Journal of Engineering Sciences
Vol 13, Issue 05, May /2022
ISSN NO:0377-9254

CNN based multi modal medical image fusion with classification using YOLO-V2

Saradha Rani Sabbava[1], Sai Praneeth Boddapati [2], Tanuja Dasari[3], Naveen Bollam[4].
1 M.Tech, Assistant Professor, Dept. of ECE, Gitam Deemed to be University, Rushikonda, Andhra Pradesh, India
2,3,4, B. Tech, Dept of ECE, Gitam Deemed to be University, Rushikonda, Andhra Pradesh.

Abstract: Various multimodalities in the medical realm, such as computed tomography (CT) and magnetic resonance imaging (MRI), are combined to provide a fused image. Medical image fusion is a technique for preserving crucial information by combining all relevant information from numerous images into a single fused image generated by using convolutional neural network (CNN). In this work, brain CT and MRI images are fused together and will be classified as normal/abnormal using YOLO-V2. If the network is detected as abnormal, the part of the tumor region is localized. These works are done by using Convolutional Neural Network, YOLO-V2 architecture and image processing techniques. Experimental results demonstrate that the proposed method can achieve promising results in terms of both visual quality and objective assessment.

Keywords: Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Convolutional Neural Network, YOLO.

1. Introduction

With the rapid development of sensor and computer technology, medical imaging has emerged as an irreplaceable component in various clinical applications including diagnosis, treatment planning and surgical navigation. To provide medical practitioners sufficient information for clinical purposes, medical images obtained with multiple modalities are usually required, such as X-ray, computed tomography (CT), magnetic resonance (MR), positron emission tomography (PET), single photon emission computed tomography (SPECT), etc.

Due to the difference in imaging mechanism, medical images with different modalities focus on different categories of organ/tissue information. For instance, the CT images are commonly used for the precise localization of dense structures like bones and implants, the MR images can provide excellent soft-tissue details with high-resolution anatomical information, while the functional information on blood flow and metabolic changes can be offered by PET and SPECT images but with low spatial resolution. Multi-modal medical image fusion aims at combining the complementary information contained in different source images by generating a composite

image for visualization, which can help physicians make easier and better decisions for various purposes [1].

In recent years, a variety of medical image fusion methods have been proposed [2]–[17]. Due to the difference in imaging mechanism, the intensities of different source images at the same location often vary significantly.

For this reason, most of these fusion algorithms are introduced in a multi-scale manner to pursue perceptually good results. In general, these multi-scale transform (MST)-based fusion methods consist of three steps, namely, decomposition, fusion and reconstruction. Multi-scale transforms which are frequently studied in image fusion include pyramids [17]–[19], wavelets [9], [20], [21], multi-scale geometrical transforms like contourlet and shearlet [2], [6], [10], [16]. In image fusion research, sparse representation is another popular image modelling approach, which has also been successfully applied to fuse multi-modal medical images [4], [5], [15], [22].

One of the most crucial issues in image fusion is calculating a weight map which integrates the pixel activity information from different sources. In most existing fusion methods, this target is achieved by two steps known as activity level measurement and weight assignment. In conventional transform domain fusion methods, the absolute value of a decomposed coefficient (or the sum of those values within a small window) is employed to measure its activity, and then a "choose-max" or "weighted-average" fusion rule is applied to assign weights to different sources based on the obtained measurement. Clearly, this kind of activity measurement and weight assignment are usually not very robust resulting from many factors like noise, mis-registration and the difference between source pixel intensities.

To improve the fusion performance, many complex decomposition approaches and elaborate weight assignment strategies have been recently proposed in the literature [6], [8]–[13], [15], [16]. However, it is actually not an easy task to design a ideal activity level measurement or weight assignment strategy which can comprehensively take all the key issues of fusion into account. Moreover, these two steps are designed individually without a strong association by many fusion methods, which may greatly limit the algorithm performance. In this paper, this issue is addressed from another viewpoint to overcome the difficulty in designing robust activity level measurements and weight assignment strategies. Specifically, a convolutional neural network (CNN) [23] is trained to encode a direct mapping from source images to the weight map.

In this way, the activity level measurement and weight assignment can be jointly achieved in an "optimal" manner via learning network parameters. Considering the different imaging

modalities of multi-modal medical images, we adopt a multi-scale approach via image pyramids to make fusion process more consistent with human visual perception. In addition, a local similarity based strategy is applied to adaptively adjust the fusion mode for the decomposed coefficients of source images.

The rest of this paper is organized as follows. In Section 2, some related work and the basic idea of the proposed fusion method are introduced. Section 3 presents the fusion method in detail. Experimental results and discussions are provided in Section 4. Finally, Section 5 concludes the paper.

2. literature survey:

In our recent work [24], a CNN-based multi-focus image fusion method which can obtain state-of-the-art results was proposed. In the method, two source images are fed to the two branches of a siamese convolutional network in which the two branches share the same architecture and weights [25], respectively. Each branch contains three convolutional layers and the obtained feature maps essentially act as the role of activity level measures. The feature maps of two branches are concatenated and then pass through two fully-connected layers (they are converted into equivalent convolutional layers in the fusion process to allow arbitrary input size [26]), which can be viewed as the weight assignment part of a fusion method. As a result, the value of each coefficient in the network output map indicates the focus property of a pair of source image patches at a corresponding location. By assigning the value as the weights of all the pixels within the patch location and then averaging the overlapped pixels, a focus map with the same size of source images is generated. The final fused image is obtained based on the focus map using the weighted-average rule along with two consistency verification techniques [20]. In [24], the feasibility and superiority of CNNs used for image fusion have been explicitly presented. Please refer to [24] for more details. The target of this paper is to extend the CNN model to medical image fusion. However, the method proposed in [24] cannot be directly used to fuse medical images primarily due to the following two reasons. 1) The fusion method [24] is performed in spatial domain. As medical images are obtained with different imaging modalities, transform domain based fusion methods are more likely to produce results with less undesirable artifacts for their good consistency with human visual perception. 2) The two inputs of the CNN model in [24] are assumed to have similar local structures. As mentioned above, the intensities of multi-modal medical images at the same location often vary significantly, so the assumption of local similarity between two source images is not always valid. To address the first problem, we apply a pyramid-based multi-scale approach [27] to pursue perceptually better results. Specifically, each source image is

decomposed into a Laplacian pyramid while the weight map obtained from the network is decomposed into a Gaussian pyramid. The fusion procedure is conducted at every decomposition level.

For the second issue, we adopt a local similarity-based fusion strategy to determine the fusion mode for the decomposed coefficients [18]. When the contents of source images have high similarity, the "weighted-average" fusion mode is applied to avoid losing useful information. In this situation, the weights obtained by the CNN is more reliable than the coefficient-based measure, so they are employed as the merging weights. When the similarity of image contents is low, the "choose-max" or "selection" fusion mode is preferred to mostly preserve the salient details from source images. In this situation, the CNN output is not reliable and the pixel activity is directly measured by the absolute values of the decomposed coefficients. Based on the above ideas, the CNN model presented in [24] can be applied to the fusion of medical images.

It is worthwhile to note that both the pyramid-based decomposition and the similarity-based fusion mode determination are just "naive" techniques which are commonly-used in the field of image fusion. Nevertheless, it will be demonstrated that a reasonable usage of these techniques incorporated with the CNN model can result in state-of-the-art fusion performance.

3. Proposed Method

Image Fusion is the process of combining information from two or more images of the same scene taken at the same instant or at different instants to provide more detailed images than the individual images separately. The image fusion techniques involve pixel-based methods, decision-based methods and feature based methods. A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. The pre-processing required in a ConvNet is much lower as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, ConvNets have the ability to learn these filters/characteristics.

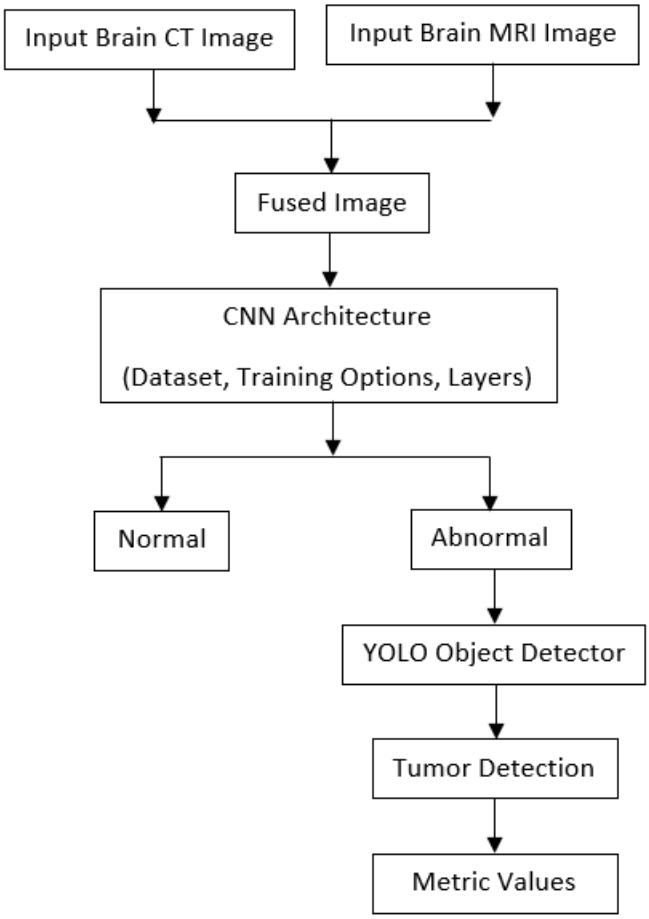


Fig. 1. Block Diagram of Proposed Method.

The 512 feature maps after concatenation are directly connected to a 2-dimensional vector. It can be calculated that the slight mode only takes up about 1.66 MB of physical memory in single precision, which is significantly less than the 33.6 MB model employed in [24]. Finally, this 2- dimensional vector is fed to a 2-way softmax layer (not shown in Fig. 1), which produces a probability distribution over two classes. The two classes correspond two kinds of normalized weight assignment results, namely, “first patch 1 and second patch 0” and “first patch 0 and second patch 1”, respectively. The probability of each class indicates the possibility of each weight assignment. In this situation, also considering that the sum of two output probabilities is 1, the probability of each class just indicates the weight assigned to its corresponding input patch. The network is trained by high-quality image patches and their blurred versions using the approach in [24]. In the training process, the spatial size of the input patch is set to 16 × 16 according to the analysis in [24].

The creation of training examples are based on multi-scale Gaussian filtering and random sampling. The softmax loss function is employed as the optimization objective and we adopt the stochastic gradient descent (SGD) algorithm to minimize it. The training process is operated

on the popular deep learning framework Caffe [28]. Please refer to [24] for the details of example generation and network training. Since the network has a fully-connected layer that have fixed dimensions (pre-defined) on input and output data, the input of the network must have a fixed size to ensure that the input data of a fully-connected layer is fixed. In image fusion, to handle source images of arbitrary size, one can divide the images into overlapping patches and input each patch pair into the network, but it will introduce a large number of repeated calculations.

To solve this problem, we first convert the fully-connected layer into a equivalent convolutional layer containing two kernels of size $8 \times 8 \times 512$ [26]. After the conversion, the network can process source images of arbitrary size as a whole to generate a dense prediction map, in which each prediction (a 2-dimensional vector) contains the relative clarity information of a source patch pair at the corresponding location. As there are only two dimensions in each prediction and their sum is normalized to 1, the output can be simplified as the weight of the first (or second) source. Finally, to obtain a weight map with the same size of source images, we assign the value as the weights of all the pixels within the patch location and average the overlapped pixels.

3.2 YOLO

In YOLO-V2 the details of each block in the visualization can be seen by hovering over the block. Each Convolution block has the BatchNorm normalization and then Leaky Relu activation except for the last Convolution block. YOLO divides the input image into an **S×S** grid. Each grid cell predicts only **one** object.

For each grid cell,

- it predicts **B** boundary boxes and each box has one **box confidence score**,
- it detects **one** object only regardless of the number of boxes B,
- it predicts **C conditional class probabilities** (one per class for the likeliness of the object class).

The boundary boxes contain box confidence score. The confidence score reflects how likely the box contains an object (**objectless**) and how accurate is the boundary box. We normalize the bounding box width *w* and height *h* by the image width and height. *x* and *y* are offsets to the corresponding cell. Hence, *x*, *y*, *w* and *h* are all between 0 and 1. Each cell has 20 conditional class probabilities. The **conditional class probability** is the probability that the detected object belongs to a particular class (one probability per category for each cell). The class confidence score for each prediction box is computed as:

class confidence score = box confidence score * conditional class probability

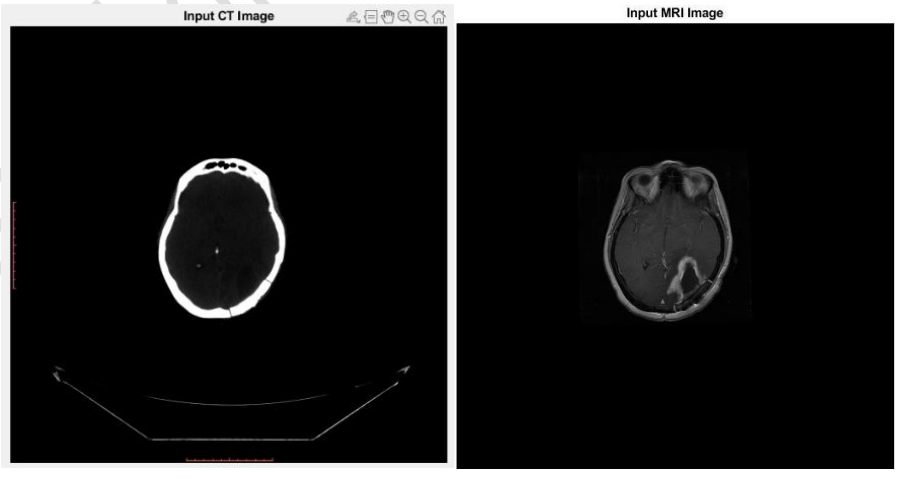
It measures the confidence on both the classification and the localization (where an object is located). We may mix up those scoring and probability terms easily. Here are the mathematical definitions for your future reference. YOLO predicts multiple bounding boxes per grid cell. To compute the loss for the true positive, we only want one of them to be **responsible** for the object. For this purpose, we select the one with the highest IoU (intersection over union) with the ground truth. This strategy leads to specialization among the bounding box predictions. Each prediction gets better at predicting certain sizes and aspect ratios.

YOLO uses sum-squared error between the predictions and the ground truth to calculate loss. The loss function composes of:

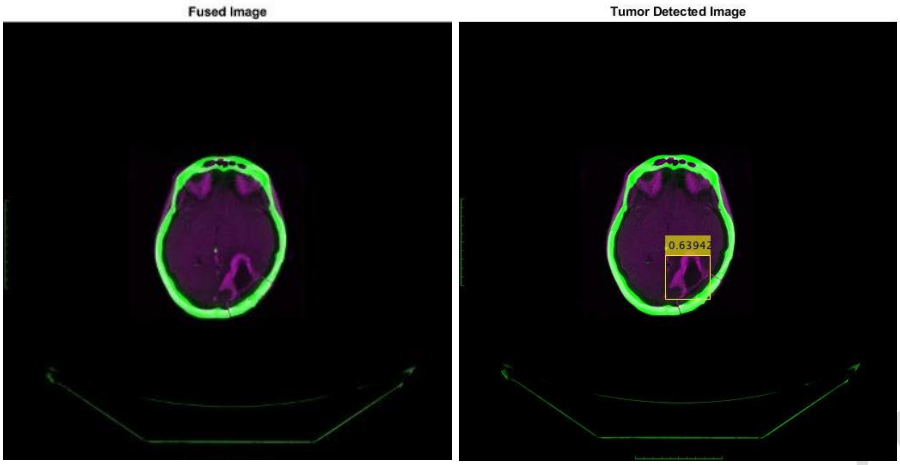
- the **classification loss**.
- the **localization loss** (errors between the predicted boundary box and the ground truth).
- the **confidence loss** (the objectness of the box).

4. Simulation Results

All the experiments have been done in MATLAB 2016b version under the high-speed CPU conditions for faster running time with test images. Aim of any fusion algorithm is to integrate required information from both source images in the output image. Fused image cannot be judged exclusively by seeing the output image or by measuring fusion metrics. It should be judged qualitatively using visual display and quantitatively using fusion metrics. In this section, The objective of any fusion algorithm is to generate a qualitative fused image. For better quality, fused image should have optimal values for all these metrics. The fusion metric with best value is highlighted in bold letter.



(a) (b)



(c) (d)

Figure 2. Fusion outcome (a) Input CT Image, (b) Input MRI Image, (c) Fused Image, (d) Tumour Detected Image

However, all the existing fusion methods outputs not good at visual perception, lack of contrast with edge information and texture preservation. Our proposed method with 3 different datasets which are presented in figure 2 looks more quality in visualization, good contrast with proper edge information and excellent texture preservation as the value of entropy is much higher.

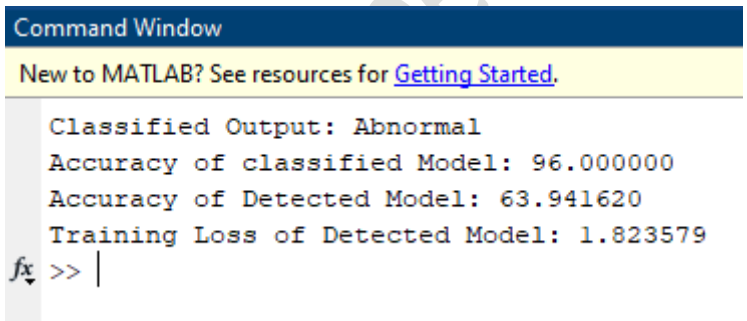


Figure 3. Performance outcomes.

5. Conclusion

The main goal of this work is to design efficient automatic brain tumour classification and detection with high accuracy, performance and low complexity. In the conventional brain tumour classification is performed by using segmentation, texture and shape feature extraction and SVM classifier which takes high computational time and high complexity. Further to improve the performance and to reduce the computation time, a convolution neural network-based classification and YOLO based detection is introduced in the proposed scheme. Also, the classification results are given as tumour (abnormal) or normal brain images. Accuracy and training loss metrics are evaluated and produces best adequate outputs when compared to existing works.

References

[1]. A. James and B. Dasarathy, "Medical image fusion: a survey of the state of the art," Information Fusion, vol. 19, pp. 4–19, 2014.

[2]. L. Yang, B. Guo, and W. Ni, "Multimodality medical image fusion based on multiscale geometric analysis of contourlet transform," Neurocomputing, vol. 72, pp. 203–211, 2008.

[3]. Z. Wang and Y. Ma, "Medical image fusion using m-pcnn," Information Fusion, vol. 9, pp. 176–185, 2008.

[4]. B. Yang and S. Li, "Pixel-level image fusion with simultaneous orthogonal matching pursuit," Information Fusion, vol. 13, pp. 10–19, 2012.

[5]. S. Li, H. Yin, and L. Fang, "Group-sparse representation with dictionary learning for medical image denoising and fusion," IEEE Transactions on Biomedical Engineering, vol. 59, pp. 3450–3459, 2012.

[6]. Z. L. G. Bhatnagar, Q. Wu, "Directive contrast based multimodal medical image fusion in nsct domain," IEEE Transactions on Multimedia, vol. 15, pp. 1014–1024, 2013.

[7]. S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," IEEE Transactions on Image Processing, vol. 22, no. 7, pp. 2864–2875, 2013.

[8]. R. Shen, I. Cheng, and A. Basu, "Cross-scale coefficient selection for volumetric medical image fusion," IEEE Transactions on Biomedical Engineering, vol. 60, pp. 1069–1079, 2013.

[9]. R. Singh and A. Khare, "Fusion of multimodal medical images using daubechies complex wavelet transform c a multiresolution approach," Information Fusion, vol. 19, pp. 49–60, 2014.

[10]. L. Wang, B. Li, and L. Tan, "Multimodal medical volumetric data fusion using 3-d discrete shearlet transform and global-to-local rule," IEEE Transactions on Biomedical Engineering, vol. 61, pp. 197–206, 2014.

[11]. Z. Liu, H. Yin, Y. Chai, and S. Yang, "A novel approach for multimodal medical image fusion," Expert Systems with Applications, vol. 41, pp. 7425–7435, 2014.

[12]. G. Bhatnagar, Q. Wu, and Z. Liu, "A new contrast based multimodal medical image fusion framework," Neurocomputing, vol. 157, pp. 143– 152, 2015.

[13]. Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," Information Fusion, vol. 24, no. 1, pp. 147–164, 2015.

[14]. Q. Wang, S. Li, H. Qin, and A. Hao, "Robust multi-modal medical image fusion via anisotropic heat diffusion guided low-rank structural analysis," Information Fusion, vol. 26, pp. 103–121, 2015.

[15]. Y. Liu and Z. Wang, "Simultaneous image fusion and denosing with adaptive sparse representation," IET Image Process., vol. 9, no. 5, pp. 347–357, 2015.

[16]. Y. Yang, Y. Que, S. Huang, and P. Lin, "Multimodal sensor medical image fusion based on type-2 fuzzy logic in nsct domain," IEEE Sensors Journal, vol. 16, pp. 3735–3745, 2016.

[17]. [17] J. Du, W. Li, B. Xiao, and Q. Nawaz, "Union laplacian pyramid with multiple features for medical image fusion," Neurocomputing, vol. 194, pp. 326–339, 2016.

[18]. [18] P. Burt and R. Kolczynski, "Enhanced image capture through fusion," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 1993, pp. 173–182.

[19]. [19] A. Toet, "A morphological pyramidal image decomposition," Pattern Recognition Letters, vol. 9, no. 4, pp. 255–261, 1989.

[20]. [20] H. Li, B. Manjunath, and S. Mitra, "Multisensor image fusion using the wavelet transform," Graphical Models and Image Processing, vol. 57, no. 3, pp. 235–245, 1995.

[21]. [21] J. Lewis, R. OCallaghan, S. Nikolov, D. Bull, and N. Canagarajah, "Pixel- and region-based image fusion with complex wavelets," Information Fusion, vol. 8, no. 2, pp. 119–130, 2007.

[22]. [22] Y. Liu, X. Chen, R. Ward, and Z. Wang, "Image fusion with convolutional sparse representation," IEEE Signal Processing Letters, vol. 23, no. 12, pp. 1882–1886, 2016.

[23]. [23] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based leaning applied to document recognition," Proceedings of The IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[24]. [24] Y. Liu, X. Chen, J. Cheng, H. Peng, and Z. Wang, "Multi-focus image fusion with a deep convolutional neural network," Information Fusion, vol. 36, pp. 191–207, 2017.

[25]. [25] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4353–4361.

[26]. [26] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. L, "Overfeat: Integrated recognition, localizaton and detection using convolutional networks," arXiv, vol. 1312.6299v4, pp. 1–16, 2014.

[27]. [27] T. Mertens, J. Kautz, and F. V. Reeth, "Exposure fusion," in Proc. Pacific Graphics, 2007, pp. 382–390.

[28]. [28] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding,” in Proceedings of the ACM International Conference on Multimedia, 2014, pp. 675–678.

[29]. [29] M. Haghighat, A. Aghagolzadeh, and H.Seyedarabi, “A non-reference image fusion metric based on mutual information of image features,” Computers and Electrical Engineering, vol. 37, pp. 744–756, 2011.

[30]. C. S. Xydeas and V. S. Petrovic, “Objective image fusion performance measure,” Electronics Letters, vol. 36, no. 4, pp. 308–309, 2000.

[31]. G. Piella and H. Heijmans, “A new quality metric for image fusion,” in Proceedings of 10th International Conference on Image Processing, 2003, pp. 173–176.

[32]. Y. Han, Y. Cai, Y. Cao, and X. Xu, “A new image fusion performance metric based on visual information fidelity,” Information Fusion, vol. 14, pp. 127–135, 2013.

[33]. http://www.med.harvard.edu/AANLIB/.