2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

CSpA-DN: Channel and Spatial Attention Dense Network for Fusing PET and MRI Images

Bicao Li[1, 3, 4], *Zhoufeng Liu*[1], *Shan Gao*[2, 3, 4], *Jenq-Neng Hwang*[3], *Jun Sun*[4, 1], *Zongmin Wang*[4]

[1]School of Electronic and Information Engineering
Zhongyuan University of Technology
Zhengzhou 450007, China
{lbc, liuzhoufeng}@zut.edu.cn

[2]College of Information Science and Engineering
Henan University of Technology
Zhengzhou 450001, China
gaoshan_zz@126.com

[3]Department of Electrical Engineering
University of Washington
Seattle, WA 98195, USA
hwang@uw.edu

[4]School of Information engineering
Zhengzhou University
Zhengzhou 450001, China
sunjun@zut.edu.cn, zmwang@ha.edu.cn

***Abstract*—In this paper, we propose a novel fusion framework based on a dense network with channel and spatial attention (CSpA-DN) for PET and MR images. In our approach, an encoder composed of the densely connected neural network is constructed to extract features from source images, and a decoder network is leveraged to yield the fused image from these features. Simultaneously, a self-attention mechanism is introduced in the encoder and decoder to further integrate local features along with their global dependencies adaptively. The extracted feature of each spatial position is synthesized by a weighted summation of those features at the same row and column with this position via a spatial attention module. Meanwhile, the interdependent relationship of all feature maps is integrated by a channel attention module. The summation of the outputs of these two attention modules is fed into the decoder and the fused image is generated. Experimental results illustrate the superiorities of our proposed CSpA-DN model compared with state-of-the-art methods in PET and MR images fusion according to both visual perception and objective assessment.**

Keywords—channel attention; dense network; image fusion; spatial attention; PET and MRI

I. INTRODUCTION

As is well known, medical imaging plays an increasingly significant role in many clinical applications including disease diagnosis, treatment planning and surgical navigation [1]. With the rapid development of sensor mechanisms and medical imaging technologies, there are many kinds of medical image modalities in clinical applications, such as Computed Tomography (CT), Magnetic Resonance imaging (MRI), and Positron Emission Tomography (PET), etc. Generally, different modalities of medical images need to be observed and analyzed separately by physicians to acquire comprehensive information for diagnosing the illness, but this separating mode of assessment is time-consuming and may result in inconvenience in many clinical applications [2]. In consequence, the goal of medical image fusion is to integrate the complementary information contained in multi-modal images [3, 4]. Specifically, fusion of PET and MRI images can not only obtain abundant anatomical structures from MRI images and but also retain rich functional information form PET images.

In recent years, a plenty of medical image fusion methods have been proposed. Due to the difference in imaging mechanism of multi-modal medical images, the pixel intensities of raw images at the same location generally vary significantly. As a result, many algorithms based on multi-scale decomposition (MSD) are introduced to pursue perceptually good fusion results. These frequently-used MSD fusion technologies include Laplacian pyramid (LP) [5], discrete wavelet transform [6] and dual-tree complex wavelet transform [7], multi-resolution singular value decomposition (MSVD) [8], non-subsampled contourlet transform (NSCT) [9] and non-subsampled shearlet transform (NSST) [10]. Nevertheless, many researches in the literature demonstrate that the performances of these MST approaches mainly depend on designing an effective fusion strategy.

In recent years, ever since the success of AlexNet in the ImageNet challenge of 2012 [11], deep learning techniques have been applied in many fields of computer vision, and the more recent success in image fusion tasks. Liu et al. [12] trained a deep convolutional neural network (CNN) by the image patches to learn a direct mapping from source raw images to the fused image. Subsequently, Liu et al. [13] also proposed a medical image fusion method using a Siamese convolutional network to create a weight map, which can synthesize the pixel activity information of the source images. Prabhakar et al. [14] introduced a deep unsupervised CNN framework to learn the fusion operation. Zhang et al. [15] introduced a general image fusion approach based on the CNN model called IFCNN. They extract the features from the source images using two convolutional layers and merge these convolutional features by a suitable fusion strategy. Finally, the fusion image was reconstructed by these merged features. Likewise, Liu et al. [16] also proposed a two-stream fusion network (TFNet) to fuse panchromatic image and multi-spectral image.

Recent researches have reported that the backpropagated gradient information of the neural network can vanish and

978-1-7281-8808-9/20/$31.00 ©2020 IEEE

2020 25th International Conference on Pattern Recognition (ICPR) | 978-1-7281-8808-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICPR48806.2021.9412543

Authorized licensed use limited to: Communication University of China. Downloaded on September 02,2023 at 10:37:03 UTC from IEEE Xplore. Restrictions apply.

"washout" as the CNN layers become increasingly deep. Many publications [17-20] create skip connections between previous layers and later layers to address this problem. In contrast to ResNets, Huang et al. [21] concatenated the features instead of summation before passed into the next layer. In their approach, the additional inputs from all previous layers are fed into the subsequent layers, which can preserve the feed-forward nature and ensure maximum information flow between layers. Therefore, each layer consists of all features of the previous layers in a simple densely connected way, referred as a Dense Convolutional Network (DenseNet), which has the strong ability of feature representation with fewer parameters than traditional CNNs. Li and Wu [22] proposed a new deep learning model based on the DenseNet architecture for infrared and visible image fusion named as DenseFuse. Xu et al. [23] introduced a weight block to get the weights of two source images by calculating image quality assessment and entropy metrics. Combining the weights, they presented a new image fusion framework for multiple tasks applying the dense connected network called FusionDN.

Nevertheless, the significant information of different modalities of source images can vary significantly. For instance, as representations of various medical images, PET images represent the functional information with high-contrast pixel intensities, whereas MRI images primarily show the abundant structural and texture information of human organs and tissues with gradient variations. Furthermore, the primary obstacle of image fusion based on deep learning is the insufficiency of ground-truth fused images. To deal with this issue, the ground-truth fusion images are synthetically created in some publications. However, the obtained ground-truth fused images in this manner are not appropriate for all fusion problems and the process is time-consuming. To deal with the above-mentioned challenges, we propose a novel fusion model based on dense network that does not require the ground-truth fusion images.

It is reported by recent researches [24, 25] that the channel and spatial attention mechanisms have been successfully applied in the scene segmentation and curvilinear structure segmentation. In clinical applications, tree-like structures are frequently encountered in biomedical contexts in clinical applications, such as the bronchial system, the vascular topology, and the breast ductal network, etc. In addition, each feature channel can be considered as the specific structure response. Motivated by [24], we introduce a self-attention network consisting of two parallel attention modules, i.e., channel attention and spatial attention, into our fusion model and further enhance the capability of feature representation. The spatial attention mechanism can integrate the global structural information into local features. The channel attention strategy can sufficiently leverage the interdependencies between different channels. To improve the computational efficiency, we employ a Criss-Cross attention block that was introduced by [26] to obtain the spatial attention.

Combining dense network and self-attention mechanism, we propose a novel Channel and Spatial Attention Dense Network (CSpA-DN) for PET and MRI images. Our approach consists of three components: an encoder based on the densely connected network, a self-attention network and a decoder. Given two source images, the encoder network is leveraged to extract the image features. Then, these features are fed into the self-attention network to adaptively combine local features as well as their global dependencies. Finally, the better feature representations are obtained from the self-attention network and fed into the decoder to generate the final fused image. Hence, our PET and MRI fusion task does not require the ground-truth fusion images. Both visual assessment and quantitative evaluation results demonstrate the superior performance of CSpA-DN compared with state-of the-art approaches.

To sum up, the primary contributions of our work consist of the following aspects:

- Taking into account the shortage of ground-truth fusion images as the stumbling block, we propose a novel deep learning network for image fusion.
- The image fusion problem is formulated as an encoder-decoder framework; thus, the proposed fusion model can be trained in an end-to-end learning mode and automatically generate the fused image without designing any activity level measurement or fusion rule.
- We propose a novel densely connected network with self-attention mechanism to improve the discriminant ability of spatial feature representations for image fusion.
- Experimental results on PET and MRI images demonstrate state-of-the-art performance.

The rest of this paper is organized as follows. In Section 2, we introduce the new proposed fusion architecture based on a densely connected network with a self-attention mechanism. Section 3 provides the training datasets and our experimental results on PET and MRI images with a comparison to the state-of-the-art approaches. Concluding remarks and perspectives are addressed in Section 4.

II. Proposed CSpA-DN Fusion Model

In this section, we will describe in details the network architecture of our proposed CSpA-DN fusion model, a typical encoder-decoder structure as shown in Fig. 1, which consists of an encoder network, an attention network and a decoder.

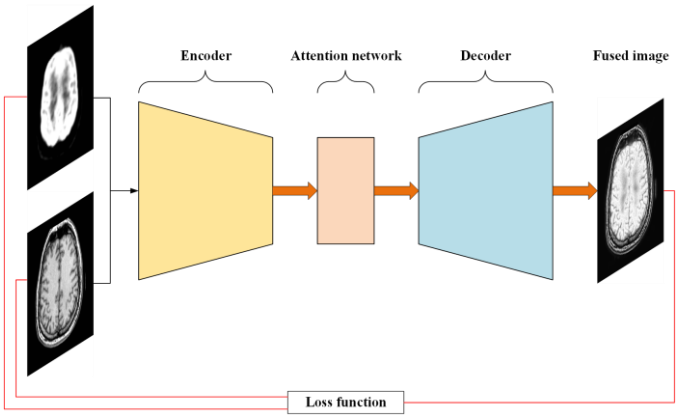

Fig. 1. Schematic architecture of our proposed CSpA-DN fusion network.

Authorized licensed use limited to: Communication University of China. Downloaded on September 02,2023 at 10:37:03 UTC from IEEE Xplore. Restrictions apply.

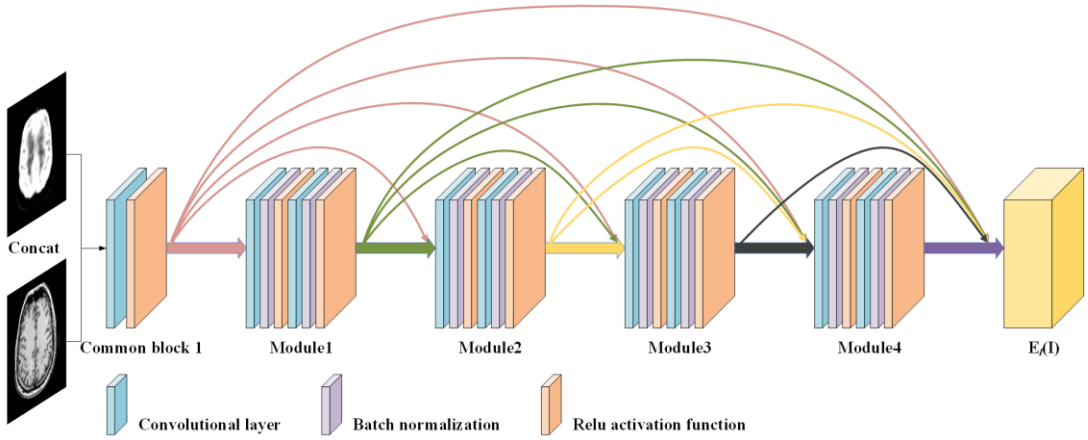

Fig. 2. Architecture of the Encoder based on the densely connected network. These color lines represent the skip connections between layers. The output is the encoded feature maps El(I).

A. *Encoder Network*

Firstly, we utilize an encoder network to extract the features of source images. Feature extraction is a significant procedure in deep learning-based image fusion approaches. To adequately employ the features of middle layers, we apply the dense connections with skip paths between layers close to the input and those close to the output. As is shown in Fig. 2, our encoder consists of two parts: the common block 1 and the dense block containing four modules. The common block 1 contains a convolutional layer with the kernel size of 3×3 and a Relu activation function. Each of these modules include two convolution layers with 3×3 filters and stride 1, and each of convolutional layer is followed by batch normalization and Relu activation function. The skip connections (those color curves in Fig. 2) are built between each module and all other modules in a feed-forward manner, which can enhance the feature representations and improve the computational efficiency. Different from the DenseNet [21], our network does not include the pooling layers that introduce down-sampling operation and drop out some detail information in the process of image fusion. It means that the input size of each layer is the same as the output size and can preserve the image features as much as possible when passing through the whole network. The input of our encoder is the concatenation of two single-channel source images I_1 and I_2. And in our network, the reflection mode is adopted to pad the input images. In this way, the input source images can be any size.

B. *Self-Attention Network*

The upper and lower parts in Fig. 3 represent the spatial and channel attention modules, respectively.

1) *Spatial attention module*

The channel number of input features of the attention network is C and the size of feature map is H×W. Firstly, these feature maps F_0 obtained from the encoder network are fed into the spatial attention module. Then, we adopt two convolutional layers with filter size of 1×1, to create two new feature maps $F_1 \in \mathbb{R}^{C'\times H\times W}$, $F_2 \in \mathbb{R}^{C'\times H\times W}$. Note that the number of output features C' is less than input feature channel number C to reduce dimension. For each spatial position in F_1, we can yield a vector $F_1(u) \in \mathbb{R}^{C'}$. At the same time, we can also collect a feature vector set $F_2(u) \in \mathbb{R}^{(H+W-1)\times C'}$ from F_2, which is located in the same row or column with position u. This operation is termed as collection. After that a multiplication (affinity operation) between F_1 and the transpose of F_2 is performed to merge these features. Subsequently, the spatial relationship can be described by a spatial attention matrix with the size of (H+W-1)×H×W utilizing a softmax function:

$$S_{i,u} = \frac{\exp\left(F_1(u)\cdot F_2^T(i,u)\right)}{\sum_{i=1}^{C'}\sum_{u}^{H+W-1}\exp\left(F_1(u)\cdot F_2^T(i,u)\right)}, \quad (1)$$

where $S_{i,u}$ represents the impact of the spatial position u in the ith channel; the symbol 'T' denotes the transpose operation; the larger $S_{i,u}$, the greater spatial attention at position u.

At the same time, we employ another convolution layer with the kernel size of 1×1 to generate the feature map $F_3 \in \mathbb{R}^{C\times H\times W}$. Likewise, we can also obtain the collection set of F_3,

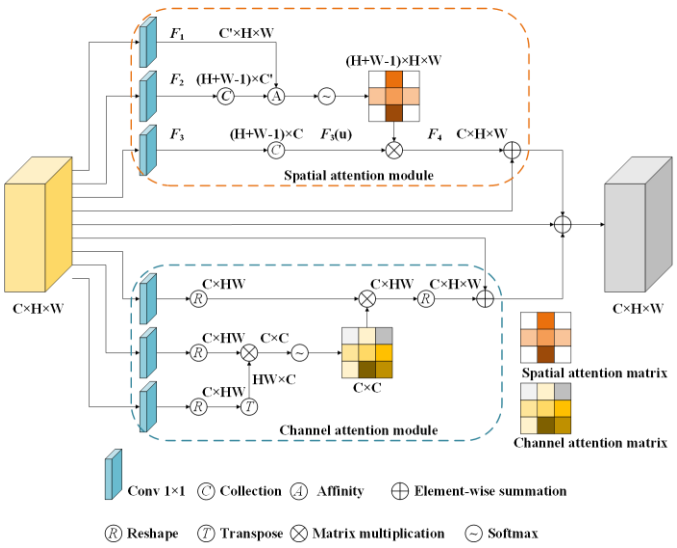

Fig. 3. The architecture of spatial attention module and channel attention module.

Authorized licensed use limited to: Communication University of China. Downloaded on September 02,2023 at 10:37:03 UTC from IEEE Xplore. Restrictions apply.

denoted by $F_3(u) \in \mathbb{R}^{(H+W-1)\times C}$. And then a matrix multiplication between $F_3(u)$ and the spatial attention matrix $S_{i,u}$ is performed and the result F_4 is summed with the input features F_0. Finally, the output F_5 of spatial attention module is obtained by performing an element-wise summation operation between F_4 and the original features F_0 as follows:

$$F_5 = \alpha F_4 + F_0(u) = \alpha \sum_{i=1}^{C} F_3(i,u) S_{i,u} + F_0(u), \quad (2)$$

were α is a scale parameter that balancing the tradeoff between F_4 and F_0. The feature at each position in F_5 is the weighted summation between the features of all positions at the same row and column with this position and original features. Hence, the global structural information is merged into local features according to the spatial attention module.

2) Channel attention module

Different from the spatial attention mechanism, channel attention strategy mainly focuses on the interdependencies between different channels of feature maps to improve the ability of feature representation. The original feature F_0 is fed into three 1×1 convolutional layers and obtain three channel attention maps C_1, C_2 and $C_3 \in \mathbb{R}^{C\times H\times W}$. Firstly, we reshape these channel maps C_1, C_2 and C_3 to C×HW. Then, the channel attention matrix can be calculated through performing a matrix multiplication between C_1 and the transpose of C_2, and expressed as the following softmax function:

$$C_{yx} = \frac{\exp\left(C_1(x)\cdot C_2^T(y)\right)}{\sum_{x=1}^{C}\exp\left(C_1(x)\cdot C_2^T(y)\right)}, \quad (3)$$

where C_{yx} measures the impact of channel x on channel y. In addition, the channel attention matrix C_{yx} is multiplied with C_3 and the result C_4 is reshaped to the same size of input features C×H×W. Ultimately, C_4 is summed with the input original feature maps F_0 via a weighted parameter to get the output of channel attention module:

$$C_5 = \beta \sum_{x=1}^{C} C_4 + F_0 \quad (4)$$

This formula demonstrates that the final feature of each channel is a weighted sum of the features of all channels and original feature maps. These operations can model the dependencies between the channels of feature maps. To sum up, the self-attention mechanism including spatial attention and channel attention can help to facilitate the discriminability of deep features.

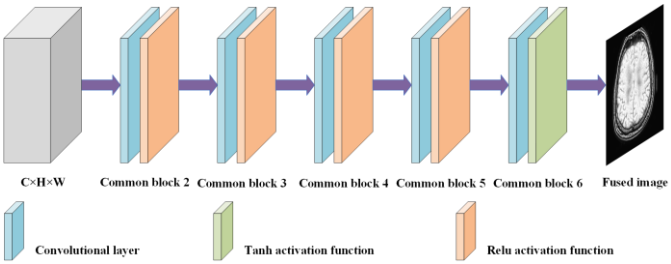

Fig. 4. The detailed structure of the decoder network. The input is feature maps obtained by channel and spatial attention network with the same size as $E_i(I)$.

C. Decoder Network

The last component of our proposed fusion model is reconstructing the final fused image from these enhanced features obtained by the self-attention network. The decoder network consists of five blocks. Each of the former four blocks includes one convolution layer and one Relu layer, and the last block is composed of one convolutional layer and Tanh as an activation function. In the decoder network, all filter size and strides are set as 3×3 and 1, along with the reflection padding mode used before convolution operation to avoid the information loss. The input of the decoder network is the feature maps with the channel number C and size of H×W and the output result is the final constructed fusion image. In the whole network, the sizes of feature maps are constant, which means that it can avoid information lost and invasion for image fusion during down-sampling and up-sampling processes.

D. Loss function

In this section, we will discuss the loss function used in our proposed fusion model. The objective of image fusion is to reconstruct the high-quality fusion image containing more information from two source images. In deep learning, mean square error (MSE) is frequently-used loss function to constrain the prediction of model close to the ground-truth output. This loss function constrains the intensity information between the fusion image and two source images. In order to reconstruct the fusion image more precisely, we introduce the structural similarity into the total loss. Therefore, we minimize the following loss function L to train our fusion model.

$$L = L_{mse} + \gamma L_{ssim} = L_{mse} + \gamma(1 - SSIM) \quad (5)$$

In formula (5), SSIM denotes the structural similarity index measure and it measures the structural similarity of two images. The total loss function is a weighted combination of the MSE loss L_{mse} and the SSIM loss L_{ssim} with the weight parameter γ.

III. Experimental Results

To evaluate the performance of the proposed image fusion model, the fusion experiments were conducted on one publicly available dataset including PET and MRI images. In Section *A*, the dataset and the setting of parameters are described. Then, we demonstrate the qualitative analyses of our fusion model in Section *B*. The quantitative performance of our model in terms of these assessment metrics is exemplified in Section *C* and compared with those of state-of-the-art methods. Our proposed image fusion model is implemented in the Pytorch framework [27]. It is worth noting that all deep learning-based methods run on the same NVIDIA GPU TESLA T4, while other fusion methods run on the same CPU i7-8565.

A. Dataset and Parameter Setting

The dataset used for medical image fusion in this work is obtained from the database of Whole Brain Atlas [28]. Among these data, 58 pairs of PET and MRI images are selected as the training data and 20 pairs are employed as test images to evaluate the performance of our fusion model. All these pairs

Authorized licensed use limited to: Communication University of China. Downloaded on September 02,2023 at 10:37:03 UTC from IEEE Xplore. Restrictions apply.

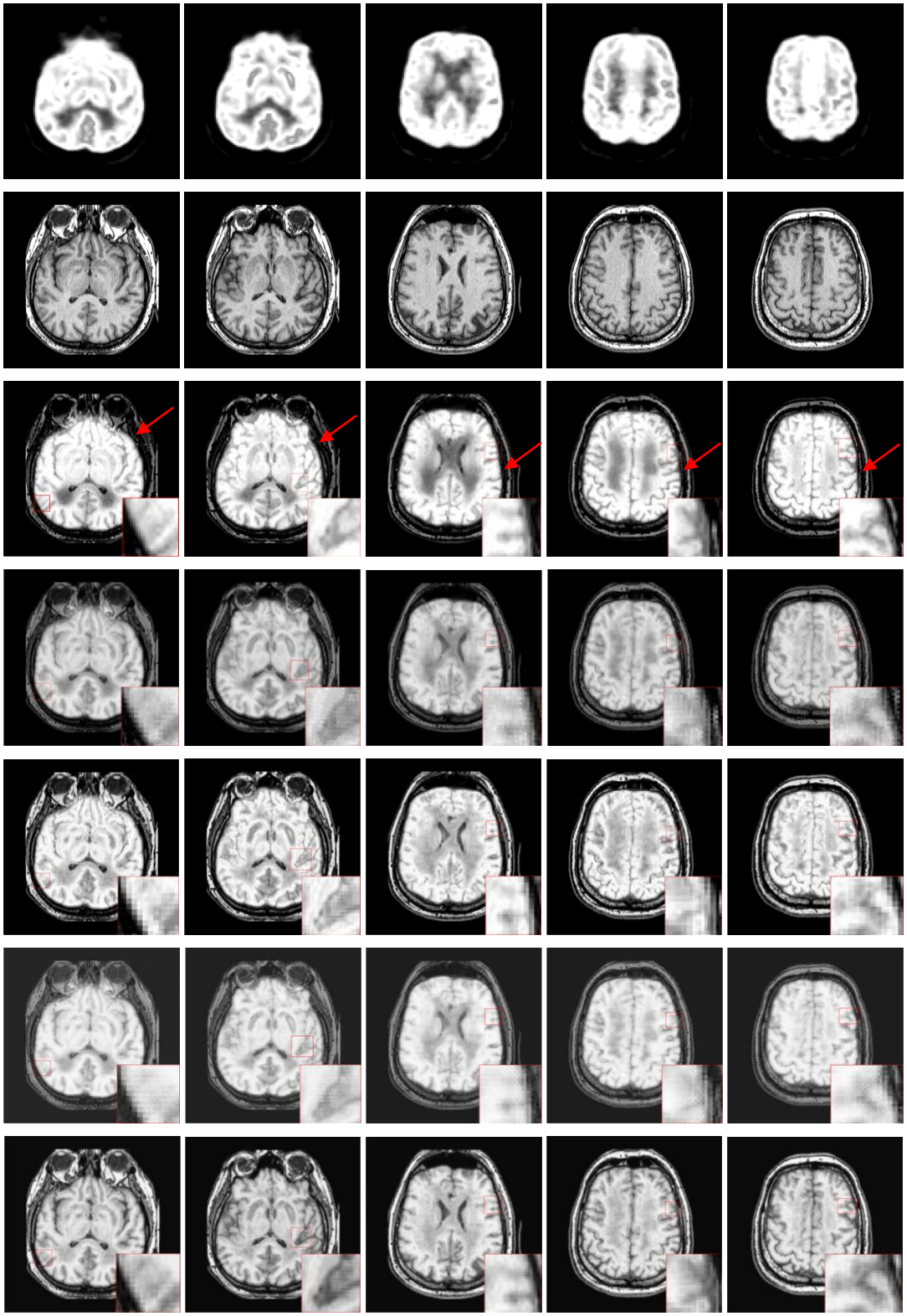

Fig. 5. Qualitative results of PET and MRI image fusion. From top to bottom: PET images, MRI images, and fusion images of GTF, MSVD, IFCNN, TFNet and our proposed CSpA-DN.

Authorized licensed use limited to: Communication University of China. Downloaded on September 02,2023 at 10:37:03 UTC from IEEE Xplore. Restrictions apply.

of images have been pre-registered with the same spatial size of 256×256. In the training process, the source images are cropped to patches of size 64×64. The batch size is set as 64 and the number of epochs is 600.

To evaluate the quality of the fused image and assess the performance of different fusion methods, four objective evaluation metrics are considered to evaluate quantitatively our fusion model and other methods in this work, i.e., entropy measure (EN) [29], Xydeas and Petrovic metric (Q_{abf}) [30], feature mutual information using image pixels (FMI_pixel) [31], FMI with discrete cosine transformation (FMI_dct).

EN is devoted to assessing the characteristics of fused images, and measure the amount of information contained in the fusion results. On the other hand, the other three metrics, Q_{abf}, FMI_pixel and FMI_dct are exploited to evaluate the correlations between the fusion images and two source images. Specifically, Qabf metric is mainly leveraged to evaluate the edge information that is transformed from the source images to fused images. FMI_pixel and FMI_dct focus on measuring structural and textural information in spatial and frequency domains, respectively. For each metric, a higher score demonstrates a better fusion performance.

B. *Qualitative Assessments*

The qualitative results are displayed in Figure 5, consisting of five typical and intuitive fusion results on five different axial planes of brain PET (see the 1st row of Fig. 5) and MRI (the 2nd row of Fig. 5) images. Among these fused images, the fusion results of gradient transfer fusion (GTF) method [32] can retain much more functional information (shown in the 3rd row of Fig. 5) of PET images than other fusion approaches, but GTF method loses much textural and structural information of MRI images (areas marked by red arrows). Although these fusion images generated by MSVD (the 4th row of Fig. 5) and TFNet (the 6th row of Fig. 5) can preserve the structural information from MRI images, these two fusion models also reduce the contrast in the MRI images. Furthermore, these results obtained

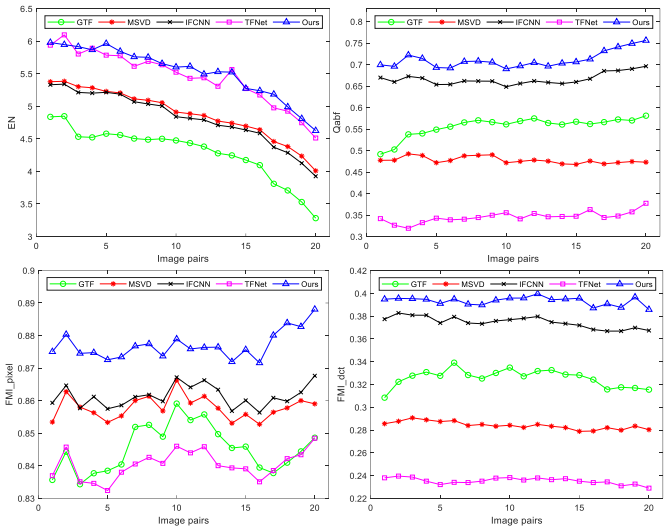

Fig. 6. Objective assessment scores of four fusion metrics using five methods on twenty pairs of PET and MRI images.

TABLE I. MEAN VALUES OF ASSESSMENT RESULTS ON PET AND MRI IMAGE FUSION. BOLD BLACK INDICATES THE BEST.

	EN	Q_{abf}	FMI_pixel	FMI_dct
GTF	4.2897	0.5562	0.8453	0.3256
MSVD	4.8821	0.4776	0.8578	0.2841
IFCNN	4.8187	0.6663	0.8613	0.3746
TFNet	5.4555	0.3457	0.8404	0.2353
Ours	**5.5289**	**0.7111**	**0.8770**	**0.3933**

created by IFCNN (the 5th row of Fig. 5) and our CSpA-DN model (the last row of Fig. 5) not only retain intensity values and significant structural information from MRI images, but also preserve rich functional information in PET images. In addition, the textural details of our CSpA-DN fusion results are more distinct than those of IFCNN, as illustrated in the highlighted regions (red boxes).

C. *Quantitative Evaluations*

The quantitative assessments of our fusion model and other four methods such as GTF [32], MSVD [8], IFCNN [15], TFNet [16]. These approaches include two traditional fusion algorithms and two deep learning-based fusion technologies. We also employ the objective metrics introduced in section *A* to evaluate our fused results, and different metrics are utilized for different evaluation emphases.

The quantitative fused results of twenty pairs of test images are shown in Fig. 6, where five color curves in each subfigure represent results of five fusion methods. The assessment score of each method over twenty pairs of source images is provided. It is observed that our proposed fusion method outperforms other four algorithms on all test image pairs except for EN value of image pair No.2.

Table I shows the mean evaluation scores of twenty tests for each metric by using five fusion methods. As can be observed from these results, our fusion method achieves the highest mean values on four metrics and indicates a better performance. The highest mean assessment scores also demonstrate that our fusion results preserve more information, structural details, stronger contrast, and a higher similarity with source images.

D. *Ablation study*

In this section, we carry out the ablation experiments to demonstrate the effectiveness of the self-attention mechanism in our proposed fusion model. In order to illustrate the effect, the following experiments are performed. The fusion framework that only contains the encoder and decoder is performed on the training and test data. The self-attention network consisting of spatial and channel attention modules is inserted into the fusion model to enhance the structural information of the fusion images. The experimental settings of two comparative experiments are the same and the results are shown in Fig. 7. The functional information of PET images is preserved in the fused results obtained in the absent and present of self-attention network (shown in the third and fourth rows of Fig. 7). Nevertheless, the fused images shown in the third row obtained in the absent of self-attention almost loss the structural information from MRI images. On the contrary, those fused results in the fourth row can effectively

Authorized licensed use limited to: Communication University of China. Downloaded on September 02,2023 at 10:37:03 UTC from IEEE Xplore. Restrictions apply.

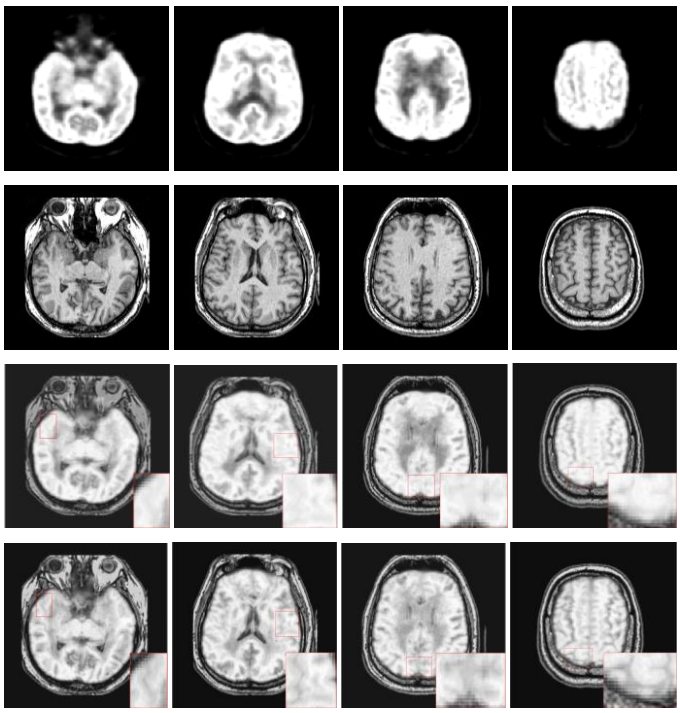

Fig. 7. Results on whether the self-attention mechanism exists in our fusion model. From top to bottom: PET images, MRI images, and the fusion images obtained in the absent and present of self-attention, respectively.

retain more structural information from MRI images. In consequence, this ablation experiment demonstrates that the self-attention mechanism can further merge the local features from the source images to the fused results.

IV. Conclusion

In this work, we propose a novel fusion framework, named CSpA-DN, based on a densely connected network with channel and spatial attention for PET and MR images. Our algorithm is an end-to-end model with the encoder-decoder architecture, in which the densely connected neural network is constructed to extract features from two source images and a decoder network is employed to generate the fused image. Moreover, a self-attention mechanism is introduced in the encoder and decoder to further integrate local features along with their global dependencies adaptively. The output features at each position of the encoder is generated by a weighted summation of those features at all positions employing a spatial attention module. At the same time, the interdependent relationship of all feature maps is integrated by a channel attention module. The summation of the outputs of these two modules is fed into the decoder and the fused image is generated. Experimental results on twenty pairs of test images demonstrate the better performance of our proposed CSpA-DN fusion model compared with other four fusion approaches for PET and MR images, according to both qualitative assessments and quantitative evaluations by using four objective metrics. In addition, the ablation experiments illustrate that the self-attention mechanism in our fusion model can effectively preserve more structural information from the source images.

Acknowledgment

This research was supported by the National Natural Science Foundation of China (No. 61901537), China Postdoctoral Science Foundation (No.2020M672274), Scientific and technological program projects of Henan Province (No. 192102210127), Science and technology guiding project of China National Textile and Apparel Council (No. 2019059), Program of Young backbone teachers in Zhongyuan University of Technology (No. 2019XQG04).

References

[1] Liu Y, Chen X, Cheng J, et al. A medical image fusion method based on convolutional neural networks[C]//2017 20th International Conference on Information Fusion (Fusion). IEEE, 2017: 1-7.

[2] Yin M, Liu X, Liu Y, et al. Medical image fusion with parameter-adaptive pulse coupled neural network in nonsubsampled shearlet transform domain[J]. IEEE Transactions on Instrumentation and Measurement, 2018, 68(1): 49-64.

[3] James A P, Dasarathy B V. Medical image fusion: A survey of the state of the art[J]. Information fusion, 2014, 19: 4-19.

[4] Du J, Li W, Lu K, et al. An overview of multi-modal medical image fusion[J]. Neurocomputing, 2016, 215: 3-20.

[5] Du J, Li W, Xiao B, et al. Union Laplacian pyramid with multiple features for medical image fusion[J]. Neurocomputing, 2016, 194: 326-339.

[6] Li H, Manjunath B S, Mitra S K. Multisensor image fusion using the wavelet transform[J]. Graphical models and image processing, 1995, 57(3): 235-245.

[7] Lewis J J, O'Callaghan R J, Nikolov S G, et al. Pixel-and region-based image fusion with complex wavelets[J]. Information fusion, 2007, 8(2): 119-130.

[8] Naidu V P S. Image fusion technique using multi-resolution singular value decomposition[J]. Defence Science Journal, 2011, 61(5): 479-484.

[9] Bhatnagar G, Wu Q M J, Liu Z. Directive contrast based multimodal medical image fusion in NSCT domain[J]. IEEE transactions on multimedia, 2013, 15(5): 1014-1024.

[10] Guorong G, Luping X, Dongzhu F. Multi-focus image fusion based on non-subsampled shearlet transform[J]. IET Image Processing, 2013, 7(6): 633-639.

[11] Alom M Z, Taha T M, Yakopcic C, et al. The history began from alexnet: A comprehensive survey on deep learning approaches[J]. arXiv preprint arXiv:1803.01164, 2018.

[12] Liu Y, Chen X, Peng H, et al. Multi-focus image fusion with a deep convolutional neural network[J]. Information Fusion, 2017, 36: 191-207.

[13] Liu Y, Chen X, Cheng J, et al. A medical image fusion method based on convolutional neural networks[C]//2017 20th International Conference on Information Fusion (Fusion). IEEE, 2017: 1-7.

[14] Prabhakar K R, Srikar V S, Babu R V. DeepFuse: A Deep Unsupervised Approach for Exposure Fusion with Extreme Exposure Image Pairs[C]//ICCV. 2017: 4724-4732.

[15] Zhang Y, Liu Y, Sun P, et al. IFCNN: A general image fusion framework based on convolutional neural network[J]. Information Fusion, 2020, 54: 99-118.

[16] Liu X, Liu Q, Wang Y. Remote sensing image fusion based on two-stream fusion network[J]. Information Fusion, 2020, 55: 1-15.

[17] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[18] Huang G, Sun Y, Liu Z, et al. Deep networks with stochastic depth[C]//European conference on computer vision. Springer, Cham, 2016: 646-661.

[19] Larsson G, Maire M, Shakhnarovich G. Fractalnet: Ultra-deep neural networks without residuals[J]. arXiv preprint arXiv:1605.07648, 2016.

Authorized licensed use limited to: Communication University of China. Downloaded on September 02,2023 at 10:37:03 UTC from IEEE Xplore. Restrictions apply.

[20] Srivastava R K, Greff K, Schmidhuber J. Training very deep networks[C]//Advances in neural information processing systems. 2015: 2377-2385.

[21] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.

[22] Li H, Wu X J. Densefuse: A fusion approach to infrared and visible images [J]. IEEE Transactions on Image Processing, 2018, 28(5): 2614-2623.

[23] Xu H, Ma J, Le Z, et al. FusionDN: A unified densely connected network for image fusion [C]//Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence. 2020.

[24] Fu J, Liu J, Tian H, et al. Dual attention network for scene segmentation [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 3146-3154.

[25] Mou L, Zhao Y, Chen L, et al. CS-Net: Channel and Spatial Attention Network for Curvilinear Structure Segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2019: 721-730.

[26] Huang Z, Wang X, Huang L, et al. CCnet: Criss-cross attention for semantic segmentation[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019: 603-612.

[27] https://pytorch.org/

[28] http://www.med.harvard.edu/AANLIB/home.html

[29] Liu Y, Liu S, Wang Z. A general framework for image fusion based on multi-scale transform and sparse representation[J]. Information fusion, 2015, 24: 147-164.

[30] Xydeas C S, Petrovic V. Objective image fusion performance measure[J]. Electronics letters, 2000, 36(4): 308-309.

[31] Haghighat M, Razian M A. Fast-FMI: non-reference image fusion metric[C]//2014 IEEE 8th International Conference on Application of Information and Communication Technologies (AICT). IEEE, 2014: 1-3.

[32] Ma J, Chen C, Li C, et al. Infrared and visible image fusion via gradient transfer and total variation minimization[J]. Information Fusion, 2016, 31: 100-109.

Authorized licensed use limited to: Communication University of China. Downloaded on September 02,2023 at 10:37:03 UTC from IEEE Xplore. Restrictions apply.