2023 The 23rd International Conference on Control, Automation and Systems (ICCAS 2023)
Yeosu Venezia Hotel & Utop Marina Hotel, Yeosu, Korea, Oct. 17∼20, 2023

RDIU-Net: Lightweight Medical Image Segmentation Network

Juon Kurosawa, Armagan Elibol, and Nak Young Chong*

School of Information Science,
Japan Advanced Institute of Science and Technology
Ishikawa, 923-1292, Japan
{s2110071,aelibol,nakyoung}@jaist.ac.jp * Corresponding author

Abstract: In recent years, medical image segmentation using deep learning methods has become more and more popular and developed with the aim of both reducing human-related errors and the time required for manual segmentation. One of the pioneers in deep learning-based biological image segmentation networks, U-Net was proposed back in 2015. Since then, several models have been proposed to extend U-Net. However, the trade-off between computational complexity and accuracy remains a major challenge. To address this trade-off, we use a new Involution kernel for spatial information and propose a model lightweight medical image segmentation network, Residual Involution U-Net (RDIU-Net). Involution, Residual, and Dense structures are incorporated into the U-Net model to extract both channel and spatial features. Evaluations have been carried out on three different datasets of ultrasound, X-ray, and dermoscopic images. The proposed model RDIU-Net showed superior results in accuracy, processing speed, training stability, and convergence compared to U-Net.

Keywords: Medical Imaging, Image Segmentation, Deep Learning.

1. INTRODUCTION

Deep learning methods are now commonly used in medical image analysis and Computer-Aided Diagnosis (CAD) systems. Expected benefits include a reduction in the risk of human errors, a reduction in examination time, and an improvement in the accuracy of image reading. A wide variety of images in the medical domain have been studied, including MR images, CT images, ultrasound images, and X-ray images, among which organ segmentation and lesion detection and classification using deep learning have appeared in many studies [1]. A fully convolutional network (FCN) was proposed in [2] as a deep network for image segmentation used in image analysis, which maintains the spatial information in the source image and performs end-to-end processing between pixels. U-Net, the most popular medical image segmentation network, was proposed in [3] as an extension of FCN, combining encoders and decoders with skip connections to enable highly accurate segmentation without using large amounts of data. In addition, many models extend U-Net, but convolution redundancy and increased computational complexity are still issues for these networks. The architecture of U-Net is shown in Fig. 1.

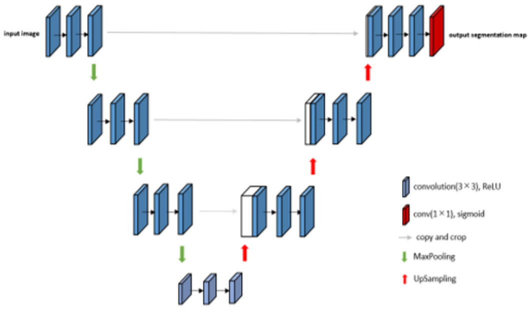

Fig. 1. U-Net Structure

This paper proposes a faster and more accurate medical image segmentation network (RDIU-Net: Residual Dense Involution U-Net). In particular, RDIU-Net propagates channel and spatial information by focusing on the Involution and Residual structure and the Involution and Dense structure, respectively. Within the Residual structure, 1 × 1 Involution and 1 × 1 Convolution reduce the spatial and channel dimensions. By separating and exchanging spatial and channel information, both features can be efficiently extracted, and a variety of feature maps can be learned. Moreover, by processing Involution and Convolution as a Dense structure, it is possible to propagate both channel and spatial information while maintaining their feature values. We present experimental and comparative results with RDIU-Net and U-Net using ultrasound nerve segmentation images [4], chest X-ray lung region segmentation images [5], and skin lesion segmentation images [6]. RDIU-Net reduced model weight significantly and showed drastic advantages in model stability and convergence.

2. RELATED WORK

2.1 Fully Convolutional Network (FCN)

FCN is an end-to-end, pixel-to-pixel learning method used in semantic segmentation tasks. The initial fully connected layer is eliminated and replaced by a new convolution layer, and all layers use the convolution layer to support the segmentation task. To restore the original image size in the Decoder, the feature map obtained in the last Convolution layer is restored to the original resolu-

This work was supported by JSPS KAKENHI Grant Number JP23K03756

tion by Up Sampling, and prediction is performed.

2.2 Involution

Convolution has the property of being independent of spatial information and specific to channel information, and CNN-based models have so far achieved remarkable improvements in the field of image recognition. However, its ability to adapt to several features at different spatial locations is limited. Convolution has difficulty in capturing long-range interactions of spatial information and has problems with the size of the receptive field. And the more complex the model, the greater the redundancy between channels and the greater the computational complexity/parameters. To overcome these drawbacks, Li et al. [7] proposed Involution, an inversion of the concept of Convolution, which is a kernel (filtering) process to reduce the redundancy and computational complexity of CNN-based problems. Unlike Convolution, which randomly generates kernels, the Involution kernel refers to the channel of each pixel and obtains the channel information when the kernel generates it. In other words, the Convolution kernel shares the same kernel weights for each pixel, while the Involution kernel differs in that the kernel weights for each pixel are different for each pixel. This allows the Involution kernel to have a larger receptive field than the Convolution kernel and to capture different spatial information in the image. The output feature map of Involution is the output $\mathbf{Y_{i,j,k}}$ of a multiply-add operation on the input using the Involution kernel.

$$\mathbf{Y_{i,j,k}} = \sum_{\mathbf{(u,v)\in\Delta K}} \mathbf{H}_{\mathbf{i,j,u+[\frac{K}{2}],v+[\frac{K}{2}],[\frac{kG}{C}]}} \mathbf{X_{i+u,j+v,k}} \tag{1}$$

The channel is divided into G, and the kernel is shared between the divided channels. Convolution is based on the same kernel shared by all pixels, with all weights determined randomly. Therefore, the receptive field depends on the kernel size. Conversely, with Involution, the weights are determined for each pixel from in the equation, so each pixel has a different kernel weight and is independent of the channel. This makes it possible to capture different spatial information without limiting the receptive field. Fig. 2 depicts a graphical description of the Involution.

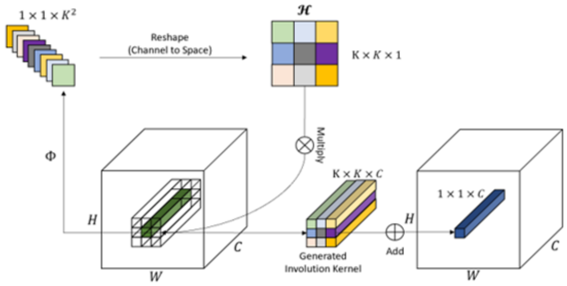

Fig. 2. A Graphical illustration of Involution

2.3 Residual Structure

Residual Structure is a model originally proposed in [8] to solve the gradient explosion/disappearance problem in deep networks in the field of image recognition. Until now, as the number of layers deepened, there has been a degradation problem where the accuracy drops at a certain point. However, ResNet succeeded in improving accuracy by enabling learning in deep networks with as many as 152 layers. ResNet's relatively simple structure and parallelized learning of residuals prevent the gradient problem from occurring and enable learning at deeper levels. The input $\mathbf{x}$ is processed in parallel by the identity map $\mathbf{x}$ and the residual map $\mathbf{F(x)}$. The residuals are added. The output is $\mathbf{H(x)}$, where $\mathbf{H(x)} = \mathbf{F(x)} + \mathbf{x}$ as denoted in Fig. 3.

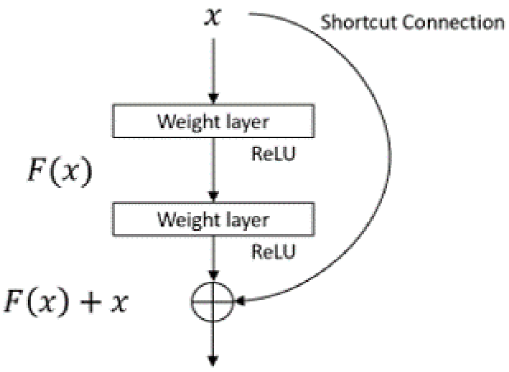

Fig. 3. Residual Structure

2.4 Dense Structure

A dense structure concept was proposed in [9] to solve the gradient explosion/loss problem in deep networks and to reuse feature maps. The model contributes to accuracy by solving the gradient problem, reducing parameters, and propagating complex feature maps by combining the previous feature map, the next feature map, and the 2nd next feature map by skip connections, respectively (see Fig. 4). The output of the l layer is $\mathbf{x_l}$, and the Dense structure is calculated by $\mathbf{H_l}([\mathbf{x_0}, \mathbf{x_1}, \cdots, \mathbf{x_{l-1}}])$ and $\mathbf{H_l}$ is composed of batch normalization, ReLU, and 3×3 Convolution operations.

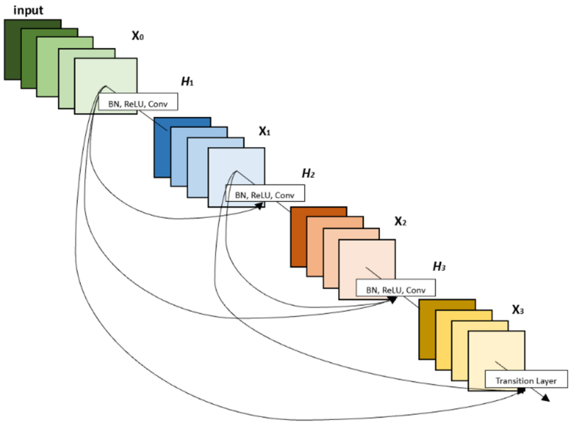

Fig. 4. Dense Structure

3. PROPOSED MODEL:RDIU-NET

Our proposed model, Residual Dense Involution U-Net (RDIU-Net) is based on the U-Net structure and makes use of Involution Kernel to become a lightweight, Dense Structure for maintaining spatial information, and Residual Structure for gradient problems and consistent learning stability. Its graphical representation is given in Fig. 5. The first layer of the encoder is a Convolutional (3x3) and Residual Involution Block (RIB), each layer incorporating a nonlinear activation function (ReLU) and batch normalization. Fig. 6 shows the RIB structure. The RIB performs residual learning of the input and feature maps propagated by Involution (1×1)/Convolution (1×1)/Dropout. The RIB has a relatively simple structure, connecting spatial and channel features in the Involution and Convolution kernels with residual connections in the Residual structure. The spatial and channel dimensions are reduced by the Involution (1×1) and Convolution (1×1) in the RIB, and the spatial and channel information are separated and exchanged to efficiently extract both features and learn various feature maps. This makes it possible to learn a wide variety of feature maps. At the bottom of the Encoder, Convolution, and RIBs are propagated as a Dense structure (Residual Dense Involution Block). This process sets the output propagated by the 3 × 3 Convolution and RIB as a single Dense Block, prepares multiple such blocks, and propagates them as a Dense structure between the blocks. This allows propagation while maintaining the feature values of both channel and spatial information. Let $\mathbf{Y_n}$ be the output of n^{th} dense block (Convolution and RIB). The input of n^{th} $(n \in 1, 2, \cdots, N)$ dense blocks are concatenated with the feature map of the previous block and propagated. The output feature map is the same for Convolution and Involution. The input and output of the RIB can be expressed via the following equations 2 and 3

$$[\mathbf{Y_1}, \mathbf{Y_2}, \cdots, \mathbf{Y_{n-1}}] \in \mathbb{R}^{\mathbf{i-1 \times H \times W \times K \times K \times C}} \quad (2)$$

$$\mathbf{Y_n} \in \mathbb{R}^{\mathbf{H \times W \times K \times K \times C}} \quad (3)$$

The proposed model combines the Involution, Residual, and Dense structures to significantly reduce the weight of the model compared to existing models. In particular, the use of Involution, which has a large receptive field, improves accuracy due to the fact that different features can be extracted.

4. EXPERIMENTAL RESULTS

We perform semantic segmentation on medical imaging datasets using the original U-Net and the proposed model RDIU-Net. We present experimental results on accuracy and processing speed, and model stability and convergence. Three datasets were used: ultrasound nerve segmentation images, chest X-ray lung region segmentation images, and skin lesion segmentation images. The data sets and input image sizes are 2,326 (128×128×1), 247 (256×256×1), and 2,594 (128×128×1) images, respectively. For the evaluation, the Dice coefficient and IoU are used, which are calculated with the use of a confusion matrix. Also, we calculate the processing speed per image (ms/per image). We applied some low-level image processing methods as pre-processing (e.g., normalization). For all datasets, we used Adam Optimization with the learning rate of $1e-4$, and the loss function used was the binary cross-entropy loss. Models were trained on a GPU A100-SXM-40GB. Experimental values for accuracy and processing speed are given in Table 1. It can be seen that the proposed model not only improved the accuracy but also reduced the computational cost significantly over all tested datasets. In particular, the accuracy of both the Dice coefficient and IoU were improved more notably in the chest X-ray lung region segmentation image and skin lesion segmentation datasets. As for processing speed, a 20% to 30% speedup was achieved. The total number of parameters was also reduced by 62%, from 15.6M for U-Net to 5.9M for RDIU-Net. In Table 2, we report the stability and convergence. We define stability as the number of times the dice coefficient decreased by 0.05 from the previous epoch while convergence is the number of the first epoch when a dice score reaches over a certain threshold (denoted as t in the column headers accordingly). Figs. 7 and 8 show the predicted label images and Dice coefficient graphs for the chest x-ray lung region segmentation dataset. Table 2 and Fig. 8 show that RDIU-Net has better model stability and convergence than U-Net.

5. CONCLUSIONS AND FUTURE WORK

In this research, we proposed Residual Involution Dense U-Net for medical image segmentation tasks. We used three datasets and found that RDIU-Net is more robust to medical images than the existing model, U-Net, with improved accuracy, lighter model weight, significantly improved training stability, and convergence speed. This is achieved thanks to the contributions made by Involution, Residual, and Dense structures. The Residual Involution Block, which is designed within Residua for Involution and Convolution, decouples the interaction of both spatial and channel information, allowing for efficient processing while maintaining accuracy. Specifically, this study makes use of 1 × 1 Involution and 1 × 1 Convolution to exchange and separate spatial and channel information. This allows for generating feature maps with both spatial and channel features. In addition, by incorporating the Dense structure in the bottom row, complex feature maps can be propagated to maintain accuracy. In future work, We plan to make more comparative evaluations with BCDU-Net [10] and RU-Net [11].

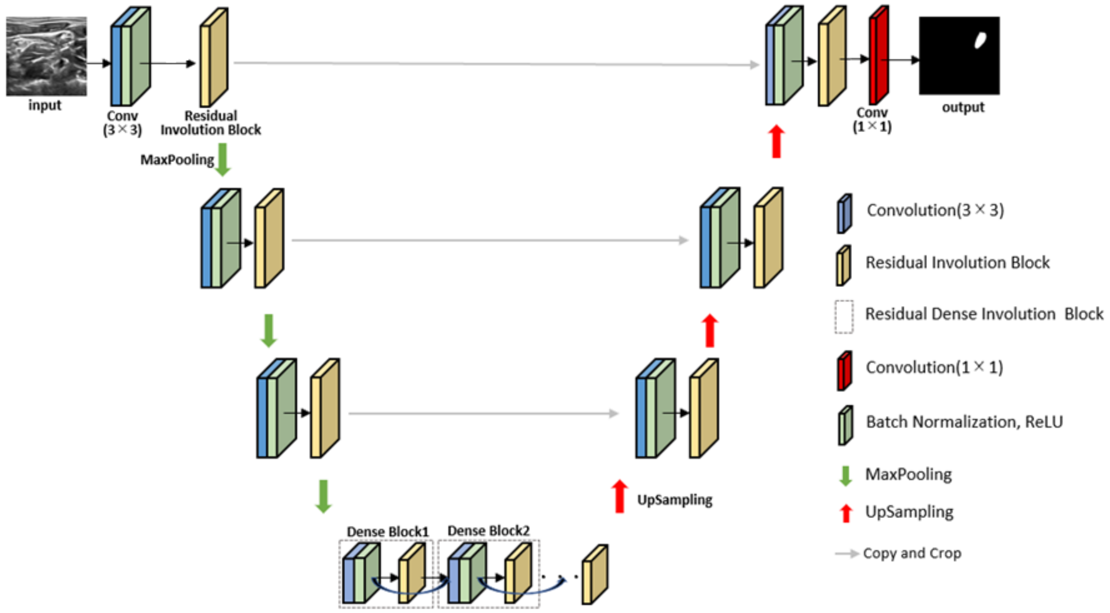

Fig. 5. A schematic representation of the proposed RDIU-Net model

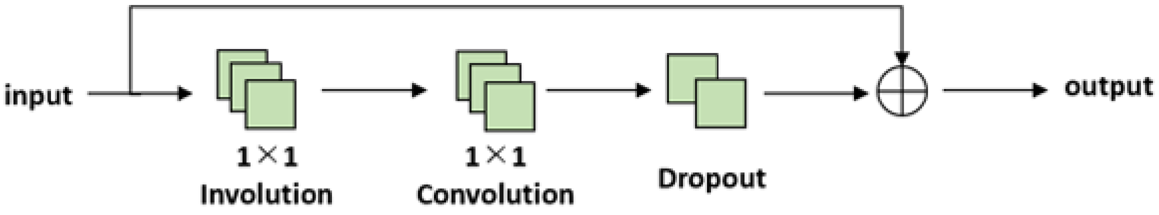

Fig. 6. Residual Involution Block (RIB)

Table 1. Experimental Results: Dice, IoU, and Processing Speed

Dataset	Ultrasound Nerve Segmentation			Chest X-ray Lung Segmentation			Skin Lesion Segmentation			Total Number of
Value / Model	Dice	IoU	Processing Speed (ms/per image)	Dice	IoU	Processing Speed (ms/per image)	Dice	IoU	Processing Speed (ms/per image)	Parameters (in M)
U-Net	0.702	0.571	32.8	0.958	0.929	106.8	0.777	0.675	31.0	15.6
RDIU-Net	0.712	0.574	22.8	0.972	0.935	83.2	0.801	0.694	22.8	5.9

Table 2. Experimental results on model stability and convergence

Dataset	Ultrasound Nerve Segmentation		Chest X-ray Lung Segmentation		Skin Lesion Segmentation	
Value / Model	Stability	Convergence (t=0.7)	Stability	Convergence (t=0.95)	Stability	Convergence (t=0.75)
U-Net	5	36	3	153	15	36
RDIU-Net	0	17	0	34	4	18

REFERENCES

[1] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[2] 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.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[4] Anna Montoya, Hasnin, kaggle446, shirzad, Will Cukierski, and yffud. Ultrasound nerve segmentation, 2016. Available online at: `https://kaggle.com/competitions/ultrasound-nerve-segmentation`, last accessed on 03.22.2023.

[5] Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-

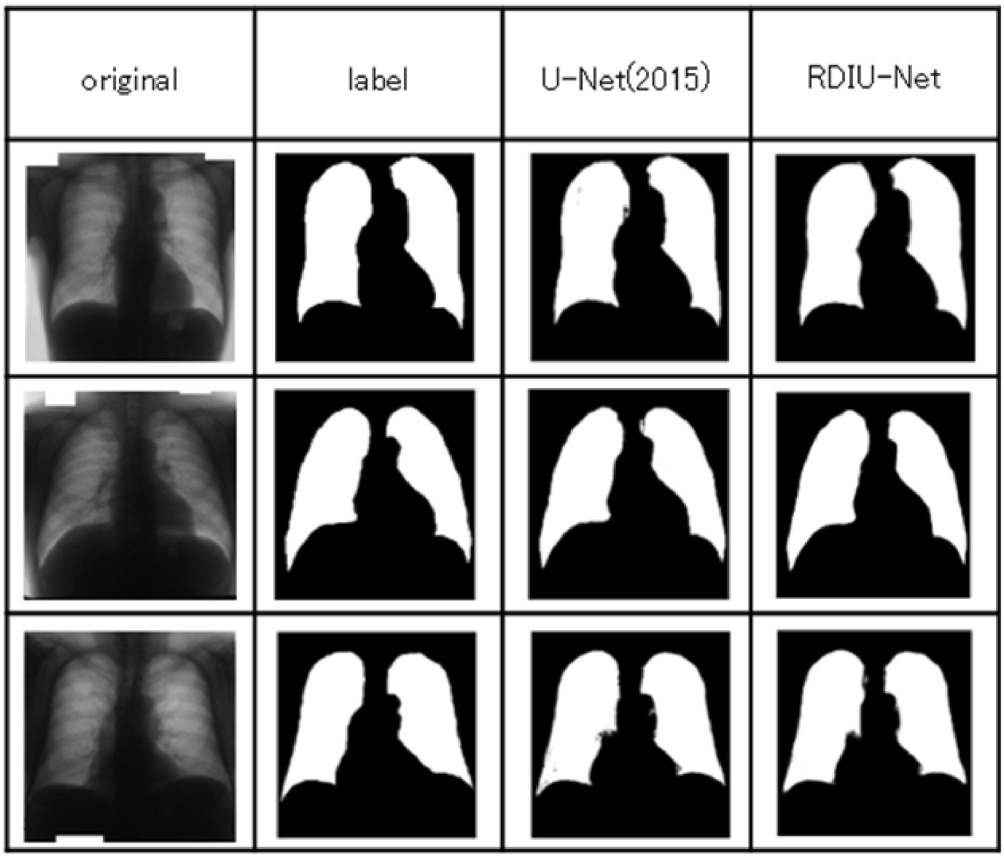

Fig. 7. An example of segmentation results for chest X-Ray lung image dataset

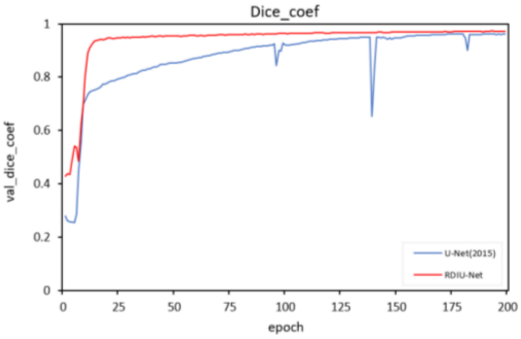

Fig. 8. Dice coefficient graph of chest X-ray lung region segmentation dataset. RDIU-Net is illustrated with a red line while U-Net is with a blue line.

ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. Development of a digital image database for chest radiographs with and without a lung nodule. *American Journal of Roentgenology*, 174(1):71–74, 2000.

[6] 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.

[7] Duo Li, Jie Hu, Changhu Wang, Xiangtai Li, Qi She, Lei Zhu, Tong Zhang, and Qifeng Chen. Involution: Inverting the inherence of convolution for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12321–12330, 2021.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[10] 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.

[11] Yu Lyu, Wei-Liang Huo, and Xiao-Lin Tian. Ru-net for heart segmentation from cxr. *Journal of Physics: Conference Series*, 1769(1):012015, jan 2021.