SPECTRALLY-ENFORCED GLOBAL RECEPTIVE FIELD FOR CONTEXTUAL MEDICAL IMAGE SEGMENTATION AND CLASSIFICATION

Yongzhi Li[†]*, Lu Chi*[§]*, Guiyu Tian*[†]*,Yadong Mu*[§*]*, Shen Ge*[‡]*,Zhi Qiao*[‡]*,Xian Wu*[‡]*,Wei Fan*[‡]

[†]Center for Data Science, Peking University; {yongzhili,wavey}@pku.edu.cn
[§]Wangxuan Institute of Computer Technology, Peking University; {chilu,myd}@pku.edu.cn
[‡] Tencent; {shenge,xiaobuqiao,kevinxwu,davidwfan}@tencent.com

ABSTRACT

Deep convolutional neural networks (CNNs) have recalibrated the state-of-the-art for a plethora of applications in medical image analyzing such as segmentation and classification. Large receptive field is crucial for modeling long-range spatial dependency in medical images. In this paper, we propose a novel architectural network design for accomplishing a full-image global receptive field, which we call spectral residual block (SRB). Specifically, we propose to utilize a unitary transform that essentially conducts a local-to-global transform. All elements are mapped to spectral domain and thus globally depend on each other. A variety of global operators are carefully devised and efficiently enforce a full-image receptive field, including spectral ReLU for frequency-sensitive filtering and spectral convolutions. The output in spectral domain is eventually converted back global-to-local via a reverse unitary transform. The proposed framework is generic and flexible, and could be applied to various network structures and tasks. Comprehensive evaluations on skin lesion segmentation and Chest X-Ray classification show that our method achieves the state-of-the-art performance, demonstrating both effectiveness and efficiency.

Index Terms— Non-local, Medical image analysis, Segmentation, Classification, Neural network

1. INTRODUCTION

Deep convolutional neural networks (CNNs) have become major workhorses for a variety of medical image classification and segmentation applications [1, 2]. The key building block in a majority of neural network architectures is the convolutional layer, which relies on local connection and learns filters capturing informative patterns across local spatial neighborhood and feature channels. For many image-oriented tasks (such as human pose estimation [3] or image segmentation [1]), investigating long-range spatial dependencies, rather than local image pixels or regions, plays a key role in contextual modeling. Typically, such long-distance dependencies are modelled by large receptive fields, whose extents can be enlarged by recurrently stacking many small-kernel convolutional layers (such as the 3×3 kernels in ResNet [4]) or using dilated or deformable convolutions [5]. More recent solutions devise various non-local connections with learnable parameters, exemplified by the self-attentive operator in [6].

All above-mentioned non-local convolutional schemes suffer from several weaknesses. Firstly, the dominant solutions like dilation [7] or deformable convolutions [8, 5] are still essentially local. To capture long-range dependencies, they need to be applied repeatedly to convey some message to far locations. It only connects distant sites indirectly, thus not effective enough. In addition, despite of fewer parameters to be tuned, it is also computationally inefficient and arguably complicates the numerical optimization. Secondly, the recent trend of directly connecting two positions within a feature map [6, 9], though avoiding recurrent multi-hop propagation, significantly increases the computational complexity and is thus beyond the scope of most practitioners.

This work novelly designs *spectral residual block*, which can be inserted into an ordinary neural network for accomplishing a global receptive field. This is achieved by two paired operations: local-to-global and global-to-local transforms. The former converts a feature map into some spectral domain, wherein updating an element globally affects all others. A number of spectral operators are defined, including spectral ReLU for frequency-sensitive filtering and spectral convolutions for channel-wise information fusion. When applied, they efficiently enforce a full-image receptive field. The output in the spectral domain is finally de-correlated via the global-to-local transform. Both cross-domain transforms are efficiently implemented by some bi-linear unitary transforms, either fixed or learnable by gradient back-propagation.

There are several advantages of spectral residual block: 1) The proposed framework is generic, instantiated by various unitary transforms and spectral operators. In particular, we mainly choose Fourier transform for demonstration pur-

*Yadong Mu is the corresponding author. Part of this work was conducted when Yongzhi Li was a research intern at Tencent Medical AI Lab. This work is supported by Beijing Natural Science Foundation (Z190001), Beijing Municipal Commission of Science and Technology (Z181100008918005) and National Natural Science Foundation of China (61772037).

978-1-7281-1331-9/20/$31.00

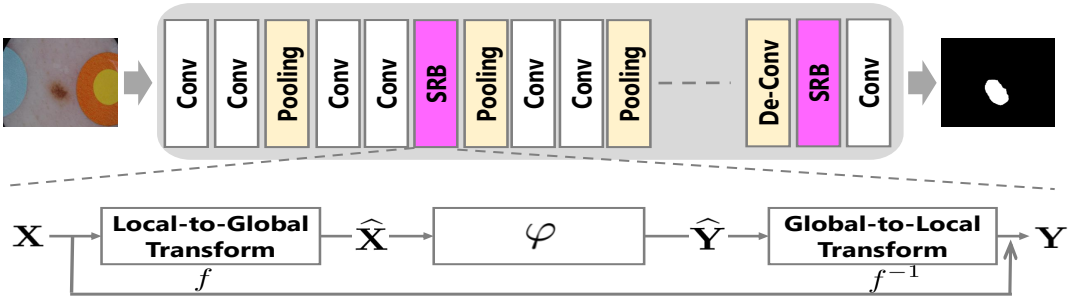

Fig. 1. Illustration of the spectral residual block (SRB). $\widehat{\mathbf{X}} = f(\mathbf{X})$ and $\mathbf{Y} = f^{-1}(\varphi(\widehat{\mathbf{X}})) + \mathbf{X}$ (note the residual link between $\mathbf{X}, \mathbf{Y}$).

pose; 2) It provides more powerful long-range dependencies than [8, 5] by spectrally computing interactions between any two positions, regardless of their positional distance. 3) It also requires significantly lower complexity and fewer parameters than [6, 9], to achieve a similar level of performance.

We experiment with two medical image analysis tasks, skin lesion segmentation and chest X-Ray classification, to showcase the effectiveness of the proposed method. For the segmentation task, the global receptive field helps to diagnose some large or indistinguishable skin lesions, while in chest X-ray classification our method allows the model to analyze a larger range area to detect the potential abnormal region, thus long-range context is regarded to be useful in both tasks. Comprehensive evaluations and ablative studies are conducted on two medical image benchmarks. Our method achieves state-of-the-art results on all experimental settings. This strongly demonstrates the empirical effectiveness of richer contextual modeling brought by global receptive field in a spectral domain.

2. FORMULATION AND INSTANTIATION

In this section, we first introduce the detailed formulation of the module we designed and some symbol definitions. Then we introduce two kinds of instantiations of the local-to-global transform. Finally, we analyze the computational complexity of each part of the proposed module.

Formulation: Fig. 1 shows a typical application scenario for spectral residual block (hereafter SRB in short). Given an ordinary deep CNN (like Unet [1] or ResNet [4]), the SRB can be flexibly inserted wherever long-range dependencies need to be tackled. Let $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{H \times W \times C}$ be the input / output tensors to an SRB, respectively. The mechanism of an SRB can be compactly described as:

$$\mathbf{Y} = f^{-1}\left(\varphi(f(\mathbf{X})) + f(\mathbf{X})\right) = f^{-1}\left(\varphi(f(\mathbf{X}))\right) + \mathbf{X}, \quad (1)$$

where f denotes the local-to-global linear transform while f^{-1} denotes the global-to-local transform and $f^{-1}(f(\mathbf{X})) = \mathbf{X}$ holds. $\varphi : \mathbb{R}^{H \times W \times C} \mapsto \mathbb{R}^{H \times W \times C}$ is some task-specific sub-network that operates on the global tensor $\widehat{\mathbf{X}} = f(\mathbf{X})$.

Here, we introduce two unitary matrices on the complex domain $\mathbf{P} \in \mathbb{C}^{H \times H}$ and $\mathbf{Q} \in \mathbb{C}^{W \times W}$. Let $concat()$ be an operator that concatenates a number of arrays along specific dimension, and $\mathbf{X}_c \in \mathbb{R}^{H \times W}$ be the c_{th} channel of $\mathbf{X}$. The global-to-local / local-to-global transforms are computed in a channel-wise fashion as below for efficacy consideration:

$$f(\mathbf{X}) = concat\left(\mathbf{P}\mathbf{X}_1\mathbf{Q}, \ldots, \mathbf{P}\mathbf{X}_C\mathbf{Q}\right), \quad (2)$$

$$f^{-1}(\widehat{\mathbf{Y}}) = concat\left(\mathbf{P}^*\widehat{\mathbf{Y}}_1\mathbf{Q}^*, \ldots, \mathbf{P}^*\widehat{\mathbf{Y}}_C\mathbf{Q}^*\right) \quad (3)$$

where * denotes conjugate transpose matrix. It is obvious that $f(\mathbf{X})$ is globalized, with each entry correlated with all others in the same feature channel. A simple derivation of Eqn. (1) shows $\varphi(f(\mathbf{X})) = f(\mathbf{Y}) - f(\mathbf{X}) = f(\mathbf{Y} - \mathbf{X})$ since f is linear. It implies that φ is actually designed to approximate the residual $\mathbf{Y} - \mathbf{X}$ in some global domain induced by f.

Instantiation of f : In this proposed framework, f can be flexibly defined. This work provides two different instantiations, one is with fixed $\mathbf{P}$ and $\mathbf{Q}$ implemented as discrete Fourier transform (DFT), while the other one with learnable matrix adapted via gradient back-propagation.

Fourier transform is widely adopted in signal processing. It converts a function of signals into the frequencies that make it up, in a reversible way. Discrete Fourier transform can be accomplished by matrix multiplication over complex numbers. The 2-D DFT can be formulated in the form of:

$$DFT(\mathbf{X}) = \mathbf{P}_{FFT}\mathbf{X}\mathbf{Q}_{FFT} \quad (4)$$

where $\mathbf{P}_{FFT}, \mathbf{Q}_{FFT}$ are some special structural unitary matrices that enable fast Fourier transform (FFT). In this senario, we set $f(\mathbf{X}) = DFT(\mathbf{X})$ in Eqn. (2) and f^{-1} to inverse DFT. It should be noted that some other commonly used spectral transforms such as discrete cosine transform (DCT), or discrete wavelet transform (DWT) can also be instantiated as f. For efficiency, we only discuss the DFT version here.

The $\mathbf{P}$ and $\mathbf{Q}$ matrices are fixed in the DFT version. We further propose a version which allows $\mathbf{P}$ and $\mathbf{Q}$ to be learned automatically in training. In this version, both $\mathbf{P}$ and $\mathbf{Q}$ are initialized as real orthogonal matrices with normal distribution. In other words: $\mathbf{P}\mathbf{P}^T = \mathbf{P}^T\mathbf{P} = \mathbf{I}_{H \times H}$ and $\mathbf{Q}\mathbf{Q}^T = \mathbf{Q}^T\mathbf{Q} = \mathbf{I}_{W \times W}$. All values in the matrix can be adapted via the gradient back-propagation process. However, simple application of back-propagation may lead to non-orthogonal matrices. To ensure the orthogonality, a QR decomposition is adopted as a post-processing after each gradient-based update. This can be

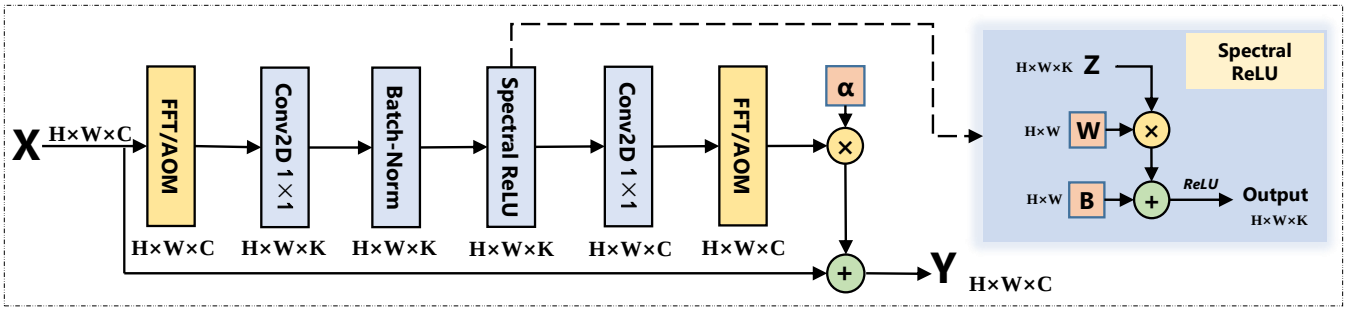

Fig. 2. The instantiation of SRB where f is FFT or AOM, f^{-1} is inverse FFT or AOM. Channel number in the blue blocks $K = C/2$ in practice. φ is comprised of spectral version of 1×1 group convolution / BN / ReLU. The peachy squares represent learnable parameters.

expressed as follows:

$$\mathbf{P}^t = \tilde{\mathbf{P}}^t \mathbf{R}_p^t, \qquad \mathbf{P}^{t+1} \leftarrow \tilde{\mathbf{P}}^t \tag{5}$$

$$\mathbf{Q}^t = \tilde{\mathbf{Q}}^t \mathbf{R}_q^t, \qquad \mathbf{Q}^{t+1} \leftarrow \tilde{\mathbf{Q}}^t \tag{6}$$

where $\tilde{\mathbf{P}}^t, \mathbf{R}_p^t$ are orthogonal and upper triangular matrix at step t, respectively. And $\mathbf{P}^{t+1}$ is simply set to $\tilde{\mathbf{P}}^t$ in the post-processing. The $\mathbf{Q}$ matrix could also be re-orthogonalized in a similar fashion. We call this adaptive orthogonal matrix (AOM) version.

Instantiation of φ : Fig. 2 shows an instance of φ inspired by the bottle-neck block in ResNet [4]. It is comprised of a pair of 1×1 convolutions which do channel reduction and promotion respectively, batch normalization (BN) and ReLU. Whenever needed, we use the prefix *spectral* to emphasize all operations are conducted in a spectral domain. The computations of spectral convolutions / BN are identical to ordinary versions, yet operating on spectral frequencies and with each update affect all image positions. For a globalized tensor $\mathbf{Z} \in \mathbb{R}^{H \times W \times K}$, spectral ReLU is controlled by learnable parameters $\mathbf{W}, \mathbf{B} \in \mathbb{R}^{H \times W}$ and defined as:

$$\mathbf{Z}_{i,j,k} \leftarrow \max\left(\mathbf{W}_{i,j} \cdot \mathbf{Z}_{i,j,k} + \mathbf{B}_{i,j},\, 0\right), \tag{7}$$
$$(i,j,k) \in [1 \dots H] \times [1 \dots W] \times [1 \dots K],$$

which essentially does the job of frequency-sensitive filtering. The learnable parameter α in Fig. 2 is introduced to balance the residual block and original $\mathbf{X}$ as defined in Eqn. (1). And $\alpha = 0$ initializes an identity mapping in practice.

3. EVALUATIONS

To evaluate the effectiveness of our proposed method, we experimented on two large public medical datasets. Extensive ablative studies are also conducted to investigate the key factors. For all the baselines, we adopt the original implementations by authors if it is available, or re-implemented by ourself using the PyTorch framework.

3.1. Skin Lesion Segmentation

The first study is conducted on the ISIC Archive dataset [10] for skin lesion segmentation, which is known as the largest

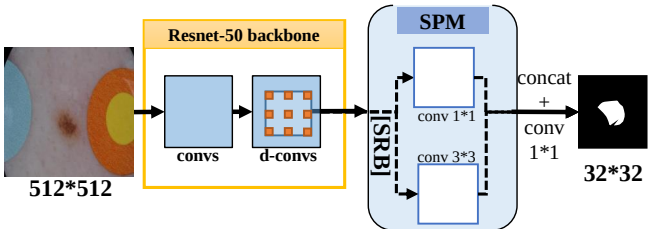

Fig. 3. The Resnet-50 backbone and the SPM inspired by [11]. One SRB block is inserted before the 3*3 conv to capture the global information. Dilated convolution is adopted to keep the feature map resolution, which is denoted as "d-convs" in the diagram.

publicly available collection of quality-controlled dermoscopic images of skin lesions. The lesion images were acquired with a variety of dermatoscope types, from all anatomic sites (excluding mucosa and nails). Ground truth masks are obtained by several means: fully automatically labeled and checked by human experts, hand-crafted flood-fill algorithm, or manual polygon tracing. All 2,594 images with mask annotations are included in our experiments.All original images are uniformly resized to 512×512 before being fed to the networks. Since the ground truth labels of original test data are confidentially kept by dataset organizer, the direct comparison between our methods and the original competing methods is infeasible. We randomly shuffle the entire data and build our own data partition: 90% for training and 10% for evaluation.

Classic U-Net [1] and a simple parallel module (SPM) inspired by the DeepLab_V3 [11] are adopted as our basic models, waiting for SRB application. The two basic models' architecture are shown in the Fig. 1 and Fig. 3 respectively. The popular ResNet-50 [4] is selected as the backbone of all basic models. For fair comparisons, all models are trained from scratch and no data augmentation is applied. The pixel-wise cross-entropy loss is used as the loss function and an Adam optimizer with default setting (learning rate=10^{-3}, momentum=0.9, weight-decay=10^{-4}) is adopted to tune all the parameters. We train all models for 150 epochs for full convergence. Following FCN [12], pixel-wise segmentation accuracy and IoU metrics on the test set are jointly used for performance comparison.

We first select three very powerful methods as

Table 1. Comparison with state-of-the-art methods on skin lesion segmentation. All models except the U-Net use the ResNet-50 as backbone. FW IoU is the frequency weighted intersection over union. Over Acc represents the Overall Accuracy. Params indicates the number of model parameters.

Model	Mean Acc	Over Acc	Mean IoU	FW IoU	Params(M)
PSPNet	92.11	95.55	87.23	91.58	68.06
DeepLabV3	92.66	**95.73**	87.76	91.93	39.76
DANet	92.84	95.60	87.51	91.72	49.49
U-Net	91.84	94.64	85.17	90.08	1.943
+4 SRB_F	92.93	95.66	87.66	91.83	2.076
SPM	93.14	95.62	87.60	91.77	29.59
+1 SRB_F	**93.25**	95.72	**87.87**	**91.95**	31.70

Table 2. Experimental results and ablation study for skin lesion segmentation. U-Net* denotes the model removed the last maxpooling layer. upN denotes the N_{th} up-sampling layer in U-Net's decoder. Among all, up4 is closest to the encoder. The sign a/ means after and b/ means before.

Model	Mean Acc	Over Acc	Mean IoU	FW IoU
U-Net	91.84	94.64	85.17	90.08
+1 SRB_F	92.21	95.41	86.92	91.35
+2 SRB_F	92.80	95.46	87.17	91.48
+3 SRB_F	92.75	95.64	87.57	91.78
+4 SRB_F	92.93	**95.66**	**87.66**	**91.83**
+1 SRB_A	92.02	95.40	86.86	91.33
+2 SRB_A	92.09	95.60	87.34	91.67
+3 SRB_A	92.65	95.40	87.01	91.38
+4 SRB_A	**92.95**	95.60	87.53	91.73
U-Net*	90.60	93.79	83.10	88.65
+1 a/ up2	92.14	95.05	86.12	90.77
+1 a/ up3	91.87	95.48	86.99	91.44
+1 a/ up4	91.87	95.54	87.14	91.55
+1 b/ up4	**92.38**	**95.61**	**87.42**	**91.70**
+1 Nonlocal	91.92	95.33	86.69	91.22
+1 A^2-Net	91.36	95.09	86.03	90.78

the baselines for comparison, which are PSPNet [13], DeepLab_V3 [11] and DANet [14] respectively. Results are shown in the Table 1. It can be seen that by inserting four SRB_Fs into the U-Net, there is a huge improvement in each indicator, especially the $\sim$2.5 points improvement on Mean IoU. And this help the U-Net achieve the comparable state-of-the-art result with much less parameters (nearly 20 times less than DeepLabV3). At the same time, only one SRB_F module insertion in SPM made it instantly exceed all the baselines and get a new state-of-the-art performance with less model parameters than the previous methods. Compared to U-Net, the promotion is not so impressive, and we believe this is because that the backbone ResNet-50 network is already a relatively deep network with a large receptive field, not to mention the use of dilated convolutions, which may lead to saturated performance.

Next, in order to better explore the influence of the numbers of SRB insertion, we choose the basic U-Net as the backbone and conduct some detailed ablation studies. First, we explore the number of SRBs and the effect of different f instantiation. The results are shown in the upper and middle parts of the Table 2. As seen from the table, inserting one SRB could bring a massive improvement, while the inclusion of more SRBs slightly improves in both FFT or AOM situations, yet the resultant gains diminished quickly. We believe that inserting one SRB can enhance the model's receptive field effectively, while the benefits of more are limited. At the same time, it can be observed that the upgrades brought by the different SRBs are similar. We believe the reason that the SRB_A did not achieve better results may due to the loss of some useful information in QR decomposition post-processing. Considering that SRB_A requires more computation, the SRB_F is thus more favored in practice.

Table 3. Comparisons with other non-local blocks on skin lesion segmentation. The increment of computation and parameters by adding one block before the 4_{th} up-sampling layer in the U-Net*.

Model	SRB_F	SRB_A	Non-local	A^2-Net
Δ GFLOPs	0.028	0.031	2.126	0.071
Δ Params(K)	8.704	8.448	33.02	16.64

To explore the importance of the receptive field and the effect of insertion locations in the network backbone, we first remove the max_pooling_4 layer in the decoder of U-Net (denoted as U-Net* in Table 2), which brings a reduced receptive field for segmentation and thus supposedly easier to observe the effect of SRB. The lower part of the Table 2 shows that after removing the topmost layers, the baseline U-Net* drops by 2 points in terms of Mean IoU. The models obtained by inserting a single SRB_F at different locations consistently improve the performance metrics. Interestingly, the deeper the inserted position is, the better performance we get. This is because the feature of the deeper position contains more high-level semantic information, which is more conducive to extract relevant information.

For fair comparison, we also compared the performance with two other non-local neural blocks the Non-local [6] and A^2-Nets [15]. Same as the above setting, we insert one block into the U-Net* before the 4_{th} up-sampling layer, respectively. And the result are shown in the bottom part of Table 2. It is obvious that our method achieved the best result with the lowest computation complexity and least parameters increment, which can be clarified in the Table 3.

To demonstrate the effectiveness more intuitively. Fig. 4 contrasts some segmentation masks obtained by U-Net and our proposed method. Guided by richer context brought by global receptive field, our model can achieve excellent results even on many challenging scenarios. Take the picture in the column (C) as an example, the lesion is very inconspicuous, and U-Net can not segment any region. Benefit from the global receptive field in SRB insertion, the overall information is fused and compared, and then the inapparent lesions are more easily to be distinguished in our method.

3.2. Thoracic Disease Classification

To evaluate the proposed method on the medical image classification task, we conduct the second experiment on the ChestX-ray14 [16] dataset, which collects 112,120 frontal-view images of 30,805 unique patients. Among the images, 51,708 are labeled with up to 14 pathologies, while others are marked as "no finding". In this way, each image is associated

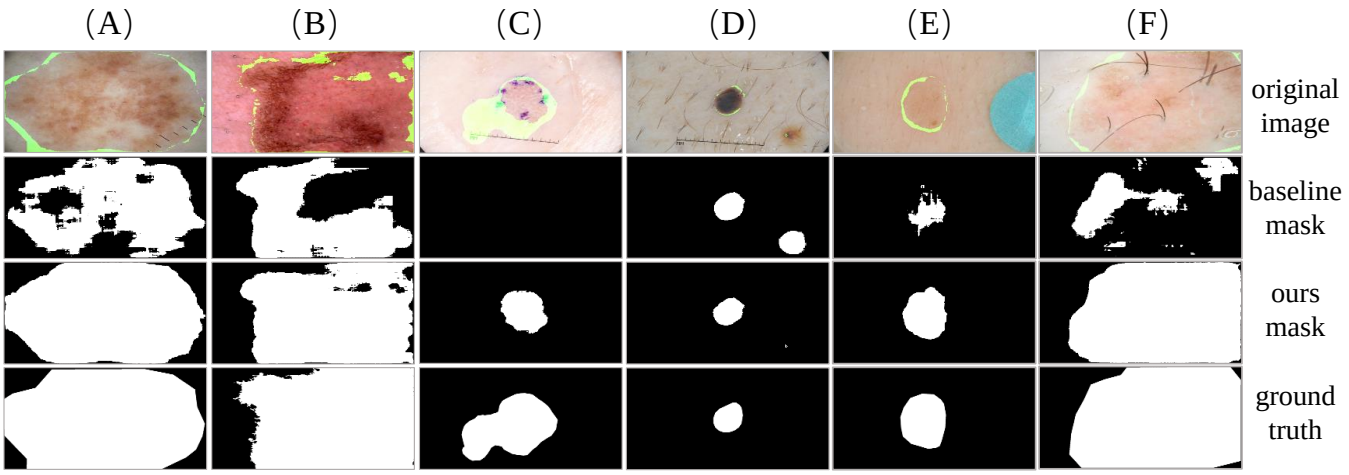

Fig. 4. Visualization of skin lesion segmentation results. Each column is an example. The green region on original images indicate the residual between our result and the ground truth.

Table 4. Comparing classification performances on ChestX-ray14 [16]. Each pathology is denoted by its first four characters, except that Pneu1 and Pneu2 are for Pneumonia and Pneumothorax respectively. Due to space limit, only a subset of pathologies are shown. The "mean" column is averaged over all 14 pathologies. R50 / D121 are the abbreviations of ResNet-50 / DenseNet-121 respectively.

Method	Atel	Card	Effu	Infi	Mass	Nodu	Pne1	Pne2	Cons	Edem	Fibr	Pleu	Mean
Wang et al.[16]	0.716	0.807	0.784	0.609	0.706	0.671	0.633	0.806	0.708	0.835	0.769	0.708	0.738
Yao et al.[17]	0.772	0.904	0.859	0.695	0.792	0.717	0.713	0.841	0.788	0.882	0.767	0.765	0.803
Rajpurkar et al.[18]	0.821	0.905	0.883	0.720	0.862	0.777	0.763	0.893	0.794	0.893	0.804	0.814	0.842
Kumar et al.[19]	0.762	0.913	0.864	0.692	0.750	0.666	0.715	0.859	0.784	0.888	0.756	0.774	0.795
Li et al.[20]	0.800	0.870	0.870	0.700	0.830	0.750	0.670	0.870	0.800	0.880	0.780	0.790	0.806
R50	0.823	0.908	0.881	0.715	0.868	0.777	0.760	0.881	0.804	0.901	0.824	0.820	0.837
R50 + 2 SRB_F	**0.830**	**0.917**	**0.887**	**0.726**	**0.881**	**0.797**	**0.766**	0.896	0.811	**0.906**	0.834	**0.828**	**0.848**
D121	0.825	0.915	0.882	0.715	0.873	0.778	0.758	0.891	**0.813**	0.901	0.833	0.820	0.841
D121 + 2 SRB_F	0.828	0.914	0.883	0.714	0.876	0.791	0.762	**0.900**	0.811	0.903	**0.835**	0.816	0.847

with a 14-D multi-hot label vector, with each dimension representing for a separate pathology. We randomly shuffle and split it into two sub-sets with a ratio of 1:4. The larger subset with 89696 images is used for model training and the rest with 22424 images is used for the evaluation purpose.

In all experiments, the original images will be first resized to 240×240. Standard engineering tricks (such as random cropping, color jittering, horizontal flipping, mulit-crop average etc.) are applied at either the training or evaluation stage. Following common practice [18], ResNet-50 and DenseNet-121 are utilized as our network backbones, and a weighted binary cross-entropy layer is appended to render the classification loss. All models are optimized in a typical setting (*e.g.*, Adam optimizer, an initial learning rate of 10^{-3} that attenuates by 1/10 for every 20 epochs, a momentum of 0.9). The early stop strategy is adopted to avoid over-fitting. AUC (the area under the receiver operating characteristic curve) score is used as the performance metric.

Table 4 summarizes the AUC scores obtained by various competing methods and several variants of our model. It can be observed that our method dominates all the baseline methods by inserting two additional SRB_Fs into the basic models (R50 and D121) in Table 4. We can also find that the same SRB could bring more promotion on the ResNet-50 than the DenseNet-121. This is also due to the fact that deeper models (DenseNet-121) have an already larger receptive field, which is consistent with the conclusion we have in the skin segmentation experiments. However, with the addition of SRB, we can achieve the comparable or even better results by simpler models with fewer parameters than larger models.

For ablative studies, we also investigate different locations in the backbone to insert SRB (*e.g.*, after different residual blocks of ResNet-50) and the number of inserted SRB (varying from 1-3 SRBs). However, the resulting changes of the performances with SRB are surprisingly almost negligible. We attribute this to the fact that classification is a single-output problem and is thus less dependent on the spatial context. Another explanation is that a quantity of thoracic disease are relatively located in a pretty small range, so that the global information has limited influences to improve the results, which can be backed up in Figure 5.

To interpret the network predictions, we also produce heatmaps to visualize the most indicative areas of the disease in the image using the class activation mappings (CAMs) [21]. Figure 5 shows several examples of the CAMs on the 14-class pathology classification task. We can clearly see that

for a wide range of lesions, our method can effectively diagnose and locate the lesions accurately, even if they are distributed in different parts of the chest, such as the effusion occurs in 1(C) and the infiltrate shown in 2(A). However, there are a lot of samples in which the pathology is just in a very small region, such as 2(B) and 2(C). In these cases, the global receptive field is difficult to play a large role, so there are some deviations on the pathologies localization.

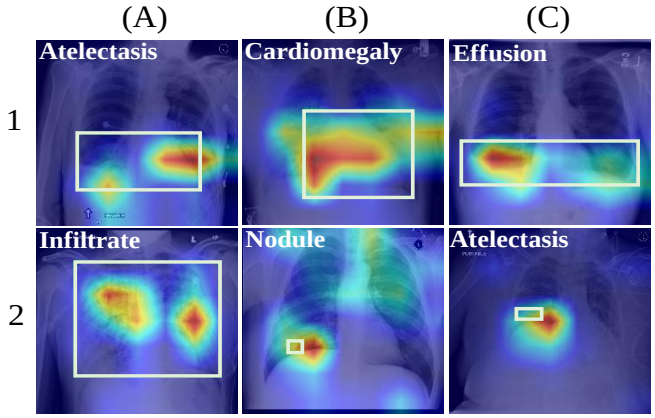

Fig. 5. Examples of pathologies localization using Class Activation Maps on the test images, which highlight the areas of the X-ray that are most important for making a particular pathology classification. All examples are positive for corresponding labels. The ground truth bounding boxes in green on the results are provided by the dataset.

4. CONCLUDING REMARKS

We proposed a novel method for learning global receptive field in deep neural networks. Our key idea is conducting residual learning in some spectral domain brought by bilinear unitary transform based local-to-global transform. We carefully design several spectral operators and empirically validate the proposed spectral residual blocks on two public large-scale datasets. Strong evidence is observed to demonstrate its effectiveness and generalisability on different medical image analysis tasks. For future work, we can extend the transform to 3D data, such as CT or MR data. By integrating information on different slices, we believe this will lead to further performance promotion.

5. REFERENCES

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

[2] Abhijit Guha Roy, Sailesh Conjeti, Debdoot Sheet, Amin Katouzian, Nassir Navab, and Christian Wachinger, "Error corrective boosting for learning fully convolutional networks with limited data," in *MICCAI*, 2017.

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[5] Mo Zhang, Xiang Li, Mengjia Xu, and Quanzheng Li, "RBC semantic segmentation for sickle cell disease based on deformable u-net," in *MICCAI*, 2018.

[6] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," in *CVPR*, 2018.

[7] Fisher Yu and Vladlen Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv: 1511.07122*, 2015.

[8] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei, "Deformable convolutional networks," in *ICCV*, 2017.

[9] Limin Wang, Wei Li, Wen Li, and Luc Van Gool, "Appearance-and-relation networks for video classification," in *CVPR*, 2018.

[10] Noel C. F. Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen W. Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael A. Marchetti, Harald Kittler, and Allan Halpern, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)," *arXiv: 1902.03368*, 2019.

[11] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv: 1706.05587*, 2017.

[12] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.

[13] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, "Pyramid scene parsing network," in *CVPR*, 2017.

[14] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu, "Dual attention network for scene segmentation," in *CVPR*, June 2019.

[15] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng, "Aˆ2-nets: Double attention networks," in *NIPS*. 2018.

[16] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers, "Chest X-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," in *CVPR*, 2017.

[17] Li Yao, Eric Poblenz, Dmitry Dagunts, Ben Covington, Devon Bernard, and Kevin Lyman, "Learning to diagnose from scratch by exploiting dependencies among labels," *arXiv: 1710.10501*, 2017.

[18] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al., "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv: 1711.05225*, 2017.

[19] Pulkit Kumar, Monika Grewal, Srivastava, and Muktabh Mayank, "Boosted cascaded convnets for multilabel classification of thoracic diseases in chest radiographs," in *International Conference Image Analysis and Recognition*. Springer, 2018, pp. 546–552.

[20] Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Li Fei-Fei, "Thoracic disease identification and localization with limited supervision," in *CVPR*, 2018, pp. 8290–8299.

[21] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *CVPR*, 2016.