AN ATTENTION-BASED MULTI-SCALE FEATURE LEARNING NETWORK FOR MULTIMODAL MEDICAL IMAGE FUSION

Meng Zhou, Xiaolan Xu, Yuxuan Zhang*

Department of Computer Science, University of Toronto, Toronto, ON, Canada

ABSTRACT

Multimodal medical images play an important role in clinical applications, images with different modalities could provide rich information about patients for radiologists to achieve accurate diagnosis. Image fusion techniques have emerged as a valuable approach for synthesizing complementary information from these multimodal images into a unified representation. In this paper, we introduce a novel Dilated Residual Attention Network for multimodel brain CT and MRI fusion. We further propose a novel Softmax Feature weighted fusion strategy, to fuse the information of two images in the latent space. Experiments show our proposed framework achieves state-of-the-art performance in both objective and subjective assessments. We made our code publicly available at `https://github.com/simonZhou86/dilran`.

***Index Terms*—** Deep Learning, Image Fusion, Attention Mechanism, Residual Learning, Brain MRI, CT

1. INTRODUCTION

Medical imaging plays an increasingly prominent role in modern clinical diagnosis. The rich information embedded in multimodal medical images equips radiologists with a multi-faceted perspective for precise decision-making and treatment planning. To acquire such information during diagnosis, radiologists may have to analyze various types of medical images, e.g., CT for bone density information and MRI for nerves and blood vessels information, which is time-consuming and requires meticulous efforts. Multimodal medical image fusion can merge the complementary information of original images and present the required information in one fused image. This is clinically significant because radiologists can now access more detailed and comprehensive information about disease-related changes in a single image, reducing the time they have to switch between different images. In recent years, deep learning-based approaches have been proven successful in image fusion tasks. Convolutional neural networks (CNNs) have the ability to capture and extract features and can be used to acquire, merge features, and reconstruct fused images from the features. CNN-based models have also proven to be useful in medical image fusion and have the potential to merge complex information in various medical image modalities [1].

Related Works. Image fusion methods have been studied for years. Traditional fusion algorithms include Possion Image Editing [2], cross bilateral filter [3], and non-subsampled contourlet transform [4]. Recent advancements in deep learning have led to the successful use of CNNs in the image fusion task. [5] proposed an improved version of pulse-coupled neural network (PCNN) to perform contourlet transform for medical image fusion. [6] applied a CNN to the multi-focus image fusion task, using the network to generate a weight map of pixel activity to improve the quality of fused images. [7] proposed DenseFuse, a CNN architecture employing dense blocks for infrared and visible image fusion. [8] improved the DenseFuse by introducing a multi-scale DenseNet (MSD-Net), which integrates a multi-scale mechanism using filters of varying sizes for effectively capturing features at different scales. Recent developments have seen a surge in the use of residual attention mechanisms [9, 10], these two methods have been pivotal in fusing anatomical and functional medical images and achieving state-of-the-art performance on various fusion tasks.

Contributions. In this work, we propose a novel Dilated Residual Attention Network (DILRAN) as a core module of the feature encoder to effectively learn the semantic representations of input images. DILRAN builds upon the residual and pyramid attention network and the dilated convolutions, enabling it to extract multi-scale deep semantic features. Furthermore, we present a novel fusion strategy termed Softmax Feature Weighted Strategy (SFNN) to fuse two feature maps based on the matrix nuclear norm. Finally, experiments show our proposed framework and fusion strategy exceed the state-of-the-art performance based on objective fusion metrics (PSNR, MI, etc.) and subjective image quality.

2. MATERIALS AND METHODS

2.1. Data

We use "The Harvard Whole Brain Atlas" [1] [11] for this work. The dataset contains 184 pairs of co-registered multimodal medical images of brain CT and MRI. All images are with

*email: simonzhou@cs.toronto.edu

[1]Available at `https://www.med.harvard.edu/aanlib/`

size 256×256, we normalize the pixel values in the range of $[0, 1]$ for both CT and MRI. We randomly hold out 20 image pairs for testing and for the remaining 164 pairs, we use 80% of the data to train and 20% for validation.

2.2. Method

Figure 1 summarizes the overall pipeline in this work. Our proposed fusion framework consists of a feature extractor, a fusion module, and a reconstruction module. The feature extractor aims to formulate deep semantic features of input image pairs, and then use them as inputs to the fusion module (Fig. 1a). We introduced the Dilated Residual Attention Network (DILRAN) as the core module to extract the rich information of the input image in the latent space, see details in Section 2.2.1. Next, a softmax-weights-based fusion strategy is proposed to fuse the two extracted feature maps into one map containing important features from both original feature maps, see details in 2.2.2. Finally, the reconstruction module, which consists of three convolution layers, is used to reconstruct the fused image based on fused features.

2.2.1. Dilated Residual Attention Network

The design principle of Dilated Residual Attention Network (DILRAN) is inspired by two mechanisms: residual attention [12] and pyramid attention [13]. The residual attention network aims to generate attention-aware local semantic features and could gradually refine the extracted features in deep layers. Furthermore, the residual attention network is capable of speeding up the model convergence without vanishing the gradient. However, it may not be able to extract multi-scale features. Hence, we utilize the pyramid attention network [13] to learn multi-scale features. We focus on the feature pyramid attention that could provide deeper and richer semantic features using hierarchical convolution blocks. To enable the model to capture features at different scales and receptive fields (RFs), we replace convolutions with larger kernel filters with a sequence of convolutions with smaller kernel filters. Figure 1b shows the proposed DILRAN architecture, our method consists of a single 3×3 convolutions (a 3×3 RFs); followed by *two* 3×3 convolutions (represents a 5×5 RFs) in the second level, and finally followed by *three* 3×3 convolutions (represents a 7×7 RFs) in the third level.

To further enhance the model to learn local information and fine details in the image, we leverage the $\{1, 3, 5\}$-dilated convolution on shallow features of the original input image to extract the multi-scale information. The receptive field is expanded using three different dilated rates to improve the discriminative multi-scale feature extraction ability of the model. Once the multi-scale features are extracted, we concatenate those features channel-wise, and then the residual-pyramid attention paradigm is used to further extract deep features.

2.2.2. Fusion Strategy

The fusion strategy in the fusion module is used to fuse the extracted features of input images into a single feature map. We introduce a novel fusion strategy termed "Softmax Feature Weighted Strategy" that exceeds the state-of-the-art performance. First, we obtain two output feature maps f_1, f_2 from the extraction module for input images I_1, I_2, respectively. The output feature map from the extraction module can be used to generate the corresponding weight map that indicates the amount of contribution of each pixel to the fused feature map [14]. To get the weight map, we take the Softmax operation on the feature map, $S(x_i) = \frac{\exp(x_i)}{\sum_i \exp(x_i)}$, where x_i is the ith channel of the output feature map x.

After we obtained the Softmax output, we computed the matrix nuclear norm ($\| \cdot \|_*$), which is the summation of its singular values. Finally, we obtain the weights for the output feature map by taking the weighted average of the maximum value of the nuclear norm along the channel i, which is given in Equation (1).

$$\begin{aligned} W_1 &= \frac{\phi(\|S(x_i)^1\|_*)}{\phi(\|S(x_i)^1\|_*) + \phi(\|S(x_i)^2\|_*)} \\ W_2 &= \frac{\phi(\|S(x_i)^2\|_*)}{\phi(\|S(x_i)^1\|_*) + \phi(\|S(x_i)^2\|_*)} \end{aligned} \tag{1}$$

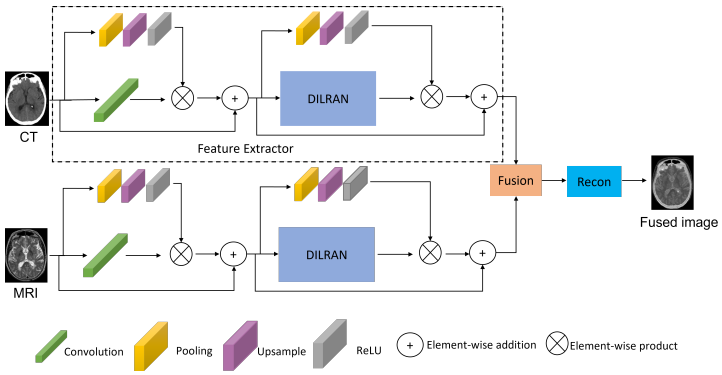
Where $S(x_i)^j, j \in [1, 2]$ is the weight map after Softmax operation for the feature map $f_j, j \in [1, 2]$, $\phi(\cdot)$ is the $max()$ function. The final fused feature map is then given by $f = W_1 * f_1 + W_2 * f_2$.

2.3. Loss function

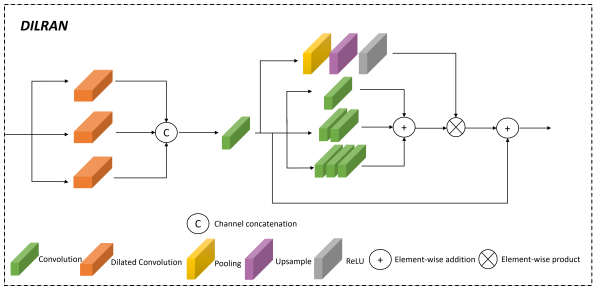
We hypothesize that in order to make the fused image as close as two input images, the individual image before and after the reconstruction should be as close as possible. Then, after we apply the fusion strategy and obtain the fused deep features, the reconstruction module will produce a more realistic and reasonable fused image. The fused image should also have a strong correlation with both input images. Therefore, we treat the overall training object as a reconstruction task, and we define x to be the input image and $\hat{x}$ to be the reconstructed image. We use the $\mathcal{L}_2$ distance to measure the pixel differences between x and $\hat{x}$; gradient loss [15] to model the fine details of textures in the reconstructed image, and perceptual loss [16] to model the high-level semantic similarity between reconstructed and input images. In detail, our loss function is defined as in Equation (2):

$$\begin{aligned} \mathcal{L}_{pixel} &= \|x - \hat{x}\|_2^2 \\ \mathcal{L}_{perp} &= \|f^i(x) - f^i(\hat{x})\|_2^2 \\ \mathcal{L}_{grad} &= \|\nabla(x) - \nabla(\hat{x})\|_2^2 \end{aligned} \tag{2}$$

The image gradient loss is realized by the $\mathcal{L}_2$ norm of the image gradient in x and y-direction. $f^{(i)}$ in $\mathcal{L}_{perp}$ is the ith



(a) The overall pipeline of the proposed method, which consists of a feature encoder based on the DILRAN module proposed in this work, a fusion module, and an image reconstruction module.



(b) Proposed Dilated Residual Attention Network architecture. Three hierarchical convolution blocks contain one, two, and three consecutive 3×3 convolutional layers, respectively.

Fig. 1: *Left:* The overall network structure introduced in this work. *Right:* The Dilated Residual Attention Network architecture proposed in this work.

layer from the pretrained VGG16 network [17]. Finally, the total loss function is given by $\mathcal{L}_{total} = \mathcal{L}_{pixel} + \lambda_1 \mathcal{L}_{perp} + \lambda_2 \mathcal{L}_{grad}$, where λ_1, λ_2 are weight balancing factors of the gradient and perceptual loss, respectively.

To train the network, we use Adam as the optimizer, the learning rate is set to 0.0001, the batch size is set to 4, and the model is trained for 100 epochs. The weight balancing factors λ_1 and λ_2 are set to 0.2.

3. EXPERIMENTS AND RESULTS

We use three baseline methods to prove the effectiveness of the proposed method, zero-shot learning for medical image fusion [14], MSRPAN [9], and MSDRA [10]. Hyperparameters of these methods are the default values suggested by the authors. We utilize six commonly used quantitative metrics for evaluating our proposed method: Peak signal-to-noise ratio (PSNR), Structural Similarity (SSIM) [18], Feature SSIM (FSIM) [19], Mutual information (MI) [20], pixel-wise feature MI (FMI-pixel) [21], and Information Entropy (EN).

3.1. Comparison between fusion strategies

As we discussed in Section 2.2.2, the selected function for $\phi(\cdot)$ is $max()$, but we also conduct experiments on $mean()$, and $sum()$. We provide detailed quantitative results when different fusion strategies are used in Table 1. FER [9] and FL1N [10] are two fusion strategies proposed previously. Our proposed fusion strategy performed well on five metrics (PSNR, SSIM, FMI, FSIM, EN). Figure 2 shows the qualitative results of different fusion strategies. For SFNN, we select two strategies that produce the best image quality to visualize and also notice that different choices of $\phi(\cdot)$ in our proposed strategy do not affect the metrics very much, demonstrating the robustness of the proposed fusion strategy. Compared with the FER strategy [9] in Figure 2c, our results have better fidelity, and the inner tissue boundary is more clear (pointed out by the orange arrow), while the FER strategy fails to show the boundary of inner tissues. Moreover, our results produce brighter edges than the FL1N strategy [10] in Figure 2d (pointed out by the blue arrow), which is better when separating between the bone edge and inner tissues.

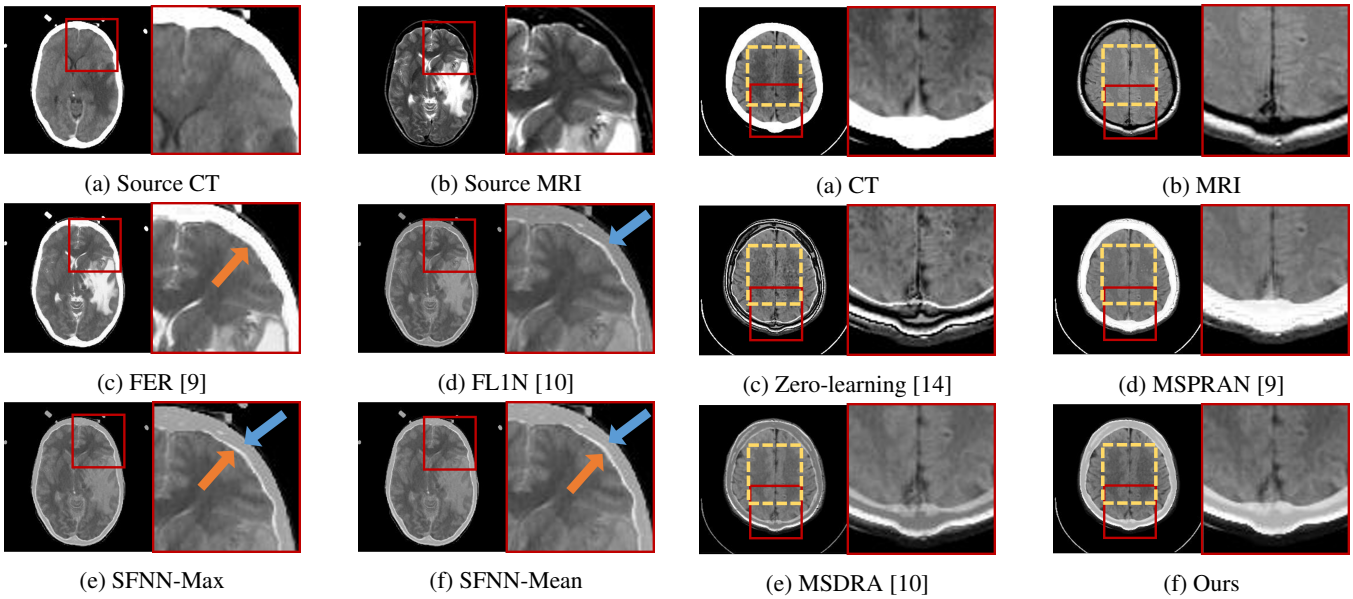
Table 1: Comparison between different fusion strategies, bold values represent the best results. Higher values indicate better performance for all metrics.

	PSNR ↑	SSIM ↑	MI ↑	FMI-Pixel ↑	FSIM ↑	Entropy ↑
FER [9]	13.944	0.736	4.551	0.876	0.806	8.720
FL1N [10]	15.979	0.739	4.569	0.878	0.813	9.782
SFNN ($mean$)	15.876	**0.740**	**4.578**	0.876	0.812	9.772
SFNN (max)	**16.413**	**0.740**	4.558	**0.891**	**0.820**	**9.816**
SFNN (sum)	15.876	**0.740**	**4.578**	0.876	0.812	9.772

3.2. Comparison with baselines

Table 2 presents the quantitative results of the brain MRI-CT fusion task between our proposed method and other baseline methods. Our method outperforms four fusion metrics compared with baselines: PSNR, pixel-wise FMI, FSIM, and Entropy, while the SSIM score is lower than that of the MSRPAN model by only 0.001. The largest PSNR value indicates our fused image contains less noise and achieves the best image quality. The largest FMI value demonstrates our fused image contains as many source image features as possible. The largest FSIM value suggests our fused image has less information loss at the feature level. Finally, the largest entropy shows our fused image contains more information and details.

From Figure 3, it is also obvious from the enlarged box that the proposed method preserves the edge and detailed in-



(a) Source CT (b) Source MRI

(c) FER [9] (d) FL1N [10]

(e) SFNN-Max (f) SFNN-Mean

Fig. 2: Visualization of different fusion strategies for brain CT-MRI fusion. (a) and (b) are source images, (c) is the FER fusion strategy in [9], (d) is the FL1N fusion strategy in [10], and (e), (f) are the strategies we proposed in this work.

(a) CT (b) MRI

(c) Zero-learning [14] (d) MSPRAN [9]

(e) MSDRA [10] (f) Ours

Fig. 3: Visualization of brain CT-MRI fusion results. (a) and (b) are source images, (c) is the zero-learning method [14], (d) is the MSPRAN method [9], (e) is the MSDRA method [10], and (f) is the method we proposed in this work using SFNN-max fusion strategy. Zoom in for a better view.

Table 2: Comparison between different methods, bold numbers represent optimal values. Higher values indicate better performance for all metrics.

	PSNR ↑	SSIM ↑	MI ↑	FMI-Pixel ↑	FSIM ↑	Entropy ↑
Zero-shot [14]	13.525	0.681	4.633	0.836	0.738	4.279
MSDRA [10]	15.693	0.697	4.586	0.867	0.797	8.167
MSRPAN [9]	14.528	**0.741**	**4.652**	0.874	0.808	8.969
Ours	**16.413**	0.740	4.558	**0.891**	**0.820**	**9.816**

formation well. Compared to other methods, Figure 3c does not retain the edge information from CT, and details of inner tissues are also distorted; Figure 3d has a favorable visual appearance, but the fidelity is low and lost the tissue boundary information from the MRI source image. Figure 3e has an uncleared boundary so it is hard to differentiate between boundaries and tissues. Take a closer look at the golden bounding box around the centrum semiovale region in Figure 3, [9, 14] fail to preserve the intensity information from the CT, and [10] results in less intensity contrast with other tissues. Finally, our proposed method has a better intensity contrast between edges and tissues, retains important information from both source images and results in better fidelity.

4. CONCLUSIONS

In this paper, we proposed a novel network architecture DILRAN to extract multi-scale features and introduced the Softmax Feature weighted fusion strategy for multimodal brain CT-MRI fusion. The original images are passed into the feature encoder to extract multi-scale deep semantic features; features are then fused together based on the fusion strategy we proposed, and finally, the fused features are decoded into the image space. Our fusion strategy is fixed, there is no parameter that needs to be updated in both the training and inference phases, which enables real-time image fusion. Our results show the proposed method is superior compared to baseline methods in both subjective visual appearance and objective fusion metrics. Our method could provide a reliable reference for disease diagnosis in the real-life clinical routine.

Future Work: There are other modalities such as SPECT and PET that could provide other useful clinical information, we plan to investigate the fusion performance between these two modalities with MRI. The fusion strategy we proposed is fixed in this work, we plan to replace it with a neural network to automatically adjust the weights to fuse two feature maps. Lastly, we aim to scale our proposed method to 3D for 3D medical image fusion.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using the human subject dataset, The Whole Brain Atlas is available in open access at `https://www.med.harvard.edu/aanlib/`. Ethical approval was not required as confirmed by the license attached with the open access data.

6. ACKNOWLEDGMENTS

No funding was received for conducting this study. The authors would like to thank Compute Canada (`https://alliancecan.ca/en`) for the GPU provided.

7. REFERENCES

[1] Bing Huang, Feng Yang, Mengxiao Yin, Xiaoying Mo, Cheng Zhong, et al., "A review of multimodal medical image fusion techniques," *Computational and mathematical methods in medicine*, vol. 2020, 2020.

[2] Patrick Pérez, Michel Gangnet, and Andrew Blake, "Poisson image editing," in *ACM SIGGRAPH 2003 Papers*, pp. 313–318. 2003.

[3] BK Shreyamsha Kumar, "Image fusion based on pixel significance using cross bilateral filter," *Signal, image and video processing*, vol. 9, no. 5, pp. 1193–1204, 2015.

[4] Arthur L Da Cunha, Jianping Zhou, and Minh N Do, "The nonsubsampled contourlet transform: theory, design, and applications," *IEEE transactions on image processing*, vol. 15, no. 10, pp. 3089–3101, 2006.

[5] Yu Tian, Yibing Li, and Fang Ye, "Multimodal medical image fusion based on nonsubsampled contourlet transform using improved pcnn," in *2016 IEEE 13th International Conference on Signal Processing (ICSP)*. IEEE, 2016, pp. 799–804.

[6] Yu Liu, Xun Chen, Hu Peng, and Zengfu Wang, "Multi-focus image fusion with a deep convolutional neural network," *Information Fusion*, vol. 36, pp. 191–207, 2017.

[7] Hui Li and Xiao-Jun Wu, "Densefuse: A fusion approach to infrared and visible images," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2614–2623, 2019.

[8] Xu Song, Xiao-Jun Wu, and Hui Li, "Msdnet for medical image fusion," in *International conference on image and graphics*. Springer, 2019, pp. 278–288.

[9] Jun Fu, Weisheng Li, Jiao Du, and Yuping Huang, "A multiscale residual pyramid attention network for medical image fusion," *Biomedical Signal Processing and Control*, vol. 66, pp. 102488, 2021.

[10] Weisheng Li, Xiuxiu Peng, Jun Fu, Guofen Wang, Yuping Huang, and Feifei Chao, "A multiscale double-branch residual attention network for anatomical–functional medical image fusion," *Computers in Biology and Medicine*, vol. 141, pp. 105005, 2022.

[11] D Summers, "Harvard whole brain atlas: www. med. harvard. edu/aanlib/home. html," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 74, no. 3, pp. 288–288, 2003.

[12] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang, "Residual attention network for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3156–3164.

[13] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang, "Pyramid attention network for semantic segmentation," *arXiv preprint arXiv:1805.10180*, 2018.

[14] Fayez Lahoud and Sabine Süsstrunk, "Zero-learning fast medical image fusion," in *2019 22th International Conference on Information Fusion (FUSION)*. IEEE, 2019, pp. 1–8.

[15] Cameron Fabbri, Md Jahidul Islam, and Junaed Sattar, "Enhancing underwater imagery using generative adversarial networks," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 7159–7165.

[16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

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

[18] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[19] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[20] Guihong Qu, Dali Zhang, and Pingfan Yan, "Information measure for performance of image fusion," *Electronics letters*, vol. 38, no. 7, pp. 1, 2002.

[21] Mohammad Bagher Akbari Haghighat, Ali Aghagolzadeh, and Hadi Seyedarabi, "A non-reference image fusion metric based on mutual information of image features," *Computers & Electrical Engineering*, vol. 37, no. 5, pp. 744–756, 2011.