Deep Residual Inception Encoder–Decoder Network for Medical Imaging Synthesis

Fei Gao, Teresa Wu, Xianghua Chu, Hyunsoo Yoon, Yanzhe Xu, and Bhavika Patel

***Abstract*—Image synthesis is a novel solution in precision medicine for scenarios where important medical imaging is not otherwise available. The convolutional neural network (CNN) is an ideal model for this task because of its powerful learning capabilities through the large number of layers and trainable parameters. In this research, we propose a new architecture of residual inception encoder–decoder neural network (RIED-Net) to learn the nonlinear mapping between the input images and targeting output images. To evaluate the validity of the proposed approach, it is compared with two models from the literature: synthetic CT deep convolutional neural network (sCT-DCNN) and shallow CNN, using both an institutional mammogram dataset from Mayo Clinic Arizona and a public neuroimaging dataset from the Alzheimer's Disease Neuroimaging Initiative. Experimental results show that the proposed RIED-Net outperforms the two models on both datasets significantly in terms of structural similarity index, mean absolute percent error, and peak signal-to-noise ratio.**

***Index Terms*—Deep learning, image synthesis, inception, medical imaging and residual net.**

I. Introduction

DURING the last decade, individualized precision medicine has emerged as a novel paradigm for diagnosis and treatment in healthcare. One cornerstone for precision medicine is medical imaging. Tremendous efforts have been dedicated to medical imaging research which can be categorized in four generalized areas: imaging-based classification [1]–[7], object detection [8]–[10], segmentation [11]–[14] and imaging synthesis [15]–[18]. The emerging convolutional neural network (CNN) has been successfully introduced into all these areas with different focuses [1]. An example of imaging detection and classification working on the object of interest is tumor detection and classification. Specifically, classification is to categorize the tumor, for example, as benign vs. malignant, in which the entire image or the extracted region of interest (ROI) is fed into a CNN, with one or more probabilities or class labels as the outputs. As early as 1996, a 4-layer CNN was implemented to classify ROIs from mammograms as either biopsy-proven masses or normal tissues [2]. Since then, different CNNs have been introduced for various classification tasks including but not limited to breast lesions [3], [4], lung patterns [5], skin lesions [6] or pulmonary peri-fissural nodules [7]. The task of detection is to derive an envelope box to enclose the given object. In the area of detection, bounding boxes or patches centered on the candidate objects are identified and CNN-based detectors are trained to find boxes that truly contain desired objects. Applications have included detecting colonic polyps on CT images [8], cerebral microbleeds from MRI scans [9], and nuclei in histopathological images [10]. Of note both classification and detection patterns are interested in the objects instead of pixel-based information, thus, the requirement of pixel-level details can be relaxed.

There is another category of problems known as dense prediction. It requires pixel-level specifics and focuses of imaging segmentation and synthesis. In segmentation, a probability map that quantifies the likelihood of each pixel being within the imaging object (e.g., tumor) is generated. Successful implementations have been reported in brain tumor/structures segmentation [11]–[13], epithelial tissue segmentation in prostatectomy [14], and others. In [11], a four-layer CNN is designed to take T1, T2 Magnetic Resonance images (MRI) and Fractional Anisotropy (FA) image as inputs and outputs are the segmentation maps for three types of tissues, namely white matter, gray matter, and cerebrospinal fluid. To do so, a local response normalization layer is implemented between the convolutional layer and the final fully connected layer to enforce competitions between features at the same spatial location across different feature maps resulting much improved segmentations. In [12], a fully convolutional neural network (FCNN) collaborated with random fields in a unified framework is proposed to segment brain tumor regions in MRI images, while in [14], the same FCNN is introduced in the task of epithelial tissue segmentation. In [13], a two-pathway CNN architecture is proposed to harvest both local features (longer pathway) and global contextual features (shorter pathway) simultaneously and improve the brain tumor segmentation. As research on exploring CNN on segmentation

Manuscript received December 31, 2018; revised March 21, 2019 and April 14, 2019; accepted April 17, 2019. Date of publication April 22, 2019; date of current version January 6, 2020. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health GrantU01 AG024904) and DOD ADNI (Department of Defense Award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from many other sources. Detailed ADNI acknowledgement information is available in http://adni.loni.usc.edu/wpcontent/uploads/how_to_apply/ADNI_Manuscript_Citations.pdf. *(Corresponding author: Xianghua Chu.)*

F. Gao, T. Wu, H. Yoon, and Y. Xu are with the Industrial Engineering Program, School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ 85281 USA (e-mail: fgao16@asu.edu; teresa.wu@asu.edu; hyoon15@asu.edu; yanzhexu@asu.edu).

X. Chu is with the College of Management, Institute of Big Data Intelligent Management and Decision, Shenzhen University, Shenzhen 518060, China (e-mail: x.chu@szu.edu.cn).

B. Patel is with the Department of Radiology, Mayo Clinic in Arizona, Scottsdale, AZ 85259, USA (e-mail: patel.bhavika@mayo.edu).

Digital Object Identifier 10.1109/JBHI.2019.2912659

progresses, a notable new architecture with "U" shape (U-Net [19]) has emerged. The novel design of a contracting path to capture context and a symmetric expanding path to enable precise localization improves the segmentation performance significantly. Following the success, U-Net and its variants are studied in a number of medical imaging segmentation problems. For instance, it is implemented in joint craniomaxillofacial bone segmentation and landmark digitization [20]. A 3D U-Net is designed in volumetric imaging segmentation for Xenopus kidney [21]. V-Net [22], an extension of U-Nets with added shortcut connections between different layers, is developed to segment prostate within 3D volumetric images.

Imaging synthesis tackles a different dense prediction problem. It is to discover the pixel-wise nonlinear associations between the input images and the output images. Imaging synthesis has great potentials in medical applications, especially in scenarios where some imaging modalities may be of limited access or missing due to various reasons such as cost, radiation or utilization of intravenous contrast [15]. As a new field, to the best of our knowledge, the very first published literature may be from Li [16]. To test the innovative idea, a 4-layer shallow network is developed to map the Fludeoxyglucose Positron Emission Tomography (FDG-PET) images from MRI. Improved clinical diagnostic accuracy is observed after using the combination of MRI and synthetic FDG-PET for Alzheimer's disease. In [17], a 4-layer CNN is designed to reconstruct dual-energy subtraction soft-tissue chest images from multi-scale gradient imaging of the original chest radiograph image. Another interesting effort is related to breast cancer research. Full Field Digital Mammography (FFDM) is the mainstay in breast cancer screening program but is known to suffer from diagnostic accuracy. Contrast Enhanced Digital Mammography (CEDM) is utilized iodinated IV contrast plus mammography (provides a low energy imaging comparable to FFDM and recombined subtracted imaging by taking advantage of Kedge of iodine digitally acquired from high-energy images [18]). While promising, as a new modality, CEDM is not yet widely available in many medical centers worldwide. To tackle this accessibility issue, a Shallow-Deep CNN is proposed in [23] to render synthetic recombined images from FFDM thus significantly improving breast cancer detection compared with the methods using FFDM alone. In this research, a 4-layer CNN (Shallow-CNN) is implemented to map the low energy (FFDM) images to the recombined images [23]. The research reviewed above is taking the proof-of-the-concept approach exploring the applicability of 4-layer network in image synthesis. The aforementioned 4-layer network is shallow and simpler compared to the deep networks used in imaging classification, detection, and segmentation. Therefore, most research only handles the images by taking small patches from the ROIs. For example, in [16], [23], the ROIs are smaller than 400×600 pixels and the size of training patches is 15×15 pixels. We contend this approach may work well for smaller images or under the condition where an ROI is available. For the later cases, the involvement from domain experts (e.g., fellowship-trained and board-certified radiologists) is required. An ideal solution for synthetic imaging is a deep CNN being capable of handling the whole image. A shallow network with limited learning power may suffer, while a deep network may be the promising network to explore [24]–[26]. This is because a deep network has many more layers and trainable parameters, is better equipped to learn the complicated associations between input and output images at the whole image scale.

Given imaging segmentation and synthesis share the common interest of pixel level details [27], [28], the satisfying performance of U-Net in segmentation makes it a potential approach for synthetic imaging research. There is an initial attempt in this direction. In [27], a 27-layer sCT-DCNN borrowing the 'copy and crop' idea from U-Net is implemented to generate synthetic CT images from MRI images of same subjects. Significantly improved synthetic results are achieved compared to the traditional atlas-based method. It is worth mentioning that in [27] as well as other segmentation architectures, max pooling is extensively used to reduce feature maps' resolution by representing each grid (e.g., a group of 4 neighboring pixels) with a single value (maximum value) in its subsequent feature maps. This maximization operation may keep the pixel-level specifics to some extent. In the applications where the input images and output images are of similar resolutions, the performance of approach in [27] may not be guaranteed.

In this research, we propose a new deep CNN, named Residual Inception Encoder-Decoder Net (RIED-Net). Since image synthesis and image segmentation all address the pixel level prediction problem, it is a good starting point to adopt existing state-of-art segmentation network structure such as U-Net [19] in image synthesis problem directly. However, it may lose some pixel level information because of the max-pooling and un-pooling layers used. Noticing that the pixel level information is very important in image synthesis tasks, we implement convolution and deconvolution layers in replacing the max-pooling and un-pooling respectively. However, the additional layers may lead to issues of gradient vanishing or degradation, which has long been criticized from very deep networks [24], [29]. Residual short-cut [24] has been proposed as a solution to this problem; but the existing residual shortcut can be only implemented between layers of the same size and is not applicable in the U-Net architecture. To address this issue, we propose the residual inception block, in which one additional inception path with 1×1 convolution is adapted for feature map resizing. So the whole RIED-Net takes the advantage of U-Net design and is improved by reserving more pixel level information; it is also robust to overfitting and gradient vanishing problem because of the residual inception block introduced.

The remaining of this paper is organized as follows. Section II provides a detailed description of the background for this research. Details about the proposed architecture are demonstrated in Section III. Experiments and discussions are presented in Section IV. Finally, Section V concludes this paper.

II. Background

A. U-Net and Dense Prediction Problem

CNNs have been successfully implemented to tackle different machine learning and computer vision problems. Improved performance is achieved in imaging classification and object detection tasks [30]–[32]. Researchers further extend this success to imaging segmentation, a dense prediction problem with

U-Net [19] being a representative model. U-Net and its variants have been applied to various segmentation problems such as joint craniomaxillofacial bone segmentation and landmark digitization [20], volumetric imaging segmentation for Xenopus kidney [21] and segment prostate from 3D volumetric images [22]. Most recently, U-Net is introduced to imaging synthesis, an example is sCT-DCNN [27].

Within U-Net, max pooling is a common approach to reduce the spatial resolution and increase the receptive fields in the CNN models. During the max pooling operation, the dimension of input representation is reduced by replacing each $n \times n$ matrix (n is the pooling size) with one single value (e.g., maximum value) as in the output representation map. After several iterations of pooling operations, the high dimensional input image is represented by a set of feature maps of reduced spatial resolution. Max pooling may be desirable for imaging classification and detection problems where the outcome is an abstracted prediction of the interested object as a whole. However, dense prediction problem differs as it requires preserving the pixel-level details [26]. As a result, max pooling used in a dense prediction problem may face the challenges of losing pixel information. Recognizing this problem, fully convolutional networks (FCNs) [33] are proposed to enlarge the feature maps through bilinear interpolation, and an unpooling layer is introduced in [34]. Specifically, when doing the max pooling operation within a grid, the locations of pixels with maximum intensity are recorded. In the corresponding unpooling layer, an output feature map is enlarged from the input map, the recorded position within output feature map is filled with corresponding value from input map, and the rest positions are placed with zeros (zero padding). As pointed out by [21], [26], unpooling suffers from the loss of information due to the excessive use of dimension reduction and zero paddings. Another potential issue with the max pooling and unpooling approach is, the max pooling operation keeps the location of the pixels with maximum contrast compared with its neighbors and then positions the pixel back to the same location in the corresponding unpooling operation. The underlying assumption is that the pixels from the input image with high contrast remain at the same positions throughout different layered feature maps and the output image. This may not be true in image synthesis where the input image and output image are from two different modalities, i.e., same region in location from two images may show different appearances [35]. One possible solution is incorporating the use of convolutional and deconvolutional layers with the learnable filters to better record the compression information during the encoding process and decompression information during the decoding process. This will be reviewed in the next subsection.

B. Convolutional and Deconvolutional Layers

The convolutional layer is the core building block of a CNN. A set of learnable filters are included in the convolutional layer to compute the convolutional values as the filters slide through all the pixels. Often, the filters slide a single pixel per step (stride = 1) to keep spatial resolution of input and output feature the same [30], [36]. By setting different strides, the filter can jump several pixels and obtain an output feature map of reduced spatial resolution, such design is implemented in the networks proposed in [18], [37]. In parallel, the deconvolution layer associates one single input with multiple outputs and is used as the reserve operation of convolution layer to enlarge and densify the outputs [34].

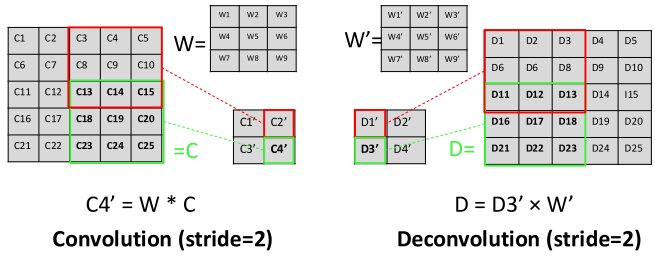

Fig. 1. Illustration of convolution and deconvolution operations ($*$: convolution operation; $\times$: multiplication of a scalar (e.g., D3′) and a matrix (e.g., W′)).

For illustration purpose, an example of convolution and deconvolution is shown in Fig. 1. Using convolution operation, the value of each pixel (e.g., C4′) in the output map is equal to the convolution of its corresponding area (C) in the input map and a learnable filter (W). As a result, the value of each pixel in the output map is a weighted summation of all corresponding input pixels. In deconvolution operation, the values of an output region equal to the pairwise multiplication of its corresponding pixel (D3′) in input map with the filter (W′). By learning the optimal filters (W and W′) while training the network model, the pixel-information is well preserved in encoding and decoding process. But, as the network is getting deeper with added convolution and deconvolution layers, gradient vanishing or degradation issue may emerge. To address these issues, we propose a residual inception structure, it will be discussed in the following section.

C. Residual Short-Cut and Inception Block

Deep networks integrate multiple level features and classifiers in an end-to-end multilayer fashion, and the levels of features are enriched by the number of stacked layers (depth) [36]. The stacked convolutional layers tend to underperform its shallower counterparts due to the problems such as gradient vanishing/exploding, as millions of parameters in deep networks are updated based on a single value of gradient. The early layers tend to be less sensitive to that gradient tends to get smaller when it moves backward through the layers [24]. This problem is even worse for the dense prediction problem, as the gradient is calculated by averaging prediction errors on all pixels, thus making it less contrasting and even less sensitive to early layers.

To address the gradient vanishing issue in imaging classification, residual shortcut connection is first introduced in [24]. In residual shortcut block, formally, let H(x) denotes the desired non-linear mapping between the input and output of the residual block, instead of directly estimating H, the residual map $F(x) = H(x) - x$ is estimated by the learnable filters within the

2 residual blocks, and the original mapping can be recast into $F(x) + x$. Different experiments have been conducted and justified so that the residual mapping is much easier to optimize, resulting a more accurate results. Residual shortcut design also achieves extended success in dense predictions such as segmentation [22], [37]. However, in these models, the residual shortcut is implemented between arbitrary layers and is used as a feature transducer. These models are not designed following the original purpose of short-cut design, which enables training deep neural networks easier and more accurate by simply stacking the short cut blocks. Moreover, implementation of residual short cut blocks requires the size of input and output feature maps to be equivalent to conduct pixel-wised summation. This is not applicable if we still want to take advantage of novel design of U-Net, as the decoding layers are doubled with feature maps copied from corresponding encoding layers.

The concept of inception is first introduced in [31] and leads to several similar variants [38]–[40]. The power and novelty of inception block lies in its multiple inception paths; with these paths, feature maps of different scales and levels are derived and combined to approximate more complicated feature maps, which is otherwise achieved through larger filters and more layers. In this way, the training efficiency and accuracy are improved. Since multiple inception paths of different scales are needed and the outputs feature maps from different inception paths are concatenated together, additional parameters are needed within each inception block.

III. RIED-Net

In image synthesis problem, assuming input image $\mathrm{I} \in \mathrm{R}^{\mathrm{m}\times\mathrm{n}}$ and $\mathrm{O} \in \mathrm{R}^{\mathrm{m}\times\mathrm{n}}$ is the corresponding output image, the relationship between them can be formulated as

$$\mathrm{O} = S\,(\mathrm{I}) \tag{1}$$

where $S : R^{m\times n} \rightarrow R^{m\times n}$ denotes the complex non-linear mapping between the input and output images. The image synthesis problem is to make an estimation of function S:

$$\arg\min_{S} \|S\,(\mathrm{I}) - \mathrm{O}_1\| \tag{2}$$

In this research, we propose the Residual Inception Encoder-Decoder Neural Network (RIED-Net) to estimate the desired S, and we use ℓ_1-norm to measure the difference between input and output images. The overall architecture of RIED-Net is shown in Fig. 2. It consists of an encoding path (left side) to compress essential information in the extracted patches from low-level to high-level and a decoding path (right side) to reconstruct the final outputs from the compressed feature maps. The 'copy and crop' idea and a symmetric expanding path are added to capture the context features from encoding path to decoding path. Convolutional and deconvolutional layers are introduced to replace the max-pooling and un-pooling layers as learnable filters so the pixel information can be traced in both the encoding procedure to reduce the feature maps' spatial resolution and the decoding procedure to increase the spatial resolution. In addition, the inception residual block is proposed to address issues raised from networks getting deeper and ensure a better accuracy.

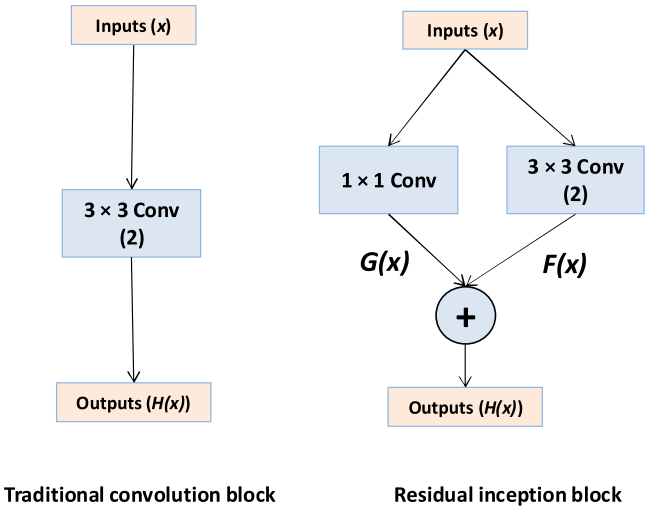

Fig. 2. Schemas for original convolution block and proposed residual inception (Note that in traditional convolution block, the input $\boldsymbol{x}$ and output $H(\boldsymbol{x})$ has different number of channels which makes the directly residual shortcut inapplicable).

The RIED has 9 residual inception blocks, with 5 blocks in the encoding path and the remaining 4 in the decoding path. The model takes gray scale images of size 128 × 128 × 1, and outputs the predicted mask of the same size. Within each block, there is a main path of 3 convolution layers and a residual inception path from the first layer to the last layer, on which a 1 × 1 convolutional layer is implemented. The channel number can be found on the top of each layer in Fig.2. Between each block in the encoding path, a convolutional layer with stride equal to 2 is implemented to reduce the resolution by half; between each block in the decoding path a deconvolution layer with stride equal to 2 is implemented to double the resolution. There are a total of 12,247,233 trainable parameters within the proposed model. Among the 12,247,233 parameters, 25% (3,052,960) are within the convolution and deconvolution layers for resolution change while reserving pixel information; 3% (348,192) are within the residual inception shortcut; the remaining 72% parameters are within the main encoding and decoding path for feature generation.

The residual inception blocks take a new architecture (Fig. 3). It consists of one traditional convolutional path with two 3 × 3 convolutional layers as sCT-DCNN or U-Net, and a unique residual inception short-cut path with a 1 × 1 convolutional layer. The 1 × 1 convolutional layer is implemented to increase (during encoding) or decrease (during decoding) the filter depth and project the input feature map into the same space as output to ensure the pixel-wise summation.

In the traditional 2-layer convolution block, given the input image/feature map x, we assume the desired mapping fitted by stacked nonlinear layers fitting is $H(\boldsymbol{x})$. After introducing a residual inception shortcut with one single convolution layer, $H(\boldsymbol{x})$ can be estimated as $F(\boldsymbol{x}) + G(\boldsymbol{x})$ in the proposed residual inception block. In this way, $H(\boldsymbol{x})$ is estimated simultaneously using features from 2 different levels, which will improve the

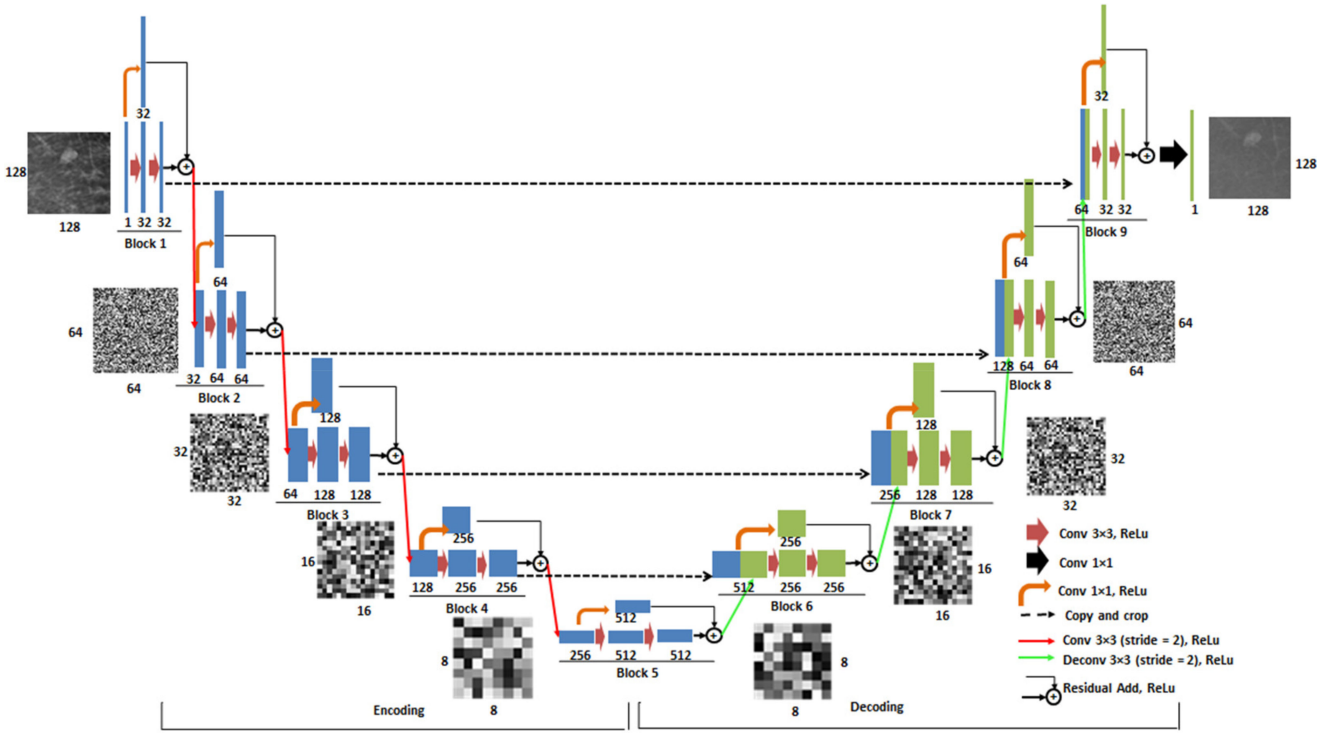

Fig. 3. Architecture of RIED-Net (Each brown arrow represents a 3 × 3 convolutional operation with a rectified linear unit (ReLu) as the activation function (Conv 3 × 3, ReLu). Each red arrow denotes a 3 × 3 convolutional operation (stride = 2, with a ReLu as the activation function), each orange arrow denotes a 1 × 1 convolutional operation (with a ReLu as the activation function) and each green arrow denotes a 3 × 3 deconvolutional operation (stride = 2, with a ReLu as the activation function). Each black dotted arrow denotes a copying operation. The final purple arrow denotes the final 1 × 1 convolutional operation that generates the output of synthetic image. The depth (number of channels) of the feature map from each convolutional layer is provided at the bottom of each box. Examples of feature maps from different levels are also displayed. There are 9 residual inception blocks (block 1 ∼ 9) in RIED-Net).

accuracy as more features are introduced [31]. Besides, $G(\boldsymbol{x})$ can be regarded as a projection/estimation of x [39], following the same logic in [24], [38], the residual mapping $F(\boldsymbol{x})$ and projection $G(\boldsymbol{x})$ are much easier to optimize, resulting in more accurate results than the original mapping $H(\boldsymbol{x})$.

Our intention to adopt the inception path is different from the original design. In this study, we do not expect the convolution layers along different reception paths to learn the complicated mapping and combine them together. Here we implement ONE single inception path with a projection layer (1 × 1 convolution layer) that changes the feature map's channel size to enable the pixel-wised summation required by residual learning. Second, the residual inception block requires much less parameters compared with inception block. In our residual inception block, only one cheapest 1 × 1 convolutional layer is added, and the outputs are combined through pixel-wised summation, which is free of parameters. While in inception block, many more parameters are needed because 1) multiple inception paths of different feature sizes are needed, and 2) feature maps from different inception paths are concatenated together for the following layers. Our proposed residual inception block addresses the problem that the input feature maps have different channels from the output feature maps. It is also simpler and easier to deploy than other state-of-art inception designs.

IV. Experiments

In this section, we conduct two experiments to validate the performance of RIED-Net using digital mammography dataset from Mayo Clinic and a public neuroimaging dataset from Alzheimer's Disease Neuroimaging Initiative (ADNI) [41]. Three commonly used metrics, mean absolute error (MAE), structural similarity index (SSIM) [42], and peak signal-to-noise ratio (PSNR) [43] from literature that quantify the similarity between the ground truth image and the synthetic image are employed to evaluate the performance of the synthesis model. The experiments are conducted on a Dell desktop with 32 GM RAM and 12 CPU cores. The models are trained with a single NVIDIA TITAN XP GPU with 12GB memory.

A. Experiment I: Case Study on Breast Cancer

Breast cancer is the leading type of cancer in women accounting for 25% of all cancer cases worldwide. Full field digital mammography (FFDM) is the only imaging modality proven to reduce mortality from breast cancer. However, using FFDM is not an optimal approach in breast cancer screening due to its relatively low detection sensitivity in many subgroups of women, who has high-risk and dense breast [44]. Using dynamic contrast-enhanced breast MRI may yield significantly higher cancer detection sensitivity, but its substantially higher

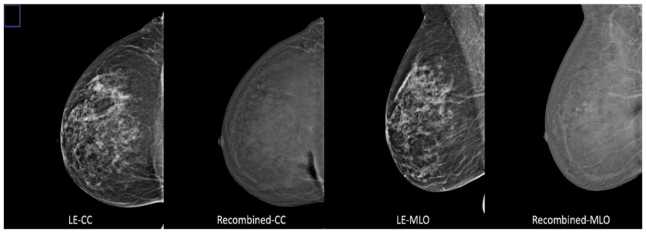

Fig. 4. Examples of images in CEDM dataset.

cost, lower accessibility, and longer imaging scanning time forbid breast MRI being used as a primary imaging modality in breast cancer screening and detection [45]. To combine the advantages of both FFDM and MRI, a new novel imaging modality namely, contrast-enhanced digital mammography (CEDM) emerges. CEDM includes low energy (LE) image, which is comparable to routine FFDM [46] and recombined contrast enhanced image similar to breast MRI. Several studies including prospective clinical trials conducted at Mayo Clinic have indicated that CEDM is a promising imaging modality that overcomes tissue overlapping ("masking") which can be seen in FFDM, provides tumor neovascularity related functional information similar to MRI, while maintaining the high image resolution of FFDM [47]–[50]. While promising, as a new modality, CEDM is not widely available in many medical centers. The motivation of this study is to explore the application of synthetic imaging to render "virtual" recombined images from the LE image in hoping to fulfill the clinical needs.

In this experiment, we evaluate the performance of RIED-Net in mapping the LE images to the recombined images. Since image synthesis is a relatively new field, two methods from literature are taken for comparison: Shallow CNN [23] and sCT-DCNN [27].

1) Dataset: Based on Institutional Review Board (IRB) approved study and data collection protocol, we reviewed 139 CEDM examinations performed using the Hologic Imaging system (Bedford, MA, USA) between August 2014 and December 2015. In CEDM dataset for each subject, there are both LE and recombined cranial-caudal (CC) and mediolateral-oblique (MLO) views of each breast. Examples for the images are shown in Fig. 4.

All images are in 2560(width) $\times$ 3328(height) with intensity ranges from 0 $\sim$ 4095. Among the dataset, 112 (80%) subjects were randomly selected as training dataset, the remaining 27 (20%) subjects were used as the test dataset. For each subject, CC view and MLO view images are treated as two separate training images, which results in a dataset of 224 (112 $\times$ 2) training images and 54 test images (27 $\times$ 2).

2) Image Processing and Training: It is a common approach to extract patches from images as the training samples to address the shortage of training dataset [23], [27]. However, the size of patches varies case by case. Larger patches require more memory for calculation, while smaller patches allow the network to see very little context. In the experiment, based on the preliminary experiments, we set the training patch size to be 128 $\times$ 128. After patch size is set, training samples are extracted from the images in the step size of 8 in each dimension, and patches outside the breast boundary are excluded. As a result, a dataset of 65,800 patches are obtained from the 112 training subjects. Among these 65,800 patches, 59,220 (90%) are used as training samples, and the remaining 6,580 (10%) are used as validation samples to tune the parameters.

The proposed RIED-Net estimates an end-to-end mapping from LE images to recombined images. Once we have decided the configuration of the network, the set of parameters $\Theta = \{\theta_i\}$ for RIED-Net should be estimated to build the mapping function S**.** The estimation can be achieved by minimizing the loss $F(D;\Theta)$ between the synthetic recombined image and the ground-truth recombined image. Given a set of paired patches $\mathrm{P} = \{(\mathrm{X}_1, \mathrm{Y}_1), (\mathrm{X}_2, \mathrm{Y}_2), \ldots, (\mathrm{X}_\mathrm{K}, \mathrm{Y}_\mathrm{K})\}$ where $\{\mathrm{X}_i\}$ and $\{\mathrm{Y}_i\}$ denote the LE and recombined image patches respectively, and *K* is the total number of training samples. The mean absolute error (MAE) is utilized as the loss function:

$$F\left(D;\Theta\right) = \frac{1}{N}\sum_{i=1}^{\mathrm{N}} \|X_i - S(Y_i)_1\| \tag{3}$$

The parameter setting is decided based on the best validation results. Specifically, the overall architecture is implemented with programming language Python, and libraries including Keras and TensorFlow. Adam [51] is used as the optimizer. Learning rate is set to be 0.002 with learning rate decay equal to 0.005. Training batch size is set to be 64 and training iteration is set to be 80. We use the default settings of Keras for all the other parameters. For the two competing models, the optimal parameters reported in the literature are used.

3) Comparison: The comparison of performance of the different models is conducted on the reserved test dataset of 54 images. For each image, we slide the 128 $\times$ 128 window from left to right, top to bottom (step size = 2) in LE image, to get the input patches. The input patches are fed into the trained model, from which we get the synthetic recombined image patches (128 $\times$ 128) as outputs. The output patches are placed at the same position as their corresponding input patches in the synthetic recombined images. For the positions with overlapping pixels, the values are replaced with mean value for all overlapping pixels. In this way, the synthetic recombined images are finally rendered. Our ultimate goal is to synthesize the whole image, so it is more desirable to evaluate the performance based on the predicted complete image and its corresponding ground truth image instead of simply comparing the individual patches. As a result, to quantify the performance of synthesis for our proposed model, a set of 54 synthetic recombined images are generated for each LE image in the test dataset with the trained model. Each individual synthetic recombined image is then compared with its corresponding ground truth image using MAE, SSIM, and PSNR. In terms of each evaluation metric, the mean value and standard deviation across the 54 pairs of synthetic-ground truth image are reported. Two state-of-art models (Shallow CNN

TABLE I
PERFORMANCES OF DIFFERENT MODELS ON THE CEDM TEST DATASET

	MAE	SSIM	PSNR
Shallow CNN	219.753(±21.563)	0.793(±0.023)	29.224(±1.462)
sCT-DCNN	11.502 (±2.187)	0.958(±0.013)	43.346(±1.462)
RIED-Net	**11.277 (±2.112)**	**0.962(±0.012)**	**43.450(±1.423)**

TABLE II
p-VALUES OF t-TESTS ON PAIRWISE COMPARISION: (A) MAE, (B) SSIM, (C) PSNR

(a)	Shallow CNN	sCT-DCNN	RIED-Net
Shallow CNN	-	<0.001	<0.001
sCT-DCNN	<0.001	-	0.003
RIED-Net	<0.001	0.003	-
(b)	Shallow CNN	sCT-DCNN	RIED-Net
Shallow CNN	-	<0.001	<0.001
sCT-DCNN	<0.001	-	0.015
RIED-Net	<0.001	0.015	-
(c)	Shallow CNN	sCT-DCNN	RIED-Net
Shallow CNN	-	<0.001	<0.001
sCT-DCNN	<0.001	-	0.004
RIED-Net	<0.001	0.004	-

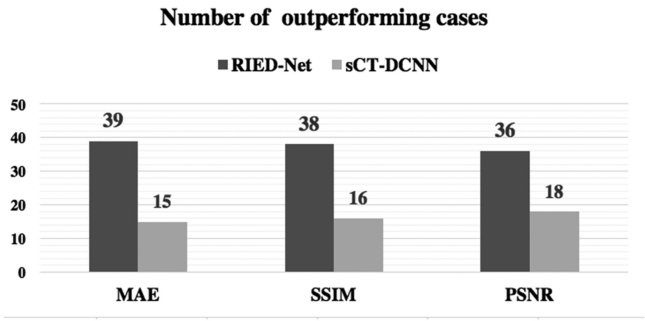

Fig. 5. Distribution of outperforming cases for MAE, SSIM and PSNR on CEDM test dataset.

[23] and sCT-DCNN [27]) are implemented for comparison (see Table I). Also, the paired t-tests are conducted (see Table II).

From the Table I, we have two conclusions. First, shallow CNN significantly underperforms both sCT-DCNN and our proposed RIED-Net on all three metrics. This confirms our argument that deep models with more trainable parameters may outperform shallow networks in the imaging synthesis problem. Comparing to sCT-DCNN, RIED-Net shows marginal performance advantages (11.277 vs. 11.502 in MAE, 0.962 vs. 0.958 in SSIM, 43.450 vs. 43.346 in PSNR). RIED-Net has a small standard deviation indicating its robust performance. These differences may occur by chance and results from few outlier samples. In order to gain more confidence about the performance difference, we conducted a paired t-test with results shown in Table II. According to the results, with all *p*-values are smaller than 0.02, we can determine the existence of significant performance differences.

To justify the marginal outperformance, we delve in details on a case by case basis. As seen in Fig. 5, among all the 54 images,

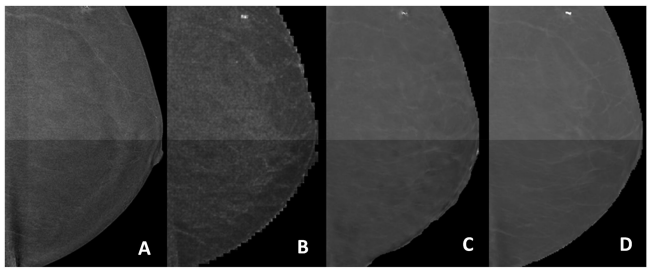

Fig. 6. Sample of one ground truth recombined image (A). Output synthetic re-combined images of Shallow-CNN (B). SCT-DCNN (C). The proposed RIED-Net (D).

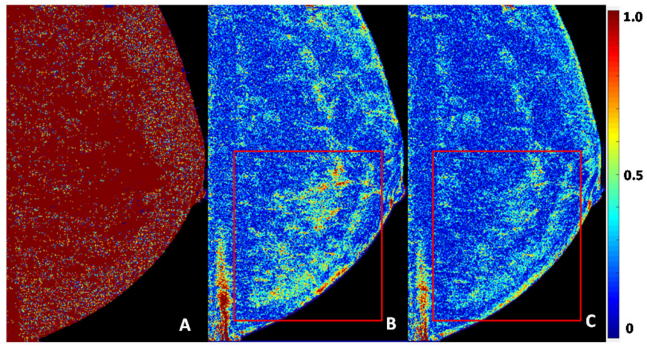

Fig. 7. Error maps of output images for Shallow-CNN (A). sCT-DCNN (B). The proposed RIED-Net (C).

our proposed RIED-Net has higher SSIMs (the higher the better) on 38 images (70.4%), higher PSNRs (the higher the better) on 36 images (66.7%), smaller MAE (the smaller the better) on 39 images (72.2%). In looking at all three metrics together, RIED-Net outperforms sCT-DCNN on 36 cases (>66.7%).

For illustration, we include one image from each model (see Fig. 6). Fig. 6A is ground truth recombined image. Fig. 6B, 6C, and 6D are predicted images of shallow CNN, sCT-DCNN, and RIED-Net respectively. The error maps for output images are shown in Fig. 7. Within the error map, the value of a pixel is the absolute value of difference between the intensities of two pixels at the same location in ground truth image and synthetic image. Each value is then divided by the same normalizer (normalizer value = 15). The values greater than 1 are assigned with 1s. The aim of this procedure is to normalize the range of difference map into between 0 and 1, while excluding the effects of outlier pixels.

First of all, as expected, limited by the learning capability, there is a very significant gap between the output of the 4-layer shallow CNN and ground truth images (high MAE values). We can focus on the comparison between sCT-DCNN and our proposed model. Comparing output images Fig. 6C and 6D, we can observe that Fig. 6C is coarser within the breast region, especially in the region close to the boundary, while in Fig. 6D, these regions are sharper and clearer. This is because, in these regions, the dense tissue is interlaced with other parts such vessels or fat, the differences among pixels from different parts are large. sCT-DCNN with max pooling loses the pixel information

TABLE III
NUMBER OF PARAMETERS WITHIN EACH METHOD

Method	Number of parameters
Shallow-CNN	664,641
sCT-DCNN	9,346,081
RIED-Net	12,247,233

and the unpooling layers fail to restore such information, as a result, these pixels cannot be differentiated well and tend to give similar predictions. The advantages of RIED-Net in this scenario are clearly shown. In looking at the error maps in Fig. 7(B and C), the red bounding box in Fig. 7B has larger high-error regions compared to Fig. 7C. This may be because in sCT-DCNN, during the prediction, if a single pixel is estimated with high error, it will first affect its 3 neighboring pixels after unpooling layers, and this effect tends to expand to more pixels after more unpooling layers. In RIED-Net, the succeeding pixels after deconvolutional layers depend not only on that specific preceding pixel, but also on the trainable parameters within the deconvolutional layers. In this way, even if a pixel is estimated with high error, this results in its following neighboring pixels to be relieved through the deconvolutional layer, thus the region of high-error in Fig. 7C tends to be small and in isolated regions. We conclude RIED-Net has promising potential for the imaging synthesis problem in breast cancer research on digital mammography (DM) modality.

The details of number of parameters required by each of the approaches are shown in Table III. There are a total of 12,247,233 trainable parameters within the proposed model. Among the 12,247,233 parameters, 25% (3,052,960) are within the convolution and deconvolution layers for resolution change while reserving pixel information; 3% (348,192) are within the residual inception shortcut; the remaining 72% (8,846,081) parameters are within the main encoding and decoding path for feature generation. Our proposed method adopts comparable number of parameters as sCT-DCNN in the main encoding and decoding path. Most extra parameters ($\sim$90%) are mainly used to reserve pixel information when changing the feature map resolution and a small portion ($\sim$10%) are used to in the inception residual shortcut for further performance improvement.

For the proposed model, the total training time (80 iterations) is 25,360 seconds. The average time to generate a synthetic image is 5.2 seconds. The time for training\generating a synthetic image for the sCT-DCNN and Shallow-CNN are 20,081/4.5 and 1,600/0.4 seconds respectively. Shallow-CNN has the fastest training and prediction time as it has the least parameters. The sCT-DCNN has less parameter than our proposed RIED-Net, its training process is faster (20.8%). However, once the model is trained, the gap of prediction time is narrowed to 13.4%.

Next, we will explore its applicability to an Alzheimer disease dataset across two imaging modalities: FDG-PET and MRI.

B. Experiment II: Case Study on Alzheimer's Disease

Alzheimer's disease (AD) is a progressive neurodegenerative disease which is the most frequent type among elderly dementia patients. In the U.S., approximately 5.5 million people over 65 are afflicted by AD (Alzheimer's Association, 2018). This drives a great amount of research investigating ways for the diagnosis and prognosis of AD. And the use of imaging has been highlighted by multiple expert consensus groups nationally and internationally, such as the working group convened by National Institute of Aging (NIA) and the Alzheimer's Association (AA) and the International Working Group [52]. It has been widely-recognized that imaging of different modalities, including but not limited to structural MRI, FDG-PET, and amyloid-PET, play important and often complementary roles. However, it is difficult for a single modality to serve all the purposes as each modality has unique strength and weakness. Combining different imaging modalities is vitally important to make accurate and early diagnosis and prognosis, a prerequisite to develop effective disease-modifying therapies. However, patients may not have all imaging modalities available due to various reasons. In this experiment, the proposed architecture is to learn the non-linear mapping between FDG-PET images and MRI images. It will be trained to render synthetic FDG-PET images given MRI images as input. Its performance is compared with the same two competing methods used in experiment I.

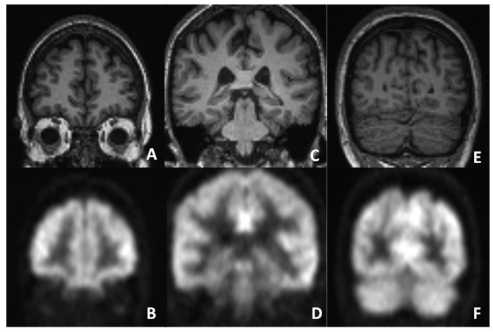

Fig. 8. Examples of MRI slices (A/C/E) in ADNI dataset and their corresponding FDG-PET images (B/D/F).

1) Dataset: The ADNI is launched aiming at finding the relationship between progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD) and biomarkers, MRI, PET or clinical and neuropsychological assessments. ADNI enrolls a large cohort (>800) of participants [53], PET, MRI images, as well as clinical information are available. In this experiment, 14 subjects are downloaded and used in the experiment. Detailed information for ADNI dataset is as following: MRI (before co-registration: 256 $\times$ 256 $\times$ 170, after co- registration: 79/79/91, intensity 0 $\sim$ 255), FDG-PET (before co-registration: 128 $\times$ 128 $\times$ 90, after co-registration: 79 $\times$ 79 $\times$ 91, intensity: 0 $\sim$ 255). Three sample images from different slices are shown in Fig. 8.

2) Image Processing: The MRI and FDG-PET images are firstly spatially normalized into a same template space to make them rigidly aligned with each other. This process is known as image co-registration, which is conducted through a MATLAB based library named Statistical Parametric Mapping (SPM 12 https://www.fil.ion.ucl.ac.uk/spm/software/spm12/). After co-registration, the size of FDG-PET and MRI become 79 $\times$ 79 $\times$ 91 (limited by the resolution of atlas used in SPM 12). In this

TABLE IV
PERFORMANCES OF DIFFERENT MODELS ON THE ADNI TEST DATASET

	MAE	SSIM	PSNR
Shallow CNN	25.312(±3.051)	0.842 (±0.061)	17.852 (±1.475)
sCT-DCNN	13.122(±3.452)	0.943 (±0.042)	21.932 (±2.919)
RIED-Net	**11.329(±3.278)**	**0.968 (±0.039)**	**23.527 (±2.826)**

TABLE V
p-VALUES OF t-TESTS ON PAIRWISE COMPARISONS: (A) MAE, (B) SSIM, (C) PSNR

(a)	Shallow CNN	sCT-DCNN	RIED-Net
Shallow CNN	-	<0.001	<0.01
sCT-DCNN	<0.001	-	0.011
RIED-Net	<0.001	0.011	-
(b)	Shallow CNN	sCT-DCNN	RIED-Net
Shallow CNN	-	<0.001	<0.001
sCT-DCNN	<0.001	-	0.027
RIED-Net	<0.001	0.027	-
(c)	Shallow CNN	sCT-DCNN	RIED-Net
Shallow CNN	-	<0.001	<0.001
sCT-DCNN	<0.001	-	0.008
RIED-Net	<0.001	0.008	-

experiment, based on the preliminary experiment results, we set the input and output patches to be 64 × 64. Training samples are extracted from each slice of the 3D image of each subject, to exclude slices with poor quality and limited brain regions, slice 1 ∼ 10 and slice 82 ∼ 91 are excluded. As a result, for each subject, 70 slices are extracted. Patches of size 64 × 64 are extracted at step size of 4 in each dimension and 3479 training patches are obtained from each training subject. In the second experiment, the parameter settings for the 3 models are the same as experiment I.

3) Comparison: For the proposed model, the total training time (80 iterations) is 6,321 seconds. The average time to generate a synthetic 3D PET image is 3.6 seconds. The training time/prediction time (one 3D synthetic PET image) for the sCT-DCNN and Shallow-CNN are 5,267/2.9 and 395/0.4 seconds respectively. Shallow-CNN has the fastest training and prediction time as it has the least parameters. The sCT-DCNN has less parameter than our proposed RIED-Net, its training process is faster (16.7%). However, once the model is trained, the gap in prediction time 24.1%.

Given the relatively small dataset we have in this experiment, the comparison of performance for different models is conducted based on the leave-one-out cross validation. Within each fold, all the other settings and procedure are the same as experiment I. Specifically, performance metric for each subject is calculated based on the average value across all slices. The final metric value is averaged across all 14 subjects and reported in Table IV and paired t-test results are summarized in Table V. From these two tables, we can conclude that significant performance differences still exist in experiment II.

Similar to the first experiment, we conclude shallow CNN underperforms sCT-DCNN and RIED-Net, and RIED-Net significantly outperforms sCT-DCNN in terms of all the three

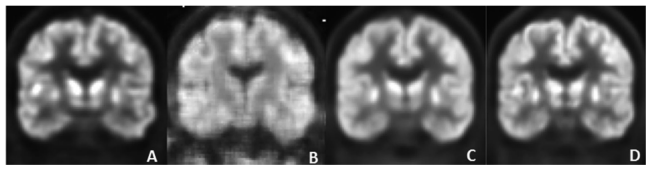

Fig. 9. Sample of one ground truth FDG-PET image (A). Output synthetic FDG-PET images of Shallow-CNN (B). sCT-DCNN(C). Our proposed RIED-Net (D).

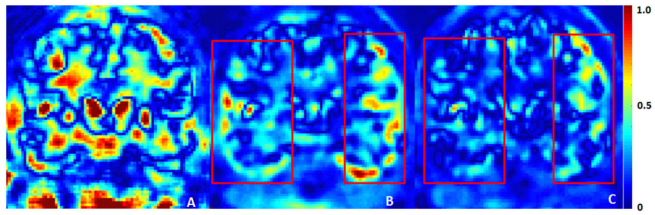

Fig. 10. Error maps of output images for Shallow-CNN (A). sCT-DCNN (B). Our proposed RIED-Net (C).

metrics. Among the 14 test cases, RIED-Net has higher SSIM and PSNRs on 11 (79%) and 12 (86%) test subjects respectively, lower MAEs on 12 images (86%). From Table IV, we can see that the improvements on MAE and PSNR is relatively significant (14% and 7%) compared with SSIM (2%). This is because MAE and PSNR are closely related to the pixel wised intensity difference; our method is proposed for better training/prediction performance and trained to get the minimal intensity wised difference between input and output images. On the other hand, SSIM is widely used to measure the perceptual quality, which considers the regional intensity mean and variation. The reduced pixel wised intensity difference does help marginally improve the perceptual quality of synthetic images.

Fig. 9 is the illustrative figure showing one image from each of the three models with Fig. 9A is ground truth FDG-PET image, Fig. 9B, 10C and 10D are output images of shallow CNN, sCT-DCNN and RIED-Net. In Fig. 10, A, B, and C are the error maps for the outputs from 3 models. The error maps are generated through the same procedure as experiment I.

As seen in Fig. 9, the output of shallow CNN (Fig. 9B) roughly restores the layout of ground true FDG-PET image (Fig. 9A) while with significant errors in details (Fig. 10A). In the error maps in Fig. 10, we can locate several regions where sCT-DCNN have higher errors in a larger area, for example, the two regions highlighted with the red bounding box in Fig. 10B, while RIED-Net shows lower errors in a smaller area in the same locations. If we map these regions back to the ground truth PET image, we can find that these regions have higher contrast, which means they are functioning more than surrounding regions and are more important for clinical use and interpretation. However, within such region, after several rounds of max pooling, only the pixels with peak intensities are kept; its neighboring pixels, though also have a relatively high intensity, are excluded. The excluded information is very difficult to reconstruct from the decoding path. As a result, the region with higher error tends

to expand. On the other hand, our proposed algorithm reserved some information for the peak point as well as its surrounding regions; the information helps the model get a better prediction for the surrounding regions. From the figure, we conclude RIED-Net has satisfying performance on synthesizing images across modalities.

V. Discussion and Conclusion

Image synthesis is becoming an important field in medical imaging research. Particularly so in scenarios where some image modalities may not be available due to accessibility/costs, radiation exposure, or need for intravenous contrast agent. To date, CNNs have shown the promise in medical imaging research mostly in imaging classification, detection, and segmentation. In this study, we propose a novel residual inception encoding-decoding network (RIED-Net) to address and enable image synthesis. There are two main contributions. First, the convolutional layers are introduced to reserve pixel information during the encoding process when the feature map size is reduced to increase receptive field size; deconvolutional layers are implemented to restore pixel information within the decoding process. Second, residual inception shortcut block is designed to address the gradient vanishing issues and improve the prediction accuracy. The performance of our proposed architecture is evaluated in two disparate imaging datasets. Comparison experiments confirm the outperformance of the proposed network model.

While promising, there is room for future work. For example, we do observe in Fig. 6 and Fig. 7 from the breast cancer study, all the models perform poorly on small regions of interest (e.g., suspicious tumor), as the ROI region is relatively small compared with the whole breast resulting in the models failing to train properly for small tumors. Future plans are to specifically study lesions <1 cm in size and train the model with this select cohort to improve performance.

References

[1] H. Greenspan, B. van Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1153–1159, May 2016.

[2] B. Sahiner *et al.*, "Classification of mass and normal breast tissue: A convolution neural network classifier with spatial domain and texture images," *IEEE Trans. Med. Imag.*, vol. 15, no. 5, pp. 598–610, Oct. 1996.

[3] B. Q. Huynh, H. Li, and M. L. Giger, "Digital mammographic tumor classification using transfer learning from deep convolutional neural networks," *J. Med. Imag.*, vol. 3, no. 3, 2016, Art. no. 034501.

[4] T. Araujo *et al.*, "Classification of breast cancer histology images using convolutional neural networks," *PLoS One*, vol. 12, no. 6, pp. 1–14, 2017.

[5] B. Microbiana *et al.*, "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1207–1216, May 2016.

[6] J. Yap, W. Yolland, and P. Tschandl, "Multimodal skin lesion classification using deep learning," *Exp. Dermatol.*, vol. 27, pp. 1261–1267, 2018.

[7] F. Ciompi *et al.*, "Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box," *Med. Image Anal.*, vol. 26, no. 1, pp. 195–202, 2015.

[8] H. Roth *et al.*, "Improving computer-aided detection using convolutional neural networks and random view aggregation," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1170–1181, May 2016.

[9] Q. Dou *et al.*, "Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1182–1195, May 2016.

[10] K. Sirinukunwattana, S. E. A. Raza, Y.-W. Tsang, D. R. J. Snead, I. A. Cree, and N. M. Rajpoot, "Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1196–1206, May 2016.

[11] W. Zhang, R. Li, H. Deng, L. Wang, W. Lin, S. Ji, and D. Shen, "Deep convolutional neural networks for multi-modality isointense infant brain image segmentation," *NeuroImage*, vol. 108, pp. 214–224, 2015.

[12] X. Zhao, Y. Wu, G. Song, Z. Li, Y. Zhang, and Y. Fan, "A deep learning model integrating FCNNs and CRFs for brain tumor segmentation," *Med. Image Anal.*, vol. 43, pp. 98–111, 2018.

[13] M. Havaei *et al.*, "Brain tumor segmentation with deep neural networks," *Med. Image Anal.*, vol. 35, pp. 18–31, 2017.

[14] W. Bulten, G. J. S. Litjens, C. A. Hulsbergen-van de Kaa, and J. van der Laak, "Automated segmentation of epithelial tissue in prostatectomy slides using deep learning," *Proc. SPIE*, vol. 10581, p. 27, 2018.

[15] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.

[16] R. Li *et al.*, "Deep learning based imaging data completion for improved brain disease diagnosis," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2014, pp. 305–312.

[17] W. Yang *et al.*, "Cascade of multi-scale convolutional neural networks for bone suppression of chest radiographs in gradient domain," *Med. Image Anal.*, vol. 35, pp. 421–433, 2017.

[18] B. K. Patel, S. A. Garza, S. Eversman, Y. Lopez-Alvarez, H. Kosiorek, and B. A. Pockaj, "Assessing tumor extent on contrast-enhanced spectral mammography versus full-field digital mammography and ultrasound," *Clin. Imag.*, vol. 46, pp. 78–84, 2017.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.

[20] S. G.-F. Shen, Z. Tang, K.-C. Chen, J. J.Xia, and D. Shen, "Joint craniomaxillofacial bone segmentation and landmark digitization by context-guided fully convolutional networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2018, vol. 3, pp. 720–728.

[21] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2016, pp. 424–432.

[22] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis.*, 2016, pp. 565–571.

[23] F. Gao *et al.*, "SD-CNN: A shallow-deep CNN for improved breast cancer diagnosis," *Comput. Med. Imag. Graph.*, vol. 70, pp. 53–62, 2018.

[24] K. He, X. Zhang, and S. Ren, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[25] Z. Wu, C. Shen, and A. van den Hengel, "Wider or deeper: Revisiting the ResNet model for visual recognition," *Pattern Recognit.*, vol. 90, pp. 119–133, 2019.

[26] H. Chen *et al.*, "Low-dose CT with a residual encoder-decoder convolutional neural network (RED-CNN)," *IEEE Trans. Med. Imag.*, vol. 36, no. 12, pp. 2524–2535, Dec. 2017.

[27] X. Han, "MR-based synthetic CT generation using a deep convolutional neural network method," *Med. Phys.*, vol. 44, no. 4, pp. 1408–1419, 2017.

[28] H. Greenspan, M. Frid-Adar, E. Klang, I. Diamant, M. Amitai, and J. Goldberger, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018.

[29] K. He and J. Sun, "Convolutional neural networks at constrained time cost," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5353–5360.

[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014. [Online]. Available: https://arxiv.org/abs/1409.1556.

[31] C. Szegedy *et al.*, "Going deeper with convolutions," 2014. [Online]. Available: https://arxiv.org/abs/1409.4842.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on Imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

[33] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[34] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 1520–1528.
[35] E. A. Morris, "Contrast-enhanced digital mammography," in *Diseases of the Brain, Head Neck, Spine 2016-2019: Diagnostic Imaging*. Berlin, Germany: Springer, 2016, vol. 69, pp. 339–342.
[36] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[37] A. Fakhry, T. Zeng, and S. Ji, "Residual deconvolutional networks for brain electron microscopy image segmentation," *IEEE Trans. Med. Imag.*, vol. 36, no. 2, pp. 447–456, Feb. 2017.
[38] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, vol. 4, p. 12, 2017.
[39] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1800–1807.
[40] M. Z. Alom, M. Hasan, C. Yakopcic, and T. M. Taha, "Inception recurrent convolutional neural network for object recognition," 2017. [Online]. Available: https://arxiv.org/abs/1704.07709.
[41] M. W. Weiner *et al.*, "Impact of the Alzheimer's disease neuroimaging initiative, 2004 to 2014," *Alzheimer's Dementia*, vol. 11, no. 7, pp. 865–884, 2016.
[42] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[43] X. Han, "MR-based synthetic CT generation using a deep convolutional neural network method," *Med. Phys.*, vol. 44, no. 4, pp. 1408–1419, 2017.
[44] J. G. Elmore, K. Armstrong, C. D. Lehman, and S. W. Fletcher, "Clinician's corner screening for breast cancer," *J. Am. Med. Assoc.*, vol. 293, no. 10, pp. 1245–1256, 2005.
[45] E. Warner *et al.*, "Surveillance of BRCA1 and BRCA2 mutation carriers with magnetic resonance imaging, ultrasound, mammography, and clinical breast examination," *J. Am. Med. Assoc.*, vol. 292, no. 11, pp. 1317–1325, 2004.
[46] M. A. Francescone *et al.*, "Low energy mammogram obtained in contrast-enhanced digital mammography (CEDM) is comparable to routine full-field digital mammography (FFDM)," *Eur. J. Radiol.*, vol. 83, no. 8, pp. 1350–1355, 2014.
[47] E. M. Fallenberg *et al.*, "Contrast-enhanced spectral mammography versus MRI: Initial results in the detection of breast cancer and assessment of tumour size," *Eur. Radiol.*, vol. 24, no. 1, pp. 256–264, 2014.
[48] Y. C. Cheung *et al.*, "Diagnostic performance of dual-energy contrast-enhanced subtracted mammography in dense breasts compared to mammography alone: Interobserver blind-reading analysis," *Eur. Radiol.*, vol. 24, no. 10, pp. 2394–2403, 2014.
[49] E. Luczyńska, S. Heinze-Paluchowska, S. Dyczek, P. Blecharz, J. Rys, and M. Reinfuss, "Contrast-enhanced spectral mammography: Comparison with conventional mammography and histopathology in 152 women," *Korean J. Radiol.*, vol. 15, no. 6, pp. 689–696, 2014.
[50] J. Gillman, H. K. Toth, and L. Moy, "The role of dynamic contrast-enhanced screening breast MRI in populations at increased risk for breast cancer," *Women's Health*, vol. 10, no. 6, pp. 609–622, 2014.
[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. [Online] Available: https://arxiv.org/abs/1412.6980.
[52] B. Dubois *et al.*, "Advancing research diagnostic criteria for Alzheimer's disease: The IWG-2 criteria," *Lancet Neurol.*, vol. 13, no. 6, pp. 614–629, 2014.
[53] M. W. Weiner *et al.*, "Impact of the Alzheimer's disease neuroimaging initiative, 2004 to 2014," *Alzheimer's Dementia*, vol. 11, no. 7, pp. 865–884, 2015.