Deep Learning based Multi-modal Medical Image Fusion

Aditya Kahol and Gaurav Bhatnagar

May 10, 2023

Contents

1 Abstract

Medical practitioners often have to work with images which come from various modalities, ranging from X-ray based Computed Tomographies (CT) to radio wave based Magnetic Resonance Imaging (MRI). Each image modality carries different information with them. Multi-modal image fusion is the process of merging images of different modalities to obtain a single image that carries almost all the complementary as well as the redundant details to form an image

Cite this chapter as:
Aditya Kahol and Gaurav Bhatnagar, "Deep learning-based multimodal medical image fusion", in Data Fusion Techniques and Applications for Smart Healthcare, Editor(s): Amit Kumar Singh and Stefano Berretti, pp. 251-279, 2024.

composed of much higher information. This process in which a single image, carrying information of different modalities is rather useful for medical practitioners and researchers for analyzing a patient's body to detect lesions (if any) and to make a correct diagnosis. Feature extraction plays the key role when it comes to image fusion for multimodal image data, and with that in mind, convolutional neural networks have been extensively used in the literature of image fusion for some time now. However, not many of the deep learning based models have been specifically designed for medical images. With that motivation the chapter is divided into two parts. The former will be all about a comprehensive review to some of the works that have been done recently in the field of multimodal image fusion. And inspired by few of the methods discussed, in later, an unsupervised deep learning based medical image fusion architecture incorporating multi-scale feature extraction will be proposed. The extensive experiments on various multi-modal medical images are finally implemented to analyze performance, stability and superiority of the proposed technique.

2 Introduction

Medical imaging refers to the process of image acquisition using some *special* imaging devices that are multimodal in nature, and which lets the viewer know about the internal parts of a human body. With an increasing advancement in the field of radiography and in particular medical imaging, health care industry has been adept at exploiting the various uses of those imaging devices for an effective treatment strategy. The human body is considered to contain structural and functional information [1], where structural information or more commonly known as anatomical information, consists of: bones, soft tissues, cartilage, tendons etc. Generally, these anatomical features are acquired through X-ray based Computed Tomographies, and radio wave based Magnetic Resonance Imaging. CT scans are used to view information having dense structures, i.e. bones, cartilage, tumors, etc. While the radio wave based MRI-scans are used to view lower density features, for example soft tissues, or fluidic components inside the body. Functional information refers to the physiological and metabolic changes within the body, and this information is gathered by the means of Positron Emmision Tomographies (PET), and Single Photon Emission Computed Tomographies (SPECT). Both PET and SPECT uses radiotracers to appraise organ and tissue functions. PET and SPECT are different in the sense that both of them uses different kinds of radiotracers, namely: positron and gamma-ray based, respectively.

With such diverse set of modalities, the information brought forth by these images are quiet complementary, and hence multimodal image fusion has been identified as a decisive solution which aims to integrate information from these images to obtain a single and more complete image, which can facilitate medical practioners to indentify the presence of lesions or any kind of anomaly within the patients body with ease. More formally *image fusion* is the process of merging or integrating complementary information of several source images such that the resultant image provides more detail and quality over any of the individual source images.

2.1 Fusion levels

On the basis of levels of abstraction, image fusion algorithms are categorized into three distinct fusion levels (Fig. 1), namely: (a)Pixel level, (b)Feature level, and (c)Decision level. Each level comes with its benefits and drawbacks when dealing with the complications put forward by multi-modal information from each source images. Several factors can aid for the comparison of each fusion level, which are based on information loss, computational complexity, senstivity to noise, and classification accuracy.

Pixel-level: Also known as low-level image fusion, algorithms in this category works directly with the pixels of the input source images, and therefore amounts to maximum information gain according to human perception. In terms of image processing jargons, fusion results produced by pixel level algorithms have maximum energy. Based on the fusion rule, algorithms in this category are further splitted into two parts: (a)spatial domain, and (b)transform domain methods.

Fusion rules in spatial domain methods are developed by smartly manipulating the input image pixels, methods such as weighted pixel avergaing [3], min-max [3], and focus measure detection [4] comes in this category. Spatial domain techniques are quiet effective when it comes to single sensor source images, however medical images comes from multiple sensors, hence spatial domain algorithms fail to capture relevant details, therefore we transform the source images to some other domain where the relevant information from them can be easily captured, fusion rules defined in this case comes under transform domain methods. It must be noted that upon applying the fusion rule, inverse transformation should also be applied to come back to the spatial domain in order to view the fused image. Pyramid based [5] and [6], and wavelet based algorithms [7], [8] and [9] comes in this category. Because of the effectiveness and ease of implementation, majority of the image fusion literature is filled with pixel-level algorithms. Though intuitive to understand and easily implemented, the algorithms in this category are prone to errors such as presence of artifacts, shift-invariance, and blurriness [10].

Feature-level: Also known as middle or intermediate level algorithms, happen to be a bit more complicated than pixel-level. In hindsight, feature level algorithms are divided into two parts- the first part consists of feature extraction from the input source images, and the second part consists of defining a fusion mapping which utilizes those extracted features to give a high quality fused image. Feature level algorithms aims at capturing the detailed parts of the source images, for example lines, edges, texture, corner, etc. Algorithms in this category are further divided into three classes, each of which has a different

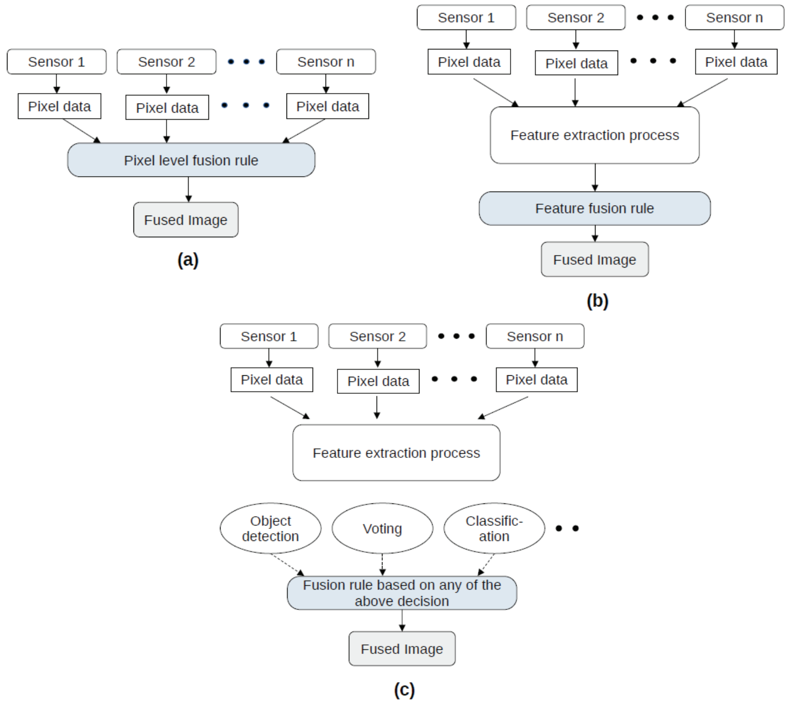


Figure 1: Image Fusion Levels: (a)Pixel Level (b)Feature Level (c)Decision Level

feature extraction procedure, namely: (a)Region-based, (b)Machine-learning based, and (c) Similarity matching based algorithms.

Region-based algorithms begins with region partition and semantic segmentation based approaches to detect salient features from the input images, and then a fusion rule is devised in such a way that the salient features can be merged appropriately. Majority of the region-based algorithms have incorporated multiscale decompositions such as Discrete Wavelet Transforms [7], Contourlet Transforms [8], Shearlet transforms [9], and many more, most of which will be discussed in section 2.

Machine learning and region-based algorithms are not too different, the only difference is that, the fusion rule in the former is defined using a machine learning classifier, additionally, saliency detection and classification can be further improved using machine learning techniques, fusion using genetic algorithms [11], Support Vector Machines [12], and Particle Swarm Optimization [13] comes in this category.

Similarity matching algorithms accounts for human visual system into the fusion rule, that is, it utilizes the visual features such as lines, texture, shape and other structural details of the input source images to design the fusion algorithm. Authors in [14], [15], and [16] have worked on similarity matching algorithms for multimodal and medical image fusion tasks.

Decision-level: These are high level information fusion algorithms, which are less explored in the literature. Feature-level fusion acts as a prerequisite for this level, which is followed by feature classification and building decision maps or indices for final fusion operation. The fusion operation is carried out based on the best decision, or the decision which has the highest probability of giving better classification in the end. Hence, algorithms in this level are the most accurate of all, however, the downside of these algorithms is it does not go well with the human visual system, that is, information loss is bound to happen. Decision-level fusion algorithms include voting, Bayes' inference [17, 18], fuzzy integrals [19], and many other methods. Tab.1 below inspired from [2] gives a brief summary over the performances for each level discussed above.

Table 1: Attribute performance summary

Attributes	**Pixel-level**	**Feature-level**	**Decision-level**
Information loss	Minimum	Medium	Maximum
Information content	Highest	Medium	Lowest
Method complexity	Easiest	Moderate	Hardest
Classification performance	Worst	Moderate	Best
Noise sensitivity	Highest	Medium	Lowest

2.2 Preprocessing pipeline

Medical images are acquired from multiple sensors, and it is possible that images obtained from any of the sensor have some kind of glitch or artifact in them, those glitches can be comfortably detected by a medical expert, however once the images with each of the modality is fused, those glitches (if undetected) would also get fused, and that will be rather undesirable. Hence, this stage is the most crucial step for any kind of image fusion task, therefore, source images must be preprocessed for noise and other kinds of artifacts such as spatial blurring and non-uniform illumination. For illumination related issues, image enhancement can also be carried out as a preprocessing step. A detailed survey on methods for noise reduction and image enhancement techniques is given in [20, 21].

Another prerequisite for any kind of image fusion is *image registration*, in which source images are geometrically aligned with respect to size, orientation and location of the reference source image. Figure 2 given below illustrates image registration technique as a preprocessing step for image fusion. A dense amount of theory and methods are available in the image processing and computer vision literature which deals with image registration, and is a separate field of research in its own right. A compresensive survey for traditional as well as modern day deep learning based image registration techniques can be seen in [22, 23].

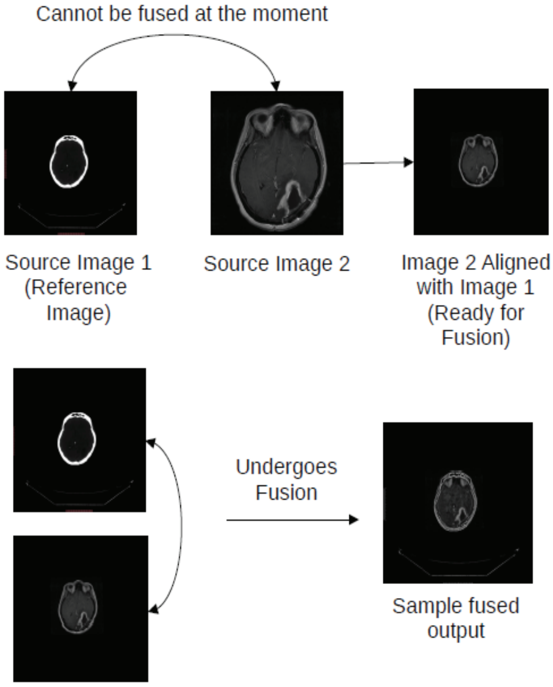

Cannot be fused at the moment
Source Image 1 (Reference Image)
Source Image 2
Image 2 Aligned with Image 1 (Ready for Fusion)
Undergoes Fusion
Sample fused output

Figure 2: Image registration as a preprocessing step for CT-MRI fusion using pixel averaging criteria

3 Literature survey and state-of-the-art

Image fusion algorithms have been in the literature for quiet some time now, hence it is imperative to first review the traditional state of the art approaches for the same, and then talk about the recent deep learning based advances in the field. Hence, this section is divided into two parts, the former is all about non-deep learning based technqiues, and latter will be on Deep Learning based advances for medical images.

3.1 Traditional techniques

When it comes to traditional state of the art techniques for medical image fusion, majority of the techniques come from: multi-scale decomposition (MSD) methods, morphological methods, and fuzzy logic based methods. Multi-Scale decomposition techniques are based on transforms such as pyramids [24], wavelets [25], contourlets [26], curvelets [27] and framelets [28]. In pyramid-based techniques for image fusion, the decomposition process involves creating a set of band-pass and low-pass versions of the original image by sequentially filtering and down sampling it, resulting in a pyramid-like structure.

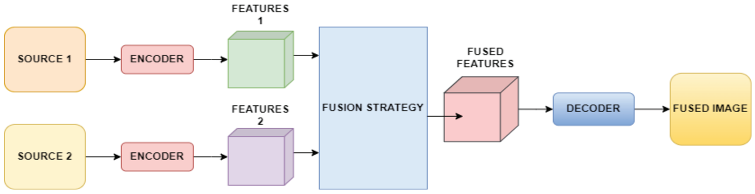
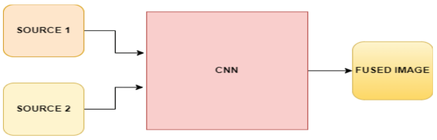
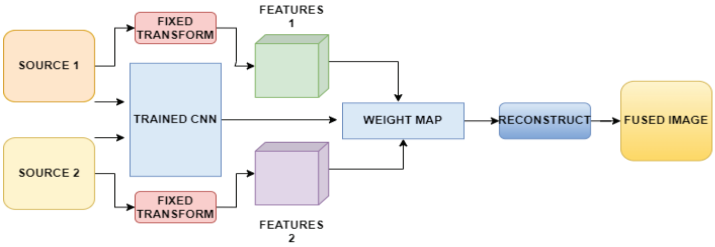
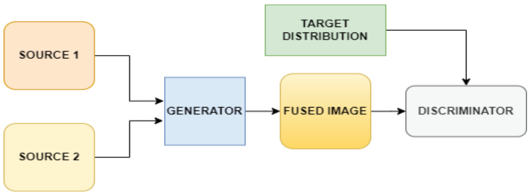
Authors in [24] have incorporated Laplacian filter for decomposition which used discrete cosine transform for data compression, and through averaging fusion rule, the authors were able to fuse CT and MRI images. Authors in [5] used gradient pyramids incorporating Gaussian filters. Similarly by varying the filters for decomposition, different pyramid based fusion techniques can be developed, for example morphological pyramids [6], ratio pyramids [29], steerable pyramids [30] and many more. In [6], a non-linear morphological approach is considered for decomposition, which shows promise when compared with linear approaches. The main issue with most of the pyramid based schemes is that it suffers from blocking effects, and edge distortion, and to overcome these problems variants of wavelet based techniques were introduced. Authors in [25] introduced gradient based discrete wavelet transform to fuse MRI-T1, T2, and FLAIR images by incorporating two separate fusion rules which involves a max and an averaging operation respectively. Authors in [31] proposed a sparse representation based method for medical image fusion in the tetrolum domain which is a new adaptive version of the Haar wavelet. The proposed scheme was able to preserve the color and contrast information, but had also introduced artifacts such as black dots. Authors in [32] proposed a medical image fusion method which incorporated principal component analysis, and had also used Intensity Hue Saturation color model to retain the color information in order to fuse MRI and PET images. The proposed method was able to retain more spatial characteristics with no color and spatial distortion, however the method was not robust enough to be tried on other types of modalities. Authors in [33] introduced a new method where the fusion coefficients were obtained using the standard deviation and density function of the shift invariant Shearlet transform (SIST). Modalities such as MRI, PET, SPECT and CT were used for the fusion process. The proposed method was able to capture both functional and spatial information quiet well. Nonsubsampled Contourlet Transform (NSCT) domain is another quiet popular domain transform technique in the medical image fusion literature, authors in [34] proposed a method in the NSCT domain where the low frequency sub band coefficient is obtained by taking the square of the maximum entropy of the coefficients within a local window. The maximum weighted sum modified laplacian is used to obtain the high frequency sub band coefficient. According to quantitative evaluations, this algorithm performs better than several existing methods and produces good contrast. Authors in [26] introduced a fusion rule for CT and MRI brain images by exploiting the major properties of the NSCT by using maximum and average masking fusion rules that were devised for the approximate and directional coefficients respectively. Authors in [35] proposed a fusion method for medical images using the curvelet transform. The method begins by converting each source image into curvelet coefficients. These coefficients are then fused using a PCA fusion rule. Finally, the inverse curvelet transform is applied to produce the final fused image.

Table 2: Summary for Traditional fusion techniques

Work	**Fusion method**	**Modalities used**	**Strengths**	**Weaknesses**
Laplacian based fusion using discrete cosine transform [24]	DCT & Laplacian Pyramids	CT & MRI	Exhibits higher edge strength & good contrast with respect to human visual systems (HVS).	Introduced blocky artifacts.
Gradient pyramids incorporating gaussian filter [5]	Fusion features calculated using local structure tensor in the gradient domain	CT & MRI	Shows superior results when compared with other pyramid methods	Unable to obtain a pristine noiseless image
Gradient based discrete wavelet transform [25]	DWT & gradient pyramids	MRI-T1, MRI-T2 & FLAIR	Improved entropy, mutual information & fusion symmetry performance.	Works well only for MRI brain scans, hence not robust.
Sparse representation in Tetrolum domain [31]	Tetrolum transform	MRI, PET & SPECT	Ensures the preservation of important details & minimal contrast loss.	Black dotted artifacts upon fusion.
Principal component analysis & IHS based fusion [32]	PCA & IHS	MRI & PET	Retains more essential data & better spatial spatial characteristics.	Narrow application scope, with notable color inaccuracies.
Fusion using shift invariant Shearlet transform [33]	Shearlet transform	MRI, PET, SPECT & CT	Obtains both functional & spatial information.	Does not include analysis of other multiscale geometric tools.
Nonsubsampled Contourlet Transform [34], [26]	NSCT	MRI & CT	Outperforms multiple existing techniques while also producing strong contrast.	Not robust, cannot be generalized to anatomical & functional medical images.
Fusion using curvelet transform [35]	Curvelet & PCA	MRI & CT	Effectively captures the curves & edges of images in its representation.	Requires more complex computation & parameter adjustment.
Fusion using fuzzy transform [36]	Fuzzy transform	MRI-T1, MRI-T2, CT & PET	Generates a fused image with improved contrast that conveys more information.	More time consuming

Authors in [36] introduced a fuzzy transform based method. The process begins by dividing the images into equal-sized blocks, which are then transformed into sub-blocks of varying sizes using the fuzzy transform. The maximum-entropy fusion rule is then applied to the sub-blocks, followed by the inverse fuzzy transform on the fused sub-blocks. The method's performance was evaluated through both subjective and objective means. Tab.2 briefly summarises the traditional non deep learning based methods discussed above.

Table 3: Different Deep Learning architectures

Autoencoder based architecture
 FEATURES 1 SOURCE 1 ENCODER FEATURES 2 FUSION STRATEGY FUSED FEATURES DECODER FUSED IMAGE SOURCE 2 ENCODER
Convolutional Neural Network based architectures

 FIXED TRANSFORM FEATURES 1 SOURCE 1 TRAINED CNN WEIGHT MAP RECONSTRUCT FUSED IMAGE SOURCE 2 FIXED TRANSFORM FEATURES 2
Generative Adversarial Network based architecture
 TARGET DISTRIBUTION SOURCE 1 GENERATOR FUSED IMAGE DISCRIMINATOR SOURCE 2

3.2 Deep learning based techniques

As the Tab.2 illustrates, traditional medical image fusion algorithms have major weaknesses, including a lack of robustness and the generation of artifacts. This raises questions about the effectiveness of the feature extraction and feature fusion processes. To address these issues, researchers have turned to deep learning (DL) based image fusion techniques. DL based techniques for image fusion are classified into four categories: autoencoder (AE)-based, convolutional neural network (CNN)-based, generative adversarial networks (GAN)-based, and transformers based architectures. Tab.3 gives a brief overview towards the image fusion process using the first three architectures mentioned.

The AE method usually pre-trains an autoencoder on a different dataset in order to learn the most precise feature extraction and reconstruction process in order to generate a fused representation in a supervised or unsupervised setting, however, AE based methods faces issues with edge distortion because it happens to loose important information while learning for latent representations of the source images. CNN based architectures are quiet flexible, as seen in Tab.3, CNN based methods can be applied in two ways, one way is to utilize just the convolutional layers to extract the useful features, and then train another set of conv layers for the fusion process. Whereas a different approach suggests to use CNNs with a transform domain technique which can inturn bring in more complexity towards the architecture and hence can extract even deeper features without having to go deep, and then use another set of conv layers for the reconstruction process, this technique is also called a hybrid method. GAN based methods incorporate an adversarial game based approach, where the game is played between a generator and a discriminator. Generator network attempts to fool the discriminator by producing fake fused images using the source images as an input, whereas the discriminator, having to know about the distribution of the real fused images attempts to put penalty on the generator whenever caught. With this generator-discriminator game, a GAN based architecture comes up with a trained network for the generator to produce realistic looking fused images. Up until recently, vision transformers have been in use for a variety of applications in computer vision, through the self-attention mechanism these models are able to solve the issues of long range dependencies and the need for augmentation put forth by CNN based models. Transformer models lets the network to learn local as well as global features of the input source images, and hence becomes very powerful for the image fusion process, no matter the modality of the source images.

In [37], the authors attempted to discontinue the manual designing of complicated fusion rules using complicated activity level measurements by the use of CNNs. They used a siamese network to propose a new framework that combines activity level measurement and weight assignment through network learning. Additionally, the authors incorporated Laplacian and Gaussian pyramid based decomposition techniques to design their fusion rule, making their scheme into a hybrid method. This approach yielded superior results when compared to several traditional state-of-the-art fusion techniques for CT, MRI and SPECT images, however, due to its simplistic fusion rule, the fused image suffered from broken edges. Authors in [38], proposed a novel encoder-decoder network for visual-infrared (VI-IR) image fusion technique (DenseFuse), their encoder module incorporated a dense network having three convolutional layers with skip connections, which made the feature learning process quicker and more precise. The decoder module used four convolutional layers with relu activation for each of them. By incorporating structural similarity based cost function, the authors were able to achieve comparable results with state-of-the art image fusion techniques. However in order to produce superior results for medical images, the same authors came up with multiscale encoder network in [39] (MSDNet), in this, they improved the encoding process by incorporating three convolutional layers of different sizes simultaneously in order to learn more complex features. With this change the authors were able to design a fusion scheme for medical, visual-infrared and multi-focus images. However, the results were just slightly improved in comparison to their previous method, hence the impact was not significant. The authors in [40] presented a novel end-to-end image fusion framework (IFCNN) that utilized the power of ResNet101 [41] pretrained over multi-focus image datasets available online. The proposed framework is a general-purpose method that can handle different types of images. The authors demonstrated its ability to elegantly fuse multi-focus, multi-modal and multi-exposure images, and had overshadowed majority of the CNN, AE and GAN based methods, however, inspight of its overwhelming performance when it came to multi-modal medical images (CT-MRI), the fused results were rather dull and were not able fully capture the important details of the source images. On the same grounds as [40], authors in [42] proposed a unified unsupervised end-to-end fusion technique for multi-focus, multi-modal and multi-exposure images. The proposed technique utilised the power of the VGG16 [43] architecture for its feature extraction process. Overall, the method produced decent results when compared with traditional fusion schemes but had issues majorly with intensity and contrast in the fused output. In [44], the authors proposed EMFusion, an enhanced unsupervised image fusion framework designed specifically for medical images. The architecture of this framework ensures the enhancement by adding constraints to both the surface and deeper levels, thereby preserving relevant features of the source images. Another encoder-decoder based fusion framework (MSENet) was proposed by authors in [45] for medical images, which had used multi-scale feature extraction process using CNNs, results obtained using this technique outperformed traditional state-of-the-art techniques, however an extensive experiment with DL based techniques was not provided. Using U-Net [46] as the backbone of the feature extraction process, authors in [47] proposed a self-supervised image fusion scheme for visible and infrared images. Their proposed architecture was able to capture relevant feature information uniformly from infrared and visible images, with that said, the proposed technique has yet to outperform the state-of-the-art GAN based techniques which are already overshadoweded by encoder-decoder based architectures.

FusionGAN [48] was one of the first GAN based image fusion framework designed to fuse multi-modal images. Inspired from the DCGAN [49] architecture, its generator uses CNNs which takes infrared and visible images as its input and produces a fused result having the descriminator to penalise the fused output if its distribution is different from the true distribution. Similarly authors in [50] and [51] came with their own taste of GAN based architectures for multi-modality images and had produced state-of-the-art results. However, authors in [52] came up with a dual stream attention based generator-descriminator architecture (DSAGAN) which outperformed most of the other generative models for medical image fusion. DSAGAN used three CNN based attention modules for its generator network in order to produce a high quality fused result, having said that, a six layer deep CNN based descriminator is used to penalise the image if found fake. DSAGAN was able to outperform most of the traditional and GAN based fusion schemes in terms of entropy and blind image quality metrics.

Though not extensively explored in the literature, transformer based image fusion architectures are rising after their overwhelming performance in terms of image classification and object detection [53]. Authors in [54] proposed a bimodal transformer based visual-infrared image fusion framework, having a complex three level architecture, where in the first level it uses multiscale feature extraction process using dense networks, in the second level it brings in two separate transformers which takes in the dense features from visual and infrared inputs respectively and prepares them for the fusion stage (level three). Level three utilizes CNNs and fast fourier transforms for the fusion to take place. With such a complicated archicture, the framework was able to outperform many traditional and deep learning based state-of-the-art techniques. Having to know this, transformer based architectures ([55], [56], [57]) are quiet heavy in terms of parameters and data required for training. Tab. 4 provides a brief summary of some of the methods discussed above.

Having to know about all the strengths and weaknesses of each of the deep learning based framworks for medical image fusion. The proposed architecture in this chapter aims to unify the strengths of most of the algorithms discussed so far, and also aims to mitigate the weaknesses put forth by each of the method. Notable weaknesses shown by most of the methods are the following:

- Encoder-Decoder based architectures: Most of the results suffered from broken edges or were unable to capture the true intensity and contrast information from the source images.

- Purely convolutional neural network based architectures: Having to train the network over databases different from medical images makes the network learn features which are not useful, and hence results in a fused outcome which is sometimes different from the original source images cand lead to ambiguity.

- Generative adversarial network based architectures: These networks are sensitive to change in hyperparameters. They also require data in abundance for training, which becomes a challenge when just medical data is concerned.

- Transformer based architectures: Though not explored in detail, the architectures of these frameworks are rather complex and the results obtained so far are comparable with encoder-decoder and/or CNN based architectures.

Table 4: Image Fusion survey for Deep learning algorithms

Architectures/ methods	**Modalities**	**Strengths**	**Shortcomings**
DenseFuse [38]	VI, IR	Results comparable to state of the art techniques	Poor performance for medical images
MSDNet [39]	VI, IR, CT, MRI, SPECT	Improvement upon DenseFuse	Results not compared with enough DL based methods.
IFCNN [40]	VI, IR, CT, MRI, SPECT	Was able to capture textural features well.	Unable to capture contrast details.
EMFusion [44] MSENet [45]	CT, MRI, PET, SPECT	Unsupervised multi--scale feature extra--ction.	Extensive experiments not provided. Poor contrast.
FusionGAN [48] DDcGAN [50] DSAGAN [52]	VI, IR, CT, MRI, PET, SPECT	Outperforms most of CNN and traditional based methods.	Sensitive to hyper--parameters, cannot generalize well.
TransFuse [56] THFuse [57]	CT, MRI VI, IR	Maintains long range dependencies and can capture local and global features well.	Requires a lot of training data. Time inefficient.

4 Proposed framework

Upon knowing about the strengths and weaknesses of traditional and deep learning based image fusion techniques. It is noteworthy to understand the significance of activity level measurement. In traditional image fusion techniques, activity level measurements are taken by first transforming the source images into a different domain and then coming up with a mathematical formulation for relevant feature extraction, and then measuring for levels of activity for each pixel in the source images using a quantifier. These activity levels helps in obtaining the best possible fused image. In deep learning based techniques the activity level measurement can be made using features extracted automatically with the help of convolutional neural networks. And from the techniques discussed in the previous section, it can be concluded that using multiscale feature extraction process turns out to be the most optimal method for extracting relevant features without having to go deep.

The proposed architecture is an unsupervised deep learning model for multimodality medical image fusion which contains the following key components:

- Siamese multi-scale feature extraction module which is able to learn more accurate features even if the training data is significantly less.
- Uses long and short skip connections for smooth training and long range dependencies.
- Uses partial reference image quality metric as the training loss function in order to capture textural and contrast information well.

The network contains eight trainable layers which are divided into three modules: (a)Siamese multiscale feature extraction, (b)Feature fusion and (c)Reconstruction. In order to solve the issues put forth by various deep learning based techniques, it is worth mentioning that the architecture aims to incorporate the following things:

- Long range dependencies without having a complicated architecture
- Little dependence on hyperparameters.
- Generalizes well.
- Have the ability to learn significant features with fewer layers.

4.1 Feature extraction process

The feature extraction module of the proposed architecture in Fig.3 is inspired by the multi-scale decomposition process used in the traditional medical image fusion algorithms. The module is comprised of a Siamese CNN architecture having two identical subnetworks with shared weights that takes as an input two multimodality images. Instead of relying on a deeper architecture with numerous convolutional layers, the module relies heavily on learning multi-scale feature representations which in turn helps in capturing low and high level features simultaneously [58]. To capture these features, kernels of size 5×5 and

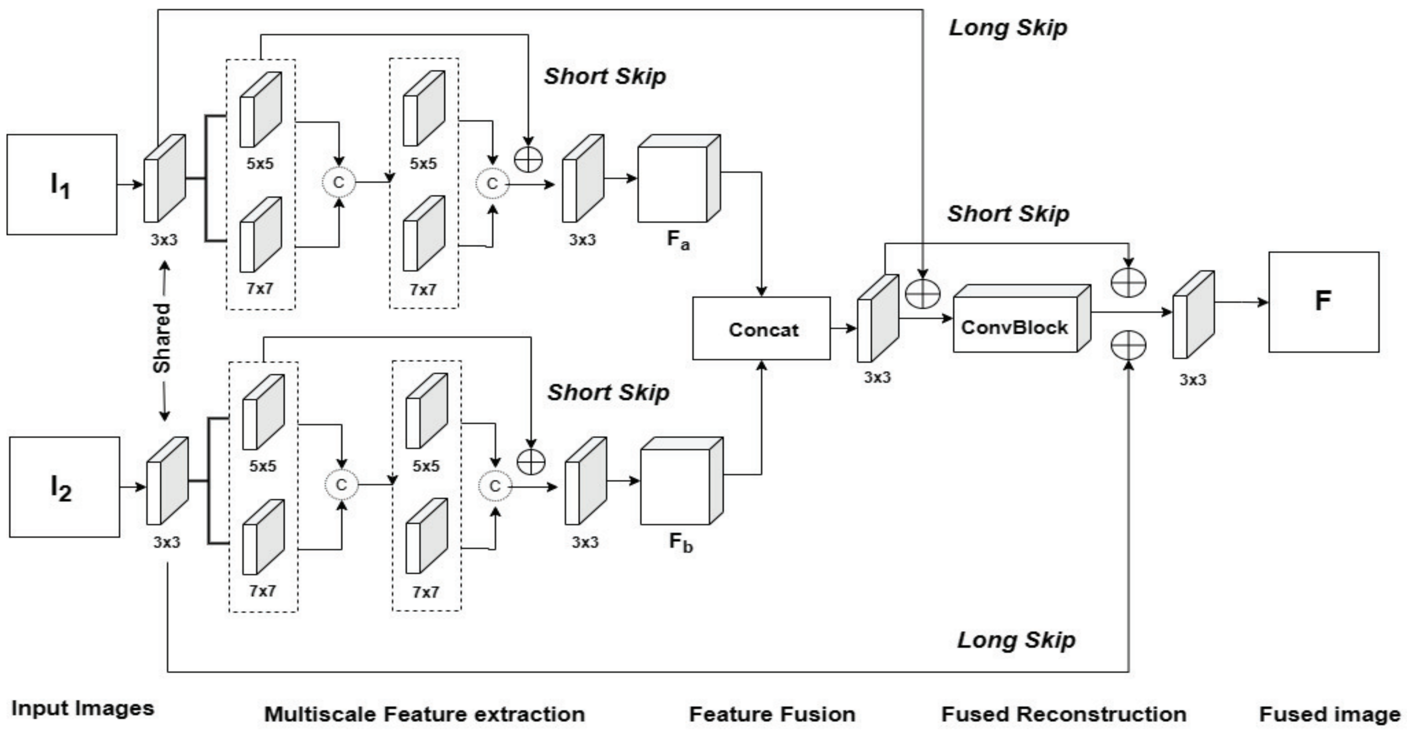


Figure 3: Proposed Fusion Architecture

7×7 are used respectively. All convolutional layers are followed by batch normalization having rectified linear unit activation in the end. In order to initialize and re-map the feature extraction process, the multi-scale layers are preceded and followed by mono scale convolutions having kernels of size 3×3. The module also supports short skip connections which provides an uninterrupted gradient flow throughout the network. The mono scale convolutions alongside batch normalization and ReLU can be expressed as:

$$A_{i+1} = max\{0, \mathcal{BN}(W_i * A_i)\}$$

Where W_i and $\mathcal{BN}$ represents the convolutional kernel and Batch normalization respectively. A_{i+1} is the output feature map of the $i+1^{th}$ convolutional layer, and the symbol $'*'$ represents the convolution operation. Likewise, for the multi-scale feature extraction, each mappings from 5×5 and 7×7 are extracted simultaneously, which can be expressed as:

$$\begin{aligned} a^1_{i+1} &= max\{0, \mathcal{BN}(W^1_i * A_i)\} \\ a^2_{i+1} &= max\{0, \mathcal{BN}(W^2_i * A_i)\} \end{aligned}$$

Where a^l_i is the output mapping at the l^{th} scale of i^{th} convolutional layer, and W^l_i is the l^{th} scale convolutional kernel. Subsequently, the mappings are concatenated for further processing, which is expressed as follows:

$$A^k_{i+1} = Concat(a^1_{i+1}, a^2_{i+1})$$

Where A^k_{i+1} represents the concatenated representation of the k^{th} multi-scale convolutional layer. For the purpose of illustration Fig. 4 represents the multi-scale block of the feature extraction process. Note that, during the feature extraction process pooling operation is not considered, as experiments suggests that it is actually removing the essential information from the source images whereby unabling to reconstruct the fused image. Also, padding of appropriate size is considered in order to keep the sizes of the feature maps same as that of the source images. Unit strides are taken in each of the convolutions for the whole architecture.

4.2 Feature fusion and Reconstruction

The feature fusion process is fairly straight forward, similar to the concatenation step during multi-scale extraction, the $Concat(F_a, F_b)$ operation completes the feature fusion process, where F_a and F_b are the final feature representations of the multi-scale extraction process as illustrated in Fig.3.

The Reconstruction phase of the proposed architecture comprises of a set of convolution operations having short and long skip connections in order to output the final fused image. It must be noted that deconvolutions or transposed convolutions was not used for this step since appropriate amount of padding was performed in each step of the feature extraction phase, and hence applying transposed convolutions was actually inhibiting the model to reconstruct a well

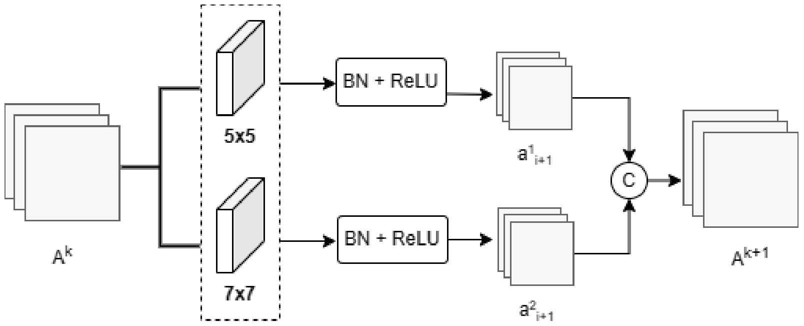


Figure 4: Proposed multi-scale fusion block.

defined fused image. The ”ConvBlock” in the architecture consists of a sequence of two back to back convolution operations of size 3 × 3 each, which are preceded and followed by 3 × 3 kernels, having short and long skip connections in between to smoothen the reconstruction process. The long skip connections are the key to making the proposed model robust, as it takes care of the essential features brought forth by each of the imaging modality.

4.3 Implementation Details

The multi-scale structural similarity index (MS-SSIM) is a modified version of the structural similarity index (SSIM) [59] that takes into account the structural information of an image at multiple scales. SSIM is a widely used image similarity metric that measures the similarity between two images by comparing their luminance, contrast, and structural information. While SSIM is effective at measuring the similarity between two images at a single scale, it can be sensitive to small changes in the images that may not be perceptually significant. MS-SSIM was developed to address this limitation by considering the structural information of the images at multiple scales, which allows it to better capture the perceived similarity between the images. In a patch P, SSIM for the center pixel $\tilde{p}$ can be expressed as:

$$\mathrm{SSIM}(\tilde{p}) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 c_2)} = l(\tilde{p}).cs(\tilde{p})$$

Where c_1 and c_2 are small constants that are used to stabilize the division and prevent the denominators from becoming too small, x, y are sliding windows in the reference image and source images respectively, with mean as μ_x, variance as σ_x^2, and the covariance of x and y is denoted as σ_{xy}.
Given a dyadic pyramid of K levels MS-SSIM can be defined as:

$$\text{MS-SSIM}(\tilde{p}) = l_K^{\alpha}(\tilde{p}).\prod_{j=1}^{K} cs_j^{\beta_j}(\tilde{p})$$

In this chapter, the researchers explore the use of MS-SSIM as a loss function for the task of multimodal medical image fusion. Loss functions are an essential component of machine learning models, as they define the measure of how well the model is able to learn from the training data. The choice of loss function can have a significant impact on the performance and convergence of the model. For patch P its center pixel $\tilde{p}$, the MS-SSIM loss is evaluated as:

$$\mathcal{L}^{\text{MS-SSIM}}(P) = 1 - \text{MS-SSIM}(\tilde{p})$$

Further details regarding MS-SSIM can be furnished from [60]. To evaluate the performance of the model, the researchers used MS-SSIM as the primary loss function, along with mean squared error loss, which can be termed as pixel loss, given by:

$$\mathcal{L}_{\text{pixel}}(F) = \sum_{i=1}^{n} \text{MSE}(I_i, F)$$

Where I_i are the input source images and F is the fused image respectively. The researchers found that using MS-SSIM as the primary loss function resulted in better fusion results. In comparison, using the pixel loss alone as the loss function resulted in intensity distortions and had introduced artifacts as well. The final loss function is the sum of both of these losses.

$$\mathcal{L} = \lambda \mathcal{L}_{\text{MS-SSIM}}(F) + \mathcal{L}_{\text{pixel}}(F)$$

where the value of λ is chosen to be 500 for faster convergence. Overall, the findings suggest that MS-SSIM is a promising loss function for multimodal medical image fusion and warrants further investigation.

The major limitation for multimodal medical image fusion is the lack of training data and ground truth availability. The proposed architecture is therefore designed in such a way that even if it is trained on fewer images, the model is able to capture relevant features to perform a high quality fusion task, which was possible due to the multi-scale feature extraction and the long-short residual connections [61]. Data used for training and testing is openly available on Harvard Brain Atlas [62]. The model was trained on 180 image pairs of CT and MRI brain scans. Each image is resized to 200×200 for training. One-cycle learning rate schedular [63] is incorporated with a maximum learning rate of 0.01. The model was trained using adam optimizer [64] for 100 epochs with a batch size of 40. Implementation was done over Google colaboratory environment and the model is built using the pytorch framework [65].

5 Experimental results and discussions

5.1 Evaluation setup

Performance of image fusion techniques is evaluated using various criteria such as entropy, mutual information, and fusion symmetry. The choice of criteria depends on the fusion task and the type of fused image. Based on the mentioned criterias, dicussed below are few image quality metrics that are used for comparative analysis.

(a) Normalized Mutual Information ($\mathcal{Q}_{\text{MI}}$): Mutual information (MI) is a statistical measure that quantifies the dependence between two variables. It is often used to measure the amount of information shared by two images. The mathematical definition of MI for two discrete random variables X and Y is as follows:

$$\text{MI}(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)}$$

Note that, $p(x, y)$ is the joint probability distribution of X and Y, with $p(x)$ and $p(y)$ being the marginal distributions of X and Y respectively. Suppose the source images are given by I_1 and I_2, with the fused image as F, then the normalized mutual information [66] is given by:

$$\mathcal{Q}_{\text{MI}} = 2\{\frac{\text{MI}(I_1, F)}{E(I_1) + E(F)} + \frac{\text{MI}(I_2, F)}{E(I_2) + E(F)}\}$$

where $E(.)$ is the entropy. Range of $\mathcal{Q}_{\text{MI}}$ lies from 0 to 1, 0 indicating poor fusion quality while 1 meaning perfectly fused image.

(b) Feature mutual Information ($\mathcal{Q}_{\text{FMI}}$): Feature mutual information is a metric used to measure the similarity between fused image and reference image. It is used to evaluate the quality of image fusion and ensures that important features of the original images are preserved in the fused image. Mathematical details regarding the metric can be seen from [67]. Note that $\mathcal{Q}_{\text{FMI}} \in [0, 1]$, and higher score indicates better fusion results.

(c) Edge based measure ($\mathcal{Q}_{\text{AB/F}}$): It measures the total information transference of the source images to the fused image using edge information. It is defined as:

$$\mathcal{Q}^{\text{AB/F}} = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} [Q^{\text{AF}} w_{i,j}^{x} + Q^{\text{BF}} w_{i,j}^{y}]}{\sum_{i=1}^{M} \sum_{j=1}^{N} [w_{i,j}^{x} + w_{i,j}^{y}]}$$

where A, B and F being the source and fused images respectively. Notice, Q^{AF} and Q^{BF} are edge preserving values which are weighted by w^x and w^y. Range of $\mathcal{Q}^{AB/F}$ lies in $[0, 1]$, where values closer to 0 indicates less information is transferred and values close to 1 indicates higher information gain. Additional details about the metric can be obtained from [68]

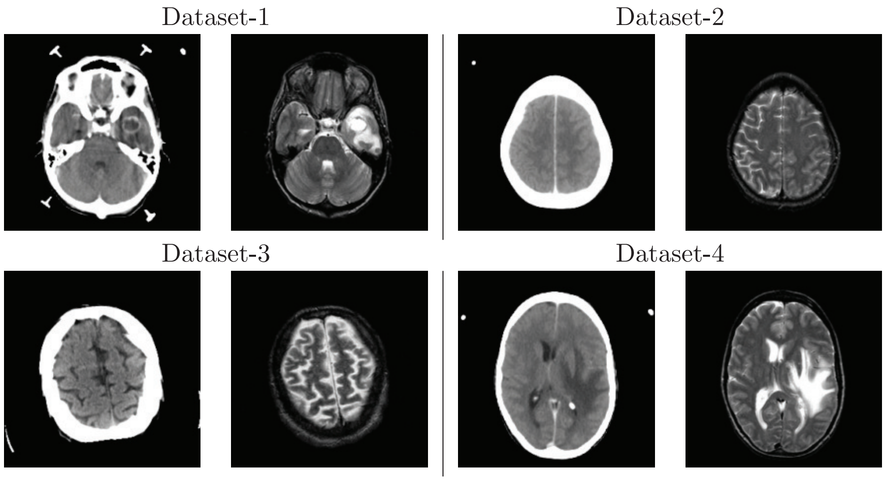
(d) Structural similarity based metric ($\mathcal{Q}_{\text{S}}$): An alternative method for using SSIM in image fusion evaluation is presented in [69], which is based on the traditional definition of SSIM and is defined as:

$$\mathcal{Q}_{\mathrm{S}} = \left\{ \begin{matrix} \lambda(w)\mathrm{SSIM}(\mathrm{A{,}F}|w) + (1-\lambda(w))\mathrm{SSIM}(\mathrm{B{,}F}|w), \\ \text{if } \mathrm{SSIM}(\mathrm{A{,}B}|w) \geq 0.75 \\ \\ \max[\mathrm{SSIM}(\mathrm{A{,}F}|w), \mathrm{SSIM}(\mathrm{B{,}F}|w)], \\ \text{if } \mathrm{SSIM}(\text{A,F—}w) < 0.75 \end{matrix} \right\}$$

where w is a window of odd size that scans from top left to bottom right with unit steps and $\lambda(w)$ is weight obtained from the local image features. A, B are the source images and F is the fused image. Similar to other metrics, $\mathcal{Q}_{\mathrm{S}} \in [0, 1]$, 0 indicating poor fusion quality, while 1 indicating perfectly fused image.

The indices/metrics mentioned above uses the input and fused images in order to evaluate the performance of the fusion algorithm, and hence these metrics can be termed as partial reference image quality metrics because the ground truth is unavailable. Moreover, the literature is further extended to no-reference image quality metrics as well. These metrics incorporates the knowledge of human visual systems, and uses statistical methods to mathematically quantify the perceptual quality of the image. One such metric used for the analysis is Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [70]. The goal of employing blind image quality metrics is to predict the naturalness of the fused images without using input or ground-truth images as reference. The metric is designed to produce a positive numerical result, with a smaller score indicating better visual quality. This allows for the comparison of the proposed result with other techniques.

Table 5: Experimental Dataset: CT(left)-MRI(right) image pairs used for comparison



5.2 Results and Discussions

To reveal the strengths of the proposed medical image fusion framework, a variety of experiments were conducted using the evaluation metrics mentioned above, and the results obtained were compared to five state-of-the-art image fusion techniques, three of which are non deep learning based which uses Stationary Wavelet Transform [71], Laplacian Pyramids [72] and Non-Subsampled Contourlet Transform [73], the remaining two are deep learning based techniques one of which relies heavily on the VGG19 [74] architecture and the other is the IFCNN [40] method. The superiority of the results obtained is discussed both subjectively and objectively in the following sections. For the purpose of evaluation, brain scans of CT and MRI modalities are chosen, the image pairs are registered beforehand, and are freely available online [62]. Table 5 represents the set of image pairs used for objective experimentation.

5.2.1 Qualitative Evaluation

Evaluating the fused results through a qualitative approach is of paramount importance, as it ensures that the images not only meet the technical requirements but also the standards of the human visual system. The proposed architecture resulted in visually pleasing fused images upon training. As shown in Table 6 and 7, the fused images not only retained texture details but also maintained contrast information from the input source images, which makes sure that if there exists an anomaly(tumor/lesion) in any of the source images, then it will be captured in the fused image as well. On the contrary results produced by LRD and NSCT based approaches are overexposed, while SWT and IFCNN have rather dull appearence. Results given by the VGG19 based methods are also visually appealing and appears to have captured most of the relevant information.

5.2.2 Quantitative Evaluation

As can be seen from Table 8, the results obtained from the proposed architecture have shown dominance in terms of the blind image quality metric ($\mathcal{Q}_{\mathrm{BRISQUE}}$) and information transference ($\mathcal{Q}^{\mathrm{AB/F}}$), which further clarifies the claims of generalizability and long range dependency. Additionally, it must also be noted that NSCT and VGG19 based method have dominated in terms of mutual information ($\mathcal{Q}_{\mathrm{MI}}$, $\mathcal{Q}_{\mathrm{FMI}}$) and structural similarity ($\mathcal{Q}_{\mathrm{S}}$), however, after VGG19 based model, the proposed results have the overall best performance as compared with the other four methods. Given that IFCNN had already outperformed several of the GAN and CNN based models, its results are far inferior when compared with the proposed architecture.

Table 6: Fusion Results of the proposed framework

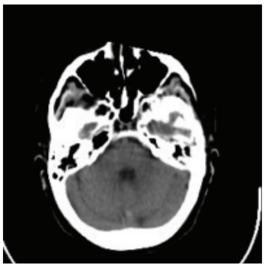
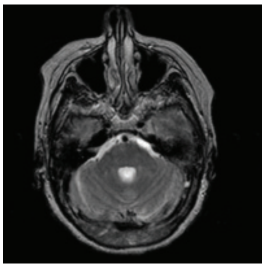
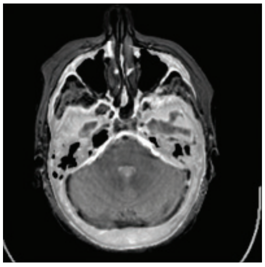
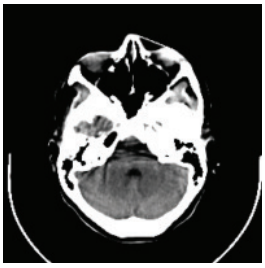
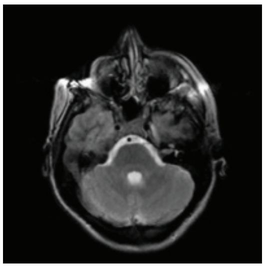
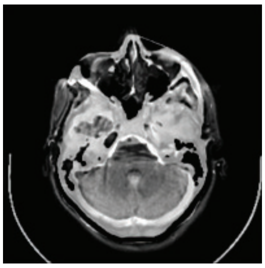
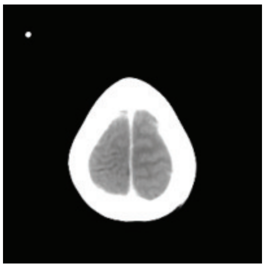
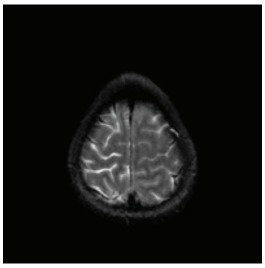
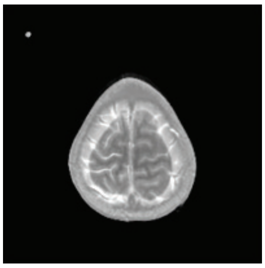
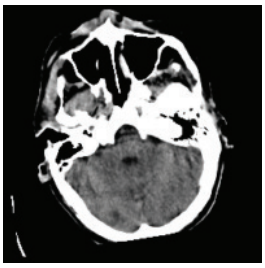
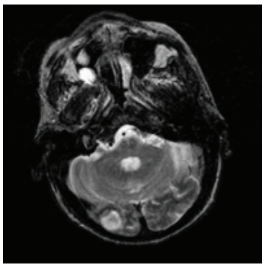
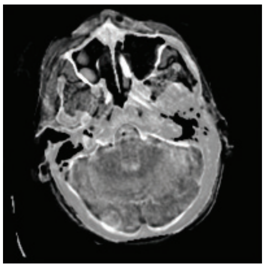
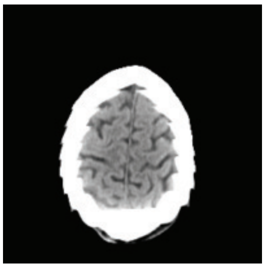
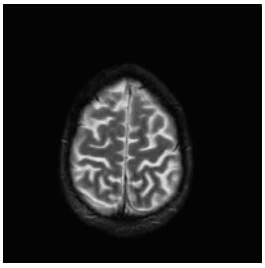
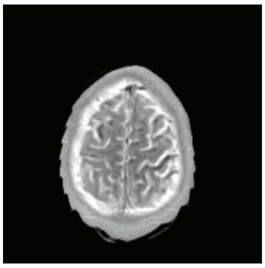
CT	MRI	Fused Images
		
		
		
		
		

Table 7: Subjective Quality evaluation

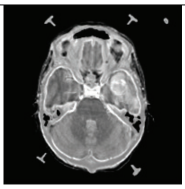
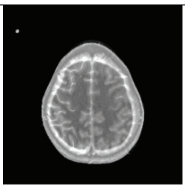
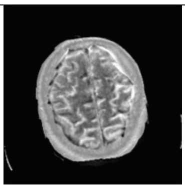
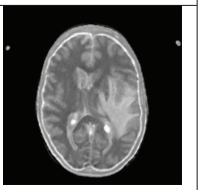
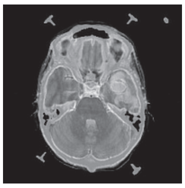
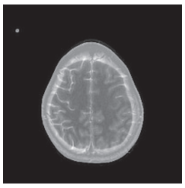
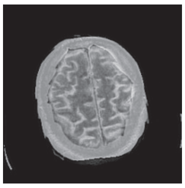
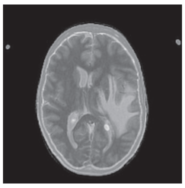
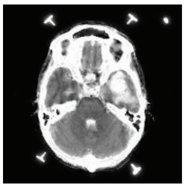
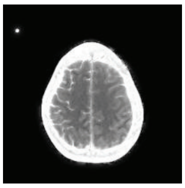
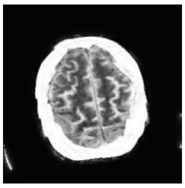
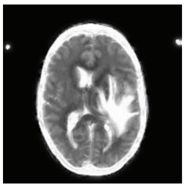
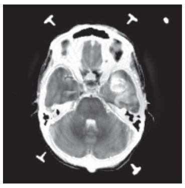
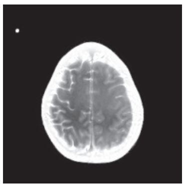
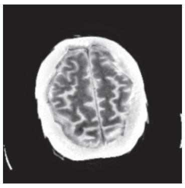
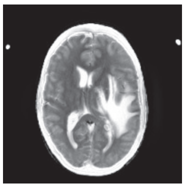
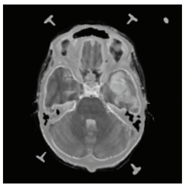
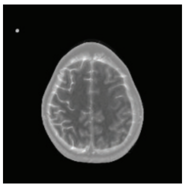
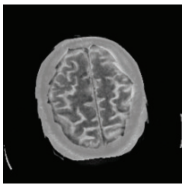
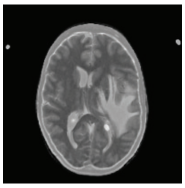
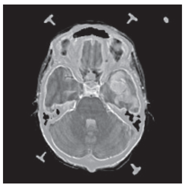
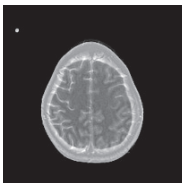
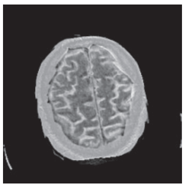
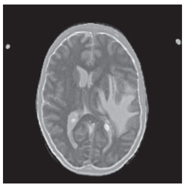
	Dataset-1	Dataset-2	Dataset-3	Dataset-4
Proposed				
SWT [61]				
LRD [62]				
NSCT [63]				
VGG19 [64]				
IFCNN [39]				

Table 8: Objective quality evaluation

		$\mathcal{Q}_{\text{MI}}$	$\mathcal{Q}_{\text{FMI}}$	$\mathcal{Q}_{\text{AB/F}}$	$\mathcal{Q}_{\text{S}}$	$\mathcal{Q}_{\text{BRISQUE}}$
Dataset-1	Proposed	0.8118	0.8367	**0.7190**	0.9957	**34.5174**
	SWT	0.8157	0.8453	0.6780	0.9961	40.6390
	LRD	0.7860	0.8521	0.5880	0.9930	35.5579
	NSCT	0.8375	**0.8639**	0.6670	0.9936	38.8441
	VGG19	**0.8773**	0.8546	0.7040	**0.9962**	41.2680
	IFCNN	0.6385	0.7715	0.5519	0.9925	42.5134
Dataset-2	Proposed	0.9428	0.8909	**0.8149**	0.9963	**47.1250**
	SWT	0.9669	0.8933	0.8000	**0.9970**	50.7880
	LRD	0.9336	0.8952	0.7101	0.9933	48.6970
	NSCT	0.9725	**0.8971**	0.7552	0.9935	49.4562
	VGG19	**0.9990**	0.8960	0.8089	0.9970	49.5364
	IFCNN	0.7159	0.8238	0.6211	0.9924	50.9619
Dataset-3	Proposed	0.8730	0.8601	**0.7601**	0.9947	**43.4630**
	SWT	0.8710	0.8738	0.7265	0.9955	46.6320
	LRD	0.8840	0.8756	0.6398	0.9901	44.4210
	NSCT	0.9094	**0.8850**	0.6776	0.9908	44.5331
	VGG19	**0.9109**	0.8775	0.7378	**0.9956**	46.3750
	IFCNN	0.8999	0.8688	0.7668	0.9953	47.5811
Dataset-4	Proposed	0.8424	0.8781	**0.7131**	0.9970	**36.7379**
	SWT	0.8658	0.8820	0.7064	0.9973	42.4839
	LRD	0.8203	0.8852	0.6313	0.9955	41.7659
	NSCT	0.8644	**0.8953**	0.6903	0.9960	41.8448
	VGG19	**0.9245**	0.8882	0.7061	**0.9974**	43.0376
	IFCNN	0.6101	0.7889	0.5586	0.9921	43.8322

6 Conclusion

Image fusion is a technique that combines multiple images with similar content but different information to create a single high-quality image. In this chapter, the authors provide an extensive review of various traditional and recent deep learning-based state-of-the-art techniques for multimodality image fusion. Upon reviewing, it was pointed out, that deep learning based techniques have outgrown majority of the traditional image fusion methods. It was also pointed out that when it came to medical image fusion deep learning based techniques suffered from issues such as long range dependency, senstivity to hyperparameters, poor generalizability and learning insignificant features. All these issues were addressed by proposing a novel unsupervised multi-scale siamese architecture utilizing the power of convolutional neural networks for medical image fusion. Upon comparing the proposed architecture with traditional and modern image fusion methods, it was concluded that the proposed scheme excels in terms of contrast, texture and information preservation. Although there is still room for improvement, it is believed that the proposed model can be extended to a robust general-purpose fusion framework for images of other modalities as well.

References

[1] Du, J., Li, W., & Xiao, B. (2018). Fusion of anatomical and functional images using parallel saliency features. Inf. Sci., 430, 567-576.

[2] Hermessi, H., Mourali, O., & Zagrouba, E. (2021). Multimodal medical image fusion review: Theoretical background and recent advances. Signal Process., 183, 108036.

[3] Gang Xiao , Durga Prasad Bavirisetti , Gang Liu , Xingchen Zhang, Image Fusion, Springer Singapore, ISBN 978-981-15-4867-3, 2020, doi: https://doi.org/10.1007/978-981-15-4867-3

[4] A. Kahol and G. Bhatnagar, “A new multi-focus image fusion framework based on focus measures,” 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2021, pp. 2083-2088, doi: 10.1109/SMC52423.2021.9659111.

[5] Z. Jin , Y. Wang , Z. Chen , S. Nie , Medical image fusion in gradient domain with structure tensor, J Med Imaging Health Inform 6 (5) (2016) 1314–1318 .

[6] G.K. Matsopoulos , S. Marshall , J.N.H. Brunt , Multiresolution morphological fu- sion of MR and CT images of the human brain, IEE Proceedings - Vision, Im- age and Signal Processing 141 (3) (1994) 137–142.

[7] Yang, Yong. “A Novel DWT Based Multi-focus Image Fusion Method.” Procedia Engineering 24 (2011): 177-181.

[8] Li, T., & Wang, Y. (2011). Biological image fusion using a NSCT based variable-weight method. Inf. Fusion, 12, 85-92.

[9] Jin, X., Chen, G., Hou, J., Jiang, Q., Zhou, D., & Yao, S. (2018). Multimodal sensor medical image fusion based on nonsubsampled shearlet transform and S-PCNNs in HSV space. Signal Process., 153, 379-395.

[10] Meher, B., Agrawal, S., Panda, R., & Abraham, A. (2019). A survey on region based image fusion methods. Inf. Fusion, 48, 119-132.

[11] A. Mumtaz, A. Majid and A. Mumtaz, “Genetic Algorithms and its application to image fusion,” 2008 4th International Conference on Emerging Technologies, Rawalpindi, Pakistan, 2008, pp. 6-10, doi: 10.1109/ICET.2008.4777465.

[12] C. Heng, L. Jie and Z. Weile, “A Novel Support Vector Machine-Based Multifocus Image Fusion Algorithm,” 2006 International Conference on Communications, Circuits and Systems, Guilin, China, 2006, pp. 500-504, doi: 10.1109/ICCCAS.2006.284686.

[13] Yuan Gao, Shiwei Ma, Jingjing Liu, Yanyan Liu, Xianxia Zhang, Fusion of medical images based on salient features extraction by PSO optimized fuzzy logic in NSST domain, Biomedical Signal Processing and Control, Volume 69, 2021, 102852, ISSN 1746-8094, https://doi.org/10.1016/j.bspc.2021.102852.

[14] Z. -S. Xiao and C. -X. Zheng, “Medical Image Fusion Based on the Structure Similarity Match Measure,” 2009 International Conference on Measuring Technology and Mechatronics Automation, 2009, pp. 491-494, doi: 10.1109/ICMTMA.2009.558.

[15] M. M. Rahman, B. C. Desai and P. Bhattacharya, “A Feature Level Fusion in Similarity Matching to Content-Based Image Retrieval,” 2006 9th International Conference on Information Fusion, 2006, pp. 1-6, doi: 10.1109/ICIF.2006.301664.

[16] Zhizhong Fu, Yufei Zhao, Yuwei Xu, Lijuan Xu, Jin Xu, Gradient structural similarity based gradient filtering for multi-modal image fusion, Information Fusion, Volume 53, 2020, Pages 251-268, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2019.06.025

[17] VJ. Kittler, Multi-Sensor Integration and Decision Level Fusion (The Institution of Electrical Engineers, London)

[18] B. Jeon, D.A. Landgrebe, Decision fusion approach for multitemporal classification. IEEE Trans. Geosci. Remote Sens. 37(3), 1227–1233 (1999)

[19] Y. Yang, J. Wu, S. Huang, Y. Fang, P. Lin and Y. Que, “Multimodal Medical Image Fusion Based on Fuzzy Discrimination With Structural Patch Decomposition,” in IEEE Journal of Biomedical and Health Informatics, vol. 23, no. 4, pp. 1647-1660, July 2019, doi: 10.1109/JBHI.2018.2869096.

[20] A. Ravishankar, S. Anusha, H. K. Akshatha, A. Raj, S. Jahnavi and J. Madhura, “A survey on noise reduction techniques in medical images,” 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2017, pp. 385-389, doi: 10.1109/ICECA.2017.8203711.

[21] K. Koonsanit, S. Thongvigitmanee, N. Pongnapang and P. Thajchayapong, “Image enhancement on digital x-ray images using N-CLAHE,” 2017 10th Biomedical Engineering International Conference (BMEiCON), Hokkaido, Japan, 2017, pp. 1-4, doi: 10.1109/BMEiCON.2017.8229130.

[22] A. Gholipour, N. Kehtarnavaz, R. Briggs, M. Devous and K. Gopinath, “Brain Functional Localization: A Survey of Image Registration Techniques,” in IEEE Transactions on Medical Imaging, vol. 26, no. 4, pp. 427-451, April 2007, doi: 10.1109/TMI.2007.892508.

[23] Y. Zhang, Z. Zhang, G. Ma and J. Wu, “Multi-Source Remote Sensing Image Registration Based on Local Deep Learning Feature,” 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 2021, pp. 3412-3415, doi: 10.1109/IGARSS47720.2021.9553142.

[24] A. Sahu, V. Bhateja, A. Krishn and Himanshi, ”Medical image fusion with Laplacian Pyramids,” 2014 International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom), 2014, pp. 448-453, doi: 10.1109/MedCom.2014.7006050.

[25] B. Deepa, M. G. Sumithra, T. D. Bharathi and S. Rajesh, “MRI Medical Image Fusion Using Gradient Based Discrete Wavelet Transform,” 2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2017, pp. 1-4, doi: 10.1109/ICCIC.2017.8524436.

[26] T. J. Reddy and S. N. Rao, “A novel fusion approach for multimodal medical images using Non-Subsampled contourlet transform,” 2016 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), 2016, pp. 838-841, doi: 10.1109/ICACCCT.2016.7831757.

[27] Himanshi , V. Bhateja , A . Krishn , A . Sahu , Medical image fusion in curvelet do- main employing PCA and maximum selection rule, in: :Satapathy S., Raju K., Mandal J., Bhateja V. (eds) Proceedings of the Second International Confer- ence on Computer and Communication Technologies. Advances in Intelligent Systems and Computing, 379, Springer, India, 2016, pp. 1–9 .

[28] Bhatnagar, G., & Wu, Q.M. (2012). An Image Fusion Framework Based on Human Visual System in Framelet Domain. Int. J. Wavelets Multiresolution Inf. Process., 10.

[29] M. Li , Y. Dong ,Review of image fusion algorithm based on multiscale decom- position, in: :Proceedings 2013 International Conference on Mechatronic Sci- ences, Electric Engineering and Computer (MEC), IEEE, 2013, pp. 1422–1425.

[30] H. Deng and Y. Ma, “Image fusion based on steerable pyramid and PCNN,” 2009 Second International Conference on the Applications of Digital Information and Web Technologies, London, UK, 2009, pp. 569-573, doi: 10.1109/ICADIWT.2009.5273861.

[31] Shahdoosti, H.R., & Mehrabi, A. (2018). Multimodal image fusion using sparse representation classification in tetrolet domain. Digit. Signal Process., 79, 9-22.

[32] Q. Liu, C. He, H. Li, H. Wang, Multimodal medical image fusion based on IHS and PCA, Proc. Eng. 7 (2010) 280–285.

[33] Wang, L., Li, B., & Tian, L. (2014). Multi-modal medical image fusion using the inter-scale and intra-scale dependencies between image shift-invariant shearlet coefficients. Inf. Fusion, 19, 20-28.

[34] Ganasala, P., & Kumar, V. (2014). CT and MR Image Fusion Scheme in Nonsubsampled Contourlet Transform Domain. Journal of Digital Imaging, 27, 407-418.

[35] P Rn, Desai U, Shetty VB (2014) Medical image fusion analysis using curvelet transform. In: Int. Conf. on Adv. in Comp., Comm., and Inf. Sci. (ACCIS-14). pp 1–8

[36] Manchanda, M., & Sharma, R. (2016). A novel method of multimodal medical image fusion using fuzzy transform. J. Vis. Commun. Image Represent., 40, 197-217.

[37] Y. Liu, X. Chen, J. Cheng and H. Peng, "A medical image fusion method based on convolutional neural networks," 2017 20th International Conference on Information Fusion (Fusion), Xi'an, China, 2017, pp. 1-7, doi: 10.23919/ICIF.2017.8009769.

[38] H. Li and X. -J. Wu, "DenseFuse: A Fusion Approach to Infrared and Visible Images," in IEEE Transactions on Image Processing, vol. 28, no. 5, pp. 2614-2623, May 2019, doi: 10.1109/TIP.2018.2887342.

[39] Song, X., Wu, XJ., Li, H. (2019). MSDNet for Medical Image Fusion. In: Zhao, Y., Barnes, N., Chen, B., Westermann, R., Kong, X., Lin, C. (eds) Image and Graphics. ICIG 2019. Lecture Notes in Computer Science(), vol 11902. Springer, Cham. https://doi.org/10.1007/978-3-030-34110-7_24

[40] Yu Zhang, Yu Liu, Peng Sun, Han Yan, Xiaolin Zhao, Li Zhang, IFCNN: A general image fusion framework based on convolutional neural network, Information Fusion, Volume 54, 2020, Pages 99-118, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2019.07.011.

[41] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.

[42] H. Xu, J. Ma, J. Jiang, X. Guo and H. Ling, "U2Fusion: A Unified Unsupervised Image Fusion Network," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 1, pp. 502-518, 1 Jan. 2022, doi: 10.1109/TPAMI.2020.3012548.

[43] Simonyan, Karen and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." CoRR abs/1409.1556 (2014): n. pag.

[44] Xu, Han and Jiayi Ma. “EMFusion: An unsupervised enhanced medical image fusion network.” Inf. Fusion 76 (2021): 177-186.

[45] Weisheng Li, Ruyue Li, Jun Fu, Xiuxiu Peng, MSENet: A multi-scale enhanced network based on unique features guidance for medical image fusion, Biomedical Signal Processing and Control, Volume 74, 2022, 103534, ISSN 1746-8094.

[46] Ronneberger, Olaf, Philipp Fischer and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation.” ArXiv abs/1505.04597 (2015): n. pag.

[47] X. Lin, G. Zhou, W. Zeng, X. Tu, Y. Huang and X. Ding, “A Self-Supervised Method for Infrared and Visible Image Fusion,” 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 2022, pp. 2376-2380, doi: 10.1109/ICIP46576.2022.9897731.

[48] Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, Junjun Jiang, FusionGAN: A generative adversarial network for infrared and visible image fusion, Information Fusion, Volume 48, 2019, Pages 11-26, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2018.09.004.

[49] Radford, Alec, Luke Metz and Soumith Chintala. “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks.” CoRR abs/1511.06434 (2015): n. pag.

[50] J. Ma, H. Xu, J. Jiang, X. Mei and X. -P. Zhang, “DDcGAN: A Dual-Discriminator Conditional Generative Adversarial Network for Multi-Resolution Image Fusion,” in IEEE Transactions on Image Processing, vol. 29, pp. 4980-4995, 2020, doi: 10.1109/TIP.2020.2977573.

[51] Yicheng Wang, Shuang Xu, Junmin Liu, Zixiang Zhao, Chunxia Zhang, Jiangshe Zhang, MFIF-GAN: A new generative adversarial network for multi-focus image fusion, Signal Processing: Image Communication, Volume 96, 2021, 116295, ISSN 0923-5965, https://doi.org/10.1016/j.image.2021.116295.

[52] Jun Fu, Weisheng Li, Jiao Du, Liming Xu, DSAGAN: A generative adversarial network based on dual-stream attention mechanism for anatomical and functional image fusion, Information Sciences, Volume 576, 2021, Pages 484-506, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2021.06.083.

[53] C. -F. R. Chen, Q. Fan and R. Panda, “CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification,” 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 347-356, doi: 10.1109/ICCV48922.2021.00041.

[54] S. Park, A. G. Vien and C. Lee, “Infrared and Visible Image Fusion Using Bimodal Transformers,” 2022 IEEE International Conference on

Image Processing (ICIP), Bordeaux, France, 2022, pp. 1741-1745, doi: 10.1109/ICIP46576.2022.9897993.

[55] W. Tang, F. He, Y. Liu and Y. Duan, "MATR: Multimodal Medical Image Fusion via Multiscale Adaptive Transformer," in IEEE Transactions on Image Processing, vol. 31, pp. 5134-5149, 2022, doi: 10.1109/TIP.2022.3193288.

[56] Qu, Linhao, Shaolei Liu, Manning Wang, Shiman Li, Siqi Yin, Qin Qiao and Zhijian Song. "TransFuse: A Unified Transformer-based Image Fusion Framework using Self-supervised Learning." ArXiv abs/2201.07451 (2022): n. pag.

[57] Jun Chen, Jianfeng Ding, Yang Yu, Wenping Gong, THFuse: An infrared and visible image fusion network using transformer and hybrid feature extractor, Neurocomputing, Volume 527, 2023, Pages 71-82, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2023.01.033.

[58] Hafiz Tayyab Mustafa, Jie Yang, Masoumeh Zareapoor, Multi-scale convolutional neural network for multi-focus image fusion, Image and Vision Computing, Volume 85, 2019, Pages 26-35, ISSN 0262-8856, https://doi.org/10.1016/j.imavis.2019.03.001.

[59] Zhou, W., A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. "Image Quality Assessment: From Error Visibility to Structural Similarity." IEEE Transactions on Image Processing. Vol. 13, Issue 4, April 2004, pp. 600–612.

[60] Wang, Zhou, Eero P. Simoncelli and Alan Conrad Bovik. "Multiscale structural similarity for image quality assessment." The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003 2 (2003): 1398-1402 Vol.2.

[61] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778

[62] Shenton, M. E., Kikinis, R., McCarley, W., Saiviroonporn, P., Hokama, H. H., Robatino, A., Jolesz, F. A. Harvard brain atlas: a teaching and visualization tool. Proceedings 1995 Biomedical Visualization. https://doi.org/10.1109/BIOVIS.1995.528700

[63] Smith, L.N., & Topin, N. (2017). Super-convergence: very fast training of neural networks using large learning rates. Defense + Commercial Sensing.

[64] Kingma, D.P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. CoRR, abs/1412.6980.

[65] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner,

B., Fang, L., Bai, J., & Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. ArXiv, abs/1912.01703.

[66] Hossny, M., Nahavandi, S., & Creighton, D.C. (2008). Comments on 'Information measure for performance of image fusion'. Electronics Letters, 44, 1066-1067.

[67] M. Haghighat and M. A. Razian, "Fast-FMI: Non-reference image fusion metric," 2014 IEEE 8th International Conference on Application of Information and Communication Technologies (AICT), Astana, Kazakhstan, 2014, pp. 1-3, doi: 10.1109/ICAICT.2014.7036000.

[68] C Xydeas and V Petroviä, "Objective Image Fusion Performance Measure", Electronic Letters, vol. 36, no. 4, pp. 308-309, 2000.

[69] Piella, G., & Heijmans, H.J. (2003). A new quality metric for image fusion. Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429), 3, III-173.

[70] Mittal, A., Moorthy, A.K., & Bovik, A.C. (2012). No-Reference Image Quality Assessment in the Spatial Domain. IEEE Transactions on Image Processing, 21, 4695-4708.

[71] Prakash, Om and Ashish Khare. "CT and MR Images Fusion Based on Stationary Wavelet Transform by Modulus Maxima." (2015).

[72] Li, X., Guo, X., Han, P., Wang, X., Li, H., & Luo, T. (2020). Laplacian Redecomposition for Multimodal Medical Image Fusion. IEEE Transactions on Instrumentation and Measurement, 69, 6880-6890.

[73] Z. Zhu, M. Zheng, G. Qi, D. Wang and Y. Xiang, "A Phase Congruency and Local Laplacian Energy Based Multi-Modality Medical Image Fusion Method in NSCT Domain," in IEEE Access, vol. 7, pp. 20811-20824, 2019

[74] Li, Hui, Xiaojun Wu and Josef Kittler. "Infrared and Visible Image Fusion using a Deep Learning Framework." 2018 24th International Conference on Pattern Recognition (ICPR) (2018): 2705-2710.