Zero-Learning Fast Medical Image Fusion

Fayez Lahoud, Sabine Süsstrunk
School of Computer and Communication Sciences
École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
{fayez.lahoud, sabine.susstrunk}@epfl.ch

***Abstract*—Clinical applications, such as image-guided surgery and noninvasive diagnosis, rely heavily on multi-modal images. Medical image fusion plays a central role by integrating information from multiple sources into a single, more understandable output. We propose a real-time image fusion method using pre-trained neural networks to generate a single image containing features from multi-modal sources. The images are merged using a novel strategy based on deep feature maps extracted from a convolutional neural network. These feature maps are compared to generate fusion weights that drive the multi-modal image fusion process. Our method is not limited to the fusion of two images, it can be applied to any number of input sources. We validate the effectiveness of our proposed method on multiple medical fusion categories. The experimental results demonstrate that our technique achieves state-of-the-art performance in both visual quality, objective assessment, and runtime efficiency.**

***Index Terms*—Image fusion, multi-modal medical fusion, neural networks, real time**

I. Introduction

Due to diverse imaging technologies, medical devices can capture different organ and tissue characteristics. For example, computed tomography (CT) captures dense structures and bones while magnetic resonance imaging (MRI) detects soft tissues. Thanks to these multi-modal visualizations, physicians can produce accurate and reliable diagnosis. However, sequential analysis of multi-modal images is inconvenient. Multi-modal image *fusion* allows physicians to visualize complementary data in a single image. The composite results are thus useful in clinical applications such as image-guided surgery, radiotherapy, patient diagnosis and treatment planning [1]–[3].

A variety of medical image fusion models have been proposed [3]–[12] over the past decade. Most of these methods consist of three steps, namely decomposition, fusion, and reconstruction. As such, the two main aspects affecting the fusion quality are the image transform and the fusion rule, i.e., a pixel level decision map. Typically, the map is generated via an activity level measurement, followed by a weight map assignment based on it. Due to the commonly used image decompositions [8], [13], [14], activity level measurements are not robust to noise, mis-registrations, nor the dynamic range of the sources. It is still difficult to design an activity level function that accounts for all the fusion requirements without limiting the algorithm performance. We address this by using convolutional neural networks (CNN) to design a robust and efficient activity level measurement and weight map generation model.

Recently, deep networks have been employed in multiple fusion problems such as multi-focus fusion [15], multi-exposure fusion [16], and visible and infrared fusion [17]. In fact, neural networks can be considered as feature extractors themselves, where intermediate maps represent salient features that can be used to reconstruct a fused image. However, although deep learning methods often achieve better performance than their classical counterparts, they still have major drawbacks. For instance, deep networks often generalize based on the datasets they are trained on, e.g., a network trained for multi-focus image fusion will only be suitable for that task. Additionally, these methods require large specialized datasets for training. Finally, neural networks, and especially convolutional neural networks (CNN), require large amounts of memory and are computationally expensive both in time and energy.

In contrast, our method is based on pre-trained neural networks. Consequently, it requires neither a specific dataset for training nor a particular network architecture. As illustrated in Fig. 1, the network is fed source images, and their intermediate layer representations are used to generate activity levels, which are compared to generate fusion weight maps. The weight maps are then refined and used to construct the fused image. We study the effect of layer depth on the quality of the fusion, and demonstrate that early layers are sufficient to obtain high quality fusion. This is an order of magnitude faster and more memory efficient than state-of-the-art techniques. Finally, our method is modality agnostic, i.e., it performs well on different input sources. We show its applicability to computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), and single photon emission computed tomography (SPECT). We also show bi- and tri-modal image fusion.

Our main contribution is a novel technique to integrate pre-trained neural networks in the image fusion task. Our method shows better performance than current state-of-the-art techniques and runs in real time without any specialized hardware. The low computational requirements make it very beneficial for continuous monitoring systems, and for deployment on

We thankfully acknowledge the support of the Hasler Foundation (grant no. 16076, S.A.V.E.) for this work.

Code available at https://github.com/IVRL/Zero-Learning-Medical-Fusion

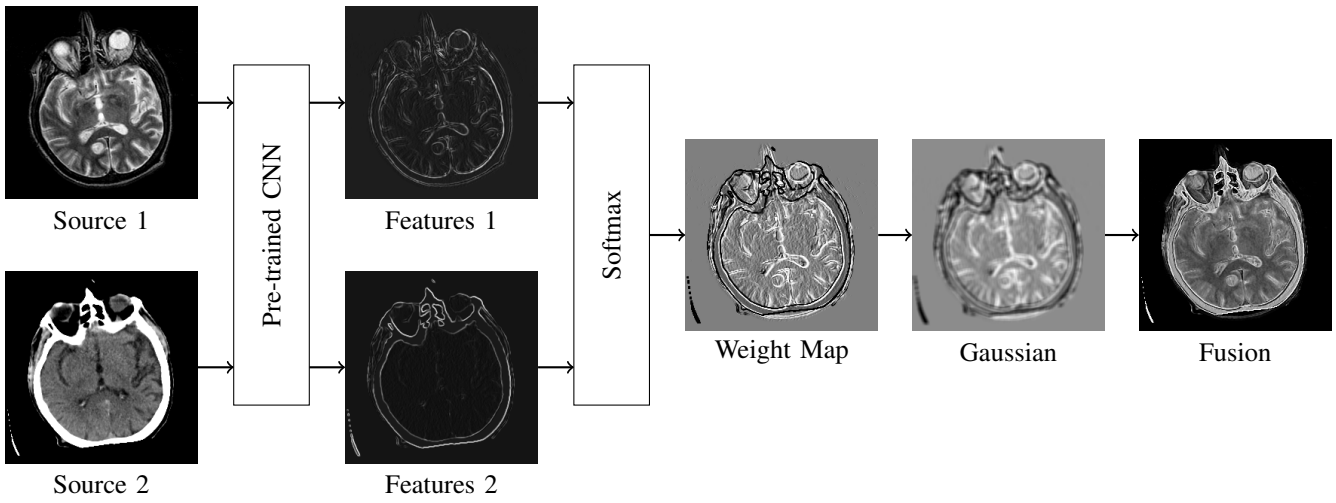

Fig. 1. Schematic diagram of the proposed image fusion algorithm.

limited hardware architectures.

II. Related Work

Image fusion is a vast field that covers an array of algorithms and applications. We first cover classical fusion approaches, and then focus on literature related to generating fusion images using neural networks or deep representations obtained via convolutional layers.

Classical fusion algorithms use multi-scale transforms (MST) to obtain perceptually good results. Frequently used MSTs include pyramids [9], [13], wavelets [8], [18], filters (GFF [19]), and geometrical transforms like contourlets and shearlets [5]–[7], [14]. Another popular approach for modeling images is sparse representation [20], [21]. Images are encoded as a sparse linear combination of samples representing salient features. This encoding is then used to fuse the source images (LP SR [22]). All these methods enforce a specific relation between inputs and outputs, i.e., they operate by a rule which fits a limited set of predefined image priors. Our approach is based on neural networks which learn by example. By training on a large amount of examples, networks are able to construct a more complete understanding of image priors, and as such generalize better on image fusion tasks.

Pulse-coupled neural networks (PCNN) are another set of models that are commonly used in image fusion. They are inspired by the cat's visual cortex (NSCT PCNN [14] and NSST PAPCNN [12]). Unlike deep learning architectures, these models are only two-dimensional, with each neuron corresponding to a single pixel in an image. They also operate without prior training, i.e., they do not form a statistical model of natural images. Unlike them, CNNs are typically trained on large datasets of images, allowing them to model and detect salient features of images, which we leverage in our method to obtain better fusion rules than methods based on PCNNs.

The earliest work involving neural networks poses multi-focus fusion as a classification task [23]. Three focus measures are computed as input features, and a shallow network is trained to output the weight maps. Due to the architecture of the network, the images are fed in patches, resulting in border issues at patch boundaries. In our work, we use CNN models. Thus, the source images are directly fed to the network. This improves the computational efficiency of the method, and the fused output does not suffer from patch border artifacts.

The authors in [21] propose a fusion method based on convolutional sparse representation. This method, while different from deep learning techniques, still uses convolutional features to generate fusion weights. Similarly, convolutional neural networks have been trained on image patches to generate decision maps for multi-focus [15] and medical image fusion [10]. Additionally, DeepFuse [16] uses a two-branch network to fuse extreme exposure image pairs, obtaining state-of-the-art performance. In these approaches, the network predicts the weights for the fusion, but requires training data and is only suitable for a specific fusion task. In contrast, our method requires no training which alleviates the necessity of collecting data. Moreover, it generalizes well to any fusion problem. We use a single pre-trained network as a feature extractor for any multi-modal image fusion.

Finally and closer to our work, in [17], images are decomposed into base and detail content, and a neural network is used to fuse the detail contents while the base parts are averaged. Similarly to these methods, we leverage the ability of deep features to represent salient regions in images. However, contrary to them, we do not run any time consuming optimization scheme, and are able to obtain fused images in real time (> 150fps on 256×256 images).

III. Method

We propose a novel fusion strategy using convolutional neural networks to extract deep features and generate weight maps. In this section, we detail how we obtain the weight maps, and construct the fused image.

Suppose that there are K pre-registered source images denoted as $I_k|k \in \{1, 2, \cdots, K\}$. Additionally, suppose there

is a pre-trained convolutional neural network with L layers, with C_l output channels per layer l. We denote $f_k^{c,l}$ as the c-th feature map of the k-th image extracted at the l-th layer of the network (taken after ReLU operation), the feature map is computed as

$$f_k^l = \max(0, F_l(I_k)) \tag{1}$$

where $F_l(.)$ is the application of the network layers onto the input image up to layer l. The $\max(0, .)$ function denotes the ReLU operation.

For every feature map, we denote $\hat{f}_k^l$ as the l_1-norm computed over the C_l channels of the feature maps of layer l as follows

$$\hat{f}_k^l = \sum_{c=0}^{C_l} ||f_k^{c,l}||_1 \tag{2}$$

This constitutes a measure of the activity level corresponding to the input image at layer l.

The feature maps are extracted for $\mathcal{L}$ layers , so we obtain, per image k, a set of features maps $\hat{\mathcal{F}}_k = \{\hat{f}_k^l | l \in \mathcal{L}\}$. For every layer l, the K feature maps can be used to generate k weight maps to indicate the amount of contribution of each image at a specific pixel. For our method, we use the softmax operator to generate said maps as follows

$$W_k^l = \frac{e^{\hat{f}_k^l}}{\sum_{j=1}^{K} e^{\hat{f}_j^l}} \tag{3}$$

where $e^{(.)}$ is the exponentiation with base e.

In order to account for small mis-registrations, and remove undesirable artifacts around the edges of both modalities, we apply a Gaussian smoothing to the weight maps with $\sigma = 0.01\sqrt{w^2 + h^2}$, where w and h are the spatial dimensions of the weight maps. Fig. 1 illustrates the process of taking two source images, extracting their representative feature maps, generating the fusion weights according to their activation levels at layer l, smoothing, and finally fusing them.

At layer l, we have a set of weights $\mathcal{W}^l = \{W_k^l | k \in \{1, 2, \cdots, K\}\}$. Using these weight maps, the image fusion at layer l is computed as

$$I_F^l = \sum_{k=1}^{K} W_k^l I_k \tag{4}$$

In order to reconstruct the fusion from multiple layers, we choose the maximum pixel from the multi-layer fusions

$$I_F = \max_{l \in \mathcal{L}} [I_F^l] \tag{5}$$

Finally, I_F is clipped to the appropriate range to remove any out of range values. If desired, a tone mapping function could be applied.

IV. Experiments

First, we explore the quality of the medical image fusion with respect to the layer l at which the feature maps are extracted, and the impact of deeper layers on the fusion run time. We then show that our method is competitive with state-of-the-art medical fusion techniques and an order of magnitude faster than prior methods. And last, we show an example of tri-modal fusion.

A. Experimental settings

In order to conduct our experiments, we collect multi-modal medical images from the Whole Brain Atlas [24]. It is a benchmark database containing CT, MRI, PET and SPECT images for normal and abnormal brain structures. All the images in the dataset are pre-registered.

We collect data for four multi-modal fusion tasks. In total, we assemble 97 images for MRI-CT fusion, 24 for T1-T2 weighted MRI fusion, 144 for MRI-PET, and 119 for MRI-SPECT corresponding to different brain diseases. The fusion of these modalities is interesting because they each capture different complementary tissue properties. As discussed previously, CT is sensitive to dense structures and bones while MRI captures softer organ tissues. T1 and T2 are relaxation times that characterize tissue, and thus capture different properties of the same substance under MRI. A PET scan contains information about the activity of the brain. As such it allows physicians to view how it is working and detect operational abnormalities. Finally, SPECT detects blood flow changes in the brain, and is often used to help doctors pinpoint the regions in the brain that are causing epileptic seizures. As both PET and SPECT capture activity information, they are best complemented with an MRI image detailing the structure of the brain where the biological processes take place.

B. Metrics

In our experiments, five commonly used objective fusion metrics are adopted to conduct quantitative evaluations for medical image fusion [10], [25]. They are entropy (*EN*), mutual information (*MI*), structural similarity index (*SSIM*) [26], Q_{abf} [27], and N_{abf} [28]. *EN* reflects the amount of information present in the fused image while *MI* estimates the amount of information transferred from the input images into the fused image. *SSIM* measures how well the structural information of input images is preserved in the fusion. Q_{abf} measures the success of edge information transferred from the sources to the fused images. Finally, N_{abf} measures the level of noise or artifacts added to the fused image that are not present in the source images. A lower value indicates that the fused image contains less artifacts and noise. For all these metrics, a higher value suggests a better performance, except N_{abf} where a lower value indicates a better result.

C. Depth of feature maps

A deep neural network consists of multiple layers, and it is interesting for our application to understand the effect of depth on the output fusion. To that end, we compute the image fusion at multiple depth levels of the network and compare them. For this experiment, we use a VGG-16 [29] network pre-trained for classification on ImageNet [30]. VGG contains multiple pooling layers which decrease the resolution of the

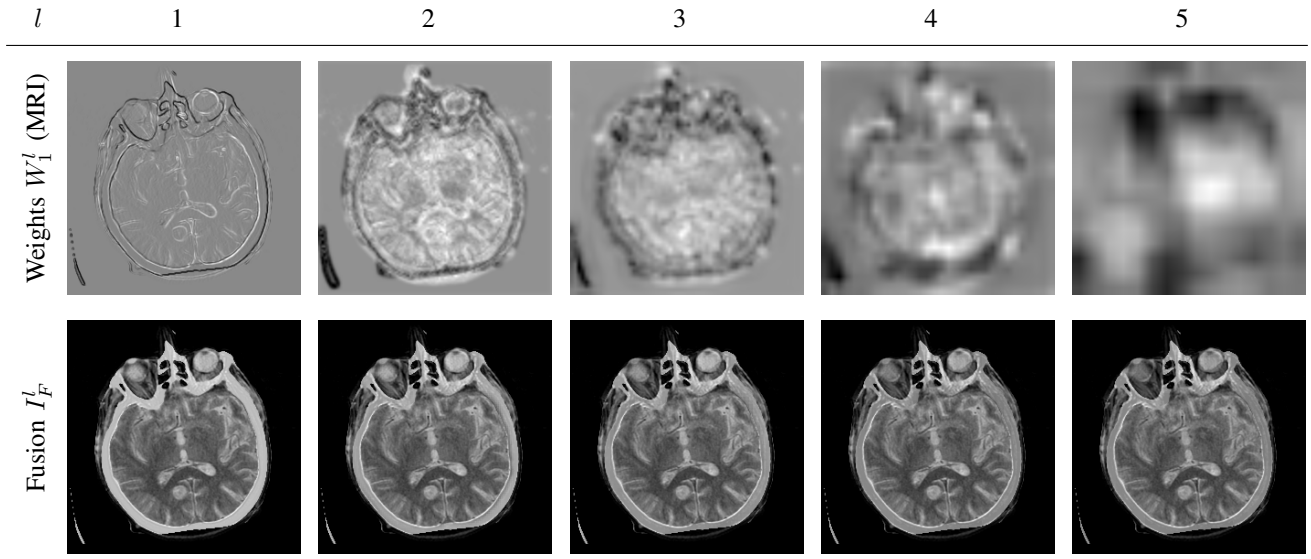

Fig. 2. MRI-CT Fusion weights and results comparison with respect to feature depth. The weights W_1^l shown correspond to the MRI image, the weights corresponding to the CT image are simply $W_2^l = 1 - W_1^l$.

TABLE I
FUSION QUALITY WITH RESPECT TO FEATURE DEPTH

Metric	I_F^1	I_F^2	I_F^3	I_F^4	I_F^5
EN	**4.49**	4.36	4.32	4.33	4.28
MI	**8.98**	8.73	8.71	8.65	8.66
SSIM	**0.73**	0.72	0.71	0.71	0.70
Q_{abf}	**0.68**	0.67	**0.68**	0.66	0.66
N_{abf}	**0.02**	**0.02**	0.03	0.05	0.08
Time (sec/image)	**0.006**	0.007	0.007	0.007	0.008

feature maps. As such, the width and height of the weight maps depend on the layer l over which they were computed. In order to use deeper layers, we need to upsample their feature maps to the size of the original image. VGG contains 5 large convolutional blocks, so we denote I_F^b the fused image taken at layers up to block $b \in \{1, \cdots, 5\}$ using the method described in the previous section.

Table I shows the image fusion quality with respect to the depth of the features considered in the weight computation. The results shown are computed on the fusion of the 97 MRI-CT image pairs. As we consider deeper features from the network, *EN*, *MI*, and *SSIM* decrease, while the amount of fusion noise increases. This means images reconstructed from deeper feature maps contain more information that does not exist in any of the source images compared to those reconstructed using shallower feature maps. Finally, the chosen depth has very little impact on the run time efficiency.

Finally, Fig. 2 shows the progressive weight maps and fusion results as the depth l increases. Due to the lower resolution of the deeper layers, the weight maps contain unwanted upsampling artifacts. These artifacts lower the quality of the fusion by introducing a higher amount of undesirable noise and distorting the intensity levels. For instance, the global contrast of the fusion decreases as the considered depth l increases, making it more difficult to perceive edges. These results show that the shallow layers of a deep neural network are sufficient to build a high quality fusion method.

D. Comparison to other image fusion methods

In this section, we compare our proposed method with other approaches on visual quality, objective assessment, and computational efficiency.

We compare against generic and medical image fusion methods on the four aforementioned modalities: MRI-CT, T1-T2, MRI-PET, and MRI-SPECT. The MRI and CT signals contain single channel images, while PET and SPECT contain false colors that are generated using a pre-defined color mapping. The MRI-CT images correspond to different brain diseases, mainly acute stroke, embolic infarctions, and fatal stroke. The T1-T2 images correspond to scans of a patient suffering a cerebral hemorrhage. MRI-PET images correspond to glioma, and the MRI-SPECT set contains images from five patients having subacute stroke, cavernous angioma, vascular dementia, AIDS, and vascular malformation.

For the color image fusion with MRI, the color image is first transformed into YUV space from the original RGB. Then, the Y component is fused with the grayscale MRI image. The final fused image is reconstructed by converting the intermediate fusion with the U and V components back to RGB.

The methods we compare against are GFF [19], NSCT [7], NSCT PCNN [14], LP SR [22], LIU CNN [10], NSST PAPCNN [12], and LI CNN [17]. GFF is based on a base and detail decomposition with guided filters to generate weights. NSCT and NSCT PCNN all use the non-subsampled contourlet transform to decompose the images into low and high

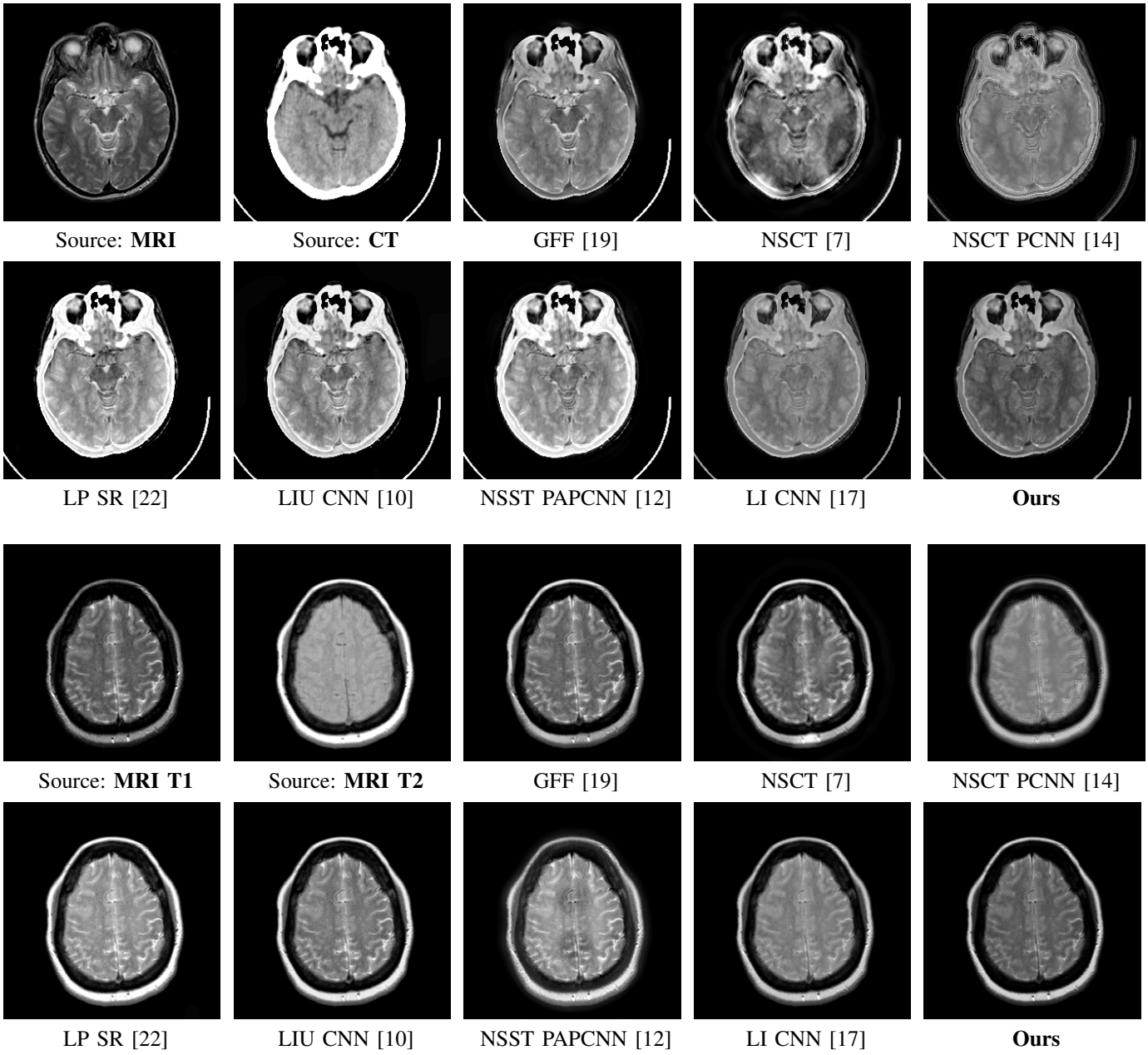

Fig. 3. Performance comparison of different methods on grayscale fusion. The first pair (MRI-CT) is taken from the acute stroke set (slice 11) and the second (T1-T2) from the cerebral hemorrhage set (slice 18). Better viewed on screen.

frequencies, and apply different fusion rules depending on the frequency. LP SR uses Laplacian pyramids to decompose the images and applies sparse representation for low frequency fusion and maximum selection rule for high frequency fusion. LIU CNN employs Siamese Neural Networks [31] to predict the fusion weights with input images in the spatial domain. NSST PAPCNN uses non-subsampled shearlet transform for image decomposition, with an energy based fusion for low frequencies, and an adaptive PCNN for high frequencies. Finally, LI CNN decomposes images using Tikhonov filters. They use a neural network to determine the fusion weights for high frequencies, and simply average the low frequencies. They present their work on infrared and visible fusion, but we include this approach because it is closest to ours. The parameters of all these methods are set to the default values from the provided code. Finally, for our method, we use a VGG-16 network pre-trained on ImageNet, with weights computed at layer $l = 1$.

1) Qualitative evaluation: Fig. 3 shows two pairs of images corresponding to grayscale fusion, MRI-CT and T1-T2, respectively. Fig. 4 shows two pairs corresponding to color fusion, MRI-PET and MRI-SPECT, respectively.

GFF, NSCT and NSCT PAPCNN are not able to capture complementary information in all the regions, which is re-

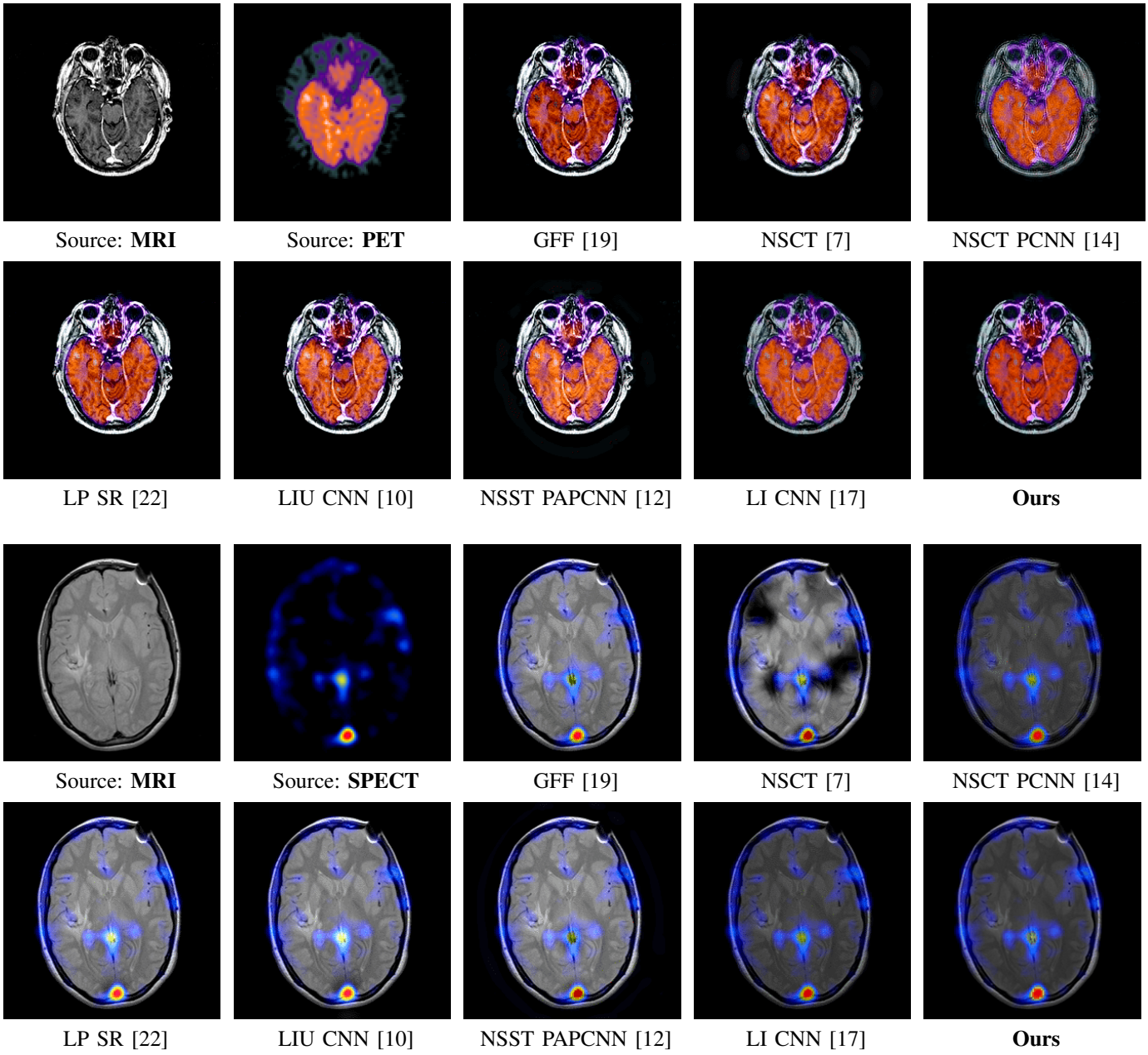

Fig. 4. Performance comparison of different methods on color fusion. The first pair (MRI-PET) is taken from the glioma set (slice 43) and the second (MRI-SPECT) from the cavernous angioma set (slice 14). Better viewed on screen.

flected in the stark contrast inside regions that should have similar intensities. This can be seen for GFF in MRI-CT fusion at the edges, which exhibit a single intensity in both sources but present different levels in the fusion. Similar artifacts can be spotted for NSCT in MRI-CT and MRI-SPECT fusion and for NSCT PAPCNN in T1-T2 fusion. In contrast, since the networks we use have been trained on large datasets with images having various intensity levels, our proposed method respects the intensity differences inside and between salient regions leading to a more intensity consistent fusion that is easier to interpret.

NSCT PCNN and NSCT PAPCNN exhibit blurry and noisy fusion results across all the modalities presented. This is mostly noticeable in the T1-T2 fusion. Since our method operates in the spatial domain, it reduces the amount of artifacts that are produced due to directional decompositions.

Finally, LP SR and LIU CNN generate weights that are highly biased towards the highest intensity signal. This results in overexposed fusions, where darker details from sources do not appear in the fusion. For example, in the MRI-SPECT fusion, the dark blue details are more difficult to spot in LP SR and LIU CNN fusions. LI CNN obtains low contrast results as it only averages the low frequencies. In comparison, our method better represents the edges between neighboring

TABLE II
OBJECTIVE ASSESSMENT OF DIFFERENT METHODS ON FOUR CATEGORIES OF MULTI-MODAL MEDICAL IMAGE FUSION.

Modality	Metric	GFF	NSCT	NSCT PCNN	LP SR	LIU CNN	NSST PAPCNN	LI CNN	Ours
MRI CT	*EN*	3.57	3.92	3.70	3.20	3.44	3.75	3.41	**4.49**
	MI	7.13	7.85	7.40	6.39	6.88	7.50	6.83	**8.98**
	SSIM	0.71	0.60	0.59	0.71	0.68	0.69	0.72	**0.73**
	Q_{abf}	**0.73**	0.52	0.48	0.71	0.66	0.66	0.67	0.68
	N_{abf}	0.06	0.23	0.17	0.07	0.07	0.14	**0.02**	**0.02**
MRT1 MRT2	*EN*	3.60	3.81	3.74	3.70	3.88	4.23	3.56	**4.68**
	MI	7.19	7.61	7.48	7.40	7.76	8.47	7.13	**9.35**
	SSIM	0.83	0.74	0.72	0.83	0.77	0.76	0.85	**0.87**
	Q_{abf}	**0.76**	0.54	0.50	**0.76**	0.69	0.66	0.70	0.71
	N_{abf}	0.04	0.14	0.11	0.05	0.06	0.12	0.01	**0.004**
MRI PET	*EN*	2.11	2.26	2.30	2.14	2.11	2.47	2.19	**2.61**
	MI	4.22	4.51	4.59	4.27	4.23	4.95	4.38	**5.21**
	SSIM	0.83	0.80	0.71	0.82	0.82	0.77	0.83	**0.85**
	Q_{abf}	0.84	0.77	0.68	0.83	0.83	0.70	0.82	**0.85**
	N_{abf}	0.07	0.09	0.06	0.08	0.07	0.11	0.02	**0.01**
MRI SPECT	*EN*	3.15	3.18	3.06	3.14	3.14	3.28	3.18	**4.27**
	MI	6.31	6.36	6.12	6.28	6.28	6.57	6.35	**8.54**
	SSIM	0.65	0.60	0.39	0.67	0.65	0.60	0.66	**0.71**
	Q_{abf}	0.65	0.53	0.32	0.66	0.65	0.60	0.65	**0.68**
	N_{abf}	0.13	0.14	0.07	0.13	0.13	0.16	0.02	**0.01**

TABLE III
RUNNING TIME OF DIFFERENT METHODS ON TWO SOURCE IMAGES OF SIZE 256×256 (IN SECONDS)

	GFF	NSCT	NSCT PCNN	LP SR	LIU CNN	NSST PAPCNN	LI CNN	Ours (CPU)	Ours (GPU)
Time	0.05	3.49	0.46	0.04	13.12	6.92	1.65	0.03	**0.006**
Std σ	0.02	1.90	0.10	0.008	1.91	2.46	0.22	0.003	**5e-4**

regions without introducing any noise or artifacts.

E. Quantitative evaluation

To quantitatively evaluate the performance of these methods against ours, we adopt the previously discussed objective metrics: *EN*, *MI*, *SSIM*, Q_{abf}, and N_{abf}.

Table II shows the performance of different methods against ours on the four tested modalities. The experiments show that our method achieves better performance in all the cases, except for Q_{abf} for MRI-CT and T1-T2 where it still shows competitive results relative to the other methods.

Additionally, the objective metrics reflect the qualitative results. For example, NSCT PCNN and NSCT PAPCNN show high N_{abf} values, caused by the blur and noise artifacts present in these fusions. Conversely, our methods achieves near zero noise levels on all modalities. Similarly, LP SR and LIU CNN have lower *MI* values, reflected by the overexposed fusions hiding detail information. Unlike them, the fusion weights generated by our technique have a higher sensitivity to the presence of information in a pixel than its absolute value.

Finally, Table III shows the average run time of each method on an image taken from the dataset. The average was computed on all 384 images. The experiments were run on a Intel Core i7-7700HQ CPU (2.8GHz), and a GeForce GTX 1050 GPU (2Gb). Our method has the fastest run time, even when run on the CPU. With the use of a GPU, our fusion technique is at least one order of magnitude faster than any other method. Note that for LIU CNN [10], the authors present a faster implementation in C++, where they report an average speed of 0.08 seconds per image. However, only their MATLAB code was published. In both cases, our method is significantly faster and more suitable for real time monitoring systems.

F. Tri-modal fusion

Most fusion methods are designed for bi-modal image fusion, and usually operate on more sources by successive fusions, which is not optimal because fusion algorithms assume image priors that might not be present in the intermediate fusion results. Additionally, neural network based approaches such as LIU CNN [10] can only fuse two images from the image space they were trained on. Successive fusions lead to multiple artifacts as fused images do not correspond to any input space. However, our method can inherently fuse any number $K \geq 2$ of images, see Eq. 4. Fig. 5 shows such an

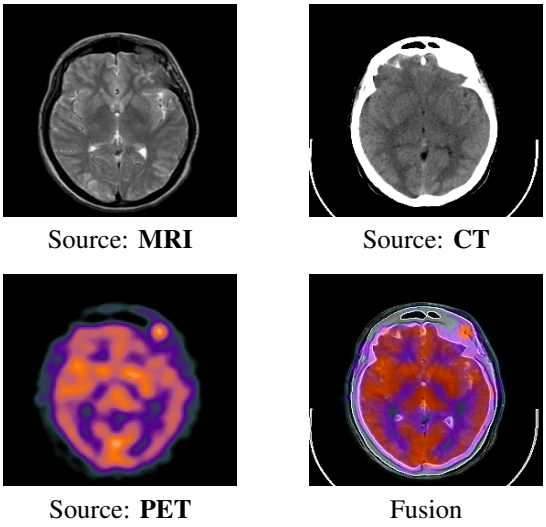

Fig. 5. Tri-modal fusion of MRI, CT, and PET slices. Best viewed on screen.

example, the skull bones present in the CT scan are noticeable in the fusion, alongside the color information from PET and the tissue information from the MRI. Note that the addition of a modality does not affect the processing time of the neural network because the feature maps are extracted in parallel.

V. Conclusion

We present a novel medical image fusion algorithm based on pre-trained neural networks. Unlike typical neural network based techniques, our method requires no prior training on the image modalities and generalizes well to cover different fusion modalities. We leverage the ability of trained networks to detect salient regions in images and extract deep feature maps that describe these regions. By comparing these feature maps, we generate fusion weights to combine the source images.

We show experimentally the robustness of our method to network depth levels. Moreover, we show that our method creates extremely fast high-quality zero-noise fusion results. We also demonstrate its applicability to bi- and tri-modal image fusion. Finally, our method is a lightweight high-quality technique with promising applications in real time systems, and on low-energy hardware.

References

[1] H. Tanaka, S. Hayashi, K. Ohtakara, H. Hoshi, and T. Iida, "Usefulness of ct-mri fusion in radiotherapy planning for localized prostate cancer," *Journal of Radiation Research*, 2011.

[2] M. Yoshino, H. Nakatomi, T. Kin, T. Saito, N. Shono, S. Nomura, D. Nakagawa, S. Takayanagi, H. Imai, H. Oyama *et al.*, "Usefulness of high-resolution 3d multifusion medical imaging for preoperative planning in patients with posterior fossa hemangioblastoma," *Journal of Neurosurgery*, vol. 127, no. 1, 2017.

[3] G. Bhatnagar, Q. J. Wu, and Z. Liu, "A new contrast based multimodal medical image fusion framework," *Neurocomputing*, vol. 157, 2015.

[4] S. Daneshvar and H. Ghassemian, "Mri and pet image fusion by combining ihs and retina-inspired models," *Information Fusion*, vol. 11, no. 2, 2010.

[5] S. Das and M. K. Kundu, "Nsct-based multimodal medical image fusion using pulse-coupled neural network and modified spatial frequency," *Medical & Biological Engineering & Computing*, vol. 50, no. 10, 2012.

[6] ——, "A neuro-fuzzy approach for medical image fusion," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 12, 2013.

[7] G. Bhatnagar, Q. J. Wu, and Z. Liu, "Directive contrast based multimodal medical image fusion in nsct domain," *IEEE Transactions on Multimedia*, vol. 15, 2013.

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

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

[10] Y. Liu, X. Chen, J. Cheng, and H. Peng, "A medical image fusion method based on convolutional neural networks," in *IEEE Fusion*, 2017.

[11] W. Zhao and H. Lu, "Medical image fusion and denoising with alternating sequential filter and adaptive fractional order total variation," *IEEE Transactions on Instrumentation and Measurement*, vol. 66, no. 9, 2017.

[12] M. Yin, X. Liu, Y. Liu, and X. Chen, "Medical image fusion with parameter-adaptive pulse coupled neural network in nonsubsampled shearlet transform domain," *IEEE Transactions on Instrumentation and Measurement*, no. 99, 2018.

[13] A. Toet, "A morphological pyramidal image decomposition," *Pattern Recognition Letters*, vol. 9, no. 4, 1989.

[14] N. Wang, Y. Ma, K. Zhan, and M. Yuan, "Multimodal medical image fusion framework based on simplified pcnn in nonsubsampled contourlet transform domain," *Journal of Multimedia*, vol. 8, no. 3, 2013.

[15] Y. Liu, X. Chen, H. Peng, and Z. Wang, "Multi-focus image fusion with a deep convolutional neural network," *Information Fusion*, vol. 36, 2017.

[16] K. Ram Prabhakar, V. Sai Srikar, and R. Venkatesh Babu, "Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs," in *IEEE International Conference on Computer Vision*, 2017.

[17] H. Li, X.-J. Wu, and J. Kittler, "Infrared and visible image fusion using a deep learning framework," in *IEEE International Conference on Pattern Recognition*, 2018.

[18] G. Qu, D. Zhang, and P. Yan, "Medical image fusion by wavelet transform modulus maxima," *Optics Express*, vol. 9, no. 4, 2001.

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

[20] Y. Liu and Z. Wang, "Simultaneous image fusion and denoising with adaptive sparse representation," *IET Image Processing*, vol. 9, no. 5, 2014.

[21] Y. Liu, X. Chen, R. K. Ward, and Z. J. Wang, "Image fusion with convolutional sparse representation," *IEEE Signal Processing Letters*, vol. 23, no. 12, 2016.

[22] 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, 2015.

[23] S. Li, J. T. Kwok, and Y. Wang, "Multifocus image fusion using artificial neural networks," *Pattern Recognition Letters*, vol. 23, no. 8, 2002.

[24] D. Summers, "Harvard whole brain atlas: www.med.harvard.edu/aanlib/home.html," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 74, no. 3, 2003.

[25] J. Du, W. Li, K. Lu, and B. Xiao, "An overview of multi-modal medical image fusion," *Neurocomputing*, vol. 215, 2016.

[26] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, 2004.

[27] C. S. Xydeas and V. S. Petrovic, "Objective pixel-level image fusion performance measure," in *Sensor Fusion: Architectures, Algorithms, and Applications IV*, vol. 4051, 2000.

[28] V. Petrovic and C. Xydeas, "Objective image fusion performance characterisation," in *IEEE International Conference on Computer Vision*, vol. 2, 2005.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009.

[31] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a siamese time delay neural network," in *Advances in Neural Information Processing Systems*, 1994.