Deep Frequency Re-calibration U-Net for Medical Image Segmentation

Reza Azad
Institute of Imaging and Computer Vision,
RWTH Aachen University, Germany
azad@lfb.rwth-aachen.de

Afshin Bozorgpour
Sharif University of Technology,
Iran
bozorgpour@ce.sharif.edu

Maryam Asadi-Aghbolaghi
School of Computer Science, Inst. for Research in
Fundamental Sciences (IPM), Iran
masadi@ipm.ir

Dorit Merhof
Institute of Imaging and Computer Vision,
RWTH Aachen University, Germany
dorit.merhof@lfb.rwth-aachen.de

Sergio Escalera
Universitat de Barcelona and
Computer Vision Center, Spain
sergio@maia.ub.es

Abstract

The human visual cortex is biased towards shape components while CNNs produce texture biased features. This fact may explain why the performance of CNN significantly degrades with low-labeled input data scenarios. In this paper, we propose a frequency re-calibration U-Net (FRCU-Net) for medical image segmentation. Representing an object in terms of frequency may reduce the effect of texture bias, resulting in better generalization for a low data regime. To do so, we apply the Laplacian pyramid in the bottleneck layer of the U-shaped structure. The Laplacian pyramid represents the object proposal in different frequency domains, where the high frequencies are responsible for the texture information and lower frequencies might be related to the shape. Adaptively re-calibrating these frequency representations can produce a more discriminative representation for describing the object of interest. To this end, we first propose to use a channel-wise attention mechanism to capture the relationship between the channels of a set of feature maps in one layer of the frequency pyramid. Second, the extracted features of each level of the pyramid are then combined through a non-linear function based on their impact on the final segmentation output. The proposed FRCU-Net is evaluated on five datasets ISIC 2017, ISIC 2018, the PH^2, lung segmentation, and SegPC 2021 challenge datasets and compared to existing alternatives, achieving state-of-the-art results.

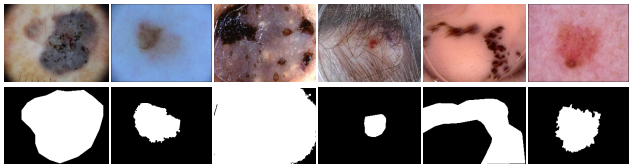

Figure 1. Skin lesion examples showing large visual variability.

1. Introduction

Medical imaging is a key element in computer-aided diagnosis and smart medicine. Obtaining accurate results from medical imaging allows to enhance diagnostic efficiency, resulting in a reduction of time, cost, and error of human-based processing. Different modern medical imaging approaches like Magnetic Resonance Imaging (MRI), or Computed Tomography (CT), are useful for the medical examination of different parts of the human body. Therefore, automated processing of these kinds of imaging data is essential to support diagnosis and treatment of diseases.

Medical image segmentation is an important and effective step in numerous medical imaging tasks. To help clinicians to make an accurate diagnosis, and shorten the time-consuming inspection and evaluation processes, it is required to pre-segment some crucial tissues or abnormal features in medical images. Image segmentation includes a large number of applications ranging from skin cancer detection in RGB images, lung tissue segmentation in CT images to pathological image analysis.

Medical image segmentation is a challenging task due to several complexities. E.g. in case of skin cancer segmentation, there are large intra-class variabilities and inter-class similarities because of differences in color, texture, shape, size, contrast, and location of the lesions (Figure 1). Low contrast and obscuration which can be observed between the affected areas and normal regions make the recognition a hard task. To overcome these issues, different approaches have been proposed for medical image segmentation. Like other fields of research in computer vision, deep learning-based networks have outperformed traditional machine learning approaches. Convolutional neural networks (CNN), inspired by the human visual cortex, have been a widely used deep network. It can learn complex feature hierarchies from the data layer by layer.

The main drawback of deep networks is their extreme hunger for annotated training data. However, in medical image segmentation, large (and annotated) datasets are hard to obtain, due to the burden of manual annotation. To deal with this issue Ronneberger et al. [31] extended the idea of fully convolutional neural network (FCN) [26] to U-Net. Compared to the previous approaches, U-Net was able to produce better performance and also leverage the need of large amounts of training data. This network includes an encoding and a decoding path. The encoder extracts a large number of feature maps by reducing the dimensionality. On the other hand, the decoder produces the segmentation maps by applying a hierarchical series of up-convolutional layers.

To improve the performance of U-Net, many extensions of this network have been proposed [3, 4, 15, 5, 14, 32, 35]. These networks aim to enhance the original U-Net by inserting attention mechanisms, recurrent residual strategies, or other non-linear functions in the convolutional layers. What all these networks have in common is their bias towards extracting features based on texture rather than shape. This fact limits the ability of these convolutional neural networks (CNNs) to leverage useful low-frequency information, e.g. shape information [17]. It has been shown that the representation power of CNNs can be improved by employing shape information through adjusting input images [17]. However, it is still an open problem to design an efficient approach for CNNs that can attenuate high-frequency local components and benefit from low-frequency information.

To address the above problems, in this paper we propose our so-called Frequency Re-calibration U-Net (FRCU-Net). We introduce a frequency level attention mechanism to control and aggregate the representation space using a weighted combination of different types of frequency information. To take advantage of both texture and shape features based on their effect on the performance, we propose to include the Laplacian pyramid in the bottleneck layers of the U-Net. The low-frequency domain from the Laplacian pyramid causes the network to learn shape information while the high-frequency level is responsible for texture-based features, resulting in a reduction of the effect of the noise on the final representation. We employ the Laplacian pyramid inspired by other successful traditional image processing tools like SIFT, where Laplacian pyramid was shown to have a high representative power for describing the object in various frequency domains by deploying Gaussian kernels.

To enhance the discriminative power of different channels of one frequency level, a channel-wise attention mechanism is exploited to re-calibrate the frequency representations, inspired by the effectiveness of the proposed squeeze and excitation modules [22]. We then propose to employ a weighted combination function to aggregate the features of all levels of the Laplacian pyramid and allow the network to learn weights of the levels based on their importance on the final result. This mechanism helps the network to focus more on the informative and meaningful features while suppressing noisy ones by using global embedding information of the channels.

We evaluate FRCU-Net on five datasets: ISIC 2017 [13], ISIC 2018 [12], PH^2 [28], Lung segmentation [1], and SegPC 2021 [16] challenge datasets. The experimental results demonstrate that the proposed network achieves superior performance compared to state-of-the-art alternatives.

2. Related Work

Research on deep learning has grown rapidly, and deep learning networks are nowadays prominent strategies for segmentation in medical imaging. FCN [26], a pixels-to-pixels network, is one of the first convolutional networks introduced for image classification. To keep the original resolution, all fully connected layers are replaced with convolution and deconvolution layers. Ronneberger et al. extend this idea and proposed U-Net [31] for biomedical image segmentation. U-Net is a fairly symmetrical U-shaped encoder-decoder architecture, in which the encoder and decoder parts are combined with skip connections at different scales to integrate deep and shallow features.

Different extensions of U-Net have been proposed for image segmentation. To process 3D volumes, 3D U-Net was been proposed [11], in which all 2D operations are replaced with their 3D counterparts. By exploiting more skip connections and convolutions, U-Net++ [35] can solve the problem that edge information and small objects are lost due to the down-sampling functions. ResUNet [23] improves the performance of the original U-Net by employing a better CNN backbone with a U-shaped structure that extracts multi-scale information. Some other extensions of U-Net have been proposed by inserting additional modules in different parts of the network. Different studies show that integrating a self-attention mechanism into U-Net by modeling global interactions of all pixels in feature maps results

in better performance. Schlemper et al. propose attention-based U-Net [32] by inserting additive attention gate into the skip-connections. Inspiring by the ideas of squeeze and excitation approaches [22] and dense connections, Asadi et al. proposed MCGU-Net [4] in which the channel-wise attention improves the performance of the original U-Net.

Deng et al. proposed PraNet [14] by adding an RBF module to the skip connection to capture visual informative features at multiple scales. Azad et al. [5] enhance the performance of the U-Net by inserting non-linearity in the skip connections through ConvLSTM for combining the features from encoder and decoder parts rather than a simple concatenation. Martin et al. [27] utilized a stacked version of BCDU-Net for myocardial pathology segmentation. Alom et al. [3] extended U-Net by adding Recurrent Convolutional Neural Network (RCNN) and Recurrent Residual Convolutional Neural Network (R2CNN) in which feature accumulation with recurrent residual convolutional layers ensures better feature representation.

Deeplab [10] utilizes the idea of atrous spatial pyramid pooling (ASPP) at several grid scales. Atrous convolution layers with different rates capture multi-scale information, resulting in better performance on several segmentation benchmarks [21]. By taking into account the advantages of both U-shape networks and pyramid spatial pooling, Chen et al. introduce Deeplabv3+ in which Atrous convolution extracts rich semantic information in the encoding path and controls the density of the decoder features. Azad et al. [6] improve Deeplabv3+ by inserting two attention modules of channel-wise attention and multi-scale attention mechanisms in the Atrous convolution.

It has been shown that CNNs have a strong texture inductive bias which limits their ability to leverage useful shape information [17]. In other words, convolutional networks have a bias towards extracting features based on texture rather than shape. In the context of few-shot learning, a set of Difference of Gaussians (DoG) is inserted into a deep network to attenuate high-frequency local components in the feature space [7]. Lai et al. [24] propose Laplacian Pyramid Super-Resolution Network (LapSRN) to progressively reconstruct the sub-band residuals of high-resolution images for image super-resolution. Their model takes coarse-resolution feature maps as input, and predicts the high-frequency residuals. To enhance the performance of a U-shaped architecture and remove the texture bias of convolutional layers, we propose to utilize the frequency domain of the extracted features to learn shape information along with texture information, reducing the amount of noise on the feature representation.

3. Proposed Method

Inspired by a recent study on texture bias [17] and squeeze and excitation module [22], we present FRCU-Net (Figure 2). We propose a frequency attention mechanism to re-calibrate the representation space within a U-Net based architecture. The proposed module is capable of re-calibrating the representation space by taking into account the informative frequency domains and reconstructing the representation by the nonlinear attention mechanism. To this end, our proposed method incorporates the frequency attention module into the latent space to re-arrange and calibrate the frequency domain for better representation. In the following subsections, we describe each network component in detail.

3.1. Encoder

The contracting path of the U-shaped architecture (encoder) aims at extracting hierarchically semantic features and capture context information. To train the encoder containing a high number of parameters, a large dataset including a large number of labeled data is necessary. The idea of transfer learning allows the network to leverage knowledge from pre-trained models and use it to solve a new problem with fewer data. We utilize Xception as the backbone of the proposed network, and therefore, we can finetune the network by using a set of parameters pre-trained on the PASCAL VOC dataset. The network with the Xception backbone converges fast and achieve accurate results.

Xception structure is a linear stack of depthwise separable convolution layers with residual connections. Channel-wise 3×3 spatial convolution, and 1×1 pointwise convolution are utilized in our FRCU-Net. We define the encoder model E with parameters θ, which takes the input sample $I \in R^{H' \times W' \times C'}$ and generate the encoded feature map $X \in R^{H \times W \times C}$ as,

$$X = E_\theta(I), \tag{1}$$

where H', W', and C' are the dimensions of the input data, H, W, and C are the dimensions of the encoded feature representation, and θ is the set of network parameters.

3.2. Frequency Re-calibration Module

A U-shaped architecture includes a sequence of regular convolutional layers in the bottleneck layer. The convolutional layers have a strong texture inductive bias. In other words, these models tend to perform the recognition task based on the object texture, while recognition in human vision is highly influenced by shape. By utilizing the extracted feature maps from the convolutional layers in the frequency domain, we can take advantage of both shape and texture information for training the network. The importance of texture and shape is different for different applications and data. The low-frequency domain of the extracted feature maps contain shape information of the input data while higher frequencies are responsible for texture information. Instead of focusing on only one of the two general

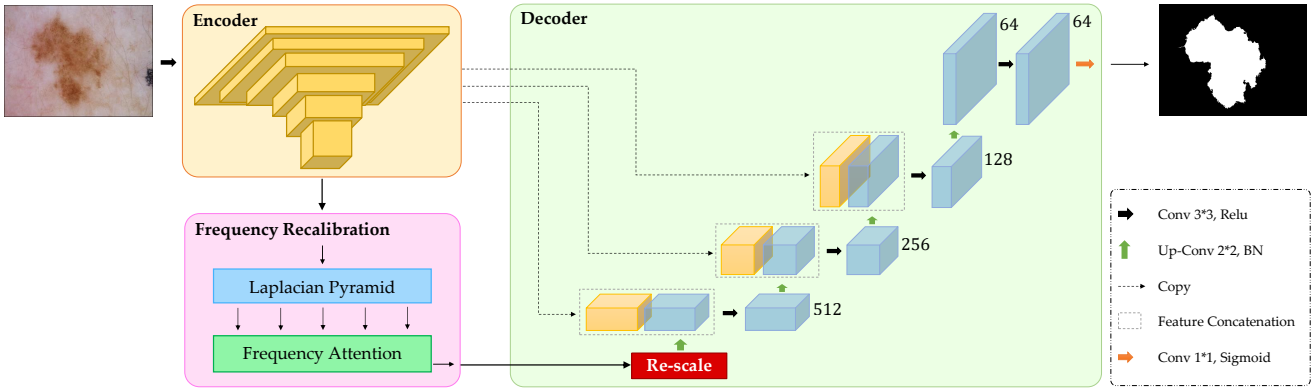

Figure 2. FRCU-Net with 1) Laplacian pyramid to take convolutional features to frequency domain and 2) frequency attention mechanism for a non-linearly weighted combination of all levels of the pramid.

types of information (i.e., shape and texture), we propose a frequency re-calibration (FRC) module which consists of a Laplacian pyramid and frequency attention (Figure 3). We compute the frequency levels of the extracted feature maps through the Laplacian pyramid. Moreover, the frequency attention mechanism allows the network to focus on the more informative frequency level of features. The FRC module is exploited in the bottleneck layer.

3.2.1 Laplacian Pyramid

The extracted feature maps from the convolutional layers are included into the frequency domain through a Laplacian pyramid mechanism. To approximate the Laplacian function we use the same strategy as [7], e.g. Difference of Gaussian (DoG) technique to generate the Laplacian pyramid. First, we extract a $(L+1)$ Gaussian representation from the encoded feature map $X \in R^{H\times W\times C}$ using different values as the variance of the Gaussian function to generate different scales,

$$G_l(X) = X * \frac{1}{\sigma_l\sqrt{2\pi}} e^{-\frac{i^2+j^2}{2\sigma_l^2}}, \tag{2}$$

where σ_l is the variance of the l^{th} Gaussian function, i and j represent the spatial location in the encoded feature space, X is the input set of encoded feature maps which consists of C channels with the size of $H \times W$, and $*$ denotes the convolution operator. To encode frequency information at different scales, we apply a pyramid of DoGs with increasing variance. The l^{th} level of the pyramid is computed as

$$LP_l = \begin{cases} G_l - G_{l+1}, & 1 \le l < L \\ G_L, & l = L \end{cases}, \tag{3}$$

where LP_l is the l^{th} level of the Laplacian pyramid, G_l is the output of the l^{th} Gaussian functions, and L is the number of levels of the pyramid.

3.2.2 Frequency Attention

Different levels of the Laplacian pyramid contain different kinds of information. For instance, the low-frequency level features include shape-based features, while the higher level ones are more related to texture. The importance of these kinds of information differs depending on the data and the task at hand. Inspired by the squeeze and excitation network [22], we propose a frequency attention mechanism to non-linearly aggregate the features of all levels of the frequency domain. In other words, the network employs the global information of each frequency level of the Laplacian pyramid. This idea helps the network to selectively empathize informative frequency levels and suppress less useful ones.

First, for each level of the Laplacian pyramid, we normalize all input channels. To this end, we utilize the global context information of the input features to generate weights for all input channels of each Laplacian pyramid. The global average pooling is calculated as,

$$GAP_l^f = \frac{1}{H\times W}\sum_{i}^{H}\sum_{j}^{W} LP_l^f(i,j), \tag{4}$$

where the LP_l^f is the f^{th} channel of the frequency features of the l^{th} Laplacian pyramid level, $H \times W$ is the size of each channel, and GAP_l^f is the output of the global average pooling function for the f^{th} channel of the l^{th} Laplacian pyramid level. Two fully connected layers (FCL) are then used to capture the channel-wise dependencies of feature maps at each level as

$$w_l^f = \sigma\left(\mathbf{W}_2\,\delta\left(\mathbf{W}_1\,GAP_l^f\right)\right), \tag{5}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are the parameters of the FCLs, δ and σ are the ReLU and Sigmoid activation functions, respectively, and w_l^f is the learnt weight for the f^{th} channel of the l^{th} layer. The final feature map in each channel is computed

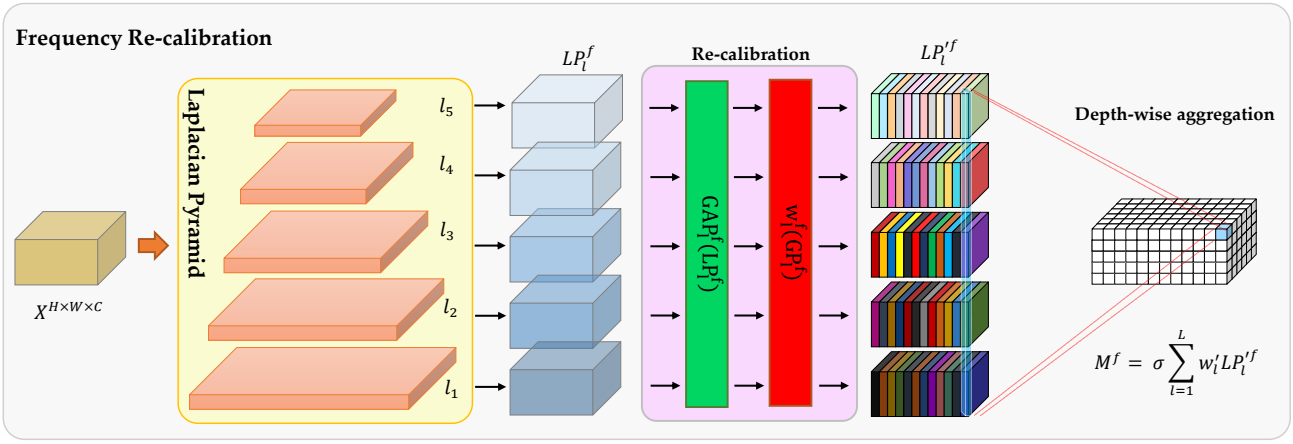

Figure 3. Frequency attention mechanism consists of 1) feature re-calibration to focus more on the informative features and 2) a non-linear depth-wise aggregation to combine features from different levels of the pyramid .

by multiplication of the learnt weight and the input channel feature $\tilde{LP}_l^f = w_l^f . LP_l^f$.

After re-calibrating all of the feature maps in each layer, we aggregate the features of all the pyramid levels taking into account their discriminative power. To do that, a weight is learned for each level and a non-linear depthwise aggregation is utilized to combine these features as,

$$M^f = \sigma\left(\sum_{l=1}^{L} w_l' \, \tilde{LP}_l^f\right), \tag{6}$$

where w_l' is the learnt weight for the l^{th} level, $\tilde{LP}_l^f$ is the f^{th} channel of the feature set from the l^{th} level, and M^f is the f^{th} channel of the output feature map.

3.3. Decoder

The same decoder as in the regular U-Net is utilized in our network. The features from the encoder part are concatenated with the up-sampled features from the previous decoder layer. The concatenated features are then passed to two 3×3 convolutional functions. We utilize the cross entropy energy function to train the network.

4. Experimental Results

We evaluate the proposed network on five datasets: ISIC 2017 [13], ISIC 2018 [12], PH^2 [28], Lung segmentation [1], and SegPC 2021 [16] challenge datasets. For implementation, we use Keras with TensorFlow backend. All experiments were performed on an NVIDIA GTX 1080 GPU with a batch size of 8 without any data augmentation. We use the Adam optimizer with a learning rate equal to 10^{-4} for training and stop the training process of the network when the validation does not change in 10 consecutive epochs[1]. To compare the proposed network with other alternatives, we consider several performance metrics, including accuracy (AC), sensitivity (SE), specificity (SP), F1-Score, and Jaccard similarity (or Jaccard index) (JS). The baseline network has the same structure as FRCU-Net, but without FRC module.

4.1. ISIC 2017 Dataset

The ISIC 2017 dataset [13] is obtained from the 2017 Kaggle competition which consisted of 3 tasks: lesion segmentation, dermoscopic feature detection, and disease classification. The skin lesion segmentation data is considered for evaluation in this paper. This dataset includes 2000 skin lesion (cancer or non-cancer) images as training set with masks for segmentation. We use 1250 samples for training, 150 samples for validation data, and 600 samples as test set. The original size of each sample is 576×767 pixels. The same pre-processing as [3] is used to resize images to 256×256 pixels.

Figure 4 shows some segmentation results of our proposed network. In Table 1 the quantitative results of the proposed network on this dataset are compared with some other related approaches. The FRCU-Net achieves better performance than the baseline network for all the metrics. The results demonstrate that, except for the sensitivity, the proposed network achieves better results than the other approaches.

4.2. ISIC 2018 Dataset

The International Skin Imaging Collaboration (ISIC) published this dataset [12] as a large-scale dataset of dermoscopy images in 2018 which includes 2594 images with their corresponding ground truth annotations (containing

[1]Source code is available on https://github.com/rezazad68/FRCU-Net.

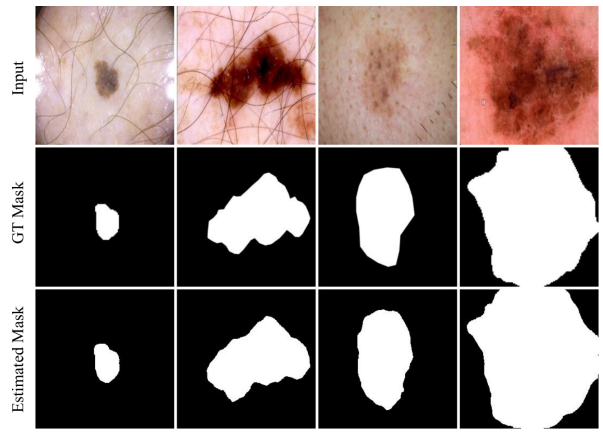

Figure 4. Segmentation result of FRCU-Net on ISIC 2017.

Table 1. Performance comparison on ISIC 2017 dataset (best results are bolded).

Methods	F1	SE	SP	AC	JS
U-net [31]	0.8682	**0.9479**	0.9263	0.9314	0.9314
Melanoma det. [13]	-	-	-	0.9340	-
Lesion Analysis [2]	-	0.8250	0.9750	0.9340	-
R2U-net [3]	0.8920	0.9414	0.9425	0.9424	0.9421
MCGU-Net [4]	0.8927	0.8502	0.9855	0.9570	0.9570
Baseline	0.9036	0.8745	0.9857	0.9647	0.9647
FRCU-Net	**0.9269**	0.9150	**0.9861**	**0.9727**	**0.9727**

cancer or non-cancer lesions). We used 1815 images for training, 259 for validation and 520 for testing, like other approaches [3]. We resize the original size of each sample, i.e., from 2016×3024, to 256×256 pixels.

Figure 5 shows some example outputs of the proposed network. Table 2 lists the quantitative results of different alternative methods and the proposed network on this dataset. It can be seen that better performance is achieved by the proposed network w.r.t. state-of-the-art alternatives for F1-score, sensitivity, accuracy and Jaccard similarities, and for all the metrics, our FRCU-Net outperform the baseline.

Table 2. Performance comparison on ISIC 2018 dataset (best results are bolded.).

Methods	F1	SE	SP	AC	PC	JS
U-net [31]	0.647	0.708	0.964	0.890	0.779	0.549
Att U-net [30]	0.665	0.717	0.967	0.897	0.787	0.566
R2U-net [3]	0.679	0.792	0.928	0.880	0.741	0.581
Att R2U-Net [3]	0.691	0.726	0.971	0.904	0.822	0.592
BCDU-Net [5]	0.851	0.785	0.982	0.937	0.928	0.937
MCGU-Net [4]	0.895	0.848	**0.986**	0.955	**0.947**	0.955
Baseline	0.892	0.871	0.978	0.954	0.914	0.954
FRCU-Net	**0.913**	**0.904**	0.979	**0.963**	0.922	**0.963**

4.3. PH^2 Dataset

The PH^2 dataset [28] is a dermoscopic image database which was introduced for both segmentation and classification. The dataset contains a total number of 200 melanocytic lesions, including 80 common nevi, 80 atypical nevi, and 40 melanomas. The manual segmentation of the skin lesions are available as the ground truth. The resolution of each input image is 768×560 pixels. We follow the experimental setting used in [25], and randomly split the dataset into two sets of 100 images, and then use one set as test data, 80% of the other set for the training, and the remaining data for validation. For this dataset, we exploit the learnt weights of ISIC 2017 as the pre-trained model and then finetune the network with the training data.

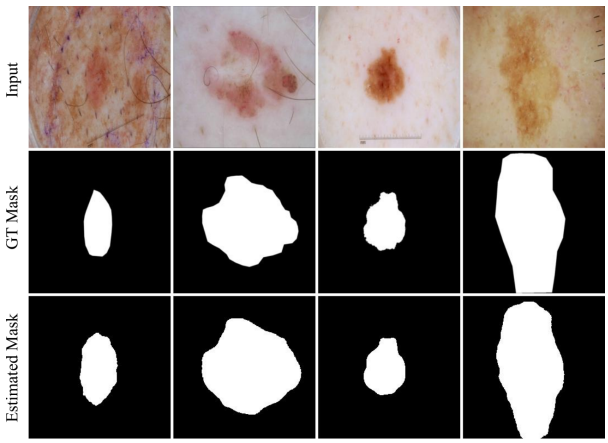

Figure 5. Segmentation result of FRCU-Net on ISIC 2018.

Some segmentation outputs of the proposed network for PH^2 dataset are depicted in Figure 6. The results of the proposed network are compared with other state-of-the art approaches in Table 3. It can be seen that except from the specificity, the proposed approach results in better performance than state-of-the-art alternatives. The performance of the FRCU-Net is also better than the baseline network.

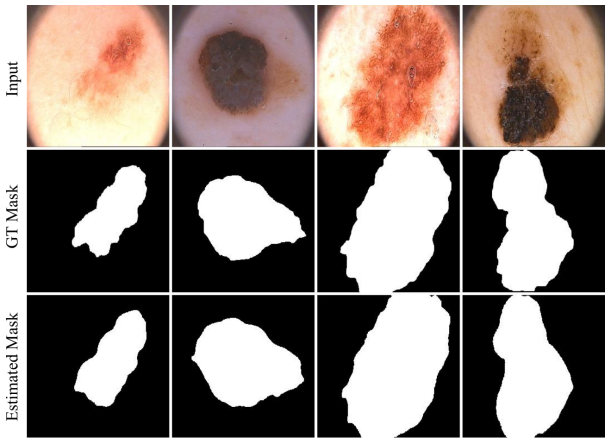

Figure 6. Segmentation result of FRCU-Net on PH^2.

Table 3. Performance comparison on PH^2 dataset (best results are bolded).

Methods	DIC	SE	SP	AC	JS
FCN [29]	0.8903	0.9030	0.9402	0.9282	0.8022
U-net [31]	0.8761	0.8163	0.9776	0.9255	0.7795
SegNet [8]	0.8936	0.8653	0.9661	0.9336	0.8077
FrCN [2]	0.9177	0.9372	0.9565	0.9508	0.8479
MCGU-Net [4]	0.9263	0.8322	0.9714	0.9537	0.9537
Baseline	0.9278	0.9071	**0.9787**	0.9568	0.9568
FRCU-Net	**0.9497**	**0.9730**	0.9689	**0.9689**	**0.9689**

4.4. Lung Segmentation Dataset

The Lung Nodule Analysis (LUNA) competition at the Kaggle Data Science Bowl in 2017 introduced a lung segmentation dataset [1]. This data includes 2D and 3D CT images with labels for lung segmentation . For our evaluation, we use 70% of the data as train set and the remaining 30% as test set. The size of each image is 512×512 pixels. The lung lesions in CT images have almost the same Hausdorff value as other structures that are not of interest, such as bone and air. We use the same strategy as [5] to estimate the lung region as a region inside the estimated surrounding tissues.

Figure 7 shows some outputs of the proposed network. In Table 4, the performance of the FRCU-Net on this dataset is compared with other state-of-the-art approaches. It can be seen that after the MCGU-Net, we have the best F1-Score and accuracy for the FRCU-Net among other approaches listed in this table. The MCGU-Net uses bidirectional ConvLSTM in the skip connection layers and dense connections in the bottleneck layer. Consequently, compared to FRCU-Net, MCGU-Net comprises a larger number of parameters for training, and it therefore needs much longer for convergence.

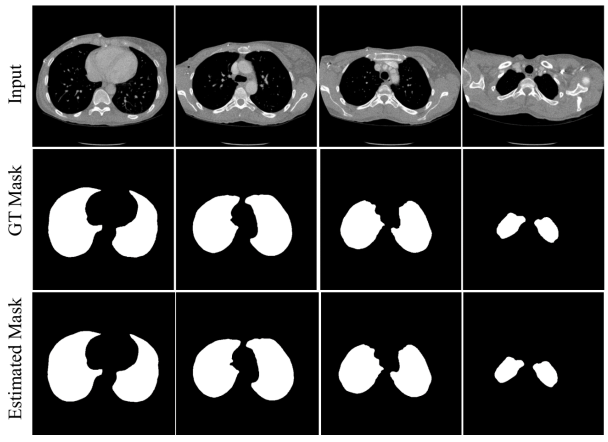

Figure 7. Segmentation result of FRCU-Net on Lung segmentation dataset.

Table 4. Performance comparison on Lung dataset (best results are bolded).

Methods	F1	SE	SP	AC
U-net [31]	0.9658	0.9696	0.9872	0.9872
RU-net [3]	0.9638	0.9734	0.9866	0.9836
R2U-Net [3]	0.9832	**0.9944**	0.9832	0.9918
MCGU-Net[4]	**0.9904**	0.9910	**0.9982**	**0.9972**
Baseline	0.9851	0.9914	0.9962	0.9954
FRCU-Net	0.9901	0.9904	**0.9982**	0.9970

Table 5. Performance comparison on SegPC dataset (best results are bolded).

Methods	mIOU
XLAB Insights [9]	0.9360
DSC-IITISM [9]	0.9356
bmdeep [9]	0.9065
Baseline	0.9215
FRCU-Net	**0.9392**

4.5. SegPC 2021 Challenge dataset

We evaluate our FRCU-Net on multiple myeloma cell segmentation grand challenges which are provided by the SegPC 2021 [16, 18, 19]. Images in this dataset were captured from bone marrow aspirate slides of patients diagnosed with Multiple Myeloma (MM), a type of white blood cancer. This dataset consists of a training set with 290 samples, a validation set with 200, and a test set with 277 samples. Since the test data is not publicly available, we split the training dataset into a training and validation set and evaluate the method on the original validation set as our test set. All the samples have been annotated by a pathologist and instance base segmentation masks are provided for each object of interest (myeloma plasmacells).

Some segmentation outputs of the FRCU-Net are shown in Figure 8. The mIOU of the proposed network is compared with the challenge winners in Table 5. The first-ranked team (XLAB Insights) utilizes a combination of three instance segmentation networks (SCNet [33], ResNeSt[34], and Mask-RCNN [20]) with a slight modification to suit the current task. The second team (DSC-IITISM) employs the Mask-RCNN model with heavy data augmentation approaches. Lastly, bmdeep [9] uses an attention deeplabv3+ method [6] with a multi-scale region-based training strategy. In our pipeline, we also use this [9] strategy and replace the attention deeplabv3+ network with our proposed model. Our experimental results demonstrate that our proposed approach improves the performance compared to all aforementioned approaches.

4.6. Effect of the FRC Module

The main modification of our proposed network compared to the U-Net is utilizing the frequency domain features in the bottleneck layer. In Figure 9, we compare the segmentation output of the proposed network with the base-

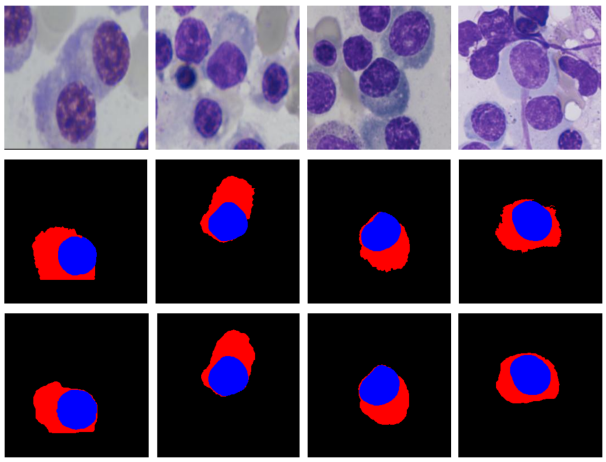

Figure 8. Segmentation result of FRCU-Net on SegPC dataset.

line model (U-Net). It shows a more precise and fine segmentation output of the proposed network by utilizing the frequency domain.

The main task of the FRC module in our proposed network is taking convolutional features from some levels of the frequency domain. Different levels of the Laplacian pyramid are responsible for different kinds of features. For instance, high level frequencies include significant shape information, while lower frequencies contain information about the texture of the input data. It is worth mentioning that we also evaluated the network without the Laplacian pyramid, i.e., with the SE block only. The SE block improves the F1-Score of the base architecture with by 1% for ISIC 2017 dataset while the performance of the FRCU-Net (Laplacian pyramid plus SE block) was about 2.3% higher than our baseline. In other words, both components are clearly responsible for the achieved gain.

The FRC module in our network is employed to combine these kinds of features so that the network learns to attend more on the kind of feature which is the most discriminative one based on each particular benchmark. This can be seen in Figure 9. Compared to U-Net, FRCU-Net results in a more accurate output segmentation, providing an accurate and smooth segmentation boundary that properly defines the shape of the skin lesion. As we can see in the third row of Figure 9, the skin lesion is not as obvious as other examples and there is an overlap between the background and the lesion. Shape-based features are relevant to segment this example. Overall, one can see that the visual performance of FRCU-Net is better than the original U-Net.

5. Conclusion

In this paper, we proposed the FRCU-Net for skin lesion segmentation. It has been shown that the regular convolutional layers tend to learn texture-based features, while in

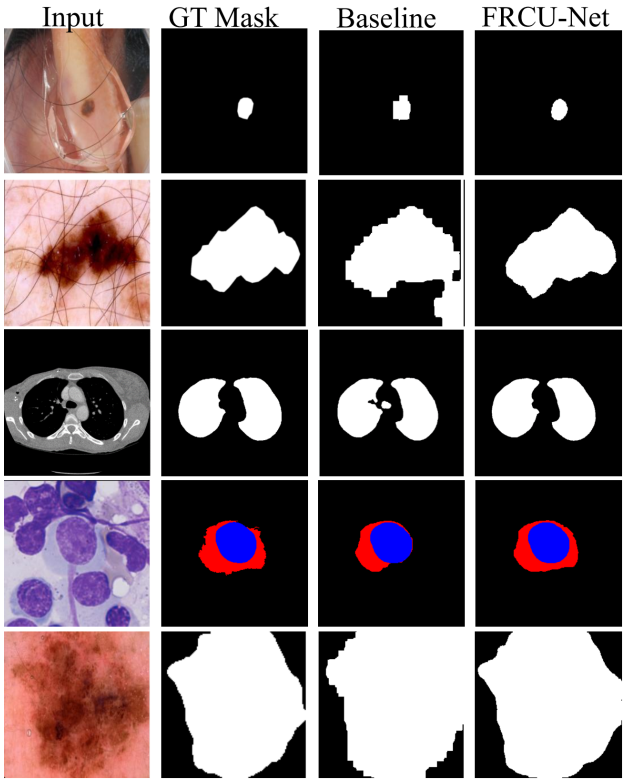

Figure 9. Visual effect of the FRC module in the FRCU-Net. From top to bottom, examples from ISIC 2018, ISIC 2017, Lung segmentation, SegPC, and PH^2 datasets.

many applications shape-based features can provide highly discriminative information. In order to consider both of these kinds of features, we proposed to extend the classical U-Net by inserting the FRC module, which comprises two parts: Laplacian pyramid and frequency attention. We represent the extracted feature maps of the convolutional layers in the frequency domain to capture both texture-based and shape-based information. To aggregate the features from all levels of the Laplacian pyramid, we proposed a frequency attention mechanism. The channels of each set of feature maps are first re-calibrated by employing the global average pooling information. The features from different levels of the pyramid were then combined by utilizing a non-linear aggregation function. Our evaluation on five public medical image segmentation datasets demonstrated that the proposed FRCU-Net outperforms state-of-the-art alternatives.

6. Acknowledgment

This work has been partially supported by the Spanish project PID2019-105093GB-I00 (MINECO/FEDER, UE) and by ICREA under the ICREA Academia programme.

References

[1] https://www.kaggle.com/kmader/finding-lungs-in-ct-data.

[2] Mohammed A Al-Masni, Mugahed A Al-Antari, Mun-Taek Choi, Seung-Moo Han, and Tae-Seong Kim. Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks. *Computer methods and programs in biomedicine*, 162:221–231, 2018.

[3] Md Zahangir Alom, Mahmudul Hasan, Chris Yakopcic, Tarek M Taha, and Vijayan K Asari. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. *arXiv preprint arXiv:1802.06955*, 2018.

[4] Maryam Asadi-Aghbolaghi, Reza Azad, Mahmood Fathy, and Sergio Escalera. Multi-level context gating of embedded collective knowledge for medical image segmentation. *arXiv preprint arXiv:2003.05056*, 2020.

[5] Reza Azad, Maryam Asadi-Aghbolaghi, Mahmood Fathy, and Sergio Escalera. Bi-directional convlstm u-net with densley connected convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[6] Reza Azad, Maryam Asadi-Aghbolaghi, Mahmood Fathy, and Sergio Escalera. Attention deeplabv3+: Multi-level context attention mechanism for skin lesion segmentation. In *European Conference on Computer Vision*, pages 251–266. Springer, 2020.

[7] Reza Azad, Abdur R Fayjie, Claude Kauffmann, Ismail Ben Ayed, Marco Pedersoli, and Jose Dolz. On the texture bias for few-shot cnn segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2674–2683, 2021.

[8] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[9] Afshin Bozorgpour, Reza Azad, Eman Showkatian, and Alaa Sulaiman. Multi-scale regional attention deeplab3+: Multiple myeloma plasma cells segmentation in microscopic images. *arXiv preprint arXiv:2105.06238*, 2021.

[10] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[11] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

[12] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

[13] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

[14] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 263–273. Springer, 2020.

[15] Abdur R Feyjie, Reza Azad, Marco Pedersoli, Claude Kauffman, Ismail Ben Ayed, and Jose Dolz. Semi-supervised few-shot learning for medical image segmentation. *arXiv preprint arXiv:2003.08462*, 2020.

[16] Shiv Gehlot, Anubha Gupta, and Ritu Gupta. Ednfc-net: Convolutional neural network with nested feature concatenation for nuclei-instance segmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1389–1393. IEEE, 2020.

[17] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

[18] Anubha Gupta, Rahul Duggal, Shiv Gehlot, Ritu Gupta, Anvit Mangal, Lalit Kumar, Nisarg Thakkar, and Devprakash Satpathy. Gcti-sn: Geometry-inspired chemical and tissue invariant stain normalization of microscopic medical images. *Medical Image Analysis*, 65:101788, 2020.

[19] Anubha Gupta, Pramit Mallick, Ojaswa Sharma, Ritu Gupta, and Rahul Duggal. Pcseg: Color model driven probabilistic multiphase level set based tool for plasma cell segmentation in multiple myeloma. *PloS one*, 13(12):e0207908, 2018.

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[21] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul J Kennedy. Atrous convolution for binary semantic segmentation of lung nodule. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1015–1019. IEEE, 2019.

[22] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[23] Debesh Jha, Sharib Ali, Håvard D Johansen, Dag D Johansen, Jens Rittscher, Michael A Riegler, and Pål Halvorsen. Real-time polyp detection, localisation and segmentation in colonoscopy using deep learning. *arXiv preprint arXiv:2011.07631*, 2020.

[24] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and

accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017.

[25] Xinhua Liu, Gaoqiang Hu, Xiaolin Ma, and Hailan Kuang. An enhanced neural network based on deep metric learning for skin lesion segmentation. In *2019 Chinese Control And Decision Conference (CCDC)*, pages 1633–1638. IEEE, 2019.

[26] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[27] Carlos Martín-Isla, Maryam Asadi-Aghbolaghi, Polyxeni Gkontra, Victor M Campello, Sergio Escalera, and Karim Lekadir. Stacked bcdu-net with semantic cmr synthesis: Application to myocardial pathology segmentation challenge. In *Myocardial Pathology Segmentation Combining Multi-Sequence CMR Challenge*, pages 1–16. Springer, 2020.

[28] Teresa Mendonça, Pedro M Ferreira, Jorge S Marques, André RS Marcal, and Jorge Rozeira. Ph 2-a dermoscopic image database for research and benchmarking. In *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 5437–5440. IEEE, 2013.

[29] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015.

[30] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[32] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention gated networks: Learning to leverage salient regions in medical images. *Medical image analysis*, 53:197–207, 2019.

[33] Thang Vu, Haeyong Kang, and Chang D Yoo. Scnet: Training inference sample consistency for instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2701–2709, 2021.

[34] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020.

[35] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer, 2018.