Neurocomputing 267 (2017) 406–416

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Deep auto-context convolutional neural networks for standard-dose PET image estimation from low-dose PET/MRI

Lei Xiang[a], Yu Qiao[b], Dong Nie[c], Le An[c], Weili Lin[c], Qian Wang[a,*], Dinggang Shen[c,d,*]

[a] Med-X Research Institute, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
[b] Shenzhen key lab of Comp. Vis. & Pat. Rec., Shenzhen Institutes of Advanced Technology, CAS, Shenzhen, China
[c] Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[d] Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, Republic of Korea

ARTICLE INFO

Article history:
Received 19 December 2016
Revised 1 May 2017
Accepted 9 June 2017
Available online 29 June 2017

Communicated by Bo Du

Keywords:
PET image restoration
Deep convolutional neural network
Auto-context strategy

ABSTRACT

Positron emission tomography (PET) is an essential technique in many clinical applications such as tumor detection and brain disorder diagnosis. In order to obtain high-quality PET images, a standard-dose radioactive tracer is needed, which inevitably causes the risk of radiation exposure damage. For reducing the patient's exposure to radiation and maintaining the high quality of PET images, in this paper, we propose a deep learning architecture to estimate the high-quality standard-dose PET (SPET) image from the combination of the low-quality low-dose PET (LPET) image and the accompanying T1-weighted acquisition from magnetic resonance imaging (MRI). Specifically, we adapt the convolutional neural network (CNN) to account for the two channel inputs of LPET and T1, and directly learn the end-to-end mapping between the inputs and the SPET output. Then, we integrate multiple CNN modules following the auto-context strategy, such that the tentatively estimated SPET of an early CNN can be iteratively refined by subsequent CNNs. Validations on real human brain PET/MRI data show that our proposed method can provide competitive estimation quality of the PET images, compared to the state-of-the-art methods. Meanwhile, our method is highly efficient to test on a new subject, e.g., spending ~2 s for estimating an entire SPET image in contrast to ~16 min by the state-of-the-art method. The results above demonstrate the potential of our method in real clinical applications.

1. Introduction

Positron emission tomography (PET) is a functional imaging technique, which produces 3D *in-vivo* observation of the metabolic process in the body. It provides molecular information on the biology of many diseases. Accordingly, PET has been increasingly recognized as an important tool for diagnosis [1,2], determination of prognosis [3,4], and response monitoring in oncology [5,6]. There are also other imaging technologies, such as computed tomography (CT) and magnetic resonance imaging (MRI). Recently the introduction of PET/CT and PET/MRI scanners enables the acquisition of both structural and functional information in a single scan session.

The high-quality PET images play a crucial role in diagnosing brain diseases and disorders [7], because they can provide detailed functional information for assessment and diagnosis. In order to obtain the high-quality PET images, a standard-dose tracer injection to tissue or organ is needed, which inevitably raises the risk of radioactive exposure. To address this problem, the well-known As Low As Reasonably Achievable (ALARA) [8] principle is adopted to minimize the radiation exposure in clinical practice. Although the principle helps to decrease the risk of radiation exposure, it also degrades the quality of PET images and potentially involves unnecessary noises and artifacts. Two examples of the low-dose PET (LPET) and their corresponding standard-dose PET (SPET) images are shown in Fig. 1. It can be observed that the quality of the LPET images is worse than that of the SPET images.

In order to improve the quality of the acquired PET images, numerous reconstruction and denoising methods have been developed. Mejia et al. [9] proposed a multi-resolution approach for noise reduction of PET images by employing specific filters to homogeneous and heterogeneous image regions. Pogam et al. [10] succeeded in addressing the issue of resolution loss with standard denoising by combining the complementary wavelet and curvelet transforms. Bagci and Mollura [11] used the singular value thresholding concept and the Stein's unbiased risk estimation method to optimize the soft thresholding rule for denoising. These techniques are mainly designed for SPET images only.

* Corresponding author.
E-mail address: xianglei_15@sjtu.edu.cn (L. Xiang).

http://dx.doi.org/10.1016/j.neucom.2017.06.048
0925-2312/

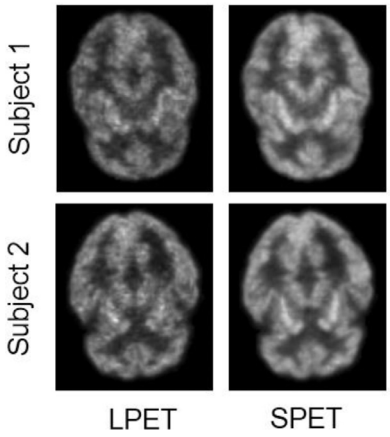


Fig. 1. Two examples of the LPET images and their corresponding SPET images.

However, our objective here is to estimate the SPET image from the corresponding LPET image, which is acquired with low-dose tracer injection. Similar works can be found for the quality enhancement of CT images. For example, Gervaise et al. [12] proposed an adaptive iterative dose reduction (AIDR) method to achieve the high-quality images, while reducing the radiation dose in CT acquisition.

Multi-modality data has been proven to provide complementary and effective information for increasing the quality of each single modality [13,14]. It is shown in the literature that the anatomical or the structural information (e.g., from CT or MRI [15,16]) contributes to better SPET image quality. In our work, we utilize both the LPET images and the corresponding structural T1 images for the estimation of the high-quality SPET images. We will detail the way to combine T1 images and LPET images using convolutional neural network (CNN) to estimate SPET images in Section 3.1.

In this paper, we first use a basic four-layer CNN to build a relatively simple model, which derives the SPET image from the LPET image and the T1 image. As an end-to-end architecture, the deep network maps the LPET and the T1 inputs to the SPET output directly without using handcrafted features. Then, we treat the tentatively estimated SPET image as the source of the context information [17]. In addition to the context information, both original LPET and T1 images are also used as inputs to a new four-layer CNN. In this way, we gradually concatenate multiple CNNs into a much deeper network. The entire network, which consists of multiple four-layer CNNs, can be optimized altogether with back-propagation. The experimental results reveal that the proposed method can effectively utilize the structural information in T1 image for the estimation of the high-quality SPET image. Meanwhile, the auto-context [18] strategy allows us to gradually improve the quality of the SPET estimation, given multiple four-layer basic CNNs. In general, our method achieves competitive performance regarding the quality of the estimated SPET images while its time cost is significantly reduced compared to the state-of-the-art methods.

The rest of this paper is organized as follows. We will review the related work in Section 2, and then describe the details of our proposed method in Section 3. Section 4 quantitatively analyzes key components of the proposed method and conducts comparisons with the state-of-the-art methods. The conclusions are drawn in Section 5.

2. Related work

Research efforts have been made in the literature to directly estimate the SPET images from the LPET images. The estimation often requires the input of the tracer-free MRI scan and relies on the sparse learning technique. For example, in [14], the mapping-based sparse representation (m-SR) was adopted for SPET image reconstruction. To speed up the process, the patch-selection-based dictionary construction method was used to build a relatively small but representative dictionary, which can heavily reduce the processing time. Subsequently, a semi-supervised tripled dictionary learning method was used for SPET image reconstruction [19]. This method can improve the prediction results by utilizing multiple modalities (i.e., T1 image, fractional diffusivity and mean diffusivity from diffusion weighted data). It also allows a certain modality to be missing, thus including huge clinical data for training. Recently, An et al. [20] proposed the data-driven multi-level canonical correlation analysis (MCCA) scheme to map the SPET and the LPET image data into a common space, where the patch-based sparse representation was then utilized to generate the coupled LPET and SPET dictionaries. These sparse-learning-based methods consist of several steps generally, including patch extraction, encoding, and reconstruction. Most of these methods are time-consuming particularly when testing new cases, which have to solve a large number of optimization problems and thus might not be applicable in real clinical practice.

CNN dates back to decades [21], and deep CNNs have shown an explosive popularity partially due to its success in image classification tasks [22,23]. This technique has been successfully applied to many computer vision fields, such as face detection [24–26], semantic segmentation [27,28], and object tracking [29–31]. There are also some successful applications in medical image fields, such as cell detection [32,33] and prostate segmentation [34,35]. There are several factors that lead to its success: (i) the efficient implementation on modern powerful GPUs to train large networks with huge number of parameters [23], (ii) the proposal of useful tricks like Rectified Linear Unit (ReLU) [36] and dropout [37] that avoid the problems of gradient vanish and overfitting, and (iii) an abundance of labeled data (like ImageNet [38]) for training deep architectures. Recently, the proposed mechanism called batch normalization [39] also helps to speed up convergence in training very deep neural networks, leading to better performance. Specifically, Li et al. [40] proposed a deep-learning-based imaging data completion method to predict PET image from structural MRI image. Our method differs from this method in two ways. First, we apply deep neural network to estimate SPET by using multiple modalities, i.e., LPET and T1 images. Second, compared to [40], which has only three convolution layers, our network is much deeper and effectively leverage the auto-context information for the purpose of SPET estimation.

Recently, Dong et al. [41] presented a method namely Super-Resolution Convolutional Neural Network (SRCNN) for single image super-resolution, which directly learns an end-to-end mapping between low-resolution and high-resolution images. This model, which takes the low-resolution image as input and outputs the high-resolution one, partly inspired our work for SPET image estimation from the LPET image. However, different from Dong's work, we propose to incorporate the structural T1 image in the input layer of the CNN architecture, and refine the estimation of the SPET image iteratively in an auto-context way based on the inputs of multiple modalities, which makes our model much deeper compared to Dong's model.

3. Method

We present the details of our deep CNNs for SPET estimation in this section. We first introduce the basic multi-modal CNN, which maps the inputs of LPET and T1 to the output of SPET within four convolution layers only. Then, we concatenate multiple basic CNN modules into a deeper network following the auto-context fashion,

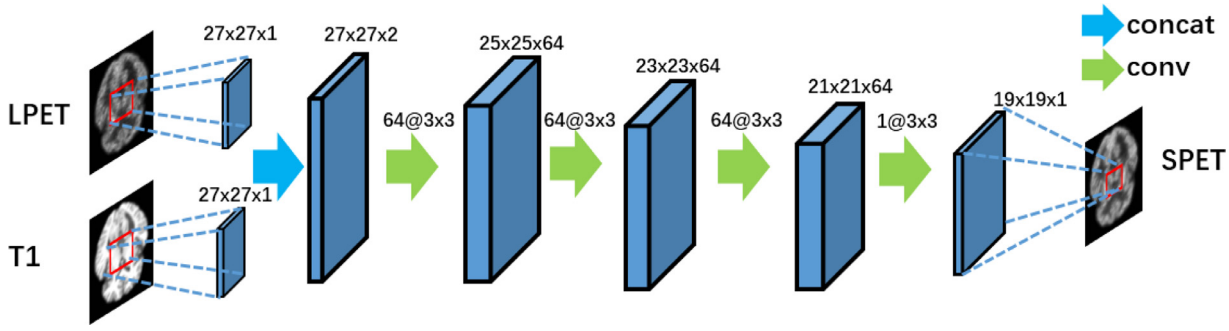


Fig. 2. The architecture of the basic four-layer CNN used to estimate SPET from LPET and T1 images. The inputs include two feature maps corresponding to LPET and T1 image patches, respectively. The output is the corresponding SPET image patch. There are four convolution layers in this basic CNN model.

such that the tentative SPET estimation can be iteratively refined with the help of the context information and the original LPET/T1 input images.

3.1. The basic multi-modality CNN architecture

In this work, we propose to use the CNN model for estimating the SPET image from LPET and T1 images. Our work is motivated by the fact that, in addition to the low-quality functional data, structural T1 images can help the estimation of the high-quality functional images. Although CNNs have been used for similar tasks in the literature, it is still challenging to fuse multiple medical image modalities. To this end, we treat multi-modality images as different feature maps, and input them to CNN after concatenation. In this way, we present a straightforward solution for combining multi-modality image data. Since T1 image contains complementary information other than the functional PET data, our CNN architecture is capable of better estimating SPET from LPET and T1 images.

Considering the limited number of training images, we switch to solve the problem slice by slice here. That is, we extract all axial slices and treat them as separate images independently in training. For a new test subject, we estimate all slices and then stack them into the 3D volume along the inferior-superior direction. Our experiments confirm that the final results are observably satisfactory along the inferior-superior direction, as shown in Fig. 9.

The multi-modality CNN, whose architecture is shown in Fig. 2, aims to learn the end-to-end mapping between the input LPET and T1 images and the output SPET image. Note that there are two input feature maps of this CNN in the input layer, corresponding to T1 image and LPET image, respectively. The network consists of four convolution layers, without using any pooling. The main reason is that pooling is commonly used in recognition and classification for reducing the dimension of feature maps and also making the network invariant to small translation of the input. Therefore, pooling might not be suitable for pixel-wise image quality enhancement in this work. On the other hand, the convolution layers provide similar functions regarding sparse coding, including patch extraction and representation, non-linear mapping, and reconstruction [42].

In our basic multi-modality CNN model, we concatenate the two patches of LPET and T1 in the input layer, followed by four convolution layers. The first convolution layer contains n_1 filters of the support $m \times f_1 \times f_1$, where m is the number of the feature maps (with $m = 2$ here), and $f_1 \times f_1$ denotes the spatial size of the filter. In general, the first layer can be expressed as

$$\max(0, W_1 * [Y, Z] + B_1), \tag{1}$$

where $*$ represents the convolutional operator, Y and Z denote the LPET and T1 image patches respectively, and $[\cdot, \cdot]$ means the concatenation operation that combines two patches. W_1 and B_1 denote the filters and the biases, respectively. Intuitively, W_1 applies n_1 convolution filters on the input image patches, each of which has a kernel size of $m \times f_1 \times f_1$. The output thus consists of n_1 feature maps.

The second, third, and fourth convolution layers can be configured in the similar way. For example, we set the second convolution layer to contain n_2 filters of the size $n_1 \times f_2 \times f_2$. So the parameters of the second layer can be represented as W_2 and B_2. After the second convolution layer, we will get n_2 feature maps as the output. Eventually, in the fourth convolution layer, there is only one filter ($n_4 = 1$). The single output of the fourth layer corresponds to the expected output of the SPET image patch, which shares the same center location with the input LPET and T1 image patches. All other parameters of individual layers are shown in Fig. 2. In particular, we set $m = 2$, $n_1 = n_2 = n_3 = 64$, and $f_1 = f_2 = f_3 = 3$. We do not use any padding in each convolution layer, so the sizes of the feature maps decrease when the layer becomes deeper. For example, as shown in Fig. 2, the original size of the input LPET in training is 27 $\times$ 27, and the size of the output is 19 $\times$ 19.

Let us denote the output image estimated by the basic four-layer CNN as $F_{basic}(Y_i, Z_i; \theta_{basic})$. Here, F indicates the end-to-end mapping, and $\theta_{basic} = \{W_1, W_2, W_3, W_4, B_1, B_2, B_3, B_4\}$ records the estimated network parameters. We term X_i as the ground-truth SPET for the ith training subject image patch. The input LPET and T1 image patches are denoted as Y_i and Z_i, respectively. θ_{basic} can thus be solved by minimizing the error between the reconstructed output $F_{basic}(Y_i, Z_i; \theta_{basic})$ and the corresponding ground-truth X_i of the same size with that of the output for training. We use the Mean Squared Error (MSE) as the loss function:

$$L_{basic}(\theta_{basic}) = \frac{1}{M} \sum_{i=1}^{M} \| F(Y_i, Z_i; \theta_{basic}) - X_i \|^2, \tag{2}$$

where M is the number of the training image patches. We use stochastic gradient descent with the standard back-propagation [43] to minimize the loss function. Using the L_2 loss function favors a high Peak Signal to Noise Ratio (PSNR). Note that PSNR is a widely used metric for quantitatively evaluating image restoration quality, as it is related to the perceptual quality. Our goal is to make the estimated SPET and the ground-truth SPET as similar as possible.

Note that the input/output sizes shown in Fig. 2 apply to the training process only. In testing, we treat the trained CNN model as fully convolutional network (FCN) [41] which can take the entire LPET and T1 images as inputs. This operation avoids to apply CNN for each patch independently and can save large computational cost. Since there is no padding in each convolution layer, we apply zero padding to the input test image to make sure that the sizes of the input image and the final output image are the same.

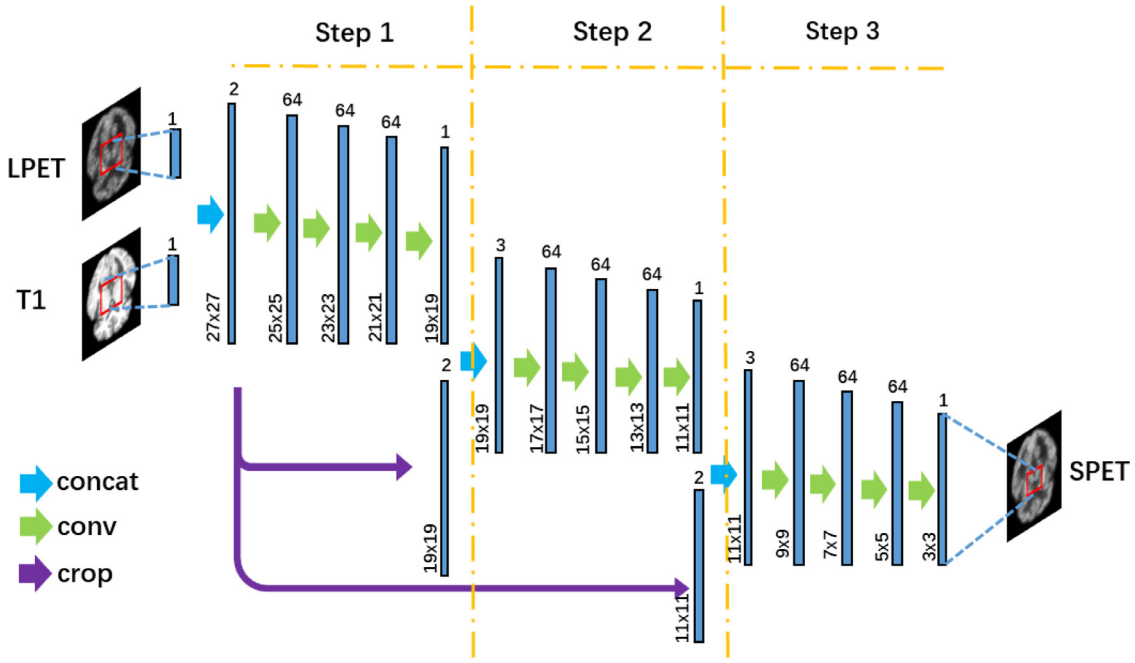


Fig. 3. Illustration of the deep auto-context CNN architecture. 'Concat' represents concatenation operation that concatenates individual feature maps. 'Conv' represents the convolutional operation. 'Crop' represents the crop operation that keeps the sizes of different feature maps consistent. In Step 1, the inputs of the basic four-layer CNN are LPET and T1 images. In Step 2 and Step 3, the tentatively estimated SPET image from the last step is also included as an additional input.

For example, if the size of the input test image is 100×100 and it intends to pass four convolution layers of many 3×3 filters, we pad the input image and augment its size to 108×108 prior to the first convolution layer. In this way, the final output image will reduce to the size of 100×100.

Meanwhile, batch normalization was recently introduced by Ioffe and Szegedy [39] to ease the training of deep neural networks. It reflects the fact that neural networks tend to learn more efficiently when their inputs are normalized to zero mean with unit variance. This strategy can be extended to the internal layer of CNNs. To this end, we apply batch normalization for every convolution layer in our implementation. For each convolution layer in Fig. 2, the output from the precedent layer can thus be processed through batch normalization and then feed as the input to the subsequent convolution layer.

3.2. Deep auto-context CNNs for SPET estimation

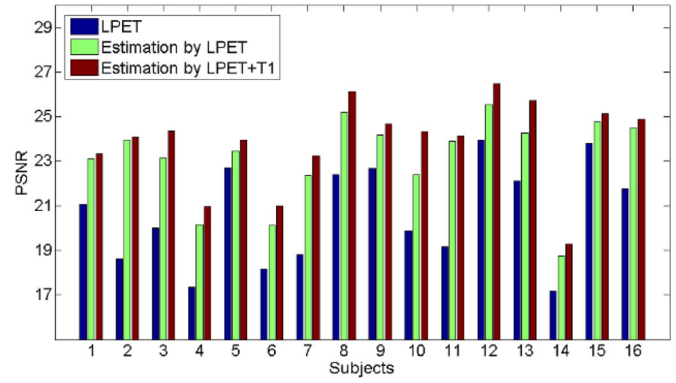
We propose to concatenate multiple CNNs to formulate a much deeper structure, to improve the quality of the estimated SPET image gradually. The concatenated CNNs, which are shown in Fig. 3, lead to a deep auto-context-like learning architecture [17,18,51–53]. First, we use the basic four-layer CNN (shown in Fig. 2) to estimate the SPET image based on both LPET and T1 images. Then, the tentatively estimated SPET, along with the original LPET and T1 images, are all input to the subsequent new four-layer CNN. That is, there are three input channels for the second and latter CNNs, i.e., the tentatively estimated SPET, LPET, and T1 images.

In our implementation, we concatenate three four-layer CNNs to formulate the deep structure. The output of the 1st CNN (namely after "Step 1") is combined with the original LPET and T1 images, which are cropped from the center to get the same size with the output of CNN 1. The 2nd (Step 2) and the 3rd (Step 3) CNNs share the same architecture with Step 1, though the numbers of the input feature maps vary slightly as in Fig. 3. The sizes of the outputs of the 2nd and the 3rd CNNs are 11×11 and 3×3 in training, respectively.

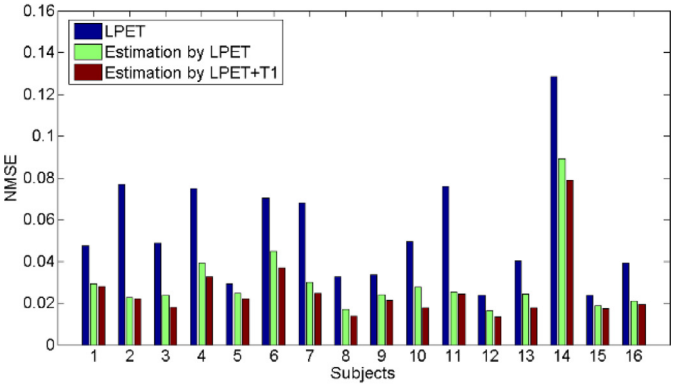
With three four-layer CNNs concatenated in our implementation, there are totally 12 convolution layers, which make our architecture deep enough to estimate SPET from LPET and T1. Note that our deep architecture is significantly different from the conventional 12-layer convolutional neural network. Specifically, the LPET and T1 image inputs are forcefully directed to Step 2 and Step 3 in our method. Concerning the high similarity between LPET and SPET, the estimation of SPET can easily be dominated by LPET while ignoring T1 image. Meanwhile, the learning may end within a few convolution layers (e.g., only four layers in Step 1) as the mapping from LPET to SPET is not complex. With concatenated CNNs, T1 images are directly used as the inputs of each step and thus can play a more important role in the estimation of SPET even though its appearance is very different from SPET (compared with LPET especially). The influence of the structural information from T1 now can arrive at the very deep layers in our architecture through the concatenated CNNs. For fair comparison, in Section 4.4, we will conduct experiments with the conventional 12-layer CNN, where our method clearly shows better SPET estimation capability than simply increasing the number of layers in CNN.

The concatenation of CNNs also leads to auto-context-like learning [18]. Specifically, the tentative estimation of each four-layer CNN (e.g., Step 1) can be further refined with the subsequent CNNs (e.g., Step 2). Moreover, the parameters of the concatenated CNNs can be optimized jointly with back-propagation. This differs from the *conventional* auto-context learning framework where the concatenated classifiers/regressors are often trained independently. In the final, we formulate the entire architecture of the concatenated CNNs into an end-to-end mapping, which estimates SPET from the combination of LPET and T1 images directly.

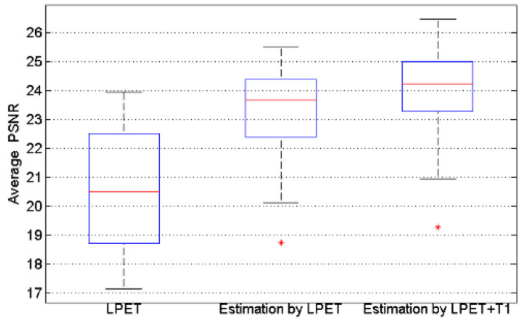
It is worth noting that direct training of the convolutional network with such a large depth may easily fall into local minima. Inspired by previous studies on training neural networks with deep supervision [44,45], the weighted auxiliary loss is also adopted in the network to further strengthen the training process. In particular, the auxiliary loss is computed from the end of each step. We



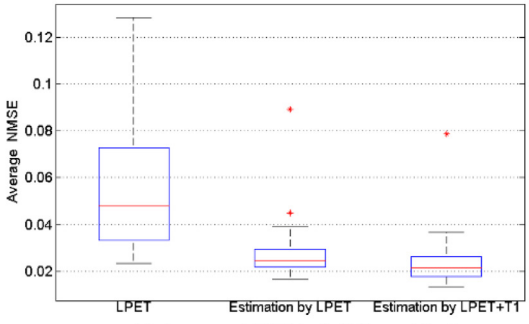
(a) PSNR of each subject using leave-one-out strategy



(b) NMSE of each subject using leave-one-out strategy



(c) Average PSNR of all the subjects



(d) Average NMSE of all the subjects

Fig. 4. The performances of using different input image settings, measured by PSNR and NMSE. (a) and (b) give the PSNR and NMSE scores of each subject in the leave-one-out validation. (c) and (d) give the average PSNR and NMSE scores of all the subjects. 'LPET' indicates the PSNR/NMSE between the original LPET and the ground-truth SPET. 'Estimation by LPET' represents the scores of the results estimated using only LPET as the input. 'Estimation by LPET+T1' represents the scores of the results estimated using both LPET and T1 images.

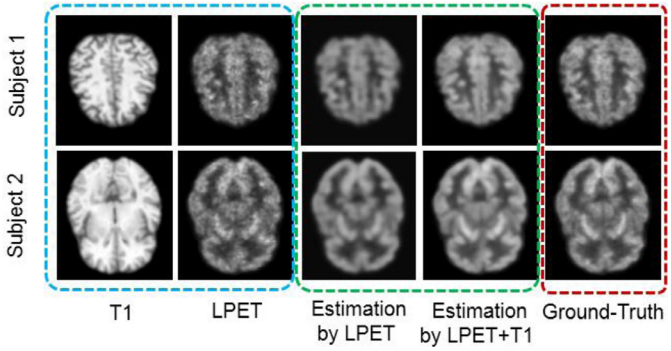


Fig. 5. Visual examples of using multi-modality inputs for SPET estimation. The two rows represent two different subjects. Original inputs of the LPET and the T1 images are in the blue dashed box. The estimated SPET images using different input settings are in the green dashed box. The ground-truth SPET images are in the red dashed box. 'T1' represents the input T1 image. 'LPET' is for the input LPET image. 'Estimation by LPET' represents the estimated SPET image by using only LPET as the input. 'Estimation by LPET+T1' represents the estimated SPET by using both LPET and T1 images as the inputs. 'Ground-Truth' is for the SPET image acquired in our dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

derive from (2) the loss after Step i and denote it as L_i. The total loss L_{total} for the entire deep auto-context CNN architecture is:

$$L_{total}(\theta_{total}) = \beta(L_1 + L_2) + L_3 + \varphi(\theta_{total}) \quad (3)$$

where $\varphi(\theta_{total})$ is the L_2-norm regularization term upon the estimated network parameters. In our experiments, β is adopted to balance the auxiliary losses among individual steps. Note that the term X_i in Eq. (2) varies for different steps when computing the auxiliary loss. We compute the loss for each step as the mean squared error between the estimated SPET and the ground-truth SPET. To this end, we need to crop the ground-truth SPET, such that it has the same size as the estimated SPET patch in each step. For example, the ground-truth SPET is cropped to 19×19 in Step 1, then 11×11 in Step 2, and 3×3 in Step 3.

4. Experimental results

We first introduce the dataset used in the experiments and discuss the parameter settings (Sections 4.1–4.2). After that, we investigate the impact of using the structural information (i.e., T1 images) for the estimation of the functional SPET data (Section 4.3). Next, we explore how our proposed deep auto-context CNNs gradually refine the SPET estimation by concatenating multiple CNNs (Section 4.4). Finally, we compare the proposed method with state-of-the-art method to demonstrate its effectiveness (Section 4.5).

4.1. Dataset

Our dataset contains 16 subjects, each of which has LPET, SPET, and T1 images. All data were acquired on a Siemens Biograph mRI PET-MR system. Their demographic information is summarized in Table 1. This study is approved by the University of North Carolina at Chapel Hill Institutional Review Board.

Before the PET scanning, each subject is administered an average of 203 MBq (from 191 MBq to 229 MBq) of ^{18}F-2-deoxyglucose (^{18}FDG). During PET scanning, an SPET image is obtained in a 12-min period within one hour of tracer injection, based on standard imaging protocols. The LPET scans are acquired in a 3-min short period, with standard-dose tracer injection, to simulate the acquisition at a reduced dose of radioactive trace. The simulation is equivalent to a quarter of the standard dose. The SPET and LPET images are acquired separately, so the noises in SPET and LPET are not correlated. All PET scans are reconstructed using standard

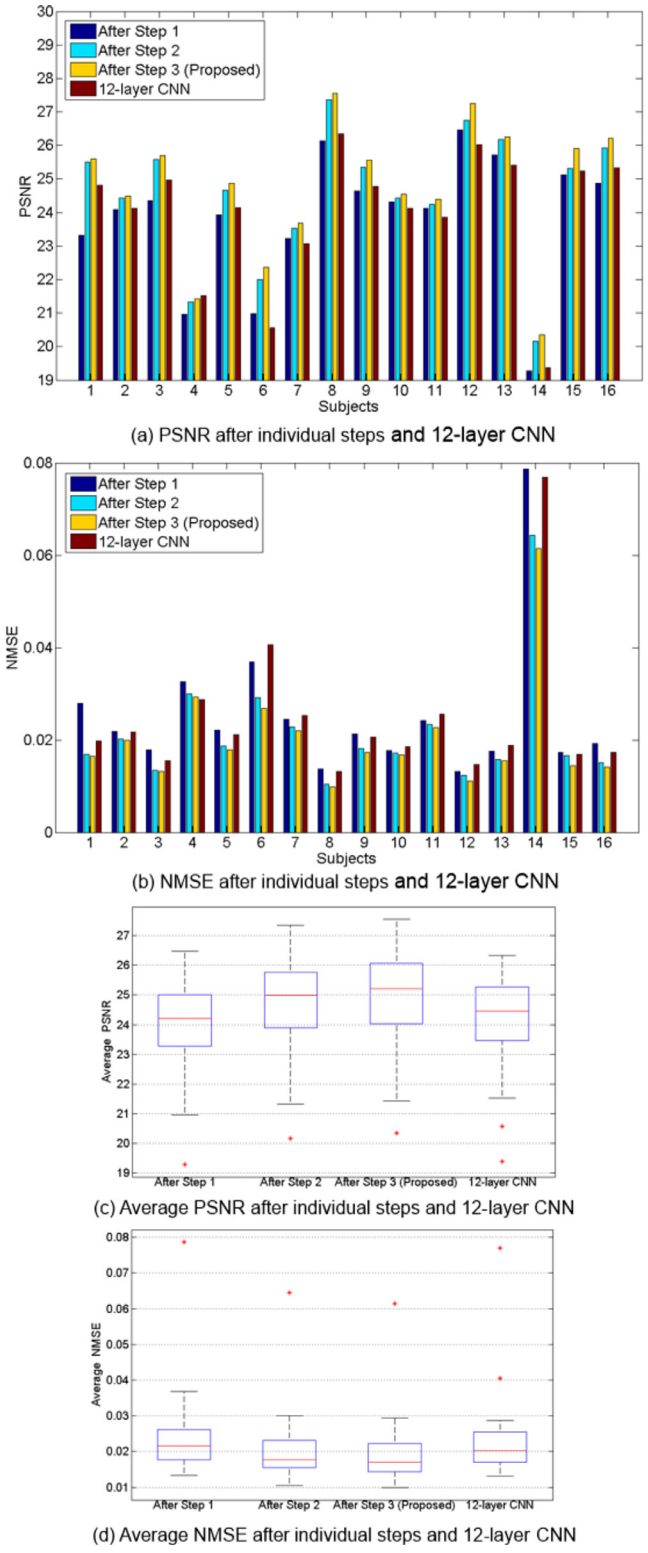


Fig. 6. The performances of concatenating multiple basic CNNs, in terms of PSNR and NMSE. (a) and (b) give the PSNR and NMSE scores of each subject by using the leave-one-out validation. (c) and (d) give the average PSNR and NMSE scores of all the subjects. Note that our method concatenates three basic CNNs, which is also indicated by 'After Step 3' in this figure. The results of the conventional 12-layer CNN are also shown in the figure.

Table 1
Demographic information of the subjects in the experiments.

Subject ID	Age	Gender	Weight (kg)
1	26	Female	50.3
2	30	Male	137.9
3	33	Female	103.0
4	25	Male	85.7
5	18	Male	59.9
6	19	Female	72.6
7	36	Female	102.1
8	28	Male	83.9
9	65	Female	68.0
10	86	Male	68.9
11	86	Female	74.8
12	66	Female	58.9
13	61	Male	83.9
14	81	Male	106.5
15	70	Female	61.2
16	72	Female	77.1

methods from the vendor. Attenuation correction, scatter and scanner uniformity are included using the vendor's standard procedure. Each PET image has a voxel size of $2.09 \times 2.09 \times 2.09\,\mathrm{mm}^3$. Meanwhile, the T1-weighted MPRAGE image is acquired with $1 \times 1 \times 1\,\mathrm{mm}^3$ resolution. For each subject, the T1 image is linearly aligned onto the corresponding PET image via affine transformation [46], followed by skull stripping [47] and intensity normalization. All images are further aligned to the space of a selected subject using FLIRT [48]. At last, we crop each image to the size of $120 \times 100 \times 100$ voxels to remove the redundant background.

4.2. Experimental configuration

The leave-one-out cross-validation strategy is employed for evaluation. That is, each time one subject is used for testing and the other images are for training. In this paper, CAFFE [49] is used to implement the CNN architecture. In the training phase, we use the same strategy with [42] that randomly selects 30,000 patches from each training image. There are totally 4.5×10^5 training patches in every leave-one-out case. The size of each patch is defined as 27×27. In Step 1, the filter sizes of the four convolution layers are set to 3×3, and the size of the output patch after Step 1 is thus 19×19. The numbers of the filters of the initial three convolution layers are the same, $n_1 = n_2 = n_3 = 64$, while there is only one filter, $n_4 = 1$, in the last layer of Step 1. Step 2 and Step 3 share similar parameters with Step 1, though their feature map sizes vary as in Fig. 3. The learning rates are 1×10^{-4} for Step 1 and Step 2, and 1×10^{-5} for Step 3. A smaller learning rate in the last four-layer CNN (i.e., Step 3) is helpful to the convergence of the network in training [50]. We adopt "SGD" as the solver for the simultaneous optimization of all steps in back-propagation. Although we use a fixed patch size in training, the deep networks can be applied to images of arbitrary sizes during testing.

To evaluate the performance of the proposed method quantitatively, we use the normalized mean squared error (NMSE) in (4) and the peak signal-to-noise ratio (PSNR) in (5):

$$NMSE = \frac{\left\|X - \hat{X}\right\|_2^2}{X^2}, \tag{4}$$

$$PSNR = 10\ln\left(\frac{ND^2}{\left\|X - \hat{X}\right\|_2^2}\right), \tag{5}$$

where X is the ground-truth SPET image, $\hat{X}$ is the estimated SPET image, D is the intensity range of image X, and N represents the

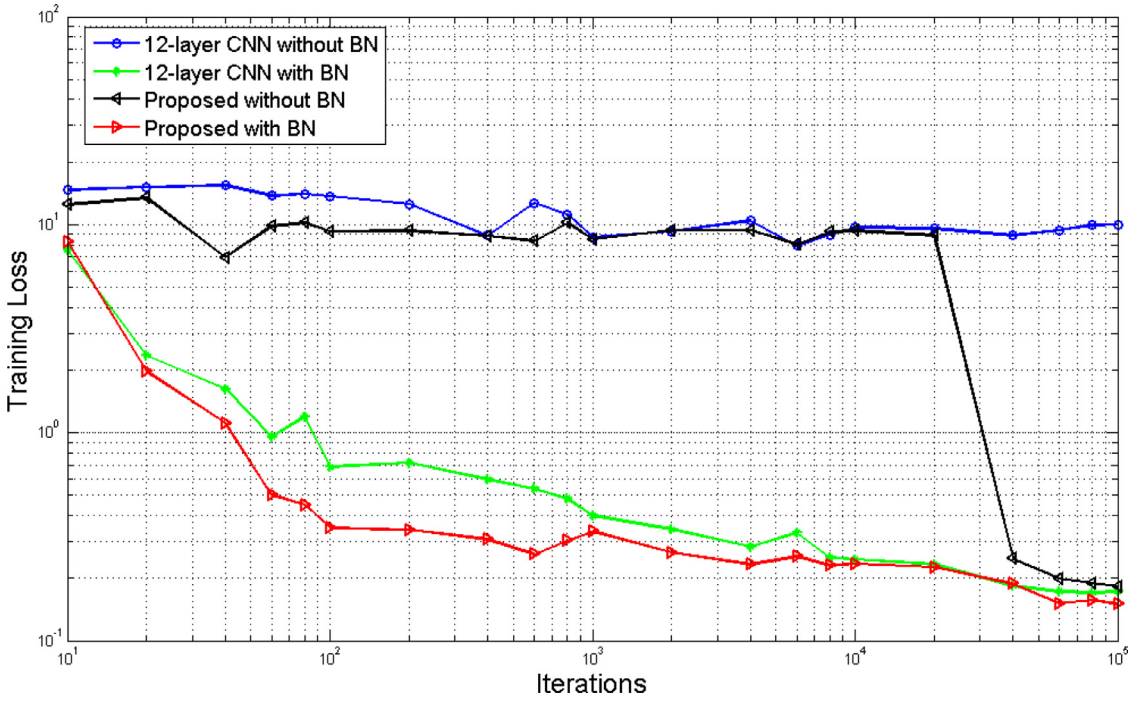


Fig. 7. Training loss with respect to the number of iterations for the 12-layer CNN and our proposed deep CNN architecture, with and without batch normalization.

total number of voxels in the image. Lower NMSE and higher PSNR indicate better quality of the estimated SPET.

4.3. Contribution of T1 in estimating SPET, using basic 4-layer CNN

To demonstrate the effectiveness of integrating multi-modality data for the estimation of SPET, we compare the performances achieved by using LPET input only and by using the combination of LPET and T1 images. When dealing with the single LPET input, we employ the same setting as in Fig. 2, but the input layer only considers the LPET image. The performances achieved by using different input settings are shown in Fig. 4. As we use the leave-one-out strategy, each subject is chosen for testing in turn. 'LPET' indicates the PSNR/NMSE scores by comparing the input LPET image with the ground-truth SPET image directly. 'Estimation by LPET' represents the estimation SPET results when using LPET as the input for our basic four-layer CNN only. 'Estimation by LPET+T1' represents the estimation SPET results when using the combined inputs of LPET and T1 as in Fig. 2.

We can observe that the results of 'Estimation by LPET' are worse than 'Estimation by LPET+T1'. In particular, the average PSNR scores of 'Estimation by LPET' and 'Estimation by LPET+T1' are 23.11 and 23.85, respectively. And the average NMSE scores of 'Estimation by LPET' and 'Estimation by LPET+T1' are 0.0299 and 0.0254, respectively. The PSNR scores and the NMSE scores are significantly different between 'Estimation by LPET' and 'Estimation by LPET+T1' (p-value < 0.01 in paired t-test). The results above imply that the structural information from T1 yields important cues for estimating the high-quality functional SPET, even though structural T1 differs from PET significantly regarding their appearances. We also provide two examples (corresponding to two rows) in Fig. 5 for visual observation, where our method yields more satisfactory estimation results regarding the ground-truth.

4.4. Concatenation of basic CNNs

Different from simply increasing the number of the layers in conventional CNN, we follow the auto-context strategy and concatenate three 4-layer basic CNNs in this work. Both LPET and T1 image patches, as well as the tentatively estimated SPET (if available), are used as the inputs to each of the three CNNs. In order to evaluate the effectiveness of concatenating multiple CNNs for auto-context-like estimation of SPET, we show the performances (measured by PSNR/NMSE) after individual steps of CNNs in Fig. 6. The average PSNR scores after Steps 1, 2, and 3 are 23.85, 24.55 and 24.76, respectively. The average NMSE scores after Steps 1, 2, and 3 are 0.0254, 0.0215 and 0.0205, respectively. The t-tests also yield p-values that are lower than 0.01 when comparing the resulted PSNRs and NMSEs between Step 2 and Step 1, and between Step 3 and Step 2. These results reveal that the estimation quality improves greatly after *refining the output of Step 1* in Step 2. The improvement of the overall PSNR/NMSE score becomes relatively limited when Step 3 is applied. To this end, we argue that the concatenation of multiple CNNs is effective to improve the quality of the estimated SPET. However, too many steps would increase the complexity of the entire network significantly, which could come with higher difficulty and more time cost for training. We have concatenated more CNNs but this fails to yield better performance. In general, we choose to concatenate three four-layer CNNs, considering both the performance and the computational efficiency.

In order to further reveal the power of our proposed method, here we compare our deep auto-context architecture (12 layers in total) with the 12-layer conventional CNN model. The results are also shown in Fig. 6. We can see that our model outperforms the 12-layer CNN. The average PSNR scores of our proposed method and the 12-layer CNN are 24.76 and 23.98, respectively. The average NMSE scores of our proposed method and the 12-layer CNN are 0.0206 and 0.0247, respectively. The differences between our method and the 12-layer CNN are statistically significant. These results show that, by concatenating multiple CNNs and forcefully directing information flows, the auto-context-like network is more effective than simply increasing the number of layers in the *conventional* CNN.

We concatenate multiple CNNs and build a deep structure, the training of which may become challenging. Therefore, we adopt

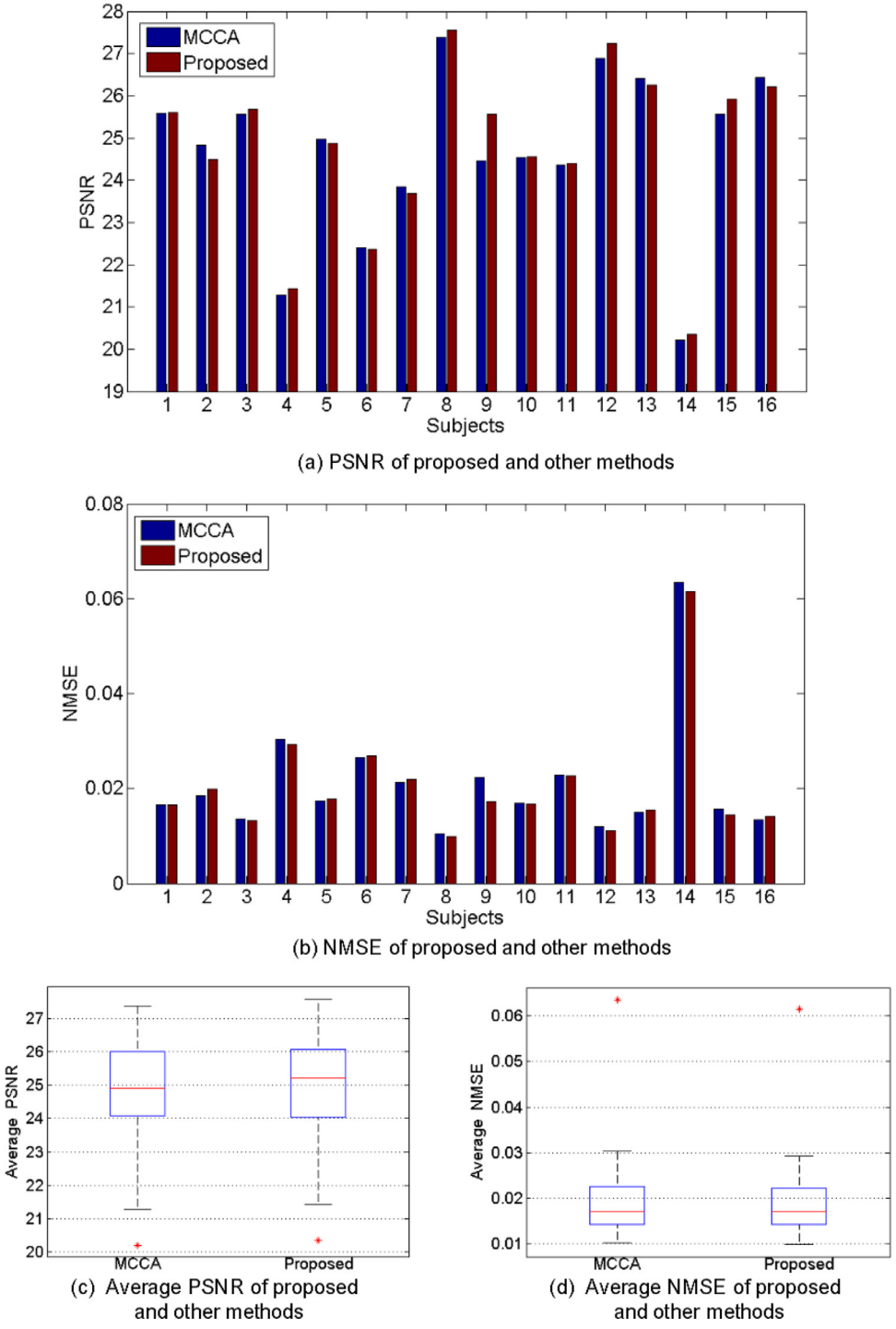


Fig. 8. Comparisons of our proposed deep auto-context CNNs with the MCCA method. (a) and (b) show the evaluation results of PNSR and NMSE scores for all 16 subjects in the leave-one-out testing. (c) and (d) give the average results of PSNR and NMSE of all the subjects.

the batch normalization strategy in modeling the network. In Fig. 7, we plot the changes of the training losses with respect to the number of iterations during training. The comparisons are conducted between the proposed method and the conventional 12-layer CNN, with and without batch normalization. Clearly, the strategy of batch normalization greatly contributes to the convergence of training. For example, without batch normalization, the conventional 12-layer CNN can hardly be trained. Meanwhile, we note that, with directed data flow in our concatenated CNNs, the training process can converge faster than the conventional CNN (i.e., by comparing the red and the green curves). The observation confirms that our method can effectively model the estimation of SPET from LPET and T1.

4.5. Comparison with sparse-learning-based MCCA method

We also compare our method with state-of-the-art MCCA method [20], which has achieved the best performance in the literature. The MCCA method, which belongs to the category of patch-based sparse learning, adopts the data-driven scheme and

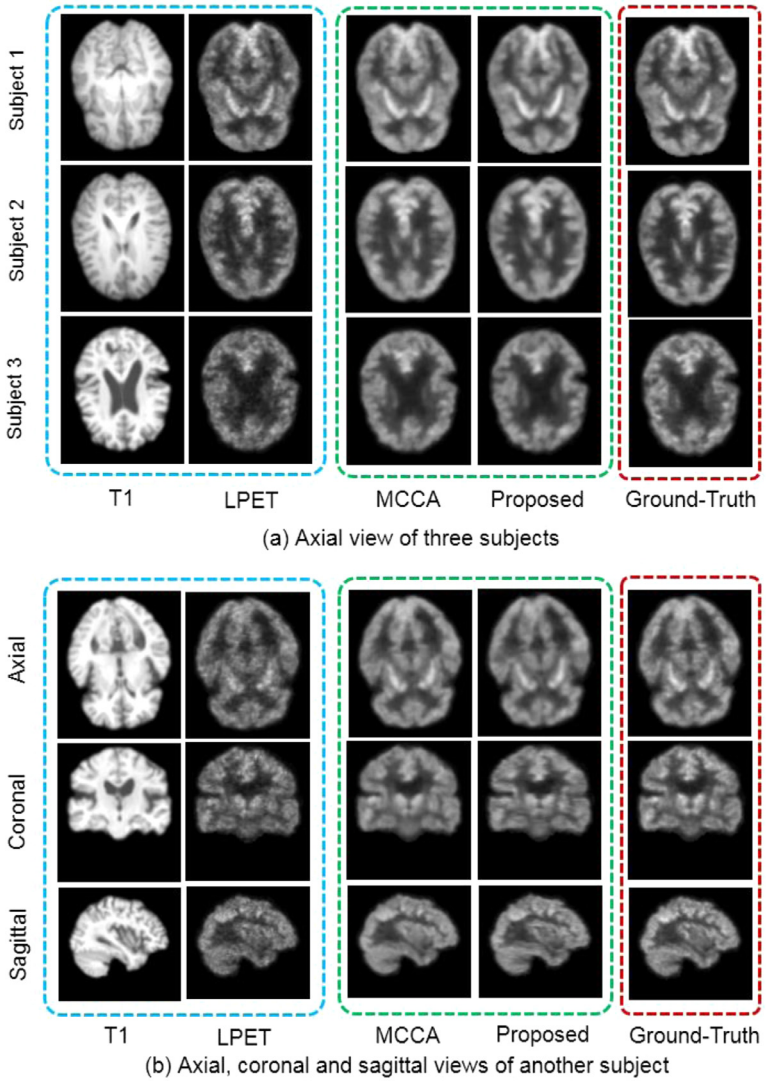


Fig. 9. Visual comparisons of our method and the MCCA method. Each row of (a) shows a subject from the axial view. (b) shows another subject from the axial, coronal and sagittal views, respectively. 'T1' represents the input T1 image, and 'LPET' represents the input LPET image. 'MCCA' represents the SPET image estimated by the MCCA method. The last column represents the ground-truth SPET images.

Table 2
Training time and testing time of two different methods.

Method	Training time	Testing time
MCCA	2.9 h	1008 s
Proposed	4.2 h	2.03 s

can iteratively refine the estimation results of the SPET images. All the PSNR/NMSE results of our method and MCCA are shown in Fig. 8. Our method performs comparatively, which yields the average PSNR of 24.76 and NMSE of 0.0206, compared to 24.67 and 0.021 by MCCA. Visual results are also provided in Fig. 9(a), where we can observe competitive results between the two methods.

Importantly, our method behaves significantly better in terms of its processing time especially in testing. Table 2 compares the time costs of our method and MCCA for both training and testing. Although our method spends more time on training, the testing procedure is much faster. Concretely, it only takes 2.03 s to test a subject by our method, while 1008 s by MCCA. The main reason is that MCCA optimizes sparse coding problems in testing, whereas our method is a completely feed-forward convolution operation without any pre-/post-processing. All the experiments are carried out on an ordinary computer with Intel Core i7 4.00 GHz processor, 16 GB RAM, and an NVIDIA Geforce GTX Titan X GPU.

Though our method carries out the computation from the axial plane slice by slice, the estimated results are still satisfactory in 3D view. In particular, after we complete the estimation upon all the slices, we stack them back to get the 3D image volume. A subject is shown in Fig. 9(b), where the axial, sagittal and coronal views are all available. We conclude that our estimation still appears to be isotropic, even though the CNN-based learning happens on the axial plane.

5. Conclusion

In this paper, we propose a novel deep auto-context CNN architecture for SPET image estimation using multi-modality data, including both LPET and T1 images. Different from previous sparse-learning-based techniques that contain time-consuming steps such as patch representation, non-linear mapping and reconstruction, our proposed method uses a deep neural network to map the

inputs to the output directly, without any pre/post-processing beyond the optimization in the training stage. When testing a subject, our method performs a single feed-forward to get the estimation result. In this way, our method can conduct the estimation of SPET very fast. Experimental results on a real human brain image dataset demonstrate that, compared to state-of-the-art method, our method has achieved competitive estimation quality, but it is up to 500 $\times$ faster.

We have also shown that our auto-context strategy is capable of building a very deep CNN architecture to further promote the estimate quality. Meanwhile, the entire network is still trained in an end-to-end way with back-propagation. Our model can be applied to other similar applications such as mapping one modality to the other. In the future, we will investigate the acceleration of the training process to make this method more efficient.

Acknowledgments

This work was supported in part by NIH grant (CA206100, EB006733, EB008374), National Natural Science Foundation of China (61473190, 61401271, 81471733), National Key Research and Development Program of China (2017YFC0107600), Science and Technology Commission of Shanghai Municipality (16511101100, 16410722400), Medical Engineering Cross Research Foundation of Shanghai Jiao Tong University (YG2014MS50).

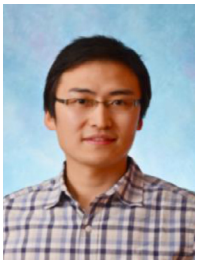
References

[1] N.E. Avril, W.A. Weber, Monitoring response to treatment in patients utilizing PET, Radiol. Clin. North Am. 43 (1) (2005) 189–204.
[2] M.E. Juweid, S. Stroobants, O.S. Hoekstra, et al., Use of positron emission tomography for response assessment of lymphoma: consensus of the Imaging Subcommittee of International Harmonization Project in Lymphoma, J. Clin. Oncol. 25 (5) (2007) 571–578.
[3] J. Vansteenkiste, S. Stroobants, The role of positron emission tomography with 18F-fluoro-2-deoxy-D-glucose in respiratory oncology, Eur. Respir. J. 17 (4) (2001) 802–820.
[4] L.F. de Geus-Oei, H.F. van der Heijden, F.H. Corstens, et al., Predictive and prognostic value of FDG-PET in nonsmall-cell lung cancer, Cancer 110 (8) (2007) 1654–1664.
[5] R. Boellaard, Standards for PET image acquisition and quantitative data analysis, J. Nucl. Med. 50 (2009) 11S–20S no. Suppl 1.
[6] W.A. Weber, Use of PET for monitoring cancer therapy and for predicting outcome, J. Nucl. Med. 46 (6) (2005) 983–995.
[7] C. Buchbender, T.A. Heusner, T.C. Lauenstein, et al., Oncologic PET/MRI, part 1: tumors of the brain, head and neck, chest, abdomen, and pelvis, J. Nucl. Med. 53 (6) (2012) 928–938.
[8] T.L. Slovis, The ALARA concept in pediatric CT: myth or reality? Radiology 223 (1) (2002) 5–6.
[9] J.M. Mejia, H.D.J. Ochoa Dominguez, O.O. Vergara Villegas, et al., Noise reduction in small-animal PET images using a multiresolution transform, IEEE Trans. Med. Imaging 33 (10) (2014) 2010–2019.
[10] A. Le Pogam, H. Hanzouli, M. Hatt, et al., Denoising of PET images by combining wavelets and curvelets for improved preservation of resolution and quantitation, Med. Image Anal. 17 (8) (2013) 877–891.
[11] U. Bagci, D.J. Mollura, Denoising PET images using singular value thresholding and Stein's unbiased risk estimate, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013, Springer, 2013, pp. 115–122.
[12] A. Gervaise, B. Osemont, S. Lecocq, et al., CT image quality improvement using adaptive iterative dose reduction with wide-volume acquisition on 320-detector CT, Eur. Radiol. 22 (2) (2012) 295–301.
[13] W. Zhang, R. Li, H. Deng, et al., Deep convolutional neural networks for multi--modality isointense infant brain image segmentation, NeuroImage 108 (2015) 214–224.
[14] Y. Wang, P. Zhang, L. An, et al., Predicting standard-dose PET image from low–dose PET and multimodal MR images using mapping-based sparse representation, Phys. Med. Biol. 61 (2) (2016) 791.
[15] F.E. Turkheimer, N. Boussion, A.N. Anderson, et al., PET image denoising using a synergistic multiresolution analysis of structural (MRI/CT) and functional datasets, J. Nucl. Med. 49 (4) (2008) 657–666.
[16] V.-G. Nguyen, S.-J. Lee, Incorporating anatomical side information into PET reconstruction using nonlocal regularization, IEEE Trans. Image Process. 22 (10) (2013) 3961–3973.
[17] T. Huynh, Y. Gao, J. Kang, et al., Estimating CT image from MRI data using structured random forest and auto-context model, IEEE Trans. Med. Imaging 35 (1) (2016) 174–183.
[18] Z. Tu, X. Bai, Auto-context and its application to high-level vision tasks and 3D brain image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 32 (10) (2010) 1744–1757.
[19] Y. Wang, G. Ma, L. An, et al., Semi-Supervised Tripled Dictionary Learning For Standard-Dose PET Image Prediction Using Low-Dose PET and Multimodal MRI, 2016.
[20] L. An, P. Zhang, E. Adeli, et al., Multi-level canonical correlation analysis for standard-dose PET image estimation, IEEE Trans. Image Process. 25 (7) (2016) 3303–3315.
[21] Y. LeCun, B. Boser, J.S. Denker, et al., Backpropagation applied to handwritten zip code recognition, Neural Comput. 1 (4) (1989) 541–551.
[22] K. Simonyan, and A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556, 2014.
[23] A. Krizhevsky, I. Sutskever, and G.E. Hinton, Imagenet Classification with Deep Convolutional Neural Networks. pp. 1097–1105.
[24] Y. Sun, X. Wang, and X. Tang, Deep Convolutional Network Cascade for Facial Point Detection. pp. 3476–3483.
[25] Y. Taigman, M. Yang, M.A. Ranzato, et al., Deepface: Closing the Gap to Human-Level Performance in Face Verification. pp. 1701–1708.
[26] Y. Sun, X. Wang, and X. Tang, Deep Learning Face Representation from Predicting 10,000 Classes. pp. 1891–1898.
[27] J. Long, E. Shelhamer, and T. Darrell, Fully Convolutional Networks for Semantic Segmentation. pp. 3431–3440.
[28] L.-C. Chen, G. Papandreou, I. Kokkinos, et al., Semantic image segmentation with deep convolutional nets and fully connected crfs, arXiv preprint arXiv:1412.7062, 2014.
[29] S. Hong, T. You, S. Kwak, et al., Online tracking by learning discriminative saliency map with convolutional neural network, arXiv preprint arXiv:1502.06796, 2015.
[30] H. Li, Y. Li, and F. Porikli, DeepTrack: Learning Discriminative Feature Representations by Convolutional Neural Networks for Visual Tracking. p. 3.
[31] C. Ma, J.-B. Huang, X. Yang, et al., Hierarchical Convolutional Features for Visual Tracking. pp. 3074–3082.
[32] A.A. Cruz-Roa, J.E.A. Ovalle, A. Madabhushi, et al., A Deep Learning Architecture For Image Representation, Visual Interpretability and Automated Basal-Cell Carcinoma Cancer Detection. pp. 403–410.
[33] J. Xu, L. Xiang, Q. Liu, et al., Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images, IEEE Trans. Med. Imaging 35 (1) (2016) 119–130.
[34] Q. Zhu, B. Du, B. Turkbey, et al., Deeply-supervised CNN for prostate segmentation, arXiv preprint arXiv:1703.07523, 2017.
[35] S. Liao, Y. Gao, A. Oto, et al., Representation Learning: a Unified Deep Learning Framework for Automatic Prostate MR Segmentation. pp. 254–261.
[36] V. Nair, and G.E. Hinton, Rectified Linear Units Improve Restricted Boltzmann Machines. pp. 807–814.
[37] N. Srivastava, G. Hinton, A. Krizhevsky, et al., Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.
[38] J. Deng, W. Dong, R. Socher, et al., Imagenet: A large-scale hierarchical image database. pp. 248–255.
[39] S. Ioffe, and C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv:1502.03167, 2015.
[40] R. Li, W. Zhang, H.-I. Suk, et al., Deep learning based imaging data completion for improved brain disease diagnosis, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014, Springer, 2014, pp. 305–312.
[41] C. Dong, C.C. Loy, K. He, et al., Image super-resolution using deep convolutional networks, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2) (2016) 295–307.
[42] C. Dong, C.C. Loy, K. He, et al., Image super-resolution using deep convolutional networks, 2015.
[43] Y. LeCun, L. Bottou, Y. Bengio, et al., Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.
[44] C.-Y. Lee, S. Xie, P. Gallagher, et al., Deeply-supervised nets, arXiv preprint arXiv:1409.5185, 2014.
[45] H. Chen, X.J. Qi, J.Z. Cheng, et al., Deep Contextual Networks for Neuronal Structure Segmentation.
[46] S.M. Smith, M. Jenkinson, M.W. Woolrich, et al., Advances in functional and structural MR image analysis and implementation as FSL, Neuroimage 23 (2004) S208–S219.
[47] F. Shi, L. Wang, Y. Dai, et al., LABEL: pediatric brain extraction using learning-based meta-algorithm, Neuroimage 62 (3) (2012) 1975–1986.
[48] B. Fischer, J. Modersitzki, FLIRT: a flexible image registration toolbox, in: Biomedical Image Registration, Springer, 2003, pp. 261–270.
[49] Y. Jia, E. Shelhamer, J. Donahue, et al., Caffe: Convolutional Architecture for Fast Feature Embedding. pp. 675–678.
[50] V. Jain, and S. Seung, Natural Image Denoising with Convolutional Networks. pp. 769–776.
[51] L. Zhang, Q. Wang, Y. Gao, G. Wu, D. Shen, Automatic labeling of MR brain images by hierarchical learning of atlas forests, Medical physics (2016).
[52] J. Zhang, L. Zhang, L. Xiang, Y. Shao, G. Wu, X. Zhou, D. Shen, Q. Wang, Brain atlas fusion from high-thickness diagnostic magnetic resonance images by learning-based super-resolution, Pattern Recognition (2017).
[53] L. Zhang, Q. Wang, Y. Gao, H. Li, G. Wu, D. Shen, Concatenated spatially-localized random forests for hippocampus labeling in adult and infant MR brain images, Neurocomputing (2017).

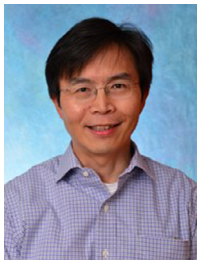
Lei Xiang is currently pursuing his Ph.D. degree at the School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China. His research interests include medical image processing, deep learning and pattern recognition.

Yu Qiao is a full professor with Shenzhen Institutes of Advanced Technology Chinese Academy of Sciences, Shenzhen, China. His research interests include computer vision, deep learning, pattern recognition, speech processing, robotics, etc.

Dong Nie is currently pursuing his Ph.D. degree at Department of Computer Science, University of North Carolina at Chapel Hill, America. His current research focuses on medical image analysis with machine learning.



Le An is currently an associate professor at School of Automation, Huazhong University of Science and Technology. His research is primarily focused on pattern recognition, computer vision, machine learning, and image processing, with applications in surveillance, biometrics, affective computing, and medical image analysis.



Weili Lin is currently the Director of the Biomedical Research Imaging Center (BRIC), The University of North Carolina, Chapel Hill. His research interests include PET and MR imaging, and their applications to cerebral ischemia and human brain development.

Qian Wang is a Research Scientist of the Med-X Research Institute, School of Biomedical Engineering, Shanghai Jiao Tong University. He received his Ph.D degree in Computer Science from the University of North Carolina at Chapel Hill in 2013. His researches focus on medical image analysis, computer vision, machine learning, artificial intelligence, and translational medical studies.

Dinggang Shen is a Professor of Radiology, Biomedical Research Imaging Center (BRIC), Computer Science, and Biomedical Engineering in University of North Carolina at Chapel Hill. Dr. Shen has published 700 articles in journals and proceedings of international conferences. Dr. Shen's research interests include: medical image analysis, computer vision, and pattern recognition.