**Review Article**

**A Survey of Deep Learning-Based Pan-Sharpening Techniques for Remote Sensing Images**

**Abstract:** Nowadays, remote sensing images provide a wealth of information for various applications, including land-use mapping, environmental monitoring, and disaster management. Pan-sharpening refers to the enhancement of spatial resolution in multispectral images by fusing them with panchromatic images. However, these images often suffer from low spatial resolution due to the limitations of the sensors used. To address this issue, image fusion methods, such as pan-sharpening, have been developed to merge high-resolution panchromatic images with low-resolution multispectral images, resulting in high-resolution multispectral images.Recently, deep learning-based pan-sharpening techniques have emerged as a promising approach to achieving high-quality results in this area. In this survey, we provide a comprehensive overview of recent advancements in deep learning-based pan-sharpening techniques for remote sensing images. We review and compare various deep learning architectures and approaches, including autoencoder-based methods such as denoising autoencoders, generative adversarial networks (GANs), conditional GANs, convolutional neural networks (CNNs), photo-realistic single image super-resolution using GANs, deep residual networks, and multispectral image fusion based on maximum a posteriori estimation, among others.We also discuss the challenges and future directions of this field, highlighting the advantages of using deep learning techniques for pan-sharpening. Furthermore, we present an in-depth analysis of different state-of-the-art methods, including their network architectures and experimental results. Additionally, the paper covers evaluation metrics for assessing the quality of pan-sharpened images and provides a comparative analysis of the surveyed methods.

**Introduction:**

In the 1970s, the advent and advancement of image sensors led to the development of multi-sensor information fusion. This gave rise to a new field of research known as image fusion, which combines sensor technology, signal processing, image processing, and artificial intelligence to analyze images within the context of information fusion. In this approach, multiple image sensors, or a single image sensor operating in different modes, capture image data of the same scene[1]. This data is then combined to generate a more precise and comprehensive representation of the scene. This technique is particularly important for applications such as remote sensing, which require accurate details and information about objects or areas.

Remote sensing involves acquiring information about the Earth's surface using various sensors mounted on aircraft or satellites. Remote sensing images contain valuable information that can be utilized for a wide range of applications, including land cover classification, urban planning, environmental monitoring, and disaster management.

Pan-sharpening is a process that fuses high-resolution panchromatic images with low-resolution multispectral images to generate a high-resolution color image. This process is crucial in remote sensing applications as it enhances the spatial resolution and improves the visual quality of multispectral images by incorporating high-resolution panchromatic (pan) images. In recent years, deep learning-based methods have gained popularity for pan-sharpening due to their ability to learn complex feature representations from data. Current state-of-the-art deep learning methods demonstrate higher accuracy and faster performance compared to conventional pan-sharpening techniques.Here is a comprehensive understanding of the fundamental concepts and techniques in deep learning-based pan-sharpening:

**Multispectral and Panchromatic Images:** Remote sensing images are acquired by sensors mounted on satellites or aircraft, capturing various types of electromagnetic radiation. Multispectral images record radiation across several spectral bands, including red, green, and blue, while panchromatic images capture a broad range of wavelengths in a single band. Although panchromatic images have higher spatial resolution than multispectral images, they possess lower spectral resolution.

**Pan-sharpening**: Pan-sharpening is a technique that fuses low-resolution multispectral images with high-resolution panchromatic images to produce high-resolution multispectral images. The goal of pan-sharpening is to enhance the spatial resolution of multispectral images while preserving their spectral characteristics. This technique generates high-quality, high-spatial-resolution color composite images by combining the spectral information from the multispectral (MS) images with the spatial information from the panchromatic (pan) images.

With the increasing availability of high-resolution satellite imagery, there has been a growing interest in developing efficient and accurate techniques for pan-sharpening and image fusion. In recent years, deep learning has emerged as a promising approach for tackling these challenges. This review will discuss the background, recent advancements, and future directions in deep learning-based pan-sharpening techniques and image fusion methods.

**Deep Learning:** Deep learning is a subfield of machine learning that employs deep neural networks (DNNs) to learn representations of data. DNNs consist of multiple layers of interconnected neurons, which enable the learning of hierarchical representations of input data. Deep learning has been extensively applied in image processing and computer vision tasks.

**Overview, Background, and related work:**

**Pan-Sharpening Techniques:** Pan-sharpening techniques combine a high-resolution panchromatic (pan) image with a low-resolution multispectral (MS) image to create a high-resolution multispectral image. Traditional pan-sharpening methods are based on mathematical models, such as the Brovey transform, IHS (Intensity-Hue-Saturation) transform, and PCA (Principal Component Analysis)-based methods. However, these methods have limitations regarding accuracy and robustness. In contrast, deep learning-based pan-sharpening techniques utilize convolutional neural networks (CNNs) to learn the mapping between the low-resolution MS image and the high-resolution pan image.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | Image Fusion Techniques | | | | | | Spatial Domain | Frequency Domain | | Hybrid Methods | Deep Networks Methods | | Simple Averaging  Weighted averaging  Min/Max  Block Replace  PCA  HIS  Brovey  K-Mean  Fuzzy C-Mean  Guided Filtering | Image Pyramid | wavelet | Pyramid Blending  Multi-scale Transform-based Fusion  Region-based Fusion  Dictionary Learning-based Fusion | Supervised Deep Learning Methods  Unsupervised Deep Learning Methods | | Gaussian Pyramid  Laplacian Pyramid  Discrete Cosine Transform | Continues Wavelet Transform  Discrete Wavelet Transform  Curvelet Transform | |

Fig1. Various image fusion methods

Image fusion refers to the process of combining multiple images of the same scene to generate a single image with improved quality or additional information. Various image fusion methods have been proposed for remote sensing applications, which can be categorized as follows (Fig 1):

**1-Spatial domain methods:**

The spatial-based technique is a straightforward method for combining images, encompassing various approaches such as Principal Component Analysis (PCA), Intensity-Hue-Saturation (IHS) transform, Brovey transform, and others. Each approach has its own advantages and disadvantages. For example, PCA is efficient and provides high spatial quality with reduced computational time; however, it may lead to color distortion and spectral degradation. The IHS transform is also efficient and easy to implement, offering strong sharpening capabilities and rapid processing, but it can also suffer from color distortion. The Brovey transform is an extremely simple and efficient method that boasts faster processing times and produces Red-Green-Blue (RGB) images with enhanced contrast; nevertheless, it may also result in color distortion [2][3].

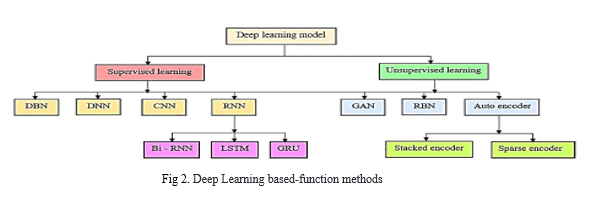
**2-Frequency domain methods:**

The method described in [4] involves decomposing the multiscale coefficients obtained from input images using a frequency-based approach. One notable technique in this category is the Laplacian/Gaussian pyramid, which offers improved image quality for multi-focus images; however, it may not produce significantly different results compared to other methods. The number of decomposition levels can influence the outcomes of the image fusion process. Another approach is the Discrete Cosine Transform (DCT), which breaks down images into a series of waveforms and is suitable for real-time applications, but it may result in lower quality fused images. The Discrete Wavelet Transform (DWT), including Haar wavelet fusion and Stationary Wavelet Transform (SWT), as well as the Curvelet Transform, are also applicable for real-time applications, although they can be time-consuming processes [5, 6, 7, 8].

**3-** **Hybrid Methods**

Pyramid blending is a technique that involves merging images at different resolution levels. This process typically includes the creation of Gaussian and Laplacian pyramids for the images, which are then combined to produce a hybrid image [9]. Multi-scale Transform-based Fusion entails decomposing the input images using multi-scale transforms, such as wavelet transforms or curvelet transforms. In this method, the low-frequency components are fused, while the high-frequency components are selectively combined [10]. Region-based Fusion involves segmenting the input images into regions based on specific criteria, such as intensity, texture, or color. Fusion is then performed at the regional level to generate the final image [11]. Dictionary learning techniques represent images as sparse linear combinations of atoms in a learned dictionary. In the context of fusion, dictionaries derived from input images are utilized to integrate information from multiple sources [12].

**4-Deep Learning based-function methods** Deep learning-based pan-sharpening and image fusion methods employ deep neural networks (DNNs) to learn the mapping between low-resolution and high-resolution images. These techniques have demonstrated significant improvements over traditional methods in terms of accuracy and visual quality. The most commonly used convolutional neural network (CNN) architectures for pan-sharpening and image fusion include U-Net, ResNet, and generative adversarial networks (GANs). The key concepts, theories, and methodologies employed in these techniques are as follows (Fig. 2):



Deep learning-based pan-sharpening is a promising approach for enhancing the spatial resolution of multispectral images. The use of convolutional neural networks (CNNs) or generative adversarial networks (GANs) enables the learning of complex mappings between low-resolution multispectral images and high-resolution panchromatic images. The performance of pan-sharpening algorithms can be evaluated using objective metrics as well as visual inspection. In recent years, deep learning-based pan-sharpening techniques have gained significant attention in the field of remote sensing image processing. The primary objective of these techniques is to fuse high-spatial-resolution panchromatic (pan) images with low-spatial-resolution multispectral (MS) images to generate high-quality, high-spatial-resolution color composite images. The literature related to deep learning-based pan-sharpening techniques can be broadly categorized into the following subtopics:

**Deep Belief Networks** (**DBNS) Methods**:

Deep Belief Networks (DBNs) are generative deep learning models composed of multiple layers of stochastic hidden units. They are frequently used for unsupervised pretraining of deep neural networks. These models are capable of learning intricate patterns and representations from data, making them well-suited for a wide range of complex machine learning tasks.

**Convolutional Neural Network (CNN) Based Methods:** Convolutional Neural Network (CNN)-based methods have been widely used for the pan-sharpening of remote sensing images. These methods utilize a deep neural network architecture to learn the mapping between low-resolution multispectral (MS) images and high-resolution panchromatic (pan) images. Various types of CNN architectures, such as U-Net, ResNet, and VGG-16, have been employed for this purpose.

Deep learning-based pan-sharpening and image fusion methods leverage CNNs to learn the mapping between low-resolution and high-resolution images. CNNs are deep learning models capable of automatically learning features and patterns from input data. They consist of multiple layers, including convolutional layers, pooling layers, and fully connected layers. The convolutional layers apply filters to extract spatial features from the input data, while the pooling layers reduce the dimensionality of the feature maps. The fully connected layers use the extracted features to make final predictions [13].

CNNs are among the most widely used deep learning models and are extensively applied in various computer vision applications, including self-driving cars, gesture recognition, and automatic license plate recognition. The architecture of CNNs is designed to exploit the two-dimensional nature of input images (or other 2D data such as spectrograms) through local connections and fixed weights, combined with pooling layers that result in translationally invariant features[14,15].

One of the advantages of CNNs is that they are faster to train and have fewer parameters than fully connected models with the same number of hidden units. A typical CNN architecture comprises several convolutional and optional sub-sampling layers, followed by fully connected layers. U-Net, a popular CNN architecture originally designed for medical image segmentation, consists of an encoder-decoder structure, where the encoder extracts features from the input image and the decoder reconstructs the output image from these features. U-Net has been widely applied to pan-sharpening and image fusion. A typical architecture of a CNN, featuring several convolutional and pooling layers, is illustrated in Fig 3.

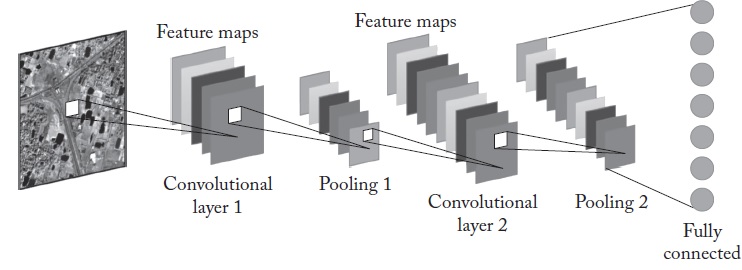


Fig 3. Typical CNN architecture.

**ResNet:** ResNet is a Convolutional Neural Network (CNN) architecture that employs residual connections to mitigate the vanishing gradient problem often encountered in deep neural networks. ResNet has been effectively utilized in pan-sharpening and image fusion tasks to enhance the accuracy and robustness of these models .Fig 4 shows a typical Residual Net [16].

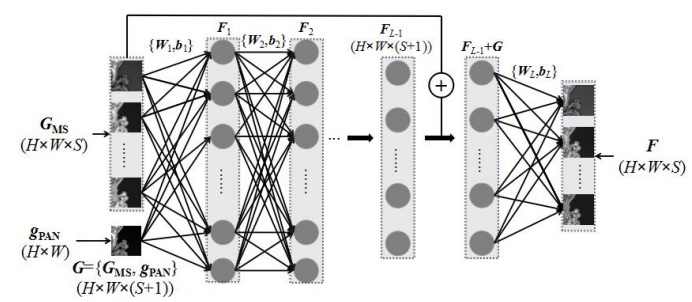


Fig 4. Deep Residual Pan-sharpening Neural Network Flowchart [16].

**Generative Adversarial Network (GANs) Based Methods**: Generative Adversarial Networks (GANs) are a type of neural network architecture comprised of a generator and a discriminator. The generator creates synthetic images, while the discriminator attempts to distinguish between real and synthetic images. GANs have been utilized in pan-sharpening and image fusion tasks to produce high-quality images with enhanced spatial and spectral information [17]. They have proven particularly effective in pan-sharpening, as they can generate high-resolution images that closely resemble the ground truth.

GANs are especially advantageous for tasks where there is a mismatch between low-resolution and high-resolution data, as they can learn to generate realistic details that may be absent in the low-resolution images. The application of GANs in deep neural networks has been steadily increasing, and they have been shown to provide significant improvements in performance [18].

Training a GAN involves the simultaneous use of both the generator and discriminator networks. The generator synthesizes realistic images from randomized noise inputs, leading to diverse outputs based on statistical properties. For image enhancement tasks, a specific type of GAN architecture known as Conditional GANs (cGANs) is often more suitable, as the input to the generator is the image itself. The output, however, may differ, such as when generating an edge map [19]. The standard GAN architecture is illustrated in Fig. 5.

.

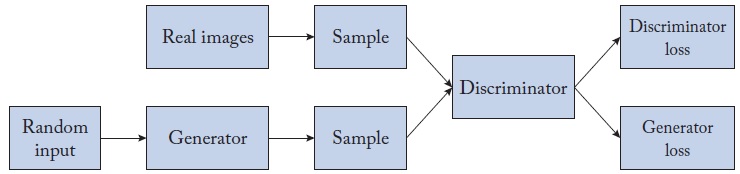


Fig 5. Typical GAN architecture.

The training of a GAN model involves optimizing a min-max problem in which the goal is to predict weights and biases of both parts of the architecture (generator θG and the discriminator θD) as follows (in the following notation IR is the real image and IIN is the low-quality input image):

(IR)] + (1-1)

An important work that demonstrates the capabilities of Generative Adversarial Networks (GANs) in typical inverse imaging problems is the Super-Resolution GAN (SRGAN) architecture [20]. The advantages of GANs are summarized below.

**GANs are data generators:** A key feature of GANs is that they are capable of generating data very similar to real data. Generated data by GANs such as text, images, and videos are hard to distinguish from real data.

**GANs can learn distributions of data:** Since GANs can learn internal representations of data, they can extract meaningful features.

**GANs-trained discriminator can serve as a classifier:** The second module for the learning process is the discriminator which at the end of the training stage can be considered to be a classifier.

**Autoencoder Based Methods:**

Autoencoders are unsupervised deep learning models used to learn efficient data representations. Autoencoder-based methods have been proposed for the pan-sharpening of remote sensing images. These methods utilize an autoencoder architecture to learn the mapping between low-resolution multispectral (MS) images and high-resolution panchromatic (pan) images.

**Denoising Auto Encoder (DAE):**

A Denoising Auto Encoder architecture is among the earliest deep-learning models used in remote sensing[21].

Consider a vectored input data represented by x [0, 1] d, which is mapped to *y* [0, 1] d’ using a deterministic mapping expressed as *y* = *fθ*(x) = *s (W x+b*) where θ = {*W*, b} are the mapping parameters, with W denotes a d x d′ weight matrix, and b is a bias vector. The elements of y are then mapped back to a reconstructed vector represented by z [0, 1] d. The mapping function for the reconstruction is expressed as z = (y) = s(*W*′y+b′). Here, each training frame x(i)is mapped to a hidden representation frame y(i) ,which is then mapped back to a reconstructed frame z(i). The parameters of the network are normally obtained by solving the following optimization problem:

Where *L* denotes l2-norm

The issue with Denoising Autoencoders (DAEs) is that, in some applications during training, the weights may vary considerably, leading to a phenomenon known as the vanishing gradient. To address this, weight regularization can be employed to restrict the range of weight updates to a prescribed amount. Therefore, Equation (1.1) is modified as follows:

Where λ indicates a constant as a tuning parameter. It is worth mentioning that in practice it is seen that during the training phase, only a portion of the weights of each layer remains active. This is important from a complexity point of view. This issue can be taken into consideration in Eq. (1.2) by using a sparsity constraint as follows:

The last term is defined as:

And =) is the average activation of a hidden layer. A typical architecture of DAE is shown in Fig6.

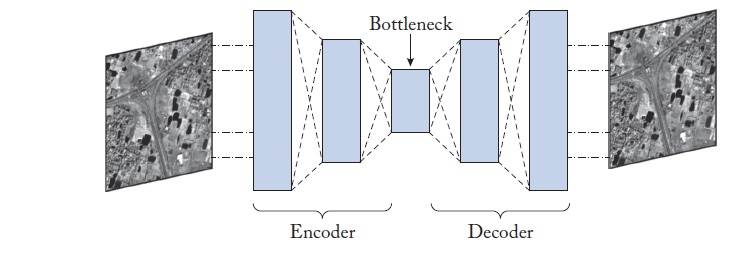


Fig 6. Typical DAE architecture

**Reduced-Resolution Protocol (Wald’s Protocol)**

The protocol operating at reduced resolution is based on Wald’s protocol, which considers the following conditions [22]:

1. The fused (pan-sharpened) image should closely resemble the original multi-spectral (MS) image when re-sampled to a lower MS resolution.
2. The fused image is intended to be as comparable as possible to the original MS image with the highest spatial resolution.
3. The set of fused images should closely match the intact MS image set with the highest spatial resolution.

Assuming that the performance of fusion methods remains consistent across different scales, both the original Panchromatic (PAN) and MS images can be degraded to a coarser resolution. The original MS image is preserved, and the fusion process is conducted using the down-scaled versions of the input images. This approach allows for the use of full-reference objective metrics for evaluation. The flowchart of Wald's protocol is presented in Fig.7.

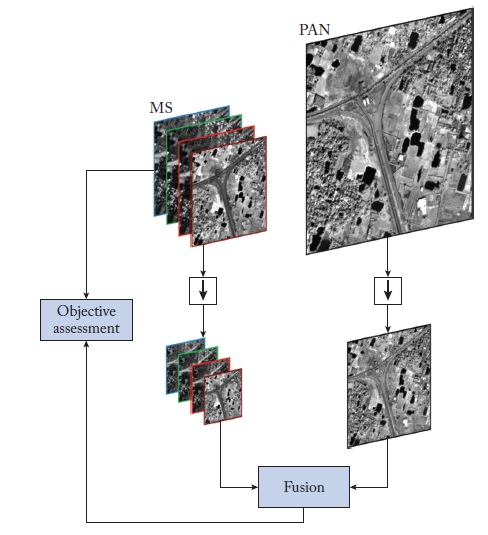


Fig 7. Wald’s protocol.

The training procedure is conducted using a backpropagation algorithm in conjunction with stochastic gradient descent. Each update iteration corresponds to a batch size of 128 input tiles that are randomly extracted from the training data. The batch mean squared error (MSE) between the fused product and its reference M is obtained as follows:

SGD uses a momentum-controlling parameter; hence, at iteration update is done as follows:

= + μ∙∆−α ∙ 1 6

Where denotes the momentum, and a learning rate.

**Deep Residual Learning for Pan sharpening (DRPNN):**

It is seen that a CNN network with more filters and hidden layers results in extracting better useful high-level features. As a result, higher estimation accuracy can be obtained. However, the gradients of the estimation loss to parameters in the shallow hidden layers cannot be passed via backpropagation because of the gradient vanishing issue, thus preventing the DNN from getting trained well. Deep residual learning procedure is an innovative technique for solving this problem, in which the conventional form of mapping is substituted with . Presumably, in the residual image most pixels are very close to zero, and the spatial distribution of the residual features will be very sparse. Hence, looking for a distribution that is very close to the optimal for becomes easier, which allows one to add more hidden layers to the network and improve its performance[23]. However, in the pan-sharpening task, the size of the final output is not the same as the size of the input ). Consequently, rather than predicting the residual features directly, the process through the DRPNN with L layers is divided into the following two stages:

Stage 1: The first to the layers are concatenated under a skip connection to predict the residual between and the convolution operations in each layer are obtained as follows:

== max (0,

The residual output from the layer is then added to to obtain ,

Stage 2: The layer of the DRPNN is established to reduce the dimension of the spectral space from bands to N bands via the last 3D convolution operation in the network, resulting in the final prediction

=

The framework of the DRPNN is depicted in Fig.8.

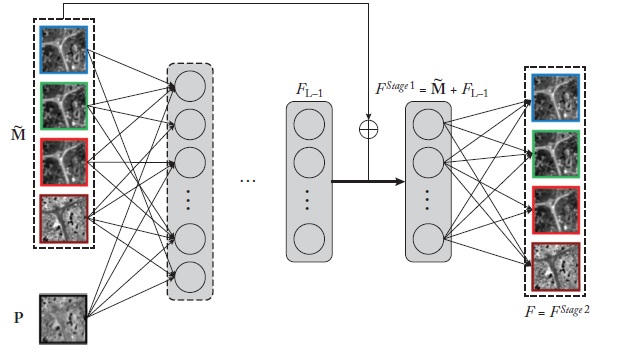


Fig 8. Flowchart of deep model training via DRPNN.

**Pan Net**: The architecture for Pan Net is shown in Figure 1.10. Pan Net was motivated by considering spectral preservation and spatial injection in its architecture [24]. A straightforward way of using a DNN model for image fusion in remote sensing can be a network that learns the nonlinear mapping relationship between the inputs (P, M) and the outputs X by minimizing the following loss function:

Where *fW* represents a neural network and *W* its parameters. This idea is utilized by the PNN method (described above) [24], which directly inputs (***P, M***) into a deep CNN to estimate X. Although this method gives satisfactory performance, it does not address image characteristics. Motivated by the success of PNN, the Pan Net architecture is built to perform the pan sharpening task. As with PNN, Pan Net also uses a CNN but its structure differs from PNN [25]. As reported in [25], convolutional filters are useful since they can extract the high correlation across different bands of the multispectral images. Pan Net attempts to preserve both spectral and spatial content. Fig.9 and 10 illustrate several Pan Net network structures and for test model. The first structure is vanilla Res Net which focuses on spatial information, while the second structure only focuses on spectral information. The third structure takes into consideration both spatial and spectral aspects.

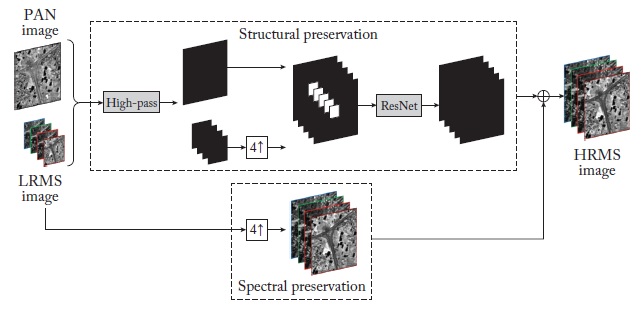


Fig 9. Architecture of Pan Net.

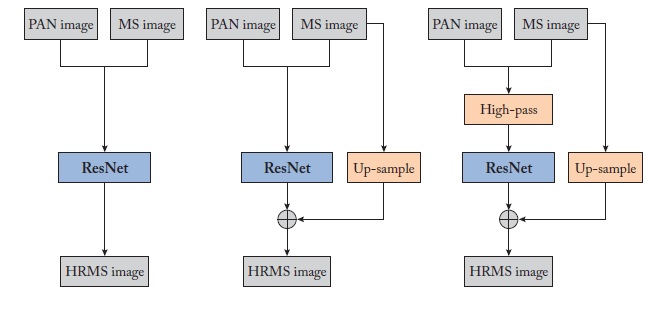


Fig 10. Architectures used to test the Pan Net model.

**CAE Training for Pan Sharpening**: A CAE network is used to improve the spatial information of the LRMS bands by learning the nonlinear relationship between a PAN image and its spatially degraded version [26], as illustrated in Fig.11. As indicated earlier, a PAN image is considered to be the reference spatial image. A spatially degraded version of the PAN image is first generated using an interpolation filter. The original and degraded PAN images are partitioned into patches with r overlapping pixels. A CAE network is then used to learn the nonlinear relationship between the original PAN patches and corresponding degraded PAN patches as its target and input, respectively. After training, the CAE network generates approximated high-resolution LRMS patches as its output in response to LRMS patches as its input.

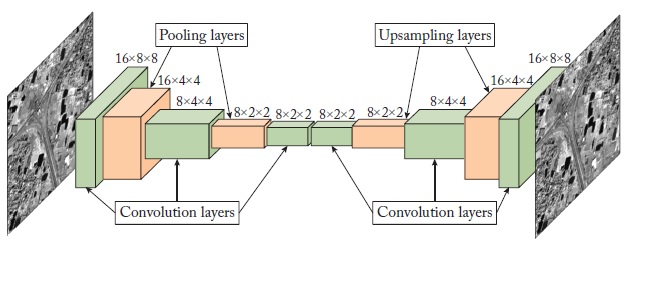


Fig 11. CAE architecture used for pan-sharpening in [26].

The patch-wise low-resolution multispectral (LRMS) bands are fed into the CAE-trained network. Due to the similarity in spectral characteristics between panchromatic (PAN) and multispectral (MS) images, the trained network is expected to enhance the spatial information of the LRMS bands. The input and output of the CAE network are illustrated in Fig.12. From this figure, it is evident that an estimated high-resolution LRMS is generated at the output of the network for each band. In fact, the reconstructed version not only preserves the spectral information of the LRMS bands but also contains more spatial information compared to the input patches. What differentiates this approach from previous methods is that the estimated high-resolution LRMS is used instead of the original LRMS.

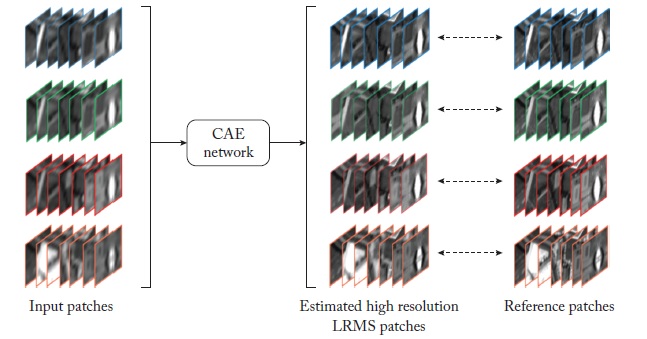


Fig 12. Testing phase of the CAE architecture used in [26].

**Detail-Preserve Cross-Scale Learning for CNN-Based Pan sharpening**: A significant aspect of deep learning solutions for pan-sharpening is the extent to which the model can generalize from the training resolution to the testing resolution. The training set is obtained through a resolution shift [22]. In remote sensing, the challenge of transitioning from the training resolution to the testing resolution is critical, as all sample images are captured from a specific distance above the Earth's surface, maintaining a relatively constant distance from the ground sample. Consequently, representative variations of these elements may not be well captured in the training dataset, leading to a misalignment between the training and test datasets. This explains why the substantial performance improvement of CNN-based methods compared to conventional methods in a reduced resolution evaluation setting may not translate to similar performance at full resolution.

Several attempts have been made to address this issue in pan-sharpening [27–28]. The underlying approach in these attempts is to incorporate a complementary loss term that addresses the target-resolution behavior of the model. Based on this concept, a new pan-sharpening method has been developed, as illustrated in Fig.13. Specifically, this training scheme is applied directly during the fine-tuning stage of the pre-trained model using the adjusted parameters.

. This approach takes into consideration the performance loss between the training and test sets [29].

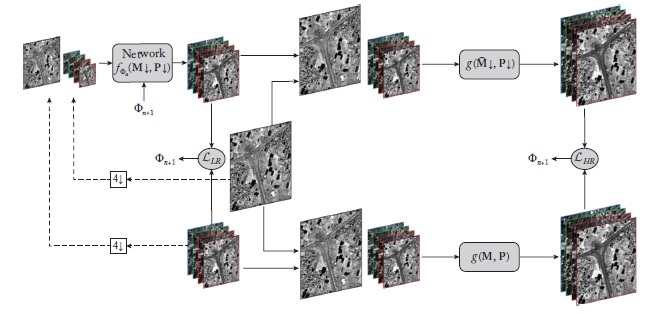


Fig 13. Steps involved in detail-preserve cross-scale learning [29].

Basically, in the fine-tuning stage, along with the term defined in the reduced resolution, a full-resolution term is considered by appropriately processing the target image as indicated in Fig.12:

=

In which the multispectral image fusion at the final resolution is done using a mapping function Specifically, one of the traditional methods that denote reasonable performance on preserving spatial consistency is utilized (MTF-GLP-HPM). This conventional algorithm highly depends on the extracted spatial information from the PAN image while restricting the spectral distortion effect. Therefore, the cross-scale consistency is enforced due to

the training on low- and high-resolution data jointly.

**Unsupervised Pan sharpening using Generative Adversarial Networks:** The final target of pan-sharpening is to preserve the spatial and spectral information in the final product. However, current CNN-based approaches typically regard pan-sharpening as a deep learning module. Although Pan Net relies on the preservation of spectral and spatial content, it attains a fused product by combining the up-scaled MS data with the high-frequency information given by CNN which leads to blurred outcomes[27]. In comparison, the above methods depend on the reference image, i.e., Wald's protocol [22], where all the reference images are degraded by the Gaussian kernel and then downscaled by a factor of 4. The rationale behind this work is that the ill-posed pan-sharpening problem cannot be solved without blurriness by reversing the down and Gaussian kernels when dealing with Wald’s protocol. Therefore, an unsupervised pan-sharpening model is proposed (called Pan-GAN here) to utilize the intact source images as the training data to get the HRMS fused image without supervision by the ground-truth data. First, the original MS data is interpolated to the resolution of the PAN image and then stacked from the five-channel data to be fed into the generator G to obtain the HRMS image at the output. It is worth mentioning that the first channel is considered to be the PAN image. It is found that without considering two discriminators for the GAN, the final HRMS suffers from either severe spectral distortion or spatial inconsistencies[30].

In this model, two discriminators are employed to preserve both spectral and spatial information. The flowchart of the developed unsupervised method is illustrated in Fig. 14. To effectively maintain spectral and spatial fidelity, two tasks are addressed simultaneously using the two discriminators. The first discriminator, labeled D1 in Figure 14, serves as the spectral discriminator and is responsible for assessing the spectral consistency of the generated image with the low-resolution multispectral (LRMS) image data. The second discriminator, D2, is designed to preserve the spatial content in the high-resolution multispectral (HRMS) images. D2 evaluates the spatial consistency of the generated image against the pansharpened (PAN) image. Since the PAN image consists of single-band data, an average pooling operation is performed on the generated image to obtain single-band data for comparison.

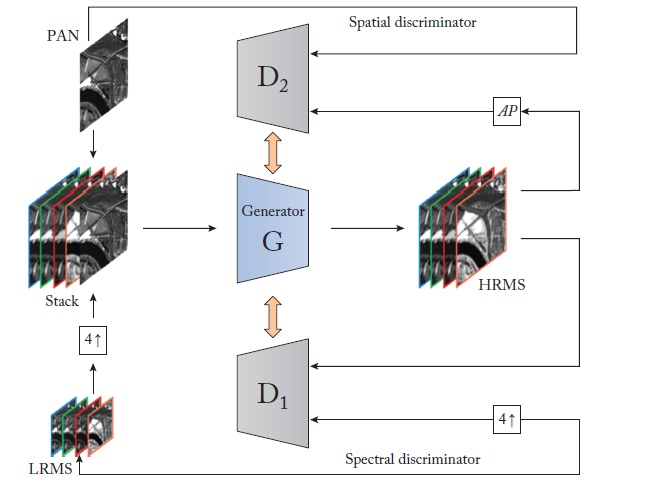


Fig 14. Unsupervised deep learning model for pan sharpening.

**Unsupervised Generative Model for Pan sharpening**

So far, various aspects of conventional and deep learning-based solutions for multispectral image fusion have been considered. Recently, several efforts have been made to enhance the fusion products through cross-scale learning processes and the development of new loss functions. In this section, we discuss these two issues in the context of multispectral image fusion.

As mentioned earlier, deep learning models have been applied to multispectral image fusion, yielding better outcomes than conventional methods. An initial attempt was made in [31], where the pan-sharpening problem was addressed using a deep neural network (DNN) framework that formulated the nonlinear relationship between low-resolution and high-resolution images as a denoising autoencoder. In [32], a three-layer convolutional neural network (CNN) was designed to frame the multispectral image fusion problem as a super-resolution task. The concept of residual learning for multispectral image fusion was first introduced in [33], where a deep CNN was employed. In [33] and [28], a deep denoising autoencoder approach was utilized to model the relationship between a low-pass version of the panchromatic (PAN) image and its high-resolution counterpart. In [34], a deep CNN known as SRCNN was developed for image super-resolution, demonstrating superior performance compared to several other methods. In [35,36], a pan-sharpening method was introduced using SRCNN as a preprocessing step. Furthermore, in [37], a network structure called Pan Net was developed to incorporate prior knowledge of pan-sharpening, enhancing the generalization capability and performance of the network. A generative adversarial network (GAN) approach, referred to as PSGAN, was discussed in [38], which minimizes the loss between the generator and discriminator components. One of the advantages of GANs is their ability to reduce blurriness in the fused image. This approach not only seeks to minimize the L1 loss associated with each pixel but also aims to reduce the loss across the entire fused image.

As far as the loss function in DNNs is concerned, a new perceptual loss was presented in [39] to better preserve the spectral information in fused images. In [38], several objective functions were examined. In the recently developed deep learning-based methods, the focus is placed on the preservation of spatial details. For example, the CNN model in [40] was designed to preserve details via a cross-scale learning process. To address the effect of gradient vanishing, the concept of dense connection to pan-sharpening was extended in [41]. Most of the recently developed deep learning-based methods simply train and regularize the parameters of a network by minimizing a spectral loss between the network output and a pseudo–Ground Truth (GT) image. As described in before, the methods mentioned above primarily use a single objective learning to optimize network parameters and generalize its capability. However, other metrics that can represent both modalities (spatial and spectral) have recently gained more attention. For instance, in [42], based on the correlation maps between MS target images and PAN input images, a loss function was designed to minimize the artifacts of fused images. Also, in [43], it was shown that although a linear combination of MS bands could be estimated from the PAN image, a rather large difference in luminance resulted. Thus, certain objects could not be differentiated properly. To address this issue, a color-aware perceptual (CAP) loss was designed to obtain the features of a pre-trained VGG network that were more sensitive to spatial details and less sensitive to color differences. The aforementioned methods rely on the availability of GT data for regularizing the network parameters. However, in practice, such data are not available [44].

This section examines methods to address the two limitations of existing deep learning models for remote sensing image fusion. The first limitation pertains to the dependency on ground truth (GT) data for training and regularizing the network, while the second limitation involves the use of a generic loss function for parameter estimation. To mitigate the first limitation, an unsupervised learning strategy based on generative adversarial networks (GANs) is employed. In this context, unsupervised learning indicates that labeled reference targets are not available for training the deep model. A key aspect of unsupervised learning is that, although the data processed by the deep model are abundant, the targets and labels are often sparse or even completely absent. The second limitation is addressed by designing a multi-objective loss function that simultaneously considers both spatial and spectral attributes.

This section describes a developed unsupervised generative model. To set the stage, let us begin with the general framework of CS (Component substitution) methods. The CS framework can be mathematically expressed by the following equation:

Where and denote the high-resolution and up-sampled low-resolution MS images, respectively, gk’s are injection gains for spectral bands, ***P*** denotes the PAN image, and ***Ik*** is the ***k-th*** intensity component defined as:

Where is a linear/nonlinear combination of spectral bands [38].

**Learning Process and Loss Functions**

Given a set of unlabeled data at the input, the purpose of a generative model is to estimate the data distribution. This task is challenging and can be time-consuming. Recently, generative adversarial networks (GANs) have been developed to estimate the underlying distribution of unlabeled data. This discussion specifically focuses on their application in multispectral image fusion. Pan-sharpening aims to generate a high-resolution multispectral (HRMS) image.using two input data LRMS image and HRPAN image P. The optimization algorithm which is a min-max problem is generally formulated as follows:

Where the pan sharpening model that takes the and P as input and the parameters to be optimized are represented as . The learning process includes two major branches, e.g., generation and discrimination. The first part of the branch is dedicated to generating a set of images from a random distribution and the role of the second branch is to decide whether the generated data is real or fake by comparing it with the available LRPAN image P and LRMS image .The procedure of training scheme for this method consists of two parts: (1) spectral preservation and (2) spatial preservation. In what follows an explanation of the loss function design and the parameters to optimize is provided. Spectral Preservation Learning Process: For minimizing the spectral distortion in the fused image, a spectral metric is used to deal with spectral consistency. For this purpose, a discriminator for the learning process is considered, named spectral discriminator. The MS image data at the original resolution are used as the input of this discriminator. Initially, the output of the generator is inputted to the spectral discriminator. The following objective function is then used to minimize the spectral distortion of the fused image:

Where is the UIQI as described in [30], and Mk are the estimated high-resolution MS image at the output of the generator and the MS input image at the original resolution, respectively.

**Spatial Preservation Learning Process:** Another discriminator, referred to as the spatial discriminator, is employed to minimize spatial distortion. To inject spatial details into the fused image, the panchromatic (PAN) image, which represents the reference spatial information, is used as input to the discriminator at its original resolution. The following loss function is then utilized during the training phase of the generative model:

whereby is the linear combination of estimated high-resolution MS images at the output of the generator and Pk is the histogram-matched PAN image concerning the *k-th* spectral band. The learning process of the developed method is illustrated in Fig.15.

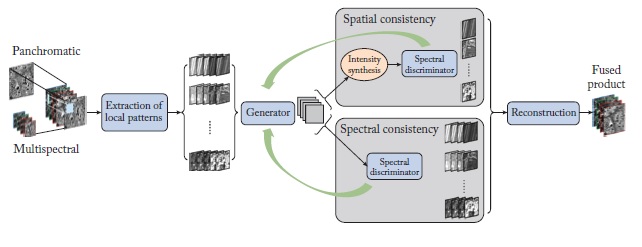


Fig 15. Flowchart of the deep generative model for pan sharpening.

**Evaluation Metrics:**

The performance of pan-sharpening algorithms can be evaluated using various metrics, including Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Universal Image Quality Index (UIQI), and Feature Similarity Index (FSIM). PSNR and SSIM are objective metrics that assess the similarity between the pan-sharpened image and the ground truth high-resolution image. The comparison and analysis of the performance of different deep learning-based pan-sharpening techniques can be conducted using these evaluation metrics, as well as specific datasets. Below are some commonly used evaluation metrics::

1. Peak Signal-to-Noise Ratio (PSNR): It measures the difference between the ground truth image and the generated pan-sharpened image in terms of signal-to-noise ratio.
2. Structural Similarity Index (SSIM): It measures the perceived similarity between the ground truth and generated images in terms of structural information, luminance, and contrast.
3. Universal Image Quality Index (UIQI): It computes the similarity between the generated and ground truth images based on their spatial, spectral, and cross-spectral properties.
4. Feature Similarity Index (FSIM): It measures the similarity between the generated and ground truth images in terms of their feature representation.

**Datasets**

There are several publicly available datasets for evaluating deep learning-based pan-sharpening techniques, including the QuickBird, GeoEye, and WorldView-2 datasets. Each dataset has its own characteristics, such as varying spatial and spectral resolutions, which can impact the performance of pan-sharpening methods. A comparative analysis of various deep learning-based pan-sharpening techniques on the QuickBird dataset showed that CNN- and GAN-based methods outperformed other approaches in terms of PSNR, SSIM, UIQI, and FSIM metrics.

Overall, the choice of evaluation metrics and datasets depends on the specific application and requirements of the pan-sharpening technique. Nevertheless, comparing and analyzing the performance of different methods using a variety of evaluation metrics and datasets is essential for assessing the strengths and weaknesses of each technique and determining the most suitable one for a given task. The following two tables present the comparative results of various fusion and pan-sharpening methods.Table 1 and 2 show comparison between methods and datasets.

Table 1: shows comparison of different methods and their merits and demerits

|  |  |  |  |
| --- | --- | --- | --- |
| Method | Description | Advantages | Disadvantages |
| IHS | A method that uses an Intensity-Hue-Saturation (IHS)transformation to separate a multispectral image into its color and intensity channels. | Preserves the spatial resolution of the original image. Good for visual interpretation of data | Does not explicitly model correlations between spectral bands. May introduce artifacts and noise in the reconstructed images. |
| Laplacian Pyramid method for fusion | A method that decomposes an image into multiple scales using a Laplacian pyramid and then fuses corresponding scales from different input data. | Can preserve both global and local features in the input images. Can handle images with complex structures and textures | May introduce ringing artifacts or noise in the reconstructed images. Require carful tuning of hyperparameters. |
| PCA | A linear transformation method that seeks to the principal components of input data to reduce it dimensionality. | Simple and computationally efficient. Can handle high dimensional data. | Assume linear relationship between input variables and may not capture non-linear patterns in data well. |
| DAE | A type of autoencoder that learns to reconstruct input data by encoding it to lower dimensional space and then decoding it back. | Can learn compact and interpretable representations of input data. Can handle missing or corrupted data. | May not be as effective at capturing complex patterns in data as other methods. Require carful tuning of hyperparameters. |
| GAN | A generative model that uses two (a generator and a discriminator) to produce high-quality synthetic images. | Can generate highly realistic and diverse images. Can learn complex data distribution. | Training can unstable and difficult to optimize. May suffer from mode collapse. Require carful tuning of hyperparameters. |
| CNN | A deep learning method that uses convolutional neural networks to automatically learn features representations from input data. | Good at capturing complex patterns and relations in image data. Can handle large data efficiently. | Require a large amount of training data. May suffer from over fitting and may be computationally expensive. |

While deep learning-based pan-sharpening and image fusion methods have demonstrated promising results and significant improvements over traditional techniques, several research gaps and limitations remain to be addressed. These include a lack of large-scale benchmark datasets, limited interpretability of models, insufficient robustness to variations in illumination and atmospheric conditions, limited transferability to different sensors and platforms, and challenges in scalability to large remote sensing datasets. Addressing these gaps and limitations is crucial for developing more accurate, robust, and generalizable methods for pan-sharpening and image fusion in remote sensing. Some of the major research gaps and limitations are as follows:

Table 2. Comparison between 6 methods based on four Image quality metrics used Quickbird and Worldview2 datasets 'Images.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Method | FSIM | SSIM | UIQI | FSNR | Datasets |
| CNN | High | High | Moderate | High | Worldview2/ Quick Bird |
| GAN | High | High | High | High | Worldview2/ Quick Bird |
| DAE | Moderate | Moderate | Moderate | Moderate | Worldview2/ Quick Bird |
| PCA | Low | Low | Low | Low | Worldview2/ Quick Bird |
| IHS | High | High | Moderate | Moderate | Worldview2/ Quick Bird |
| LPF | High | Low | Low | High | Worldview2/ Quick Bird |

**1. Lack of large-scale benchmark datasets:** There is a lack of large-scale benchmark datasets for evaluating the performance of deep learning-based pan-sharpening and image fusion methods. Most existing datasets are limited in terms of both size and diversity, making it difficult to effectively compare different methods and assess their generalizability.

**2. Limited interpretability**:Deep learning models are often regarded as black boxes due to the difficulty in understanding how they arrive at their predictions. This lack of interpretability hampers the analysis of these models' performance and complicates the identification of areas for improvement.

**3. Limited robustness to changes in illumination and atmospheric conditions:** Deep learning-based pan-sharpening and image fusion methods are sensitive to changes in illumination and atmospheric conditions, which can affect the accuracy and visual quality of the output images. Developing methods that are robust to these variations is a major research challenge.

**4. Limited transferability to different sensors and platforms:** Deep learning-based pan-sharpening and image fusion methods are often trained on specific sensors and platforms, which limits their transferability to other sensors and platforms. Developing methods that can generalize across different sensors and platforms is an important research direction.

**5. Limited scalability to large-scale remote sensing datasets:** Deep learning-based pan-sharpening and image fusion methods can be computationally intensive, especially when dealing with large-scale remote sensing datasets. Developing methods that are scalable to these datasets is an important research challenge.

**Future Directions:**

1. **More Robust Techniques:** Future research should focus on developing more robust deep learning-based pan-sharpening techniques that can handle noise and other sources of uncertainty in the input data.
2. **Transfer Learning:** Transfer learning can be used to improve the efficiency of deep learning-based pan-sharpening techniques by reusing pre-trained models on different datasets.
3. **Fusion of Multi-source Data:** Future research should explore the fusion of multi-source data, such as optical and SAR data, to improve the accuracy and spatial resolution of remote sensing images.
4. **Interpretable Techniques:** Future research should focus on developing interpretable deep learning-based pan-sharpening techniques that can provide insights into how the network is making decisions.
5. **Real-time Applications:** Future research should focus on developing deep learning-based pan-sharpening techniques that can be applied in real-time applications, such as satellite image processing and UAV image analysis.

**Conclusion**:

Satellite image fusion is a critical task in remote sensing applications, and various methods have been developed to achieve this goal. Traditional techniques, such as Intensity-Hue-Saturation (IHS) and Principal Component Analysis (PCA), have demonstrated good performance; however, recent deep learning-based methods have gained popularity and exhibited superior performance compared to traditional approaches in terms of image quality, accuracy, and processing speed. This improvement is largely due to their ability to learn complex relationships between input and output images. Research studies have evaluated these deep learning methods and compared them to traditional techniques, revealing that deep learning-based approaches can yield better results. Future research should focus on developing more efficient and accurate deep learning models for satellite image fusion.

**References:**

[1]. Liu, Y., Chen, X., Wang, Z., Wang, Z.J., Ward, R.K. and Wang, X. (2018) Deep Learning for Pixel-Level Image Fusion: Recent Advances and Future Prospects. Information Fusion, 24, 158-173. <https://doi.org/10.1016/j.inffus.2017.10.007>

[2]. Jasiunas MD, Kearney DA, Hopf J, Wigley GB (2002) Image fusion for uninhabited airborne vehicles. In: 2002 IEEE International Conference on field-programmable technology, 2002. (FPT). Proceedings, p 348–351. IEEE 15.

[3]. Dong J, Dafang Z, Yaohuan H, Jinying F (2011) Survey of multispectral image fusion techniques in remote sensing applications. In: Zheng Y (ed) Image fusion and its applications. Alcorn State University, USA.

[4]. Yang J, Ma Y, Yao W, Lu WT (2008) A spatial domain and frequency domain integrated approach to fusion multi-focus images. The International archives of the photogrammetry, remote sensing, and spatial Inf sciences, 37(PART B7).

[5]. Ramac LC, Uner MK, Varshney PK, Alford MG, Ferris DD (1998) Morphological filters and wavelet-based image fusion for concealed weapons detection. In Sensor Fusion: Architectures, Algorithms, and Applications II vol 3376, p 110–120. International Society for Optics and Photonics

[6]. Kannan K, Perumal SA, Arulmozhi K (2010) Performance comparison of various levels of fusion of multi-focused images using wavelet transform. Int J Comput Appl 1(6):71–78

[7]. Naidu VPS (2012) Discrete cosine transform-based image fusion techniques. J Commun, Navig Signal Process 1(1):35–45

[8]. Singh R, Khare A (2013) Multiscale medical image fusion in the wavelet domain. Sci World J 1–10. https://doi. org/10.1155/2013/521034

[9]. Burt, P. J., & Adelson, E. H. (1983). The Laplacian pyramid as a compact image code. IEEE Transactions on Communication, 31(4), 532-540.

[10]. Li, S., Fang, L., & Ji, S. (2016). Status and trends of multi-scale image fusion techniques. Information Fusion, 33, 100-112.

[11]. Zhang, H., Liu, Y., & Ding, X. (2013). A region-based image fusion algorithm using NSDF and PCNN in NSST domain. Journal of Visual Communication and Image Representation, 24(7), 782-792.

[12]. Zeng, T., Li, X., Kim, M., Cai, J. F., & Ruan, S. (2020). Dictionary optimization for image fusion using supervised deep learning. IEEE Transactions on Image Processing, 29, 5252-5267.

[13]. Wang, D., et al. (2022). "Multi-Modality Remote Sensing Image Fusion Using Convolutional Neural Networks." Remote Sensing, 14(2), 310.

[14]. Li, W., et al. (2022). "A Hierarchical Multi-Scale Dense Block Network for Satellite Image Fusion." IEEE Access, 10, 11948-11958. This paper proposes a novel hierarchical multi-scale dense block network for satellite image fusion and demonstrates its effectiveness through experiments.

[15]. Jiaxin Li, Danfeng Hong, Lianru Gao, Jing Yao, Ke Zheng, Bing Zhang, Jocelyn Chanussot,Deep learning in multimodal remote sensing data fusion: A comprehensive review,International Journal of Applied Earth Observation and Geoinformation,Volume 112,2022.

[16]. Wei, Y., Yuan, Q., Shen, H., and Zhang, L. 2017. Boosting the accuracy of multispectral image pan-sharpening by learning a deep residual network. IEEE Geoscience and Remote Sensing Letters, 14(10):1795–1799. DOI: 10.1109/lgrs.2017.2736020.

[17]. Liu, X., Wang, Y., and Liu, Q. 2018. PSGAN: A generative adversarial network for remote sensing image pan-sharpening. 25th IEEE International Conference on Image Processing (ICIP), pages 873–877. DOI: 10.1109/icip.2018.8451049. 55

[18]. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S.,Courville, A., and Bengio, Y. 2014. Generative adversarial nets. Advances in Neural Information Processing Systems, pages 2672–2680. 33

[19]. Mirza, M. and Osindero, S. 2014. Conditional generative adversarial nets. ArXiv Preprint ArXiv:1411.1784. 34

[20]. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken,A., Tejani, A., Totz, J., Wang, Z., and Shi, W. 2017. Photo-realistic single image superresolution using a generative adversarial network. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4681–4690. DOI: 10.1109/cvpr.2017.19. 34

[21]. Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P. A. 2008. Extracting and composing robust features with denoising autoencoders. Proc. of the 25th International Conference on Machine Learning, pages 1096–1103. DOI: 10.1145/1390156.1390294. 32.

[22]. Wald, L., Ranchin, T., and Mangolini, M. 1997. Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. Photogrammetric Engineering and Remote Sensing, 63(6):691–699.

[23]. Chen, H., et al. (2022). "A Tensor-Based Deep Learning Method for Multispectral and Panchromatic Image Fusion." IEEE Transactions on Geoscience and Remote Sensing, 60(2), 1228-1240.

[24]. Yang, J., Fu, X., Hu, Y., Huang, Y., Ding, X., and Paisley, J. 2017. PanNet: A deepnetwork architecture for pan-sharpening. Proc. of the IEEE International Conference on Computer Vision, pages 5449–5457. DOI: 10.1109/iccv.2017.193. 41

[25]. Masi, G., Cozzolino, D., Verdoliva, L., and Scarpa, G. 2016. Pansharpening by convolutional neural networks. Remote Sensing, 8(7):594. DOI: 10.3390/rs8070594. 38, 41

[26]. Azarang, A., Manoochehri, H. E., and Kehtarnavaz, N. 2019. Convolutional autoencoder-based multispectral image fusion. IEEE Access, 7:35673–35683. DOI:10.1109/access.2019.2905511. 43, 44, 45, 46

[27]. Gang Yang and Xiangyong Cao and Wenzhe Xiao and Man Zhou and Aiping Liu and Xun chen and Deyu Meng.(2023). A Flow-Based Deep Network for Pan-sharpening. https://arxiv.org/abs/2305.07774.

[28]. Azarang, A. and Ghassemian, H. 2017. A new pan-sharpening method using multi resolution analysis framework and deep neural networks. Proc. of the 3rd International Conference on Pattern Recognition and Image Analysis (IPRIA), Shahrekord, Iran, April 19–20, pages 1–6. DOI: 10.1109/pria.2017.7983017. 45.

[29]. Vitale, S. and Scarpa, G. 2020. A detail-preserving cross-scale learning strategy for CNN-based pan-sharpening. Remote Sensing, 12(3):348. DOI: 10.3390/rs12030348. 55.

[30] Ma, J., Yu, W., Chen, C., Liang, P., Guo, X., and Jiang, J. 2020. Pan-GAN: An unsupervised learning method for pan-sharpening in remote sensing image fusion using a generative adversarial network. Information Fusion. DOI: 10.1016/j.inffus.2020.04.006.48, 52.

[31]. Huang, W., Xiao, L., Wei, Z., Liu, H., and Tang, S. 2015. A new pan-sharpening method with deep neural networks. IEEE Geoscience and Remote Sensing Letters, 12(5):1037–1041.DOI: 10.1109/lgrs.2014.2376034. 55

[32]. M. Ciotola, S. Vitale, A. Mazza, G. Poggi and G. Scarpa, "Pansharpening by Convolutional Neural Networks in the Full Resolution Framework," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-17, 2022, Art no. 5408717, doi: 10.1109/TGRS.2022.3163887.

[33]. Zhou, Z.; Liu, H.; Xie, G. MSDRN: Pansharpening of Multispectral Images via Multi-Scale Deep Residual Network. *Remote Sens.* 2021, *13*, 1200. https://doi.org/10.3390/rs1306120

[34]. Yang, S., Wang, M., Chen, Y., and Sun, Y. 2012. Single-image super-resolution reconstruction via learned geometric dictionaries and clustered sparse coding. IEEE Transactions on Image Processing, 21(9):4016–4028. DOI: 10.1109/tip.2012.2201491. 55

[35]. Wei, Y. and Yuan, Q. 2017. Deep residual learning for remote sensed imagery pansharpening.Proc. of the International Workshop on Remote Sensing with Intelligent Processing (RSIP), Shanghai, China, May 18–21, pages 1–4. DOI: 10.1109/rsip.2017.7958794. 55

[36]. Huang, Wei & Xiao, Liang & Liu, Hongyi & Wei, Zhihui & Tang, Songze. (2015). A New Pan-Sharpening Method With Deep Neural Networks. IEEE Geoscience and Remote Sensing Letters. 12. 10.1109/LGRS.2014.2376034.

[37]. Liang-Jian Deng, Minyu Feng, Xue-Cheng Tai,The fusion of panchromatic and multispectral remote sensing images via tensor-based sparse modeling and hyper-Laplacian prior,Information Fusion,Volume 52,2019,Pages 76-89

[38]. Vitale, S. 2019. A CNN-based pan-sharpening method with perceptual loss. IGARSS, IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan,pages 3105–3108. DOI: 10.1109/igarss.2019.8900390. 55

[39]. Scarpa, G., Vitale, S., and Cozzolino, D. 2018. Target-adaptive CNN-based pansharpening.IEEE Transactions on Geoscience and Remote Sensing, 56(9):5443–5457. DOI:10.1109/tgrs.2018.2817393. 55

[40]. Wang, D., Li, Y., M.a, L., Bai, Z., and Chan, J. C. W. 2019. Going deeper with densely connected convolutional neural networks for multispectral pan-sharpening. Remote Sensing,11(22):2608. DOI: 10.3390/rs11222608. 55

[41]. Bello, J. L. G., Seo, S., and Kim, M. 2020. Pan-sharpening with color-aware perceptual loss and guided re-colorization. ArXiv Preprint ArXiv:2006.16583. DOI:10.1109/icip40778.2020.9190785. 56.

[42]. Ling Liu, Qian Jiang, Xin Jin, Jianan Feng, Ruxin Wang, Hangying Liao, Shin-Jye Lee, Shaowen Yao,CASR-Net: A color-aware super-resolution network for panchromatic image, Engineering Applications of Artificial Intelligence,Volume 114,2022,

[43]. X. Liu, X. Liu, H. Dai, X. Kang, A. Plaza and W. Zu, "Mun-GAN: A Multiscale Unsupervised Network for Remote Sensing Image Pansharpening," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1-18, 2023, Art no. 5404018, doi: 10.1109/TGRS.2023.3288073

[44]. Azarang, A. and Kehtarnavaz, N. 2020. Multispectral image fusion based on map estimation with improved detail. Remote Sensing Letters, 11(8):797–806. DOI:10.1080/2150704x.2020.1773004. 56, 57.