity Index (SSIM) [37]	[-1, 1]	Based on UQI and measures local differences in luminance, contrast and structure	FR
Multi-Scale Structural Similarity Index (MS-SSIM) [40]	[-1, 1]	Combination of SSIM at various scales	FR
Information Fidelity Criterion (IFC) [41]	$[0, \infty)$	Utilizes Natural Scene Statistics defined as Gaussian Scale Mixtures in the wavelet domain	FR
Visual Information Fidelity (VIF) [42]	$[0, \infty)$	Extension of IFC by normalizing over reference image content	FR
Noise Quality Measure (NQM) [43]	$(-\infty, \infty)$	SNR based on contrast pyramid variations	FR
Feature Similarity Index (FSIM) [44]	[0, 1]	Similar to SSIM utilizing phase congruency and gradient magnitude	FR
Gradient Similarity Measure (GSM) [45]	[0, 1]	Similar to SSIM, measures gradient similarity	FR
Spectral Angle Mapper (SAM) [46]	$[0, \pi]$	Compares the angle between the two spectra	FR
Erreur Relative Globalle Adimensionelle de Synthese (ERGAS) [47]	$[0, \infty)$	Mean of the normalized average error of each band	FR
Most Apparent Distortion (MAD) [48]	$[0, \infty)$	Weighted geometric mean of the local error in the luminance domain and the subband local statistics	FR
VGG loss [49]	$[0, \infty)$	MSE between feature maps extracted from intermediate layers of a VGG network for both prediction and target images	FR
Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [50]	$[0, \infty)$	Support Vector Regression Model trained on Natural Scene Statistics of locally normalized luminance coefficients accompanied with Differential Mean Opinion Scores (for different distortions)	NR
Natural Image Quality Evaluator (NIQE) [51]	$[0, \infty)$	Multivariate Gaussian model trained on Natural Scene Statistics similar to BRISQUE (but for non-distorted images only)	NR
Perception based Image Quality Evaluator (PIQE) [52]	[0, 1]	Natural Scene Statistics similar to BRISQUE extracted from blocks of the distorted image and then pooled based on variance	NR
Q_{MA} [53]	$[0, \infty)$	Linear regression on the outputs of three independent regression forests trained on extracted features of local frequency, global frequency and spatial discontinuity along with the corresponding perceptual scores	NR
Perception Index (PI) [54]	$[0, \infty)$	Linear combination of Q_{MA} and NIQE	NR
Learned Perceptual Image Patch Similarity (LPIPS) [55]	$[0, \infty)$	L2 norm and averaging between features extracted from machine learning models on supervised, self-supervised or unsupervised settings	NR
Quality with No Reference (QNR) [56]	[0, 1]	One's complements of two spectral and spatial distortion indices based on band correlation, each raised to a real-valued exponent	NR

TABLE I: Most popular metrics for image quality assessment. FR = Full Reference, NR = No Reference.

possible HR solutions to this one-to-many problem, whereas a sharper, more naturally-looking result is usually not consistent with the initial LR image. It has also been proven that there is an unattainable region in the Perception-Distortion plane whose boundary is monotonic. This means that any reconstruction method can never achieve both a low distortion error and a high perceptual quality at the same time, but attempts are made to design an algorithm as close to the boundary as possible. Figure 3 illustrates the Perception-Distortion plane and the aforementioned boundary.

An interesting conclusion of [57] is that the method that converges closer to the Perception-Distortion bound is the Generative Adversarial Network (GAN) [58]. They show that such models are usually trained to minimize a weighted sum of a distortion and a perceptual quality metric, by modifying the loss function of the Generator as:

$$l_G = \mathbb{E}[\Delta(I_{HR}, \hat{I}_{HR})] + \lambda d(p_{I_{HR}}, p_{\hat{I}_{HR}}) \tag{4}$$

where λ is the weight of the perception quality factor and $d(p_{I_{HR}}, p_{\hat{I}_{HR}})$ is usually approximated by the standard adversarial loss. Therefore, GANs are usually able to produce images of a low distortion error and with the highest perceptual quality possible for this distortion error.

V. Standard DL methods for Downscaling in Computer Vision

Resolution enhancement has been thoroughly investigated in the field of general Computer Vision over the past decades.

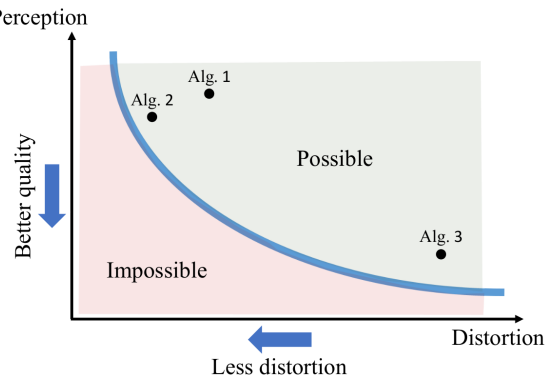


Fig. 3: The Perception-Distortion plane and the monotonic boundary separating the unattainable region. Image taken from [57] (©2018 IEEE).

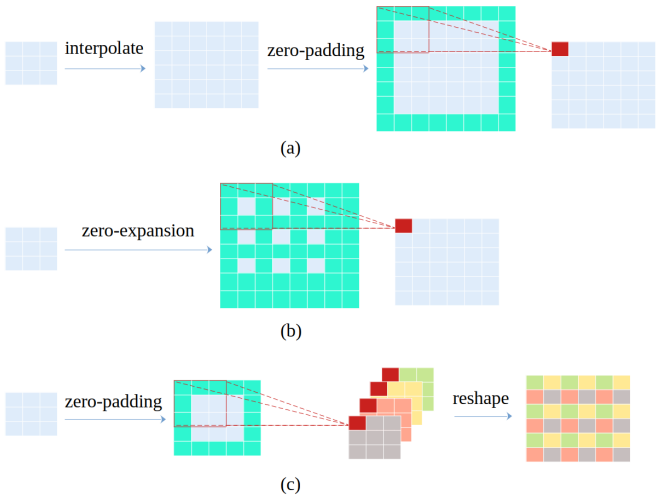


Fig. 4: Illustrated example of the three basic convolution schemes for upsampling a single-channel 3×3 feature map by a $\times 2$ factor. The red dashed lines refer to a simple 3×3 convolution. (a) Resize convolution, (b) Transposed convolution, (c) Sub-pixel convolution.

Certain methods and algorithms have been established and often serve as the basis of further investigation and improvements when developing novel approaches for RS downscaling. We present these methods in this Section and then use them throughout our paper as core modules.

A. Building blocks

In this paragraph we will briefly present some of the most fundamental building blocks of downscaling DL architectures.

Upsampling layers

- **Resize convolution**
  One of the first techniques proposed for feature downscaling. This operation involves upsampling the input by a traditional interpolation method, such as Nearest-Neighbour, bilinear or bicubic interpolation, and then performing a convolution on the result (Fig. 4(a)). Although it is a simple approach, it has been successfully applied to a number of studies in the field of CV.

- **Transposed convolution**
  This layer is also called *deconvolutional layer* [59], which is a quite inaccurate term since deconvolution in Computer Vision aims to revert the operation of a normal convolution and is rarely used in DL. Conversely, transposed convolution aims to produce a feature map of higher dimensions by first expanding the input with zero insertions and then performing a convolution (Fig. 4(b)). The transposed convolutional layer is widely used in downscaling architectures but caution is required since it is quite susceptible to producing checkerboard artifacts affecting the overall quality of the output [60].

- **Sub-pixel convolution**
  Also called *pixel-shuffle* [61], this layer comprises a convolution operation followed by a specific image reshape which rearranges the input features of shape $H \times W \times Cr^2$ to $rH \times rW \times C$ (Fig. 4(c)). This layer achieves a larger receptive field than transposed convolution and causes less artifacts in the final output [62].

Residual learning

The aim of downscaling is to learn a mapping between one (or multiple) LR image(s) and an HR image. This formulates an image-to-image translation task where the input (LR) is highly correlated with the output (HR) regardless of the scaling factor. In order to simplify this task and avoid learning such a complex translation, several studies employ global residual learning architectures [63] which focus on learning solely the residual, or difference, between input and output. Provided that a considerable part of the image remains basically unchanged, such a model is tasked to retrieve only the high-frequency details needed for the reconstruction of the HR counterpart so it generally converges faster and avoids bad minima.

In addition to global residual learning, local residual learning connections [64] are also commonly employed in downscaling architectures in order to alleviate vanishing gradients as the model gets deeper and more complex. Local residual learning shortcuts are inserted between intermediate layers while a global residual learning connection is used between input and output.

Laplacian Pyramid structure

First proposed in [65], the Laplacian Pyramid Structure is a feature extractor based on the Gaussian pyramid structure, which operates simultaneously at different scales and exploits the image difference (residuals) between levels. Applied on a DL setting, an input LR image is progressively upsampled s times through convolutional and upsampling layers, and the residual of each consecutive pair of upsampled outputs is computed. This results in the production of s residual images at different scales which contain features at different

levels of abstraction. Such structures have been extensively used in image downscaling since they split the problem into smaller manageable tasks of smaller scale and help the model converge to better optima.

Attention mechanism

Through the attention mechanism, the underlying neural network manages to isolate and focus on the most important feature details for the task at hand. Multiple types of attention mechanisms have been proposed over the years and can be categorized based on the dimension on which they operate. For example, channel attention considers the interdependence of the feature maps between channels and attributes a different weight on each one, while spatial attention emphasizes interesting regions in the spatial domain. Popular implementations of the channel attention mechanism include the *Squeeze-and-Excitation (SE) block* [66] and the *Efficient Channel Attention (ECA)* [67], while a spatial attention mechanism commonly used in practice is the *Coordinate Attention Module (CAM)* [68]. Several studies also use a combination of channel and spatial attention, such as the *Bottleneck Attention Module (BAM)* [69], the *Convolutional Block Attention Module (CBAM)* [70] and the *Triplet Attention* [71]. An interesting overview of the attention mechanisms used in downscaling architectures is presented in [72].

B. Upsampling frameworks

Although different DL architectures can vary greatly, four basic downscaling frameworks that describe all approaches present in the literature can be discerned. These frameworks are outlined in Fig. 5 and represent the possible ways to design a downscaling DL model with convolutional and upsampling/downsampling layers as basic components.

Pre-upsampling framework

This is the first framework explored in the literature for image downscaling via DL approaches. In its most common form, a traditional upsampling algorithm, e.g. bicubic interpolation, is utilized in order to upsample the image to the required scale. Then a CNN model is applied which refines the upsampled image and produces the HR result. Such an approach provides a simpler learning pipeline since the network is relieved of the burden to properly upsample the image and is only tasked to sharpen and cleanse the input. Another advantage of the pre-upsampling framework is the ability to handle images of arbitrary size and scale. On the other hand, the computational cost is increased since all operations are performed in a higher dimensional space while the preceding upsampling procedure often amplifies noise and significantly increases blurring.

Post-upsampling framework

Mitigating the complexity and high cost of the pre-upsampling approach, in the post-upsampling framework an end-to-end model undertakes the upsampling task via trainable layers located at the end of the architecture. In the most common approach, a DL network performs feature extraction on the low dimensional space of the LR image and finally increases the resolution to obtain the HR output. A disadvantage of this framework is the fixed scaling factor which forms an integral part of the architecture, thus a different model must be designed and trained for different scales. In addition, performance is highly affected by the magnitude of the scaling factor. Since upsampling is performed in a single step, high factors (e.g. $\times 8$, $\times 10$) increase the learning difficulty and make the models considerably harder to train.

Progressive upsampling framework

In this framework, a model upsamples the image in a progressive manner through consecutive convolutional and upsampling layers. At each stage the input is upsampled to a higher resolution, finally obtaining the required scale at the output. This approach facilitates the learning process since the downscaling task is decomposed into much simpler steps. Such architectures are also able to handle requirements for multiscale output since each stage produces an upsampled image of intermediate scale. However, progressive upsampling models require more complex architectures and are thus harder to design and train.

Iterative up- and down-sampling framework

This framework exploits consecutive up- and down-sampling layers which refine the reconstruction error on HR to LR projections thus extracting more information on the relationship and correlations between the two spaces. Such models usually achieve higher quality results and are able to handle higher scaling factors successfully.

C. Models

One of the first robust DL methods for downscaling was presented in [73] (*SRCNN*) where a two-layer CNN was fed an upsampled version of an image and produced a sharpened HR output. It was trained and tested on subsets of ImageNet and outperformed equivalent non-DL methods. A similar approach was adopted by Kim et al. [74] (*VDSR*) who designed a deeper, VGG-like architecture [75] with a global residual connection and managed to outperform SRCNN on the test set.

Shi et al. [61], [62] (*ESPCN*) subsequently introduced the *Sub-Pixel Convolution*, which later became a popular upsampling technique for DL models. This trick helps reduce the model's number of parameters without compromising its representational power.

The next landmark paper [76] (*LapSRN*) introduced a multiscale architecture which integrates the Laplacian Pyramid structure and produces intermediate images downscaled by smaller factors ($\times 2$, $\times 4$, $\times 8$) in a single pass. The intermediate outputs are supervised via separate Charbonnier loss functions and this progressive upsampling scheme helps the model retain high accuracy in higher scales.

Ledig et al. [77] (*SRGAN*) introduced an adversarial approach to spatially enhance natural images. The Generator,

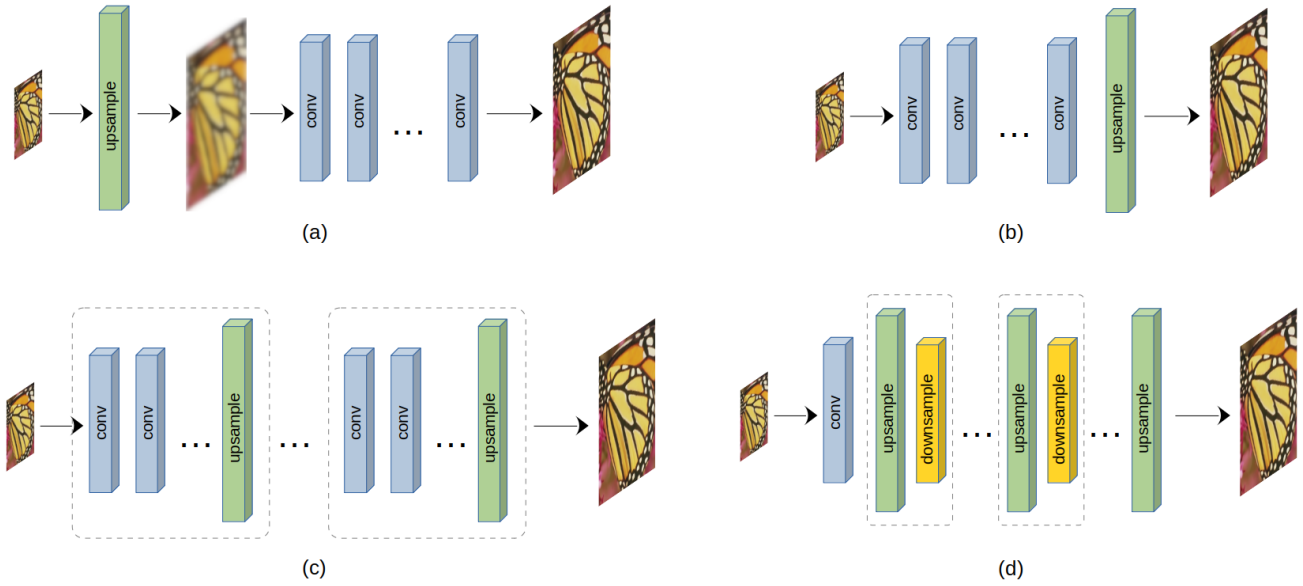


Fig. 5: The possible downscaling frameworks present in the DL literature. (a) Pre-upsampling, (b) Post-upsampling, (c) Progressive upsampling, (d) Iterative up- and down-sampling. *conv* represents a convolutional layer, *upsample* an upsampling layer and *downsample* a downsampling layer, all of which are trainable. Layers enclosed by dashed boxes denote stackable blocks.

named *SRResNet*, consists of a series of residual blocks, local and global residual connections and sub-pixel convolutional layers for downscaling. The Discriminator is a VGG-like network which performs the real/fake binary classification. The Generator's loss function is a combination of the adversarial loss and a term comparing the produced downscaled and the target HR image. Based on this model, Wang et al. [78] (*ESRGAN*) propose a number of improvements to achieve sharper results. They replace the residual blocks with novel *residual-in-residual dense blocks* which actually comprise of dense blocks with global residual connections, as seen in Fig. 6 and use the Relativistic average Discriminator introduced in [79].

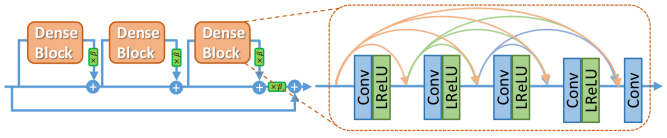


Fig. 6: Illustration of the residual-in-residual block (RIRB). It contains multiple dense blocks and residual connections both between blocks and between the input and output of the RIRB. β here refers to the residual scaling parameter. Image taken from [80].

Following the success of the baseline SRGAN, Lim et al. [81] (*EDSR/MDSR*) extend the SRResNet architecture by removing the ReLU activations outside the residual blocks and deepening the model. The authors name this architecture EDSR and train it separately for the scaling factors $\times 2$, $\times 3$ and $\times 4$. They also noted that by fine-tuning a pretrained $\times 2$ model when training for $\times 3$ or $\times 4$ downscaling, the entire training process is accelerated and the algorithm converges much faster. Based on this observation, the authors argue that downscaling at multiple scales are inter-related tasks, so they design an alternative model, namely MDSR, which handles multiple scales simultaneously. Subsequently, Yu et al. [82] (*WDSR*) introduce two novel residual blocks to the EDSR architecture. These blocks employ a *wide activation* approach by constricting the features of the identity mapping pathway and widening the features before activation.

Another robust technique has been proposed in [83] (*RDN*). The authors present a *Residual Dense Block* (RDB) which comprises a dense block with three novelties: (i) *Contiguous memory* (CM) where the output of an RDB is fed to each layer of the next RDB, (ii) *Local feature fusion* (LFF) which is a concatenation and a 1×1 convolution layer at the end of an RDB which adaptively controls the output information making the network easier to train, and finally (iii) *Local residual learning* (LRL) which is a residual connection between the input and output of the RDB. Utilizing a sequence of such RDB blocks and sub-pixel upsampling layers, the final RDN architecture is formed and then trained with the MAE loss function.

A number of methods, such as [85] (*DBPN and D-DBPN*) and [86] (*SRFBN*), opt for an iterative up- and downsampling strategy in the main core of their model. Particularly, several consecutive layers alternatively perform up- and down-projection operations learning different types of image degradation, which then contribute to the construction of the final HR image. This procedure provides an error feedback mechanism for projection errors at each stage and manages to extract better representations of the various features.

Some methods ([87] (*DRCN*), [84] (*DRRN*)) propose the use of recursive structures inside the model. Arguing that the addition of more layers make a network inefficient and

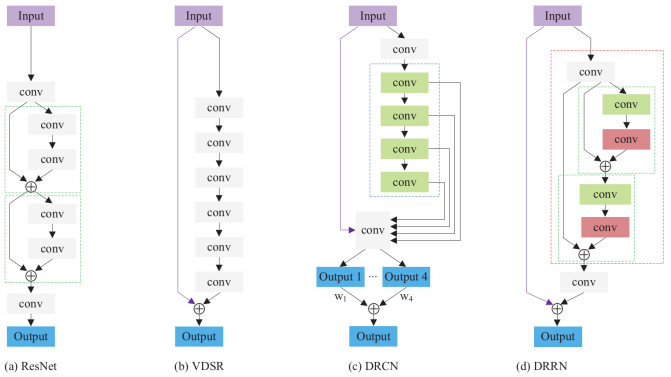


Fig. 7: Overview of the classic ResNet, VDSR, DRCN and DRRN architectures. Global residual connections are marked by a purple line, $\oplus$ refers to element-wise addition, and outputs with blue color are supervised. (a) ResNet: The green dashed box signifies a residual block. (b) VDSR. (c) DRCN: The blue dashed box refers to a recursive layer whose convolutional layers are marked in green and share the same weights. (d) DRRN: The red dashed box refers to a recursive block and the green dashed box marks the residual units. The corresponding convolutional layers marked in green and red share the same weights. Image taken from [84] (©2017 IEEE).

more likely to overfit, the aforementioned studies introduce recursive convolutional layers which apply the same convolution multiple times. Therefore weights are shared between consecutive convolutional operations and more stable convergence is achieved. Fig. 7 displays the structural differences between DRCN and DRRN for better understanding. A similar extension is also proposed for the LapSRN model in [88] (*MS-LapSRN*). In particular, the network parameters across pyramid levels are shared since they perform a similar task via a similar structure, and the feature embedding sub-network of each pyramid level is replaced by a series of recursive convolutional layers in order to increase the robustness of the model without increasing the number of parameters accordingly.

Finally, Zhang et al. [89] (*RCAN*) propose a channel attention module which consists of a Global Average Pooling layer and a gating mechanism that adds attention to the pooled features and enables the model to focus on the informative feature maps. Multiple such attention modules are incorporated inside Residual-in-Residual blocks and the final downscaling is performed by Sub-Pixel convolutions. When combined with a self-ensembling strategy, RCAN outperforms several robust DL methods.

Table II summarises the most popular models in CV for image downscaling via DL, vis-à-vis the building blocks employed, the upsampling framework adopted, whether a GAN pipeline is used or not, and the number of the model parameters. The latter attribute is useful to assess the complexity of each model, therefore weigh its proneness to overfit given the training data available.

VI. Downscaling Taxonomy in RS

Based on the dimensions and modalities to be combined, a variety of downscaling schemes have been proposed in the context of EO. Fig. 8 provides a simple, yet complete taxonomy of the methodological approaches used in literature according to our review.

Given this taxonomy, one can discern three fundamental groups of satellite image downscaling approaches for RS, depending on whether spectral, temporal on no external information is used:

1) *Spatiospectral Fusion*: images of different spatial and spectral resolutions are fused in order to produce an image of the highest possible spatial resolution in the coarser bands.
2) *Spatiotemporal Fusion*: images of high spatial but low temporal resolution are fused with images of low spatial but high temporal resolution in order to produce images of the highest resolution in both dimensions.
3) *Super-resolution*: A single or multiple images are downscaled without any additional external information.

In more detail, when the downscaling process is assisted with information on different spectra, then Spatiospectral Fusion techniques are used. These techniques are further discriminated based on the type of input spectra at hand, resulting in multispectral fusion (two multispectral images with different spectral information), pansharpening (a multispectral image and a panchromatic image) and multispectral/hyperspectral fusion (a multispectral and a hyperspectral image). In contrast, when the same spectra are available at different time steps and different spatial resolutions then Spatiotemporal Fusion methods come into play where temporal differences are additionally exploited for the spatial downscaling. This family of methods includes two sub-families depending on the time points of the input data. Finally, when no external information is available and downscaling can only be performed directly on the initial LR data, then Super-resolution techniques can be employed. There are three method sub-families depending on the number of input images and whether additional features extracted from the same LR data are used as auxiliary input.

Fig. 9 and 10 present an overview of the aforementioned method families highlighting graphically the different approaches, whereas Fig. 11, 12 and 13 show downscaling examples of each family. In the following sections we base our review on this discrimination and provide a detailed examination of the approaches shaping each method family.

VII. Spatiospectral Fusion

Satellites are equipped with various different sensors which operate in different parts of the electromagnetic spectrum and capture information on different features of the scanned location. These features can have variable spatial resolution, thus an advanced method called Spatiospectral Fusion (SSF) is usually employed to elaborately blend the fine spatial resolution of a band B_{HR} into the coarser spatial resolution of a target band B_{LR} and obtain a new image in the target band of much higher quality.

We discern three families of SSF: multispectral image fusion, pansharpening, and hyperspectral image downscaling. These are presented next, while in Table III we summarise the main DL models developed for SSF.

Model	Building blocks used	Upsampling framework	GAN	# Parameters
SRCNN [73] [1]	simple CNN	pre-upsampling	No	57k
VDSR [74]	VGG-based, residual connections	pre-upsampling	No	665k
ESPCN [61]	simple CNN, sub-pixel convolution	post-upsampling	No	20k
LapSRN [76] [2]	Laplacian pyramid structure	progressive upsampling	No	821k
SRGAN [77]	sub-pixel convolution, residual connections	post-upsampling	Yes	Generator: 734k Discriminator: 5.2m
ESRGAN [78] [3]	sub-pixel convolution, residual-in-residual blocks	post-upsampling	Yes	Generator: 16.7m Discriminator: 14.5m
EDSR [81] [4]	sub-pixel convolution, residual connections, pretraining	post-upsampling	No	43m
MDSR [81] [4]	multiscale EDSR	post-upsampling	No	8m
WDSR [82] [5]	EDSR with wide activation modules	post-upsampling	No	small model: 1.2m big model: 37.9m
RDN [83] [6]	residual dense blocks, local residual connections, sub-pixel convolution	post-upsampling	No	22.3m
DBPN [85] [7]	residual connections, transposed convolution	iterative up- and down-sampling	No	188k - 2.2m
D-DBPN [85] [7]	residual connections, transposed convolution	iterative up- and down-sampling	No	10.3m
SRFBN [86] [8]	residual connections, transposed convolution, recurrent layers	iterative up- and down-sampling	No	3.6m
DRCN [87]	recursive convolutions, residual connections	pre-upsampling	No	1.8m
DRRN [84] [9]	DRCN with recursive blocks and added local residual connections	pre-upsampling	No	297k
MS-LapSRN [88] [2]	LapSRN with shared weights and recursive blocks	progressive upsampling	No	222k
RCAN [89] [10]	channel attention, sub-pixel convolution, residual-in-residual blocks, residual connection	post-upsampling	No	16m

[1] http://mmlab.ie.cuhk.edu.hk/projects/SRCNN.html
[2] https://github.com/phoenix104104/LapSRN
[3] https://github.com/xinntao/ESRGAN
[4] https://github.com/LimBee/NTIRE2017
[5] https://github.com/JiahuiYu/wdsr_ntire2018
[6] https://github.com/yulunzhang/RDN
[7] https://www.toyota-ti.ac.jp/Lab/Denshi/iim/members/muhammad.haris/projects/DBPN.html
[8] https://github.com/Paper99/SRFBN_CVPR19
[9] https://github.com/tyshiwo/DRRN_CVPR17
[10] https://github.com/yulunzhang/RCAN

TABLE II: Overview of the most popular downscaling models in CV. Parameters are an estimation for the $\times 4$ scaling factor and links to the official code repositories are provided where possible.

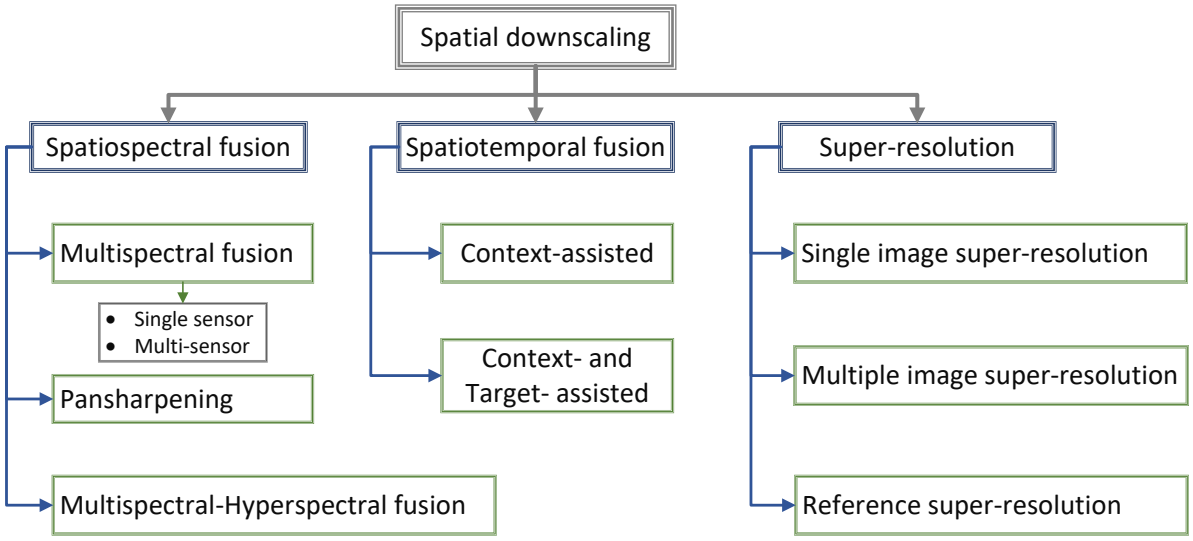


Fig. 8: Proposed taxonomy of DL downscaling methods in the literature.

A. *Multispectral image fusion*

Using information from a single satellite source has the advantage of consistent satellite orbit characteristics (e.g. altitude, inclination etc.) and atmospheric conditions. Some satellites carry multiple sensors that allow simultaneous capture of multiresolution images thus providing an ideal setting for SSF

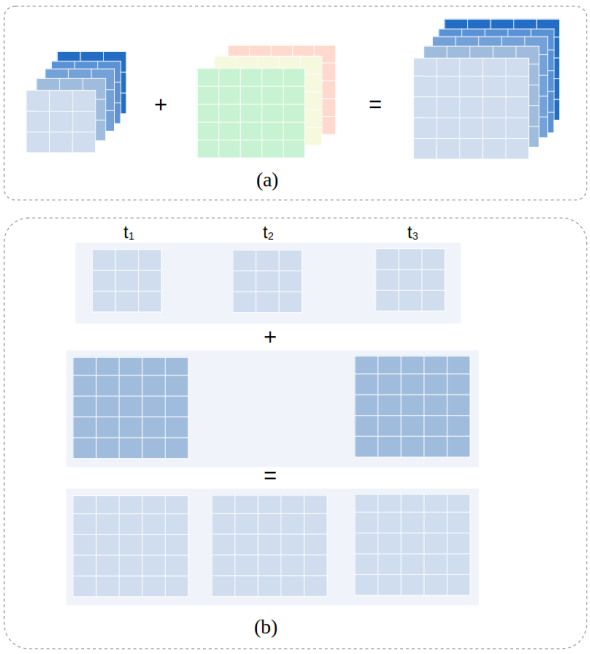


Fig. 9: (a) Spatiospectral Fusion: an image of coarse spatial resolution is fused with an image of fine spatial resolution containing different bands. The result is a version of the former image downscaled to the spatial resolution of the latter. (b) Spatiotemporal Fusion: an image of high temporal (t_1, t_2, t_3) but low spatial resolution is fused with an image of low temporal (t_1, t_3) but high spatial resolution. The result is an image of the highest spatial resolution in time t_2.

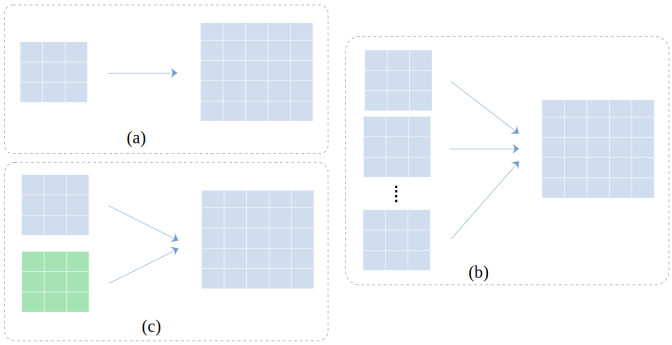


Fig. 10: (a) Single Image Super-resolution: A single LR image is downscaled without using any external information. (b) Multiple Image Super-resolution: Multiple LR images of the same scene are used to acquire an image of higher spatial resolution of that scene. (c) Reference Super-resolution: An LR image is downscaled by combining information from features extracted from it.

and a common data source. For example, the constellation of Sentinel-2 satellites (A/B) launched by the European Space Agency (ESA) acquires an image with 13 discrete bands, four of which have 10m spatial resolution, six have 20m and three have 60m [93]. Several methods ([94] (*DSen2*, *VDSen2*), [95], [96] (*FUSE*), [97] (*SPRNet*)) use two input sets, one for the B_{HR} and one for the B_{LR} resampled to match the target resolution, as input to CNN models which aim to transfer high-frequency details from B_{HR} to B_{LR} in order to spatially enhance the latter accordingly. DSen2, VDSen2 and the model proposed by Palsson et al. [95] use a concatenation of both sets in the input while FUSE and SPRNet process each set in parallel and then fuse the results. In a similar setting, Luo et al. [98] (*FusGAN*) propose a GAN framework consisting of an ESRGAN Generator and a PatchGAN Discriminator [99], which takes as input a downsampled concatenation of HR and LR Sentinel-2 bands, to recover the original LR bands (Fig. 14). On the other hand, Nguyen et al. [100] (*S2SUCNN*) propose a multi-scale model which takes as input the bands in their original resolution and progressively upsamples the lower resolution ones guided by the extracted features of the higher resolution bands to finally obtain all Sentinel-2 bands in a 10m spatial resolution. The final result is subsequently degraded to be compared with the original input in a MAE loss function. Finally, an interesting approach is presented in [101], where the FUSE model is evaluated under an unsupervised training scheme. Contrary to the original FUSE study which employs a pre-upsampling framework and thus relies upon the primary creation of synthetic training data, the authors propose a reversed pipeline, where the model is applied on the original images and its output is then downsampled and compared with the initial input. Subsequently, a second term is added to the loss function which is calculated on the local correlation between the B_{HR} and B_{LR} bands and accounts for the preservation of high-frequency details. The preliminary results showcase the potential of this approach, which however is still below the level of the supervised learning scheme.

Shao et al. [102] (*ESRCNN*) propose a framework that extends the SRCNN architecture (Table II) and utilizes auxiliary information from Sentinel-2 in order to downscale Landsat-8 images. The Landsat-8 satellite provides observations in the visible, NIR and SWIR spectra at 30m and a panchromatic band at 15m spatial resolution every 16 days [103] so the goal of this study is to produce the equivalent Landsat images at 10m spatial resolution. The whole process can be broken down into two separate steps. First, is the self-adaptive fusion of Sentinel-2, where the 20m Sentinel-2 bands (11 and 12) are resampled to 10m using k-Nearest Neighbours (k-NN) interpolation and are then concatenated with the native 10m bands as input to the proposed ESRCNN model. The output are the bands 11 and 12 downscaled to 10m resolution. Following this is the multi-temporal fusion of Landsat-8 and Sentinel-2, where the 30m Landsat bands (1-7) and the panchromatic band are resampled to 10m again using k-NN interpolation and are concatenated with the native 10m Sentinel-2 bands and the downscaled 20m Sentinel-2 bands. These are fed to the ESRCNN which outputs a downscaled version of the Landsat bands 1-7. A distinct advantage of this method against traditional approaches is the ability to fuse Sentinel-2 and Landsat data obtained on different, albeit close, dates. Using the same satellite sources, Chen et al. [2] propose the fusion of Sentinel-2 and Landsat images in order to enhance the latter to a spatial resolution of 10m. They prove that an adversarial approach is superior to a non-adversarial one and the proposed model resembles the architecture of the ESRGAN trained on a

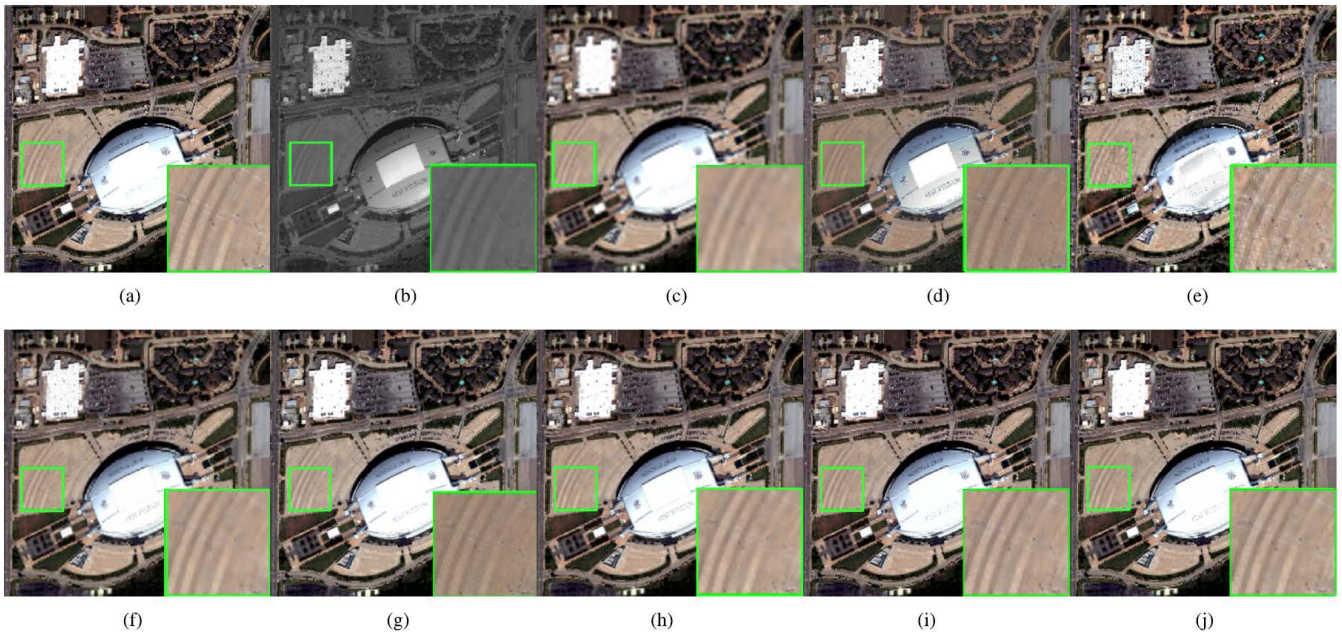


Fig. 11: Example of pansharpening on WorldView-3 data. (a) High-resolution, "ground truth" image, (b) Panchromatic, (c) Low-resolution multispectral, (d) - (j) Pansharpening results obtained by different DL approaches. Image taken from [90] (©2021 IEEE).

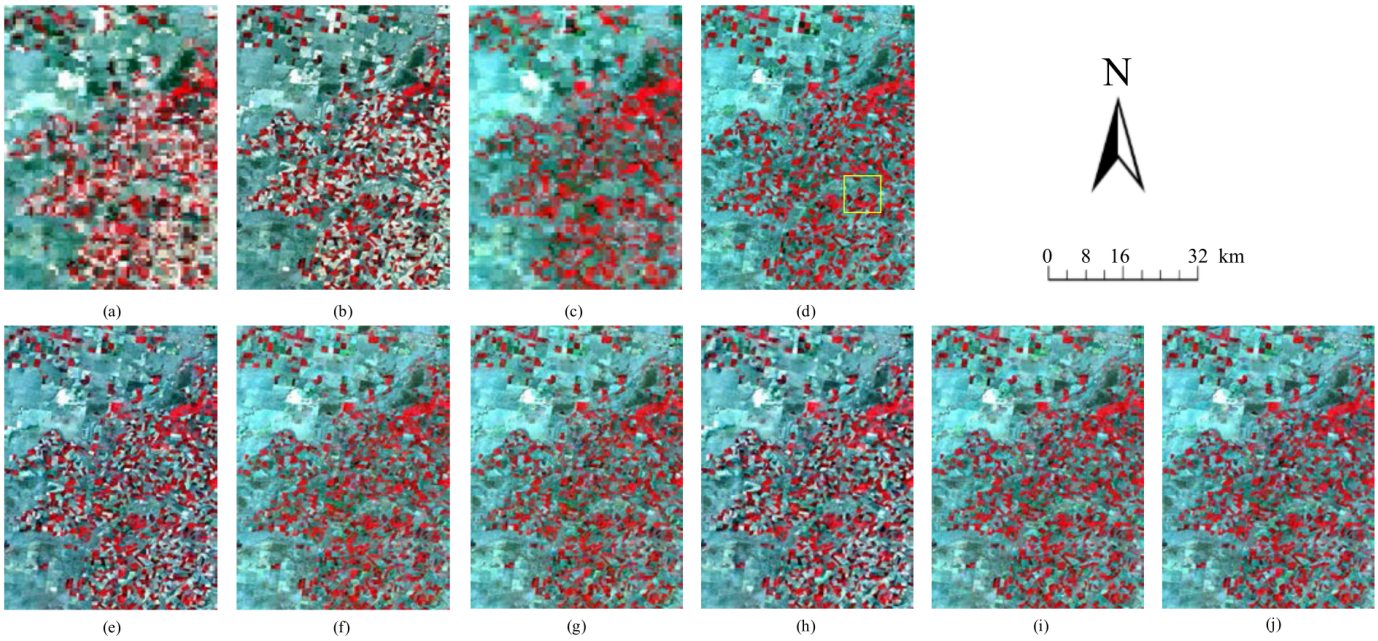
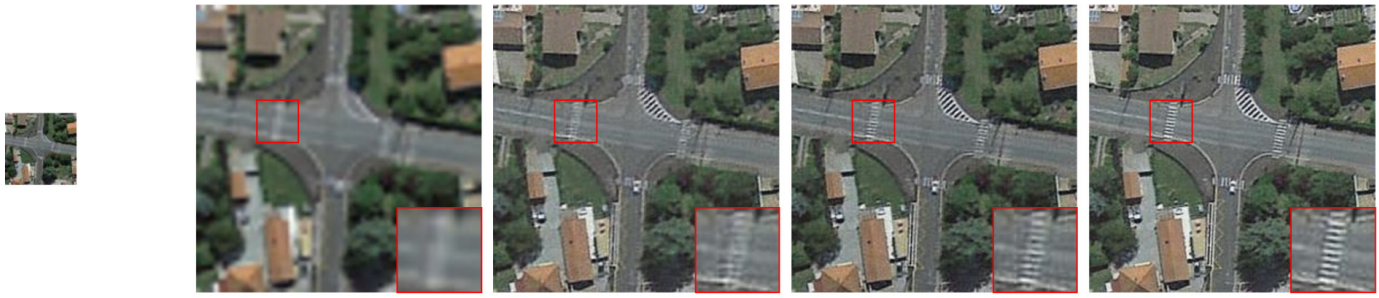


Fig. 12: Example of Spatiotemporal Fusion. (a) Low resolution image on time t_1, (b) High resolution image on time t_1, (c) Low resolution image on time t_2, (d) High resolution image on time t_2 which is the target, (e) - (j) Prediction results on time t_2 obtained by different approaches. Image taken from [91] (©2021 IEEE).

composite of the Red, Green and Blue bands for both satellites. The authors also tested whether the GAN model could be improved by pretraining on natural instead of satellite images using the DIV2K dataset (Section X) but the results were not favourable.

In their study, Dong et al. [104] (*RRSGAN* and *RRSNet*) argue that Remote Sensing images coming from different sources must be carefully aligned before processing due to differences in altitude, viewpoint or angle. They form a dataset consisting of WorldView-2 (0.5m) and GaoFen-2 (0.8m) observations as well as the corresponding images from Google Earth (0.6m). The proposed model is a GAN where image alignment is assisted by the extraction of gradients. In particular, a CNN is fed the input images and their gradients, and proceeds



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

Fig. 13: Example of Single-Image Super-resolution. (a) Low resolution image, (b) - (e) Prediction results obtained by different approaches for a scaling factor $\times 4$. Image taken from [92] (©2020 IEEE).

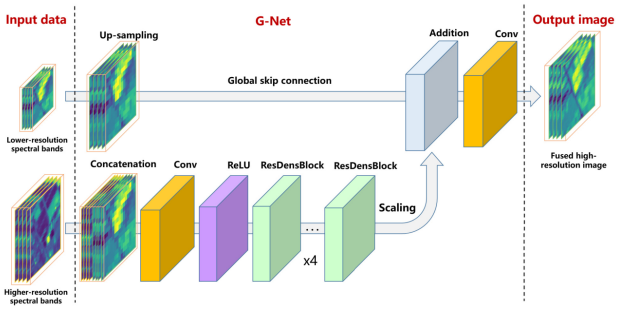


Fig. 14: The FusGAN Generator network. *Conv* indicates the convolutional layer, *ResDensBlock* the Residual Dense Block as described in ESRGAN. Image taken from [98] (©2020 IEEE).

to extract features which are then aligned via a pyramid with deformable convolutional layers [105]. Subsequently, a Relevance Attention Module is proposed in order to combine the aligned features by focusing on the relevant information, and a series of upsampling blocks performs the final downscaling. For the adversarial training two Discriminators are employed, one for the downscaled image and one for the gradient of the downscaled image produced by the Generator. The loss function is a weighted sum of: (i) the MAE between downscaled and HR image, (ii) the adversarial loss for downscaled and HR image, (iii) the VGG loss between downscaled and HR image, (iv) the MAE between the gradients of the downscaled and HR images, and (v) the adversarial loss for the gradients of the downscaled and HR images. Results show that both the adversarial RRSGAN and the non-adversarial RRSNet perform better than numerous other DL methods, with RRSGAN producing more high frequency details.

In conclusion, considering single-source data for multispectral image fusion, the available solutions cover a variety of needs. For example, when all LR input images have the same spatial resolution (e.g. 20m) then SPRNet seems to be a more suitable and robust approach. On the other hand, when hardware and/or time restrictions apply, FUSE provides a lightweight candidate since it contains very few trainable parameters ($\sim$ 28k) compared to other methods but has only be applied with a $\times 2$ scaling factor. Finally, for an end-to-end approach where all multiresolution input bands are downscaled in a single forward pass, FusGAN seems to produce more accurate and sharp results. In the case of multi-source input data, ESRCNN tackles the lack of clear, cloudless HR input images on the required date by enabling the use of multiple HR images acquired at arbitrarily close dates. The authors observe that specially when using more than 3 Sentinel-2 images, the model is able to additionally capture Land Use/Land Cover changes in the landscape. On the contrary, when the HR input images are inevitably contaminated by clouds or even absent in some cases, RRSGAN is able to overcome the loss of information and produce downscaled results of acceptable quality thanks to its robust feature extraction and attention mechanisms.

B. Pansharpening

Pansharpening refers to a downscaling process aided by a panchromatic band. This special type of band allows the acquisition of a single measurement for the total intensity of visible light in a single pixel, thus panchromatic sensors are able to detect brightness changes at quite small spatial scales.

The first work to introduce convolutional neural networks to pansharpening is [106] (*PNN*). Inspired by the super-resolution field of computer vision, Masi et al. build upon SRCNN and improve it by augmenting the input with a number of radiometric indices tailored to features relevant for Remote Sensing applications (NDVI, NDWI, etc.). Following the three steps of sparse coding super-resolution [107], they make use of a three layer convolutional neural network named *PNN* as shown in Fig. 15. Their method follows the pre-upsampling framework.

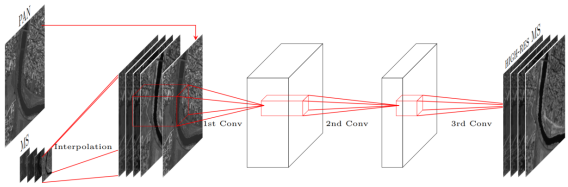


Fig. 15: Outline of PNN. The network comprises of three layers that are expected to match the three steps of sparse coding super-resolution. Image taken from [106].

Motivated by the high non-linearity of deeper networks and inspired by SRCNN and PNN, Wei et al. propose a deep residual network named *DRPNN* [108], in which they add some pansharpening specific improvements. Yang et al. also propose a deep residual network named *PanNet* [109] that preserves both spatial and spectral resolution. For spectral preservation, they directly add the upsampled multispectral images to the network output, while for spatial preservation, they train the network in the high-pass filtering domain rather than the image domain, as this is expected to generalize better among different satellites (Fig. 16). Starting from PNN too, Scarpa et al. [110] explore a number of variations to improve its performance and robustness. They propose the use of MAE loss, which boosts performance and allows fast convergence, exploit skip connections and add a target-adaptive fine-tuning phase. Their ablation study shows that shallow architectures are able to perform as well as the deeper ones, thus they use a three layer CNN (*L1-RL-FT*) with residuals.

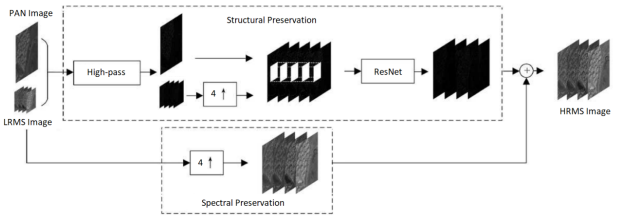


Fig. 16: Outline of PanNet. The network decouples structural from spectral preservation. Image taken from [109] (©2017 IEEE).

A different approach inspired by metric learning that makes use of stacked autoencoders is introduced in [111]. Up-scaled panchromatic images are divided into patches, grouped according to their geometry and fed as input to autoencoders that are utilized to map them into hierarchical feature spaces that accurately capture non linear manifolds, while at the same time preserve their local geometry in the embedding space. Based on the assumption that multispectral and their corresponding panchromatic patches form the same geometric manifolds, the geometric multimanifold embedding model (*DML-GMME*) using a metric learning loss function is trained to estimate high resolution multispectral image patches.

A two-branch network named *MSDCNN* is proposed in [112]. While the one branch is a three layer convolutional neural network, the other one is a deep residual network with multiscale convolutional blocks. Multi-scale refers to the fact that authors use convolutional filters with different sizes to extract feature maps. The two subnetworks are jointly trained and the final estimation is a sum over the estimation of each subnetwork.

In [113] (*DiCNN*), a general detail injection formulation of pansharpening is proposed. DiCNN comprises two convolutional neural networks, *DiCNN1* and *DiCNN2*, both utilizing the pre-upsampling framework. *DiCNN1* adds a skip connection to the PNN architecture, while *DiCNN2* works under the assumption that ideally, the multispectral spatial details should match and be relevant only to the panchromatic image. Thus, it utilizes only the panchromatic image as an input to the network, while the pre-interpolated multispectral image is used only at its end. Structural comparison between PNN, DRPNN, DiCNN1 and DiCNN2 can be seen in Fig. 17.

Liu et al. [115] propose a method named *MIPSM* that combines a shallow–deep convolutional network (*SDCN*) and a spectral discrimination-based detail injection (*SDDI*) model. SDCN consists of a three layer shallow network and a deep residual network, which can capture mid-level and high-level spatial features from panchromatic images. SDCN works on the high-pass filtering domain. SDDI is developed to merge the spatial details extracted by SDCN into multispectral images with minimal spectral distortion. SDCN and SDDI are jointly trained.

Inspired by component substitution and multiresolution analysis, Deng et al. [116] design two deep residual networks named *CS-Net* and *MRA-Net* respectively that extract details and have a solid physical justification. They also design a network that is directly fed with details extracted by differencing the single panchromatic image with each multispectral band. This network is called *Fusion-Net*. They make use of the pre-upsampling framework using a polynomial kernel. Cai et al. [117] propose a progressive downscaling pansharpening neural network named *SRPPNN*, which includes three components: (a) a downscaling process that extracts inner spatial detail that is present in multispectral image and combines it with the spatial detail of panchromatic image to generate fused results, (b) progressive pansharpening to separate the spatial resolution improvement process, which achieves a gradual and stable pansharpening process and (c) a high-pass residual module that helps by directly injecting spatial detail from panchromatic images and achieves better spatial preservation. Dong et al. [118] propose a Laplacian pyramid network called *LPPNet* that has a clear physical interpretation of pansharpening, follows the general idea of multiresolution analysis and divides pansharpening into two processes: (a) detail extraction and (b) reconstruction. For (a), they use the Laplacian pyramid to decompose the panchromatic image into multiple levels that can distinguish the details of different scales. They build a simple detail extraction subnetwork for each level, which can help fully extract the depth of different levels. For (b), the subband residuals estimated at each level are injected into the respective level of the multispectral image, while they are upsampled and fed as input to the next subnetwork, which can help make full use of complementary details between different levels.

Instead of focusing on the architecture, Jiang et al. [119] focus on the input/output of the network. They introduce three novelties: (a) the differential information mapping strategy, (b) the auxiliary gradient information strategy and (c) the combination of an attention module with residual blocks. Taking into account the under-utilization of the panchromatic image in the input, they propose to copy and assign the panchromatic image to each band of the downscaled multispectral image.

Motivated by the existence of mixed pixels in satellite images, where each pixel tends to cover more than one constituent material, Qu et al. [120] propose a method based on self-attention mechanism (*SAM*) [121] that works at the sub-

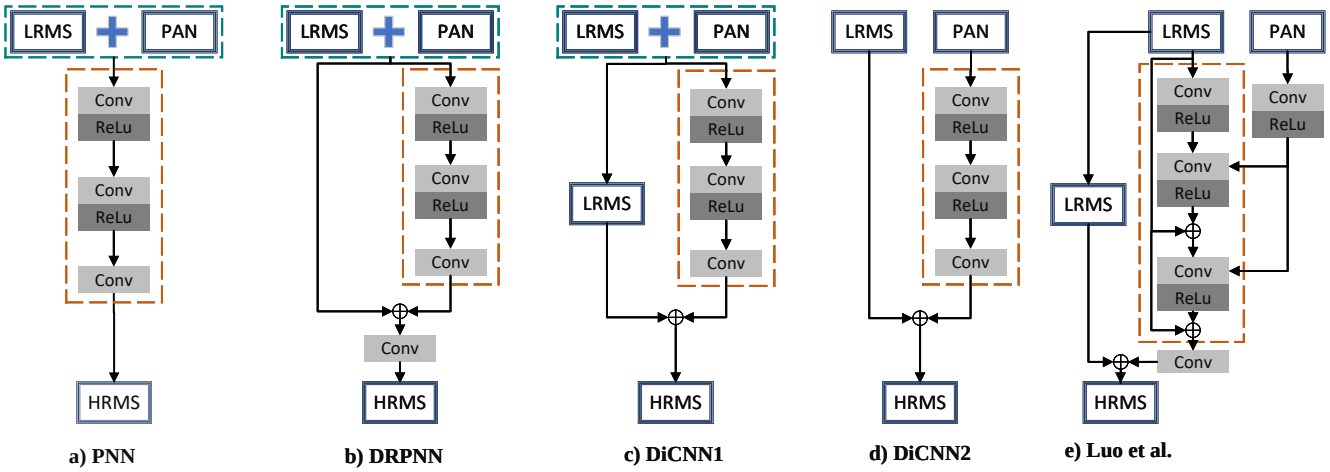


Fig. 17: Structural comparison between (a) PNN, (b) DRPNN, (c) DiCNN1 and (d) DiCNN2, modified from [113] (©2019 IEEE), and e) the model of Luo et al., modified from [114] (©2020 IEEE).

pixel level. A method using skip connections inspired by [122] is introduced in [114]. Luo et al. propose a novel loss function that utilizes spatial constraints, spectral consistency and the QNR index (Section IV). Instead of using simple stacked convolutional layers and separate the feature extraction way, their network architecture adopts an iterative way to jointly extract and fuse the features. Outline of their method can be seen in Fig. 17e.

Zhang et al. [123] propose a model comprising two networks: gradient information network (*TNet*) and pansharpening network (*PNet*). TNet is a residual network committed to seeking the nonlinear mapping between gradients of panchromatic and high resolution multispectral images, which essentially is a spatial relationship regression of imaging bands in different ranges. PNet is a spatial attention residual network used to generate high resolution multispectral images, which is not only supervised by the high resolution multispectral reference image, but also constrained by the trained TNet.

Inspired by the learned iterative soft thresholding algorithm, Yin et al. [124] propose a deep pansharpening network that integrates the detail injection, variational optimization and Deep Learning schemes into a single framework. It consists of the input convolutional layer, *Conv-ISTA* module (deep unfolded network), fusion module and the output convolutional layer. The weighted use of variational optimization with Deep Learning is proposed in *VO+Net* [125] too. For the variational optimization modeling, a general details injection term inspired by the classical multiresolution analysis is proposed as a spatial fidelity term and a spectral fidelity employing the multispectral sensor's modulation transfer functions is also incorporated. For the Deep Learning injection, a weighted regularization term is designed to introduce Deep Learning into the variational model. The final convex optimization problem is efficiently solved by the designed alternating direction method of multipliers.

Zhang et al. [126] (*SC-PNN*) propose a saliency cascade convolutional neural network that consists of two parts: (a) a dilated deformable fully convolutional network (*DDCN*) for saliency analysis and (b) a saliency cascade residual dense network (*SC-RDN*) for pansharpening. DDCN is a network based on hybrid and deformable convolution aiming to separate salient regions like residential areas from nonsalient areas like mountains and vegetation areas. SC-RDN is composed of three stages: (a) detail maps of multispectral and panchromatic images are extracted via dual-tree complex wavelet transform (*DT-CWT*) [127], (b) a deep regression network based on residual dense blocks takes those detail maps as input and produces the primarily sharpened image with high spatial and spectral quality and (c) a saliency enhancement module emphasizes the impact of the obtained saliency map via the saliency-weighted region convolution (*SW-RC*). More details about this method can be seen in Fig. 18.

Given that the convolution operation is focused on the local region and thus position-independent global information is difficult to obtain, Lei et al. [90] propose an efficient non-local attention residual network (*NLRNet*) to capture the similar contextual dependencies of all pixels.

Motivated by the unavoidable absence of ground truth, which often results in networks trained solely in a reduced resolution domain, Vitale et al. [128] propose a new learning strategy involving a loss function with terms computed both at reduced and full resolution images, thus enforcing cross-scale consistency. Their method is based on *A-PNN* [129], an advanced version of PNN with: (a) a different loss function for training (MAE instead of MSE), (b) a residual learning configuration and (c) a target adaptive scheme. In the same direction, Ciotola et al. [130] introduce a full-resolution training framework, in which training takes place in the high-resolution domain, relying only on the original panchromatic and multispectral pairs (with no downgrading), thus avoiding any loss of information. They design a new compound loss function with two components accounting separately for spatial and spectral consistency.

Apart from convolutional neural networks, one of the first attempts to utilize generative adversarial networks for producing high quality pansharpened images is introduced by

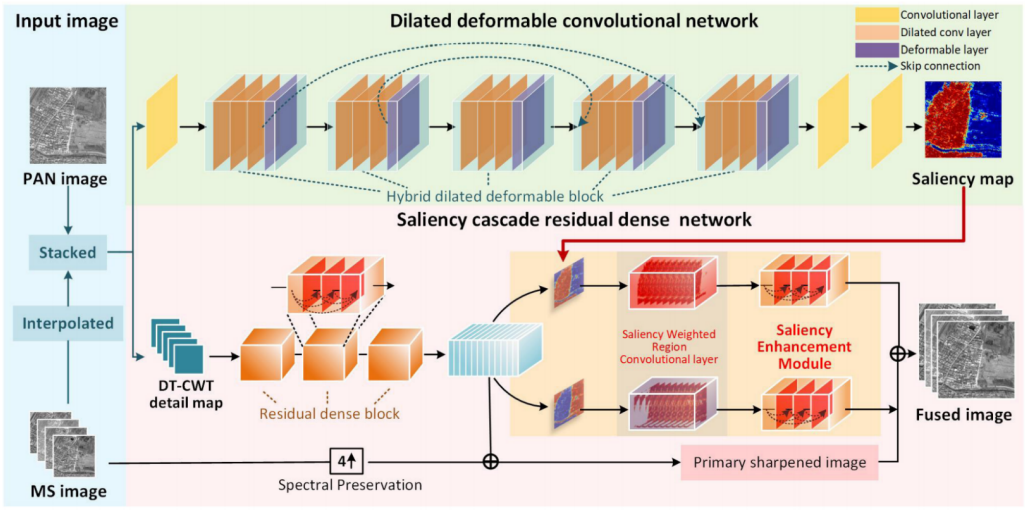


Fig. 18: Outline of SC-PNN. Image taken from [126] (©2021 IEEE).

Liu et al. in [131] (*PSGAN*). PSGAN comprises a Generator that takes as input panchromatic images and maps them to the desired high resolution multispectral images and a Discriminator that implements the adversarial training strategy for generating higher fidelity pansharpened images. Making the assumption that (a) the spectral distribution of the fused image should be consistent with that of the LR multispectral and (b) the spatial distribution of the fused image should be consistent with that of the panchromatic image with the same resolution, Ma et al. propose the use of a generative adversarial network with two discriminators in [132] (*Pan-GAN*). The Generator of Pan-GAN attempts to generate a high resolution multispectral image containing major spectral information of the LR multispectral image together with additional image gradients of the panchromatic image.

A similar generative adversarial network architecture called *MDSSC-GAN SAM* exploiting jointly the spatial and spectral information sources is proposed in [133]. Gastineau et al. make use of two Discriminators too, one to preserve the texture and geometry of images by taking as input the luminance Y and near-infrared band of images and the other to preserve the color and the spectral resolution by comparing the chroma components Cb and Cr. A different approach in which pan-sharpening is treated as a colorization problem is introduced by Ozcelink et al. in [134] (*PanColorGAN*). In contrast with the ordinary, the authors give as input the grayscale transformed multispectral image and train the model to learn the colorization of it. The model learns to generate an original multispectral image by taking as input the corresponding reduced-resolution and grayscale ones. PanColorGAN is trained using both a reconstruction (MAE) and an adversarial loss. This can be interpreted as that the model learns to separate the spectral and spatial components of the multispectral image during training.

In conclusion, when hardware and/or time restrictions apply, L1-RL-FT is a great solution, as it is lightweight and trains very fast. It also seems to have good generalization ability and to solve the problem of insufficient data with its target-adaptive tuning phase. DML-GMME is a unique approach that utilizes deep metric learning and autoencoders. Having a rich ablation and being a lightweight model, a researcher would gain useful insight experimenting with it. Accurate and sharp results seem to be produced by LPPNet, a network that simplifies the pansharpening problem into several pyramid-level learning problems. LPPNet makes use of the Laplacian pyramid decomposition technique to decompose the image into different levels that can differentiate large- and small-scale details, thus achieving great visual appearance. Novel ideas that a researcher might want to consider are presented in Zhang et al. and Luo et al. Zhang et al. design a special gradient transformation network that searches the nonlinear mapping between gradients of panchromatic and multispectral images. Luo et al. propose a panchromatic-guided strategy that continuously extracts and fuses features from the panchromatic image. VO+Net is a framework that can be put on top of other approaches to improve the end result. Finally, SC-PNN is a solution that successfully makes use of saliency maps and provides great visual results.

C. Hyperspectral/Multispectral fusion

Hyperspectral image sharpening aims at fusing an observable low spatial resolution hyperspectral image with a high spatial resolution multispectral of the same scene in order to acquire a high resolution hyperspectral image.

One of the first works to utilize CNNs for hyperspectral/multispectral fusion is introduced by Palsson et al. in [135]. Authors propose the use of a 3D CNN with three layers for the hyperspectral/multispectral fusion. The dimensionality of the hyperspectral image is reduced using PCA in order to constrain the computational cost and increase robustness. Dian et al. [136] propose a deep hyperspectral image sharpening method called *DHSIS* that directly learns the priors of the high resolution hyperspectral image via CNN-based residual learning. They first initialize the HR hyperspectral image by solving a Sylvester equation. Then, to learn the priors, they utilize the initialized HR hyperspectral image as the input

of the CNN to map the residuals between reference HR hyperspectral image and initialized HR hyperspectral image. This initialization can fully utilize the constraints of the fusion framework, thus improving the quality of the input data. The learned priors of HR hyperspectral image are returned to the fusion framework to reconstruct the final estimated HR hyperspectral image, which can further improve performance (Fig. 19).

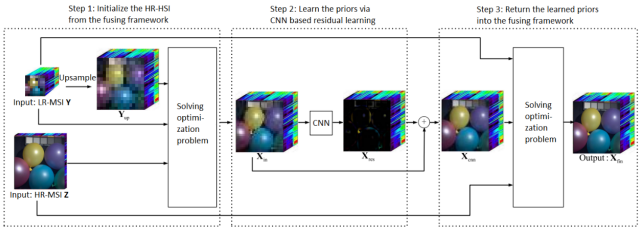


Fig. 19: Outline of DHSIS, a deep hyperspectral image sharpening method. Image taken from [136] (©2018 IEEE).

Zhou et al. [137] introduce a pyramid fully convolutional network (*PFCN*) consisting of two subnetworks: (a) an encoder aiming to encode the low resolution hyperspectral image into a latent image and (b) a pyramid fusion that utilizes this latent image together with a high resolution multispectral pyramid image to progressively reconstruct the high resolution hyperspectral image in a global-to-local way. More details about the method can be seen in Fig. 20.

Instead of formulating the task of hyperspectral/multispectral fusion as the spatial downscaling of a low resolution hyperspectral image, Han et al. [138] formulate it as the spectral downscaling of a high resolution multispectral image. Their method *CF-BPNN* consists of three stages: (a) the fusion problem is formulated as a nonlinear spectral mapping from a high resolution multispectral image to a high resolution hyperspectral image with the help of a low resolution hyperspectral image, (b) a cluster-based learning method using multi-branch neural networks is utilized to ensure a more reasonable spectral mapping for each cluster and (c) an associative spectral clustering is proposed to ensure that training and fusion clusters are consistent.

He et al. introduce *HyperPNN* [139], a hyperspectral image sharpening method via spectrally predictive CNNs, exploiting spectral convolution structure to strengthen spectral prediction. Li et al. propose a detail-based deep Laplacian pansharpening model (*DDLPS*) [140] to improve the spatial resolution of hyperspectral imagery. Their method includes three main components: downscaling, detail injection and optimization. They make use of the well-known Laplacian pyramid super-resolution network LapSRN (Section V) to improve the resolution of each band. Then, a guided image filter and a gain matrix are used to combine the spatial and spectral details with an optimization problem which is formed to adaptively select an injection coefficient.

Shen et al. [141] propose a twice optimizing net with matrix decomposition (*TONWMD*). They first decouple the fusion problem into a spectral and a spatial optimization task with the help of matrix decomposition. These two problems are handled sequentially by solving a linear (Sylvester) equation. Then, they train a deep residual network to establish the mapping between the initial and reference images. Finally, the predicted result is returned to the optimization procedure to get the final fusion image. Xie et al. [142] propose *MHF-Net*, a network having clear physical meaning and great interpretability. They first construct a hyperspectral/multispectral fusion model which merges the generalization models of low resolution images and the low rankness prior knowledge of high resolution hyperspectral image into a concise formulation. Then, they build the network by unfolding the proximal gradient algorithm to solve the proposed model. Liu et al. [143] propose *UMAG-Net*, a network comprising a multi-attention autoencoder network (*MAE*) and a multiscale feature-guided network (*MSFG*). First, MAE extracts deep multiscale features of the multispectral image and then a loss function containing a pair of hyperspectral and multispectral images is used to iteratively update the parameters of the network and learn prior knowledge of the fused image. MSFG is used to construct the final high resolution hyperspectral image. Non-local blocks are used to better retain spectral and spatial details of the image. Laplacian blocks are used to connect the MAE with the MSFG to achieve better fusion results while ensuring feature alignment. Although UMAG-Net does not use satellite hyperspectral data, the expansion into them is straightforward. Fig. 21 shows the method.

Zhang et al. [144] propose *SSR-Net*, an interpretable spatial-spectral reconstruction network that consists of three components: (a) cross-mode message inserting (*CMMI*); an operation producing a preliminary fused high resolution hyperspectral image, (b) a spatial reconstruction network (*SpatRN*) that focuses on reconstructing the lost spatial information of the low resolution hyperspectral image with the guidance of a spatial edge loss and (c) a spectral reconstruction network (*SpecRN*) that aims to reconstruct the lost spectral information of the high resolution multispectral image under the constraint of a spectral edge loss.

In conclusion, even though the architectures proposed in hyperspectral/multispectral fusion are limited in number, they exhibit remarkable variability (CNNs, 3D CNNs, GANs, etc.). MHF-Net is an interpretable network showing superiority both visually and quantitatively. A bright idea that a researcher should take into account is presented in PFCN. Authors propose to encode the spectral information of the low resolution hyperspectral image into a latent image and then decode this image with a high resolution multispectral image pyramid into a sharp high resolution hyperspectral image. The drawback of this method is the fact that experiments are conducted on simulated images. SSR-Net treats hyperspectral/multispectral fusion as a spatial-spectral reconstruction problem. Authors provide a good ablation study and useful insights. Finally, a complete solution that has not yet been tested on Remote Sensing data is proposed in UMAG-Net. This solution combines great ideas like the use of multi-attention, non-local blocks, Laplacian blocks and a loss function that measures both spectral and spatial similarity between pairs of images.

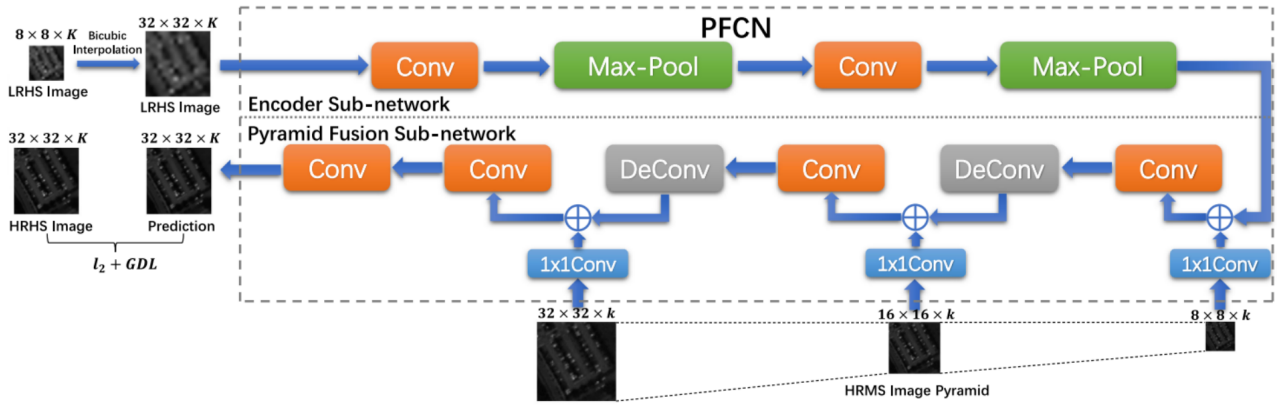


Fig. 20: Outline of PFCN comprising an encoder sub-network and a pyramid fusion sub-network. Image taken from [137] (©2019 IEEE).

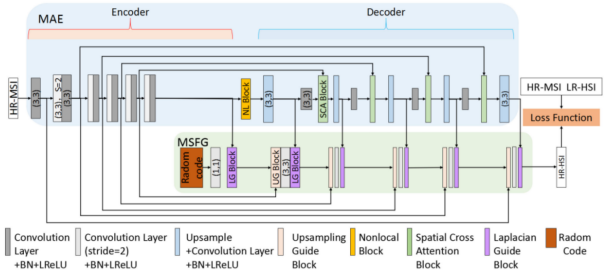


Fig. 21: Outline of UMAG-Net comprising an encoder and a decoder with spatial cross attention mechanism. Image taken from [143] (©2021 IEEE).

VIII. Spatiotemporal Fusion

Apart from their spectral signatures, satellites are also characterized by their unique revisit times. Spatiotemporal Fusion (STF) aims to integrate images of high spatial but low temporal resolution (HSLT) with images of low spatial but high temporal resolution (LSHT). A typical dataset for the STF problem consists of LSHT-HSLT image pairs at one or multiple time steps and the aim is to predict an HR image on a future or intermediate target time t_{target}. All images must contain similar spectral information, including the number of bands and the bandwidths. For example, MODIS (Moderate Resolution Imaging Spectroradiometer) captures images daily (high temporal resolution) at a 250m to 1km scale (low spatial resolution) [146], whereas Landsat-8's OLI captures images every 16 days (low temporal resolution) at a 30m scale (high spatial resolution) [103]. Both sensors operate on the visible and infrared spectra, therefore one could combine pairs of MODIS (LSHT) and Landsat-8 OLI (HSLT) images on different dates in order to produce high spatial resolution images on a prediction date t_{target}.

The various STF methods present in the literature follow a context-assisted (C-A) or context- and target-assisted (CT-A) scheme depending on the availability of target data during the training phase. CT-A approaches use additional LSHT information on t_{target} whereas C-A approaches exploit LSHT-HSLT pairs from non-target times only (Fig. 22). We must note here that a couple of other discriminant factors can also be observed among STF studies. First, some methods perform a pre-processing step where time difference images defined as $I_{ij} = I_j - I_i$ for the time steps t_i and t_j are computed and used as additional input to the model. Such an approach is followed by [91], [147]–[153]. Secondly, whereas the most common strategies involve data from times prior to t_{target}, there are cases where future observations are also required, as in [147], [150]–[158]. For simplicity, in this work we will solely employ the C-A vs. CT-A classification and separately describe each category in the following subsections, while in Table IV we provide an overview of all STF methods. Note that we refer to the HSLT images on time t as F_t and the LSHT images as C_t respectively.

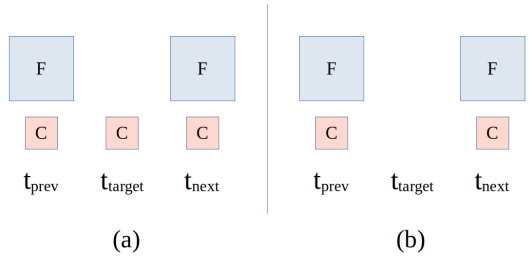


Fig. 22: Data used for (a) CT-A and (b) C-A Spatiotemporal Fusion during training. t_{prev} and t_{next} are one or multiple dates before and after the target date t_{target} respectively. F refers to the HSLT image, C to the LSHT image.

A. Context- and target-assisted (CT-A)

Several researchers argue that the spatial resolution gap between certain sensors, such as those carried by MODIS and Landsat, is quite large and data coming from both sources undergo different atmospheric and geometric corrections. Therefore, they design models which produce intermediate images enhanced by a smaller scaling factor in order to facilitate the downscaling process. For example, Song et al. [154] (*STFDCNN*, Fig. 23) propose a two-stage model which takes as input an arbitrary pair of Landsat-5/7 (25m) and MODIS

Model	Fusion type	Fusion data	CV model	Building blocks	Upsampling framework	Architecture	Code available / # params
DSen2 [94]	MS	Sentinel-2	-	Residual learning	pre-upsampling	CNN	yes / 1.8m
VDsen2 [94]	MS	Sentinel-2	-	Residual learning	pre-upsampling	CNN	yes / 37.8m
Palsson et al. [95]	MS	Sentinel-2	-	Residual learning	pre-upsampling	CNN	no / -
FUSE [96]	MS	Sentinel-2	-	Residual learning	pre-upsampling	CNN	no / 28k
FusGAN [98]	MS	Sentinel-2	ESRGAN	Residual learning, sub-pixel convolution	post-upsampling	GAN	no / -
S2SUCNN [100]	MS	Sentinel-2	-	Residual learning	progressive upsampling	CNN	yes / -
Ciotola et al. [101]	MS	Sentinel-2	-	Residual learning	-	CNN	no / -
SPRNet [97]	MS	Sentinel-2	-	Residual learning	pre-upsampling	CNN	no / -
ESRCNN [102]	MS	Multitemporal Landsat-8, Sentinel-2	SRCNN	-	pre-upsampling	CNN	yes / -
Chen et al. [2]	MS	Landsat-8, Sentinel-2	ESRGAN	Residual learning, sub-pixel convolution	post-upsampling	GAN	no / -
RRSGAN [104]	MS	WorldView-2, GaoFen-2	-	Residual learning, sub-pixel convolution, attention mechanism	progressive upsampling	GAN	yes / 7.47m
PNN [106]	PAN + MS	Ikonos, GeoEye-1, WorldView-2	SRCNN	-	pre-upsampling	CNN	yes / 310k
PanNet [109]	PAN + MS	Ikonos, WorldView-2, WorldView-3	-	Residual learning, high-pass filtering	progressive upsampling	CNN	no / 250k
DRPNN [108]	PAN + MS	Ikonos, WorldView-2, Quickbird	SRCNN	Residual learning	pre-upsampling	CNN	no / 1.6m
DML-GMME [111]	PAN + MS	Ikonos, WorldView-2, Quickbird, GaoFen-2	-	Stacked Sparse Autoencoders [145]	pre-upsampling	CNN	no / 8k
MSDCNN [112]	PAN + MS	Ikonos, WorldView-2, Quickbird	-	Residual learning	pre-upsampling	2 CNNs	no / -
L1-RL-FT [110]	PAN + MS	WorldView-2, WorldView-3	SRCNN	Residual learning	pre-upsampling	CNN	yes / -
DiCNN [113]	PAN + MS	WorldView-2 Washington, Ikonos Hobart, Quickbird Sundarbans	SRCNN	-	pre-upsampling	2 CNNs	no / 180k
DIRCNN [119]	PAN + MS	Ikonos, Quickbird, Gaofen-1, Gaofen-2	-	Residual learning, attention mechanism, auxiliary gradient data	pre-upsampling	CNN	no / 1.6m
MIPSM [115]	PAN + MS	Ikonos, Quickbird	-	Residual learning, high-pass filtering	pre-upsampling	2 CNNs	no / -
Fusion-Net [116]	PAN + MS	WorldView-2, WorldView-3, Quickbird, Gaofen-2	-	Residual learning	pre-upsampling	CNN	yes / 230k
SRPPNN [117]	PAN + MS	Quickbird, WorldView-3, Landsat-8	-	Residual learning, high-pass filtering	pre-upsampling	CNN	no / -
UP-SAM [120]	PAN + MS	GeoEye-1, IKONOS, WorldView-2, WorldView-3	-	Residual learning, attention mechanism, sub-pixel accuracy	pre-upsampling	CNN	no / -
Luo et al. [114]	PAN + MS	Gaofen-2, WorldView-2	-	Residual learning, attention mechanism	pre-upsampling	CNN	no / -
GTP-PNet [123]	PAN + MS	WorldView-2, Gaofen-2, Quickbird	-	Residual learning, gradient information	pre-upsampling	2 CNNs	no / -
PSCSC-Net [124]	PAN + MS	GeoEye-1, Ikonos, WorldView-2	-	Deep unfolding, variational optimization	pre-upsampling	CNN	no / 1.1m
VO+Net [125]	PAN + MS	WorldView-3, WorldView-2, QuickBird	-	Variational optimization	pre-upsampling	CNN	no / -
SC-PNN [126]	PAN + MS	WorldView-3, GeoEye-1, SPOT5	-	Saliency analysis, hybrid and deformable convolution	pre-upsampling	CNN + FCN	no / -
NLRNet [90]	PAN + MS	WorldView-3, QuickBird	-	Residual learning, attention mechanism	pre-upsampling	CNN	no / -
LPPNet [118]	PAN + MS	Pavia Center, Houston, Los Angeles	-	Laplacian pyramid decomposition	pre-upsampling	CNN	no / -
Vitale et al. [129]	PAN + MS	GeoEye-1, WorldView-2	-	Residual learning	pre-upsampling	CNN	no / -
Ciotola et al. [130]	PAN + MS	GeoEye-1, WorldView-2, WorldView-3	-	-	-	CNN	no / -
PSGAN [131]	PAN + MS	QuickBird, GaoFen-2, WorldView-2	-	-	pre-upsampling	GAN	yes / 1.88m
Pan-GAN [132]	PAN + MS	GaoFen-2, WorldView-2	-	2 discriminators: spatial and spectral	pre-upsampling	GAN	no / -
MDSSC-GAN SAM [133]	PAN + MS	Pléiades, WorldView-3	-	2 discriminators: spatial and spectral, residual learning, attention mechanism	pre-upsampling	GAN	yes / -
PanColorGAN [134]	PAN + MS	Pléiades, WorldView-2, WorldView-3	-	Self-supervised, noise/color injection	pre-upsampling	GAN	no / -
Palsson et al. [135]	MS + HS	Pavia Center, Ikonos	-	-	pre-upsampling	3D CNN	no / -
DHSIS [136]	MS + HS	CAVE, Harvard	-	Self-supervised, noise injection	pre-upsampling	GAN	yes / -
PFCN [137]	MS + HS	Botswana, Washington DC, Pavia Center	-	Residual learning	pre-upsampling	CNN	no / -
CF-BPNN [138]	MS + HS	AVIRIS, Pavia Center	-	k-means clustering	pre-upsampling	NN	no / -

HyperPNN [139]	MS + HS	Washington DC Mall, Moffett Field, Salinas Scene	-	-	pre-upsampling	CNN	no / -
DDLPS [140]	MS + HS	Moffett Field, Chikusei, Salinas Scene	LapSRN	-	pre-upsampling	CNN	no / -
TONWMD [141]	MS + HS	CAVE, Harvard, Pavia Center	-	Residual learning, matrix decomposition	pre-upsampling	CNN	no / -
MHF-Net [142]	MS + HS	CAVE, Chikusei, Houston, Pavia Center	-	-	pre-upsampling	CNN	yes / -
UMAG-Net [143]	MS + HS	CAVE, Harvard	-	Attention mechanism	pre-upsampling	CNN, AE	no / -
SSR-Net [144]	MS + HS	Pavia Center, Botswana, Washington DC Mall	-	-	pre-upsampling	CNN	yes / -

TABLE III: Summary of the state-of-the-art Deep Learning models for spatiospectral fusion for image downscaling in Remote Sensing. CV model refers to the models presented in Table II. PAN: panchromatic, MS: multispectral, HS: hyperspectral.

(500m) images and learns to predict an intermediate enhanced image of 250m spatial resolution. The intermediate image is computed in a pre-upsampling fashion, while the final 25m image via a post-upsampling SRCNN structure. During inference, features are extracted from MODIS images on times t_1, t_2 and t_3 (where t_2 is the prediction date) which are linearly combined with the corresponding Landsat images on t_1 and t_3 to produce the final HR result. Building upon this, Zheng et al. [158] (*VDCNSTF*) propose deeper network architectures and redesign the SRCNN stage as a multiscale model producing images at 125m and 25m.

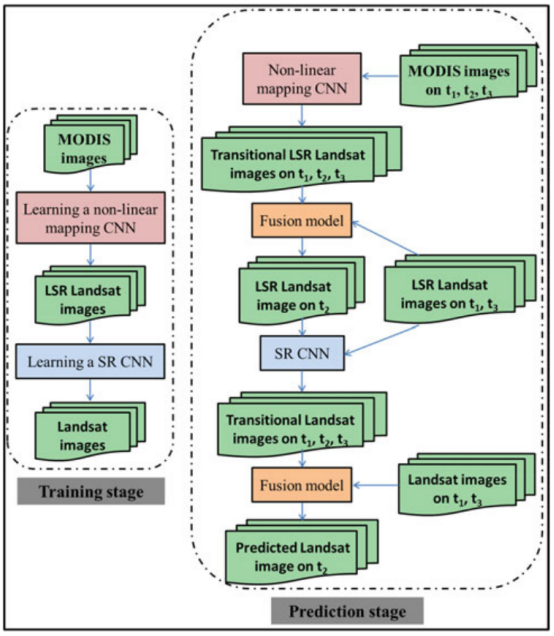


Fig. 23: Outline of the STFDCNN method. Image taken from [154] (©2018 IEEE).

A slightly different approach is followed by Liu et al. [147] (*StfNet*) who argue that the temporal changes expressed by a time difference image are highly correlated with the contents of the original images. Therefore, they design a model which takes as input a LSHT MODIS image (250-300m) at the prediction date t_2, a date before (t_1) and a date after (t_3) the prediction date, and a corresponding HSLT Landsat image at dates t_1 and t_3, produces time difference images and then reconstructs the HR image on date t_2 by transferring information from these temporal relations. More specifically, they propose two CNNs which take as input a concatenation of the MODIS time difference image and the Landsat image and produce a time difference Landsat. They employ these networks to learn the following mappings: (I) $(C_{13}, F_1) \rightarrow F_{13}$ and $(C_{13}, F_3) \rightarrow F_{13}$, (II) $(C_{12}, F_1) \rightarrow F_{12}$ and $(C_{23}, F_3) \rightarrow F_{23}$. Mapping (I) can be supervised by the label F_{13} which is available in the training data, forming the time difference reconstruction term of the loss function. The results of mapping (II) are summed to obtain a predicted F_{13} which is compared to the label F_{13}, forming the temporal consistency term of the loss function. The total loss function is a weighted sum of these two terms. Finally, the predicted F_{12} and F_{23} are combined with F_1 and F_3 through an adaptive local weighting strategy to obtain the target image F_2. A schematic outline of the method is presented in Fig. 24. Compared with non-DL and DL approaches, the proposed StfNet achieves sharper results will less visible artifacts.

Tan et al. [159] (*DCSTFN*) propose a two-branch CNN which takes as input the LSHT MODIS image on the prediction date t_2 along with a pair of HSLT Landsat-8 and LSHT MODIS (500m) images on a date prior but close to the prediction date, t_1. The first branch of the model learns a mapping from LSHT to HSLT images in a post-upsampling scheme, while the second one extracts information from the HSLT with a sequence of convolutional layers. The three outputs, which share the same width and height, are then concatenated following the assumption of the traditional STARFM algorithm [160]: $F_2 = C_2 - F_1 - C_1$ for dates $t1$ and $t2$ and enter a series of convolutions for the final reconstruction. In a subsequent publication [161] (*EDCSTFN*), the authors propose an enhancement over the DCSTFN model which instead of processing solely the LSHT images on the first branch, it takes as input both the LSHT images and the HSLT image concatenated along the channels dimension and extracts information on their spectrum differences. Finally, the authors describe a novel, flexible training scheme where more than one reference pairs can be used as input either during the training or inference phase, depending on data availability. The proposed EDCSTFN model manages to outperform DCSTFN and StfNet on most cases, while displaying a more stable and consistent behaviour.

Li et al. [148] (*DMNet*) propose a complex CNN architecture with two multiscale mechanisms including parallel convolutions with either different kernel sizes or different dilation rates for a more efficient feature extraction. The model takes as input the MODIS time difference image C_{12}

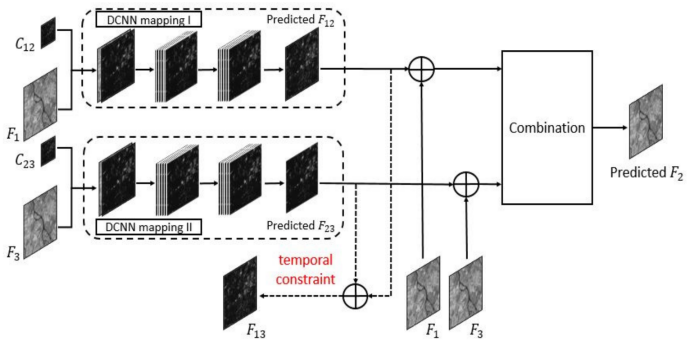


Fig. 24: Outline of the StfNet method. *DCNN* refers to a 3-layer deep CNN. Image taken from [147] (©2019 IEEE).

and the Landsat image F_1 and learns to predict F_2. In a follow-up study [149] (*AMNet*), the authors propose progressive upsampling at three scales ($\times 4$, $\times 8$ and $\times 16$) through deconvolutional layers, while a third model segment combines the feature maps at each scale to extract more spatial details and temporal dependencies. The output of this segment is then fed to a channel attention mechanism and a spatial attention mechanism in sequence. The final results respect the spatial and temporal changes of the data but are significantly blurred.

A number of studies have also focused on the application of GANs to the CT-A spatiotemporal fusion problem. For example, Shang et al. [91] (*GASTFN*) propose an adversarial version of the DCSTFN model, where an EDSR-like Generator performs the spatial enhancement task. Experiments showed that the proposed model yields sharper and more accurate results compared to the non-adversarial DCSTFN. Bouabid et al. [162] propose a model similar to the popular *pix2pix* GAN [163] which comprises a conditional GAN with a U-Net architecture for the Generator and a PatchGAN architecture for the Discriminator.

Chen et al. [155] (*CycleGAN-STF*) employ a cycle GAN architecture [164] in order to enhance the traditional FSDAF algorithm [165]. The main framework consists of the following four stages: (I) Generation: a cycle GAN takes as input the HSLT image pair (F_{t-1}, F_{t+1}) and produces a F_t^{GAN} in the output. The GAN produces a single image each time, so an iterative generation scheme is introduced in order to generate multiple in-between images. (II) Selection: a single F_t^{GAN} image is selected based on mutual information metrics of the HSLT and LSHT images. (III) Enhancement: the discrete wavelet transform is used to enhance the quality of the selected image, borrowing information from C_t (IV) Fusion: The result of the previous steps along with C_t and C_{t-1} are inserted in the FSDAF algorithm to obtain the final prediction. The model was only compared with traditional non-DL algorithms. Experiments showed that CycleGAN-STF outperformed the other approaches in preserving spatial details but resulted in loss of spectral information.

Zhang et al. [156] (*STFGAN*) propose a cascade of two SRGAN-like structures which learn to produce an HR Landsat image for a target date t_2 based on Landsat-5/7 data from dates t_1 and t_3 and MODIS data from dates t_1, t_2 and t_3. The first GAN takes as input the two Landsat and all the corresponding MODIS images and produces an intermediate Landsat image, $\hat{F}_2^{int}$. Due to the limited ability of the SRGAN for spatial enhancement to such a large scaling factor ($\times 16$), this image is far from optimal. Therefore, a second GAN is used which takes as input the Landsat images along with a downsampled version of these Landsat images and the intermediate $\hat{F}_2^{int}$, to produce the final F_2 image.

A different approach is followed by Tan et al. [166] (*GAN-STFM*) who propose a conditional GAN architecture for downscaling MODIS images with a Landsat reference. The Generator follows a U-Net architecture and the input is the coarse MODIS image at the prediction date t, C_t, and a fine Landsat image at a different date t^* arbitrarily close to the target, F_{t^*}. Similarly, the Discriminator takes as input a concatenation of either the coarse C_t and the corresponding ground truth F_t or the coarse C_t and the predicted F_t^{pred} in order to perform a fake/real classification. All convolutional blocks in both networks are replaced by custom residual blocks with Switchable Normalization [167] in the Generator and Spectral Normalization [168] in the Discriminator. The authors further propose the use of a multiscale Discriminator where all inputs are additionally downsampled with factors /2 and /4 and are used to train three different discriminators with similar architectures at different scales. The proposed method is compared with non-DL approaches and EDCSTFN, showing the superiority of the random Landsat reference selection against the temporal proximity imposed by STF in terms of computational cost, without compromising the downscaling quality.

Different DL approaches for blending Landsat-8 with Formosat-2 (8m) images to increase the number of cloud-free observations have been studied by Teo et al. [169]. First, Landsat images were resampled to 8m and then blended with the rest via a simple STARFM algorithm. Secondly, pairs of Formosat and Landsat images obtained on the same date were fed to a VDSR model which learnt to predict the residual between the LR and HR features. This prediction was then used to estimate the final spatially enhanced image. The last two experiments, nicknamed *Blend-then-SR* and *SR-then-blend*, tested the hybrid approaches of applying STARFM for blending and then VDSR for downscaling, or applying VDSR for downscaling and then STARFM for blending, respectively. The study concludes that the SR-then-blend approach yielded best results overall which implies that spatially enhancing the LR images before fusion can reduce the variation between the two image sets.

B. Context-assisted (C-A)

A C-A approach which aims to integrate temporal change to an end-to-end model is proposed by Jia et al. [150] (*DL-SDFM*). They design a two-stream CNN, with one branch (M_1) learning a temporal change-based mapping and the other (M_2) learning a spatial change-based mapping. Each branch consists of inception modules containing dilated convolutions with different dilation factors and the overall model is trained with two types of input data: in a time-forward pass the time

Model	Input assistance	Time Difference images	Prior dates only	CV model	Architecture	Code available / # params
STFDCNN [154]	CT-A	no	no	SRCNN	CNN	no / -
VDCNSTF [158]	CT-A	no	no	VDSR	CNN	no / -
StfNet [147]	CT-A	yes	no	-	CNN	no / -
DCSTFN [159]	CT-A	no	yes	-	CNN	yes / 409k
EDCSTFN [161]	CT-A	no	yes	-	CNN	yes / 282k
DMNet [148]	CT-A	yes	yes	-	CNN	no / 327k
AMNet [149]	CT-A	yes	yes	-	CNN	no / -
GASTFN [91]	CT-A	yes	no	EDSR	GAN	no / -
Bouabid et al. [162]	CT-A	no	yes	-	GAN	yes / -
CycleGAN-STF [155]	CT-A	no	no	-	GAN	no / -
STFGAN [156]	CT-A	no	no	SRGAN	GAN	no / -
GAN-STFM [166]	CT-A	no	yes	-	GAN	yes / 578k + 3.6m
Teo et al. [169]	CT-A	no	yes	VDSR	GAN	no / -
DL-SDFM [150]	C-A	yes	no	-	CNN	no / -
HDLSFM [170]	C-A	no	yes	LapSRN	CNN	no / -
STF3DCNN [152]	C-A	yes	no	-	CNN	no / -
BiaSTF [153]	C-A	yes	no	-	CNN	no / -

TABLE IV: Summary of the state-of-the-art Deep Learning models for spatiotemporal fusion for image downscaling in Remote Sensing. CV model refers to the models presented in Table II.

differences are computed forward in time, whereas in a time-backward pass they are computed backwards in time. In the former case, the learnt mappings are: $M_1 : (C_{13}, F_1) \rightarrow \hat{F}_3^1$ and $M_2 : (C_3, F_1 - C_1) \rightarrow \hat{F}_3^2$, and in the latter case they are: $M_1\prime : (C_{31}, F_3) \rightarrow \hat{F}_1^1\prime$ and $M_2\prime : (C_1, F_3 - C_3) \rightarrow \hat{F}_1^2\prime$. All outputs are supervised by the given labels. Then, in the prediction phase the model produces the following mappings: $M_1 : (C_{12}, F_1) \rightarrow \hat{F}_2^1$ and $M_2 : (C_2, F_1 - C_1) \rightarrow \hat{F}_2^2$ for the forward pass and $M_1\prime : (C_{32}, F_3) \rightarrow \hat{F}_2^1\prime$ and $M_2\prime : (C_2, F_3 - C_3) \rightarrow \hat{F}_2^2\prime$ for the backward pass. Fig. 25 presents the entire pipeline. The authors compared DL-SDFM with two traditional approaches and the DL-based STFDCNN model and argue that their method manages to capture phenological change and achieve results closer to the ground truth but slightly inferior to STFDCNN visually.

Jia et al. [170] (*HDLSFM*) propose a hybrid approach which involves a LapSRN model for spatial downscaling and a linear model for extracting temporal changes. To alleviate the problem of large radiation differences between LR and HR images, the LapSRN is trained on MODIS-Landsat pairs to produce an intermediate output at $\times 2$ scale following the progressive upsampling scheme. During inference, temporal changes are captured by a linear model which extracts information from both F_1 and the intermediate output of LapSRN for images C_1 and C_2. In the final donwscaled image, considerable blurring was observed in heterogeneous areas of the underlying scene.

Downscaling a time series of MODIS images based on Landsat observations captured on sparser dates is addressed by Peng et al. [152] (*STF3DCNN*). The proposed approach takes as input the time difference MODIS images between each consecutive pair of dates and a 3D CNN model is trained to produce the corresponding time difference Landsat images of the in-between dates. The output is added to the original Landsat series to produce the final prediction. The presented method manages to capture abrupt changes in the observed scene.

A novel idea was presented in [153] (*BiaSTF*), where it is argued that when different sensors capture changes with differences in spectral and spatial viewpoints, a considerable bias between these sensors is introduced. No previously published method accounts for this bias so the authors propose a pipeline with two CNNs, one for learning the spectral/spatial changes and the other for learning the sensor bias. Both networks are trained with a separate MSE loss and take as input pairs of MODIS and Landsat observations. The final prediction is obtained by summing the output of the two networks along with the initial HSLT image. The results showed that this inclusion of the sensor bias lets the model converge to a lower minimum and its predictions exhibit fewer spatial and spectral distortions.

Concluding, the studies presented in this section provide a variety of methods for tackling the spatiotemporal variation of the observed landscape. The lack of a common benchmark dataset again renders the direct comparison of all methods infeasible but certain useful characteristics can be discerned. First, models such as EDCSTFN, GASTFN and GAN-STFM require a minimal number of input images, thus facilitating the downscaling task in areas with severe cloud contamination. Among these approaches, GAN-STFM has the additional advantage of using fine images at arbitrary dates prior to the target date, which provides an extra level of freedom concerning the selection of images for training and/or inference. Secondly, EDCSTFN, DMNet, STF3DCNN and BiaSTF employ simple architectures with a limited number of trainable parameters, which makes them ideal candidates for quick experimentation and testing. Finally, considering the spectral correlation between the different bands enables the model to exploit complementary information in order to better uncover land cover and phenological changes. The models accepting multi-band input are EDCSTFN, GASTFN, STFGAN, GAN-STFM, DL-SDFM and STF3DCNN.

IX. Super-resolution

Super-resolution is a broad family of methods which aim to enhance the spatial resolution of an image without the need to blend information from auxiliary sources either in the spectral or the temporal dimension. For better assessment, they can be categorized into *Single image Super-resolution*

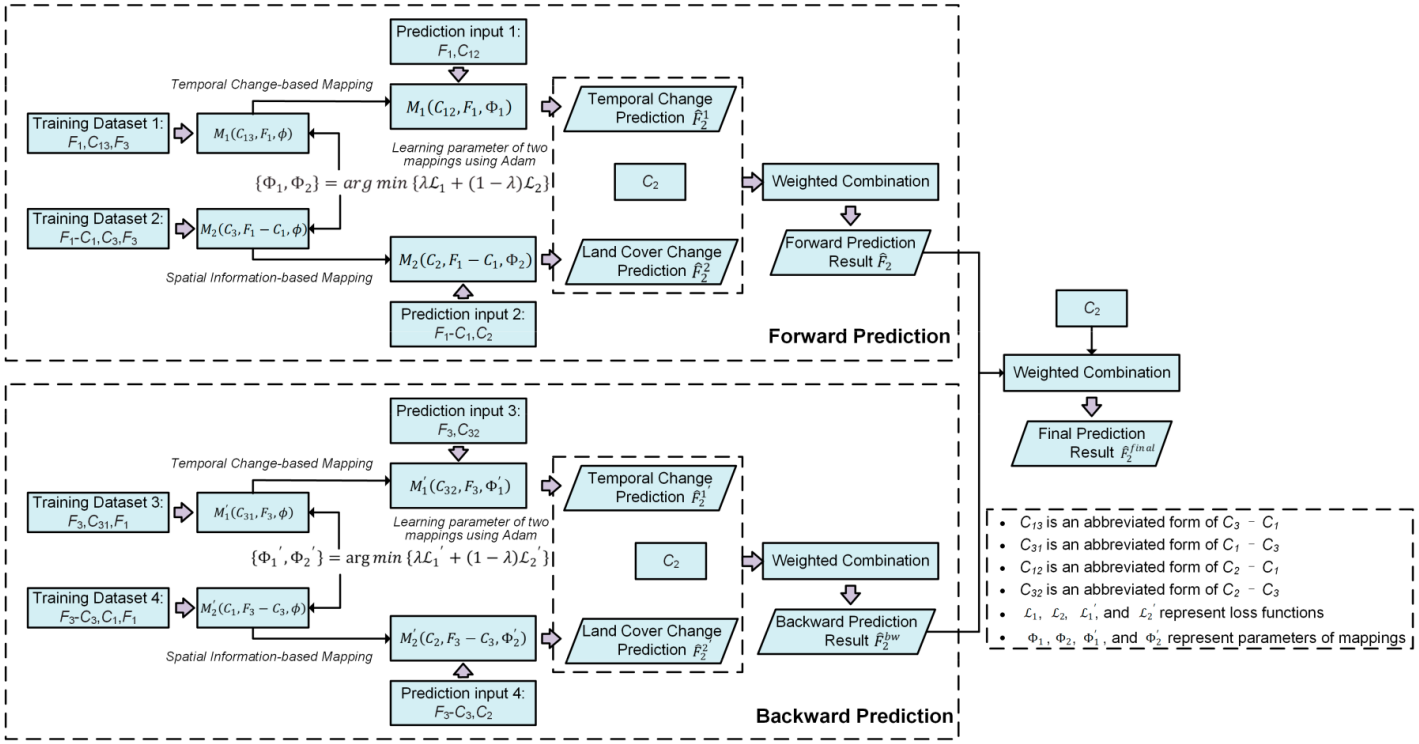


Fig. 25: The DL-SDFM pipeline. Image taken from [150].

(*SISR*), *Multiple image Super-resolution* (*MISR*) and *Reference Super-resolution* (*RefSR*). These are presented next, while in Table V we summarise the main DL models developed for SR. In Section IX-D we examine SR architectures that are specific for SAR and aerial imagery.

A. Single Image Super-resolution (SISR)

SISR aims to recover an HR version of a single LR input image. However, lost pixel information in the LR image can never be fully retrieved but only hallucinated, which means that multiple possible HR images can be constructed from one LR source. This renders the SISR problem mathematically ill-posed and non-invertible, but it often comprises the only viable approach when only a single LR input is available. Therefore, several attempts have been made to employ DL techniques in the SISR domain for Remote Sensing.

Multi-scale approaches

Lei et al. [171] (*LGCNet* - Fig. 26) design a CNN model which combines feature maps produced by previous layers in order to extract information at different scales and level of detail. The model was evaluated on the UC Merced dataset and selected Gaofen-2 images, and managed to outperform traditional image enhancement methods such as bicubic interpolation and sparse coding, but showed only marginal improvements compared to other established DL models. Haut et al. [172] experiment on the same data with a residual model containing a sequence of convolutional layers for feature extraction and an inception module followed by upsampling layers for the final downscaling. Their method achieved performance similar to that of LGCNet.

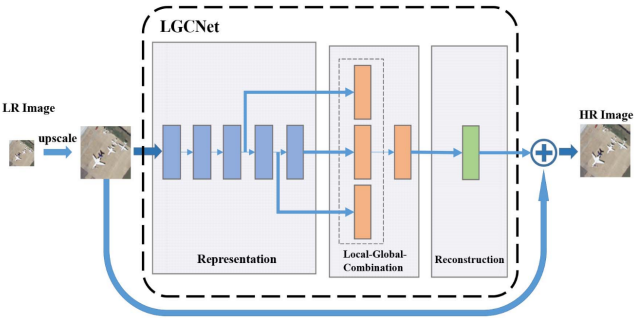


Fig. 26: A high-level overview of the LGCNet. Blue boxes represent convolutional layers followed by ReLU activation, orange boxes represent the concatenation of selected feature maps via a convolutional layer and the green box represents the last convolutional layer for the final reconstruction. Image taken from [171] (©2017 IEEE).

Lu et al. [173] (*MRNN*) propose a pre-upsampling architecture with parallel convolutional layers and design a network with three parallel branches containing residual blocks of different convolutional kernel sizes. Each branch is initially trained separately with interpolated versions of the original LR image varying in size and then all branches are combined for the final image reconstruction and finetuned in an end-to-end setting. Experimental results show promising improvements over other state-of-the-art DL methods, especially for larger scaling factors. In another multiscale approach [174], Xu et al. employ a U-Net resembling architecture, adding a module with sequential dilated convolutions at the bottleneck section,

a global residual connection and pixel shuffle operations before the final output. The dilated convolutions have different dilation rates, allowing the model to extract information using different receptive fields and scales.

Multi-task learning

In their study, Yan et al. [175] (*MSF*) exploit a multi-task learning procedure in order to improve the generalization of the underlying network to different degradation models. According to the standard approach, an image is downsampled by convolving with a Gaussian blur kernel, applying bicubic interpolation and then adding some noise. The authors argue that a model trained on images degraded by a single Gaussian kernel may perform quite well on such images but fail to generalize to different kernels. Therefore they propose a model trained in a multi-task setting where each task represents a separate Gaussian kernel and is learnt by a dedicated CNN.

Additional post-processing

A study by Qin et al. [176] (*DGANet-ISE*) presented a custom post-processing pipeline for the improvement of the output of an SR model. Their architecture is heavily based on EDSR (Section V) and is trained with a custom loss function which additionally considers the gradient similarity between prediction and target. The model's output is then iteratively improved via a proposed *Image-Specific Enhancement* (ISE) algorithm which back-projects the error between the SR output and the LR input image and accordingly updates the prediction. This algorithm alleviates the possible variation between the training and testing datasets which might occur from different sensing platforms, light conditions, etc.

Different sources for I/O

Contrary to most approaches in this category that exploit Wald's protocol, a number of methods have been proposed which utilize different sources for the input and output. Galar et al. [177] (*S2PS*) propose the use of PlanetScope images as target in order to downscale the four Sentinel-2 10m bands. They train a modified version of the EDSR separately for each of the NIR and Red bands accounting also for the style transfer loss [178] between prediction and target. Pouliot et al. [179] (*DCR-SRCNN*) use Sentinel-2 observations to downscale the corresponding Landsat-8 and Landsat-5 images from three regions in Canada through an SRCNN architecture with denser residual connections trained to predict a single band. Landsat-Sentinel training pairs were selected based on a minimum change vector across time and the authors noted that better results were obtained for Sentinel observations closest to the prediction date due to the dynamic behaviour of land cover types such as croplands. Finally, Collins et al. [180] apply an SRCNN on the two Resourcesat sensors. The constellation of Indian Resourcesat satellites (1/2) provide multitemporal and multiresolution observations in the same spectra with coincident captures enabling the use of SISR techniques. Both satellites carry the sensors LISS III, which captures information in the Green, Red, NIR and SWIR bands with 24m spatial resolution and a 24-day revisit cycle, and AWiFS, which captures the same bands with 56m spatial resolution and a 5-day revisit cycle. The authors used a training set with coincident images from the two satellites in order to downscale the AWiFS data to match the spatial resolution of the corresponding LISS III data. The model was evaluated only against simple baselines and produced better PSNR and SSIM scores.

Different degradations

Sheikholeslami et al. [181] (*EUSR*) employ a dense network with a bilinear upsampling layer for the reconstruction. Contrary to the majority of studies in literature, the authors downsample the initial dataset via the Lanczos3 kernel [182] to be used in the model's training, following Wald's protocol. The resulting image is then downsampled again with the same kernel and compared with the initial LR image in a PSNR-based loss function. Experiments show that results are similar to other methods, but the proposed approach prevails when larger input images are used.

Arguing that most published studies following the Wald's protocol produce synthetic LR images through a specific distortion model and develop methods which focus solely on the enhancement of such LR images, Zhang et al. [183] propose an unsupervised model to handle multi-degradation schemes. In particular, their approach involves a post-upsampling Generator network which produces an SR image and a Degrader network which distorts this SR result. The final loss function is the MSE between the degraded image and the original LR, thus alleviating the need to compare the result to an HR ground truth. For the Degrader the authors adopt the same pipeline as in [184]. Results on the UC Merced, NWPU-RESIS45 datasets (Section X) and Jilin-1 satellite images showed that the proposed method outperformed state-of-the-art DL approaches when distortions other that bicubic interpolation were used for the LR input. It managed to produce results closer to the ground truth and retain edges and object shapes more correctly.

Wavelets

A large family of traditional non-DL approaches perform the super-resolution task in the frequency domain, usually through the Wavelet Transform. The general pipeline is to analyze the image into a number of frequency components, separately enhance the components and then apply the inverse transformation to obtain the final SR image. A number of DL methods have been proposed ([185] (*WTCRR*), [186] (*DWTSR*), [187] (*RRDGAN*)) which use the 2D Discrete Wavelet Transform and design a DL network to undertake the task of component enhancement. In WTCRR residual blocks of a ResNet are replaced with recurrent blocks in order to reduce the number of parameters and increase the network depth without overfitting. On the other hand, DWTSR uses a simpler architecture but employs the 2D Stationary Wavelet Transform along with the 2D Discrete Wavelet Transform for richer features. Finally, RRDGAN enhances the ESRGAN architecture with denser connections, a Relativistic Discriminator and a Total Variation loss [188] in order to separately enhance the four components of the Haar Wavelet Transform. All of the aforementioned studies achieve good results indicating that the frequency domain may offer more useful information to a DL model and is thus worth exploring further.

Attention mechanism

Several studies also employ attention mechanisms in order to aid the downscaling process and help the model focus on the high-frequency details of the image. For example,

Dong et al. [189] (*MPSR*) and Gu et al. [190] (*DRSEN*) design architectures with various residual connectivity schemes and channel attention modules similar to the *Squeeze-and-Excitation* blocks proposed in [66]. Haut et al. [191] utilize the RCAB attention module [89] inside convolutional blocks with residual connections at multiple levels. RCAB is also adopted by Zhang et al. [192] (*MSAN*, *SAMSAN*) who additionally propose a scene-adaptive learning framework where a separate model is finetuned on each possible scene depicted in an RS image, and Dong et al. [193] (*DSSR*) who also present a chain learning strategy where a $\times k^2$ model is based on a pretrained $\times k$ model. A similar architecture to DSSR is proposed by Wang et al. [194] (*AMFFN*) where both Squeeze-and-Excitation and RCAB modules are applied on a multiscale feature extraction framework containing parallel convolutions with varying kernel sizes. Lei et al. [195] (*IRAN*) propose a network comprising a series of inception modules followed by channel (Squeeze-and-Excitation) and spatial attention mechanisms. Similarly, Wang et al. [196] (*NLASR*) design a model with non-local blocks [197] which follows the iterative up- and down-sampling scheme with channel and spatial attention modules. Finally, based on the popular EDSR architecture, Peng et al. [198] (*PGCNN*) propose a gated residual block which encourages the model to focus on high-frequency details, whereas Lei et al. [199] (*HSENet*) employ custom attention modules which aim to discover information recurring in multiple scales inside the image. All of the aforementioned studies show that the inclusion of such attention mechanisms boosts the model's performance and helps achieve a sharper downscaled result closer to the HR ground truth.

Recursion

Chang et al. [200] (*BCLSR*) present a novel approach by employing a recursive framework on images obtained from the GaoFen-2 satellite. Their model comprises multiple densely connected convolutional blocks which share their parameters and feed their outputs to a BiConvLSTM layer. The output is then downscaled via a sub-pixel convolution. Results showed that this method outperformed several established DL models and produced sharper results without losing substantial high-frequency details.

Generative Networks

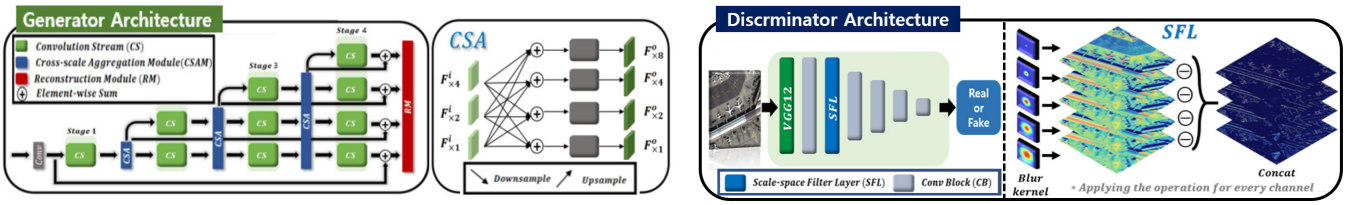
A multitude of studies have also explored the adaptation of GAN models for SR. In an interesting approach, Lei et al. [201] (*CDGAN*) present the *"discrimination-ambiguity" problem*, which states that Remote Sensing images contain more low-frequency components than natural images thus impairing the Discriminator's ability to decide whether a given input is real or fake. To tackle this issue, they propose a "Coupled Discriminator" that takes as input both the predicted SR image and its corresponding HR ground truth shuffled by a random gate, and is then tasked to decide whether the input constitutes a real-fake pair (1) or a fake-real pair (0). The Generator architecture is based on ESRGAN. The model competed against a number of DL methods on the UCMerced and WHU-RS19 datasets (Section X) as well as selected GaoFen-2 images and produced less blurry results and with fewer artifacts.

A number of studies have also proposed minor adjustments of popular SR architectures to fit the needs of the RS domain. For example, Ma et al. [202] (*DRGAN*) utilize an RDN-like architecture for the Generator with sub-pixel convolution for downscaling and a VGG loss function. Their model was evaluated on the NWPU-RESISC45 dataset (Section X) and several other computer vision benchmarks and achieved sharper images with cleaner object boundaries as compared with other state-of-the-art DL methods. Salgueiro Romero et al. [203] (*RS-ESRGAN*) adapt the ESRGAN model in a pre-upsampling framework and train the Generator in three stages; first, it is trained on a set of WorldView images only, then finetuned on pairs of WorldView and Sentinel-2 images and finally trained in an adversarial manner with WorldView and Sentinel-2 pairs. The final image is formed by a linear combination of the Generator's output trained with and without the adversarial scheme, which helps the user calibrate the Perception-Distortion Trade-off.

Multi-scale Generators

Dense and multi-level connections have also been introduced to different Generator architectures with the aim to extract more accurate representations of both small- and large-scale objects. For example, Wang et al. [204] (*udGAN*) design a novel Ultra-Dense Residual Block (UDRB) which contains parallel convolutions and additional diagonal connections, while features at each level are concatenated through a bottleneck 1×1 convolution to limit the channel size. Their study illustrates the value of this new connectivity scheme by surpassing several other established DL methods in the sharpness and quality of the produced images. Shin et al. [205] propose a multiscale Generator comprising multiple parallel streams in a pyramidal fashion, each of which is formed by a series of residual dense blocks. A reconstruction module fuses the output of all streams and produces the final SR image. Before entering the Discriminator, an HR or SR image is first fed to a pretrained VGG network and a number of intermediate feature maps are selected. A set of blurring Gaussian kernels are applied on these feature maps and the results are then fed to a Discriminator model with PatchGAN architecture. Both networks are illustrated in Fig. 27. The proposed method achieved highly better results compared to EEGAN and CDGAN, and managed to capture and recover even small-scale details in the produced images which the other techniques failed to do.

Another multiscale approach was introduced by [206] (*Enlighten-GAN*) which improves on the ESRGAN by adding an *"enlighten block"* to the Generator. This block outputs an intermediate SR image and helps the Generator learn high-frequency information in a progressive manner. The loss function has a *Self-Supervised Hierarchical Perceptual Loss* component, where an autoencoder is trained from scratch on RS images and the distance between the corresponding feature maps of the SR and HR images is computed. Finally, the authors present a novel large image tiling and batching approach for downscaling overlapping satellite image patches separately (Fig. 28). Experimental results showed that Enlighten-GAN produces sharper images with much fewer artifacts than other GAN-based methods, while at the same time retains the true hue and shapes of the objects.



(a) Generator (b) Discriminator

Fig. 27: The Generator and Discriminator for the GAN proposed in [205]. Image taken from [205] (©2020 IEEE).

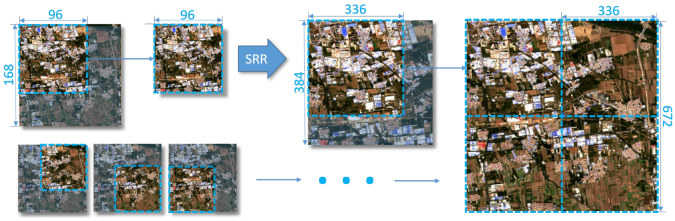


Fig. 28: An example of the clipping-and-merging method pipeline. The input image has size 168×168 and is cropped into four overlapping patches, each of size 96×96. The patches are independently downscaled by a Super-resolution algorithm (denoted SRR here) producing four 384×384 images. Half of the overlap region of each patch is then clipped, ending up with four 336×336 images which are then joined to produce the final SR prediction. Image taken from [206].

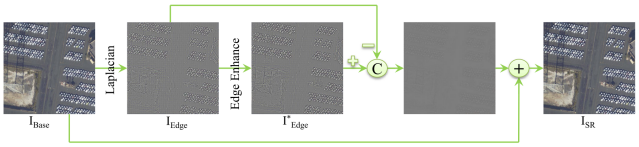


Fig. 29: The pipeline of the edge enhancement procedure for EEGAN. Image taken from [207] (©2019 IEEE).

GANs and Attention

Attempting to improve the output of an SR GAN model, multiple studies exploit attention mechanisms. Jiang et al. [207] (*EEGAN*) propose a Generator which first enhances the input and then extracts and sharpens its edges (Fig. 29). A mask branch with attention mechanism is also employed during the edge enhancement step to focus on the useful information. The model outperforms SRGAN, VDSR and SR-CNN on the Kaggle Draper Satellite Image Chronology dataset (Section X). In addition, Yu et al. [92] (*E-DBPN*) propose an extension of the popular DBPN model in a GAN setting. The Generator adopts the DBPN architecture where each up-projection unit is followed by a Squeeze-and-Excitation channel attention mechanism and the features extracted from multiple levels of the network are fused in a sequential manner. The authors pretrain the Generator with the MSE loss and then finetune it in an adversarial setting. Results showed that the proposed model produces sharper results closer to the ground truth, with fewer blurring effects and artifacts. Finally, Li et al. [208] (*SRAGAN*) design a complex GAN with local and global channel and spatial attention modules both in the Generator and the Discriminator network in order to capture short- as well as long-range dependencies between pixels. Several experiments proved the superiority of the proposed model, especially in higher scaling factors.

B. Multiple Image Super-resolution (MISR)

In a MISR setting, a model takes as input multiple LR images of the same scene taken from different angles/viewpoints and aims to synthesize a single HR image. The main advantage of this approach is the fact that the minor geometric displacements and distortions among the LR images offer a richer source of information for a candidate downscaling model than any individual LR image alone, thus usually obtaining better results than SISR. Also, a key difference from spatiotemporal or spatiospectral fusion is the fact that both LR and HR images contain information on the same spectra, whereas their acquisition times are never coincident.

Such a MISR method is described in [209] (*EvoNet*) where a number of shifted LR images are used to produce a single HR image. In the proposed model, each LR image is independently enhanced through a ResNet and then the individual SR outputs are co-registered and fed to the Evolutionary Image Model (EvoIM) algorithm [210] which constructs the final output. One experiment employed artificially shifting and downsampling images for the creation of training data, whereas another experiment utilized a number of Sentinel-2 images in order to produce a SPOT-like HR output downscaled by a $\times 2$ factor. EvoNet achieved higher results against several traditional SISR and MISR approaches in both distortion and perceptual quality metrics at the expense of higher computational time. On a qualitative basis, EvoNet produced results similar to SRGAN but less blurry and with more artifacts.

A common source of data for the MISR problem is the PROBA-V satellite, which is able to capture multispectral images in 300m spatial resolution every day, and 100m spatial resolution every 5 days. Since both observations lie in the same spectral bands and are never paired on the same date, a number of studies exploit the LR images for the construction of the corresponding HR image in a MISR approach, with the authors in [211] proposing a PROBA-V dataset exclusively for this problem setting. They also design a simple 4-layer CNN for benchmarking and propose a custom metric which takes into account spatial displacements between the prediction and the ground truth.

In their study [212] (*DeepSUM* - Fig. 30), Molini et al.

design a network that downscales a NIR or Red band of PROBA-V data. The model takes as input a single image and performs feature extraction. All extracted features are then co-registered and fused in the feature space. Before the final fusion, a Mutual Inpainting process is employed in order to replace unreliable pixels in a feature map (such as clouds, shadows, etc) with values taken from corresponding feature maps of other images. The authors claim that end to end training of this model leads to many local optima so they choose to train each step separately. Evaluated against other MISR methods, the proposed model achieved better results and sharper output scoring first in the PROBA-V Super-resolution challenge issued by the European Space Agency [211]. In a subsequent publication [213] (*DeepSUM++*), the authors extend the feature extraction part with graph convolutional operations in order to exploit non-local correlations among pixels.

Another popular method for the PROBA-V dataset was proposed by Deudon et al. [214] (*HighRes-Net*). The authors argue that the set of LR images contain redundant low-frequency information so they select the median LR image as reference and pair each LR image with this. Then they train a model to extract a shared representation for each pair which allows it to highlight differences in multiple LR views and focus on the important high-frequency features. The extracted embeddings are then recursively fused using a mechanism with shared weights and the common representation is downscaled to predict the final SR image. Another model called *ShiftNet* is also proposed which registers the SR with the target HR image in order to properly calculate the loss function. Without such registration, the model outputs blurry results to compensate for the misalignment between the SR and the target HR. The architecture follows HomographyNet proposed by [215] but is trained cooperatively with HighRes-Net in an end to end setting and achieves results similar to DeepSUM.

Rifat Arefin et al. [216] (*MISR-GRU* - Fig. 31) choose to tackle the MISR problem in a time-series setting by regarding the LR input images as a temporal sequence. At each time step, their model takes as input one LR image and the median of all LR inputs, co-registers them and produces a unified feature map. The output of this stage is then fed to a stack of ConvGRU modules [217] and the output is globally averaged across the temporal dimension and downscaled. The final prediction is also registered following the *ShiftNet* strategy introduced by [214] and the loss function is a custom negative PSNR which involves a brightness bias. MISR-GRU achieved the highest score compared with FSRCNN, SRResNet, DeepSUM and HighRes-Net, and the authors concluded that the proposed model's accuracy was highly affected by the number of LR inputs and the amount of occlusion observed in the LR images.

A more complex model was proposed by Salvetti et al. [218] (*RAMS*) which employs 3D convolutions and attention mechanisms on both temporal and spatial domains in order to downscale a single band of PROBA-V data. The 3D convolutions are able to assess the inter-relations across the different dimensions, whereas the attention modules focus on the similarity between the input LR images (temporal attention) or the useful high-frequency details to retain on the spatial domain of the LR feature maps (feature attention). The model performed quite similarly to MISR methods such as HighRes-Net and DeepSUM. The authors also experimented with temporal self-ensembling strategy and observed a significant increase in the output accuracy but at the expense of computational speed.

C. Reference Super-resolution (RefSR)

In RefSR the input of the model is accompanied by an auxiliary (reference) image which provides additional information to assist the downscaling process. A number of studies have explored using features extracted from the original data as reference input and hereafter we highlight a selection of the most promising attempts in the literature.

An adversarial RefSR approach is proposed by a series of publications ([219], [220], [221]) that focus on the saliency information of the input images. In [219] (*SD-GAN* - Fig. 32) the authors discriminate the highly salient areas of an image as foreground and the less salient as background, and they argue that by applying different reconstruction principles based on the level of saliency the GAN will be able to produce more realistic images stripped of hallucinated pseudotextures. For that reason, they propose the extraction of a saliency map for each input image through a Weakly Supervised Learning scheme [222] and design a Generator which takes as input the LR image concatenated with its corresponding saliency map along the channel dimension and produces an SR output. Additionally, a paired Discriminator is used for the adversarial learning, one for the salient (foreground) and one for the non-salient (background) areas. Experimentation on GeoEye-1 panchromatic images showed that SD-GAN outperformed other DL approaches such as SRCNN, ESPCN, VDSR and SRGAN. Qualitative analysis proved that it managed to produce less pseudotextures in salient areas than SRGAN. Extending their previous work in a subsequent study [220] (*SG-FBGAN*), the same research group proposes a recursive Generator architecture and a triplet of Discriminators. More precisely, the Generator performs parallel processing of salient and non-salient information in a recursive fashion and the final output of the network is the output of the last iteration. Similarly to SD-GAN, a salient area Discriminator and a non-salient area Discriminator are employed, along with a global Discriminator which takes as input the SR or HR image and learns to classify them. Then the outputs of all Discriminators over all iterations is averaged to calculate an overall Discriminator loss. When compared with VDSR, RDN, EDSR, SRFBN, SRGAN, SD-GAN and D-DBPN, the proposed method achieved superior results, producing more realistic images with fewer pseudotextures and artifacts. The authors also experiment with Curriculum Learning and more complex degradation schemes and the results were superior to the other DL approaches, especially for higher scaling factors ($\times 3$ and $\times 4$).

To summarize the above analysis, there are two main approaches a researcher can take depending on the number of available images in the dataset at hand. When only a single LR image can be acquired per occasion, SISR and

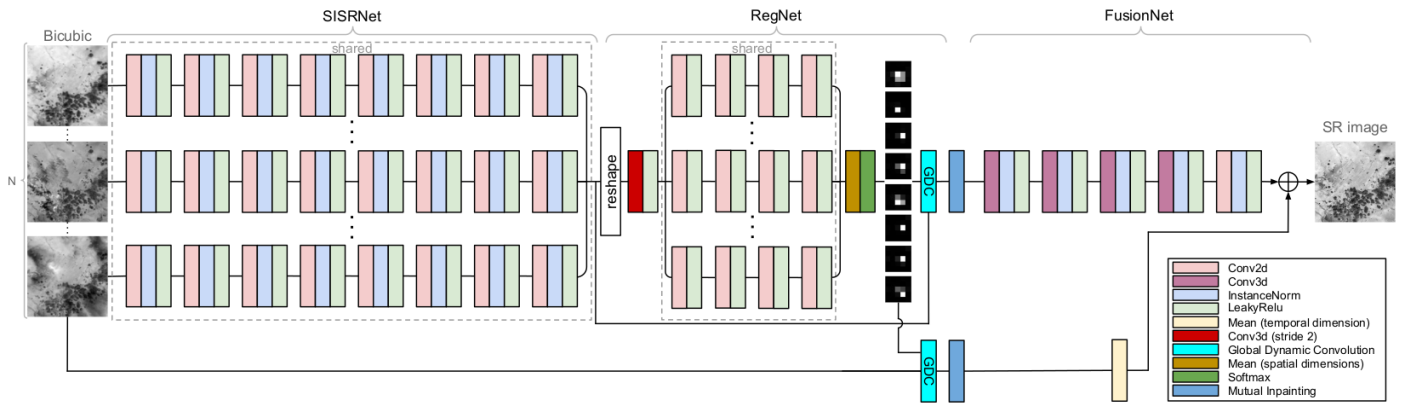


Fig. 30: Overview of the DeepSUM model. *SISRNet* performs feature extraction, *RegNet* feature registration and *FusionNet* the final feature fusion and reconstruction. The *Global Dynamic Convolution* (GDC) is a convolution between an image and the corresponding learnt filter for image registration. Image taken from [212] (©2020 IEEE).

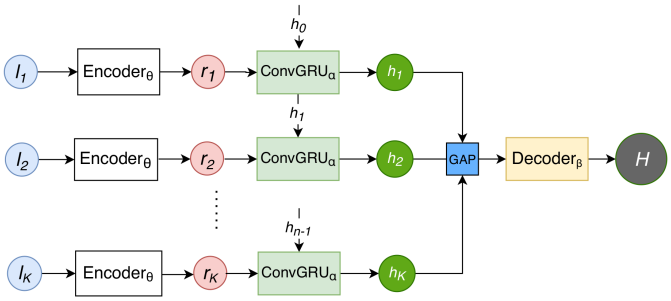


Fig. 31: Overview of the MISR-GRU model. l_i is the i^{th} LR input image, H is the predicted downscaled image, h_i is the i^{th} hidden state of the ConvGRU layer and GAP is the Global Average Pooling layer. In the original paper, the Encoder comprises two convolutional layers and two residual blocks (each with two convolutional layers and Parametric ReLU activation), while the Decoder consists of a deconvolutional layer and two 1×1 convolutional layers. Image taken from [216] (©2020 IEEE).

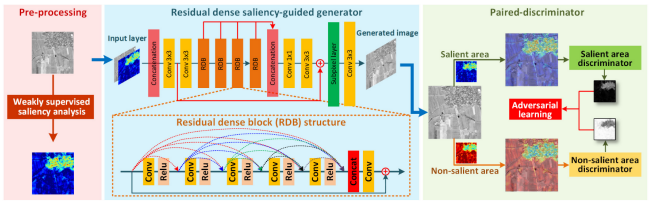


Fig. 32: The SD-GAN model. Image taken from [219] (©2020 IEEE).

RefSR methods can be applied. In particular, several of the aforementioned models offer a robust solution to the downscaling problem, proving that certain mechanisms and modules can further boost performance and achieve sharp results. For example, attention mechanisms (e.g. MPSR, DRSEN, DSSR, Haut et al. II, NLASR) can always assist the discovery and preservation of high frequency components, whereas multi-scale feature extraction structures (e.g. NLASR, Shin et al.) can unravel non-local correlations inside the image and expand the receptive field of basic convolutional layers. Furthermore, a number of novel techniques seem to leverage the efficiency of the underlying model, e.g. the diagonal connectivity scheme proposed in udGAN or the clipping-and-merging post-processing technique and the autoencoder loss proposed in Enlighten-GAN. Finally, certain methods (EUSR, DWTSR, DRSEN, DSSR, DGANet-ISE, NLASR, Shin et al., SG-FBGAN) manage to perform better at larger scaling factors whereas Zhang et al. provide an interesting candidate when different distortions have taken place during the LR image acquisition. Unfortunately, up to this point in time only a handful of RefSR methods have been developed and none seems to match the efficiency and robustness of the SISR domain.

On the other hand, when multiple LR images can be obtained for each training/testing sample, then MISR models can be employed. In this family of methods, MISR-GRU and RAMS in particular seem to prevail in terms of both the resulting image quality and the number of trainable parameters. It is worth noting that a common challenge faced by all MISR approaches is the co-registration of the input LR images which is handled differently by each proposed model, either inside the network or as a separate pre-processing step in the pipeline. In addition, this co-registration may incur minor shifts in the output, which in turn can potentially affect the computation of the loss function during training and encourage a blurry result. This phenomenon has been successfully handled through the ShiftNet module which has been proposed in HighRes-Net and been subsequently used in other studies. Finally, it is again proven that attention mechanisms enhance the downscaled output and also that the number and clarity of the input LR images can greatly affect the final result.

D. SR for SAR and aerial imagery

Synthetic Aperture Radar (SAR)

Most of the SAR spatial resolution enhancement techniques related to deep neural networks use the SISR approach which makes the data collection, processing and experimentation

Model	SR type	Description/novelty	CV model	Building blocks	Upsampling framework	Architecture	Code available / # params
LGCNet [171]	SISR	Multi-scale approach, features from different layers	-	Residual learning	pre-upsampling	CNN	no / -
Haut et al. [172]	SISR	Multi-scale approach with inception module	-	Residual learning, sub-pixel convolution	post-upsampling	CNN	no / -
MRNN [173]	SISR	Multi-scale approach, parallel feature extraction from different scales of the LR input	-	Residual learning	pre-upsampling	CNN	no / -
Xu et al. [174]	SISR	Multi-scale approach, U-Net model with dilation module at the bottleneck	-	Residual learning, sub-pixel convolution	post-upsampling	CNN	no / -
MSF [175]	SISR	Multi-task learning, different model for each Gaussian kernel	-	Residual learning	pre-upsampling	CNN	no / -
DGANet-ISE [176]	SISR	Post-processing algorithm and gradient loss term	EDSR	Residual learning, sub-pixel convolution	post-upsampling	CNN	no / -
S2PS [177]	SISR	Downscaling of Sentinel-2 images using PlanetScope as target	EDSR	Residual learning, sub-pixel convolution	post-upsampling	CNN	no / -
DCR-SRCNN [179]	SISR	Downscaling of Landsat-5/8 images using Sentinel-2 as target	SRCNN	Residual learning	pre-upsampling	CNN	no / 993k
Collins et al. [180]	SISR	Downscaling of coarser AWiFS images using sharper LISS III images from Resourcesat	SRCNN	-	pre-upsampling	CNN	no / -
Zhang et al. [183]	SISR	Unsupervised model which learns multiple image degradations	-	Residual learning, bilinear upsampling layers	post-upsampling	GAN	no / -
EUSR [181]	SISR	Dense network, the resulting image is downsampled and compared with LR input	-	Bilinear upsampling layers	post-upsampling	CNN	no / -
WTCRR [185]	SISR	Approach assisted by Discrete Wavelet transform, recurrent blocks are used	DRRN	Residual learning	pre-upsampling	CNN	no / -
DWTSR [186]	SISR	Approach assisted by Discrete Wavelet transform and Stationary Wavelet transform	-	Residual learning	pre-upsampling	CNN	no / -
RRDGAN [187]	SISR	Approach assisted by Discrete Wavelet Transform and the TV loss function	ESRGAN	Residual learning, sub-pixel convolution	post-upsampling	GAN	no / -
MPSR [189]	SISR	Multi-scale approach with residual connections and channel attention	-	Residual learning, sub-pixel convolution, attention mechanism	post-upsampling	CNN	no / -
DRSEN [190]	SISR	Approach with channel attention	EDSR	Residual learning, sub-pixel convolution, attention mechanism	post-upsampling	CNN	no / 8.6m
Haut et al. II [191]	SISR	Approach with channel attention	-	Residual learning, sub-pixel convolution, attention mechanism	post-upsampling	CNN	no / -
MSAN, SAMSAN [192]	SISR	Approach with channel attention and scene-adaptive learning	WDSR	Residual learning, sub-pixel convolution, attention mechanism	post-upsampling	CNN	no / -
DSSR [193]	SISR	Approach with channel attention and chain training	WDSR	Residual learning, sub-pixel convolution, attention mechanism	post-upsampling	CNN	no / 9.1m
AMFFN [194]	SISR	Multi-scale approach with channel attention	-	Residual learning, sub-pixel convolution, attention mechanism	post-upsampling	CNN	no / -
IRAN [195]	SISR	Approach with inception modules and both channel and spatial attention	-	Residual learning, sub-pixel convolution, attention mechanism	post-upsampling	CNN	no / 1.88m
NLASR [196]	SISR	Multi-scale approach with non-local modules and both channel and spatial attention	-	Residual learning, sub-pixel convolution, attention mechanism	iterative up- and down-sampling	CNN	no / 10.7m
PGCNN [198]	SISR	Approach with channel attention	EDSR	Residual learning, sub-pixel convolution, attention mechanism	post-upsampling	CNN	no / 1.44m
HSENet [199]	SISR	Attention for multi-scale recurring features	-	Residual learning, sub-pixel convolution, attention mechanism	post-upsampling	CNN	yes / -
BCLSR [200]	SISR	Recurrent convolutional model	-	Residual learning, sub-pixel convolution	post-upsampling	CNN	yes / 170k
CDGAN [201]	SISR	Coupled Discriminator	ESRGAN	Residual learning, sub-pixel convolution	post-upsampling	GAN	no / 1.4m
DRGAN [202]	SISR	RDN-like Generator	RDN	Residual learning, sub-pixel convolution	post-upsampling	GAN	no / -

RS-ESRGAN [203]	SISR	Multiple training phases on different datasets	ESRGAN	Residual learning	pre-upsampling	GAN	yes / -
udGAN [204]	SISR	Multi-scale Generator with ultra-dense residual blocks	-	Residual learning, sub-pixel convolution	post-upsampling	GAN	no / 2.4m
Shin et al. [205]	SISR	Multi-scale Generator with pyramidal structure, Discriminator with difference of Gaussian kernels on feature maps	-	Residual learning, sub-pixel convolution	progressive upsampling	GAN	no / -
Enlighten-GAN [206]	SISR	Multi-scale Generator with intermediate output, clipping-and-merging method	ESRGAN	Residual learning, sub-pixel convolution	progressive upsampling	GAN	no / -
EEGAN [207]	SISR	Downscaling is assisted by edge enhancement and attention	-	Residual learning, sub-pixel convolution , attention mechanism	progressive upsampling	GAN	yes / -
E-DBPN [92]	SISR	DBPN-like Generator with channel attention on multiple layers	DBPN	Residual learning, transposed convolution, attention mechanism	iterative up- and down-upsampling	GAN	no / -
SRAGAN [208]	SISR	Generator and Discriminator with local and global channel and spatial attention modules	-	Residual learning, attention mechanism, sub-pixel convolution	post-upsampling	GAN	no / 4.8m
EvoNet [209]	MISR	Approach assisted by EvoIM algorithm	-	Residual learning	pre-upsampling	CNN	no / -
Märtens et al. [211]	MISR	Simple CNN for PROBA-V images which takes as input a concatenation of the LR images	-	-	pre-upsampling	CNN	no / 119k
DeepSUM [212]	MISR	Super-resolution of each input separately and fusion of results	-	Residual learning	pre-upsampling	CNN	yes / -
DeepSUM++ [213]	MISR	Extension of DeepSUM with graph convolutional operations	-	Residual learning	pre-upsampling	CNN	no / -
HighRes-Net [214]	MISR	Paired super-resolution of an LR image and the chosen reference LR, ShiftNet for registration of results	-	Residual learning, transposed convolution	post-upsampling	CNN	yes / 600k + 34m
MISR-GRU [216]	MISR	LR images are regarded as a time series, paired super-resolution is performed at each time step similar to HighRes-Net, ConvGRU layers and ShiftNet are used	-	Residual learning, transposed convolution	post-upsampling	CNN	yes / 900k
RAMS [218]	MISR	Approach assisted by 3D convolutions and attention modules	-	Residual learning, sub-pixel convolution, attention mechanism	post-upsampling	3D CNN	yes / 1m
SD-GAN [219]	RefSR	Saliency information used as reference	-	Residual learning, sub-pixel convolution	post-upsampling	GAN	no / -
SG-FBGAN [220]	RefSR	Extension of SD-GAN with a triplet of Discriminators and recursive layers in the Generator, curriculum learning also used	-	Residual learning, sub-pixel convolution	post-upsampling	GAN	yes / -
SR-GAN [223]	SISR	-	SRGAN	Residual learning, sub-pixel convolution	post-upsampling	GAN	no / -
NF-GAN [224]	SISR	Generator based on residual encoder-decoder, Discriminator based on ResNet50, embodies de-speckling component	-	Residual learning, transposed convolution	pre-upsampling	GAN	no / -
Di-GAN [225]	SISR	Generator based on U-Net, Discriminator based on PatchGAN-like network	-	Residual learning, transposed convolution	pre-upsampling	GAN	no / -
FSRCNN [226]	SISR	-	-	Residual learning	pre-upsampling	CNN	no / -
PSSR [227]	SISR	Learnable pre-upsampling, uses a complex structure block for complex numbers, uses residual compensation approach, uses fully polSAR	-	Residual learning, transposed convolution	pre-upsampling	CNN	no / -
WDCCN [228]	SISR	Import weighted dense connections	DRCN	Residual learning	pre-upsampling	CNN	no / -
MSSRRC [229]	SISR	Uses residual compensation, uses fully polSAR data	VDSR	Residual learning	pre-upsampling	CNN	no / -

TABLE V: Summary of the state-of-the-art Deep Learning models for Super-resolution in Remote Sensing. CV model refers to the models presented in Table II.

fairly straightforward and easier compared to optical data. However SAR data inherently introduce speckle noise, which few authors explicitly consider when building SR pipelines.

Wang et al. [223] used a SISR approach by applying an SRGAN on TerraSAR-X images after having been de-speckled using a CNN as described in [230]. The high resolution image is downsampled by a factor of 4 using a Gaussian kernel, while both Generator and Discriminator elements are CNN-

based. The Generator element produces the SR image using the low resolution image, while the Discriminator compares the SR image with the high resolution image. The loss function comprises a perceptual loss with a content (pixel-wise MSE) and a weighted adversarial (probability-based) component of the Discriminator. Gu et al. [224] propose a transfer learning GAN-based paradigm in dealing with speckle noise using a so-called Noise-Free GAN (*NF-GAN*) to preserve the high-frequency image details as much as possible. They experiment with the HH polarization channel of Airborne SAR (AIRSAR) data. The Generator element consists of a de-speckling network and the reconstruction network, while the Discriminator element is ResNet based. The de-speckling network is pretrained using optical imagery with speckle noise added on them and it uses an MSE loss. Its input is a low resolution (downsampled HR version by a factor of 2) noise-full image. As with the previous case, the NF-GAN objective function is defined by an adversarial and a pixel-wise (MSE) component. The authors train their network pipeline with and without the de-speckling component and show that the former indeed works better.

Li et al. [225] tried to solve the problem of increased system's integration time and low azimuth resolution of the Geosynchronous SAR (GEO SAR) using a CNN-based GAN approach. GEO SAR is an active area of research in developing a SAR satellite system in geosynchronous orbit which will significantly assist in operational disaster monitoring by increasing the temporal resolution compared to Low Earth Orbit (LEO) satellite systems. In particular, the authors generated synthetic geosynchronous SAR data based on ALOS PALSAR characteristics. They use a Dialectical-GAN (*Di-GAN*) [231] with the Generator element comprising a U-net while the Discriminator a PatchGAN-like network. The Generator takes the low resolution simulated GEO SAR image as input whose SR produced image is compared with the ALOS PALSAR high resolution in the Discriminator. The authors claim a noticeable improvement of resolution which is mostly based on qualitative comparison.

Cen et al. [226] propose a three-module CNN-based network named *FSRCNN* for downscaling bistatic SAR images. The first module is used for feature extraction in various scales of the low resolution images. The second module adds together the resulting feature maps that were learned from the first module. The third module consists of a reconstruction CNN that computes the final SR image. The authors compare their results with bilinear, bicubic and SRCNN approaches using PSNR and SSIM and show an overall best performance of the proposed FSRCNN.

Helal-Kelany et al. [232] aimed to enhance the co-registration accuracy between two Single-Look Complex images of ERS-1/2 data. They train a Scale-Invariant SR CNN (*SINV CNN*) model using both amplitude and phase which mainly takes advantage of feature extraction and residual blocks components. Their result is evaluated based on descriptive statistics of the coherence between SINV CNN and sinc interpolation instead of commonly used metrics used in Computer Vision which may make their output difficult to compare with other approaches

Shen et al. [227] present a rather complete work where they apply their technique (*PSSR*) to full Polarimetric SAR (PolSAR) images. On the contrary to [232] they do not treat real and imaginary image parts separately but utilize them with a separate structure block since information is lost because of separation. They use various satellite sensors such as Radarsat-2, ESAR and PiSAR whose data they de-speckle first. They compare their approach (along with residual compensation strategy) with conventional non-DL approach and Multi-channel SAR SR (MSSR) using PSNR and MAE. They also use Equivalent Number of Looks (ENL) which is used to spot whether artifacts are introduced after SR. Notably, they experiment with the presence of speckle noise and show that their approach is superior to the traditional methods. Lin et al. [229] also uses PolSAR data and propose a Residual Compensated MSSR (*MSSRRC*) to tackle issues of the conventional (non DL-based) super-resolution approaches such as insufficient use of polarimetric information and decreased reconstruction of details. Their network is a VDSR adjusted for multi-channel (full PolSAR) input applied on RadarSat-2 data which is compensated by residuals between low resolution reconstructed and original images. Prior to training all data are de-speckled. PSNR, SSIM and qualitative evaluation show better performance with and without residual compensation compared to conventional SR approaches.

Yu et al. [228] propose a Weighted Dense Connected Convolutional Network (*WDCCN*) which claim to be a better alternative to Fast Super-resolution Convolutional Neural Networks and DRCN. Their network is based on DRCN, as well as the notion on weighted dense connections and tries to combat the restricted feature propagation issue. They compare their approach with SRCNN and DRCN using PSNR which suggests a better performance.

In conclusion, before one starts searching for baseline models for SAR image downscaling based on the currently published literature, there are certain decisions that must be made. For example, the processing level of the input data ranging from Single Look Complex to coregistered and/or geometrically corrected, speckle filtered, etc., all play a role in designing fit-to-purpose downscaling models. Similarly, the preferred type of products (e.g. Fully-PolSAR, Interferometric Wide Swath mode, etc) is important. We then provide some general directions that need to be seen with care and do not discourage authors from further experimentation, since SAR image downscaling is at its research infancy. Results from architectures such as NF-GAN and PSSR indicate that speckle noise needs special treatment that should be integrated in the overall architecture, thus leading to end-to-end approaches. As baseline, researchers could begin with general noise suppression architectures established the CV field or dive deeper by adapting architectures dedicated to speckle noise reduction that already exist in the literature. Residual block components seem to also add value in the overall learning. In addition, if one decides to experiment with Single-Look Complex images, using a dedicated structure block would be more fruitful (e.g. PSSR) compared to the opposite (e.g. SINV CNN), as well as adapting activations other than ReLU (e.g. PRelu, Leaky ReLU etc.) that will not freeze the filters' weight update.

Finally, we suggest that more focus can be placed on GAN-based architectures in SAR downscaling since they can exploit more types of inputs and explicitly take into consideration SAR imaging unique characteristics.

Aerial imagery from UAVs/Drones

By their initial mass production and market distribution, Unmanned Aerial Vehicles (UAV) comprise one of the most applicable and simple mean of data acquisition influencing a plethora of applications, Remote Sensing included. Simple architectures, easy-to-use and low-cost solutions contributed to increasing their usage and expanding their applicability for various objectives. The simplicity in integrating widely used sensory systems such as optronics played a significant role in substituting core Remote Sensing systems as they overcome many applicability limitations. Nonetheless, despite their efficiency and robustness as data acquisition systems, simple cameras mounted on a UAV cannot entirely substitute satellite alternatives as the latter exhibit enhanced payload sensor technical specifications such as higher spatial resolution.

Aiming at exploiting UAV systems in specific Remote Sensing applications and higher spatial resolution for the acquired images, numerous super-resolution approaches have been proposed and validated in real use cases. Depending on the availability of the input images, resolution enhancement techniques are typically divided into MISR and SISR methods, as for satellite imagery SR. However, no DL models have been developed for the MISR case, therefore hereafter we will only focus on the SISR approach.

Targeting on identifying higher frequencies on images, wavelet multiscale representations have been used for training a CNN and thus, vice versa for their estimation [233]. A shallower CNN architecture was proposed in Gonzalez et al. [234] to be integrated on-board of a UAV so that computational resources and power requirements could be retained at low levels. The combination of two sequential CNNs along with a bicubic up-sampling stage produce sufficient spatial imagery data. A similar technique was also deployed in Truong et al. [235] where the LR image is inserted in a deep CNN with a residual skip connection and network-in-network for generating the higher resolution images. To reduce resource consumption by decreasing the total number of network parameters, a deep recursive dense network [236] (*DRDN*) has been proposed. The recursive dense block can extract abundant local features and adaptively combine different hierarchical features of the input image. A dedicated implementation of SRGAN (V) for UAV operations has been incorporated as an initial processing step in Zhou et al. [237] (*SAIC*). The main target of the proposed pipeline was to deliver a high precision detection framework. Nonetheless, the spatial increment of the aerial image's resolution as an initial processing step is considered imperative to attain high detection performances. Similar objective was shared in Chen et al. [238] where a synergistic CNN for spatial resolution enhancement along with a modified object detection algorithm, which processes the enhanced image, were established. Finally, dedicated CNN-based models were utilized in Aslahishahri et al. [239] targeting the enhancement of aerial spatial resolution for producing details in plant phenotyping showcasing that such models could be application oriented depending on the dataset availability.

In conclusion, most approaches applied in resolution increment of aerial images follow similar schemes as the problem is translated into a Computer Vision counterpart. The majority of the corresponding architectures rely on the extraction of features from pre-trained models which eventually limits the necessity of dedicated models apart from the application-driven solutions. Due to the fundamental operational nature of UAV systems, the overall performance is meaningful mostly in near real-time operations which eventually is a prerequisite in many cases. Hence, dedicated lightweight architectures for specific drone applications exhibit better performance both in terms of accuracy and execution time, with respect to more universal, generic and heavyweight, modeling solutions.

X. Datasets

Despite the abundance of RS images, there is still a noticeable gap in the availability of public benchmark datasets for the evaluation of downscaling methods. This is hardly surprising since such a benchmark dataset would require extremely careful handling and elaborate preprocessing pipelines during assembly in order to meet the following basic conditions:

- Each HR image must be paired with one or more LR images.
- All LR/HR pairs must share the same scaling factor.
- All LR/HR pairs must be aligned and co-registered.
- All images must contain minimum obstructions (e.g. clouds, haze, corrupt pixels, etc).
- The depicted scenes must be as diverse as possible. Especially for STF: temporal/phenological changes must be as diverse as possible.
- A large number of images are required to avoid overfitting DL models with thousands/millions of trainable parameters.

Apart from a handful of datasets proposed specifically for the task of spatial downscaling, several datasets addressing different RS problems, such as object detection or scene classification, have been systematically used by most downscaling studies since they offer a ready-to-use collection of high quality satellite images. In the following list we present the most popular of such datasets and their corresponding characteristics.

- **UC Merced** [240]: contains 2,100 aerial RGB images coming from the USGS National Map Urban Area Imagery depicting 21 different land use classes at 0.3 m resolution from several US regions.
- **WHU-RS19** [241]: contains 950 aerial RGB images from Google Earth depicting 19 classes of land use at different spatial resolutions reaching up to 0.5m. Images originate from different regions around the world.
- **WHU-RS20** [242]: an extension of the WHU-RS19 dataset with an extra land use class and a total of 5,000 aerial RGB images.
- **RSSCN7** [243]: contains 2,800 aerial RGB images from Google Earth depicting 7 land use classes.
- **RSC11** [244]: contains 1,232 aerial RGB images from Google Earth depicting 11 land use classes at 0.2m spatial resolution. Images come from several US cities.

- **Aerial Image Dataset (AID)** [245]: contains 10,000 aerial RGB images coming from Google Earth at resolutions ranging from 0.5m to 8m. They depict 30 land use classes from different countries around the world and at different time and seasons.
- **NWPU-RESISC45** [246]: contains 31,500 aerial RGB images from Google Earth depicting 45 land use classes with spatial resolution ranging from 0.2m to 30m. Images come from several different regions around the world.
- **SIRI-WHU** [247]: 2,400 aerial RGB images from Google Earth depicting 12 land use classes at a spatial resolution of 2m. The images mainly cover urban areas in China.
- **Brazilian coffee scene dataset** [248]: contains 2,876 SPOT images (Green, Red, NIR bands) over 4 regions in Brazil for binary image classification based on the presence or absence of coffee crops.
- **SEN1-2** [249]: contains 282,384 pairs of Sentinel-1 and Sentinel-2 RGB images at 10m spatial resolution from around the world at different seasons.
- **SEN12MS** [250]: contains 180,662 triplets of Sentinel-1 dual-polarization SAR, Sentinel-2 multispectral and MODIS land cover images at 10m spatial resolution coming from all around the globe and at different times.
- **DOTA** [251]: contains 2,806 aerial images from different sensors along with GaoFen-2 and Jilin-1 satellite images. This dataset is targeted towards object detection and includes labels spanning over 15 object categories.
- **DIOR** [33]: contains 23,463 aerial RGB images from Google Earth with spatial resolutions ranging from 0.5m to 30m. The images cover several regions around the globe and their labels span over 20 object categories.
- **CIA** [252]: contains 17 Landsat/MODIS pairs from Coleambally Irrigation Area, Australia, at 25m spatial resolution. Images were obtained during a single summer season but have strong spatial heterogeneity.
- **LGC** [252]: contains 14 Landsat/MODIS pairs from Lower Gwydir Catchment, Australia, at 25m spatial resolution. Images were obtained during a whole year which also included a major flood. This renders the dataset ideal for the study of abrupt and unpredictable changes in time series.
- **AHB** [253]: contains 27 Landsat/MODIS pairs from Ar Horqin Banner, China, over a span of 5 years. It is intended for the study of phenological changes in rural areas.
- **Tianjin** [253]: contains 27 Landsat/MODIS pairs from Tianjin, China, over a span of 6 years. It is intended for the study of phenological changes in urban areas.
- **Daxing** [253]: contains 29 Landsat/MODIS pairs from Daxing, China, over a span of 6 years. It is intended for the study of land cover changes.
- **Gaofen Image Dataset (GID)** [254]: contains 150 Gaofen-2 images (Red, Green, Blue, NIR bands) from many regions in China with 4m spatial resolution. It is intended for scene classification and land cover segmentation.
- **Kelvin's PROBA-V Super-resolution dataset** [211]: contains 1,160 images from PROBA-V satellite (Red, NIR bands) from several locations around the globe at different points in time. Each data point contains an HR image of 100m resolution and several LR images of 300m resolution.
- **Kaggle's Draper Satellite Image Chronology** [255]: contains 1,720 aerial RGB images from California, USA, over a period of 5 days.
- **DRealSR (Diverse Real-world image Super-resolution)** [256]: contains 31,970 of low resolution image patches including aerial images.
- **Pavia Center** [118]: acquired by ROSIS over the city of Pavia, Italy, in the wavelength range of 430–860nm. It contains 115 spectral bands and is of size 1096 × 1096.
- **Houston** [118]: acquired by an ITRES-CASI 1500 HS sensor over the campus of the University of Houston and its neighboring urban areas. Each HS image comprises 144 bands covering the spectral range of 380–1050nm, and each band contains 349 × 1905 pixels with a spatial resolution of 2.5m
- **Los Angeles** [118]: acquired over a port in the city of Los Angeles by the Hyperion sensor mounted on the Earth Observing One (EO-1) satellite. The HS image contains 242 spectral bands with a spatial resolution of 30m.
- **Botswana** [257]: acquired over the Okavango Delta in Botswana by the Hyperion sensor mounted on the Earth Observing One (EO-1) satellite. The HS image contains 242 spectral bands with a spatial resolution of 30m.
- **Hobart** [113]: acquired by the IKONOS sensor, represents an urban and harbor area of Hobart, Australia. The MS sensor is characterized by four bands (Red, Green, Blue, and NIR) and also a PAN channel with band range from 450nm to 900nm. The resolution of MS is 4m and PAN is 1m.
- **Sundarbans** [113]: obtained by the QuickBird sensor, represents a forest area of Sundarbans in India. Provides a high resolution PAN image with spectral cover range from 760nm to 850nm and a resolution of 0.6m, and a four-band (Red, Green, Blue and NIR) MS image with a resolution of 2.4m.
- **Washington DC Mall** [139]: covers an urban area in Washington DC Mall. The size of the degraded HS image is 256 × 60 and that of the PAN image is 1280 × 300.
- **Moffett Field** [139]: covers a mixed urban/rural area in Moffett Field, California. The size of the degraded HS image is 79 × 37 with 100m resolution and that of the PAN image is 395 × 185 with 20m resolution.
- **Salinas Scene** [139]: covers a rural area in Salinas Valley, California. The size of the degraded HS image is 102 × 43 and that of the PAN image is 510 × 215.
- **Chikusei** [258]: captured by the Headwall's Hyperspec Visible and Near-Infrared, series C (VNIRC) imaging sensor over Chikusei, Ibaraki, Japan, on July 29, 2014. Contains 128 bands in the spectral range of 363–1018nm. The PAN image has 300 × 300 pixels with a spatial resolution of 2.5m.
- **Foster** [258]: has 33 spectral channels from 400–720nm with 10nm per band. The original size of each HS image

in the Foster data set is 1341 × 1022.

XI. Advancements in Computer Vision

Spatial enhancement or Super-resolution is being thoroughly investigated in general Computer Vision and a great number of methods have been proposed which build on previous research and expand the state-of-the-art. Hence, in the CV field, some informative review papers have been published in the last couple of years focusing on CV DL algorithms for image downscaling, such as [12] and [16]. In this section we present some of the most promising and innovative studies in CV published over the last few years, which to the best of our knowledge have not yet been used in an RS context, hoping to provide a source of inspiration for further applications in the RS field.

Most of the studies found in literature train models on synthetic datasets where LR counterparts are synthetically constructed usually via a single predefined degradation algorithm such as bicubic interpolation. This raises the question of whether such a model can properly generalize to real-world images that have undergone arbitrary degradation processes. To that end, a number of publications (e.g. [259] (*SFTMD*), [260], [261] (*DAN*)) explore deep networks which are trained to jointly handle the downscaling task and learn the appropriate blur kernel in an end-to-end fashion. This family of methods is usually referred to as *Blind Super-resolution*.

In some cases, the available dataset comprises LR images which need to be downscaled, along with a number of HR reference images of the same domain which however do not correspond to the LR data. A family of methods attempt to exploit such HR information through domain translation approaches and the adaptation of the *CycleGAN* [164] idea. For example, [262] (*CinCGAN*), [263] (*DDGAN*), [264] (*UISRPS*) and [265] (*MCinCGAN*) propose GAN architectures which are trained to translate the LR images to cleaned, synthetic LR counterparts and then further downscale the result to an HR output. The use of cycle-consistency loss circumvents the need for paired data so any HR data of the same domain can be used.

An emerging trend in the field of Super-resolution approaches are the diffusion models. Initially proposed in [266] diffusion models employ a Markov chain to slowly add Gaussian noise to the input data and a trainable model to stochastically learn the reverse process of gradually removing this noise. Saharia et al. [267] (*SR3*) adapt this idea to image super-resolution of faces and natural images by training a U-Net to iteratively refine Gaussian noise conditioned on the LR image. Their method achieved results of remarkable sharpness and realism while remaining true to the LR input. In addition, by cascading multiple such models, higher scaling factors can be targeted (e.g. $\times 8$, $\times 16$) without compromising the final image quality. This breakthrough study showed that diffusion models can overcome GANs and set an interesting research field for future exploration.

XII. Discussion

A number of key findings have emerged from the present literature review which showcase the limitations of the current approaches. In the following paragraphs we highlight some essential topics for further exploration and research in the task of image downscaling, focused especially on the field of Remote Sensing.

Universal metrics. An important conclusion of Section IV is the fact that there exist no established evaluation metrics for downscaling models. Surely, a limited subset of the metrics presented in Table I have become more popular and widely used in recent studies, however none of them can entirely capture and assess the quality of a produced SR image. The design of a universal metric (or set of metrics) able to account for both low distortion and high perceptual quality of an image is still an open field of research and the DL community will greatly benefit from any advancement in this area.

Model interpretability. The definition of universal quality indices for EO image downscaling contributes to the robustness against the inherent super-resolved image hallucinations, and increase in the trust and interpretability of proposed SR models. Indeed, generative networks, widely used for image downscaling and thoroughly presented in this review, while they are able to achieve impressive aesthetic results, however they are prone to creating hallucinations and/or artifacts. Controlling and quantifying the trade-off between SR performance vis-à-vis the expected hallucinations level remains an open issue. In addition, it may be that a single metric characterising overall model performance is not enough, but an additional gridded output with uncertainty estimates should be produced. Therefore, we consider critical to develop algorithms that will help both ML practitioners and end-users to better understand, interpret and trust the DL model outputs. Explainable AI (xAI) algorithms [268] are essential tools towards an enhanced understanding and transparency of the developed DL models, especially for facilitating the operational uptake of EO image downscaling models.

Benchmark datasets. The availability and abundance of Remote Sensing images has greatly facilitated the formulation of datasets which satisfy the needs of complex DL models. Many researchers choose to directly download RS images from the respective providers, perform the pre-processing pipeline that best suits their analysis and subsequently evaluate the model output on a held-out subset. However, there is an urgent need for specific, carefully designed benchmark datasets tailored to the downscaling task, which will help to objectively evaluate and compare different models, thus gaining a more concrete insight on their generalization and applicability.

Model performance. In addition to the point above, the adoption of best practices during and after model-building procedures is also necessary. In the former case, ablation studies can be adopted more widely, while in the latter case, results can be followed by some sort of evidence of statistical strength when comparing models. As a result, practices such as these, among others, may lead to more understandable architectures and transparent results, as well as less biased and weak inference regarding model performance.

Open source code and reproducibility. During our study we observed a glaring lack of source code availability for the presented methods. This prevents objective evaluation and hinders quick advancements in the field. Transparency, reproducibility

and testability of the reported results and comparison with novel approaches require publicly accessible source code of the whole pipeline as well as a permissive license of use (e.g. MIT, BSD, GNU, etc). In this way, faster scientific progress can be achieved which from a model's perspective means that it can go up the Technology Readiness Level (TRL) faster. To this end, a possible contribution from the authors in addition to open source code, would be to explicitly make reference to the number of trainable parameters of their models. This information provides intuition to data scientists. Depending on the problem at hand, the available data for training and the computing resources, model size provides useful indications for training time and effectiveness, although other factors, such as the use of recursive architectures, can affect these.

Beyond a single degradation scheme. When the acquisition of LR-HR image pairs is too expensive or overall impossible, Wald's protocol often comes to the rescue. Even though it offers an outlet for the formulation of an appropriate training dataset, LR images are usually constructed with a single degradation algorithm. Consequently, a model trained on such a dataset learns to "reverse" this particular degradation scheme and therefore may fail to generalize on different degradation/distortion operations. Further study is required for the development of models able to handle diverse types of image distortion, applicable in real world scenarios during a sensor capture of an image.

Multimodal fusion. The spectral fusion of images can greatly assist the downscaling process (Section VII). But apart from captures lying in the visible and infrared spectra, new approaches can be investigated for the fusion of other spectral ranges. For example, radar imaging can provide complementary information to optical imaging, such as surface topography, and is also able to penetrate canopies and clouds/smoke. Therefore, an interesting topic of study would be the fusion of Synthetic-aperture radar (SAR) and optical data for the purpose of downscaling, which to our knowledge has not yet been investigated in the DL field.

GANs or else. GANs manage to better approximate the boundary of the Perception-Distortion plane and achieve more realistic and perceptually convincing results (see Section IV). Therefore, a further study of the GAN framework is needed in order to exploit its potential to the full extend. Additionally, an exploration of novel architectures and training schemes may lead to performances even closer to the boundary. For example, recent studies have unveiled the great power of diffusion models and future research may possibly establish them as the successor of GANs to the downscaling state-of-the-art.

Unsupervised learning. Acquiring ground truth HR labels in the training dataset is often a time consuming and expensive task, while in some cases it may also be practically infeasible. On the other hand, a synthetic training dataset can be developed through Wald's protocol, but this process requires additional degradation and high-frequency information loss. To tackle this problem, some studies employ a completely unsupervised learning scheme with specially designed loss functions. Even though these models still struggle to match the performance of their supervised competitors, they tend to preserve high-frequency details and stay faithful to the spectral content of the LR input. Therefore, we believe that unsupervised learning offers a potential outlet for handling the lack of training targets in downscaling and further research will only achieve fruitful results.

CV paradigm. The field of general Computer Vision has made a lot more progress on the task of downscaling and novel architectures and ideas have been recently introduced. We believe that the RS domain could greatly benefit from an adaptation and expansion of these developments. We introduce some of these methods in Section XI. However, caution is needed when directly applying such approaches since scaling factors in the RS domain are usually considerably larger and may hinder the model's performance. For example, Super-resolution in natural images usually involves a magnification factor much smaller than those in the Remote Sensing domain (ranging from $\times 2$ to $\times 4$ compared with $\times 8$ to $\times 16$) where texture information is severely distorted and high-frequency details are almost impossible to retrieve. Therefore a simple transfer learning approach is not possible and specialized architectures must be designed when it comes to RS data.

Downscaling SAR imagery. The techniques proposed in the literature for SAR image enhancement are few and they compare well-established techniques borrowed from computer vision research on SISR. However, special care is needed to downscale SAR data, since they present properties that need to be either taken explicitly into account by tailored model architectures or to be eliminated beforehand. For example, few authors use fully PolSAR data and even fewer incorporate the complex number nature of SAR data in their models. In addition, preprocessing steps need to be presented in a clearer way, while in our review only a number of authors apply SR techniques on data of the same level of preprocessing. This may lead to SAR unique properties such as speckle noise and geometric distortions (e.g. foreshortening, layover) affecting the model performance or resulting in misleading outcomes. Therefore, we believe that there is room for significant improvement in SAR imagery SR modeling, focusing on the unique SAR properties and designing proper model architectures, loss functions and accuracy metrics. Last but not least, other potential future research orientations could be towards adaptation of MISR and expansion of SISR approaches using SAR data acquired from different SAR imaging sensors. This will provide new external information to assist the downscaling process, exploiting different view geometries through incidence angle diversity, radar frequency bands (e.g. C-, X-, L-band), imaging modes (e.g. StripMap, Wide Swath, Spotlight, etc.), and availability of polarimetric data.

XIII. Conclusion

In this survey we offer a detailed overview of the methods available in the literature for spatial downscaling of Remote Sensing imagery. We explore the different types of spatial enhancement and introduce a comprehensive taxonomy of the various approaches. Additionally, we conduct a thorough investigation on the most popular metrics and datasets for this task, and analyze the trade-off between Perception and Distortion as a key factor for the selection of an appropriate

loss function and training scheme. Finally, we discuss the weaknesses and shortcomings of the current state of the art in the field and briefly present recent advancements in the general Computer Vision community as a source of inspiration.

As seen from our analysis, although there is a strong presence of the Deep Learning paradigm in RS and the publication rates are ever increasing, there is still plenty of room for improvement and exploration. Various facets of the downscaling problem could benefit from new contributions, such as universal evaluation metrics and model interpretability algorithms towards xAI, multi-modal datasets, innovative upsampling layers/frameworks, novel training schemes, original architectures, and many more. Due to the wide range of RS data and applicability, there is and will be an incessant need for better, more efficient and trustworthy DL models. We hope that this survey will further stimulate the research community and assist in avoiding common pitfalls in the design, development and assessment of new DL techniques.

REFERENCES

[1] P. Ghamisi, B. Rasti, N. Yokoya, Q. Wang, B. Hofle, L. Bruzzone, F. Bovolo, M. Chi, K. Anders, R. Gloaguen, P. M. Atkinson, and J. A. Benediktsson, "Multisource and Multitemporal Data Fusion in Remote Sensing: A Comprehensive Review of the State of the Art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 7, no. 1, pp. 6–39, Mar. 2019, conference Name: IEEE Geoscience and Remote Sensing Magazine.

[2] B. Chen, J. Li, and Y. Jin, "Deep Learning for Feature-Level Data Fusion: Higher Resolution Reconstruction of Historical Landsat Archive," *Remote Sensing*, vol. 13, no. 2, p. 167, Jan. 2021. [Online]. Available: https://www.mdpi.com/2072-4292/13/2/167

[3] A. O. Onojeghuo, G. A. Blackburn, Q. Wang, P. M. Atkinson, D. Kindred, and Y. Miao, "Mapping paddy rice fields by applying machine learning algorithms to multi-temporal Sentinel-1A and Landsat data," *International Journal of Remote Sensing*, vol. 39, no. 4, pp. 1042–1067, Feb. 2018. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/01431161.2017.1395969

[4] Y. Zhang, P. M. Atkinson, X. Li, F. Ling, Q. Wang, and Y. Du, "Learning-Based Spatial–Temporal Superresolution Mapping of Forest Cover With MODIS Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 1, pp. 600–614, Jan. 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7586093/

[5] Y. Feng, D. Lu, E. Moran, L. Dutra, M. Calvi, and M. de Oliveira, "Examining Spatial Distribution and Dynamic Change of Urban Land Covers in the Brazilian Amazon Using Multitemporal Multisensor High Spatial Resolution Satellite Imagery," *Remote Sensing*, vol. 9, no. 4, p. 381, Apr. 2017. [Online]. Available: http://www.mdpi.com/2072-4292/9/4/381

[6] A. Y. Sun and G. Tang, "Downscaling Satellite and Reanalysis Precipitation Products Using Attention-Based Deep Convolutional Neural Nets," *Frontiers in Water*, vol. 2, p. 536743, Nov. 2020. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frwa.2020.536743/full

[7] I. K. Lee, J. C. Trinder, and A. Sowmya, "APPLICATION OF U-NET CONVOLUTIONAL NEURAL NETWORK TO BUSHFIRE MONITORING IN AUSTRALIA WITH SENTINEL-1/-2 DATA," *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLIII-B1-2020, pp. 573–578, Aug. 2020. [Online]. Available: https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLIII-B1-2020/573/2020/

[8] M. M. Pinto, R. Libonati, R. M. Trigo, I. F. Trigo, and C. C. DaCamara, "A deep learning approach for mapping and dating burned areas using temporal sequences of satellite images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 160, pp. 260–274, Feb. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0924271619303089

[9] D. Garcia, G. Mateo-Garcia, H. Bernhardt, R. Hagensieker, I. G. L. Francos, J. Stock, G. Schumann, K. Dobbs, and F. Kalaitzis, "Pix2Streams: Dynamic Hydrology Maps from Satellite-LiDAR Fusion," *arXiv:2011.07584 [cs, eess, stat]*, Nov. 2020, arXiv: 2011.07584. [Online]. Available: http://arxiv.org/abs/2011.07584

[10] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue, and Q. Liao, "Deep Learning for Single Image Super-Resolution: A Brief Review," *IEEE Transactions on Multimedia*, vol. 21, no. 12, pp. 3106–3121, Dec. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8723565/

[11] J. J. Danker Khoo, K. H. Lim, and J. T. Sien Phang, "A Review on Deep Learning Super Resolution Techniques," in *2020 IEEE 8th Conference on Systems, Process and Control (ICSPC)*. Melaka, Malaysia: IEEE, Dec. 2020, pp. 134–139. [Online]. Available: https://ieeexplore.ieee.org/document/9305806/

[12] H. Chen, X. He, L. Qing, Y. Wu, C. Ren, and C. Zhu, "Real-World Single Image Super-Resolution: A Brief Review," *arXiv:2103.02368 [cs, eess]*, Mar. 2021, arXiv: 2103.02368. [Online]. Available: http://arxiv.org/abs/2103.02368

[13] S. M. A. Bashir, Y. Wang, and M. Khan, "A Comprehensive Review of Deep Learning-based Single Image Super-resolution," *arXiv:2102.09351 [cs, eess]*, Feb. 2021, arXiv: 2102.09351. [Online]. Available: http://arxiv.org/abs/2102.09351

[14] K. Nasrollahi and T. B. Moeslund, "Super-resolution: a comprehensive survey," *Machine Vision and Applications*, vol. 25, no. 6, pp. 1423–1468, Aug. 2014. [Online]. Available: http://link.springer.com/10.1007/s00138-014-0623-4

[15] H.-I. Kim and S. B. Yoo, "Trends in Super-High-Definition Imaging Techniques Based on Deep Neural Networks," *Mathematics*, vol. 8, no. 11, p. 1907, Oct. 2020. [Online]. Available: https://www.mdpi.com/2227-7390/8/11/1907

[16] Z. Wang, J. Chen, and S. C. H. Hoi, "Deep Learning for Image Super-Resolution: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3365–3387, Oct. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9044873/

[17] F. Dadrass Javan, F. Samadzadegan, S. Mehravar, A. Toosi, R. Khatami, and A. Stein, "A review of image fusion techniques for pan-sharpening of high-resolution satellite imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 171, pp. 101–117, Jan. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0924271620303002

[18] G. Kaur, K. S. Saini, D. Singh, and M. Kaur, "A Comprehensive Study on Computational Pansharpening Techniques for Remote Sensing Images," *Archives of Computational Methods in Engineering*, Feb. 2021. [Online]. Available: http://link.springer.com/10.1007/s11831-021-09565-y

[19] X. Meng, H. Shen, H. Li, L. Zhang, and R. Fu, "Review of the pansharpening methods for remote sensing images based on the idea of meta-analysis: Practical discussion and challenges," *Information Fusion*, vol. 46, pp. 102–113, Mar. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1566253517306036

[20] R. Fernandez-Beltran, P. Latorre-Carmona, and F. Pla, "Single-frame super-resolution in remote sensing: a practical overview," *International Journal of Remote Sensing*, vol. 38, no. 1, pp. 314–354, Jan. 2017. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/01431161.2016.1264027

[21] Q. Yuan, H. Shen, T. Li, Z. Li, S. Li, Y. Jiang, H. Xu, W. Tan, Q. Yang, J. Wang, J. Gao, and L. Zhang, "Deep learning in environmental remote sensing: Achievements and challenges," *Remote Sensing of Environment*, vol. 241, p. 111716, May 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0034425720300857

[22] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, Dec. 2017. [Online]. Available: http://ieeexplore.ieee.org/document/8113128/

[23] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 152, pp. 166–177, Jun. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0924271619301108

[24] Xiaolin Zhu, Fangyi Cai, Jiaqi Tian, and Trecia Williams, "Spatiotemporal Fusion of Multisource Remote Sensing Data: Literature Survey, Taxonomy, Principles, Applications, and Future Directions," *Remote Sensing*, vol. 10, no. 4, p. 527, Mar. 2018. [Online]. Available: http://www.mdpi.com/2072-4292/10/4/527

[25] G. Tsagkatakis, A. Aidini, K. Fotiadou, M. Giannopoulos, A. Pentari, and P. Tsakalides, "Survey of Deep-Learning Approaches for Remote

Sensing Observation Enhancement," *Sensors*, vol. 19, no. 18, p. 3929, Sep. 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/18/3929

[26] "Web of science. [Online]." https://webofknowledge.com, accessed: 2021-09-09.

[27] A. W. Wood, L. R. Leung, V. Sridhar, and D. P. Lettenmaier, "Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs," *Climatic Change*, vol. 62, no. 1, pp. 189–216, Jan 2004. [Online]. Available: https://doi.org/10.1023/B:CLIM.0000013685.99609.9e

[28] P. M. Atkinson, "Downscaling in remote sensing," *International Journal of Applied Earth Observation and Geoinformation*, vol. 22, pp. 106–114, 2013, spatial Statistics for Mapping the Environment. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0303243412000918

[29] W. Sun and Z. Chen, "Learned image downscaling for upscaling using content adaptive resampler," *IEEE Transactions on Image Processing*, vol. 29, pp. 4027–4040, 2020.

[30] W. Zhan, Y. Chen, J. Zhou, J. Wang, W. Liu, J. Voogt, X. Zhu, J. Quan, and J. Li, "Disaggregation of remotely sensed land surface temperature: Literature survey, taxonomy, issues, and caveats," *Remote Sensing of Environment*, vol. 131, pp. 119–139, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425712004804

[31] Y. Li, H. Zhang, X. Xue, Y. Jiang, and Q. Shen, "Deep learning for remote sensing image classification: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 6, Nov. 2018. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1264

[32] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote Sensing Image Scene Classification Meets Deep Learning: Challenges, Methods, Benchmarks, and Opportunities," *arXiv:2005.01094 [cs]*, Jun. 2020, arXiv: 2005.01094. [Online]. Available: http://arxiv.org/abs/2005.01094

[33] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 159, pp. 296–307, Jan. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0924271619302825

[34] W. Ma, J. Zhang, Y. Wu, L. Jiao, H. Zhu, and W. Zhao, "A Novel Two-Step Registration Method for Remote Sensing Images Based on Deep and Local Features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 4834–4843, Jul. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8648479/

[35] N. Merkle, W. Luo, S. Auer, R. Müller, and R. Urtasun, "Exploiting Deep Matching and SAR Data for the Geo-Localization Accuracy Improvement of Optical Satellite Images," *Remote Sensing*, vol. 9, no. 6, p. 586, Jun. 2017. [Online]. Available: http://www.mdpi.com/2072-4292/9/6/586

[36] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogrammetric engineering and remote sensing*, vol. 63, no. 6, pp. 691–699, 1997, publisher: Asprs American Society for Photogrammetry and. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00365304

[37] 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, Apr. 2004. [Online]. Available: http://ieeexplore.ieee.org/document/1284395/

[38] C. R. Helmrich, S. Bosse, M. Siekmann, H. Schwarz, D. Marpe, and T. Wiegand, "Perceptually Optimized Bit-Allocation and Associated Distortion Measure for Block-Based Image or Video Coding," in *2019 Data Compression Conference (DCC)*. Snowbird, UT, USA: IEEE, Mar. 2019, pp. 172–181. [Online]. Available: https://ieeexplore.ieee.org/document/8712674/

[39] Zhou Wang and A. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, Mar. 2002. [Online]. Available: http://ieeexplore.ieee.org/document/995823/

[40] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Pacific Grove, CA, USA: IEEE, 2003, pp. 1398–1402. [Online]. Available: http://ieeexplore.ieee.org/document/1292216/

[41] H. Sheikh, A. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on Image Processing*, vol. 14, no. 12, pp. 2117–2128, Dec. 2005. [Online]. Available: https://ieeexplore.ieee.org/document/1532311

[42] H. Sheikh and A. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, Feb. 2006. [Online]. Available: https://ieeexplore.ieee.org/document/1576816

[43] N. Damera-Venkata, T. Kite, W. Geisler, B. Evans, and A. Bovik, "Image quality assessment based on a degradation model," *IEEE Transactions on Image Processing*, vol. 9, no. 4, pp. 636–650, Apr. 2000. [Online]. Available: http://ieeexplore.ieee.org/document/841940/

[44] Lin Zhang, Lei Zhang, Xuanqin Mou, and D. Zhang, "FSIM: A Feature Similarity Index for Image Quality Assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011. [Online]. Available: http://ieeexplore.ieee.org/document/5705575/

[45] Anmin Liu, Weisi Lin, and M. Narwaria, "Image Quality Assessment Based on Gradient Similarity," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1500–1512, Apr. 2012. [Online]. Available: http://ieeexplore.ieee.org/document/6081939/

[46] R. H. Yuhas, A. F. Goetz, and J. W. Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm," in *Proceedings of the 3rd Annual JPL Airborne Earth Science Workshop*, Pasadena, CA, USA, Jun. 1992. [Online]. Available: https://aviris.jpl.nasa.gov/proceedings/workshops/92_docs/52.PDF

[47] L. Wald, "Quality of high resolution synthesised images: Is there a simple criterion ?" 2000.

[48] D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, p. 011006, Jan. 2010. [Online]. Available: http://electronicimaging.spiedigitallibrary.org/article.aspx?doi=10.1117/1.3267105

[49] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual Losses for Real-Time Style Transfer and Super-Resolution," *arXiv:1603.08155 [cs]*, Mar. 2016, arXiv: 1603.08155. [Online]. Available: http://arxiv.org/abs/1603.08155

[50] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-Reference Image Quality Assessment in the Spatial Domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012. [Online]. Available: http://ieeexplore.ieee.org/document/6272356/

[51] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "Completely Blind" Image Quality Analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, Mar. 2013. [Online]. Available: http://ieeexplore.ieee.org/document/6353522/

[52] Venkatanath N, Praneeth D, Maruthi Chandrasekhar Bh, S. S. Channappayya, and S. S. Medasani, "Blind image quality evaluation using perception based features," in *2015 Twenty First National Conference on Communications (NCC)*. Mumbai, India: IEEE, Feb. 2015, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/document/7084843/

[53] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang, "Learning a No-Reference Quality Metric for Single-Image Super-Resolution," *arXiv:1612.05890 [cs]*, Dec. 2016, arXiv: 1612.05890. [Online]. Available: http://arxiv.org/abs/1612.05890

[54] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, "The 2018 PIRM Challenge on Perceptual Image Super-resolution," *arXiv:1809.07517 [cs]*, Jan. 2019, arXiv: 1809.07517. [Online]. Available: http://arxiv.org/abs/1809.07517

[55] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," *arXiv:1801.03924 [cs]*, Apr. 2018, arXiv: 1801.03924. [Online]. Available: http://arxiv.org/abs/1801.03924

[56] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, and M. Selva, "Multispectral and Panchromatic Data Fusion Assessment Without Reference," *Photogrammetric Engineering & Remote Sensing*, vol. 74, no. 2, pp. 193–200, Feb. 2008. [Online]. Available: http://openurl.ingenta.com/content/xref?genre=article\&issn=0099-1112\&volume=74\&issue=2\&spage=193

[57] Y. Blau and T. Michaeli, "The Perception-Distortion Tradeoff," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6228–6237, Jun. 2018, arXiv: 1711.06077. [Online]. Available: http://arxiv.org/abs/1711.06077

[58] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, p. 2672–2680.

[59] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. San Francisco, CA, USA: IEEE, Jun. 2010, pp. 2528–2535. [Online]. Available: http://ieeexplore.ieee.org/document/5539957/

[60] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and Checkerboard Artifacts," *Distill*, 2016. [Online]. Available: http://distill.pub/2016/deconv-checkerboard

[61] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network," *arXiv:1609.05158 [cs, stat]*, Sep. 2016, arXiv: 1609.05158. [Online]. Available: http://arxiv.org/abs/1609.05158

[62] W. Shi, J. Caballero, L. Theis, F. Huszar, A. Aitken, C. Ledig, and Z. Wang, "Is the deconvolution layer the same as a convolutional layer?" *arXiv:1609.07009 [cs]*, Sep. 2016, arXiv: 1609.07009. [Online]. Available: http://arxiv.org/abs/1609.07009

[63] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017, arXiv: 1608.03981. [Online]. Available: http://arxiv.org/abs/1608.03981

[64] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv:1512.03385 [cs]*, Dec. 2015, arXiv: 1512.03385. [Online]. Available: http://arxiv.org/abs/1512.03385

[65] P. Burt and E. Adelson, "The Laplacian Pyramid as a Compact Image Code," *IEEE Transactions on Communications*, vol. 31, no. 4, pp. 532–540, Apr. 1983. [Online]. Available: http://ieeexplore.ieee.org/document/1095851/

[66] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," *arXiv:1709.01507 [cs]*, May 2019, arXiv: 1709.01507. [Online]. Available: http://arxiv.org/abs/1709.01507

[67] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," *arXiv:1910.03151 [cs]*, Apr. 2020, arXiv: 1910.03151. [Online]. Available: http://arxiv.org/abs/1910.03151

[68] Q. Hou, D. Zhou, and J. Feng, "Coordinate Attention for Efficient Mobile Network Design," *arXiv:2103.02907 [cs]*, Mar. 2021, arXiv: 2103.02907. [Online]. Available: http://arxiv.org/abs/2103.02907

[69] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "BAM: Bottleneck Attention Module," *arXiv:1807.06514 [cs]*, Jul. 2018, arXiv: 1807.06514. [Online]. Available: http://arxiv.org/abs/1807.06514

[70] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," *arXiv:1807.06521 [cs]*, Jul. 2018, arXiv: 1807.06521. [Online]. Available: http://arxiv.org/abs/1807.06521

[71] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, "Rotate to Attend: Convolutional Triplet Attention Module," *arXiv:2010.03045 [cs]*, Nov. 2020, arXiv: 2010.03045. [Online]. Available: http://arxiv.org/abs/2010.03045

[72] H. Zhu, C. Xie, Y. Fei, and H. Tao, "Attention Mechanisms in CNN-Based Single Image Super-Resolution: A Brief Review and a New Perspective," *Electronics*, vol. 10, no. 10, p. 1187, May 2021. [Online]. Available: https://www.mdpi.com/2079-9292/10/10/1187

[73] C. Dong, C. C. Loy, K. He, and X. Tang, "Image Super-Resolution Using Deep Convolutional Networks," *arXiv:1501.00092 [cs]*, Jul. 2015, arXiv: 1501.00092. [Online]. Available: http://arxiv.org/abs/1501.00092

[74] J. Kim, J. K. Lee, and K. M. Lee, "Accurate Image Super-Resolution Using Very Deep Convolutional Networks," *arXiv:1511.04587 [cs]*, Nov. 2016, arXiv: 1511.04587. [Online]. Available: http://arxiv.org/abs/1511.04587

[75] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv:1409.1556 [cs]*, Apr. 2015, arXiv: 1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556

[76] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution," *arXiv:1704.03915 [cs]*, Oct. 2017, arXiv: 1704.03915. [Online]. Available: http://arxiv.org/abs/1704.03915

[77] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," *arXiv:1609.04802 [cs, stat]*, May 2017, arXiv: 1609.04802. [Online]. Available: http://arxiv.org/abs/1609.04802

[78] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, C. C. Loy, Y. Qiao, and X. Tang, "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks," *arXiv:1809.00219 [cs]*, Sep. 2018, arXiv: 1809.00219. [Online]. Available: http://arxiv.org/abs/1809.00219

[79] A. Jolicoeur-Martineau, "The relativistic discriminator: a key element missing from standard GAN," *arXiv:1807.00734 [cs, stat]*, Sep. 2018, arXiv: 1807.00734. [Online]. Available: http://arxiv.org/abs/1807.00734

[80] Xintao, "xinntao/ESRGAN," Sep. 2021, original-date: 2018-08-31T08:18:41Z. [Online]. Available: https://github.com/xinntao/ESRGAN

[81] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced Deep Residual Networks for Single Image Super-Resolution," *arXiv:1707.02921 [cs]*, Jul. 2017, arXiv: 1707.02921. [Online]. Available: http://arxiv.org/abs/1707.02921

[82] J. Yu, Y. Fan, J. Yang, N. Xu, Z. Wang, X. Wang, and T. Huang, "Wide Activation for Efficient and Accurate Image Super-Resolution," *arXiv:1808.08718 [cs]*, Dec. 2018, arXiv: 1808.08718. [Online]. Available: http://arxiv.org/abs/1808.08718

[83] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual Dense Network for Image Super-Resolution," *arXiv:1802.08797 [cs]*, Mar. 2018, arXiv: 1802.08797. [Online]. Available: http://arxiv.org/abs/1802.08797

[84] Y. Tai, J. Yang, and X. Liu, "Image Super-Resolution via Deep Recursive Residual Network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE, Jul. 2017, pp. 2790–2798. [Online]. Available: http://ieeexplore.ieee.org/document/8099781/

[85] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep Back-Projection Networks For Super-Resolution," *arXiv:1803.02735 [cs]*, Mar. 2018, arXiv: 1803.02735. [Online]. Available: http://arxiv.org/abs/1803.02735

[86] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback Network for Image Super-Resolution," *arXiv:1903.09814 [cs]*, Jun. 2019, arXiv: 1903.09814. [Online]. Available: http://arxiv.org/abs/1903.09814

[87] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-Recursive Convolutional Network for Image Super-Resolution," *arXiv:1511.04491 [cs]*, Nov. 2016, arXiv: 1511.04491. [Online]. Available: http://arxiv.org/abs/1511.04491

[88] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Fast and Accurate Image Super-Resolution with Deep Laplacian Pyramid Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 11, pp. 2599–2613, Nov. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8434354/

[89] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image Super-Resolution Using Very Deep Residual Channel Attention Networks," *arXiv:1807.02758 [cs]*, Jul. 2018, arXiv: 1807.02758. [Online]. Available: http://arxiv.org/abs/1807.02758

[90] D. Lei, H. Chen, L. Zhang, and W. Li, "Nlrnet: An efficient nonlocal attention resnet for pansharpening," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–13, 2021.

[91] C. Shang, X. Li, Z. Yin, X. Li, L. Wang, Y. Zhang, Y. Du, and F. Ling, "Spatiotemporal Reflectance Fusion Using a Generative Adversarial Network," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–15, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9383451/

[92] Y. Yu, X. Li, and F. Liu, "E-DBPN: Enhanced Deep Back-Projection Networks for Remote Sensing Scene Image Superresolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 8, pp. 5503–5515, Aug. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8974409/

[93] "Sentinel-2 - Overview." [Online]. Available: https://sentinel.esa.int/web/sentinel/missions/sentinel-2/overview

[94] 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, Dec. 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0924271618302636

[95] F. Palsson, J. Sveinsson, and M. Ulfarsson, "Sentinel-2 Image Fusion Using a Deep Residual Network," *Remote Sensing*, vol. 10, no. 8, p. 1290, Aug. 2018. [Online]. Available: http://www.mdpi.com/2072-4292/10/8/1290

[96] 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, no. 22, p. 2635, Nov. 2019. [Online]. Available: https://www.mdpi.com/2072-4292/11/22/2635

[97] J. Wu, Z. He, and J. Hu, "Sentinel-2 Sharpening via Parallel Residual Network," *Remote Sensing*, vol. 12, no. 2, p. 279, Jan. 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/2/279

[98] X. Luo, X. Tong, and Z. Hu, "Improving Satellite Image Fusion via Generative Adversarial Training," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–14, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9212572/

[99] C. Li and M. Wand, "Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks," *arXiv:1604.04382 [cs]*, Apr. 2016, arXiv: 1604.04382. [Online]. Available: http://arxiv.org/abs/1604.04382

[100] H. V. Nguyen, M. O. Ulfarsson, J. R. Sveinsson, and M. D. Mura, "Sentinel-2 Sharpening Using a Single Unsupervised Convolutional Neural Network With MTF-Based Degradation Model," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 6882–6896, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9464640/

[101] M. Ciotola, M. Ragosta, G. Poggi, and G. Scarpa, "A Full-Resolution Training Framework for Sentinel-2 Image Fusion," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. Brussels, Belgium: IEEE, Jul. 2021, pp. 1260–1263. [Online]. Available: https://ieeexplore.ieee.org/document/9553199/

[102] Z. Shao, J. Cai, P. Fu, L. Hu, and T. Liu, "Deep learning-based fusion of Landsat-8 and Sentinel-2 images for a harmonized surface reflectance product," *Remote Sensing of Environment*, vol. 235, p. 111425, Dec. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0034425719304444

[103] "Landsat 8 Overview | Landsat Science." [Online]. Available: https://landsat.gsfc.nasa.gov/landsat-8/landsat-8-overview

[104] R. Dong, L. Zhang, and H. Fu, "RRSGAN: Reference-Based Super-Resolution for Remote Sensing Image," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–17, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9328132/

[105] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable Convolutional Networks," *arXiv:1703.06211 [cs]*, Jun. 2017, arXiv: 1703.06211. [Online]. Available: http://arxiv.org/abs/1703.06211

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

[107] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[108] Y. Wei, Q. Yuan, H. Shen, and L. Zhang, "Boosting the accuracy of multi-spectral image pan-sharpening by learning a deep residual network," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 10, pp. 1795–1799, Oct. 2017, arXiv: 1705.07556. [Online]. Available: http://arxiv.org/abs/1705.07556

[109] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, "PanNet: A Deep Network Architecture for Pan-Sharpening," in *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice: IEEE, Oct. 2017, pp. 1753–1761. [Online]. Available: http://ieeexplore.ieee.org/document/8237455/

[110] G. Scarpa, S. Vitale, and D. Cozzolino, "Target-adaptive cnn-based pansharpening," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 9, pp. 5443–5457, 2018.

[111] Y. Xing, M. Wang, S. Yang, and L. Jiao, "Pan-sharpening via deep metric learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, pp. 165–183, 2018, deep Learning RS Data. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271618300212

[112] Q. Yuan, Y. Wei, X. Meng, H. Shen, and L. Zhang, "A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 3, pp. 978–989, 2018.

[113] L. He, Y. Rao, J. Li, J. Chanussot, A. Plaza, J. Zhu, and B. Li, "Pansharpening via detail injection based convolutional neural networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 4, pp. 1188–1204, 2019.

[114] S. Luo, S. Zhou, Y. Feng, and J. Xie, "Pansharpening via unsupervised convolutional neural networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 4295–4310, 2020.

[115] L. Liu, J. Wang, E. Zhang, B. Li, X. Zhu, Y. Zhang, and J. Peng, "Shallow–deep convolutional network and spectral-discrimination-based detail injection for multispectral imagery pan-sharpening," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 1772–1783, 2020.

[116] L.-J. Deng, G. Vivone, C. Jin, and J. Chanussot, "Detail injection-based deep convolutional neural networks for pansharpening," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–16, 2020.

[117] J. Cai and B. Huang, "Super-resolution-guided progressive pansharpening based on a deep convolutional neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 6, pp. 5206–5220, 2021.

[118] W. Dong, T. Zhang, J. Qu, S. Xiao, J. Liang, and Y. Li, "Laplacian pyramid dense network for hyperspectral pansharpening," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–13, 2021.

[119] M. Jiang, H. Shen, J. Li, Q. Yuan, and L. Zhang, "A differential information residual convolutional neural network for pansharpening," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 163, pp. 257–271, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092427162030068X

[120] Y. Qu, R. K. Baghbaderani, H. Qi, and C. Kwan, "Unsupervised pansharpening based on self-attention mechanism," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 4, pp. 3192–3208, 2021.

[121] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[122] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 11 057–11 066.

[123] H. Zhang and J. Ma, "Gtp-pnet: A residual learning network based on gradient transformation prior for pansharpening," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 172, pp. 223–239, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092427162030352X

[124] H. Yin, "Pscsc-net: A deep coupled convolutional sparse coding network for pansharpening," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–16, 2021.

[125] Z.-C. Wu, T.-Z. Huang, L.-J. Deng, J.-F. Hu, and G. Vivone, "Vo+net: An adaptive approach using variational optimization and deep learning for panchromatic sharpening," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–16, 2021.

[126] L. Zhang, J. Zhang, J. Ma, and X. Jia, "Sc-pnn: Saliency cascade convolutional neural network for pansharpening," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–19, 2021.

[127] I. Selesnick, R. Baraniuk, and N. Kingsbury, "The dual-tree complex wavelet transform," *IEEE Signal Processing Magazine*, vol. 22, no. 6, pp. 123–151, 2005.

[128] S. Vitale and G. Scarpa, "A detail-preserving cross-scale learning strategy for cnn-based pansharpening," *Remote Sensing*, vol. 12, no. 3, 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/3/348

[129] G. Scarpa, S. Vitale, and D. Cozzolino, "Target-adaptive cnn-based pansharpening," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 9, pp. 5443–5457, 2018.

[130] M. Ciotola, S. Vitale, A. Mazza, G. Poggi, and G. Scarpa, "Pansharpening by convolutional neural networks in the full resolution framework," *arXiv preprint arXiv:2111.08334*, 2021.

[131] X. Liu, Y. Wang, and Q. Liu, "Psgan: A generative adversarial network for remote sensing image pan-sharpening," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 873–877.

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

[133] A. Gastineau, J.-F. Aujol, Y. Berthoumieu, and C. Germain, "Generative adversarial network for pansharpening with spectral and spatial discriminators," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–11, 2021.

[134] F. Ozcelik, U. Alganci, E. Sertel, and G. Unal, "Rethinking cnn-based pansharpening: Guided colorization of panchromatic images via gans," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[135] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "Multispectral and hyperspectral image fusion using a 3-d-convolutional neural network," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, p. 639–643, May 2017.

[136] R. Dian, S. Li, A. Guo, and L. Fang, "Deep hyperspectral image sharpening," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 11, pp. 5345–5355, 2018.

[137] F. Zhou, R. Hang, Q. Liu, and X. Yuan, "Pyramid fully convolutional network for hyperspectral and multispectral image fusion," *IEEE*

Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 12, no. 5, p. 1549–1558, May 2019.

[138] X. Han, J. Yu, J. Luo, and W. Sun, "Hyperspectral and multispectral image fusion using cluster-based multi-branch bp neural networks," *Remote Sensing*, vol. 11, no. 1010, p. 1173, Jan 2019. [Online]. Available: https://www.mdpi.com/2072-4292/11/10/1173

[139] L. He, J. Zhu, J. Li, A. Plaza, J. Chanussot, and B. Li, "Hyperpnn: Hyperspectral pansharpening via spectrally predictive convolutional neural networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 8, pp. 3092–3100, 2019.

[140] K. Li, W. Xie, Q. Du, and Y. Li, "Ddlps: Detail-based deep laplacian pansharpening for hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 10, pp. 8011–8025, 2019.

[141] D. Shen, J. Liu, Z. Xiao, J. Yang, and L. Xiao, "A twice optimizing net with matrix decomposition for hyperspectral and multispectral image fusion," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, p. 4095–4110, 2020.

[142] Q. Xie, M. Zhou, Q. Zhao, Z. Xu, and D. Meng, "Mhf-net: An interpretable deep network for multispectral and hyperspectral image fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1–1, 2020.

[143] S. Liu, S. Miao, J. Su, B. Li, W. Hu, and Y.-D. Zhang, "Umag-net: A new unsupervised multiattention-guided network for hyperspectral and multispectral image fusion," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, p. 7373–7385, 2021.

[144] X. Zhang, W. Huang, Q. Wang, and X. Li, "Ssr-net: Spatial–spectral reconstruction network for hyperspectral and multispectral image fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, p. 5953–5965, Jul 2021.

[145] Y. Bengio, "Learning deep architectures for ai," *Found. Trends Mach. Learn.*, vol. 2, no. 1, p. 1–127, Jan. 2009. [Online]. Available: https://doi.org/10.1561/2200000006

[146] "MODIS Technical specifications." [Online]. Available: https://modis.gsfc.nasa.gov/about/specifications.php

[147] X. Liu, C. Deng, J. Chanussot, D. Hong, and B. Zhao, "StfNet : A Two-Stream Convolutional Neural Network for Spatiotemporal Image Fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6552–6564, Sep. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8693668/

[148] W. Li, X. Zhang, Y. Peng, and M. Dong, "DMNet: A Network Architecture Using Dilated Convolution and Multiscale Mechanisms for Spatiotemporal Fusion of Remote Sensing Images," *IEEE Sensors Journal*, vol. 20, no. 20, pp. 12 190–12 202, Oct. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9109314/

[149] ——, "Spatiotemporal Fusion of Remote Sensing Images using a Convolutional Neural Network with Attention and Multiscale Mechanisms," *International Journal of Remote Sensing*, vol. 42, no. 6, pp. 1973–1993, Mar. 2021. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/01431161.2020.1809742

[150] D. Jia, C. Song, C. Cheng, S. Shen, L. Ning, and C. Hui, "A Novel Deep Learning-Based Spatiotemporal Fusion Method for Combining Satellite Images with Different Resolutions Using a Two-Stream Convolutional Neural Network," *Remote Sensing*, vol. 12, no. 4, p. 698, Feb. 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/4/698

[151] S. Yang and X. Wang, "Sparse Representation and SRCNN based Spatio-temporal Information Fusion Method of Multi-sensor Remote Sensing Data," *Journal of Network Intelligence*, vol. 6, no. 1, pp. 40–53, 2021.

[152] M. Peng, L. Zhang, X. Sun, Y. Cen, and X. Zhao, "A Fast Three-Dimensional Convolutional Neural Network-Based Spatiotemporal Fusion Method (STF3DCNN) Using a Spatial-Temporal-Spectral Dataset," *Remote Sensing*, vol. 12, no. 23, p. 3888, Nov. 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/23/3888

[153] Y. Li, J. Li, L. He, J. Chen, and A. Plaza, "A new sensor bias-driven spatio-temporal fusion model based on convolutional neural networks," *Science China Information Sciences*, vol. 63, no. 4, p. 140302, Apr. 2020. [Online]. Available: http://link.springer.com/10.1007/s11432-019-2805-y

[154] H. Song, Q. Liu, G. Wang, R. Hang, and B. Huang, "Spatiotemporal Satellite Image Fusion Using Deep Convolutional Neural Networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 3, pp. 821–829, Mar. 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8291042/

[155] J. Chen, L. Wang, R. Feng, P. Liu, W. Han, and X. Chen, "CycleGAN-STF: Spatiotemporal Fusion via CycleGAN-Based Image Generation," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–15, 2020.

[156] H. Zhang, Y. Song, C. Han, and L. Zhang, "Remote Sensing Image Spatiotemporal Fusion Using a Generative Adversarial Network," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–14, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9159647/

[157] X. Wang and X. Wang, "Spatiotemporal Fusion of Remote Sensing Image Based on Deep Learning," *Journal of Sensors*, vol. 2020, pp. 1–11, Jun. 2020. [Online]. Available: https://www.hindawi.com/journals/js/2020/8873079/

[158] Y. Zheng, H. Song, L. Sun, Z. Wu, and B. Jeon, "Spatiotemporal Fusion of Satellite Images via Very Deep Convolutional Networks," *Remote Sensing*, vol. 11, no. 22, p. 2701, Nov. 2019. [Online]. Available: https://www.mdpi.com/2072-4292/11/22/2701

[159] Z. Tan, P. Yue, L. Di, and J. Tang, "Deriving High Spatiotemporal Remote Sensing Images Using Deep Convolutional Network," *Remote Sensing*, vol. 10, no. 7, p. 1066, Jul. 2018. [Online]. Available: http://www.mdpi.com/2072-4292/10/7/1066

[160] Feng Gao, J. Masek, M. Schwaller, and F. Hall, "On the blending of the Landsat and MODIS surface reflectance: predicting daily Landsat surface reflectance," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 8, pp. 2207–2218, Aug. 2006. [Online]. Available: http://ieeexplore.ieee.org/document/1661809/

[161] Z. Tan, L. Di, M. Zhang, L. Guo, and M. Gao, "An Enhanced Deep Convolutional Model for Spatiotemporal Image Fusion," *Remote Sensing*, vol. 11, no. 24, p. 2898, Dec. 2019. [Online]. Available: https://www.mdpi.com/2072-4292/11/24/2898

[162] S. Bouabid, M. Chernetskiy, M. Rischard, and J. Gamper, "Predicting Landsat Reflectance with Deep Generative Fusion," *arXiv:2011.04762 [cs, eess]*, Nov. 2020, arXiv: 2011.04762. [Online]. Available: http://arxiv.org/abs/2011.04762

[163] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," *arXiv:1611.07004 [cs]*, Nov. 2018, arXiv: 1611.07004. [Online]. Available: http://arxiv.org/abs/1611.07004

[164] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," *arXiv:1703.10593 [cs]*, Aug. 2020, arXiv: 1703.10593. [Online]. Available: http://arxiv.org/abs/1703.10593

[165] X. Zhu, E. H. Helmer, F. Gao, D. Liu, J. Chen, and M. A. Lefsky, "A flexible spatiotemporal method for fusing satellite images with different resolutions," *Remote Sensing of Environment*, vol. 172, pp. 165–177, Jan. 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0034425715302042

[166] Z. Tan, M. Gao, X. Li, and L. Jiang, "A Flexible Reference-Insensitive Spatiotemporal Fusion Model for Remote Sensing Images Using Conditional Generative Adversarial Network," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–13, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9336033/

[167] P. Luo, J. Ren, Z. Peng, R. Zhang, and J. Li, "Differentiable Learning-to-Normalize via Switchable Normalization," *arXiv:1806.10779 [cs]*, Apr. 2019, arXiv: 1806.10779. [Online]. Available: http://arxiv.org/abs/1806.10779

[168] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral Normalization for Generative Adversarial Networks," *arXiv:1802.05957 [cs, stat]*, Feb. 2018, arXiv: 1802.05957. [Online]. Available: http://arxiv.org/abs/1802.05957

[169] T.-A. Teo and Y.-J. Fu, "Spatiotemporal Fusion of Formosat-2 and Landsat-8 Satellite Images: A Comparison of "Super Resolution-Then-Blend" and "Blend-Then-Super Resolution" Approaches," *Remote Sensing*, vol. 13, no. 4, p. 606, Feb. 2021. [Online]. Available: https://www.mdpi.com/2072-4292/13/4/606

[170] D. Jia, C. Cheng, C. Song, S. Shen, L. Ning, and T. Zhang, "A Hybrid Deep Learning-Based Spatiotemporal Fusion Method for Combining Satellite Images with Different Resolutions," *Remote Sensing*, vol. 13, no. 4, p. 645, Feb. 2021. [Online]. Available: https://www.mdpi.com/2072-4292/13/4/645

[171] S. Lei, Z. Shi, and Z. Zou, "Super-Resolution for Remote Sensing Images via Local–Global Combined Network," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 8, pp. 1243–1247, Aug. 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7937881/

[172] J. M. Haut, M. E. Paoletti, R. Fernandez-Beltran, J. Plaza, A. Plaza, and J. Li, "Remote Sensing Single-Image Superresolution Based on a Deep Compendium Model," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 9, pp. 1432–1436, Sep. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8660433/

[173] T. Lu, J. Wang, Y. Zhang, Z. Wang, and J. Jiang, "Satellite Image Super-Resolution via Multi-Scale Residual Deep Neural Network,"

Remote Sensing, vol. 11, no. 13, p. 1588, Jul. 2019. [Online]. Available: https://www.mdpi.com/2072-4292/11/13/1588

[174] W. Xu, C. Zhang, and M. Wu, "Multi-scale Deep Residual Network for Satellite Image Super-Resolution Reconstruction," in *Pattern Recognition and Computer Vision*, Z. Lin, L. Wang, J. Yang, G. Shi, T. Tan, N. Zheng, X. Chen, and Y. Zhang, Eds. Cham: Springer International Publishing, 2019, vol. 11859, pp. 332–340, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-030-31726-3_28

[175] L. Yan and K. Chang, "A New Super Resolution Framework Based on Multi-Task Learning for Remote Sensing Images," *Sensors*, vol. 21, no. 5, p. 1743, Mar. 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/5/1743

[176] M. Qin, S. Mavromatis, L. Hu, F. Zhang, R. Liu, J. Sequeira, and Z. Du, "Remote Sensing Single-Image Resolution Improvement Using A Deep Gradient-Aware Network with Image-Specific Enhancement," *Remote Sensing*, vol. 12, no. 5, p. 758, Feb. 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/5/758

[177] M. Galar, R. Sesma, C. Ayala, L. Albizua, and C. Aranda, "LEARNING SUPER-RESOLUTION FOR SENTINEL-2 IMAGES WITH REAL GROUND TRUTH DATA FROM A REFERENCE SATELLITE," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. V-1-2020, pp. 9–16, Aug. 2020. [Online]. Available: https://www.isprs-ann-photogramm-remote-sens-spatial-inf-sci.net/V-1-2020/9/2020/

[178] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image Style Transfer Using Convolutional Neural Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016.

[179] D. Pouliot, R. Latifovic, J. Pasher, and J. Duffe, "Landsat Super-Resolution Enhancement Using Convolution Neural Networks and Sentinel-2 for Training," *Remote Sensing*, vol. 10, no. 3, p. 394, Mar. 2018. [Online]. Available: http://www.mdpi.com/2072-4292/10/3/394

[180] C. B. Collins, J. M. Beck, S. M. Bridges, J. A. Rushing, and S. J. Graves, "Deep learning for multisensor image resolution enhancement," in *Proceedings of the 1st Workshop on Artificial Intelligence and Deep Learning for Geographic Knowledge Discovery*. Los Angeles California: ACM, Nov. 2017, pp. 37–44. [Online]. Available: https://dl.acm.org/doi/10.1145/3149808.3149815

[181] M. M. Sheikholeslami, S. Nadi, A. A. Naeini, and P. Ghamisi, "An Efficient Deep Unsupervised Superresolution Model for Remote Sensing Images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 1937–1945, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9086776/

[182] K. Turkowski, "FILTERS FOR COMMON RESAMPLING TASKS," in *Graphics Gems*. Elsevier, 1990, pp. 147–165. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/B9780080507538500425

[183] N. Zhang, Y. Wang, X. Zhang, D. Xu, X. Wang, G. Ben, Z. Zhao, and Z. Li, "A Multi-Degradation Aided Method for Unsupervised Remote Sensing Image Super Resolution With Convolution Neural Networks," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–14, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9301235/

[184] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3262–3271.

[185] W. Ma, Z. Pan, J. Guo, and B. Lei, "Achieving Super-Resolution Remote Sensing Images via the Wavelet Transform Combined With the Recursive Res-Net," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 6, pp. 3512–3527, Jun. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8600724/

[186] Q. Qin, J. Dou, and Z. Tu, "Deep ResNet Based Remote Sensing Image Super-Resolution Reconstruction in Discrete Wavelet Domain," *Pattern Recognition and Image Analysis*, vol. 30, no. 3, pp. 541–550, Jul. 2020. [Online]. Available: http://link.springer.com/10.1134/S1054661820030232

[187] X. Feng, W. Zhang, X. Su, and Z. Xu, "Optical Remote Sensing Image Denoising and Super-Resolution Reconstructing Using Optimized Generative Network in Wavelet Transform Domain," *Remote Sensing*, vol. 13, no. 9, p. 1858, May 2021. [Online]. Available: https://www.mdpi.com/2072-4292/13/9/1858

[188] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[189] X. Dong, Z. Xi, X. Sun, and L. Gao, "Transferred Multi-Perception Attention Networks for Remote Sensing Image Super-Resolution," *Remote Sensing*, vol. 11, no. 23, p. 2857, Dec. 2019. [Online]. Available: https://www.mdpi.com/2072-4292/11/23/2857

[190] J. Gu, X. Sun, Y. Zhang, K. Fu, and L. Wang, "Deep Residual Squeeze and Excitation Network for Remote Sensing Image Super-Resolution," *Remote Sensing*, vol. 11, no. 15, p. 1817, Aug. 2019. [Online]. Available: https://www.mdpi.com/2072-4292/11/15/1817

[191] J. M. Haut, R. Fernandez-Beltran, M. E. Paoletti, J. Plaza, and A. Plaza, "Remote Sensing Image Superresolution Using Deep Residual Channel Attention," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 11, pp. 9277–9289, Nov. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8770258/

[192] S. Zhang, Q. Yuan, J. Li, J. Sun, and X. Zhang, "Scene-Adaptive Remote Sensing Image Super-Resolution Using a Multiscale Attention Network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 7, pp. 4764–4779, Jul. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8978758/

[193] X. Dong, X. Sun, X. Jia, Z. Xi, L. Gao, and B. Zhang, "Remote Sensing Image Super-Resolution Using Novel Dense-Sampling Networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 2, pp. 1618–1633, Feb. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9107103/

[194] X. Wang, Y. Wu, Y. Ming, and H. Lv, "Remote Sensing Imagery Super Resolution Based on Adaptive Multi-Scale Feature Fusion Network," *Sensors*, vol. 20, no. 4, p. 1142, Feb. 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/4/1142

[195] P. Lei and C. Liu, "Inception residual attention network for remote sensing image super-resolution," *International Journal of Remote Sensing*, vol. 41, no. 24, pp. 9565–9587, Dec. 2020. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/01431161.2020.1800129

[196] H. Wang, Q. Hu, C. Wu, J. Chi, and X. Yu, "Non-Locally up-Down Convolutional Attention Network for Remote Sensing Image Super-Resolution," *IEEE Access*, vol. 8, pp. 166 304–166 319, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9189886/

[197] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[198] Y. Peng, X. Wang, J. Zhang, and S. Liu, "Pre-training of gated convolution neural network for remote sensing image super-resolution," *IET Image Processing*, vol. 15, no. 5, pp. 1179–1188, Apr. 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1049/ipr2.12096

[199] S. Lei and Z. Shi, "Hybrid-Scale Self-Similarity Exploitation for Remote Sensing Image Super-Resolution," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–10, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9400474/

[200] Y. Chang and B. Luo, "Bidirectional Convolutional LSTM Neural Network for Remote Sensing Image Super-Resolution," *Remote Sensing*, vol. 11, no. 20, p. 2333, Oct. 2019. [Online]. Available: https://www.mdpi.com/2072-4292/11/20/2333

[201] S. Lei, Z. Shi, and Z. Zou, "Coupled Adversarial Training for Remote Sensing Image Super-Resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 5, pp. 3633–3643, May 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8946581/

[202] W. Ma, Z. Pan, F. Yuan, and B. Lei, "Super-Resolution of Remote Sensing Images via a Dense Residual Generative Adversarial Network," *Remote Sensing*, vol. 11, no. 21, p. 2578, Nov. 2019. [Online]. Available: https://www.mdpi.com/2072-4292/11/21/2578

[203] L. Salgueiro Romero, J. Marcello, and V. Vilaplana, "Super-Resolution of Sentinel-2 Imagery Using Generative Adversarial Networks," *Remote Sensing*, vol. 12, no. 15, p. 2424, Jul. 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/15/2424

[204] Z. Wang, K. Jiang, P. Yi, Z. Han, and Z. He, "Ultra-dense GAN for satellite imagery super-resolution," *Neurocomputing*, vol. 398, pp. 328–337, Jul. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0925231219314602

[205] C. Shin, S. Kim, and Y. Kim, "Satellite Image Target Super-Resolution With Adversarial Shape Discriminator," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9302745/

[206] Y. Gong, P. Liao, X. Zhang, L. Zhang, G. Chen, K. Zhu, X. Tan, and Z. Lv, "Enlighten-GAN for Super Resolution Reconstruction in Mid-Resolution Remote Sensing Images," *Remote Sensing*, vol. 13, no. 6, p. 1104, Mar. 2021. [Online]. Available: https://www.mdpi.com/2072-4292/13/6/1104

[207] K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, and J. Jiang, "Edge-Enhanced GAN for Remote Sensing Image Superresolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57,

no. 8, pp. 5799–5812, Aug. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8677274/

[208] Y. Li, S. Mavromatis, F. Zhang, Z. Du, J. Sequeira, Z. Wang, X. Zhao, and R. Liu, "Single-Image Super-Resolution for Remote Sensing Images Using a Deep Generative Adversarial Network With Local and Global Attention Mechanisms," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–24, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9479919/

[209] M. Kawulok, P. Benecki, S. Piechaczek, K. Hrynczenko, D. Kostrzewa, and J. Nalepa, "Deep Learning for Multiple-Image Super-Resolution," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 6, pp. 1062–1066, Jun. 2020, arXiv: 1903.00440. [Online]. Available: http://arxiv.org/abs/1903.00440

[210] M. Kawulok, P. Benecki, D. Kostrzewa, and L. Skonieczny, "Towards Evolutionary Super-Resolution," in *Applications of Evolutionary Computation*, K. Sim and P. Kaufmann, Eds. Cham: Springer International Publishing, 2018, vol. 10784, pp. 480–496, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-319-77538-8_33

[211] M. Märtens, D. Izzo, A. Krzic, and D. Cox, "Super-resolution of PROBA-V images using convolutional neural networks," *Astrodynamics*, vol. 3, no. 4, pp. 387–402, Dec. 2019. [Online]. Available: http://link.springer.com/10.1007/s42064-019-0059-8

[212] A. B. Molini, D. Valsesia, G. Fracastoro, and E. Magli, "DeepSUM: Deep neural network for Super-resolution of Unregistered Multitemporal images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 5, pp. 3644–3656, May 2020, arXiv: 1907.06490. [Online]. Available: http://arxiv.org/abs/1907.06490

[213] ——, "Deepsum++: Non-Local Deep Neural Network for Super-Resolution of Unregistered Multitemporal Images," in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*. Waikoloa, HI, USA: IEEE, Sep. 2020, pp. 609–612. [Online]. Available: https://ieeexplore.ieee.org/document/9324418/

[214] M. Deudon, A. Kalaitzis, I. Goytom, M. R. Arefin, Z. Lin, K. Sankaran, V. Michalski, S. E. Kahou, J. Cornebise, and Y. Bengio, "HighRes-net: Recursive Fusion for Multi-Frame Super-Resolution of Satellite Imagery," *arXiv:2002.06460 [cs, eess, stat]*, Feb. 2020, arXiv: 2002.06460. [Online]. Available: http://arxiv.org/abs/2002.06460

[215] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Deep Image Homography Estimation," *arXiv:1606.03798 [cs]*, Jun. 2016, arXiv: 1606.03798. [Online]. Available: http://arxiv.org/abs/1606.03798

[216] M. Rifat Arefin, V. Michalski, P.-L. St-Charles, A. Kalaitzis, S. Kim, S. E. Kahou, and Y. Bengio, "Multi-Image Super-Resolution for Remote Sensing using Deep Recurrent Networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Seattle, WA, USA: IEEE, Jun. 2020, pp. 816–825. [Online]. Available: https://ieeexplore.ieee.org/document/9150720/

[217] N. Ballas, L. Yao, C. Pal, and A. Courville, "Delving Deeper into Convolutional Networks for Learning Video Representations," *arXiv:1511.06432 [cs]*, Mar. 2016, arXiv: 1511.06432. [Online]. Available: http://arxiv.org/abs/1511.06432

[218] F. Salvetti, V. Mazzia, A. Khaliq, and M. Chiaberge, "Multi-Image Super Resolution of Remotely Sensed Images Using Residual Attention Deep Neural Networks," *Remote Sensing*, vol. 12, no. 14, p. 2207, Jul. 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/14/2207

[219] J. Ma, L. Zhang, and J. Zhang, "SD-GAN: Saliency-Discriminated GAN for Remote Sensing Image Superresolution," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 11, pp. 1973–1977, Nov. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8933080/

[220] H. Wu, L. Zhang, and J. Ma, "Remote Sensing Image Super-Resolution via Saliency-Guided Feedback GANs," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–16, 2020.

[221] L. Zhang, D. Chen, J. Ma, and J. Zhang, "Remote-Sensing Image Superresolution Based on Visual Saliency Analysis and Unequal Reconstruction Networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 6, pp. 4099–4115, Jun. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8954940/

[222] L. Zhang, J. Ma, X. Lv, and D. Chen, "Hierarchical Weakly Supervised Learning for Residential Area Semantic Segmentation in Remote Sensing Images," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 1, pp. 117–121, Jan. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8720000/

[223] L. Wang, M. Zheng, W. Du, M. Wei, and L. Li, "Super-resolution sar image reconstruction via generative adversarial network," in *2018 12th International Symposium on Antennas, Propagation and EM Theory (ISAPE)*, 2018, pp. 1–4.

[224] F. Gu, H. Zhang, C. Wang, and F. Wu, "Sar image super-resolution based on noise-free generative adversarial network," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 2575–2578.

[225] Y. Li, D. Ao, C. O. Dumitru, C. Hu, and M. Datcu, "Super-resolution of geosynchronous synthetic aperture radar images using dialectical gans," *Science China Information Sciences*, vol. 62, no. 10, p. 209302, Apr 2019. [Online]. Available: https://doi.org/10.1007/s11432-018-9668-6

[226] X. Cen, X. Song, Y. Li, and C. Wu, "A deep learning-based super-resolution model for bistatic sar image," in *2021 International Conference on Electronics, Circuits and Information Engineering (ECIE)*, 2021, pp. 228–233.

[227] H. Shen, L. Lin, J. Li, Q. Yuan, and L. Zhao, "A residual convolutional neural network for polarimetric sar image super-resolution," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 161, pp. 90–108, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092427162030006X

[228] J. Yu, W. Li, Z. Li, J. Wu, H. Yang, and J. Yang, "Sar image super-resolution base on weighted dense connected convolutional network," in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, 2020, pp. 2101–2104.

[229] L. Lin, J. Li, Q. Yuan, and H. Shen, "Polarimetric sar image super-resolution via deep convolutional neural network," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 3205–3208.

[230] P. Wang, H. Zhang, and V. M. Patel, "Sar image despeckling using a convolutional neural network," *IEEE Signal Processing Letters*, vol. 24, no. 12, p. 1763–1767, Dec 2017. [Online]. Available: http://dx.doi.org/10.1109/LSP.2017.2758203

[231] D. Ao, C. O. Dumitru, G. Schwarz, and M. Datcu, "Dialectical gan for sar image translation: From sentinel-1 to terrasar-x," *Remote Sensing*, vol. 10, no. 10, 2018. [Online]. Available: https://www.mdpi.com/2072-4292/10/10/1597

[232] K. A. H. Kelany, A. Baniasadi, N. Dimopoulos, and M. Gara, "Improving insar image quality and co-registration through cnn-based super-resolution," in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2020, pp. 1–5.

[233] T. Wang, W. Sun, H. Qi, and P. Ren, "Aerial image super resolution via wavelet multiscale convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 769–773, 2018.

[234] D. González, M. A. Patricio, A. Berlanga, and J. M. Molina, "A super-resolution enhancement of uav images based on a convolutional neural network for mobile devices," *Personal and Ubiquitous Computing*, pp. 1–12, 2019.

[235] N. Q. Truong, P. H. Nguyen, S. H. Nam, and K. R. Park, "Deep learning-based super-resolution reconstruction and marker detection for drone landing," *IEEE Access*, vol. 7, pp. 61 639–61 655, 2019.

[236] F. Liu, Q. Yu, L. Chen, G. Jeon, M. K. Albertini, and X. Yang, "Aerial image super-resolution based on deep recursive dense network for disaster area surveillance," *Personal and Ubiquitous Computing*, pp. 1–10, 2021.

[237] J. Zhou, C.-M. Vong, Q. Liu, and Z. Wang, "Scale adaptive image cropping for uav object detection," *Neurocomputing*, vol. 366, pp. 305–313, 2019.

[238] H. Chen, Z. He, B. Shi, and T. Zhong, "Research on recognition method of electrical components based on yolo v3," *IEEE Access*, vol. 7, pp. 157 818–157 829, 2019.

[239] M. Aslahishahri, K. G. Stanley, H. Duddu, S. Shirtliffe, S. Vail, and I. Stavness, "Spatial super resolution of real-world aerial images for image-based plant phenotyping," *Remote Sensing*, vol. 13, no. 12, p. 2308, 2021.

[240] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '10*. San Jose, California: ACM Press, 2010, p. 270. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1869790.1869829

[241] G. Sheng, W. Yang, T. Xu, and H. Sun, "High-resolution satellite scene classification using a sparse coding based multiple feature combination," *International Journal of Remote Sensing*, vol. 33, no. 8, pp. 2395–2412, Apr. 2012. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/01431161.2011.608740

[242] J. Hu, T. Jiang, X. Tong, G.-S. Xia, and L. Zhang, "A benchmark for scene classification of high spatial resolution remote sensing imagery," in *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. Milan, Italy: IEEE, Jul. 2015, pp. 5003–5006. [Online]. Available: http://ieeexplore.ieee.org/document/7326956/

[243] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep Learning Based Feature Selection for Remote Sensing Scene Classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 11, pp. 2321–2325, Nov. 2015. [Online]. Available: http://ieeexplore.ieee.org/document/7272047/

[244] L. Zhao, P. Tang, and L. Huo, "Feature significance-based multibag-of-visual-words model for remote sensing image scene classification," *Journal of Applied Remote Sensing*, vol. 10, no. 3, p. 035004, Jul. 2016. [Online]. Available: http://remotesensing.spiedigitallibrary.org/article.aspx?doi=10.1117/1.JRS.10.035004

[245] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7907303/

[246] G. Cheng, J. Han, and X. Lu, "Remote Sensing Image Scene Classification: Benchmark and State of the Art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017, arXiv: 1703.00121. [Online]. Available: http://arxiv.org/abs/1703.00121

[247] B. Zhao, Y. Zhong, G.-S. Xia, and L. Zhang, "Dirichlet-Derived Multiple Topic Scene Classification Model for High Spatial Resolution Remote Sensing Imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 4, pp. 2108–2123, Apr. 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7329997/

[248] O. A. B. Penatti, K. Nogueira, and J. A. dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Boston, MA, USA: IEEE, Jun. 2015, pp. 44–51. [Online]. Available: http://ieeexplore.ieee.org/document/7301382/

[249] M. Schmitt, L. H. Hughes, and X. X. Zhu, "The SEN1-2 Dataset for Deep Learning in SAR-Optical Data Fusion," *arXiv:1807.01569 [cs]*, Jul. 2018, arXiv: 1807.01569. [Online]. Available: http://arxiv.org/abs/1807.01569

[250] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, "SEN12MS – A Curated Dataset of Georeferenced Multi-Spectral Sentinel-1/2 Imagery for Deep Learning and Data Fusion," *arXiv:1906.07789 [cs]*, Jun. 2019, arXiv: 1906.07789. [Online]. Available: http://arxiv.org/abs/1906.07789

[251] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "DOTA: A Large-scale Dataset for Object Detection in Aerial Images," *arXiv:1711.10398 [cs]*, May 2019, arXiv: 1711.10398. [Online]. Available: http://arxiv.org/abs/1711.10398

[252] I. V. Emelyanova, T. R. McVicar, T. G. Van Niel, L. T. Li, and A. I. van Dijk, "Assessing the accuracy of blending Landsat–MODIS surface reflectances in two landscapes with contrasting spatial and temporal dynamics: A framework for algorithm selection," *Remote Sensing of Environment*, vol. 133, pp. 193–209, Jun. 2013. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0034425713000473

[253] J. Li, Y. Li, L. He, J. Chen, and A. Plaza, "Spatio-temporal fusion for remote sensing data: an overview and new benchmark," *Science China Information Sciences*, vol. 63, no. 4, p. 140301, Apr. 2020. [Online]. Available: https://link.springer.com/10.1007/s11432-019-2785-y

[254] X.-Y. Tong, G.-S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sensing of Environment*, vol. 237, p. 111322, Feb. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0034425719303414

[255] "Draper Satellite Image Chronology." [Online]. Available: https://kaggle.com/c/draper-satellite-image-chronology

[256] P. Wei, Z. Xie, H. Lu, Z. Zhan, Q. Ye, W. Zuo, and L. Lin, "Component divide-and-conquer for real-world image super-resolution," in *European Conference on Computer Vision*. Springer, 2020, pp. 101–117.

[257] Y. Zheng, J. Li, Y. Li, J. Guo, X. Wu, and J. Chanussot, "Hyperspectral pansharpening using deep prior and dual attention residual network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 11, pp. 8059–8076, 2020.

[258] W. Xie, Y. Cui, Y. Li, J. Lei, Q. Du, and J. Li, "Hpgan: Hyperspectral pansharpening using 3-d generative adversarial networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 463–477, 2021.

[259] J. Gu, H. Lu, W. Zuo, and C. Dong, "Blind Super-Resolution With Iterative Kernel Correction," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[260] V. Cornillère, A. Djelouah, W. Yifan, O. Sorkine-Hornung, and C. Schroers, "Blind image super resolution with spatially variant degradations," *ACM Transactions on Graphics (proceedings of ACM SIGGRAPH ASIA)*, vol. 38, no. 6, 2019.

[261] Z. Luo, Y. Huang, S. Li, L. Wang, and T. Tan, "Unfolding the Alternating Optimization for Blind Super Resolution," *arXiv:2010.02631 [cs]*, Nov. 2020, arXiv: 2010.02631. [Online]. Available: http://arxiv.org/abs/2010.02631

[262] Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, and L. Lin, "Unsupervised Image Super-Resolution using Cycle-in-Cycle Generative Adversarial Networks," *arXiv:1809.00437 [cs]*, Sep. 2018, arXiv: 1809.00437. [Online]. Available: http://arxiv.org/abs/1809.00437

[263] G. Kim, J. Park, K. Lee, J. Lee, J. Min, B. Lee, D. K. Han, and H. Ko, "Unsupervised Real-World Super Resolution with Cycle Generative Adversarial Network and Domain Discriminator," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Seattle, WA, USA: IEEE, Jun. 2020, pp. 1862–1871. [Online]. Available: https://ieeexplore.ieee.org/document/9150730/

[264] S. Maeda, "Unpaired Image Super-Resolution using Pseudo-Supervision," *arXiv:2002.11397 [cs, eess]*, Feb. 2020, arXiv: 2002.11397. [Online]. Available: http://arxiv.org/abs/2002.11397

[265] Y. Zhang, S. Liu, C. Dong, X. Zhang, and Y. Yuan, "Multiple Cycle-in-Cycle Generative Adversarial Networks for Unsupervised Image Super-Resolution," *IEEE Transactions on Image Processing*, vol. 29, pp. 1101–1112, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8825849/

[266] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning*. PMLR, 2015, pp. 2256–2265.

[267] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *CoRR*, vol. abs/2104.07636, 2021. [Online]. Available: https://arxiv.org/abs/2104.07636

[268] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253519308103