# High- and Low-Level Feature Enhancement for Medical Image Segmentation

Huan Wang, Guotai Wang(✉), Zhihan Xu, Wenhui Lei,
and Shaoting Zhang

University of Electronic Science and Technology of China, Chengdu, China
guotai.wang@uestc.edu.cn

**Abstract.** The fully convolutional networks (FCNs) have achieved state-of-the-art performance in numerous medical image segmentation tasks. Most FCNs typically focus on fusing features in different levels to improve the learning ability to multi-scale features. In this paper, we explore an alternative direction to improve network performance by enhancing the encoding quality of high- and low-level features, so as to introduce two feature enhancement modules: (i) high-level feature enhancement module (HFE); (ii) low-level feature enhancement module (LFE). HFE utilizes attention mechanism to selectively aggregate the optimal feature information in high- and low-levels, enhancing the ability of high-level features to reconstruct accurate details. LFE aims to use global semantic information of high-level features to adaptively guide feature learning of bottom networks, so as to enhance the semantic consistency of high- and low-level features. We integrate HFE and LFE into a typical encoder-decoder network, and propose a novel medical image segmentation framework (HLF-Net). On two challenging datasets of skin lesion segmentation and spleen segmentation, we prove that the proposed modules and network can improve the performance considerably.

**Keywords:** Medical image · Segmentation · Feature enhancement · Convolutional networks

## 1 Introduction

Accurate and reliable segmentation of various anatomies from medical images is essential to improve diagnosis and assessment of related diseases. However, it is rather time-consuming to label a large amount of medical images manually. Thus, with the development of fully convolutional network (FCNs) [1], they have achieved state-of-the-art performance for many medical image segmentation tasks [2–4].

Most FCNs have a typical encoder-decoder framework [4]. The high-level semantic information of input images is embedded into the feature maps, and then the decoder uses multiple up-sampling components to restore the original resolution and generate segmentation results. However, when encoding semantic features of images, it is difficult for the encoder to effectively capture global context features of targets because of small local receptive field of bottom networks. Additionally, although the top-level feature maps from encoder may be highly semantic, the ability of decoder to

H.-I. Suk et al. (Eds.): MLMI 2019, LNCS 11861, pp. 611–619, 2019.
https://doi.org/10.1007/978-3-030-32692-0_70

reconstruct accurate details is severely limited to the low feature resolution. Therefore, much work has recently attempted to fuse low-level but high-resolution features from the bottom layers with high-level but low-resolution features from the top layers, which makes decoder generate more accurate segmentation results. Ronneberger et al. [2] proposed Unet which is one of the most representative frameworks of this idea, providing state-of-the-art performance for medical image segmentation tasks.

Although Unet has achieved great success, it also exists some problems. There is a huge gap of semantic level and spatial resolution between high- and low-level features, and the low-level features have complex background noise [5]. Therefore, it is inefficient to integrate the low-level features into high-level features by simple skip connection as used in [2]. Inspired by [6, 7], we introduce high-level features enhancement block (HFE) to optimize the encoding quality of high-level features. HFE adaptively aggregates the optimal feature information in different levels by utilizing the complementary feature information of high- and low-levels. The attention mechanism used in HFE can recalibrate the spatial and channel features of feature maps respectively, and suppress noise from the bottom layers, so as to improve encoding quality of target-related features.

In addition, we believe that not only the detail reconstruction of high-level features requires the high-resolution information from bottom layers, but also the feature encoding of bottom layers requires the guidance of high-level semantic information. With this idea, we construct a semantic embedding module (LFE). LTE adaptively guides the bottom layers to learn the features of effective regions by the global information perception abilities of high-level features, and enhances the semantic consistency of the high- and low-level features. As far as we know, this is the first time to introduce high-level semantic information into bottom layers in the field of medical image segmentation, and to guide the feature learning of the bottom-layer network through global information.

We integrate the proposed feature enhancement modules (HFE & LFE) into a typical encoder-decoder network for medical image segmentation to demonstrate that these two modules are a generic network component to boost performance, so as to propose a novel medical image segmentation framework (HLE-Net). We evaluated our HLE-Net and proposed modules on two challenging datasets of skin lesion segmentation and spleen segmentation. The results show that the proposed methods can achieve competitive performance, and improve segmentation performance considerably.

## 2 Method

We first define a set of convolutional transformations $F_{tr} : X \rightarrow X', X \in R^{H \times W \times C}$, $X' \in R^{H' \times W' \times C'}$, here $H$ and $W$ are spatial height and width, with $C$ and $C'$ being the input and output channels, respectively. The convolution transforms $F_{tr}(\cdot)$ fuse the spatial and channel information in the input feature maps $X$ in the local receptive field, thereby outputting a richer feature representation $X'$. By stacking the convolutional layers and the nonlinear activation function layers, the feature $X'$ will be encoded into higher-level semantic information $U$. In the FCNs, researchers directly fuse low-level features $X$ and high-level features $U$ into ($X + U$), or directly concatenate $(F_{tr}(X, U))$

by skip connections to obtain high-resolution information from the bottom layers. Although good results have been achieved, there are few studies on further optimizing the encoding quality of high- and low-level features in FCNs. In this study, we focus on using complementary information between high- and low-level features in FCN to enhance the encoding quality of high- and low-level features respectively, and achieve accurate and robust segmentation. We first embed high-level semantic information in the encoder of FCN, then embed the high-resolution features from bottom layers in the decoder, and use the idea of attention mechanism to enhance the efficiency of high- and low-level feature fusion in an adaptive learning way. We will detail the feature enhancement modules (HFE & LFE) proposed in this paper and the corresponding segmentation framework (HLE-Net) in the following parts.

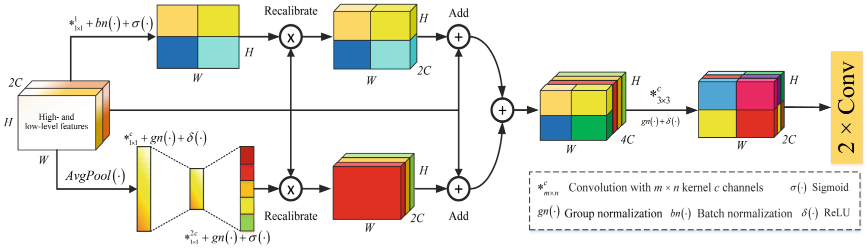
$*^{1}_{1\times1}+bn(\cdot)+\sigma(\cdot)$
Recalibrate
Add
H
W
2C
High- and low-level features
$*^{c}_{1\times1}+gn(\cdot)+\delta(\cdot)$
$AvgPool(\cdot)$
$*^{2c}_{1\times1}+gn(\cdot)+\sigma(\cdot)$
Recalibrate
Add
$*^{c}_{3\times3}$
$gn(\cdot)+\delta(\cdot)$
4C
2 × Conv
$*^{c}_{m\times n}$ Convolution with $m\times n$ kernel $c$ channels
$\sigma(\cdot)$ Sigmoid
$gn(\cdot)$ Group normalization
$bn(\cdot)$ Batch normalization
$\delta(\cdot)$ ReLU

**Fig. 1.** The framework of HFE with spatial and channel attentions.

## 2.1 High-Level Feature Enhancement Module (HFE)

We introduce a high-level feature enhancement module (HFE), which adaptively learns the feature information related to a task through complementary semantic information in high- and low-level features. In addition, HFE emphasizes that different feature channels or different spatial regions in feature maps of different semantic level have different help for tasks. By enhancing relative features and suppressing irrelative features, the encoding quality of high-level features can be greatly improved.

HFE consists of a channel recalibration and a spatial recalibration step. Firstly, we consider the adaptive channel recalibration. Assume that the high- and low-level features $Y=(X,U)=[y_1,y_2,\cdots,y_c,\cdots,y_{2c}]$ are a combination of channels $y_i\in\mathbb{R}^{H\times W}$. We use a global average pooling to compress $y_i$ into a channel descriptor, and generate a channel-wise statistics vector $z\in\mathbb{R}^{1\times1\times2c}$. The $t$-th element of $z$ is calculated by $z_t=Avgpool(y_t)$. $z$ is processed by a block of two 1 × 1 convolution layers that are followed by ReLU and Sigmoid respectively. The output of the Sigmoid is the channel-wise attention coefficient $\tilde{z}=\sigma(z')$. $\tilde{z}$ is used to recalibrate $Y$ to $\tilde{Y}_c=\tilde{z}\otimes Y$, where $\otimes$ denotes element-wise multiplication. Then, we consider spatial adaptive recalibration. By a feature transformation, we use a 1 × 1 convolution to compress $Y$ into a single channel feature map $s$, which is followed by Sigmoid to obtain pixel-wise attention coefficient $\tilde{s}=\sigma(s)$. Finally, the feature $\left(\tilde{Y}_s=\tilde{s}\otimes Y\right)$ of spatial recalibration is obtained by element-wise multiplication.

In HFE, the recalibration of spatial and channel features fully considers the guidance of different levels (high- and low-) of semantic information. By stacking HFE and up-sampling components, FCN can gradually refine and reconstruct high-resolution target details and generate accurate segmentation results. However, since the values of the attention coefficient are in the range of 0 to 1, repeated superposition of the HFE will result in a decrease in the value of the deep feature response, thereby affecting the segmentation performance. Here, we use residual connection [8] to improve the feasibility of optimization based on the preservation of original information. Therefore, the output features of channel recalibration and spatial recalibration in HFE is $\tilde{Y}_c = (1 + \tilde{z}) \otimes Y, \tilde{Y}_s = (1 + \tilde{s}) \otimes Y$. Finally, $\tilde{Y}_c$ and $\tilde{Y}_s$ are concatenated and sent to a $3 \times 3$ convolution layer to fuse their respective feature information, and the channel dimension is reduced. The framework of HFE is illustrated in Fig. 1.

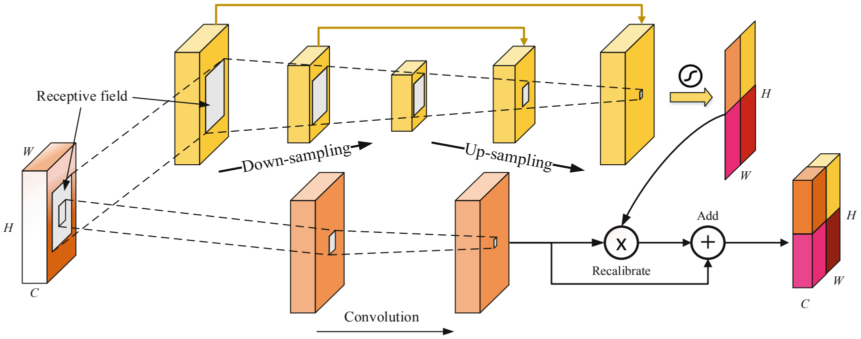

**Fig. 2.** The schematic illustration of the low-level features enhancement module.

## 2.2 Low-Level Feature Enhancement Module (LFE)

In FCN, features in different layers encode information at different levels. The features from bottom layers have rich spatial information, but they suffer from the problem of background noise and semantic ambiguity due to the small local receptive field and lack of guidance of the global context information. The detail reconstruction of the high-level features requires the help of high-resolution information from bottom layers. At the same time, we believe that the feature encoding of the bottom network in FCN requires the guidance of global semantic information to enhance the semantic consistency of high- and low-level features and suppress irrelative background noise. Therefore, the proposed LFE encodes the prior global semantic information of targets into the low-level features in an adaptive learning manner to enhance the semantic encoding ability of the bottom network.

The LFE consists of two branches: a semantic embedded branch and a trunk branch. The trunk branch is responsible for encoding and learning the features associated with the task. The semantic embedded branch is inspired by the excellent segmentation framework [2] and uses a mini encoder-decoder structure. The encoding stage quickly expands the receptive field and encodes global context information by down-sampling. The decoding stage restores spatial resolution through up-sampling and obtains high-level semantic features. Then the global semantic information of high-level features is embedded into the trunk branch to guide its feature encoding and optimize its encoding quality. At the same time, the trunk branch is enhanced to learn the features in the effective region. Specifically, as shown in Fig. 2, in the semantic embedded branch, the input feature maps $X$ obtain a more global receptive field and higher semantic features after two consecutive down-sampling and up-sampling operations. We also add skip connections between down-sampling and up-sampling to fuse information at different scales. Through a $1 \times 1$ convolutional layer, high-level semantic information is encoded into spatial projection map $s', s' \in \mathbb{R}^{H \times W}$. The final semantic embedded map is obtained from a Sigmoid layer. The value of each element in $\sigma(s')$ represents the relative importance of spatial information on the corresponding feature maps. Afterwards, this prior global information is embedded in the trunk branch to optimize its encoding quality through element-wise multiplication. In addition, in order to prevent the decrease of the feature response value of the trunk branch, we also introduce a residual connection. Therefore, the final output feature of LFE is expressed as $\tilde{X} = (1 + \sigma(s')) \otimes F_{tr}(X)$. In the literature [9], a similar structure with LFE is used to introduce a feature attention mechanism throughout the network. However, unlike [9], LFE aims to embed the global context information of the segmentation targets into low-level features through a lighter high-level semantic encoding module to improve the semantic encoding ability of the bottom network.

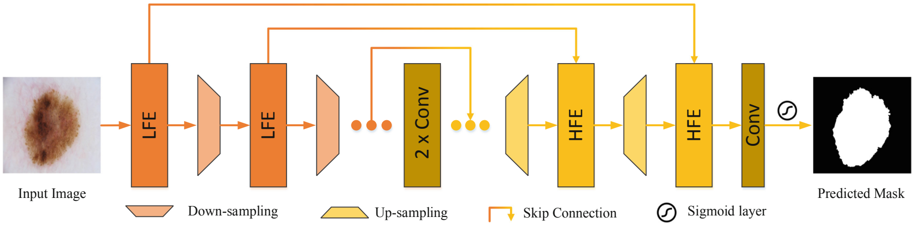

**Fig. 3.** Illustration of the framework of our proposed HLF-Net.

### 2.3 Segmentation Framework Based on Feature Enhancement

The proposed feature enhancement modules can be integrated into the existing segmentation framework to improve their feature learning abilities by replacing standard convolutional layers and skip connection operations. In this work, we integrate HFE and LFE into a typical encoder and decoder structure, and propose a new medical image segmentation network (HLE-Net). As shown in Fig. 3, the encoder network of HLE-Net is composed by superimposed LFEs. Each LFE provides semantic guidance of different levels for the encoding of the bottom network, which gradually enhances and refines the attention to complex targets. Then, the target details are reconstructed by the multi-layer HFEs and the original resolution is restored. Each convolution module consists of a $3 \times 3$ convolutional layer, a group normalization layer [10] and a ReLU layer. In this paper, HLF-Net contains 4 down-sampling and 4 up-sampling operations, and finally obtains the segmentation probability map through the Sigmoid function.

# 3 Experiments

## 3.1 Data and Experimental Setups

We extensively evaluated the proposed approach on ISIC 2017[1] skin lesion segmentation dataset [11] and the spleen segmentation dataset of CT volume images from Memorial Sloan Kettering Cancer Center[2]. In the skin lesion dataset, 2750 dermoscopic images from different clinical centers around the world were included, where 2000 for training, 150 for validation and the last 600 for testing. Our second dataset includes a total of 41 patient data. Due to memory limitations, we split the CT volume images into 512 × 512 slices to train the network. We performed data splitting at patient level and used images from 25, 4, 12 patients for training, validation and testing, respectively. Finally, by discarding some slices containing only background from the CT volume images, we obtained a total of 882 training images, 135 validation images, and 380 testing images.

Our HLE-Net is implemented using Pytorch on a Linux system with an Nvidia 1080Ti GPU. During training, we used the dice loss, the Adam optimizer with a learning rate of $1 \times 10^{-4}$ and the batch size of 6, with a learning rate reduction of 0.1 times after every 15 epochs. In each experiment, we saved the model that performed best on the validation set during training as the final test model. We used data augmentation including random copping and flipping to improve the robustness of the model. For the skin lesion dataset, we first re-scaled all images to 256 × 192 pixels and normalized the pixel values of each RGB channel to between 0 and 1. Besides the original RGB channels, we added an additional grayscale channel. For the spleen dataset, we first normalized all images to 0 to 1 and resized the images to 256 × 256.

In order to verify the effectiveness of the proposed method, we performed ablation studies on the two datasets, and compared HLE-Net with Unet-28 [2], Res-Unet-28. Res-Unet-28 is a modified Unet where each convolution block is replaced by the bottleneck building block used in the ResNet [8]. In order to evaluate our method fairly, the number of basic channels of Unet-28 and Res-Unet-28 is 28 to ensure that the number of parameters is similar to that of HLE-Net. We use the Dice coefficient, the Jaccard index and the Accuracy to evaluate the segmentation performance. Because Accuracy has very little discrimination on spleen dataset, we do not show the Accuracy of spleen segmentation in Table 1.

## 3.2 Results and Discussion

Table 1 shows the results of the different variants of the proposed method (only LFE, only HFE and HLE-Net) on the skin lesion dataset and the spleen dataset, respectively.

[1] https://challenge2017.isic-archive.com/.

[2] http://medicaldecathlon.com/.

In addition, the performance of Unet-28, Res-Unet-28 and the scores of the top three in the 2017 Skin Lesion Challenge leaderboard are also shown. It can be seen that the proposed modules considerably improve the segmentation performance of the network on both datasets. This indicates that both LFE and HFE can effectively enhance the encoding quality of the network. We further observe that LFE achieves higher performance than HFE, which confirms our hypothesis that it is more necessary for the feature encoding of the bottom network to require guidance from the global semantic information of high-level features. HLE-Net integrated LFE and HFE has the best performance among all methods. Compared with Unet-28, HLE-Net increases the jaccard index on the spleen dataset and the lesion dataset by 4.5% and 3.4%, respectively. It is also 2.3% higher than the best score on the leaderboard [13].

**Table 1.** Quantitative evaluation of different networks on spleen dataset and ISIC 2017.

| ISIC 2017 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Method | MResNet-Seg [11] | Berseth et al. [12] | Yuan et al. [13] | Unet-28 | Res-Unet-28 | HLE-Net (only HFE) | HLE-Net (only LFE) | HLE-Net |
| Dice | 0.844 | 0.847 | 0.849 | 0.838 | 0.841 | 0.859 | 0.862 | **0.866** |
| Jaccard | 0.760 | 0.762 | 0.765 | 0.754 | 0.755 | 0.777 | 0.783 | **0.788** |
| Accuracy | 0.934 | 0.932 | 0.934 | 0.930 | 0.931 | 0.935 | 0.935 | **0.939** |
| Spleen | | | | | | | | |
| Dice | – | – | – | 0.937 | 0.942 | 0.957 | 0.960 | **0.964** |
| Jaccard | – | – | – | 0.886 | 0.894 | 0.919 | 0.923 | **0.931** |
| Parameters | – | – | – | $5.9 \times 10^6$ | $6.2 \times 10^6$ | $3.6 \times 10^6$ | $3.8 \times 10^6$ | $5.5 \times 10^6$ |

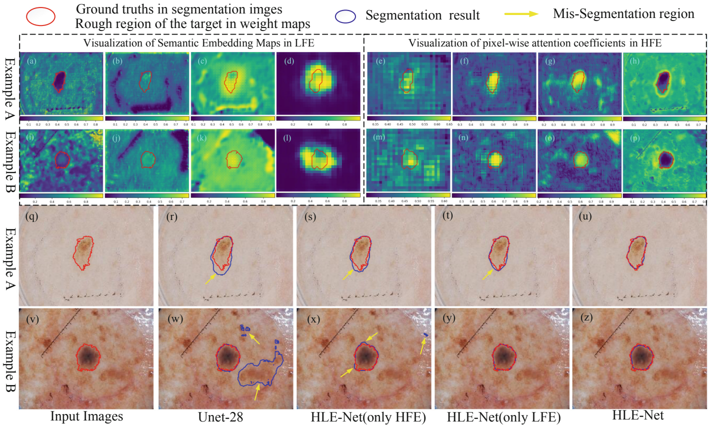

**Fig. 4.** The qualitative segmentation results of two examples (A, B) on ISIC 2017. Each example contains different network segmentation results and the visualization of the weight maps in LFE and HFE. From left to right (a–e–h, i–m–p), feature resolution goes from high to low, then from low to high, and finally restore the original resolution.

The qualitative segmentation results from two examples with different appearance on ISIC 2017 dataset are shown in Fig. 4. For the example A, the lesion area is close to normal skin, so Unet-28 incorrectly predicts normal skin as the lesion area, but HFE improves this situation. LFE further obtains a more accurate segmentation, which proves that improving the encoding quality of the bottom network can improve the segmentation result more effectively. For the example B, the background is very close to the lesion area. Unet-28 cannot accurately locate the lesion area. Gradually refining the attention on the segmentation targets through the attention mechanism can effectively solve this problem. Both LFE and HFE can accurately identify the lesion area, and LFE has a more accurate segmentation result. In addition, we also visualize the semantic embedding maps in LFE and the pixel-wise attention coefficient in HFE. It can be clearly seen that different LFE and HFE exert attention of different degrees on the segmentation targets, and as the network goes from shallow to deep, the concerning areas of LFE and HFE are gradually becoming more refined from blur.

## 4 Conclusion

This paper introduces two modules for feature enhancement for better medical image segmentation performance. LFE aims to encode high-level semantic information into the low-level features to improve the encoding ability of the bottom network. HFE optimizes the fusion efficiency of high- and low-level features using attention mechanism, which provides more high-resolution semantic guidance for high-level features. Based on these two modules, we propose a new medical image segmentation network (HLE-Net). The proposed method has achieved very competitive results in two very different tasks, skin lesion segmentation and spleen segmentation. This proves the effectiveness and wide adaptability of the proposed method. Future work aims to apply the proposed model to 3D segmentation or other segmentation tasks.

## References

1. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR, pp. 3431–3440 (2015)
2. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
3. Wang, H., Gu, R., Li, Z.: Automated segmentation of intervertebral disc using fully dilated separable deep neural networks. In: Zheng, G., Belavy, D., Cai, Y., Li, S. (eds.) CSI 2018. LNCS, vol. 11397, pp. 66–76. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-13736-6_6
4. Wang, G., Li, W., Ourselin, S., Vercauteren, T.: Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. In: Crimi, A., Bakas, S., Kuijf, H., Menze, B., Reyes, M. (eds.) BrainLes 2017. LNCS, vol. 10670, pp. 178–190. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75238-9_16

5. Zhang, Z., Zhang, X., Peng, C., Xue, X., Sun, J.: ExFuse: enhancing feature fusion for semantic segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11214, pp. 273–288. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01249-6_17
6. Roy, A.G., Navab, N., Wachinger, C.: Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks. In: Frangi, A., Schnabel, J., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 421–429. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_48
7. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M.: Attention U-Net: learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
9. Wang, F., et al.: Residual attention network for image classification. In: CVPR, pp. 3156–3164 (2017)
10. Wu, Y., He, K.: Group normalization. arXiv preprint arXiv:1803.08494 (2018)
11. Bi, L., Kim, J., Ahn, E., Feng, D.: Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks. arXiv preprint arXiv:1703.04197 (2017)
12. Berseth, M.: ISIC 2017-skin lesion analysis towards melanoma detection. arXiv preprint arXiv:1703.00523 (2017)
13. Yuan, Y.: Automatic skin lesion segmentation with fully convolutional-deconvolutional networks. arXiv:1803.08494 (2017)