Deep Learning for Downscaling Remote Sensing Images

Fusion and super-resolution

MARIA SDRAKA, IOANNIS PAPOUTSIS, BILL PSOMAS,
KONSTANTINOS VLACHOS, KONSTANTINOS IOANNIDIS,
KONSTANTINOS KARANTZALOS, ILIAS GIALAMPOUKIDIS,
AND STEFANOS VROCHIDIS

xxxxx

The past few years have seen an accelerating integration of deep learning (DL) techniques into various remote sensing (RS) applications, highlighting their power to adapt and achieving unprecedented advancements. In the present review, we provide an exhaustive exploration of the DL approaches proposed specifically for the spatial downscaling of RS imagery. A key contribution of our work is the presentation of the major architectural components and models, metrics, and data sets available for this task as well as 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 (CV) practitioners and researchers as a source of inspiration and constructive insight.

MOTIVATION

Recent technological advances have significantly increased the volume and distribution rate of 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 with 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 are many operational satellites orbiting 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 resolutions [1]. The *spatial resolution* (or the *ground sample distance*) refers to 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* refers to the range of the electromagnetic (EM) spectrum (wavebands) that the sensor acquires observations in, while *temporal resolution* (or *revisit time*) refers to 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 are usually tradeoffs among these factors, and no available sensor can capture information at the highest possible spatial and temporal resolution across all wavebands.

Digital Object Identifier 10.1109/MGRS.2022.3171836
Date of current version: 2 June 2022

Therefore, one of the hottest topics in RS is the fusion of multisource 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 DL methods and makes up an essential part of 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 articles were published recently that, 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 as well as a valuable summary of the most recent literature on enhancing the spatial resolution of satellite imagery data, specifically, using advanced DL architectures. These DL models are tailored to EO data with their unique and heterogeneous spatial, temporal, and spectral characteristics, which differ significantly from the imagery traditionally used by the CV community.

In fact, research on CV applications has motivated the production of valuable review articles, mainly for (nonsatellite) image super-resolution (SR), like [10]–[15] and [16]. Our work exclusively targets the RS field and provides a broader overview of methods and applications than [17]–[19], which focus solely on pansharpening approaches, or [20], which only examines single-image SR (SISR) 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] and [24]) focus on the state of the art of multimodal data fusion, partially addressing image resolution enhancement without focusing on DL techniques. Finally, a study similar 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 publications on RS image SR have been steadily increasing, the ratio of studies that use DL has blown up, from 5% in 2017 to almost 40% in 2020 (Figure 1). Similarly, in CV, publications on DL for image SR [26] have exhibited a steady increase.

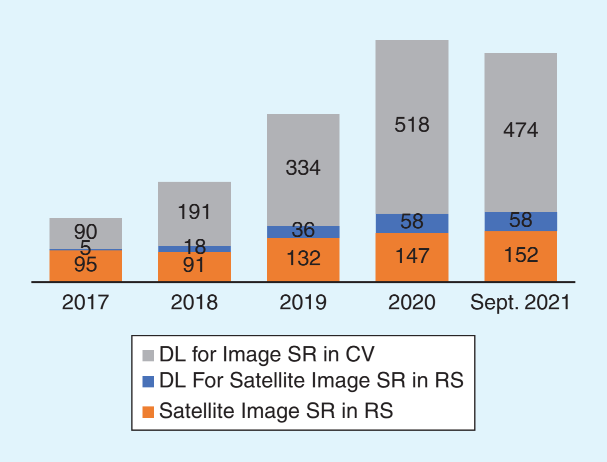

FIGURE 1. The number of published papers related to image SR for traditional and DL-based techniques for the satellite RS and CV fields [26].

In this review article, we present the recent advancements (up to July 2021) of spatial downscaling on satellite imaging through DL approaches and analyze 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 indexes, and so on.

TERMINOLOGY

Before moving forward, we need to clarify which terminology is used in this article as far as spatial resolution increase/decrease is concerned. In climate and meteorological (e.g., [27]) as well as RS [28] studies, the term *downscale* refers to the transition from low to high resolution, i.e., less to more detail representation. However, in the CV field, it is the term *upscale* that refers to the increase of (spatial) resolution, and *downscale* refers to the decrease of it (e.g., [29]); these terms are synonymous with *upsample* and *downsample*, respectively. Zhan et al. [30] conducted research on LST downscaling terminology, among others, and found that terms such as *enhancement*, *sharpening*, *fusion*, *SR*, *unmixing*, *subpixel*, and *disaggregation* are also relevant to spatial resolution increase. In this article, we use the term *downscale*.

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, semisupervised, 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 nonlinear dependencies in the input.

One critical factor for the success of any DL method is the existence of a large and diverse data set to train on. The abundance and availability of data in EO, therefore, provide 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 that exploit deeper architectures have recently been published and achieve impressive results in problems such as land use and land cover classification

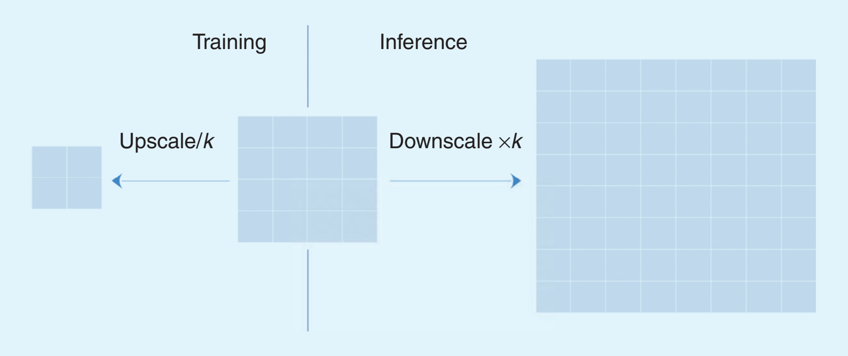

FIGURE 2. The Wald's protocol pipeline. The original image (middle) is upscaled by a $/k$ factor, and the resulting pair is used for model training. The trained model is then transferred to downscale the original image by a $\times k$ factor.

[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 EM spectrum, differentiating themselves from typical CV data, which lie mostly in the red, green, blue (RGB) range. In addition, time is quite an 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, due to either hardware failure or atmospheric conditions that are difficult for certain sensor types to penetrate (e.g., cloud coverage, haze, and so on are common obstacles for optical sensors). Therefore, any researcher willing to design and implement novel DL algorithms for EO must take all of these points into consideration.

PROBLEM DEFINITION

Given a set of n low-resolution (LR) images $(x_1, x_2, \dots, x_n)$, where $x_i \in X^{H\times W}$, and their corresponding high-resolution (HR) images $(y_1, y_2, \dots, y_n)$, 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 nonlinear 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 that account for parameters like the motion blur, quantization error of the compression process, zooming effects, exposure time, white balancing, and so on. 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 Wald's protocol [36] is employed (Figure 2). 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. Note that, throughout this document, we refer to the LR images as *coarse* (C) and the HR images as fine (F).

METRICS

Several quality metrics have been proposed 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]:

- *Full reference*: A complete HR reference image is required for comparison with the reconstructed image.
- *No reference*: Only the reconstructed image is required.
- *Reduced reference*: Only a set of features extracted from an HR image is available and used for comparison.

Table 1 presents some of the most popular quality metrics found in the literature for the task of spatial enhancement.

PERCEPTION–DISTORTION TRADEOFF

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

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

TABLE 1. THE MOST POPULAR METRICS FOR IMAGE QUALITY ASSESSMENT.

METRIC	RANGE	DESCRIPTION	CATEGORY
Mean square error (MSE)	$[0,\infty)$	Pixel-based mean square error	FR
Root-mean-square error (RMSE)	$[0,\infty)$	Pixel-based root-mean-square 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)	$[0,\infty)$	Error relative to the mean image intensity	FR
Peak signal-to-noise ratio (PSNR)	$(-\infty,\infty)$	Peak SNR based on the MSE and expressed in decibels	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 the UQI and measures local differences in luminance, contrast, and structure	FR
Multiscale structural similarity index (MS-SSIM) [40]	[–1, 1]	A combination of the 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]	$[0,\infty)$	An extension of the IFC obtained by normalizing over the reference image content	FR
Noise quality measure (NQM) [43]	$(-\infty,\infty)$	The SNR based on contrast pyramid variations	FR
Feature similarity index (FSIM) [44]	[0, 1]	Similar to the SSIM and utilizes phase congruency and gradient magnitude	FR
Gradient similarity measure (GSM) [45]	[0, 1]	Similar to the SSIM and measures gradient similarity	FR
Spectral angle mapper (SAM) [46]	$[0,\pi]$	Compares the angle between the two spectra	FR
Erreur relative globale 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]	$[0,\infty)$	The 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 nondistorted 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]	$[0,\infty)$	The linear combination of Q_{MA} and NIQE	NR
Learned perceptual image patch similarity (LPIPS) [55]	$[0,\infty)$	L2 (Euclidean) 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 indexes based on the band correlation, each raised to a real-valued exponent	NR

FR = full reference; NR = no reference.

where Δ is a similarity metric, I_{HR} is the HR image, and $\hat{I}_{\mathrm{HR}}$ 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 and can be formulated as

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

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

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

It was empirically observed and then mathematically proven [57] that distortion and perceptual quality metrics act in a complementary yet competitive manner. The perception–distortion tradeoff theorem dictates that, as the distortion error of an algorithm decreases, the visual quality must also decrease, and vice versa. In practice, pursuing a low distortion rate results in more blurry and oversmoothed images because the produced output approximates the statistical average of possible HR solutions to this one-to-many problem, whereas a sharper, more natural-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]. Researchers 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 follows:

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

where λ is the weight of the perception quality factor, and $d(p_{I_{\mathrm{HR}}}, p_{\hat{I}_{\mathrm{HR}}})$ 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.

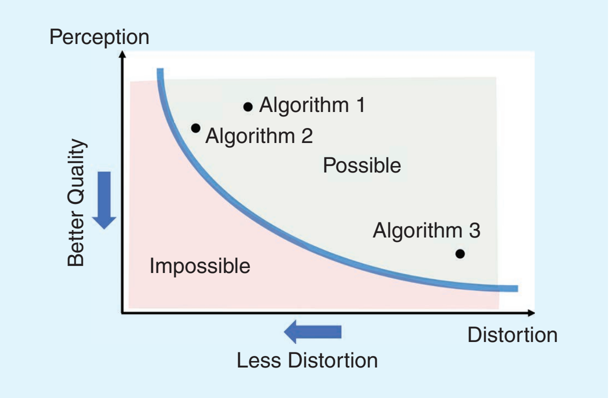

FIGURE 3. The perception–distortion plane and the monotonic boundary separating the unattainable region. (Source: [57]; used with permission.)

STANDARD DEEP LEARNING METHODS FOR DOWNSCALING IN COMPUTER VISION

Resolution enhancement has been thoroughly investigated in the field of general CV over the past decades. 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 article as core modules.

BUILDING BLOCKS

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

UPSAMPLING LAYERS

- *Resize convolution*: This was one of the first techniques proposed for feature downscaling. This operation involves upsampling the input by a traditional interpolation method, such as nearest neighbor, bilinear, or bicubic interpolation, and then performing a convolution on the result [Figure 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 the *deconvolutional layer* [59], which is a quite inaccurate term since deconvolution in CV 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 [Figure 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].
- *Subpixel convolution*: Also called *pixel shuffle* [61], this layer comprises a convolution operation followed by a specific image reshape that rearranges the input features of shape $H \times W \times Cr^2$ to $rH \times rW \times C$ [Figure 4(c)]. This layer achieves a larger receptive field than transposed convolution and causes fewer 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. To simplify this task and avoid learning such a complex translation, several studies employ global residual learning architectures [63] that focus on learning solely the residual, or difference, between the input and output. Provided that a considerable part of the image remains basically unchanged, such a model is tasked with retrieving 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 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 the 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 to 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 that 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].

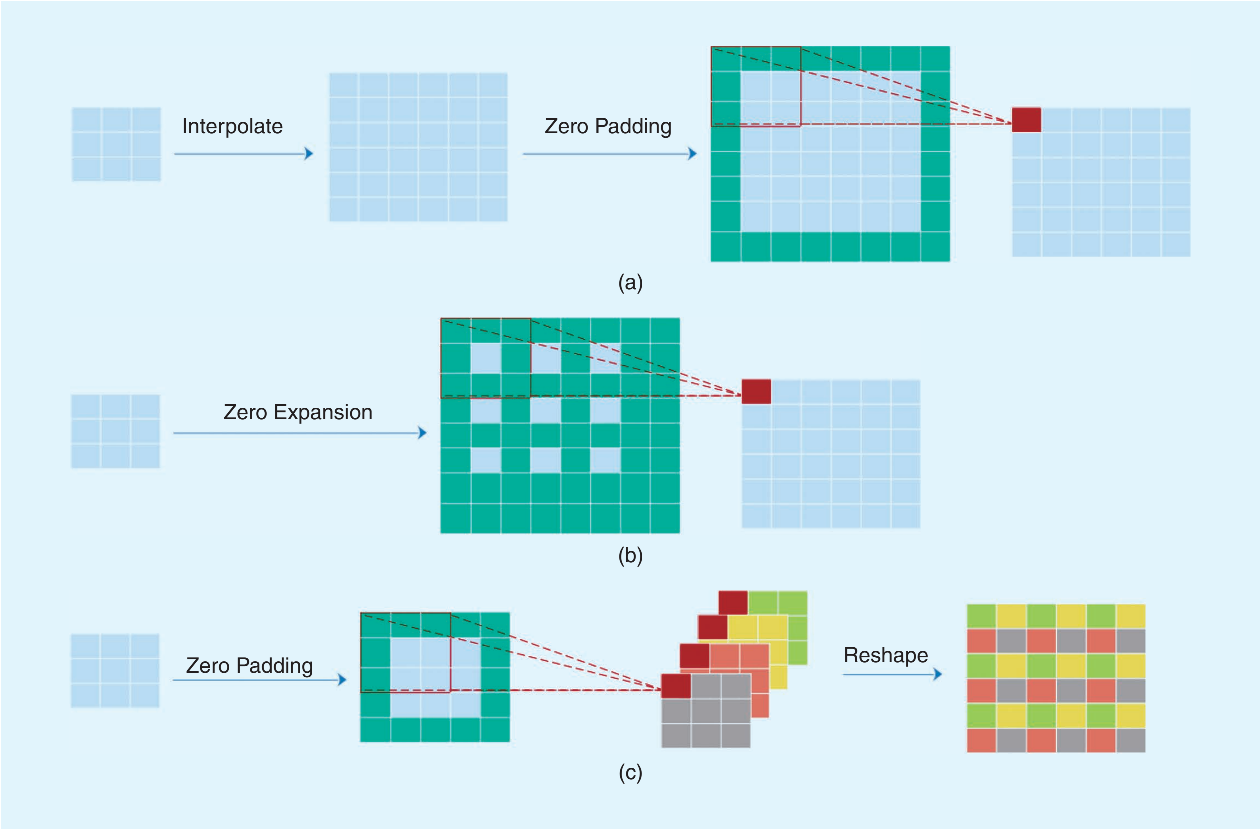

FIGURE 4. An example of the three basic convolution schemes for upsampling a single-channel 3 × 3 feature map by a ×2 factor: the (a) resize, (b) transposed, and (c) subpixel convolutions. The red dashed lines refer to a simple 3 × 3 convolution.

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 Figure 5 and represent the possible ways to design a downscaling DL model with convolutional and upsampling/downsampling layers as basic components.

PREUPSAMPLING 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 to upsample the image to the required scale. Then, a convolutional neural network (CNN) model is applied that 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 preupsampling 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.

POSTUPSAMPLING FRAMEWORK

Mitigating the complexity and high cost of the preupsampling approach, in the postupsampling 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., ×8, ×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 DOWNSAMPLING FRAMEWORK

This framework exploits consecutive up- and downsampling layers, which refine the reconstruction error on HR-to-LR projections, thus extracting more information on the

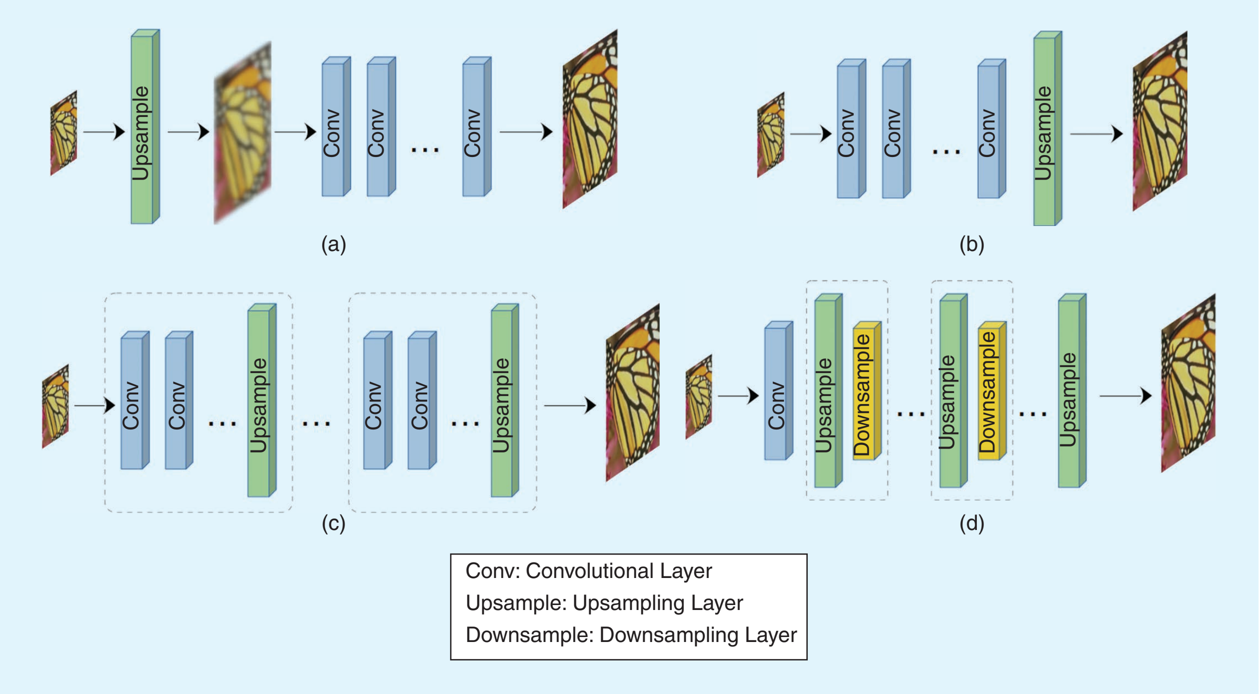

FIGURE 5. The possible downscaling frameworks present in the DL literature: (a) preupsampling, (b) Postupsampling, (c) progressive upsampling, and (d) iterative up- and downsampling. The convolutional, upsampling, and downsampling layers are all trainable. Layers enclosed by dashed boxes denote stackable blocks.

relationship and correlations between the two spaces. Such models usually achieve higher-quality results and are able to handle higher scaling factors successfully.

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 subpixel 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 article [76] (*LapSRN*) introduced a multiscale architecture that integrates the Laplacian pyramid structure and produces intermediate images downscaled by smaller factors (×2, × 4, and ×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, named *SRResNet*, consists of a series of residual blocks, local and global residual connections, and subpixel convolutional layers for downscaling. The discriminator is a VGG-like network that 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 dense blocks with global residual connections, as seen in Figure 6, and use the relativistic average discriminator introduced in [79].

Following the success of the baseline SRGAN, Lim et al. [81] (*EDSR*/*MDSR*) extend the SRResNet architecture by removing the rectified linear unit (ReLU) activations outside the residual blocks and deepening the model. The authors name this architecture *EDSR* and train it separately for the scaling factors ×2, ×3, and ×4. They also noted that, by fine-tuning a pretrained ×2 model when training for ×3 or ×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 involves interrelated 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 was proposed in [83] (*RDN*). The authors present a residual dense block (RDB) that comprises a dense block with three novelties:

- contiguous memory, where the output of an RDB is fed to each layer of the next RDB
- local feature fusion, which is a concatenation and a 1 × × 1 convolution layer at the end of an RDB that adaptively controls the output information, making the network easier to train
- local residual learning, which is a residual connection between the input and output of the RDB.

Utilizing a sequence of such RDB blocks and subpixel 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. Specifically, several consecutive layers alternatively perform up- and downprojection 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.

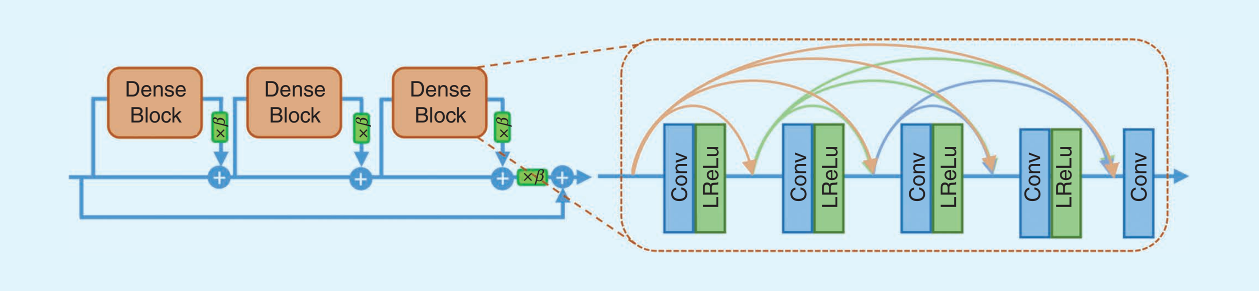

FIGURE 6. The residual-in-residual block (RIRB). It contains multiple dense blocks and residual connections both between blocks and between the input and output of the RIRB. Here, β refers to the residual scaling parameter. (Source [80]; used with permission.) Conv: convolutional layer; LReLU: leaky rectified linear unit.

Some methods (*DRCN* [87] and *DRRN* [84]) propose the use of recursive structures inside the model. Arguing that the addition of more layers makes a network inefficient and 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. Figure 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 subnetwork of each pyramid level is replaced by a series of recursive convolutional layers 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 that 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 subpixel convolutions. When combined with a self-ensembling strategy, RCAN outperforms several robust DL methods.

Table 2 summarizes 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 last attribute is useful to assess the complexity of each model and therefore weigh its proneness to overfit given the training data available.

DOWNSCALING TAXONOMY IN REMOTE SENSING

Based on the dimensions and modalities to be combined, a variety of downscaling schemes have been proposed in the context of EO. Figure 8 provides a simple yet complete

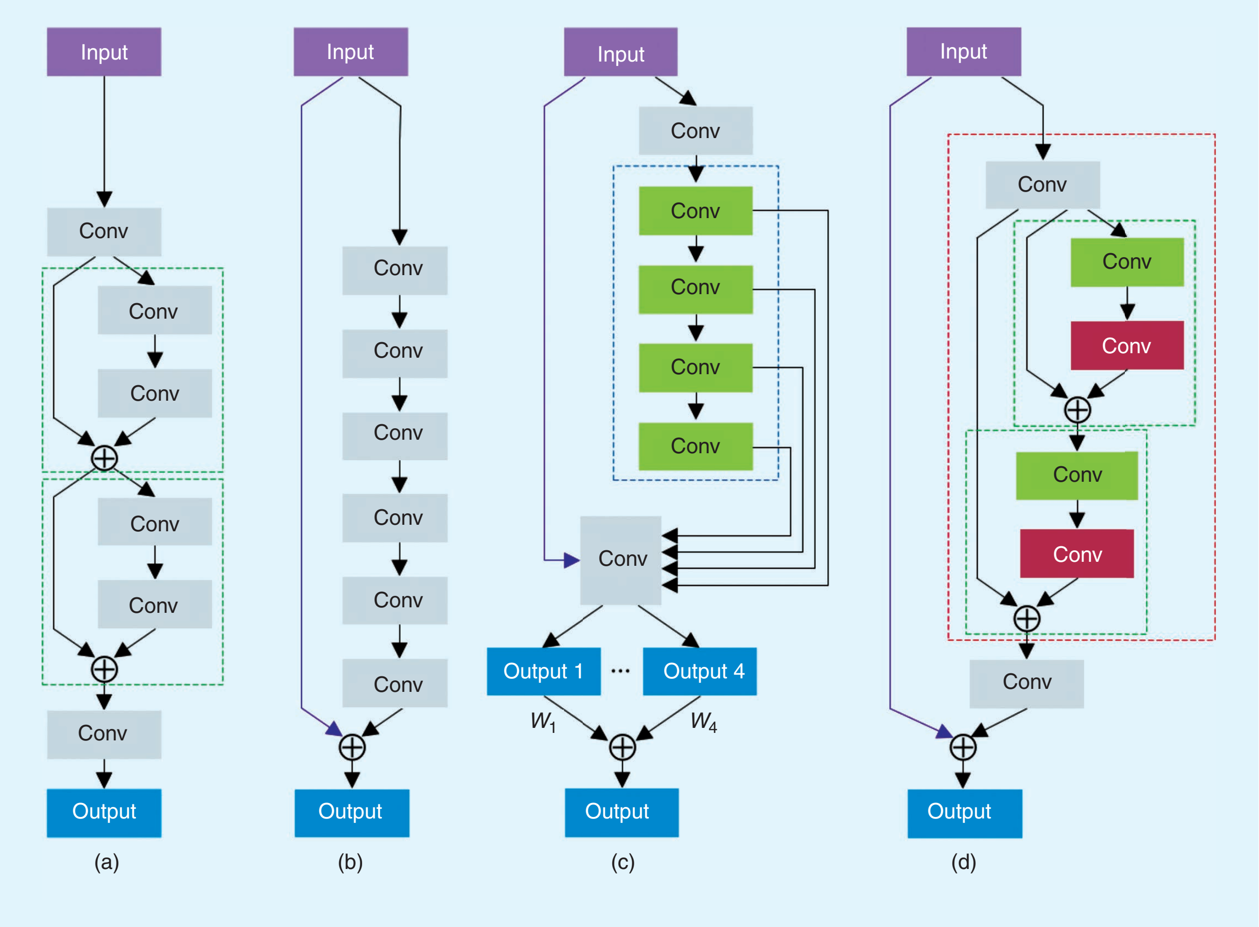

FIGURE 7. An overview of the classic ResNet, VDSR, DRCN, and DRRN architectures. Global residual connections are marked by a purple line, ⊕ refers to elementwise addition, and outputs in blue are supervised. (a) ResNet. The green dashed box signifies a residual block. (b) VDSR. (c) DRCN. The blue dashed 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. W_1 and W_4 are learnable weights assigned to each intermediate hidden state output during recursion. (Source: [84]; used with permission.)

taxonomy of the methodological approaches used in the 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, or no external information is used:

- *Spatiospectral fusion (SSF)*: Images of different spatial and spectral resolutions are fused to produce an image of the highest possible spatial resolution in the coarser bands.
- *Spatiotemporal fusion (STF)*: Images of high spatial but low temporal resolution (HSLT) are fused with images of low spatial but high temporal resolution (LSHT) to produce images of the highest resolution in both dimensions.
- *SR*: A single image or multiple images is/are downscaled without any additional external information.

In more detail, when the downscaling process is assisted with information on different spectra, then SSF techniques are used. These techniques are further discriminated based on the type of input spectra at hand, resulting in multispectral (MS) fusion (two MS images with different spectral information), pansharpening [an MS image and a panchromatic (PAN) image], and MS/hyperspectral (HS) fusion (an MS and an HS image).

TABLE 2. AN OVERVIEW OF THE MOST POPULAR DOWNSCALING MODELS IN CV.

MODEL	BUILDING BLOCKS USED	UPSAMPLING FRAMEWORK	GAN	NUMBER OF PARAMETERS
SRCNN [73][1]	Simple CNN	Preupsampling	No	57,000
VDSR [74]	VGG based and residual connections	Preupsampling	No	665,000
ESPCN [61]	Simple CNN and subpixel convolution	Postupsampling	No	20,000
LapSRN [76][2]	Laplacian pyramid structure	Progressive upsampling	No	821,000
SRGAN [77]	Subpixel convolution and residual connections	Postupsampling	Yes	Generator: 734,000 Discriminator: 5.2 m
ESRGAN [78][3]	Subpixel convolution and residual-in-residual blocks	Postupsampling	Yes	Generator: 16.7 m Discriminator: 14.5 m
EDSR [81][4]	Subpixel convolution, residual connections, and pretraining	Postupsampling	No	43 m
MDSR [81][4]	Multiscale EDSR	Postupsampling	No	8 m
WDSR [82][5]	EDSR with wide activation modules	Postupsampling	No	Small model: 1.2 m Big model: 37.9 m
RDN [83][6]	RDBs, local residual connections, and subpixel convolution	Postupsampling	No	22.3 m
DBPN [85][7]	Residual connections and transposed convolution	Iterative up- and downsampling	No	188,000–2.2 m
D-DBPN [85][7]	Residual connections and transposed convolution	Iterative up- and downsampling	No	10.3 m
SRFBN [86][8]	Residual connections, transposed convolution, and recurrent layers	Iterative up- and downsampling	No	3.6 m
DRCN [87]	Recursive convolutions and residual connections	Preupsampling	No	1.8 m
DRRN [84][9]	DRCN with recursive blocks and added local residual connections	Preupsampling	No	297,000
MS-LapSRN [88][2]	LapSRN with shared weights and recursive blocks	Progressive upsampling	No	222,000
RCAN [89][10]	Channel attention, subpixel convolution, residual-in-residual blocks, and residual connection	Postupsampling	No	16 m

Parameters are an estimation for the ×4 scaling factor, and links to the official code repositories are provided where possible.

[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.

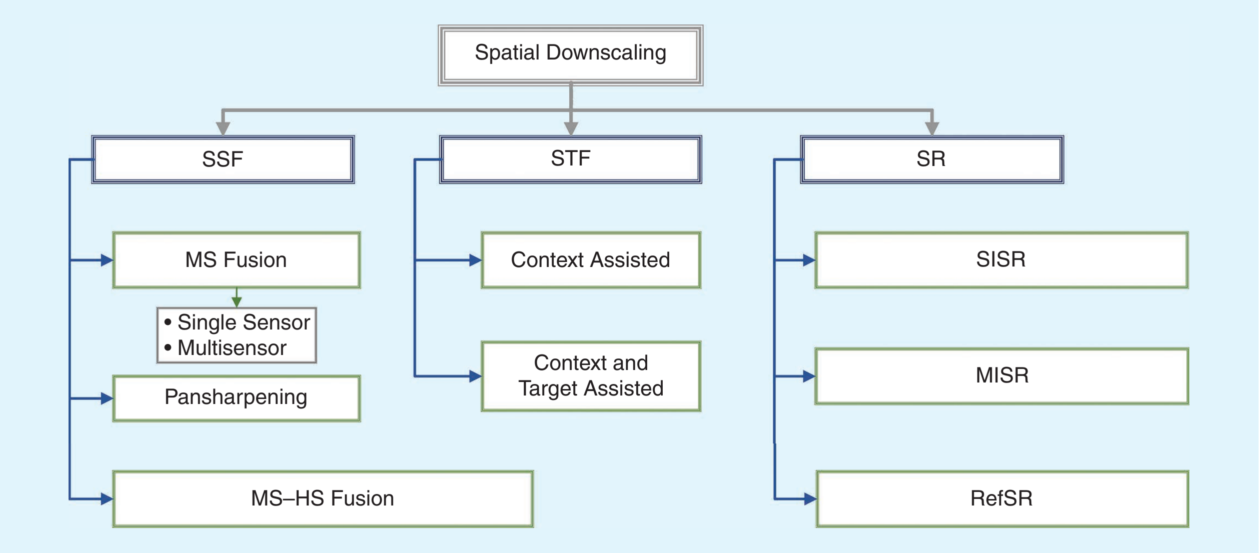

FIGURE 8. The proposed taxonomy of DL downscaling methods in the literature. MISR: multiple-image SR; RefSR: reference SR.

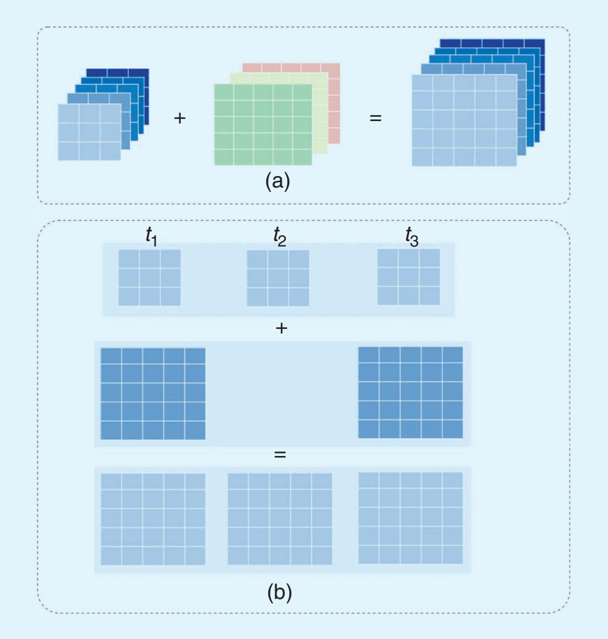

FIGURE 9. (a) SSF: an image of coarse spatial resolution is fused with an image of fine spatial resolution containing different bands. The result is a version of the former image downscaled to the spatial resolution of the latter. (b) STF: an image of high temporal (t_1, t_2, and t_3) but low spatial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

In contrast, when the same spectra are available at different time steps and different spatial resolutions, then STF methods come into play, where temporal differences are additionally exploited for the spatial downscaling. This family of methods includes two subfamilies 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 SR techniques can be employed. There are three method subfamilies depending on the number of input images and whether additional features extracted from the same LR data are used as auxiliary input.

Figures 9 and 10 present an overview of the aforementioned method families, graphically highlighting the different approaches, whereas Figures 11–13 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.

SPATIOSPECTRAL FUSION

Satellites are equipped with various different sensors that operate in different parts of the EM spectrum and capture information on different features of the scanned location. These features can have variable spatial resolution; thus, an advanced method called *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: MS image fusion, pansharpening, and HS image downscaling. These are presented next, while, in Table 3, we summarize the main DL models developed for SSF.

MULTISPECTRAL IMAGE FUSION

Using information from a single satellite source has the advantage of consistent satellite orbit characteristics (e.g., the altitude, inclination, and so on) and atmospheric conditions. Some satellites carry multiple sensors that allow simultaneous capture of multiresolution images, thus providing an ideal setting for SSF and a common data source. For example, the constellation of *Sentinel-2* satellites (A/B) launched by the European Space Agency acquires an image with 13 discrete bands, four of which have a 10-m spatial resolution; six have 20 m, and three have 60 m [93]. Several methods (*DSen2* and *VDSen2* [94], *FUSE* [95], [96], and *SPRNet* [97]) use two input sets, one for the B_{HR} and one for the B_{LR} resampled to match the target resolution, as input to the CNN models, which aim to transfer high-frequency details from B_{HR} to B_{LR} 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 the input, a downsampled concatenation of HR and LR *Sentinel-2* bands to recover the original LR bands (Figure 14). On the other hand, Nguyen et al. [100] (*S2SUCNN*) propose a multiscale model that takes as the 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 10-m spatial resolution. The final result is subsequently degraded to be compared with the original input in an 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 preupsampling 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.

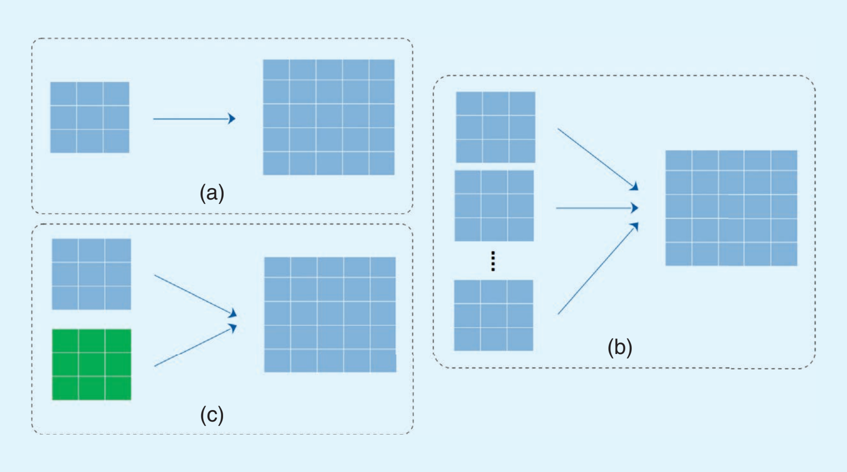

FIGURE 10. (a) SISR: a single LR image is downscaled without using any external information. (b) MISR: multiple LR images of the same scene are used to acquire an image of higher spatial resolution of that scene. (c) RefSR: an LR image is downscaled by combining information from features extracted from it.

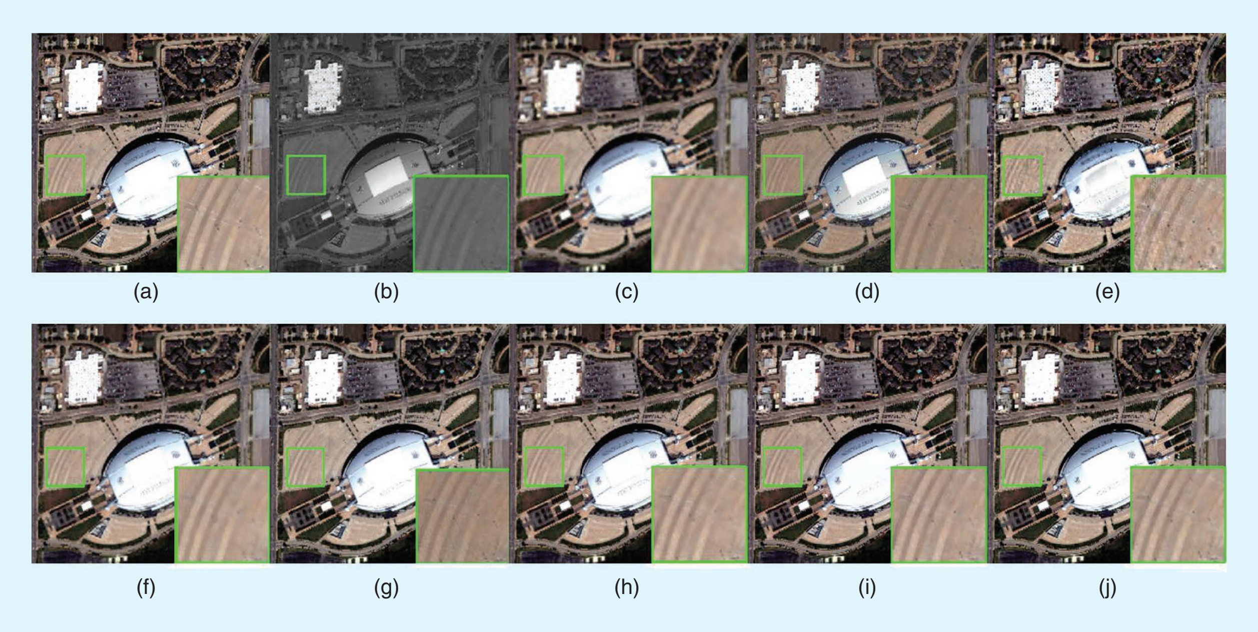

FIGURE 11. An example of pansharpening on *WorldView-3 data*: (a) an HR “ground-truth” image, (b) panchromatic, (c) LR multispectral image, and (d)–(j) the pansharpening results obtained by different DL approaches. (Source: [90]; used with permission.)

Shao et al. [102] (*ESRCNN*) propose a framework that extends the SRCNN architecture (Table 2) and utilizes auxiliary information from *Sentinel-2* to downscale *Landsat-8* images. The *Landsat-8* satellite provides observations in the visible, near-infrared (NIR), and shortwave infrared (SWIR) spectra at 30 m and a PAN band at 15-m spatial resolution every 16 days [103], so the goal of this study is to produce the equivalent Landsat images at 10-m spatial resolution.

The whole process can be broken down into two separate steps. First is the self-adaptive fusion of *Sentinel-2*, where the 20-m *Sentinel-2* bands (11 and 12) are resampled to 10 m using *k*-nearest neighbors (k-NN) interpolation and are then concatenated with the native 10-m bands as the input to the proposed ESRCNN model. The output is bands 11 and 12 downscaled to 10-m resolution. Following this is the multitemporal fusion of *Landsat-8* and *Sentinel-2*, where the 30-m Landsat bands (1–7) and the PAN band are resampled to 10 m, again using k-NN interpolation, and are concatenated with the native 10-m *Sentinel-2* bands and the downscaled 20-m *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 versus 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 to enhance the latter to a spatial resolution of 10 m. They proved that an adversarial approach is superior to a nonadversarial one, and the proposed model resembles the architecture of the ESRGAN trained on a composite of the RGB 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 data set (see the "Data Sets" section), but the results were not favorable.

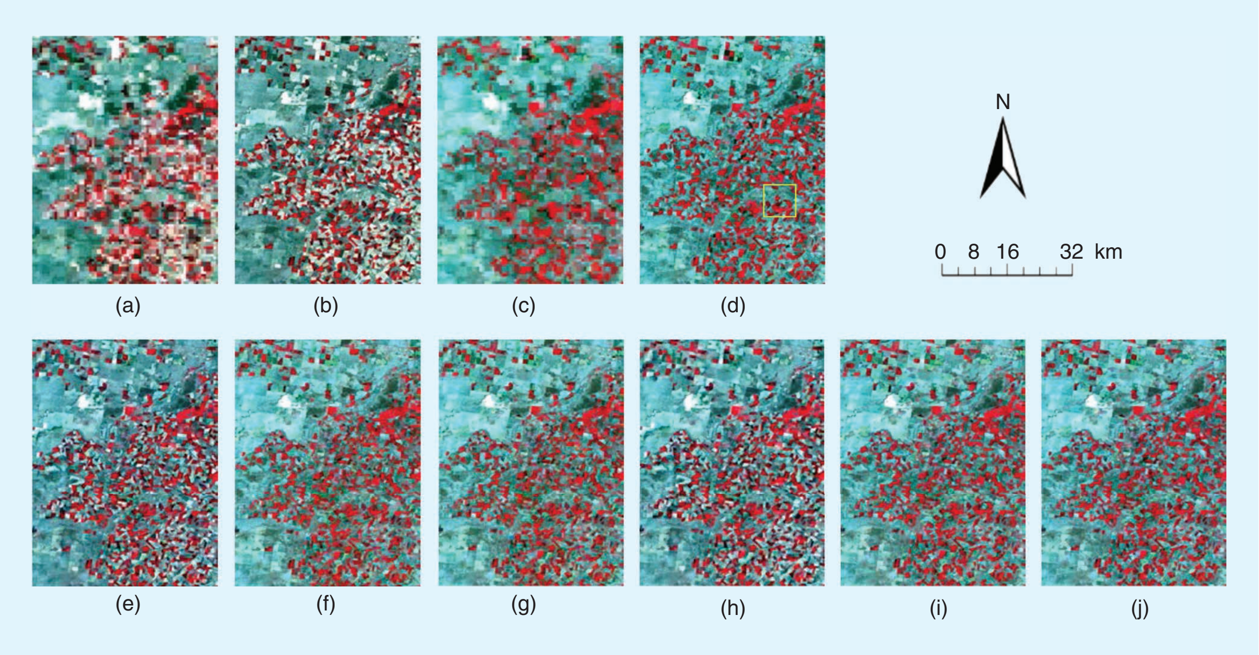

FIGURE 12. An example of STF. (a) An LR image at time t_1. (b) An HR image at time t_1. (c) An LR image at time t_2. (d) An HR image at time t_2, which is the target. (e)–(j) The prediction results at time t_2 obtained by different approaches. (Source: [91]; used with permission.)

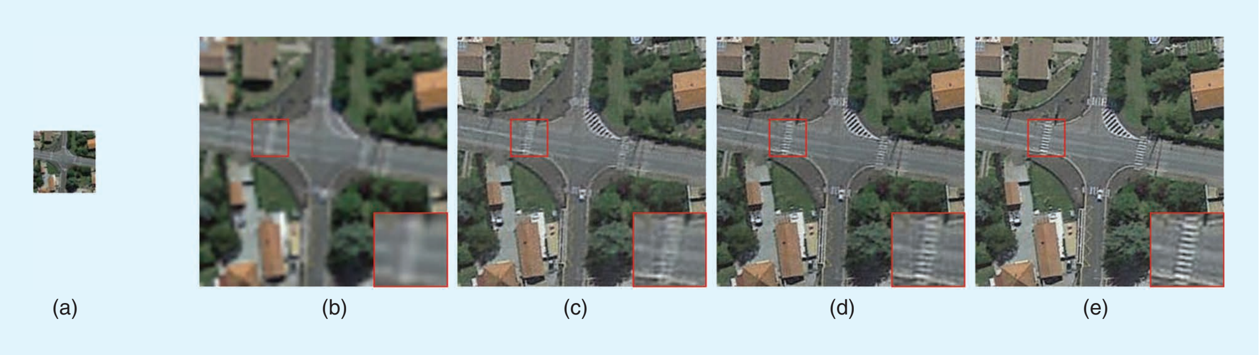

FIGURE 13. An example of SISR: (a) an LR image and (b)–(e) the prediction results obtained by different approaches for a scaling factor of $\times 4$. (Source: [92]; used with permission.)

TABLE 3. SUMMARY OF THE STATE-OF-THE-ART DL MODELS FOR SSF FOR IMAGE DOWNSCALING IN RS.

MODEL	FUSION TYPE	FUSION DATA	CV MODEL	BUILDING BLOCKS	UPSAMPLING FRAMEWORK	ARCHITECTURE	CODE AVAILABLE/ NUMBER OF PARAMETERS
DSen2 [94]	MS	*Sentinel-2*	—	Residual learning	Preupsampling	CNN	Yes/1.8 m
VDsen2 [94]	MS	*Sentinel-2*	—	Residual learning	Preupsampling	CNN	Yes/37.8 m
Palsson et al. [95]	MS	*Sentinel-2*	—	Residual learning	Preupsampling	CNN	No/—
FUSE [96]	MS	*Sentinel-2*	—	Residual learning	Preupsampling	CNN	No/28,000
FusGAN [98]	MS	*Sentinel-2*	ESRGAN	Residual learning and subpixel convolution	Postupsampling	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	Preupsampling	CNN	No/—
ESRCNN [102]	MS	Multitemporal *Landsat-8* and *Sentinel-2*	SRCNN	—	Preupsampling	CNN	Yes/—
Chen et al. [2]	MS	*Landsat-8* and *Sentinel-2*	ESRGAN	Residual learning and subpixel convolution	Postupsampling	GAN	No/—
RRSGAN [104]	MS	*WorldView-2* and *GaoFen-2*	—	Residual learning, subpixel convolution, and attention mechanism	Progressive upsampling	GAN	Yes/7.47 m
PNN [106]	PAN + MS	*IKONOS*, *GeoEye-1*, and *WorldView-2*	SRCNN	—	Preupsampling	CNN	Yes/310,000
PanNet [109]	PAN + MS	*IKONOS*, *WorldView-2*, and *WorldView-3*	—	Residual learning and high-pass filtering	Progressive upsampling	CNN	No/250,000
DRPNN [108]	PAN + MS	*IKONOS*, *WorldView-2*, and *QuickBird*	SRCNN	Residual learning	Preupsampling	CNN	No/1.6 m
DML-GMME [111]	PAN + MS	*IKONOS*, *WorldView-2*, *QuickBird*, and *GaoFen-2*	—	Stacked sparse autoencoders [145]	Preupsampling	CNN	No/8,000
MSDCNN [112]	PAN + MS	*IKONOS*, *WorldView-2*, and *QuickBird*	—	Residual learning	Preupsampling	2 CNNs	No/—
L1-RL-FT [110]	PAN + MS	*WorldView-2* and *WorldView-3*	SRCNN	Residual learning	Preupsampling	CNN	Yes/—
DiCNN [113]	PAN + MS	*WorldView-2* Washington, *IKONOS* Hobart, and *QuickBird* Sundarbans	SRCNN	—	Preupsampling	2 CNNs	No/180,000
DIRCNN [119]	PAN + MS	*IKONOS*, *QuickBird*, *Gaofen-1*, and *Gaofen-2*	—	Residual learning, attention mechanism, and auxiliary gradient data	Preupsampling	CNN	No/1.6 m
MIPSM [115]	PAN + MS	*IKONOS* and *QuickBird*	—	Residual learning and high-pass filtering	Preupsampling	2 CNNs	No/—
Fusion-Net [116]	PAN + MS	*WorldView-2*, *WorldView-3*, *QuickBird*, and *Gaofen-2*	—	Residual learning	Preupsampling	CNN	Yes/230,000
SRPPNN [117]	PAN + MS	*QuickBird*, *WorldView-3*, and *Landsat-8*	—	Residual learning and high-pass filtering	Preupsampling	CNN	No/—
UP-SAM [120]	PAN + MS	*GeoEye-1*, *IKONOS*, *WorldView-2*, and *WorldView-3*	—	Residual learning, attention mechanism, and subpixel accuracy	Preupsampling	CNN	No/—
Luo et al. [114]	PAN + MS	*Gaofen-2* and *WorldView-2*	—	Residual learning and attention mechanism	Preupsampling	CNN	No/—
GTP-PNet [123]	PAN + MS	*WorldView-2*, *Gaofen-2*, and *QuickBird*	—	Residual learning and gradient information	Preupsampling	2 CNNs	No/—
PSCSC-Net [124]	PAN + MS	*GeoEye-1*, *IKONOS*, and *WorldView-2*	—	Deep unfolding and variational optimization	Preupsampling	CNN	No/1.1 m

(*Continued*)

TABLE 3. SUMMARY OF THE STATE-OF-THE-ART DL MODELS FOR SSF FOR IMAGE DOWNSCALING IN RS. (*Continued*)

MODEL	FUSION TYPE	FUSION DATA	CV MODEL	BUILDING BLOCKS	UPSAMPLING FRAMEWORK	ARCHITECTURE	CODE AVAILABLE/ NUMBER OF PARAMETERS
VO+Net [125]	PAN + MS	*WorldView-3*, *WorldView-2*, and *QuickBird*	—	Variational optimization	Preupsampling	CNN	No/—
SC-PNN [126]	PAN + MS	*WorldView-3*, *GeoEye-1*, and *SPOT5*	—	Saliency analysis and hybrid and deformable convolution	Preupsampling	CNN + fully convolutional network	No/—
NLRNet [90]	PAN + MS	*WorldView-3* and *QuickBird*	—	Residual learning and attention mechanism	Preupsampling	CNN	No/—
LPPNet [118]	PAN + MS	Pavia Center, Houston, and Los Angeles	—	Laplacian pyramid decomposition	Preupsampling	CNN	No/—
Scarpa et al. [110]	PAN + MS	*GeoEye-1* and *WorldView-2*	—	Residual learning	Preupsampling	CNN	No/—
Ciotola et al. [130]	PAN + MS	*GeoEye-1*, *WorldView-2*, and *WorldView-3*	—	—	—	CNN	No/—
PSGAN [131]	PAN + MS	*QuickBird*, *GaoFen-2*, and *WorldView-2*	—	—	Preupsampling	GAN	Yes/1.88 m
Pan-GAN [132]	PAN + MS	*GaoFen-2* and *WorldView-2*	—	Two discriminators: spatial and spectral	Preupsampling	GAN	No/—
MDSSC-GAN SAM [133]	PAN + MS	Pléiades and *WorldView-3*	—	Two discriminators: spatial and spectral; residual learning; and attention mechanism	Preupsampling	GAN	Yes/—
PanColorGAN [134]	PAN + MS	Pléiades, *WorldView-2*, and *WorldView-3*	—	Self-supervised and noise/color injection	Preupsampling	GAN	No/—
Palsson et al. [135]	MS + HS	Pavia Center and *IKONOS*	—	—	Preupsampling	3D CNN	No/—
DHSIS [136]	MS + HS	CAVE and Harvard	—	Self-supervised and noise injection	Preupsampling	GAN	Yes/—
PFCN [137]	MS + HS	Botswana; Washington, D.C.; and Pavia Center	—	Residual learning	Preupsampling	CNN	No/—
CF-BPNN [138]	MS + HS	AVIRIS and Pavia Center	—	*k*-Means clustering	Preupsampling	NN	No/—
HyperPNN [139]	MS + HS	Washington, D.C. National Mall; Moffett Field; and Salinas Scene	—	—	Preupsampling	CNN	No/—
DDLPS [140]	MS + HS	Moffett Field, Chikusei, and Salinas Scene	LapSRN	—	Preupsampling	CNN	No/—
TONWMD [141]	MS + HS	CAVE, Harvard, and Pavia Center	—	Residual learning and matrix decomposition	Preupsampling	CNN	No/—
MHF-Net [142]	MS + HS	CAVE, Chikusei, Houston, and Pavia Center	—	—	Preupsampling	CNN	Yes/—
UMAG-Net [143]	MS + HS	CAVE and Harvard	—	Attention mechanism	Preupsampling	CNN and AE	No/—
SSR-Net [144]	MS + HS	Pavia Center; Botswana; and Washington, D.C. National Mall	—	—	Preupsampling	CNN	Yes/—

CV Model refers to the models presented in Table 2. AE: autoencoder; AVIRIS: airborne visible/infrared imaging spectrometer; CAVE: Columbia computer vision laboratory; NN: neural network; NLRNet: nonlocal attention residual network.

In their study, Dong et al. [104] (*RRSGAN* and *RRSNet*) argue that RS images coming from different sources must be carefully aligned before processing due to differences in the altitude, viewpoint, or angle. They form a data set consisting of *WorldView-2* (0.5-m) and *GaoFen-2* (0.8-m) observations as well as the corresponding images from Google Earth (0.6 m). 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 to extract features that are then aligned via a pyramid with deformable convolutional layers [105]. Subsequently, a relevance attention module is proposed to

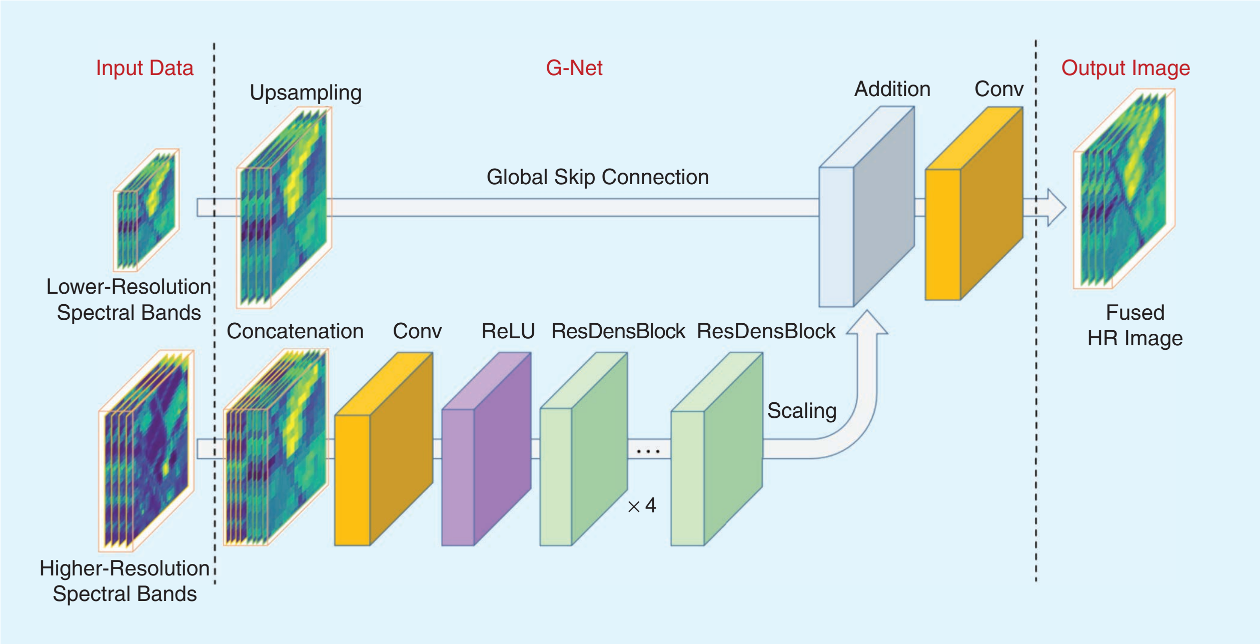

FIGURE 14. The FusGAN generator network. *ResDensBlock* is the RDB as described in ESRGAN. (Source: [98]; used with permission.) G-Net: generator of the model.

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: 1) the MAE between the downscaled and HR images, 2) the adversarial loss for the downscaled and HR images, 3) the VGG loss between the downscaled and HR images, 4) the MAE between the gradients of the downscaled and HR images, and 5) the adversarial loss for the gradients of the downscaled and HR images. The results show that both the adversarial RRSGAN and the nonadversarial RRSNet perform better than numerous other DL methods, with RRSGAN producing more high-frequency details.

In conclusion, considering single-source data for MS image fusion, the available solutions cover a variety of needs. For example, when all LR input images have the same spatial resolution (e.g., 20 m), 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 (~28,000) compared to other methods but has only be applied with a ×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 multisource 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, especially when more than three *Sentinel-2* images are used, 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.

PANSHARPENING

Pansharpening refers to a downscaling process aided by a PAN 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, PAN sensors are able to detect brightness changes at quite small spatial scales.

The first work to introduce CNNs to pansharpening is [106] (*PNN*). Inspired by the SR field of CV, Masi et al. [106] build upon the SRCNN and improve it by augmenting the input with a number of radiometric indexes tailored to features relevant for RS applications [the normalized difference vegetation index (NDVI), the normalized difference water index (NDWI), and so on]. Following the three steps of sparse coding SR [107], they make use of a three-layer CNN named *PNN*, as shown in Figure 15. Their method follows the preupsampling framework.

Motivated by the high nonlinearity 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 MS 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 (Figure 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 the 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.

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

A two-branch network named *MSDCNN* is proposed in [112]. While the one branch is a three-layer CNN, the other one is a deep residual network with multiscale convolutional blocks. *Multiscale* refers to the fact that the 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 CNNs, *DiCNN1* and *DiCNN2*, both utilizing the preupsampling framework. *DiCNN1* adds a skip connection to the PNN architecture, while *DiCNN2* works under the assumption that, ideally, the MS spatial details should match and be relevant only to the PAN image. Thus, it utilizes only the PAN image as an input to the network, while the preinterpolated MS image is used only at its end. Structural comparisons among PNN, DRPNN, DiCNN1, and DiCNN2 can be seen in Figure 17.

Liu et al. [115] propose a method named *MIPSM* that combines a shallow–deep convolutional network (*SDCN*)

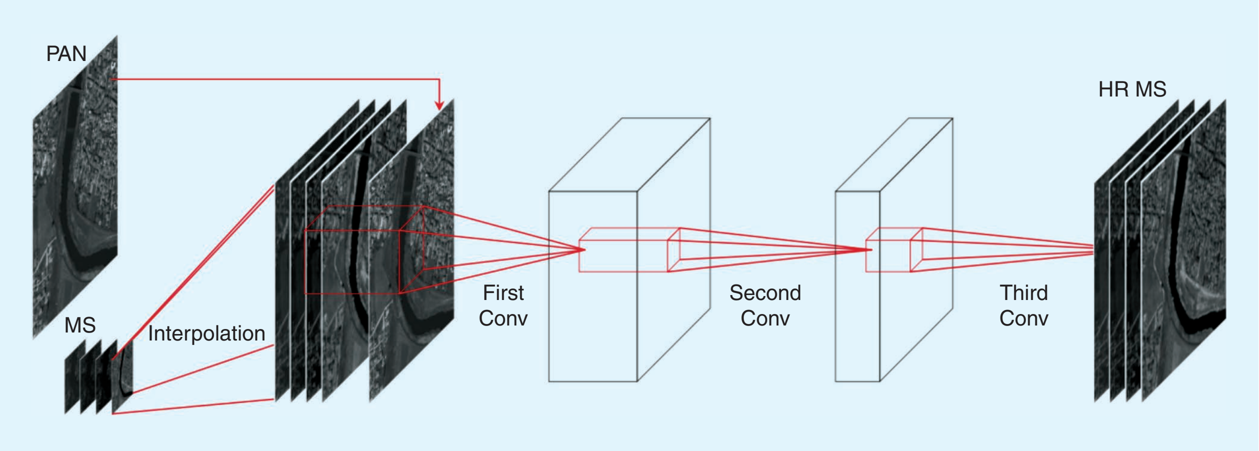

FIGURE 15. An outline of PNN. The network comprises three layers that are expected to match the three steps of sparse coding SR. (Source: [106]; used with permission.)

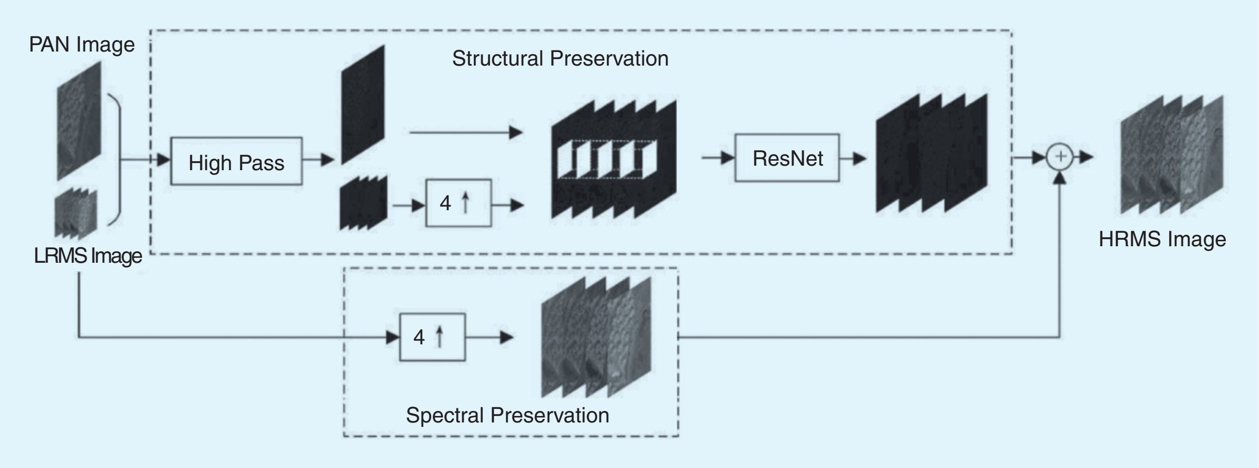

FIGURE 16. An outline of PanNet. The network decouples the structural from the spectral preservation. (Source: [109]; used with permission.)

and a spectral discrimination-based detail injection (*SDDI*) model. The SDCN consists of a three-layer shallow network and a deep residual network that can capture midlevel and high-level spatial features from PAN images. The SDCN works on the high-pass filtering domain. The SDDI is developed to merge the spatial details extracted by the SDCN into MS images with minimal spectral distortion. The 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* 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 PAN image with each MS band. This network is called *Fusion-Net*. They make use of the preupsampling framework using a polynomial kernel.

Cai et al. [117] propose a progressive downscaling pansharpening neural network named *SRPPNN*. It includes three components: 1) a downscaling process that extracts the inner spatial detail that is present in the MS image and combines it with the spatial detail of the PAN image to generate fused results; 2) progressive pansharpening to separate the spatial resolution improvement process, which achieves a gradual and stable pansharpening process; and 3) a high-pass residual module that helps by directly injecting spatial detail from PAN 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: detail extraction and reconstruction. For the detail extraction, they use the Laplacian pyramid to decompose the PAN image into multiple levels that can distinguish the details of different scales. They build a simple detail extraction subnetwork for each level that can help fully extract the depth of different levels. For reconstruction, the subband residuals estimated at each level are injected into the respective level of the MS image, while they are upsampled and fed as the 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: 1) the differential information mapping strategy, 2) the auxiliary gradient information strategy, and 3) the combination of an attention module with residual blocks. Taking into account the underutilization of the PAN image in the input, they propose copying and assigning the PAN image to each band of the downscaled MS 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 the self-attention mechanism (*SAM*) [121] that works at the subpixel level. A method using skip connections inspired by [122] is introduced in [114], in which Luo et al. propose a novel loss function that utilizes spatial constraints, spectral consistency, and the quality with no reference (QNR) index (see the "Metrics" section). Instead of using simple stacked convolutional layers and separating the feature extraction, their network architecture adopts an iterative way to jointly extract and fuse the features. An outline of their method can be seen in Figure 17(e).

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

Inspired by the learned iterative soft-thresholding algorithm, Yin [124] proposes a deep PNet that integrates the

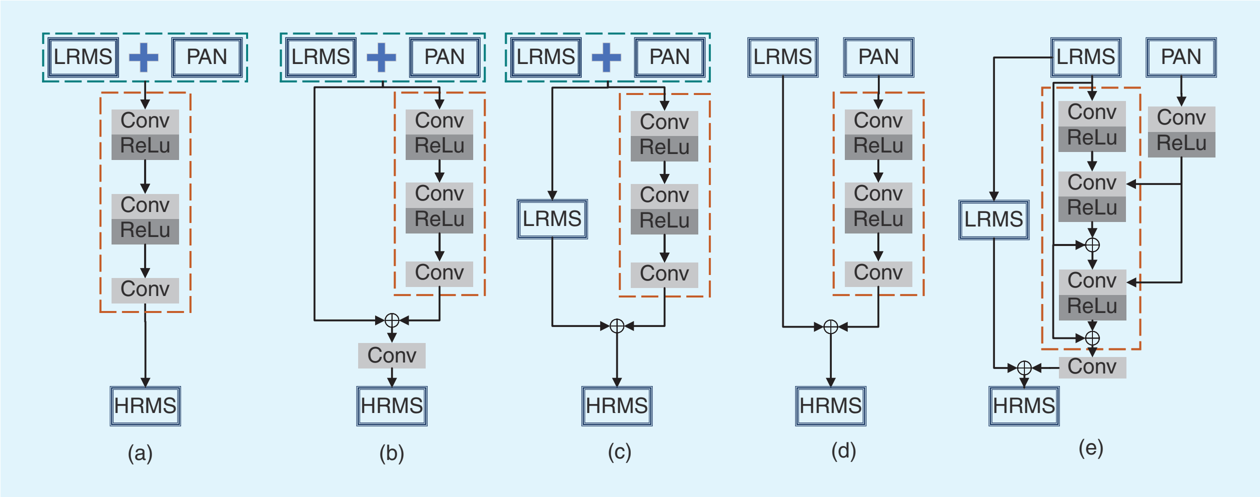

FIGURE 17. A structural comparison between (a) PNN, (b) DRPNN, (c) DiCNN1, and (d) DiCNN2 (source: [113]; used with permission) as well as e) the model of Luo et al. (source: [114]; used with permission).

detail injection, variational optimization, and DL schemes into a single framework. It consists of the input convolutional layer, *Conv-ISTA* module (deep unfolded network), fusion module, and output convolutional layer. The weighted use of variational optimization with DL is proposed in *VO+Net* [125], too. For the variational optimization modeling, a general detail injection term inspired by the classical multiresolution analysis is proposed as a spatial fidelity term, and a spectral fidelity employing the MS sensor's modulation transfer functions is also incorporated. For the DL injection, a weighted regularization term is designed to introduce DL 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 CNN that consists of two parts: 1) a dilated deformable fully convolutional network (*DDCN*) for saliency analysis and 2) 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: 1) detail maps of MS and PAN images are extracted via dual-tree complex wavelet transform (*DT-CWT*) [127], 2) a deep regression network based on RDBs takes those detail maps as the input and produces the primarily sharpened image with high spatial and spectral quality, and 3) 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 Figure 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 nonlocal attention residual network (called *NLRNet*) to capture the similar contextual dependencies of all pixels. Motivated by the unavoidable absence of the ground truth, which often results in networks trained solely in a reduced-resolution domain, Vitale and Scarpa [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* [110], an advanced version of the PNN with 1) a different loss function for training (the MAE instead of the mean square error [MSE]), 2) a residual learning configuration, and 3) 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 HR domain, relying only on the original PAN and MS 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 CNNs, one of the first attempts to utilize GANs for producing high-quality pansharpened images is introduced by Liu et al. in [131] (*PSGAN*). PSGAN comprises a generator, which takes PAN images as the input and maps them to the desired HRMS images, and a

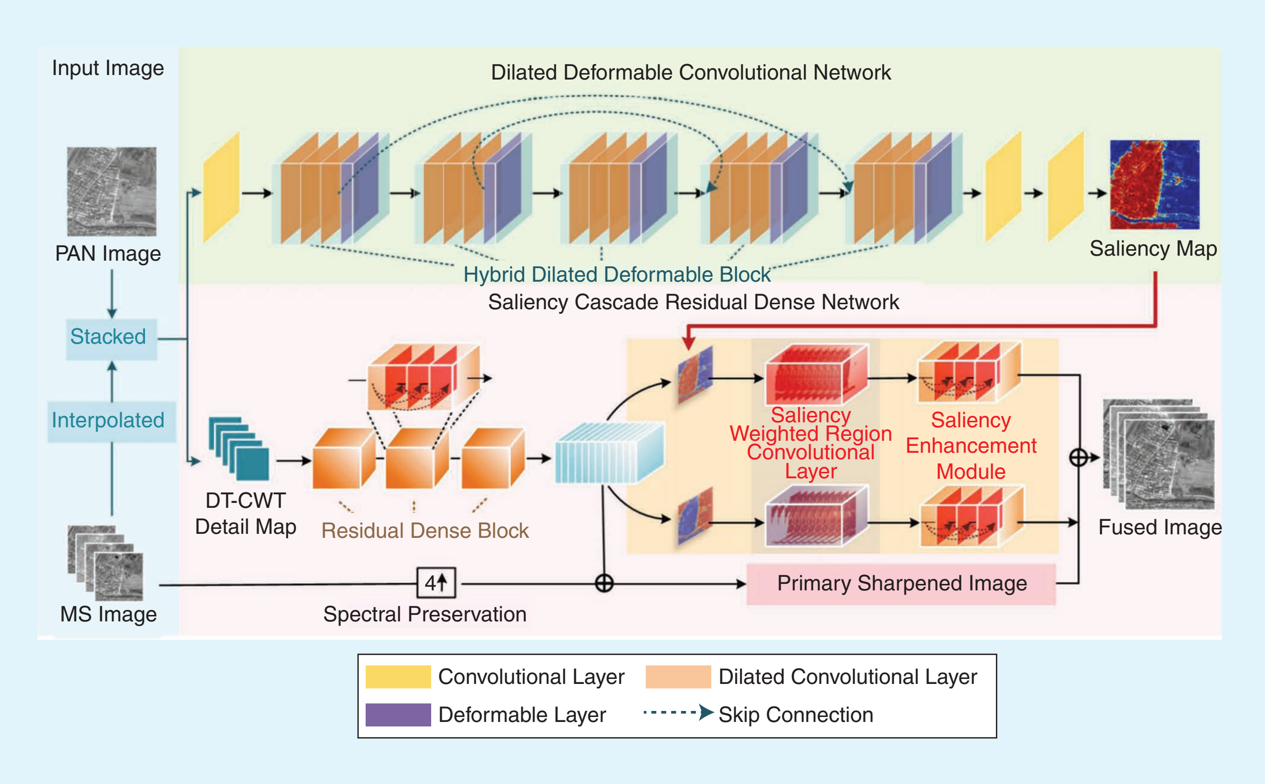
Input Image
Dilated Deformable Convolutional Network
PAN Image
Saliency Map
Hybrid Dilated Deformable Block
Saliency Cascade Residual Dense Network
Stacked
Interpolated
DT-CWT
Detail Map
Residual Dense Block
Saliency Weighted Region Convolutional Layer
Saliency Enhancement Module
Fused Image
MS Image
4↑
Spectral Preservation
Primary Sharpened Image
Convolutional Layer
Dilated Convolutional Layer
Deformable Layer
Skip Connection

FIGURE 18. An outline of SC-PNN. (Source: [126]; used with permission.)

discriminator, which implements the adversarial training strategy for generating higher-fidelity pansharpened images. Making the assumptions that 1) the spectral distribution of the fused image should be consistent with that of the LRMS image and 2) the spatial distribution of the fused image should be consistent with that of the PAN image with the same resolution, Ma et al. propose the use of a GAN with two discriminators in [132] (*Pan-GAN*). The generator of Pan-GAN attempts to generate a HRMS image containing major spectral information of the LRMS image together with additional image gradients of the PAN image.

A similar GAN architecture called *MDSSC-GAN SAM* that jointly exploits the spatial and spectral information sources is proposed in [133], in which Gastineau et al. make use of two discriminators, too: one to preserve the texture and geometry of the images by taking as the input the luminance Y and NIR 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 pansharpening is treated as a colorization problem, is introduced by Ozcelink et al. in [134] (*PanColorGAN*). In contrast with the ordinary method, the authors give, as the input, the gray-scale-transformed MS image and train the model to learn the colorization of it. The model learns to generate an original MS image by taking, as the input, the corresponding reduced-resolution and gray-scale ones. PanColorGAN is trained using both a reconstruction (MAE) and an adversarial loss. This can be interpreted as meaning that the model learns to separate the spectral and spatial components of the MS 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 a 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. Because it has a rich ablation and is a lightweight model, a researcher would gain useful insights 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 by Zhang et al. [123] and Luo et al. [114] Zhang et al. design a special gradient transformation network that searches the nonlinear mapping between the gradients of PAN and MS images. Luo et al. propose a PAN -guided strategy that continuously extracts and fuses features from the PAN 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.

HYPERSPECTRAL/MULTISPECTRAL FUSION

HS image sharpening aims at fusing an observable low-spatial-resolution HS image with a high-spatial-resolution MS image of the same scene to acquire a HRHS image. One of the first works to utilize CNNs for HS/MS fusion is introduced by Palsson et al. in [135], where the authors propose the use of a 3D CNN with three layers for the HS/MS fusion. The dimensionality of the HS image is reduced using principal component analysis to constrain the computational cost and increase robustness.

Dian et al. [136] propose a deep HS image-sharpening method called *DHSIS* that directly learns the priors of the HRHS image via CNN-based residual learning. They first initialize the HRHS image by solving a Sylvester equation. Then, to learn the priors, they utilize the initialized HRHS image as the input of the CNN to map the residuals between the reference HRHS image and initialized HRHS image. This initialization can fully utilize the constraints of the fusion framework, thus improving the quality of the input data. The learned priors of the HRHS image are returned to the fusion framework to reconstruct the final estimated HRHS image, which can further improve the performance (Figure 19).

Zhou et al. [137] introduce a pyramid fully convolutional network (*PFCN*) consisting of two subnetworks: 1) an encoder aiming to encode the LRHS image into a latent image and 2) a pyramid fusion that utilizes this latent image together with an HRMS pyramid image to progressively reconstruct the HRHS image in a global-to-local way. More details about the method can be seen in Figure 20.

Instead of formulating the task of HS/MS fusion as the spatial downscaling of an LRHS image, Han et al. [138] formulate it as the spectral downscaling of an HRMS image. Their method, *CF-BPNN*, consists of three stages: 1) the fusion problem is formulated as a nonlinear spectral mapping from an HRMS image to and HRHS image with the help of an LRHS image, 2) a cluster-based learning method using multibranch neural networks is utilized to ensure a more reasonable spectral mapping for each cluster, and 3) an associative spectral clustering is proposed to ensure that training and fusion clusters are consistent.

He et al. [139] introduce *HyperPNN*, an HS image-sharpening method via spectrally predictive CNNs, exploiting the spectral convolution structure to strengthen the spectral prediction. Li et al. [140] propose a detail-based deep Laplacian pansharpening model (*DDLPS*) to improve the spatial resolution of HS imagery. Their method includes three main components: downscaling, detail injection, and optimization. They make use of the well-known Laplacian pyramid SR network LapSRN (see the "Standard Deep Learning Methods for Downscaling in Computer Vision" section) 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.

In [142], Xie et al. propose *MHF-Net*, a network having clear physical meaning and great interpretability. They first construct an HS/MS fusion model that merges the generalization models of LR images and the low-rankness prior knowledge of an HRHS 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 multiattention autoencoder network and a multiscale feature-guided network (*MSFG*). First, the multiattention autoencoder network extracts deep multiscale features of the MS image, and, then, a loss function containing a pair of HS and MS images is used to iteratively update the parameters of the network and learn prior knowledge of the fused image. The MSFG is used to construct the final HRHS image. Nonlocal blocks are used to better retain spectral and spatial details of the image. Laplacian blocks are used to connect the multiattention autoencoder network with the MSFG to achieve better fusion results while ensuring feature alignment. Although UMAG-Net does not use satellite HS data, the expansion into them is straightforward. Figure 21 shows the method.

Zhang et al. [144] propose *SSR-Net*, an interpretable spatial–spectral reconstruction network that consists of three components: 1) cross-mode message inserting (*CMMI*), an operation producing a preliminary fused HRHS image; 2) a spatial reconstruction network (*SpatRN*) that focuses on reconstructing the lost spatial information of the LRHS image with the guidance of a spatial edge loss; and 3) a spectral reconstruction network (*SpecRN*) that aims to reconstruct the lost spectral information of the HRMS image under the constraint of a spectral edge loss.

In conclusion, even though the architectures proposed in HS/MS fusion are limited in number, they exhibit remarkable variability (CNNs, 3D CNNs, GANs, and so on). The MHF-Net is an interpretable network showing superiority both visually and quantitatively. A bright idea that researchers should take into account is presented in the PFCN. The authors propose encoding the spectral information of the LRHS image into a latent image and then decoding this image with an HRMS image pyramid into a sharp HRHS image. The drawback of this method is the fact that experiments are conducted on simulated images. The SSR-Net treats HS/MS fusion as a spatial–spectral reconstruction problem. The authors provide a good ablation study and useful insights.

Finally, a complete solution that has not yet been tested on RS data is proposed in the UMAG-Net. This solution combines great ideas like the use of multiattention, nonlocal blocks, Laplacian blocks, and a loss function that measures both the spectral and the spatial similarity between pairs of images.

SPATIOTEMPORAL FUSION

Apart from their spectral signatures, satellites are also characterized by their unique revisit times. STF aims to integrate images of HSLT with images of LSHT. A typical data set 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, the Moderate-Resolution Imaging Spectroradiometer (MODIS) captures images daily (high temporal resolution) at a scale of 250 m to 1 km (low spatial resolution) [146], whereas *Landsat-8*'s Operational Land Imager (OLI) captures images every 16 days (low temporal resolution) at a 30-m scale

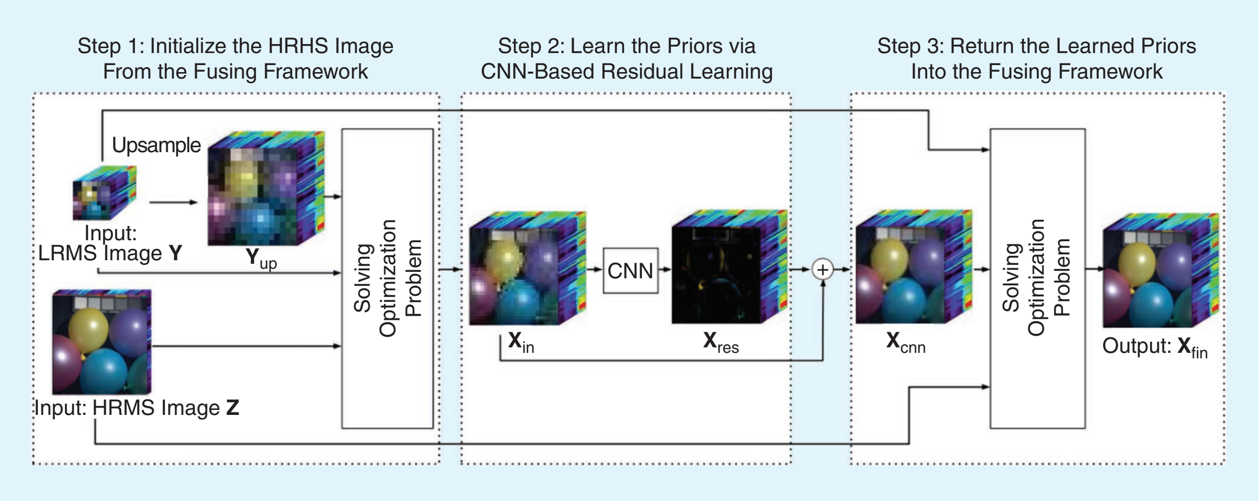
Step 1: Initialize the HRHS Image From the Fusing Framework
Step 2: Learn the Priors via CNN-Based Residual Learning
Step 3: Return the Learned Priors Into the Fusing Framework
Upsample
Input: LRMS Image Y
Y_{up}
Input: HRMS Image Z
Solving Optimization Problem
X_{in}
CNN
X_{res}
+
X_{cnn}
Solving Optimization Problem
Output: X_{fin}

FIGURE 19. An outline of DHSIS, a deep HS image-sharpening method. (Source: [136]; used with permission.)

(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 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 nontarget times only (Figure 22).

We must note here that a couple of other discriminant factors can also be observed among STF studies. First, some methods perform a preprocessing 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 inputs to the model. Such an approach is followed by [91] and [147]–[153]. Second, 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] and [150]–[158]. For simplicity, in this work, we solely employ the C-A versus CT-A classification and separately describe each category in the following sections, while, in Table 4, 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.

CONTEXT- AND TARGET-ASSISTED METHODS

Several researchers argue that the spatial resolution gap between certain sensors, such as those carried by MODIS and Landsat, is quite large and that data coming from both sources undergo different atmospheric and geometric corrections. Therefore, they design models that produce intermediate images enhanced by a smaller scaling factor to facilitate the downscaling process. For example, Song et al. [154] (*STFDCNN*) (Figure 23) propose a two-stage model that takes as the input an arbitrary pair of *Landsat-5/7* (25-m) and MODIS (500-m) images and learns to predict an intermediate enhanced image of 250-m spatial resolution. The intermediate image is computed in a preupsampling fashion, while the final 25-m image is computed via a post-upsampling SRCNN structure. During the inference, features are extracted from MODIS images at 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 on this, Zheng et al. [158] (*VDCNSTF*) propose deeper network architectures and redesign the SRCNN stage as a multiscale model producing images at 125 m and 25 m.

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 that takes as the input an LSHT MODIS image (250–300 m) at prediction date t_2, a date before (t_1) and a date after (t_3) the prediction date, and a corresponding HSLT Landsat

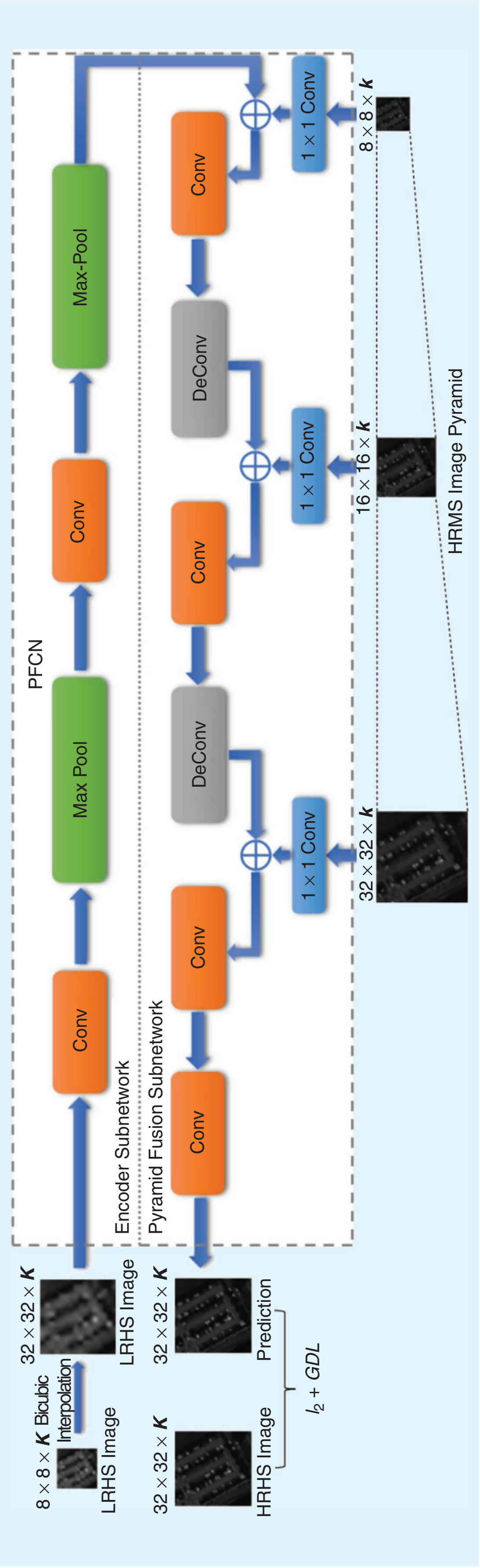

FIGURE 20. An outline of PFCN comprising an encoder subnetwork and a pyramid fusion subnetwork. (Source: [137]; used with permission.) DeConv: deconvolutional layer; GDL: gradient difference loss.

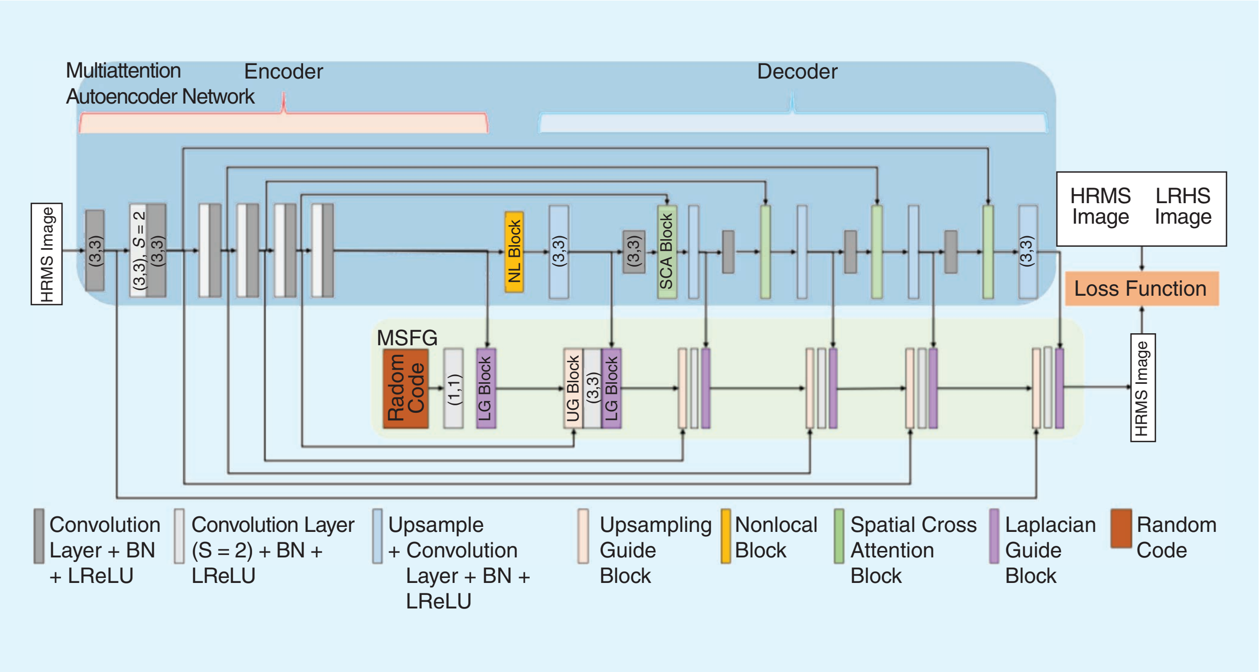

FIGURE 21. An outline of UMAG-Net comprising an encoder and a decoder with spatial cross attention mechanism. (Source: [143]; used with permission.) BN: batch normalization; LG: Laplacian guide; NL: nonlocal block; S: stride; SCA: spatial cross attention; UG: upsampling guide.

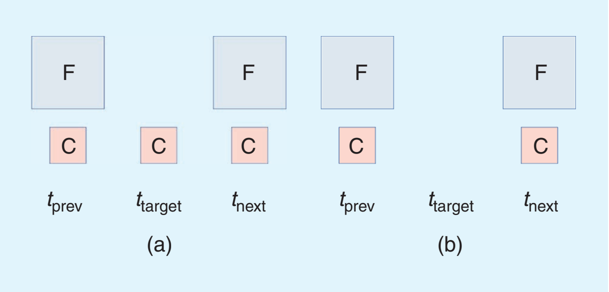

FIGURE 22. The data used for (a) CT-A and (b) C-A STF during training. *F* refers to the HSLT image, *C* refers to the LSHT image, and t_{prev} and t_{next} are one or multiple dates before and after the target date t_{target}, respectively.

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 that take as the 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: 1) $(C_{13}, F_1) \rightarrow F_{13}$ and $(C_{13}, F_3) \rightarrow F_{13}$ and 2) $(C_{12}, F_1) \rightarrow F_{12}$ and $(C_{23}, F_3) \rightarrow F_{23}$. Mapping 1 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 2 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 Figure 24. Compared with non-DL and DL approaches, the proposed StfNet achieves sharper results will fewer visible artifacts.

Tan et al. [159] (*DCSTFN*) < propose a two-branch CNN that takes as the input the LSHT MODIS image on prediction date t_2 along with a pair of HSLT *Landsat-8* and LSHT MODIS (500-m) 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 postupsampling 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 spatial and temporal adaptive reflectance fusion model (STARFM) algorithm [160], $F_2 = C_2 - F_1 - C_1$, for dates t_1 and t_2 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: instead of processing solely the LSHT images on the first branch, it takes as the input both the LSHT images and the HSLT image concatenated along the channel dimension and extracts information on their spectrum differences. Finally, the authors describe a novel flexible training scheme where more than one reference pair can be used as the input during either the training or the inference phase, depending on data availability. The proposed EDCSTFN

TABLE 4. A SUMMARY OF THE STATE-OF-THE-ART DL MODELS FOR STF FOR IMAGE DOWNSCALING IN RS.

MODEL	INPUT ASSISTANCE	TIME DIFFERENCE IMAGES	PRIOR DATES ONLY	CV MODEL	ARCHITECTURE	CODE AVAILABLE/ NUMBER OF PARAMETERS
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/409,000
EDCSTFN [161]	CT-A	No	Yes	–	CNN	Yes/282,000
DMNet [148]	CT-A	Yes	Yes	–	CNN	No/327,000
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/578,000 + 3.6 m
Teo and Fu [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/–

CV Model refers to the models presented in Table 2.

model manages to outperform DCSTFN and StfNet in most cases while displaying more stable and consistent behavior.

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 the input the MODIS time difference image C_{12} 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 (×4, ×8, and ×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 STF 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 nonadversarial 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] to enhance the traditional flexible spatiotemporal data fusion (FSDAF) algorithm [165]. The main framework consists of the following four stages:

1) *Generation*: A cycle GAN takes as the input the HSLT image pair (F_{t-1}, F_{t+1}) and produces an F_t^{GAN} in the output. The GAN produces a single image each time, so an iterative generation scheme is introduced to generate multiple in-between images.
2) *Selection*: A single F_t^{GAN} image is selected based on mutual information metrics of the HSLT and LSHT images.
3) *Enhancement*: The discrete wavelet transform is used to enhance the quality of the selected image, borrowing information from C_t.
4) *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 a loss of spectral information.

Zhang et al. [156] (*STFGAN*) propose a cascade of two SRGAN-like structures that 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 as well as MODIS data from dates t_1, t_2, and t_3. The first GAN takes as the input the two Landsat and all of the corresponding MODIS images and produces an intermediate Landsat image $\hat{F}_2^{\mathrm{int}}$. Due to the limited ability of the SRGAN for spatial enhancement to such a large scaling

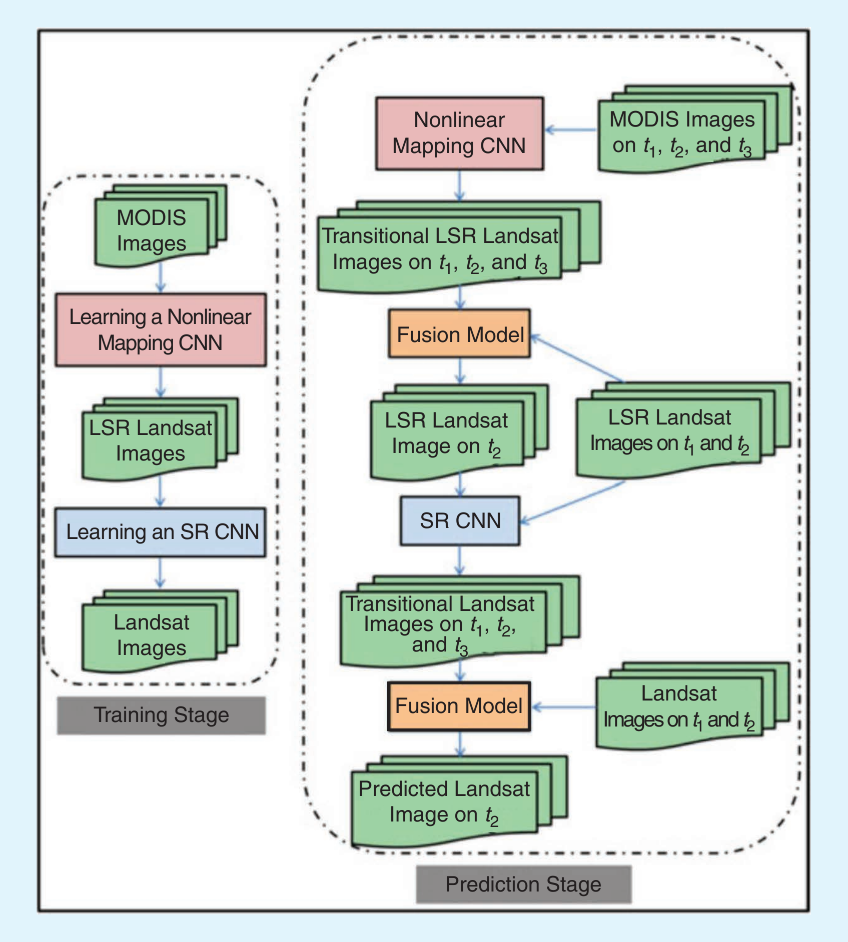
Nonlinear Mapping CNN
MODIS Images on t_1, t_2, and t_3
MODIS Images
Transitional LSR Landsat Images on t_1, t_2, and t_3
Learning a Nonlinear Mapping CNN
Fusion Model
LSR Landsat Images
LSR Landsat Image on t_2
LSR Landsat Images on t_1 and t_2
Learning an SR CNN
SR CNN
Landsat Images
Transitional Landsat Images on t_1, t_2, and t_3
Training Stage
Fusion Model
Landsat Images on t_1 and t_2
Predicted Landsat Image on t_2
Prediction Stage

FIGURE 23. An outline of the STFDCNN method. (Source: [154]; used with permission.)

factor (×16), this image is far from optimal. Therefore, a second GAN is used that takes as the input the Landsat images along with a downsampled version of these Landsat images and the intermediate $\hat{F}_2^{\text{int}}$ 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 inputs are 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 the 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} 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 EDC-STFN, 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* (8 m) images to increase the number of cloud-free observations have been studied by Teo and Fu [169]. First, Landsat images were resampled to 8 m and then blended with the rest via a simple STARFM algorithm. Second, pairs of Formosat and Landsat images obtained on the same date were fed to a VDSR model that learned 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 the best results overall, which implies that spatially enhancing the LR images before fusion can reduce the variation between the two image sets.

CONTEXT-ASSISTED METHODS

A C-A approach that 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 differences are computed forward in time, whereas in a time-backward pass, they are computed backward in time. In the former case, the learned 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 $M_1{}' : (C_{31}, F_3) \rightarrow \hat{F}_1^{1'}$ and $M_2{}' : (C_1, F_3 - C_3) \rightarrow \hat{F}_1^{2'}$. 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 $M_1{}' : (C_{32}, F_3) \rightarrow \hat{F}_2^{1'}$ and $M_2{}' : (C_2, F_3 - C_3) \rightarrow \hat{F}_2^{2'}$ for the backward pass. Figure 25 presents the entire pipeline.

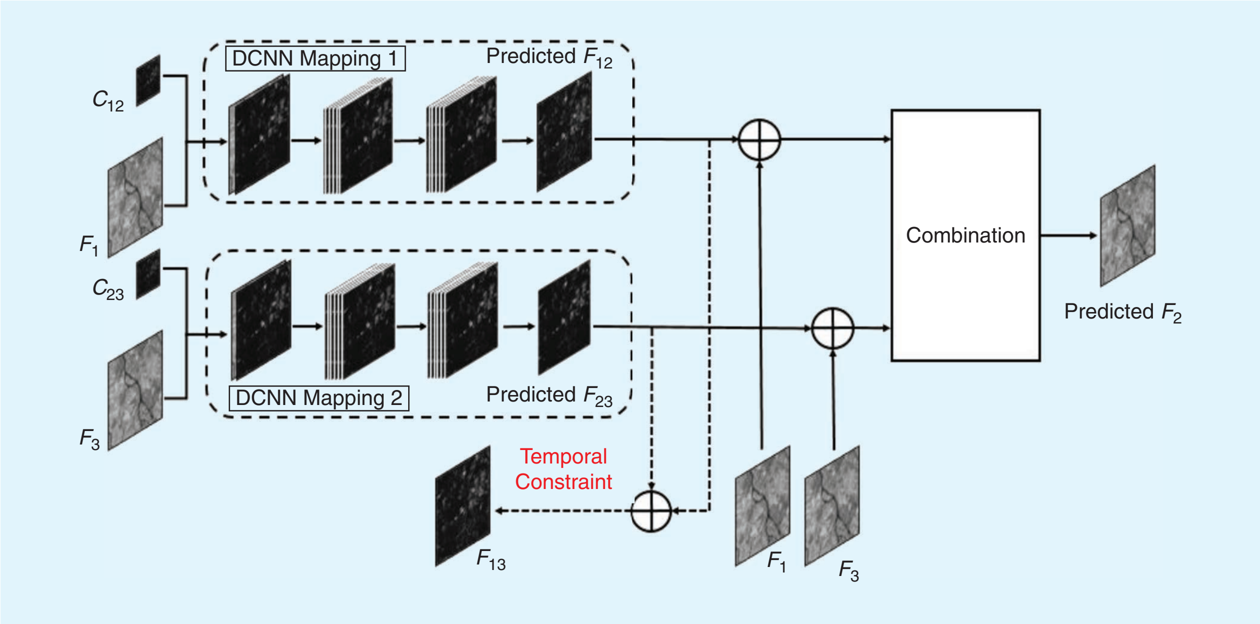

FIGURE 24. An outline of the StfNet method. *DCNN* refers to a three-layer deep CNN. (Source: [147]; used with permission.)

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 that involves an 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 the ×2 scale following the progressive upsampling scheme. During inference, temporal changes are captured by a linear model that extracts information from both F_1 and the intermediate output of LapSRN for images C_1 and C_2. In the final downscaled 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 the 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 the 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.

In conclusion, 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 data set, 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.

Second, 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 to better uncover land cover and phenological changes. The models accepting multiband input are EDCSTFN, GASTFN, STFGAN, GAN-STFM, DL-SDFM, and STF3DCNN.

SUPER-RESOLUTION

SR is a broad family of methods that aim to enhance the spatial resolution of an image without the need to blend information from auxiliary sources in either the spectral or

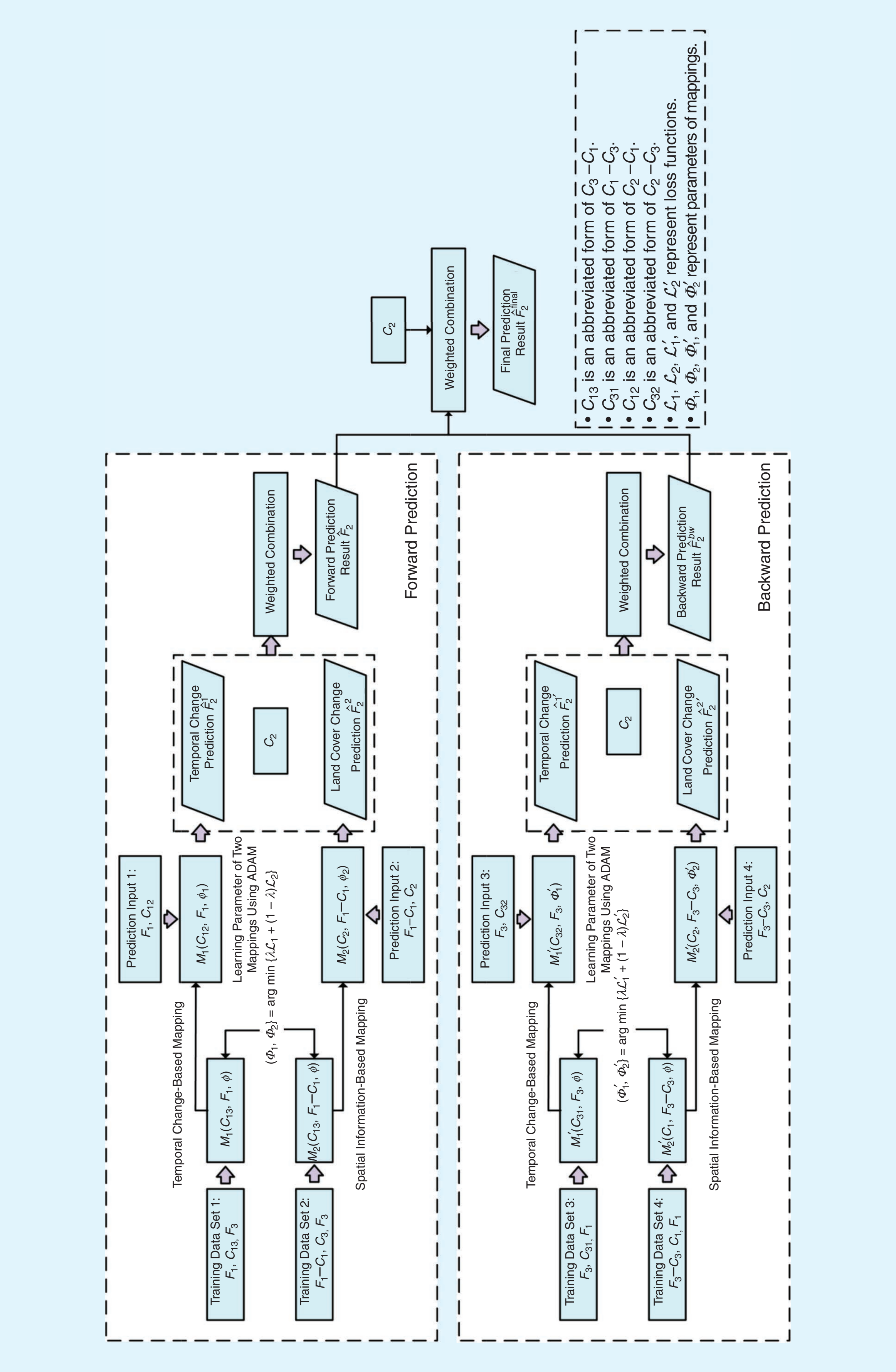
Training Data Set 1: F_1, C_{13}, F_3
Training Data Set 2: F_1-C_1, C_3, F_3
Temporal Change-Based Mapping
$M_1(C_{13}, F_1, \phi)$
$M_2(C_{13}, F_1-C_1, \phi)$
Spatial Information-Based Mapping
(Φ_1, Φ_2) = arg min $\{\lambda\mathcal{L}_1 + (1-\lambda)\mathcal{L}_2\}$
Learning Parameter of Two Mappings Using ADAM
Prediction Input 1: F_1, C_{12}
$M_1(C_{12}, F_1, \phi_1)$
$M_2(C_2, F_1-C_1, \phi_2)$
Prediction Input 2: F_1-C_1, C_2
Temporal Change Prediction $\hat{F}_2^1$
C_2
Land Cover Change Prediction $\hat{F}_2^2$
Weighted Combination
Forward Prediction Result $\hat{F}_2$
Forward Prediction
Training Data Set 3: F_3, C_{31}, F_1
Training Data Set 4: F_3-C_3, C_1, F_1
Temporal Change-Based Mapping
$M_1'(C_{31}, F_3, \phi)$
$M_2'(C_1, F_3-C_3, \phi)$
Spatial Information-Based Mapping
(Φ_1', Φ_2') = arg min $\{\lambda\mathcal{L}_1' + (1-\lambda)\mathcal{L}_2'\}$
Learning Parameter of Two Mappings Using ADAM
Prediction Input 3: F_3, C_{32}
$M_1'(C_{32}, F_3, \Phi_1')$
$M_2'(C_2, F_3-C_3, \Phi_2')$
Prediction Input 4: F_3-C_3, C_2
Temporal Change Prediction $\hat{F}_2^{1'}$
C_2
Land Cover Change Prediction $\hat{F}_2^{2'}$
Weighted Combination
Backward Prediction Result $\hat{F}_2^{bw}$
Backward Prediction
C_2
Weighted Combination
Final Prediction Result $\hat{F}_2^{\text{final}}$
• C_{13} is an abbreviated form of $C_3 - C_1$.
• C_{31} is an abbreviated form of $C_1 - C_3$.
• C_{12} is an abbreviated form of $C_2 - C_1$.
• C_{32} is an abbreviated form of $C_2 - C_3$.
• $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_1'$, and $\mathcal{L}_2'$ represent loss functions.
• Φ_1, Φ_2, Φ_1', and Φ_2' represent parameters of mappings.

FIGURE 25. The DL-SDFM pipeline. (Source: [150]; used with permission.) Adam: adaptive movement estimation algorithm.

the temporal dimensions. For better assessment, they can be categorized into *SISR*, *multiple-image SR* (*MISR*), and *reference SR* (*RefSR*). These are presented next, while, in Table 5, we summarize the main DL models developed for SR. In the "Super-resolution for Synthetic Aperture Radar and Aerial Imagery" section, we examine SR architectures that are specific for synthetic aperture radar (SAR) and aerial imagery.

SINGLE-IMAGE SUPER-RESOLUTION

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 noninvertible, but it is often 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 RS.

MULTISCALE APPROACHES

Lei et al. [171] (*LGCNet*) (Figure 26) design a CNN model that combines feature maps produced by previous layers to extract information at different scales and levels of detail. The model was evaluated on the University of California (UC), Merced data set and selected *Gaofen-2* images, and it 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 a performance similar to that of LGCNet.

Lu et al. [173] (*MRNN*) propose a preupsampling 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 fine-tuned 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, Xu et al. [174] 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.

MULTITASK LEARNING

In their study, Yan and Chang [175] (*MSF*) exploit a multitask learning procedure 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 multitask setting where each task represents a separate Gaussian kernel and is learned by a dedicated CNN.

ADDITIONAL POSTPROCESSING

A study by Qin et al. [176] (*DGANet-ISE*) presented a custom postprocessing pipeline for the improvement of the output of an SR model. Their architecture is heavily based on EDSR (see the "Standard Deep Learning Methods for Downscaling in Computer Vision" section) and is trained with a custom loss function that additionally considers the gradient similarity between the prediction and target. The model's output is then iteratively improved via a proposed image-specific enhancement (ISE) algorithm that 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 data sets that might occur from different sensing platforms, light conditions, and so on.

DIFFERENT SOURCES FOR THE INPUT AND OUTPUT

Contrary to most approaches in this category that exploit Wald's protocol, a number of methods have been proposed that utilize different sources for the input and output. Galar et al. [177] (*S2PS*) propose the use of PlanetScope images as the target to downscale the four *Sentinel-2* 10-m 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 the 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 behavior 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 linear imaging self scanning (LISS) III, which captures information in the green, red, NIR, and SWIR bands with 24-m spatial resolution and a 24-day revisit cycle, and advanced wide field sensor (AWiFS), which captures the same bands with 56-m spatial resolution and a five-day revisit cycle. The authors used a training set

TABLE 5. A SUMMARY OF THE STATE-OF-THE-ART DL MODELS FOR SR IN RS.

MODEL	SR TYPE	DESCRIPTION/NOVELTY	CV MODEL	BUILDING BLOCKS	UPSAMPLING FRAMEWORK	ARCHITECTURE	CODE AVAILABLE/ NUMBER OF PARAMETERS
LGCNet [171]	SISR	Multiscale approach and features from different layers	—	Residual learning	Preupsampling	CNN	No/—
Haut et al. [172]	SISR	Multiscale approach with inception module	—	Residual learning and subpixel convolution	Postupsampling	CNN	No/—
MRNN [173]	SISR	Multiscale approach and parallel feature extraction from different scales of the LR input	—	Residual learning	Preupsampling	CNN	No/—
Xu et al. [174]	SISR	Multiscale approach and U-Net model with dilation module at the bottleneck	—	Residual learning and subpixel convolution	Postupsampling	CNN	No/—
MSF [175]	SISR	Multitask learning and a different model for each Gaussian kernel	—	Residual learning	Preupsampling	CNN	No/—
DGANet-ISE [176]	SISR	Postprocessing algorithm and gradient loss term	EDSR	Residual learning and subpixel convolution	Postupsampling	CNN	No/—
S2PS [177]	SISR	Downscaling of *Sentinel-2* images using PlanetScope as the target	EDSR	Residual learning and subpixel convolution	Postupsampling	CNN	No/—
DCR-SRCNN [179]	SISR	Downscaling of *Landsat-5/8* images using *Sentinel-2* as the target	SRCNN	Residual learning	Preupsampling	CNN	No/993,000
Collins et al. [180]	SISR	Downscaling of coarser AWiFS images using sharper LISS III images from Resourcesat	SRCNN	—	Preupsampling	CNN	No/—
Zhang et al. [183]	SISR	Unsupervised model that learns multiple image degradations	—	Residual learning and bilinear upsampling layers	Postupsampling	GAN	No/—
EUSR [181]	SISR	Dense network, with the resulting image downsampled and compared with the LR input	—	Bilinear upsampling layers	Postupsampling	CNN	No/—
WTCRR [185]	SISR	Approach assisted by the discrete wavelet transform and use of recurrent blocks	DRRN	Residual learning	Preupsampling	CNN	No/—
DWTSR [186]	SISR	Approach assisted by the discrete wavelet transform and stationary wavelet transform	—	Residual learning	Preupsampling	CNN	No/—
RRDGAN [187]	SISR	Approach assisted by the discrete wavelet transform and the total variation loss function	ESR-GAN	Residual learning and subpixel convolution	Postupsampling	GAN	No/—
MPSR [189]	SISR	Multiscale approach with residual connections and channel attention	—	Residual learning, subpixel convolution, and attention mechanism	Postupsampling	CNN	No/—
DRSEN [190]	SISR	Approach with channel attention	EDSR	Residual learning, subpixel convolution, and attention mechanism	Postupsampling	CNN	No/8.6 m
Haut et al. II [191]	SISR	Approach with channel attention	—	Residual learning, subpixel convolution, and attention mechanism	Postupsampling	CNN	No/—
MSAN and SAMSAN [192]	SISR	Approach with channel attention and scene-adaptive learning	WDSR	Residual learning, subpixel convolution, and attention mechanism	Postupsampling	CNN	No/—
DSSR [193]	SISR	Approach with channel attention and chain training	WDSR	Residual learning, subpixel convolution, and attention mechanism	Postupsampling	CNN	No/9.1 m
AMFFN [194]	SISR	Multiscale approach with channel attention	—	Residual learning, subpixel convolution, and attention mechanism	Postupsampling	CNN	No/—
IRAN [195]	SISR	Approach with inception modules and both channel and spatial attention	—	Residual learning, subpixel convolution, and attention mechanism	Postupsampling	CNN	No/1.88 m

(Continued)

TABLE 5. A SUMMARY OF THE STATE-OF-THE-ART DL MODELS FOR SR IN RS. (*Continued*)

MODEL	SR TYPE	DESCRIPTION/NOVELTY	CV MODEL	BUILDING BLOCKS	UPSAMPLING FRAMEWORK	ARCHITECTURE	CODE AVAILABLE/ NUMBER OF PARAMETERS
NLASR [196]	SISR	Multiscale approach with nonlocal modules and both channel and spatial attention	—	Residual learning, subpixel convolution, and attention mechanism	Iterative up- and downsampling	CNN	No/10.7 m
PGCNN [198]	SISR	Approach with channel attention	EDSR	Residual learning, subpixel convolution, and attention mechanism	Postupsampling	CNN	No/1.44 m
HSENet [199]	SISR	Attention for multiscale recurring features	—	Residual learning, subpixel convolution, and attention mechanism	Postupsampling	CNN	Yes/—
BCLSR [200]	SISR	Recurrent convolutional model	—	Residual learning and subpixel convolution	Postupsampling	CNN	Yes/170,000
CDGAN [201]	SISR	Coupled discriminator	ESR-GAN	Residual learning and subpixel convolution	Postupsampling	GAN	No/1.4 m
DRGAN [202]	SISR	RDN-like generator	RDN	Residual learning and subpixel convolution	Postupsampling	GAN	No/—
RS-ESRGAN [203]	SISR	Multiple training phases on different data sets	ESR-GAN	Residual learning	Preupsampling	GAN	Yes/—
udGAN [204]	SISR	Multiscale generator with ultradense residual blocks	—	Residual learning and subpixel convolution	Postupsampling	GAN	No/2.4 m
Shin et al. [205]	SISR	Multiscale generator with pyramidal structure and discriminator with difference of Gaussian kernels on feature maps	—	Residual learning and subpixel convolution	Progressive upsampling	GAN	No/—
Enlighten-GAN [206]	SISR	Multiscale generator with intermediate output and the clipping-and-merging method	ESR-GAN	Residual learning and subpixel convolution	Progressive upsampling	GAN	No/—
EEGAN [207]	SISR	Downscaling assisted by edge enhancement and attention	—	Residual learning, subpixel convolution, and attention mechanism	Progressive upsampling	GAN	Yes/—
E-DBPN [92]	SISR	DBPN-like generator with channel attention on multiple layers	DBPN	Residual learning, transposed convolution, and 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, and subpixel convolution	Postupsampling	GAN	No/4.8 m
EvoNet [209]	MISR	Approach assisted by evolutionary image model algorithm	—	Residual learning	Preupsampling	CNN	No/—
Märtens et al. [211]	MISR	Simple CNN for *PROBA-V* images that takes as the input a concatenation of the LR images	—	—	Preupsampling	CNN	No/119,000
DeepSUM [212]	MISR	SR of each input separately and fusion of results	—	Residual learning	Preupsampling	CNN	Yes /—
Deep-SUM++ [213]	MISR	Extension of DeepSUM with graph convolutional operations	—	Residual learning	Preupsampling	CNN	No/—
HighRes-Net [214]	MISR	Paired SR of an LR image and the chosen reference LR as well as Shift-Net for registration of results	—	Residual learning and transposed convolution	Postupsampling	CNN	Yes/600,000 + 34 m
MISR-GRU [216]	MISR	LR images regarded as a time series; paired SR performed at each time step, similar to HighRes-Net; and uses ConvGRU layers and ShiftNet	—	Residual learning and transposed convolution	Postupsampling	CNN	Yes/900,000
RAMS [218]	MISR	Approach assisted by 3D convolutions and attention modules	—	Residual learning, subpixel convolution, and attention mechanism	Postupsampling	3D CNN	Yes/1 m
SD-GAN [219]	Ref-SR	Saliency information used as reference	—	Residual learning and subpixel convolution	Postupsampling	GAN	No/—

(Continued)

TABLE 5. A SUMMARY OF THE STATE-OF-THE-ART DL MODELS FOR SR IN RS. (*Continued*)

MODEL	SR TYPE	DESCRIPTION/NOVELTY	CV MODEL	BUILDING BLOCKS	UPSAMPLING FRAMEWORK	ARCHITECTURE	CODE AVAILABLE/ NUMBER OF PARAMETERS
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 and subpixel convolution	Postupsampling	GAN	Yes/—
SR-GAN [223]	SISR	—	SRGAN	Residual learning and subpixel convolution	Postupsampling	GAN	No/—
NF-GAN [224]	SISR	Generator based on residual encoder–decoder, discriminator based on ResNet50, and embodies despeckling component	—	Residual learning and transposed convolution	Preupsampling	GAN	No/—
Di-GAN [225]	SISR	Generator based on U-Net and discriminator based on PatchGAN-like network	—	Residual learning and transposed convolution	Preupsampling	GAN	No/—
FSRCNN [226]	SISR	—	—	Residual learning	Preupsampling	CNN	No/—
PSSR [227]	SISR	Learnable preupsampling, uses a complex structure block for complex numbers, uses residual compensation approach, and uses fully polSAR	—	Residual learning and transposed convolution	Preupsampling	CNN	No/—
WDCCN [228]	SISR	Import weighted dense connections	DRCN	Residual learning	Preupsampling	CNN	No/—
MSSRRC [229]	SISR	Uses residual compensation and uses fully polSAR data	VDSR	Residual learning	Preupsampling	CNN	No/—

CV Model refers to the models presented in Table 2. ConvGru: convolutional gated recurrent unit; LISS: linear imaging self scanning; polSAR: polarimetric synthetic aperture radar; SAR: synthetic aperture radar.

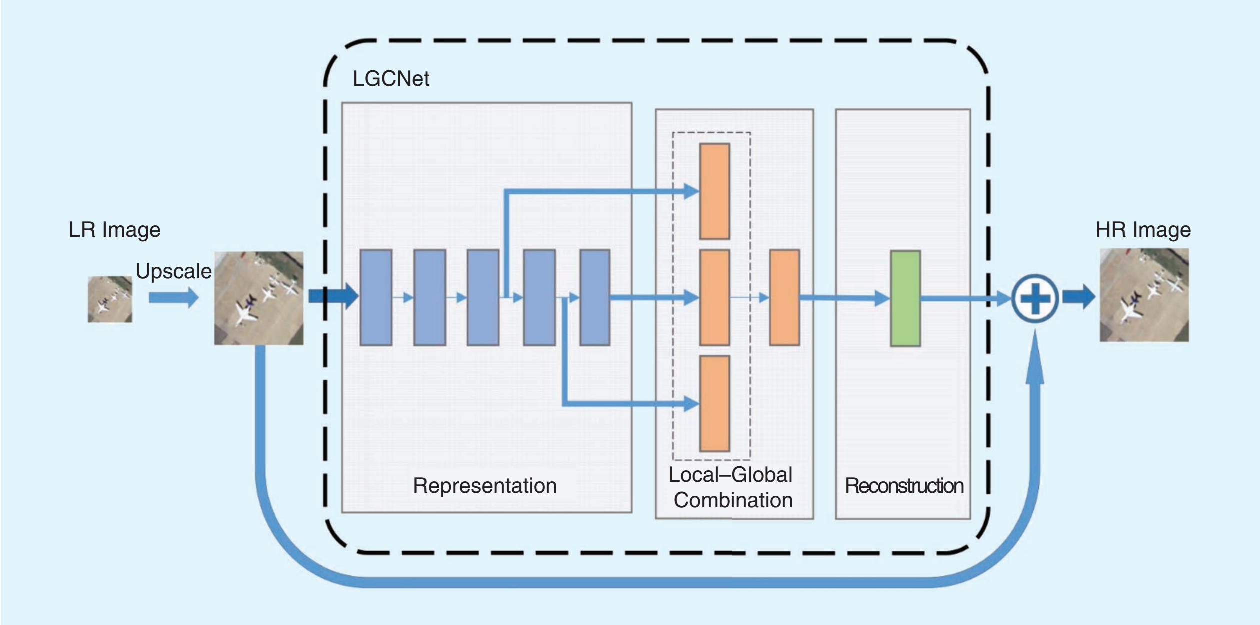

FIGURE 26. A high-level overview of LGCNet. Blue boxes represent convolutional layers followed by ReLU activation, orange boxes represent the concatenation of selected feature maps via a convolutional layer, and the green box represents the last convolutional layer for the final reconstruction. (Source: [171]; used with permission.)

with coincident images from the two satellites 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 peak signal-to-noise ratio (PSNR) and structural similarity index (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 the literature, the authors downsample the initial data set 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 Wald's protocol produce synthetic LR images through a specific distortion model and develop methods that focus solely on the enhancement of such LR images, Zhang et al. [183] propose an unsupervised model to handle multidegradation schemes. In particular, their approach involves a post-upsampling generator network that produces an SR image and a degrader network that 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 data sets (see the "Data Sets" section), 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 SR 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 (*WTCRR* [185], *DWTSR* [186], and *RRDGAN* [187]) that 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 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] 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 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 residual channel attention block (RCAB) attention module [89] inside convolutional blocks with residual connections at multiple levels. RCAB is also adopted by Zhang et al. [192] (*MSAN* and *SAMSAN*), who additionally propose a scene-adaptive learning framework where a separate model is fine-tuned on each possible scene depicted in an RS image, and Dong et al. [193] (*DSSR*) 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 and Liu [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 nonlocal blocks [197] that follows the iterative up- and downsampling scheme with channel and spatial attention modules.

Finally, based on the popular EDSR architecture, Peng et al. [198] (*PGCNN*) propose a gated residual block that encourages the model to focus on high-frequency details, whereas Lei and Shi [199] (*HSENet*) employ custom attention modules that aim to discover information recurring at 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 and Luo [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 that share their parameters and feed their outputs to a bidirectional convolutional long short-term Memory layer (BiConvLSTM). The output is then downscaled via a subpixel convolution. The results show 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 RS 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 the input both the predicted SR image and its corresponding HR ground truth shuffled by a random gate and is then tasked with deciding whether the input constitutes a real–fake pair (one) or a fake–real pair (zero). The generator architecture is based on ESRGAN. The model competed against a number of DL methods on the UC Merced and

Wuhan University-Remote Sensing (WHU-RS19) data sets (see the "Data Sets" section) as well as selected *GaoFen-2* images and produced less blurry results 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 subpixel convolution for downscaling and a VGG loss function. Their model was evaluated on the NWPU-RESISC45 data set (see the "Data Sets" section) and several other CV 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 preupsampling framework and train the generator in three stages: first, it is trained on a set of WorldView images only; then, it is fine-tuned on pairs of WorldView and *Sentinel-2* images; and, finally, it is 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 tradeoff.

MULTISCALE GENERATORS

Dense and multilevel connections have also been introduced to different generator architectures with the aim of extracting more accurate representations of both small- and large-scale objects. For example, Wang et al. [204] (*udGAN*) design a novel ultradense residual block that 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 RDBs. 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 is applied on these feature maps, and the results are then fed to a discriminator model with a PatchGAN architecture. Both networks are illustrated in Figure 27. The proposed method achieved much better results compared to EEGAN and CDGAN, and it 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*) that 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

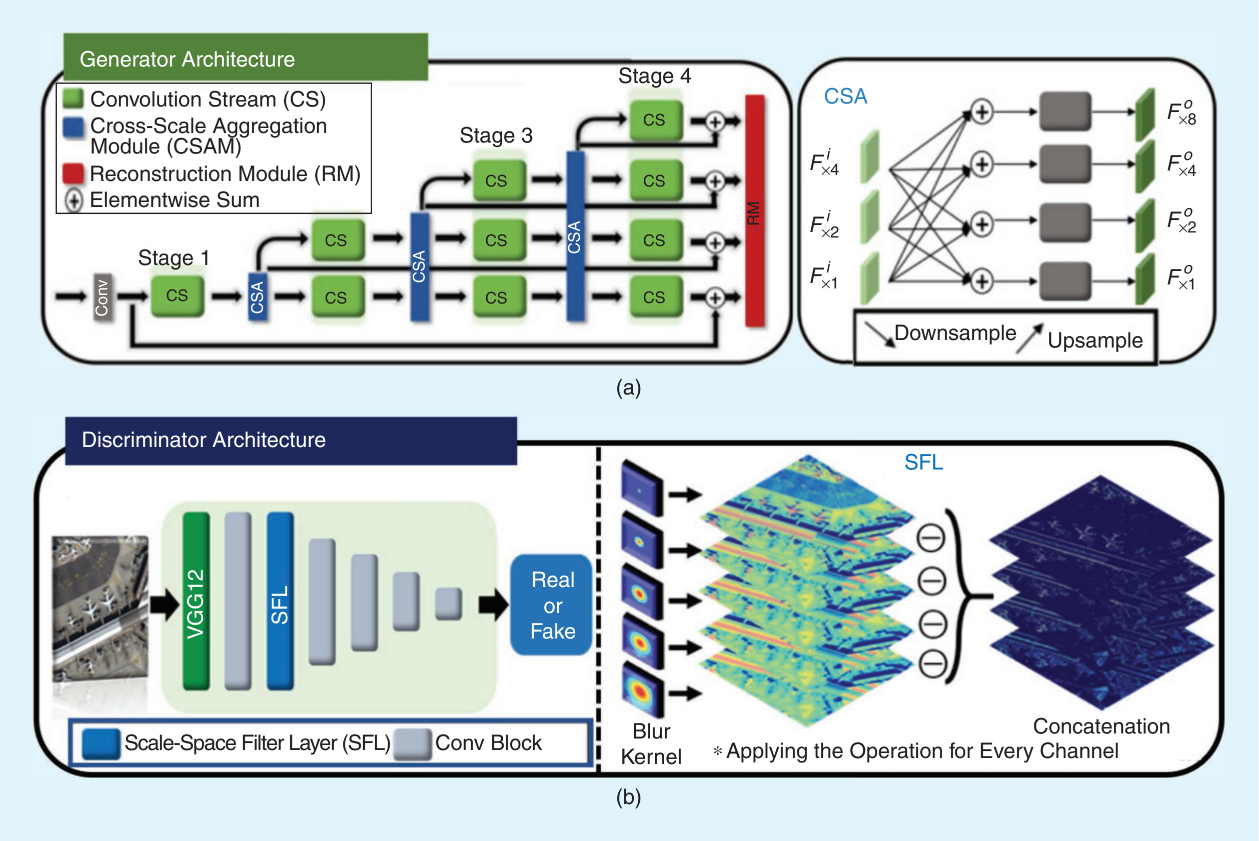

FIGURE 27. The (a) generator and (b) discriminator for the GAN proposed in [205]. (Source: [205]; used with permission.)

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 (Figure 28). Experimental results showed that Enlighten-GAN produces sharper images with much fewer artifacts than other GAN-based methods while, at the same time, retaining the true hues and shapes of the objects.

GENERATIVE ADVERSARIAL NETWORKS 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 that first enhances the input and then extracts and sharpens its edges (Figure 29). A mask branch with an attention mechanism is also employed during the edge-enhancement step to focus on the useful information. The model outperforms SRGAN, VDSR, and SRCNN on the Kaggle Draper Satellite Image Chronology data set (see the "Data Sets" section).

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, fine-tune it in an adversarial setting. The results show 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 to capture short- as well as long-range dependencies between pixels. Several experiments proved the superiority of the proposed model, especially at higher scaling factors.

MULTIPLE-IMAGE SUPER-RESOLUTION

In an MISR setting, a model takes as the 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

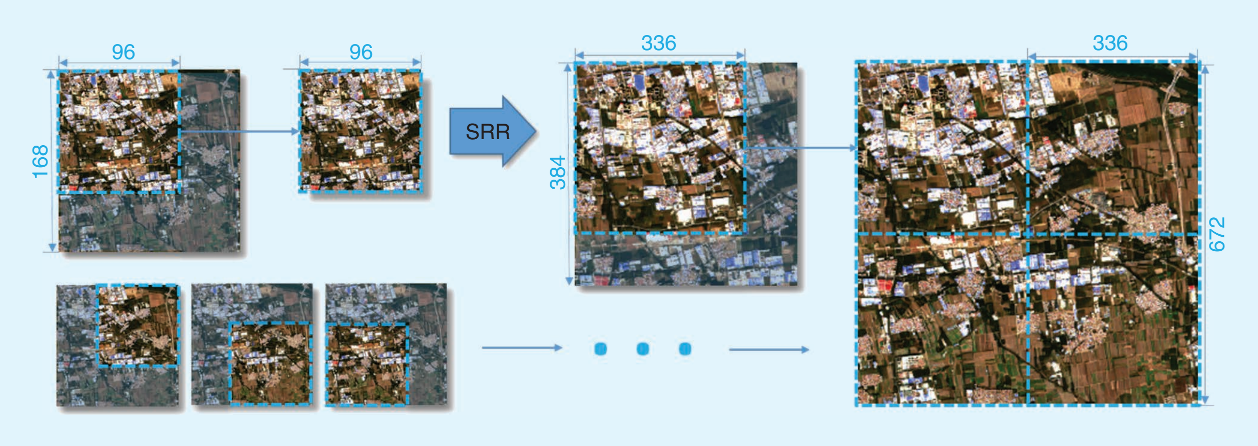

FIGURE 28. An example of the clipping-and-merging method pipeline. The input image has a size of 168 × 168 and is cropped into four overlapping patches, each with a size of 96 × 96. The patches are independently downscaled by an SR algorithm (denoted *SRR* here), producing four 384 × 384 images. Half 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. (Source: [206]; used with permission.)

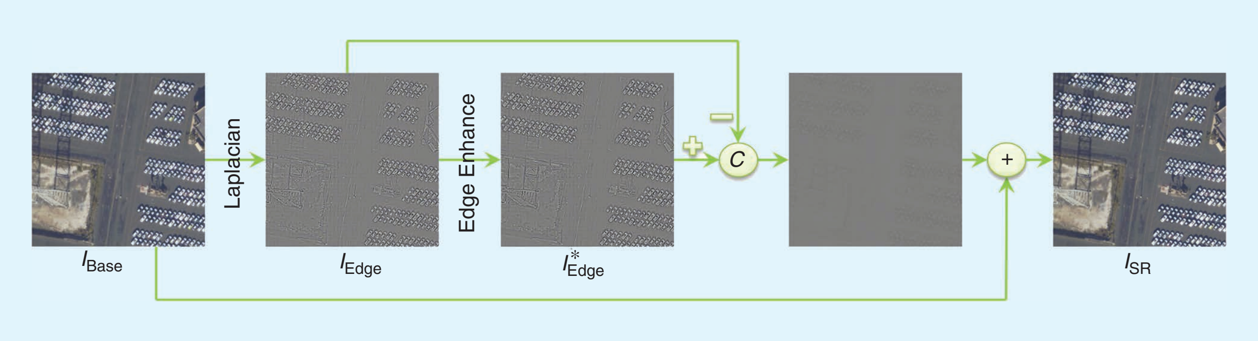

FIGURE 29. The pipeline of the edge-enhancement procedure for EEGAN. (Source: [207]; used with permission.) *I*: image; *C*: combination; I_{Base}: the intermediate SR result; I_{Edge}: edge map extracted from the intermediate SR result; I^*_{Edge}: edge map extracted from the final SR result.

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 STF or SSF is the fact that both LR and HR images contain information on the same spectra, whereas their acquisition times are never coincident.

Such an 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 coregistered and fed to the evolutionary image model 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 to produce a *Satellite pour l'Observation de la Terre* (*SPOT*)-like HR output downscaled by a ×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 *Project for On-Board Autonomy-Vegetation* (*PROBA-V*) satellite, which is able to capture MS images at 300-m spatial resolution every day and 100-m spatial resolution every five 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 an MISR approach, with the authors in [211] proposing a *PROBA-V* data set exclusively for this problem setting. They also design a simple four-layer CNN for benchmarking and propose a custom metric that takes into account spatial displacements between the prediction and the ground truth.

In their study [212] (*DeepSUM*) (Figure 30), Molini et al. design a network that downscales an NIR or red band of *PROBA-V* data. The model takes as the input a single image and performs feature extraction. All extracted features are then coregistered and fused in the feature space. Before the final fusion, a mutual inpainting process is employed to replace unreliable pixels in a feature map (such as clouds, shadows, and so on) with values taken from the 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* SR 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 to exploit nonlocal correlations among pixels.

Another popular method for the *PROBA-V* data set 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 the 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; it registers the SR with the target HR image to properly calculate

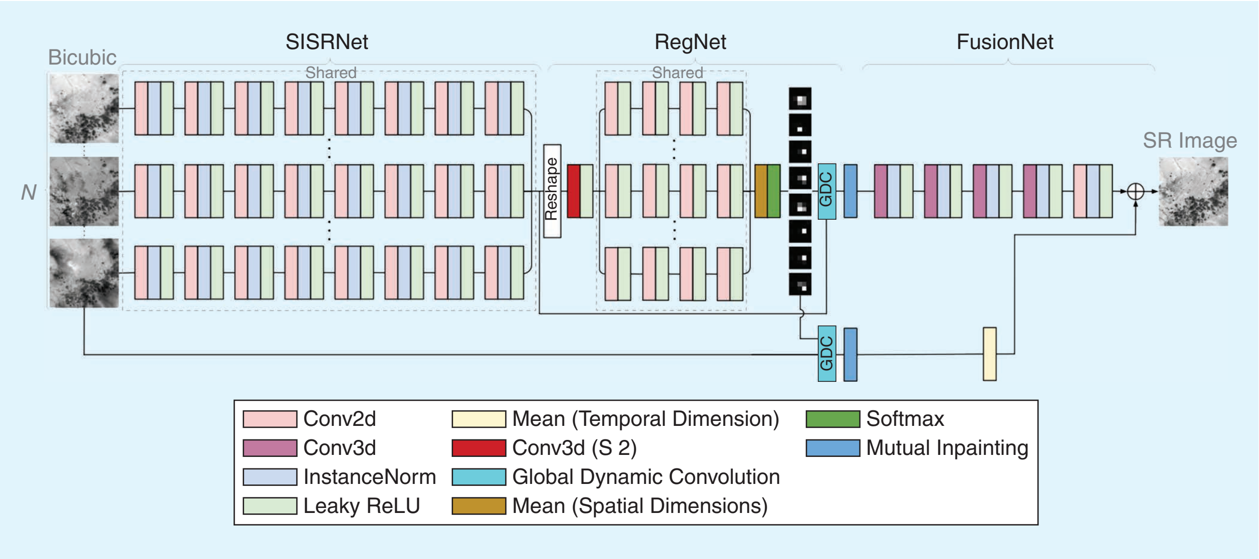

FIGURE 30. An overview of the DeepSUM model. *SISRNet* performs the feature extraction, *RegNet* the feature registration, and *FusionNet* the final feature fusion and reconstruction. The global dynamic convolution (GDC) is a convolution between an image and the corresponding learned filter for the image registration. (Source: [212]; used with permission.) *N*: number of input images.

the loss function. Without such a 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*) (Figure 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 the input one LR image and the median of all LR inputs, coregisters them, and produces a unified feature map. The output of this stage is then fed to a stack of convolutional gated recurrent (ConvGRU) unit 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 that involves a brightness bias. MISR-GRU achieved the highest score compared with FSRCNN, SRResNet, DeepSUM, and HighRes-Net, and the authors conclude that the proposed model's accuracy is 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*); it employs 3D convolutions and attention mechanisms on both the temporal and spatial domains to downscale a single band of *PROBA-V* data. The 3D convolutions are able to assess the interrelations 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 a temporal self-ensembling strategy and observed a significant increase in the output accuracy but at the expense of computational speed.

REFERENCE SUPER-RESOLUTION

In RefSR, the input of the model is accompanied by an auxiliary (reference) image, which provides additional information to assist in the downscaling process. A number of studies have explored using features extracted from the original data as the 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]–[221]) that focus on the saliency information of the input images. In [219] (*SD-GAN*) (Figure 32), the authors discriminate the highly salient areas of an image as the foreground and the less salient as the 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 that takes as the 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 nonsalient (background) areas. Experimentation on *GeoEye-1* PAN images showed that SD-GAN outperformed other DL approaches, such as SRCNN, ESPCN, VDSR, and SRGAN. A qualitative analysis proved that it managed to produce fewer 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 nonsalient information in a recursive fashion, and the final output of the network is the output of the last iteration. Similar to SD-GAN, a salient area discriminator and a nonsalient area discriminator are employed along with a global discriminator that takes as the input SR or HR images and learns to classify them. Then, the outputs of all discriminators over all iterations are 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 those of the other DL approaches, especially for higher scaling factors (×3 and ×4).

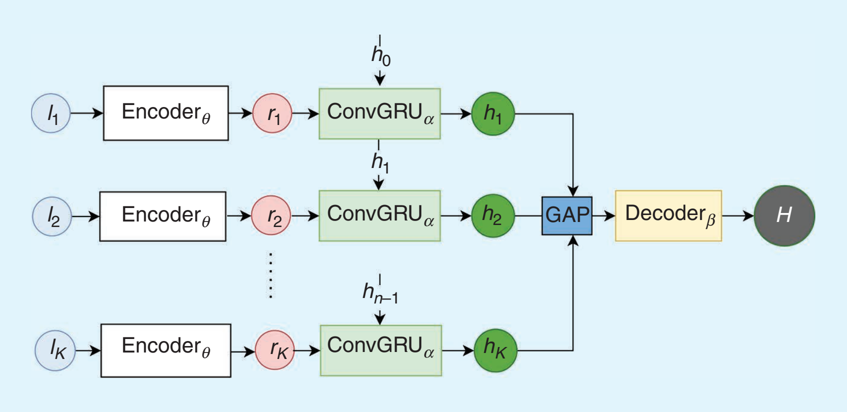

FIGURE 31. An overview of the MISR-GRU model, where I_i is the ith LR input image, H is the predicted downscaled image, h_i is the ith hidden state of the ConvGRU layer. In the original article, the encoder comprises two convolutional layers and two residual blocks (each with two convolutional layers and parametric ReLU activation), while the decoder consists of a deconvolutional layer and two 1 × 1 convolutional layers. GAP: global average pooling layer. (Source: [216]; used with permission.)

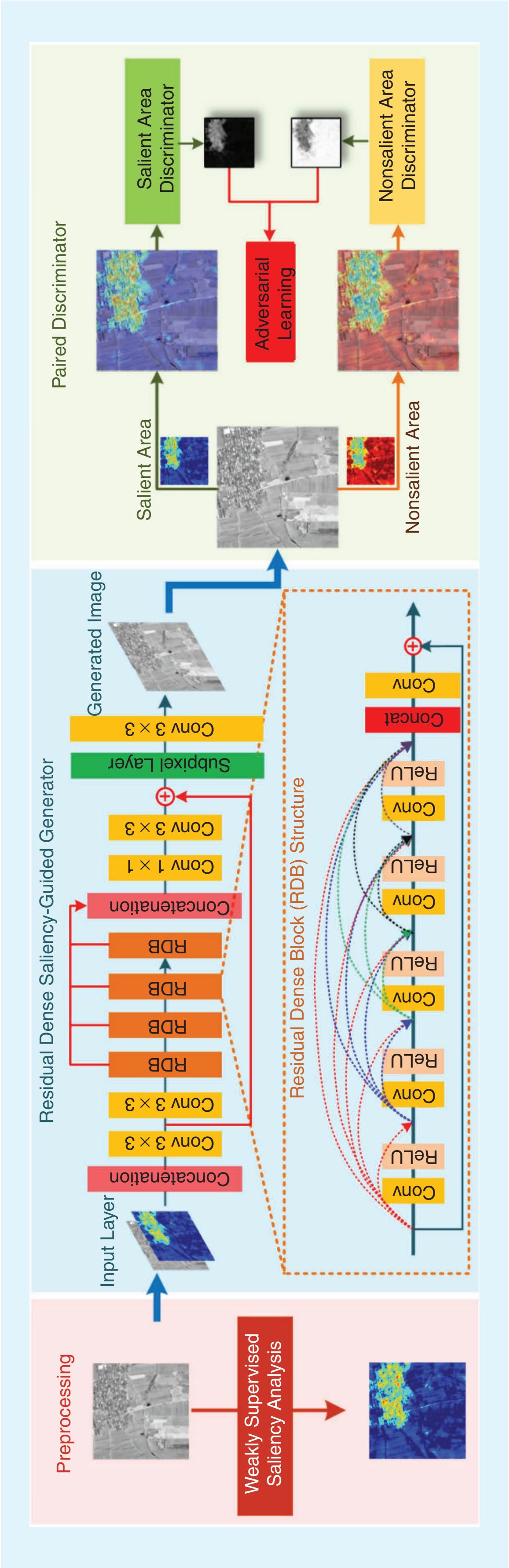

FIGURE 32. The SD-GAN model. (Source: [219]; used with permission.)

To summarize this analysis, there are two main approaches a researcher can take, depending on the number of available images in the data set at hand. When only a single LR image can be acquired per occasion, SISR and 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 [191], and NLASR) can always assist the discovery and preservation of high-frequency components, whereas multiscale feature extraction structures (e.g., NLASR and Shin et al. [205]) can unravel nonlocal 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 postprocessing technique and the autoencoder loss proposed in Enlighten-GAN. Finally, certain methods (EUSR, DWTSR, DRSEN, DSSR, DGANet-ISE, NLASR, Shin et al. [205], and SG-FBGAN) manage to perform better at larger scaling factors, whereas Zhang et al. [183] 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 coregistration of the input LR images, which is handled differently by each proposed model, either inside the network or as a separate preprocessing step in the pipeline. In addition, this coregistration 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 was proposed in HighRes-Net and, 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.

SUPER-RESOLUTION FOR SYNTHETIC APERTURE RADAR AND AERIAL IMAGERY

SYNTHETIC APERTURE RADAR

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 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 an SISR approach by applying an SRGAN on *TerraSAR-X* images after having been despeckled using a CNN, as described in [230]. The HR image is downsampled by a factor of four using a Gaussian kernel, while both the generator and discriminator elements are CNN based. The generator element produces the SR image using the LR image, while the discriminator compares the SR image with the HR image. The loss function comprises a perceptual loss with a content (pixelwise 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 horizontal–horizontal (HH) polarization channel of Airborne SAR data. The generator element consists of a despeckling network and the reconstruction network, while the discriminator element is ResNet based. The despeckling network is pretrained using optical images with speckle noise added on them, and it uses an MSE loss. Its input is an LR (downsampled HR version by a factor of two) noise-full image. As with the previous case, the NF-GAN objective function is defined by an adversarial and a pixelwise (MSE) component. The authors train their network pipeline with and without the despeckling component and show that the former, indeed, works better.

Li et al. [225] tried to solve the problem of increased system integration time and low azimuth resolution of 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 satellite systems. In particular, the authors generated synthetic GEO SAR data based on advanced land observing satellite phased array type L-band SAR (ALOS PALSAR) characteristics. They use a dialectical GAN (*Di-GAN*) [231] with the generator element comprising a U-net and the discriminator a PatchGAN-like network. The generator takes the LR simulated GEO SAR image as the input, whose SR-produced image is compared with the ALOS PALSAR HR in the discriminator. The authors claim a noticeable improvement of the resolution, which is mostly based on a 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 LR 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 coregistration accuracy between two single-look complex images of *European Remote Sensing-1/2 (ERS-1/2)* data. They train a scale-invariant SR CNN (*SINV CNN*) model using both the amplitude and phase, which mainly takes advantage of the feature extraction and residual block 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 CV, 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. Unlike [232], they do not treat the real and imaginary image parts separately but utilize them with a separate structure block since the information is lost because of separation. They use various satellite sensors, such as *Radarsat-2*, experimental SAR (ESAR), and polarimetric and interferometric SAR (PiSAR) whose data they despeckle first. They compare their approach (along with the residual compensation strategy) with the conventional non-DL approach and multichannel SAR SR (MSSR) using the PSNR and MAE. They also use equivalent number of looks, 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 use PolSAR data and propose a residual compensated MSSR (*MSSRRC*) to tackle issues of the conventional (non-DL-based) SR approaches, such as the insufficient use of polarimetric information and decreased reconstruction of details. Their network is a VDSR adjusted for multichannel (full-PolSAR) input applied on *RadarSat-2* data that is compensated for by residuals between LR reconstructed and original images. Prior to the training, all data are despeckled. 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 they claim is a better alternative to fast SR CNNs and DRCN. Their network is based on DRCN as well as the notion of weighted dense connections, and it 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, and so on, 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, and so on) 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 a baseline, researchers could begin with general noise suppression architectures established in 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., parametric ReLU, leaky ReLU, and so on) 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 UNMANNED AERIAL VEHICLES/DRONES

By their initial mass production and market distribution, unmanned aerial vehicles (UAVS) represent one of the most applicable and simple means of data acquisition influencing a plethora of applications, including RS. Simple architectures as well as 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 RS 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 for satellite alternatives, as the latter exhibit enhanced payload sensor technical specifications, such as higher spatial resolution.

Aiming at exploiting UAV systems in specific RS applications and higher spatial resolution for the acquired images, numerous SR 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 focus on only 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 onboard 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 upsampling 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 HR 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 (see the "Standard Deep Learning Methods for Downscaling in Computer Vision" section) for UAV operations has been incorporated as an initial processing step by 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.

A 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 by 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 data set availability.

In conclusion, most approaches applied in the resolution increment of aerial images follow similar schemes, as the problem is translated into a CV counterpart. The majority of the corresponding architectures rely on the extraction of features from pretrained 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 in terms of both the accuracy and the execution time with respect to more universal, generic, and heavyweight modeling solutions.

DATA SETS

Despite the abundance of RS images, there is still a noticeable gap in the availability of public benchmark data sets for the evaluation of downscaling methods. This is hardly surprising since such a benchmark data set would require extremely careful handling and elaborate preprocessing pipelines during assembly 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 coregistered.
- All images must contain minimum obstructions (e.g., clouds, haze, corrupt pixels, and so on).

- The depicted scenes must be as diverse as possible. Especially for STF, the 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 data sets proposed specifically for the task of spatial downscaling, several data sets 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 data sets and their corresponding characteristics.

- UC Merced [240] contains 2,100 aerial RGB images coming from the U.S. Geological Survey National Map Urban Area Imagery depicting 21 different land use classes at 0.3-m resolution from several U.S. 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 0.5 m. Images originate from different regions around the world.
- WHU-RS20 [242] is an extension of the WHU-RS19 data set with an extra land use class and a total of 5,000 aerial RGB images.
- Remote sensing Scene classification (RSSCN7) [243] contains 2,800 aerial RGB images from Google Earth depicting seven land use classes.
- Remote scene classification (RSC11) [244] contains 1,232 aerial RGB images from Google Earth depicting 11 land use classes at 0.2-m spatial resolution. Images come from several U.S. cities.
- The Aerial Image Dataset (AID) [245] contains 10,000 aerial RGB images coming from Google Earth at resolutions ranging from 0.5 to 8 m. 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 resolutions ranging from 0.2 to 30 m. Images come from several different regions around the world.
- RS-IDEA Research Group-WU (SIRI-WHU) [247] contains 2,400 aerial RGB images from Google Earth depicting 12 land use classes at a spatial resolution of 2 m. The images mainly cover urban areas in China.
- The Brazilian coffee scene data set [248] contains 2,876 SPOT images (green, red, and NIR bands) over four regions in Brazil for binary image classification based on the presence or absence of coffee crops.
- Sentinel 1-2 (SEN1-2) [249] contains 282,384 pairs of *Sentinel-1* and *Sentinel-2* RGB images at 10-m spatial resolution from around the world at different seasons.
- Sentinel-1/2 MODIS (SEN12MS) [250] contains 180,662 triplets of *Sentinel-1* dual-polarization SAR, *Sentinel-2* MS, and MODIS land cover images at 10-m spatial resolution coming from all around the globe and at different times.
- Dataset of Object deTection in Aerial images (DOTA) [251] contains 2,806 aerial images from different sensors along with *GaoFen-2* and *Jilin-1* satellite images. This data set is targeted toward object detection and includes labels spanning more than 15 object categories.
- DIOR [33] contains 23,463 aerial RGB images from Google Earth with spatial resolutions ranging from 0.5 to 30 m. The images cover several regions around the globe, and their labels span more than 20 object categories.
- Coleambally irrigation area (CIA) [252] contains 17 Landsat/MODIS pairs from Coleambally Irrigation Area, Australia, at 25-m spatial resolution. Images were obtained during a single summer season but have strong spatial heterogeneity.
- Lower Gwydir Catchment (LGC) [252] contains 14 Landsat/MODIS pairs from LGC, Australia, at 25-m spatial resolution. Images were obtained during a whole year, which also included a major flood. This renders the data set ideal for the study of abrupt and unpredictable changes in time series.
- Ar Horqin Banner (AHB) [253] contains 27 Landsat/MODIS pairs from ARB, China, over a span of five 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 six 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 six years. It is intended for the study of land cover changes.
- The Gaofen Image Data Set [254] contains 150 *Gaofen-2* images (RGB and NIR bands) from many regions in China with 4-m spatial resolution. It is intended for scene classification and land cover segmentation.
- Kelvin's *PROBA-V* SR Data Set [211] contains 1,160 images from the *PROBA-V* satellite (red and NIR bands) from several locations around the globe at different points in time. Each data point contains an HR image of 100-m resolution and several LR images of 300-m resolution.
- Kaggle's Draper Satellite Image Chronology [255] contains 1,720 aerial RGB images from California, United States, over a period of five days.
- Diverse Real-World Image SR [256] contains 31,970 LR image patches including aerial images.
- Pavia Center [118] was acquired by reflective optics system imaging spectrometer (ROSIS) over the city of Pavia, Italy, in the wavelength range of 430 to 860 nm. It contains 115 spectral bands and is of size 1,096 × 1,096.
- Houston [118] was acquired by an ITRES-compact airborne spectrographic imager (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 to 1,050 nm,

and each band contains 349 × 1,905 pixels with a spatial resolution of 2.5 m

- Los Angeles [118] was acquired over a port in the city of Los Angeles by the Hyperion sensor mounted on the *Earth Observing-1* (*EO-1*) satellite. The HS image contains 242 spectral bands with a spatial resolution of 30 m.
- Botswana [257] was acquired over the Okavango Delta in Botswana by the Hyperion sensor mounted on the *EO-1* satellite. The HS image contains 242 spectral bands with a spatial resolution of 30 m.
- Hobart [113], acquired by the *IKONOS* sensor, represents an urban and harbor area of Hobart, Australia. The MS sensor is characterized by four bands (RGB and NIR) and also a PAN channel with band range from 450 to 900 nm. The resolution of MS is 4 m and of PAN is 1 m.
- Sundarbans [113], obtained by the *QuickBird* sensor, represents a forest area of Sundarbans in India. This data set provides an HR PAN image with a spectral cover range from 760 to 850 nm and a resolution of 0.6 m as well as a four-band (RGB and NIR) MS image with a resolution of 2.4 m.
- Washington DC Mall [139] covers an urban area in the Washington, D.C., National Mall. The size of the degraded HS image is 256 × 60 and that of the PAN image is 1,280 × 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 10-m resolution and that of the PAN image is 395 × 185 with 20-m 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] was captured by Headwall's Hyperspec Visible and Near-Infrared, series C imaging sensor over Chikusei, Ibaraki, Japan, on 29 July 2014. The data set contains 128 bands in the spectral range of 363–1,018 nm. The PAN image has 300 × 300 pixels with a spatial resolution of 2.5 m.
- Foster [258] has 33 spectral channels from 400 to 720 nm with 10 nm per band. The original size of each HS image in the Foster data set is 1,341 × 1,022.

ADVANCEMENTS IN COMPUTER VISION

Spatial enhancement, or SR, is being thoroughly investigated in general CV, and a great number of methods have been proposed that build on previous research and expand the state of the art. Hence, in the CV field, some informative review articles 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 that, 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 the literature train models on synthetic data sets 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., *SFTMD* [259] and *DAN* [260], [261]) explore deep networks that 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 SR*.

In some cases, the available data set comprises LR images that need to be downscaled, along with a number of HR reference images of the same domain that, however, do not correspond to the LR data. A family of methods attempts 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 that 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 SR approaches is 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 the image SR 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., ×8 and ×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.

DISCUSSION

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

UNIVERSAL METRICS

An important conclusion of the "Metrics" section is the fact that there exist no established evaluation metrics for downscaling models. To be sure, a limited subset of the metrics presented in Table 1 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 indexes for EO image downscaling contributes to the robustness against the inherent superresolved 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, are able to achieve impressive aesthetic results; however, they are prone to creating hallucinations and/or artifacts. Controlling and quantifying the tradeoff between SR performance vis-à-vis the expected hallucination level remains an open issue. In addition, it may be that a single metric characterizing the overall model performance is not enough, but an additional gridded output with uncertainty estimates should be produced.

Therefore, we consider it critical to develop algorithms that will help both ML practitioners and end users to better understand, interpret, and trust the DL model outputs. explainable artificial intelligence (xAI) algorithms [129] are essential tools toward an enhanced understanding and transparency of the developed DL models, especially for facilitating the operational uptake of EO image downscaling models.

BENCHMARK DATA SETS

The availability and abundance of RS images has greatly facilitated the formulation of data sets that satisfy the needs of complex DL models. Many researchers choose to directly download RS images from the respective providers; perform the preprocessing 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 data sets tailored to the downscaling task, which will help to objectively evaluate and compare different models, thus gaining more concrete insight into their generalization and applicability.

MODEL PERFORMANCE

In addition to the point discussed, 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 the 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 an 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., Massachusetts Institute of Technology (MIT), Berkley Source Distribution (BSD), GNU, and so on). 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 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, the 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 data set, LR images are usually constructed with a single degradation algorithm. Consequently, a model trained on such a data set 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 that are applicable in real-world scenarios during the sensor capture of an image.

MULTIMODAL FUSION

The spectral fusion of images can greatly assist the downscaling process (see the "Spatiospectral Fusion" section). However, 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 SAR and optical data for the purpose of downscaling, which, to our knowledge, has not yet been investigated in the DL field.

GENERATIVE ADVERSARIAL NETWORKS OR ELSE

GANs manage to better approximate the boundary of the perception–distortion plane and achieve more realistic and perceptually convincing results (see the "Metrics" section). Therefore, a further study of the GAN framework is needed to exploit its potential to the full extent. 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 data set 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 data set 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.

COMPUTER VISION PARADIGM

The field of general CV 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 the "Advancements in Computer Vision" section. 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, SR in natural images usually involves a magnification factor much smaller than those in the RS domain (ranging from ×2 to ×4 compared with ×8 to ×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 SYNTHETIC APERTURE RADAR IMAGERY

The techniques proposed in the literature for SAR image enhancement are few, and they compare well-established techniques borrowed from CV 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 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, a number of authors apply SR techniques only 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 and 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 by 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 toward the 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., the C , X , and L bands), imaging modes (e.g., StripMap, wide swath, spotlight, and so on), and the availability of polarimetric data.

CONCLUSION

In this survey, we offer a detailed overview of the methods available in the literature for the spatial downscaling of RS 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 data sets for this task, and we analyze the tradeoff 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 CV community as a source of inspiration.

As seen from our analysis, although there is a strong presence of the DL 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 toward xAI, multimodal data sets, 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 further stimulates the research community and assists in avoiding common pitfalls in the design, development, and assessment of new DL techniques.

ACKNOWLEDGMENTS

This work received funding from the European Union's Horizon2020 research and innovation projects 1) DeepCube, under grant agreement 101004188 (Maria Sdraka and Ioannis Papoutsis); 2) NEANIAS, under grant agreement 863448 (Bill Psomas and Konstantinos Karantzalos); and 3) CALLISTO, under grant agreement 101004152 (Konstantinos Vlachos, Konstantinos Ioannidis, Ilias Gialampoukidis, and Stefanos Vrochidis).

AUTHOR INFORMATION

Maria Sdraka (masdra@noa.gr) received her M.Sc. degree in electrical and computer engineering from the National Technical University of Athens, Greece, in 2016. She is currently working toward a Ph.D. degree from the Institute of

Astronomy, Astrophysics, Space Applications, and Remote Sensing, National Observatory of Athens, Athens, 15236, Greece. Her research interests include the application of artificial intelligence techniques on remote sensing data for earth observation tasks, especially damage assessment of forest wildfires. She has worked on signal processing, data fusion, image enhancement and segmentation as well as change detection through the assistance of deep learning algorithms.

Ioannis Papoutsis (ipapouts@gmail.com) received his diploma in electrical and computer engineering from the National Technical University of Athens, Greece, in 2002, his M.Sc. degree in technologies for broadband communications from the Department of Electronic and Electrical Engineering, University College London, London, U.K., in 2003, and his Ph.D. degree in remote sensing from the National Technical University of Athens in 2014. In 2019, he was elected associate researcher with the Institute of Astronomy, Astrophysics, Space Applications, and Remote Sensing, National Observatory of Athens, Athens, 15236, Greece, where he leads OrionLab, a research unit of artificial intelligence for big earth observation data (ESA). He has been the Operations Manager with the Greek node of European Space Agency Hubs that distribute Sentinel data. He has also acted as the Copernicus Emergency Management Services Manager for Risk and Recovery activations. He has participated and coordinated several research projects funded by the European Commission and ESA. His research interests include the exploitation, management and processing of big satellite data, and machine learning for knowledge extraction and fusion of multimodal EO data. He is a Member of IEEE.

Bill Psomas (psomasbill@mail.ntua.gr) received his integrated master in rural, surveying and geoinformatics engineering from the National Technical University of Athens, Greece, in 2018. He continued his studies with a M.Sc. degree in data science and information technologies, specializing in big data and artificial intelligence at National and Kapodistrian University of Athens, Greece, where he graduated in 2020. He is currently a Ph.D. student at the National Technical University of Athens, Athens, 15780, Greece working on representation learning. Previously, he worked at Inria Rennes Bretagne-Atlantique, France and Athena Research Center, Greece. His research interests lie in the intersection of deep learning with computer vision. He has worked on metric learning, self-supervised learning, and continual learning.

Konstantinos Vlachos (kostasvlachosgrs@iti.gr) received his B.Sc. degree in geology from the University of Patras, Greece, in 2015, where he specialized in quantitative spatial analysis. In 2019, he received his M.Sc. and engineering degrees in applied earth sciences from the Geoscience and Remote Sensing Department at Delft University of Technology, The Netherlands, where he specialized in the fusion of multi-sensor satellite data using machine/deep learning for sea level estimation—funded by Deltares, an Institute for Applied Research, The Netherlands. Since 2021, he has been a research associate at the Information Technologies Institute, Center for Research and Technology Hellas, Thessaloniki, Thessaloniki, 57001, Greece, and a member of the Multimodal Data Fusion and Analytics (M4D) Group of the Multimedia Knowledge and Social Media Analytics Lab. His current research interests lie in the interdisciplinary domains of Earth science, Earth observation and artificial intelligence focusing on spatiotemporal analysis and fusion for downscaling and change detection.

Konstantinos Ioannidis (kioannid@iti.gr) received his diploma and Ph.D. degrees from the Department of Electrical and Computer Engineering, Democritus University of Thrace, Greece, in 2006 and 2013, respectively. Currently, he is a senior researcher at the Information Technologies Institute, Center for Research and Technology Hellas, Thessaloniki, Thessaloniki, 57001, Greece. He is a member of the Multimodal Data Fusion and Analytics (M4D) Group of the Multimedia Knowledge and Social Media Analytics Lab. His research interests mainly include the areas of path planning, collective behavior in swarm robotics, autonomous navigation and formation control as well as a variety of computer vision techniques indicatively, object detection, 3D representation, aerial imagery, image enhancement, aerial imagery, photogrammetry, SLAM and many others using both learning-based (machine and deep learning) models and fundamental approaches.

Konstantinos Karantzalos (karank@central.ntua.gr) received his diploma degree in engineering from the National Technical University of Athens, Greece, in 2000, and his Ph.D. degree from the National Technical University of Athens in collaboration with Ecole Nationale de Ponts et Chaussees, Champs-sur-Marne, France, in 2007. In 2007, he joined the Department of Applied Mathematics, Ecole Centrale de Paris, Gifsur-Yvette, France as a postdoc. He is an associate professor of remote sensing with the National Technical University of Athens, Athens, 15780, Greece. His teaching and research interests include geoscience and earth observation, geospatial data analytics, spectral data analysis, and machine learning with applications in, e.g., environmental monitoring and precision agriculture. He has several publications in top-rank international journals and conferences and a number of awards and honors for his research contributions. He serves on the board of directors of the Greek Space Center. He is a Senior Member of IEEE.

Ilias Gialampoukidis (heliasgj@iti.gr) received his bachelor's degree in mathematics and his M.Sc. degree in statistics and modeling from the Aristotle University of Thessaloniki, Greece. He also received a Ph.D. degree in mathematics, with a special interest in applied mathematics, time series analysis, stochastic modelling, and network analytics. He is a senior postdoctoral researcher at the Information Technologies Institute, Center for Research and Technology Hellas, Thessaloniki, Thessaloniki, 57001, Greece. He has extensive experience in EC-funded research projects through work package leaderships and critical roles in several projects. His research interests involve multimodal information retrieval, Earth observation, big data analytics, multimodal fusion, supervised (deep) and unsupervised learning, and social media mining and network analytics. He has coauthored more than 60 publications in international journals and conferences.

Stefanos Vrochidis (stefanos@iti.gr) received his diploma degree in electrical engineering from Aristotle University of Thessaloniki, Greece, his M.Sc. degree in radio frequency communication systems from the University of Southampton, and his Ph.D. degree in electronic engineering from Queen Mary University of London, U.K. Currently, he is a senior researcher (grade C) at the Information Technologies Institute, Center for Research and Technology Hellas, Thessaloniki, Thessaloniki, 57001, Greece, and the head of the Multimodal Data Fusion and Analytics (M4D) Group of the Multimedia Knowledge and Social Media Analytics Lab. His research interests include multimedia analysis and retrieval, multimodal fusion, computer vision, multimodal analytics, and artificial intelligence, as well as media and arts, and environmental and security applications. He has participated in more than 50 European and National projects (in more than 15 as project coordinator, scientific or technical manager) and has been member of the organization team of several conferences and workshops relevant to the aforementioned research areas. He has edited three books and authored more than 250 related scientific journal, conference and book chapter publications.

REFERENCES

[1] P. Ghamisi *et al.*, "Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art," *IEEE Geosci. Remote Sens. Mag. (replaces Newslett.)*, vol. 7, no. 1, pp. 6–39, Mar. 2019, doi: 10.1109/MGRS.2018.2890023.

[2] B. Chen, J. Li, and Y. Jin, "Deep learning for feature-level data fusion: Higher resolution reconstruction of historical landsat archive," *Remote Sens.*, vol. 13, no. 2, p. 167, Jan. 2021, doi: 10.3390/rs13020167.

[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," *Int. J. Remote Sens.*, vol. 39, no. 4, pp. 1042–1067, Feb. 2018, doi: 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 Trans. Geosci. Remote Sens.*, vol. 55, no. 1, pp. 600–614, Jan. 2017, doi: 10.1109/TGRS.2016.2613140.

[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 Sens.*, vol. 9, no. 4, p. 381, Apr. 2017, doi: 10.3390/rs9040381.

[6] A. Y. Sun and G. Tang, "Downscaling satellite and reanalysis precipitation products using attention-based deep convolutional neural nets," *Front. Water*, vol. 2, p. 536,743, Nov. 2020, doi: 10.3389/frwa.2020.536743.

[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 – Int. Arch. Photogram., Remote Sens. Spatial Inf. Sci.*, vol. XLIII-B1-2020, pp. 573–578, Aug. 2020, doi: 10.5194/isprs-archives-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 J. Photogram. Remote Sens.*, vol. 160, pp. 260–274, Feb. 2020, doi: 10.1016/j.isprsjprs.2019.12.014.

[9] D. Garcia *et al.*, "Pix2Streams: Dynamic hydrology maps from satellite-LiDAR fusion," Nov. 2020, *arXiv: 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 Trans. Multimedia*, vol. 21, no. 12, pp. 3106–3121, Dec. 2019, doi: 10.1109/TMM.2019.2919431.

[11] J. J. Danker Khoo, K. H. Lim, and J. T. Sien Phang, "A review on deep learning super resolution techniques," in *Proc. IEEE 8th Conf. Syst., Process Control (ICSPC)*, Dec. 2020, pp. 134–139, doi: 10.1109/ICSPC50992.2020.9305806.

[12] H. Chen, X. He, L. Qing, Y. Wu, C. Ren, and C. Zhu, "Real-world single image super-resolution: A brief review," Mar. 2021. [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," Feb. 2021, *arXiv: 2102.09351*.

[14] K. Nasrollahi and T. B. Moeslund, "Super-resolution: A comprehensive survey," *Mach. Vis. Appl.*, vol. 25, no. 6, pp. 1423–1468, Aug. 2014, doi: 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, doi: 10.3390/math8111907.

[16] Z. Wang, J. Chen, and S. C. H. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3365–3387, Oct. 2021, doi: 10.1109/TPAMI.2020.2982166.

[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 J. Photogram. Remote Sens.*, vol. 171, pp. 101–117, Jan. 2021, doi: 10.1016/j.isprsjprs.2020.11.001.

[18] G. Kaur, K. S. Saini, D. Singh, and M. Kaur, "A comprehensive study on computational pansharpening techniques for remote sensing images," *Arch. Comput. Methods Eng.*, vol. 28, no. 7, Feb. 2021, doi: 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," *Inf. Fusion*, vol. 46, pp. 102–113, Mar. 2019, doi: 10.1016/j.inffus.2018.05.006.

[20] R. Fernandez-Beltran, P. Latorre-Carmona, and F. Pla, "Single-frame super-resolution in remote sensing: A practical

overview," *Int. J. Remote Sens.*, vol. 38, no. 1, pp. 314–354, Jan. 2017, doi: 10.1080/01431161.2016.1264027.

[21] Q. Yuan *et al.*, "Deep learning in environmental remote sensing: Achievements and challenges," *Remote Sens. Environ.*, vol. 241, p. 111,716, May 2020, doi: 10.1016/j.rse.2020.111716.

[22] X. X. Zhu *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag. (replaces Newslett.)*, vol. 5, no. 4, pp. 8–36, Dec. 2017, doi: 10.1109/MGRS.2017.2762307.

[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 J. Photogram. Remote Sens.*, vol. 152, pp. 166–177, Jun. 2019, doi: 10.1016/j.isprsjprs.2019.04.015.

[24] X. Zhu, F. Cai, J. Tian, and T. Williams, "Spatiotemporal fusion of multisource remote sensing data: Literature survey, taxonomy, principles, applications, and future directions," *Remote Sens.*, vol. 10, no. 4, p. 527, Mar. 2018, doi: 10.3390/rs10040527.

[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, doi: 10.3390/s19183929.

[26] "Web of science." Accessed: Sep. 9, 2021. [Online]. Available: https://webofknowledge.com

[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, nos. 1–3, pp. 189–216, Jan. 2004, doi: 10.1023/B:CLIM.0000013685.99609.9e.

[28] P. M. Atkinson, "Downscaling in remote sensing," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 22, pp. 106–114, Jun. 2013, doi: 10.1016/j.jag.2012.04.012.

[29] W. Sun and Z. Chen, "Learned image downscaling for upscaling using content adaptive resampler," *IEEE Trans. Image Process.*, vol. 29, pp. 4027–4040, Feb. 2020, doi: 10.1109/TIP.2020.2970248.

[30] W. Zhan *et al.*, "Disaggregation of remotely sensed land surface temperature: Literature survey, taxonomy, issues, and caveats," *Remote Sens. Environ.*, vol. 131, pp. 119–139, Apr. 2013, doi: 10.1016/j.rse.2012.12.014.

[31] Y. Li, H. Zhang, X. Xue, Y. Jiang, and Q. Shen, "Deep learning for remote sensing image classification: A survey," *Wiley Interdisciplinary Rev.: Data Mining 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," Jun. 2020, *arXiv: 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 J. Photogram. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020, doi: 10.1016/j.isprsjprs.2019.11.023.

[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 Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4834–4843, Jul. 2019, doi: 10.1109/TGRS.2019.2893310.

[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 Sens.*, vol. 9, no. 6, p. 586, Jun. 2017, doi: 10.3390/rs9060586.

[36] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogram. Eng. Remote Sens.*, vol. 63, no. 6, pp. 691–699, 1997. [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 Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: 10.1109/TIP.2003.819861.

[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 *Proc. Data Compress. Conf. (DCC)*, Snowbird, UT, USA, Mar. 2019, pp. 172–181, doi: 10.1109/DCC.2019.00025.

[39] Z. Wang and A. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002, doi: 10.1109/97.995823.

[40] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, 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 Trans. Image Process.*, vol. 14, no. 12, pp. 2117–2128, Dec. 2005, doi: 10.1109/TIP.2005.859389.

[42] H. Sheikh and A. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006, doi: 10.1109/TIP.2005.859378.

[43] N. Damera-Venkata, T. Kite, W. Geisler, B. Evans, and A. Bovik, "Image quality assessment based on a degradation model," *IEEE Trans. Image Process.*, vol. 9, no. 4, pp. 636–650, Apr. 2000, doi: 10.1109/83.841940.

[44] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011, doi: 10.1109/TIP.2011.2109730.

[45] A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1500–1512, Apr. 2012, doi: 10.1109/TIP.2011.2175935.

[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 *Proc. 3rd Annu. JPL Airborne Earth Sci. 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?" in *Proc. 3rd Conf. Fusion Earth Data:*

Merging Point Meas., Raster Maps Remotely Sensed Images, Jan. 2000, pp. 99–103.

[48] D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imaging*, vol. 19, no. 1, p. 11,006, Jan. 2010, doi: 10.1117/1.3267105.

[49] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," Mar. 2016, *arXiv: 1603.08155*.

[50] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012, doi: 10.1109/TIP.2012.2214050.

[51] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a completely blind image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013, doi: 10.1109/LSP.2012.2227726.

[52] N. Venkatanath, D. Praneeth, B. Maruthi Chandrasekhar, S. S. Channappayya, and S. S. Medasani, "Blind image quality evaluation using perception-based features," in *Proc. 21st Nat. Conf. Commun. (NCC)*, Feb. 2015, pp. 1–6, doi: 10.1109/NCC.2015.7084843.

[53] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang, "Learning a no-reference quality metric for single-image super-resolution," Dec. 2016, *arXiv:1612.05890*.

[54] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, "The 2018 PIRM challenge on perceptual image super-resolution," Jan. 2019, *arXiv: 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," Apr. 2018, *arXiv: 1801.03924*.

[56] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, and M. Selva, "Multispectral and panchromatic data fusion assessment without reference," *Photogram. Eng. Remote Sens.*, vol. 74, no. 2, pp. 193–200, Feb. 2008, doi: 10.14358/PERS.74.2.193.

[57] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, *arXiv: 1711.06077*.

[58] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst. – vol. 2 (NIPS' 14)*. Cambridge, MA, USA: MIT Press, 2014, pp. 2672–2680.

[59] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2528–2535.

[60] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, 2016, doi: 10.23915/distill.00003.

[61] W. Shi *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," Sep. 2016, *arXiv: 1609.05158*.

[62] W. Shi *et al.*, "Is the deconvolution layer the same as a convolutional layer?" Sep. 2016, *arXiv: 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 Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017, doi: 10.1109/TIP.2017.2662206.

[64] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Dec. 2015, *arXiv: 1512.03385*.

[65] P. Burt and E. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. 31, no. 4, pp. 532–540, Apr. 1983, doi: 10.1109/TCOM.1983.1095851.

[66] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," May 2019, *arXiv: 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," Apr. 2020, *arXiv: 1910.03151*.

[68] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," Mar. 2021, *arXiv: 2103.02907*.

[69] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "BAM: Bottleneck attention module," Jul. 2018, *arXiv: 1807.06514*.

[70] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," Jul. 2018, *arXiv: 1807.06521*.

[71] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, "Rotate to attend: Convolutional triplet attention module," Nov. 2020, *arXiv: 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, doi: 10.3390/electronics10101187.

[73] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," Jul. 2015, *arXiv: 1501.00092*.

[74] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," Nov. 2016, *arXiv: 1511.04587*.

[75] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Apr. 2015, *arXiv: 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," Oct. 2017, *arXiv: 1704.03915*.

[77] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," May 2017, *arXiv: 1609.04802*.

[78] X. Wang *et al.*, "ESRGAN: Enhanced super-resolution generative adversarial networks," Sep. 2018, *arXiv: 1809.00219*.

[79] A. Jolicoeur-Martineau, "The relativistic discriminator: A key element missing from standard GAN," Sep. 2018, *arXiv: 1807.00734*.

[80] Xintao, "xinntao/ESRGAN," Aug. 31, 2018. [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," Jul. 2017, *arXiv: 1707.02921*.

[82] J. Yu *et al.*, "Wide activation for efficient and accurate image super-resolution," Dec. 2018, *arXiv: 1808.08718*.

[83] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," Mar. 2018, *arXiv: 1802.08797*.

[84] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2790–2798, doi: 10.1109/CVPR.2017.298.

[85] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," Mar. 2018, *arXiv: 1803.02735*.

[86] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback network for image super-resolution," Jun. 2019, *arXiv: 1903.09814*.

[87] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," Nov. 2016, *arXiv: 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 Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2599–2613, Nov. 2019, doi: 10.1109/TPAMI.2018.2865304.

[89] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," Jul. 2018, *arXiv: 1807.02758*.

[90] D. Lei, H. Chen, L. Zhang, and W. Li, "NLRnet: An efficient nonlocal attention ResNet for pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, Mar. 2021, doi: 10.1109/TGRS.2021.3067097.

[91] C. Shang *et al.*, "Spatiotemporal reflectance fusion using a generative adversarial network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, Mar. 2021, doi: 10.1109/TGRS.2021.3065418.

[92] Y. Yu, X. Li, and F. Liu, "E-DBPN: Enhanced deep back-projection networks for remote sensing scene image superresolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5503–5515, Aug. 2020, doi: 10.1109/TGRS.2020.2966669.

[93] "Sentinel-2 – Overview." 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 J. Photogram. Remote Sens.*, vol. 146, pp. 305–319, Dec. 2018, doi: 10.1016/j.isprsjprs.2018.09.018.

[95] F. Palsson, J. Sveinsson, and M. Ulfarsson, "Sentinel-2 image fusion using a deep residual network," *Remote Sens.*, vol. 10, no. 8, p. 1290, Aug. 2018, doi: 10.3390/rs10081290.

[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 Sens.*, vol. 11, no. 22, p. 2635, Nov. 2019, doi: 10.3390/rs11222635.

[97] J. Wu, Z. He, and J. Hu, "Sentinel-2 sharpening via parallel residual network," *Remote Sens.*, vol. 12, no. 2, p. 279, Jan. 2020, doi: 10.3390/rs12020279.

[98] X. Luo, X. Tong, and Z. Hu, "Improving satellite image fusion via generative adversarial training," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 1–14, 2020, doi: 10.1109/TGRS.2020.3025821.

[99] C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," Apr. 2016, *arXiv: 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 J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6882–6896, Jun. 2021, doi: 10.1109/JSTARS.2021.3092286.

[101] M. Ciotola, M. Ragosta, G. Poggi, and G. Scarpa, "A full-resolution training framework for Sentinel-2 image fusion," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.(IGARSS)*, Jul. 2021, pp. 1260–1263, doi: 10.1109/IGARSS47720.2021.9553199.

[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 Sens. Environ.*, vol. 235, p. 111,425, Dec. 2019, doi: 10.1016/j.rse.2019.111425.

[103] "Landsat 8," NASA, Washington, DC, USA. [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 Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, Jan. 2021, doi: 10.1109/TGRS.2020.3046045.

[105] J. Dai *et al.*, "Deformable convolutional networks," Jun. 2017, *arXiv: 1703.06211*.

[106] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sens.*, vol. 8, no. 7, p, 594, 2016, doi: 10.3390/rs8070594.

[107] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, 2010, doi: 10.1109/TIP.2010.2050625.

[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 Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1795–1799, Oct. 2017, doi: 10.1109/LGRS.2017.2736020.

[109] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, "PanNet: A deep network architecture for pan-sharpening," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1753–1761, doi: 10.1109/ICCV.2017.193.

[110] G. Scarpa, S. Vitale, and D. Cozzolino, "Target-adaptive CNN-based pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5443–5457, 2018, doi: 10.1109/TGRS.2018.2817393.

[111] Y. Xing, M. Wang, S. Yang, and L. Jiao, "Pan-sharpening via deep metric learning," *ISPRS J. Photogram. Remote Sens.*, vol. 145, pp. 165–183, Nov. 2018, doi: 10.1016/j.isprsjprs.2018.01.016.

[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 J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 3, pp. 978–989, 2018, doi: 10.1109/JSTARS.2018.2794888.

[113] L. He *et al.*, "Pansharpening via detail injection based convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 4, pp. 1188–1204, 2019, doi: 10.1109/JSTARS.2019.2898574.

[114] S. Luo, S. Zhou, Y. Feng, and J. Xie, "Pansharpening via unsupervised convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4295–4310, Jul. 2020, doi: 10.1109/JSTARS.2020.3008047.

[115] L. Liu *et al.*, "Shallow–deep convolutional network and spectral-discrimination-based detail injection for multispectral imagery pan-sharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1772–1783, Mar. 2020, doi: 10.1109/JSTARS.2020.2981695.

[116] L.-J. Deng, G. Vivone, C. Jin, and J. Chanussot, "Detail injection-based deep convolutional neural networks for pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 1–16, 2020, doi: 10.1109/TGRS.2020.3031366.

[117] J. Cai and B. Huang, "Super-resolution-guided progressive pansharpening based on a deep convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5206–5220, 2021, doi: 10.1109/TGRS.2020.3015878.

[118] W. Dong, T. Zhang, J. Qu, S. Xiao, J. Liang, and Y. Li, "Laplacian pyramid dense network for hyperspectral pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, May 2021, doi: 10.1109/TGRS.2021.3076768.

[119] M. Jiang, H. Shen, J. Li, Q. Yuan, and L. Zhang, "A differential information residual convolutional neural network for pansharpening," *ISPRS J. Photogram. Remote Sens.*, vol. 163, pp. 257–271, May 2020, doi: 10.1016/j.isprsjprs.2020.03.006.

[120] Y. Qu, R. K. Baghbaderani, H. Qi, and C. Kwan, "Unsupervised pansharpening based on self-attention mechanism," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3192–3208, 2021, doi: 10.1109/TGRS.2020.3009207.

[121] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS 2017)*, 2017.

[122] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 11,057–11,066, doi: 10.1109/CVPR.2019.01132.

[123] H. Zhang and J. Ma, "GTP-PNET: A residual learning network based on gradient transformation prior for pansharpening," *ISPRS J. Photogram. Remote Sens.*, vol. 172, pp. 223–239, 2021, doi: 10.1016/j.isprsjprs.2020.12.014.

[124] H. Yin, "PSCSC-Net: A deep coupled convolutional sparse coding network for pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, Jun. 2021, doi: 10.1109/TGRS.2021.3088313.

[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 Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, Mar. 2021, doi: 10.1109/TGRS.2021.3066425.

[126] L. Zhang, J. Zhang, J. Ma, and X. Jia, "SC-PNN: Saliency cascade convolutional neural network for pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 1–19, 2021, doi: 10.1109/TGRS.2021.3054641.

[127] I. Selesnick, R. Baraniuk, and N. Kingsbury, "The dual-tree complex wavelet transform," *IEEE Signal Process. Mag.*, vol. 22, no. 6, pp. 123–151, 2005, doi: 10.1109/MSP.2005.1550194.

[128] S. Vitale and G. Scarpa, "A detail-preserving cross-scale learning strategy for CNN-based pansharpening," *Remote Sens.*, vol. 12, no. 3, p. 348, 2020, doi: 10.3390/rs12030348.

[129] A. Barredo Arrieta *et al.*, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

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

[131] X. Liu, Y. Wang, and Q. Liu, "PSGAN: A generative adversarial network for remote sensing image pan-sharpening," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, 2018, pp. 873–877, doi: 10.1109/ICIP.2018.8451049.

[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," *Inf. Fusion*, vol. 62, pp. 110–120, Oct. 2020, doi: 10.1016/j.inffus.2020.04.006.

[133] A. Gastineau, J.-F. Aujol, Y. Berthoumieu, and C. Germain, "Generative adversarial network for pansharpening with spectral and spatial discriminators," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, Mar. 2021, doi: 10.1109/TGRS.2021.3060958.

[134] F. Ozcelik, U. Alganci, E. Sertel, and G. Unal, "Rethinking CNN-based pansharpening: Guided colorization of panchromatic images via GANs," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, 2020, doi: 10.1109/TGRS.2020.3010441.

[135] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "Multispectral and hyperspectral image fusion using a 3-d-convolutional neural network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 639–643, May 2017, doi: 10.1109/LGRS.2017.2668299.

[136] R. Dian, S. Li, A. Guo, and L. Fang, "Deep hyperspectral image sharpening," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5345–5355, 2018, doi: 10.1109/TNNLS.2018.2798162.

[137] F. Zhou, R. Hang, Q. Liu, and X. Yuan, "Pyramid fully convolutional network for hyperspectral and multispectral image fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 5, pp. 1549–1558, May 2019, doi: 10.1109/JSTARS.2019.2910990.

[138] X. Han, J. Yu, J. Luo, and W. Sun, "Hyperspectral and multispectral image fusion using cluster-based multi-branch BP neural networks," *Remote Sens.*, vol. 11, no. 10, p. 1173, Jan. 2019, doi: 10.3390/rs11101173.

[139] L. He *et al.*, "HyperPNN: Hyperspectral pansharpening via spectrally predictive convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 3092–3100, 2019, doi: 10.1109/JSTARS.2019.2917584.

[140] K. Li, W. Xie, Q. Du, and Y. Li, "DDLPS: Detail-based deep laplacian pansharpening for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 8011–8025, 2019, doi: 10.1109/TGRS.2019.2917759.

[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 J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4095–4110, Jul. 2020, doi: 10.1109/JSTARS.2020.3009250.

[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 Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1457–1473, 2020, doi: 10.1109/TPAMI.2020.3045010.

[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 J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 7373–7385, Jul. 2021, doi: 10.1109/JSTARS.2021.3097178.

[144] X. Zhang, W. Huang, Q. Wang, and X. Li, "SSR-net: Spatial–spectral reconstruction network for hyperspectral and

multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5953–5965, Jul. 2021, doi: 10.1109/TGRS.2020.3018732.

[145] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, Jan. 2009, doi: 10.1561/2200000006.

[146] "MODIS technical specifications," NASA, Washington, DC, USA. Accessed: Jul. 8, 2021. [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 Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6552–6564, Sep. 2019, doi: 10.1109/TGRS.2019.2907310.

[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 J.*, vol. 20, no. 20, pp. 12,190–12,202, Oct. 2020, doi: 10.1109/JSEN.2020.3000249.

[149] W. Li, X. Zhang, Y. Peng, and M. Dong, "Spatiotemporal fusion of remote sensing images using a convolutional neural network with attention and multiscale mechanisms," *Int. J. Remote Sens.*, vol. 42, no. 6, pp. 1973–1993, Mar. 2021, doi: 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 Sens.*, vol. 12, no. 4, p. 698, Feb. 2020, doi: 10.3390/rs12040698.

[151] S. Yang and X. Wang, "Sparse representation and SRCNN based spatio-temporal information fusion method of multi-sensor remote sensing data," *J. Network Intell.*, 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 Sens.*, vol. 12, no. 23, p. 3888, Nov. 2020, doi: 10.3390/rs12233888.

[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," *Sci. China Inf. Sci.*, vol. 63, no. 4, p. 140,302, Apr. 2020, doi: 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 J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 3, pp. 821–829, Mar. 2018, doi: 10.1109/JSTARS.2018.2797894.

[155] J. Chen, L. Wang, R. Feng, P. Liu, W. Han, and X. Chen, "CycleGAN-STF: Spatiotemporal fusion via CycleGAN-based image generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 1–15, 2020, doi: 10.1109/TGRS.2020.3023432.

[156] H. Zhang, Y. Song, C. Han, and L. Zhang, "Remote sensing image spatiotemporal fusion using a generative adversarial network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 1–14, 2021, doi: 10.1109/TGRS.2020.3010530.

[157] X. Wang and X. Wang, "Spatiotemporal fusion of remote sensing image based on deep learning," *J. Sensors*, vol. 2020, pp. 1–11, Jun. 2020, doi: 10.1155/2020/8873079.

[158] Y. Zheng, H. Song, L. Sun, Z. Wu, and B. Jeon, "Spatiotemporal fusion of satellite images via very deep convolutional networks," *Remote Sens.*, vol. 11, no. 22, p. 2701, Nov. 2019, doi: 10.3390/rs11222701.

[159] Z. Tan, P. Yue, L. Di, and J. Tang, "Deriving high spatiotemporal remote sensing images using deep convolutional network," *Remote Sens.*, vol. 10, no. 7, p. 1066, Jul. 2018, doi: 10.3390/rs10071066.

[160] F. Gao, J. Masek, M. Schwaller, and F. Hall, "On the blending of the Landsat and MODIS surface reflectance: Predicting daily Landsat surface reflectance," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 8, pp. 2207–2218, Aug. 2006, doi: 10.1109/TGRS.2006.872081.

[161] Z. Tan, L. Di, M. Zhang, L. Guo, and M. Gao, "An enhanced deep convolutional model for spatiotemporal image fusion," *Remote Sens.*, vol. 11, no. 24, p. 2898, Dec. 2019, doi: 10.3390/rs11242898.

[162] S. Bouabid, M. Chernetskiy, M. Rischard, and J. Gamper, "Predicting landsat reflectance with deep generative fusion," Nov. 2020, *arXiv: 2011.04762*.

[163] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," Nov. 2018, *arXiv: 1611.07004*.

[164] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," Aug. 2020, *arXiv: 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 Sens. Environ.*, vol. 172, pp. 165–177, Jan. 2016, doi: 10.1016/j.rse.2015.11.016.

[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 Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, Jan. 2021, doi: 10.1109/TGRS.2021.3050551.

[167] P. Luo, J. Ren, Z. Peng, R. Zhang, and J. Li, "Differentiable learning-to-normalize via switchable normalization," Apr. 2019, *arXiv: 1806.10779*.

[168] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," Feb. 2018, *arXiv: 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 Sens.*, vol. 13, no. 4, p. 606, Feb. 2021, doi: 10.3390/rs13040606.

[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 Sens.*, vol. 13, no. 4, p. 645, Feb. 2021, doi: 10.3390/rs13040645.

[171] S. Lei, Z. Shi, and Z. Zou, "Super-resolution for remote sensing images via local–global combined network," *IEEE Geosci.*

Remote Sens. Lett., vol. 14, no. 8, pp. 1243–1247, Aug. 2017, doi: 10.1109/LGRS.2017.2704122.

[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 Geosci. Remote Sens. Lett.*, vol. 16, no. 9, pp. 1432–1436, Sep. 2019, doi: 10.1109/LGRS.2019.2899576.

[173] T. Lu, J. Wang, Y. Zhang, Z. Wang, and J. Jiang, "Satellite image super-resolution via multi-scale residual deep neural network," *Remote Sens.*, vol. 11, no. 13, p. 1588, Jul. 2019, doi: 10.3390/rs11131588.

[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 (*Lecture Notes in Computer Science*)*, Z. Lin *et al.*, Eds. Cham: Springer International Publishing, 2019, vol. 11859, pp. 332–340. [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, doi: 10.3390/s21051743.

[176] M. Qin *et al.*, "Remote sensing single-image resolution improvement using a deep gradient-aware network with image-specific enhancement," *Remote Sens.*, vol. 12, no. 5, p. 758, Feb. 2020, doi: 10.3390/rs12050758.

[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 Ann. Photogram., Remote Sens. Spatial Inf. Sci.*, vol. V-1-2020, pp. 9–16, Aug. 2020, doi: 10.5194/isprs-annals-V-1-2020-9-2020.

[178] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2414–2423.

[179] D. Pouliot, R. Latifovic, J. Pasher, and J. Duffe, "Landsat super-resolution enhancement using convolution neural networks and Sentinel-2 for training," *Remote Sens.*, vol. 10, no. 3, p. 394, Mar. 2018, doi: 10.3390/rs10030394.

[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 *Proc. 1st Workshop on Artif. Intell. Deep Learning Geographic Knowledge Discovery*. Los Angeles, CA, USA: ACM, Nov. 2017, pp. 37–44, 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 J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1937–1945, May 2020, doi: 10.1109/JSTARS.2020.2984589.

[182] K. Turkowski, "Filters for common resampling tasks," in *Graphics Gems*. Amsterdam, The Netherlands: Elsevier, 1990, pp. 147–165.

[183] N. Zhang *et al.*, "A multi-degradation aided method for unsupervised remote sensing image super resolution with convolution neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, Dec. 2020, doi: 10.1109/TGRS.2020.3042460.

[184] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3262–3271, doi: 10.1109/CVPR.2018.00344.

[185] W. Ma, Z. Pan, J. Guo, and B. Lei, "Achieving super-resolution remote sensing images via the wavelet transform combined with the recursive ResNet," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3512–3527, Jun. 2019, doi: 10.1109/TGRS.2018.2885506.

[186] Q. Qin, J. Dou, and Z. Tu, "Deep ResNet based remote sensing image super-resolution reconstruction in discrete wavelet domain," *Pattern Recognit. Image Anal.*, vol. 30, no. 3, pp. 541–550, Jul. 2020, doi: 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 Sens.*, vol. 13, no. 9, p. 1858, May 2021, doi: 10.3390/rs13091858.

[188] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5188–5196, doi: 10.1109/CVPR.2015.7299155.

[189] X. Dong, Z. Xi, X. Sun, and L. Gao, "Transferred multi-perception attention networks for remote sensing image super-resolution," *Remote Sens.*, vol. 11, no. 23, p. 2857, Dec. 2019, doi: 10.3390/rs11232857.

[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 Sens.*, vol. 11, no. 15, p. 1817, Aug. 2019, doi: 10.3390/rs11151817.

[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 Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9277–9289, Nov. 2019, doi: 10.1109/TGRS.2019.2924818.

[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 Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4764–4779, Jul. 2020, doi: 10.1109/TGRS.2020.2966805.

[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 Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1618–1633, Feb. 2021, doi: 10.1109/TGRS.2020.2994253.

[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, doi: 10.3390/s20041142.

[195] P. Lei and C. Liu, "Inception residual attention network for remote sensing image super-resolution," *Int. J. Remote Sens.*, vol. 41, no. 24, pp. 9565–9587, Dec. 2020, doi: 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, Sep. 2020, doi: 10.1109/ACCESS.2020.3022882.

[197] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7794–7803, doi: 10.1109/CVPR.2018.00813.

[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 Process.*, vol. 15, no. 5, pp. 1179–1188, Apr. 2021, doi: 10.1049/ipr2.12096.

[199] S. Lei and Z. Shi, "Hybrid-scale self-similarity exploitation for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–10, Apr. 2021, doi: 10.1109/TGRS.2021.3069889.

[200] Y. Chang and B. Luo, "Bidirectional convolutional LSTM neural network for remote sensing image super-resolution," *Remote Sens.*, vol. 11, no. 20, p. 2333, Oct. 2019, doi: 10.3390/rs11202333.

[201] S. Lei, Z. Shi, and Z. Zou, "Coupled adversarial training for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3633–3643, May 2020, doi: 10.1109/TGRS.2019.2959020.

[202] W. Ma, Z. Pan, F. Yuan, and B. Lei, "Super-resolution of remote sensing images via a dense residual generative adversarial network," *Remote Sens.*, vol. 11, no. 21, p. 2578, Nov. 2019, doi: 10.3390/rs11212578.

[203] L. Salgueiro Romero, J. Marcello, and V. Vilaplana, "Super-resolution of Sentinel-2 imagery using generative adversarial networks," *Remote Sens.*, vol. 12, no. 15, p. 2424, Jul. 2020, doi: 10.3390/rs12152424.

[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, doi: 10.1016/j.neucom.2019.03.106.

[205] C. Shin, S. Kim, and Y. Kim, "Satellite image target super-resolution with adversarial shape discriminator," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2020, doi: 10.1109/LGRS.2020.3042238.

[206] Y. Gong *et al.*, "Enlighten-GAN for super resolution reconstruction in mid-resolution remote sensing images," *Remote Sens.*, vol. 13, no. 6, p. 1104, Mar. 2021, doi: 10.3390/rs13061104.

[207] K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, and J. Jiang, "Edge-enhanced GAN for remote sensing image superresolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5799–5812, Aug. 2019, doi: 10.1109/TGRS.2019.2902431.

[208] Y. Li *et al.*, "Single-image super-resolution for remote sensing images using a deep generative adversarial network with local and global attention mechanisms," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–24, Jul. 2021, doi: 10.1109/TGRS.2021.3093043.

[209] M. Kawulok, P. Benecki, S. Piechaczek, K. Hrynczenko, D. Kostrzewa, and J. Nalepa, "Deep learning for multiple-image super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 6, pp. 1062–1066, Jun. 2020, doi: 10.1109/LGRS.2019.2940483.

[210] M. Kawulok, P. Benecki, D. Kostrzewa, and L. Skonieczny, "Towards evolutionary super-resolution," in *Applications of Evolutionary Computation (*Lecture Notes in Computer Science*)*, K. Sim and P. Kaufmann, Eds. Cham: Springer International Publishing, 2018, vol. 10784, pp. 480–496, doi: 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, doi: 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 Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3644–3656, May 2020, *arXiv: 1907.06490*, doi: 10.1109/TGRS.2019.2959248.

[213] A. B. Molini, D. Valsesia, G. Fracastoro, and E. Magli, "Deepsum++: Non-local deep neural network for super-resolution of unregistered multitemporal images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.* Waikoloa, HI, USA, Sep. 2020, pp. 609–612, doi: 10.1109/IGARSS39084.2020.9324418.

[214] M. Deudon *et al.*, "HighRes-Net: Recursive fusion for multi-frame super-resolution of satellite imagery," Feb. 2020, *arXiv: 2002.06460*.

[215] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Deep image homography estimation," Jun. 2016, *arXiv: 1606.03798*.

[216] M. Rifat Arefin *et al.*, "Multi-image super-resolution for remote sensing using deep recurrent networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 816–825, doi: 10.1109/CVPRW50498.2020.00111.

[217] N. Ballas, L. Yao, C. Pal, and A. Courville, "Delving deeper into convolutional networks for learning video representations," Mar. 2016, *arXiv: 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 Sens.*, vol. 12, no. 14, p. 2207, Jul. 2020, doi: 10.3390/rs12142207.

[219] J. Ma, L. Zhang, and J. Zhang, "SD-GAN: Saliency-discriminated GAN for remote sensing image superresolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 11, pp. 1973–1977, Nov. 2020, doi: 10.1109/LGRS.2019.2956969.

[220] H. Wu, L. Zhang, and J. Ma, "Remote sensing image super-resolution via saliency-guided feedback GANs," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, Dec. 2020, doi: 10.1109/TGRS.2020.3042515.

[221] L. Zhang, D. Chen, J. Ma, and J. Zhang, "Remote-sensing image superresolution based on visual saliency analysis and unequal reconstruction networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 4099–4115, Jun. 2020, doi: 10.1109/TGRS.2019.2960781.

[222] L. Zhang, J. Ma, X. Lv, and D. Chen, "Hierarchical weakly supervised learning for residential area semantic segmentation in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 1, pp. 117–121, Jan. 2020, doi: 10.1109/LGRS.2019.2914490.

[223] L. Wang, M. Zheng, W. Du, M. Wei, and L. Li, "Super-resolution SAR image reconstruction via generative adversarial network," in *Proc. 12th Int. Symp. Antennas, Propag. EM Theory (ISAPE)*, 2018, pp. 1–4, doi: 10.1109/ISAPE.2018.8634345.

[224] F. Gu, H. Zhang, C. Wang, and F. Wu, "SAR image super-resolution based on noise-free generative adversarial network," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS 2019)*, 2019, pp. 2575–2578, doi: 10.1109/IGARSS.2019.8899202.

[225] Y. Li, D. Ao, C. O. Dumitru, C. Hu, and M. Datcu, "Super-resolution of geosynchronous synthetic aperture radar images using dialectical GANs," *Sci. China Inf. Sci.*, vol. 62, no. 10, p. 209,302, Apr. 2019, doi: 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 *Proc. Int. Conf. Electron., Circuits Inf. Eng. (ECIE)*, 2021, pp. 228–233, doi: 10.1109/ECIE52353.2021.00056.

[227] H. Shen, L. Lin, J. Li, Q. Yuan, and L. Zhao, "A residual convolutional neural network for polarimetric SAR image super-resolution," *ISPRS J. Photogram. Remote Sens.*, vol. 161, pp. 90–108, 2020, doi: 10.1016/j.isprsjprs.2020.01.006.

[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 *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS 2020)*, 2020, pp. 2101–2104, doi: 10.1109/IGARSS39084.2020.9324079.

[229] L. Lin, J. Li, Q. Yuan, and H. Shen, "Polarimetric SAR image super-resolution via deep convolutional neural network," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS 2019)*, 2019, pp. 3205–3208, doi: 10.1109/IGARSS.2019.8898160.

[230] P. Wang, H. Zhang, and V. M. Patel, "SAR image despeckling using a convolutional neural network," *IEEE Signal Process. Lett.*, vol. 24, no. 12, pp. 1763–1767, Dec. 2017, doi: 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 Sens.*, vol. 10, no. 10, 2018, doi: 10.3390/rs10101597.

[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 *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, 2020, pp. 1–5, doi: 10.1109/ISCAS45731.2020.9180733.

[233] T. Wang, W. Sun, H. Qi, and P. Ren, "Aerial image super resolution via wavelet multiscale convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 769–773, 2018, doi: 10.1109/LGRS.2018.2810893.

[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 Ubiquitous Comput.*, pp. 1–12, 2019, doi: 10.1007/s00779-019-01355-5.

[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, May 2019, doi: 10.1109/ACCESS.2019.2915944.

[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 Ubiquitous Comput.*, pp. 1–10, 2021, doi: 10.1007/s00779-020-01516-x.

[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, Nov. 2019, doi: 10.1016/j.neucom.2019.07.073.

[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, Oct. 2019, doi: 10.1109/ACCESS.2019.2950053.

[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 Sens.*, vol. 13, no. 12, p. 2308, 2021, doi: 10.3390/rs13122308.

[240] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst. (GIS '10)*, 2010, p. 270, doi: 10.1145/1869790.1869829.

[241] G. Sheng, W. Yang, T. Xu, and H. Sun, "High-resolution satellite scene classification using a sparse coding based multiple feature combination," *Int. J. Remote Sens.*, vol. 33, no. 8, pp. 2395–2412, Apr. 2012, doi: 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 *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2015, pp. 5003–5006, doi: 10.1109/IGARSS.2015.7326956.

[243] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2321–2325, Nov. 2015, doi: 10.1109/LGRS.2015.2475299.

[244] L. Zhao, P. Tang, and L. Huo, "Feature significance-based multibag-of-visual-words model for remote sensing image scene classification," *J. Appl. Remote Sens.*, vol. 10, no. 3, p. 35,004, Jul. 2016, doi: 10.1117/1.JRS.10.035004.

[245] G.-S. Xia *et al.*, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017, doi: 10.1109/TGRS.2017.2685945.

[246] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017, doi: 10.1109/JPROC.2017.2675998.

[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 Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 2108–2123, Apr. 2016, doi: 10.1109/TGRS.2015.2496185.

[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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 44–51, doi: 10.1109/CVPRW.2015.7301382.

[249] M. Schmitt, L. H. Hughes, and X. X. Zhu, "The SEN1-2 dataset for deep learning in SAR-optical data fusion," Jul. 2018, *arXiv: 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," Jun. 2019, *arXiv: 1906.07789*.

[251] G.-S. Xia *et al.*, "DOTA: A large-scale dataset for object detection in aerial images," May 2019, *arXiv: 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 Sens. Environ.*, vol. 133, pp. 193–209, Jun. 2013, doi: 10.1016/j.rse.2013.02.007.

[253] J. Li, Y. Li, L. He, J. Chen, and A. Plaza, "Spatio-temporal fusion for remote sensing data: An overview and new benchmark," *Sci. China Inf. Sci.*, vol. 63, no. 4, p. 140,301, Apr. 2020. https://link.springer.com/10.1007/s11432-019-2785-y, doi: 10.1007/s11432-019-2785-y.

[254] X.-Y. Tong *et al.*, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sens. Environ.*, vol. 237, p. 111,322, Feb. 2020, doi: 10.1016/j.rse.2019.111322.

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

[256] P. Wei *et al.*, "Component divide-and-conquer for real-world image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 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 Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 8059–8076, 2020, doi: 10.1109/TGRS.2020.2986313.

[258] W. Xie, Y. Cui, Y. Li, J. Lei, Q. Du, and J. Li, "HPGAN: Hyperspectral pansharpening using 3-d generative adversarial networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 463–477, 2021, doi: 10.1109/TGRS.2020.2994238.

[259] J. Gu, H. Lu, W. Zuo, and C. Dong, "Blind super-resolution with iterative kernel correction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1604–1613, doi: 10.1109/CVPR.2019.00170.

[260] V. Cornillère, A. Djelouah, W. Yifan, O. Sorkine-Hornung, and C. Schroers, "Blind image super resolution with spatially variant degradations," *ACM Trans. Graph.*, vol. 38, no. 6, pp. 1–13, 2019, doi: 10.1145/3355089.3356575.

[261] Z. Luo, Y. Huang, S. Li, L. Wang, and T. Tan, "Unfolding the alternating optimization for blind super resolution," Nov. 2020, *arXiv: 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," Sep. 2018, *arXiv: 1809.00437.*

[263] G. Kim *et al.*, "Unsupervised real-world super resolution with cycle generative adversarial network and domain discriminator," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 1862–1871, doi: 10.1109/CVPRW50498.2020.00236.

[264] S. Maeda, "Unpaired image super-resolution using pseudo-supervision," Feb. 2020, *arXiv: 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 Trans. Image Process.*, vol. 29, pp. 1101–1112, Sep. 2019, doi: 10.1109/TIP.2019.2938347.

[266] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. Int. Conf. Mach. Learn.*, 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," 2021. [Online]. Available: https://arxiv.org/abs/2104.07636

GRS