Multimedia Tools and Applications
https://doi.org/10.1007/s11042-021-10536-5

1176: ARTIFICIAL INTELLIGENCE AND DEEP LEARNING FOR BIOMEDICAL APPLICATIONS

Res2-UNeXt: a novel deep learning framework for few-shot cell image segmentation

Sixian Chan[1] · Cheng Huang[1] · Cong Bai[1] · Weilong Ding[1] · Shengyong Chen[1]

Received: 14 May 2020 / Revised: 5 January 2021 / Accepted: 13 January 2021 /
Published online: 08 May 2021

Abstract
Recently, developing more accurate and more efficient deep learning algorithms for medical images segmentation attracts more and more attentions of researchers. Most of methods increase the depth of the network to replace with acquiring multi-information. The costs of training images annotation are too expensive to label by hand. In this paper, we propose a multi-scale and better performance deep architecture for medical image segmentation, named Res2-UNeXt. Our architecture is an encoder-decoder network with Res2XBlocks. The Res2XBlocks aim at acquiring multi-scale information better in images. To cooperate with Res2-UNeXt, we put forward a simple and efficient method of data augmentation. The data augmentation method, based on the process of cell movement and deformation, has biological implications in away. We evaluate Res2-UNeXt in comparison with recent variants of U-Net: UNet++, CE-Net and LadderNet and the method that different from U-Net architecture: FCN and DFANet on the dataset of ISBI cell tracking challenge 2019 via four different cell images. The experimental results demonstrate that Res2-UNeXt can achieve better performance than both recent variants of U-Net and non-U-Net architecture methods. Besides, the proposed architecture and the data augmentation method have been proven efficiently by the ablation experiments.

Keywords Segmentation of medical images · Data augmentation · Deep learning · Image processing

1 Introduction

Image segmentation is the technology of dividing the image into several specific and unique regions and proposing the interested objects. It is a key step from image processing to image

✉ Weilong Ding
wlding@zjut.edu.cn

Sixian Chan
sxchan@zjut.edu.cn

[1] College of Computer Science, Zhejiang University of Technology, No. 288 Road LiuHe, Hangzhou, Zhejiang, 310023, China

analysis. For medical images, they are very important in clinical diagnosis. By processing and observing medical images, doctors can determine the cause of disease more quickly and accurately [17, 18]. Of course, image segmentation is the most common and significant processing. The development of medical image segmentation technology not only affects the development of other related technologies in medical image processing, such as visualization [20, 25] and 3D reconstruction [8, 9], but also plays an extremely important role in the analysis of biomedical images [30]. In recent years, due to the application of deep learning algorithm in medical image segmentation [5, 19, 33, 36, 43], medical image segmentation technology has made remarkable progress. The simplest approach to train a deep network for segmentation is the patch-based architecture [34]. It selects small patches around each pixel with a class label to train a network. Some popular network architectures for segmentation are designed by using this approach [14, 16, 27, 41]. However a patch only represents the label of one pixel in the whole image. Thus, this approach usually costs a long time to train a desirable model. The encoder-decoder architecture like U-Net [31] and FCN [24] maybe a nice choice for image segmentation. It is composed of locally connected layers such as convolution, pooling and up pooling (up sampling) [34] and outputs the segmentation map with the same size as the input image. In the encoder part, it extracts the size of the feature map with pooling or big stride convolution. During decoding, it recovers the size of the feature map while deconvolving higher-level features extracted from the encoder part. There are also many studies which adopt the encoder-decoder architecture for segmentation [11, 12, 15, 23, 39, 44, 45]. However, most of them do not pay more attentions to the ability of obtaining the multi-scale feature. To make up for this shortcoming, we are inspired by Res2Net [10] and propose a new multi-scale backbone architecture into U-Net. We create a multi-scale U-Net to obtain coarse to fine-grained information better, for improving the performance of segmentation.

In this paper, we introduce a novel framework for medical image segmentation, termed Res2-UNeXt. This framework combines the inspirations distilled from U-Net [31] backbone architecture, ResNeXt [40] and Res2Net [10] and demonstrates competitive performance. In addition, we describe data augmentation inspired by image registration that behaves well for medical image few-shot segmentation problems with label imbalance. In summary, the main contributions of this paper are as follows:

- To address the problem that U-Net and its variants cannot acquire multi-scale information perfectly, we propose Res2-UNeXt. The Res2-UNeXt can obtain both coarse-grained and fine-grained features to achieve a better performance at the segmentation of the cell edge.
- To overcome the problem of few-shot segmentation, we explicit the image registration method to augment cell image datasets. Besides, this method is also universal for data augmentation in medical image segmentation with time information.
- We evaluate Res2-UNeXt in comparison with recent variants of U-Net [31]: UNet++ [44], CE-Net [11] and LadderNet [45] and the method that different from U-Net architecture: FCN [24] and DFANet [22] on the dataset of ISBI cell tracking challenge 2019 via four different cell images. The experimental results demonstrate that Res2-UNeXt achieves the best performance in the above architectures.

The rest of this paper is organized as follows. In Section 2, we review related work in past studies. In Section 3, we present the data augmentation method based on image registration. In Section 4, we detail the framework of Res2-UNeXt. In Section 5, we analyze and discuss the experiment results. The work is concluded in Section 6.

2 Related work

2.1 Network architecture

For semantic segmentation, Fully Convolutional Networks (FCN) [24] was a watershed with milestone significance. Since it improved the state-of-the-art performance of segmentation by a significant margin (about 20% relative improvement over the state-of-the-art on the PASCAL VOC). The authors built "fully convolutional" networks that took the input of arbitrary size and produced correspondingly-sized output with efficient inference and learning. At the same time, they adapted contemporary classification networks into fully convolutional networks and transferred their learned representations by fine-tuning to the segmentation task. There were additional improvements that had been proposed with the use of Deeplab [3] models. It was the first time that showcased the importance of atrous convolutions for semantic segmentation. To refine the final segmentation, a conditioned random field (CRF) was also utilized as a post processing step. Under such background, Ronneberger et al. [31] proposed U-Net architecture which was presented for dealing with the biomedical images. It introduced the encoder-decoder paradigm via up-sampling gradually from lower size features to the original image size. After U-Net [31], there were many variants of this architecture had been proposed. TernausNet [15] replaced the encoder of U-Net with the VGG11 encoder and got the champion of the Kaggle Carvana Image Masking Challenge. UNet++ [44] designed a nested U-Net architecture with dense skip connections. In addition, inspired by residual connections and dense connections, Res-UNet [39] and Dense-UNet [12] used residual connections and dense connections to improve the blocks of U-Net. All of these variants did not pay attentions to the acquisition of multi-scale information. However, medical image segmentation usually required high precision. For instance, the size of the cells in cell images was different. In consequence, an excellent medical image segmentation model was supported to have the ability of obtaining both coarse-grained and fine-grained information. In other words, this model must be a multi-scale architecture. Gao et al. [10] proposed a new multi-scale backbone architecture, called Res2Net. It was proved to have better performance in the field of image segmentation and object detection. Inspired by this architecture, we inset Res2Net [10] backbone to U-Net [31] architecture and name it as Res2-UNeXt. Then, we apply Res2-UNeXt for the cell image segmentation in three datasets of ISBI cell tracking challenge 2019 and achieve the best results.

2.2 Few-shot segmentation and data augmentation

In the semantic segmentation of medical images, few-shot segmentation is a challenging task. Making the label in medical images is so difficult that we have to hire a couple of professional doctors to do this work. However, it is expensive. Unfortunately, the existing methods were mainly focused on natural images. Approaches for few-shot semantic segmentation utilized information from prototypical examples of the classes to be designed [6, 35]. Other methods use large labeled datasets of supplementary information such as object appearances [2], or additional information was leveraged such as human input [28]. Medical images present different challenges from natural images. For example, compared with the differences between objects in natural images, the visual differences between tissue classes of human beings were very small. Meanwhile, they suffered from the influence of imaging equipment itself.

Data enhancement was usually performed using by simple parametric transformations, such as rotation and scaling, in image-based supervised learning tasks. For medical images,

random smooth flow field was a common method to simulate anatomical changes [1, 26, 32]. These simple parametric transformations could avoid overfitting and improve algorithm performance [1, 26, 32]. However, the performance gains were caused by transforming uncontrollably with the selection of conversion functions and parameter settings [7].

Recent works had proposed learning data augmentation transformations from data. Hauberg et al. [13] focused on data augmentation to classifying MNIST digits. They aligned image pairs within each class under the assumption that the spatial transformation between images belonged to a large class of diffeomorphisms. Then, a class-specific probabilistic generative model of the transformations was learned in a Riemannian sub-manifold of the Lie group of diffeomorphisms. Other works paid attention to learn the combinations of different simple transformation functions (e.g., rotation and contrast enhancement) to perform data augmentation for natural images [4, 29]. Cubuk et al. [4] described a simple procedure called AutoAugment to automatically search for improved data augmentation policies which maximized classification accuracy. Ratner et al. [29] proposed a method for automating this process by learning a generative sequence model over user-specified transformation functions which used a generative adversarial approach. Yet these simple transformations were insufficient for capturing many of the subtle variations in MRI data. Zhao et al. [42] trained an appearance model in addition to a spatial model and used these two models to generate synthetic data. This approach had gotten a good performance in the segmentation of brain MR images. Combining with the characteristics of the cell image dataset of ISBI cell tracking challenge 2019, we draw lessons from this method. The 'Demons' algorithm [38] is explicated to generate new and reality cell images. It can improve the segmentation performance of cell images by augmenting the training datasets.

3 Data augmentation base on image registration

Few-shot segmentation is a challenging task in semantic segmentation for medical images. The deep learning methods usually need a great number of data with labels. In the dataset of ISBI cell tracking challenge 2019, there are only a few numbers of images are labeled. Thus, we have to augment the data. Tagging labels to the data by humans is a good way to expand our training data. But it is too expensive. So an efficient, sample and automatic approach is required to be explored urgently to augment the dataset.

3.1 'Demons' algorithm for augmentation

As we mentioned, the cost of obtaining medical images, especially corresponding labels of them, is too expensive. This makes it impractical to hire experts to do a lot of tagging jobs. As a result, it is a hot study that using semi-supervised approaches to generate new training images and their ground-truth in medical image segmentation, which can be treated as data augmentation. Most of these studies are combined nonrigid image registration with deep learning. Whereas they only focus on human tissue images, like brain, stomach, liver and so on. Why nobody has yet applied nonrigid image registration to the cell image dataset? There are two main reasons:

- The main contents of human tissue images may not change obviously. Only the details will change. As for phase contrast microscope cell images, the cells are alive. The quantity and shape of cells change dramatically in different images. This problem cannot be addressed by a registration model that is only trained with only one template image.

- The richer the image contents, the easier to match. The tissues occupy most of the foreground in human tissue images. The foreground is consecutive. It is a benefit for the methods of image registration based on deep learning to get a good performance on augmentation. As for phase contrast microscope cell images, the most foreground is discrete and suffers from the cell division and great deformation. It is a big challenge for the deep learning model to train by a single template image.

For those two reasons, the previous studies cannot be applied to the phase contrast microscope cell image dataset of ISBI cell tracking challenge 2019 directly. Fortunately, through observation, we find some characteristics of the preprocessed cell images. Phase contrast microscope cell image dataset is a continuous sequence of image frames with time information. The images between two consecutive frames are very similar but not identical. Demons algorithm [38] is originally derived from the optical flow algorithm, which is adept at estimating the displacement of the target of two adjacent frames in a video such as the cell image dataset of ISBI cell tracking challenge 2019. According to these properties, we apply 'Demons' algorithm [38], a traditional method of image registration, to the cell image dataset and generate new cell images and their labels.

3.2 'Demons' algorithm

Demons is a famous non-parametric deformable registration method. It is originally derived from the optical flow algorithm, which is used to estimate the displacement of the target of two adjacent frames in a video image. It means the velocity of the target moving. Thus, the velocity is computed as

$$v = \frac{(s-m) \bigtriangledown s}{|\bigtriangledown s|^2} \tag{1}$$

Where v is the velocity. m is the first frame. s is the next frame. $\bigtriangledown s$ is the gradient of s, $s-m$ is the gray difference of s and m. In image registration, s is named Static Image and m is named Moving Image. The target of registration is to determine the correspondence between pixels in s and m through finding a transformation from m to s. Thirion et al. [37] introduced the optical flow algorithm into image registration. In order to prevent the calculation problem when the image gradient is zero, an image gray difference is added in the denominator of formula (1):

$$u = \frac{(s-m) \bigtriangledown s}{|\bigtriangledown s|^2 + (s-m)^2} \tag{2}$$

After each iteration of formula (2), a small displacement field u can be calculated. To get better performance of registration, Thirion [37] smoothed the displacement field with gaussian filter. On this basis, Wang et al. [38] added both $\bigtriangledown s$ and $\bigtriangledown m$ to get displacement filed:

$$u = (s-m) \times \left(\frac{\bigtriangledown s}{|\bigtriangledown s|^2 + \alpha^2 (s-m)^2} + \frac{\bigtriangledown m}{|\bigtriangledown m|^2 + \alpha^2 (s-m)^2} \right) \tag{3}$$

$$new_{img} = m \bigoplus u \tag{4}$$

Where $\bigtriangledown m$ is the gradient of m. α is a hyperparameter to adjust the amplitude of deformation. $\bigoplus$ is an operation that Moving Image m do defamation base on displacement filed u. new_{img} is the target image.

Standing on the shoulders of giants, we come up with an idea of data augmentation of cell image datasets. There are a few numbers of labeled images in the dataset. All of the labeled images could be Moving Image (m). Their next frame could be a Static Image (s).

When Moving Image adds a displacement field from formula (3), the label will do the same deformation as Moving Image. As a result, a new image is different from the original data, and its label is generated.

3.3 Synthesizing new example

As showed in Fig. 1, Our Aug Image is different from both static image and moving image. Firstly, we obtain the Displacement Field from Moving Image to Static Image through 'Demons' algorithm [38] by formula (3). Then, the Displacement Filed is added to Moving Image and its label to get the Aug Image and the corresponding label. As we knew, the new label is fit to the new cell image and the new cell image inherits not only the textural feature and the shape feature of Moving Image but also the spatial information of Static Image. The process of generating new cell images simulates the process of cells' movement and deformation. This has some biological implications and proves that our method of data augmentation is effective. Above it is most beneficial to train of our image segmentation model.

4 The Res2-UNeXt

In this Section, we introduce the architecture of Res2-UNeXt with details in Sections 4.1 and 4.2. An efficient loss function to achieve class-imbalance problem and cell-closing problem is detailed in Section 4.3.

4.1 Architecture

Our architecture consists of several modules, described as follows:

- A U-Net [31] backbone architecture belongs the encoder-decoder paradigm. We apply it for smoothly and gradually shifting from the image to the segmentation mask.

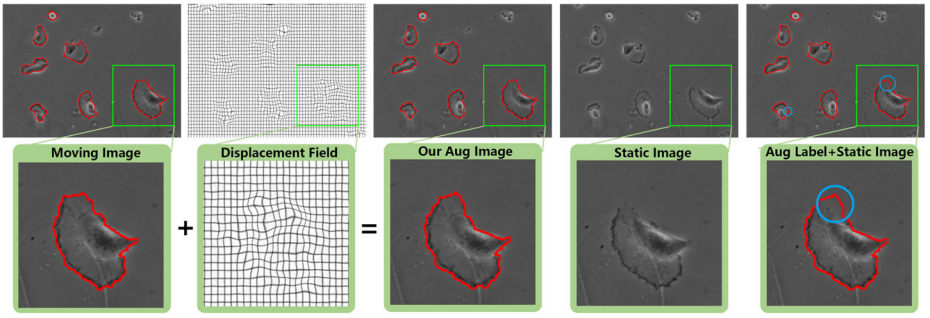

Fig. 1 Data Augmentation comparison diagram, Moving Image with label (red boundary), Displacement Field is generated by formula (3), Our Aug Image(image and label(red boundary)) is generated by formula (4), Static Image is an image next frame of Moving Image in the dataset. The last is Static Image with the label of Our Aug Image. Blue circles are the significant difference area between Our Aug Image and Static Image

- With the increase of network depth, the uniform training becomes hard but necessary. It is a good choice that the residual blocks of convolutional layers with split-transform-merge named ResNeXt [40]. To a great extent, the residual blocks eliminate the problems of gradual disappearance and explosion in the deep structures. The group convolution in ResNeXt [40] can improve the performance of the network.
- To understand the across scales, we take advantage of the Res2Net [10] for each residual building block with split-transform-merge. Res2net [10] reveals a new dimension, "scale". In addition to the existing depth, width and cardinality dimensions, the scale is a more important and effective factor. It has the ability to improve the multi-scale of CNN on a more fine-grained scale.

We term our network Res2-UNeXt because it consists of hierarchical residual-like connections within one single residual block with a separate transform merge and a U-Net backbone framework. In Res2-UNeXt the encoder part forms from six Res2XBlocks followed by the decoder which is constitutive of four Res2XBlocks. Next, we will introduce the Res2-UNeXt in detail.

4.2 The Res2-UNeXt framework

The Res2-UNeXt also belongs to an encoder-decoder style like the U-Net. The stacked layers of modified residual building blocks, which are named Res2XBlock, form the core of the network, as show in Fig. 2. In each Res2XBlock, a 3×3 convolution layer to initiate the future map of this block is employed. Figure 3 shows a Res2NeXt block. Firstly, the feature maps are evenly split into four feature map subsets denoted by x_i, where $i \in \{1, 2, 3, 4\}$. Each feature subset x_i has the same spatial scale within 1/4 number of channels of the input feature map. Secondly, except for x_1, we apply a corresponding 3×3 group convolution to each x_i, denoted by $m_i()$. The result of $m_i()$ is denoted by y_i. Then, the subset x_i of feature map is added correspondingly with the result of $m_{i-1}()$ and input to $m_i()$. In addition, skipping 3×3 group convolution for x_1 is the method of decrease of parameters. Thus, y_i is formulated as follows:

$$y_i = \begin{cases} x_i & i = 1; \\ m_i(x_i) & i = 2; \\ m_i(x_i + y_{i-1}) & 2 < i \leq 4 \end{cases} \tag{5}$$

It is noteworthy that each 3×3 convolutional operator $m_i()$ can potentially achieve feature merging from all splits x_j, $j \leq i$. When a split x_j passes a 3×3 group convolutional operator, the receptive of the output will be expanded. In conclusion, the feature map has richer scales information after passing the Res2XBlock.

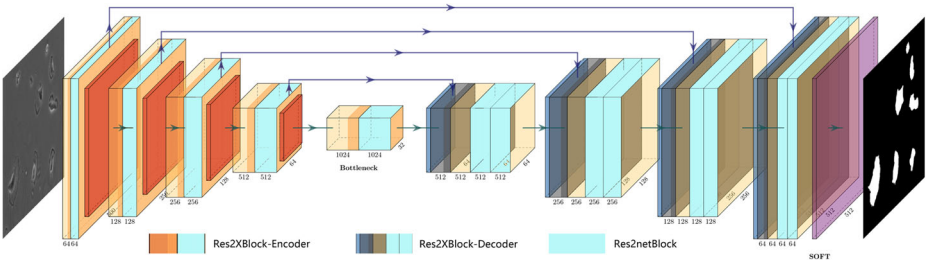

Fig. 2 Res2-UNext

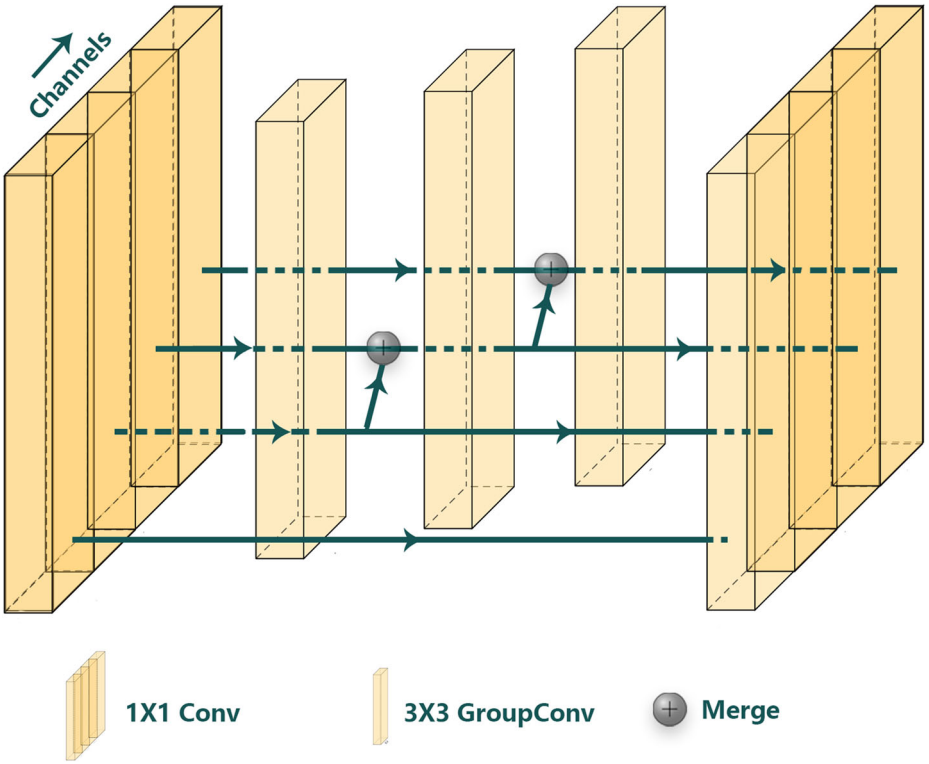

Fig. 3 Res2net Block

In Res2XBlock, feature map splits are used to extract the coarse-grained and fine-grained information. In order to merge the feature of different scales better, we connect all the splits and pass them to through a 1×1 convolution, as shown in Fig. 3. This operation can force convolutions to process features more efficiently. To reduce the number of parameters, the convolution of the first split is ignored. This can also be considered as a form of feature reused.

In the encoder state of the network, the output of each remaining block is down-sampled, where the maximum step of the maximum buffer layer is only two strides. It can further enlarge the receptive field and reduce the number of parameters.

In the decoder state, the upsampling is being done with the use of a 2×2 convolution ("up-convolution") followed by each Res2XBlock. The combination of layers from the encoder and decoder states can concatenate the two inputs and make up for the loss of feature caused by downsampled.

4.3 Loss function

The loss function is computed by a pixel-wise soft-max over the final feature map combined with the cross-entropy. The soft-max is defined as :

$$p_n(x) = exp(k_n(x))/(\Sigma_{n'=1}^{N} exp(k_{n'}(x))) \tag{6}$$

where $k_n(x)$ denotes the activation in feature channel n at the pixel position $x \in \Omega$ with $\Omega \subset Q^2$. N is the number of classes and $p_n(x)$ is the approximated maximum-function. The cross-entropy penalizes at each position the deviation of $p_{\lambda(x)}(x)$ from formula (6) by:

$$Loss = -\sum_{x \in \Omega} w(x) log(p_{\lambda(x)}(x)) \tag{7}$$

where $\lambda : \Omega \rightarrow \{1, ..., N\}$ is the true label of each pixel and $w : \Omega \rightarrow R$ is a weight map that introduced to give some pixels more importance in the training.

The weight maps are pre-computed for each ground-truth segmentation to compensate for the different frequencies of pixels of a certain category in the training dataset. For example, in a couple of cell images, the pixels of the cell are the majority or the pixels of the background are the majority. This imbalance is not a benefit for training. And another function of weight maps is making the network learn smaller separation boundaries introduced between closed cells. But our weight map is not similar with the map in U-Net, as shown in Fig. 4d and e. Our task is semantic segmentation, not instance segmentation. Giving the area between every touching cell a high weight will make the segmentation ability of the model worse. As a result, we only give the high weight in the area between closing cells, rather than the touching cells, as shown in Fig. 4d. The erosion operations are also not used to get the separation border between touching cells, as shown in Fig. 4c. The computation of w can be reduced and computed as:

$$w(x) = w_{cb}(x) + w_0 \times exp(-\frac{(dist_1(x) + dist_2(x))^2}{2\sigma^2}) \tag{8}$$

where $w_{cb} : \Omega \rightarrow R$ is the weight map to balance the class frequencies. $dist_1 : \Omega \rightarrow R$ means the distance to the border of the nearest cell. $dist_2 : \Omega \rightarrow R$ is the distance to the border of the second nearest cell. In the experiments, we set $w_0 = 9$ and $\sigma^2 \approx 25$ pixels.

5 Experiments

Datasets: As shown in Table 1, four cell image datasets from ISBI cell tracking challenge 2019 are applied for model evaluation. There are a few images with labels in those four datasets. Firstly, data enhancement is explicitly exploited to make the number of labeled images double. Then, random deformation and flip to 80x is operated for the datasets, as shown in Table 2.

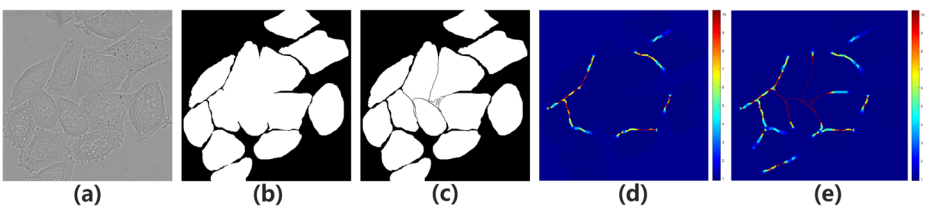

Fig. 4 Weight comparison diagram, **a** raw HeLa cells Image. **b** ground truth for semantic segmentation. **c** ground truth with erosion operations for instance segmentation(U-Net). **d** map with a pixel-wise loss weight for closing cells. **e** map with a pixel-wise loss weight for closing and touching cells.(U-Net)

Table 1 The image segmentation datasets used in our experiments

Dataset	Image number	Input size	Modality	Provider
DIC-C2DH-HeLa	168	512*512	microscopy	ISBI cell tracking challenge 2019
Fluo-N2DH-GOWT1	184	1024*1024	microscopy	ISBI cell tracking challenge 2019
PhC-C2DH-U373	230	689*520	microscopy	ISBI cell tracking challenge 2019
Fluo-C2DL-MSC	96	1200*782	microscopy	ISBI cell tracking challenge 2019

Evaluation index: We monitor the Intersection over Union (IoU) which can measure the performance of segmentation models. The IoU is computed as:

$$IOU(S, R) = \frac{|R \cap S|}{|R \cup S|} \tag{9}$$

where R denotes the set of pixels belonging to a reference object and S denotes the set of pixels belonging to its matching segmented object.

Baseline models: For comparison, we used the original U-Net [31] architecture as a baseline as it was a common performance baseline for image segmentation.

Implementation Details: We trained Res2-UNeXt with a batch size of 2 for 40 epochs using Adaptive moment estimation(Adam) [21] with an initial learning rate of 5e-4 on NVIDIA GeForce RTX 2080Ti(11GB). The learning rate would drop to a fifth of its original level with every 15 epochs. Besides, the architecture details for Res2-UNeXt were shown in Figss. 2 and 3.

Results: Table 3 demonstrated the comparison of U-Net [31] and Res2-UNeXt in terms of IoU for the task of cell segmentation in DIC-C2DH-HeLa dataset, Fluo-N2DH-GOWT1 dataset, PhC-C2DH-U373 dataset and Fluo-C2DL-MSC dataset. As it could be seen, the model only adding Res2XBlock or 'Demons' augmentation outperformed U-Net.

Table 2 The number of images in our experiments

Dataset	labeled images	Train/Val	Augment w/o 'Demons'	Augment w/ 'Demons'
DIC-C2DH-HeLa	18	Train	1079	2158
		Val	249	249
Fluo-N2DH-GOWT1	8	Train	498	996
		Val	166	166
PhC-C2DH-U373	34	Train	2241	4482
		Val	581	581
Fluo-C2DL-MSC	51	Train	2656	5229
		Val	581	581

Table 3 Segmentation results (IoU: %) for U-Net and our suggested architecture Res2-UNeXt with and without 'Demons' augment

Architecture	Dataset			
	DIC-C2DH-HeLa	Fluo-N2DH-GOWT1	PhC-C2DH-U373	Fluo-C2DL-MSC
U-Net w/o 'Demons'	87.05	93.95	86.19	77.97
U-Net w/ 'Demons'	87.77	94.06	87.45	78.02
Res2-UNeXt w/o 'Demons'	87.49	94.11	88.04	78.00
Res2-UNeXt w/ 'Demons'	**88.43**	**94.29**	**88.90**	**79.03**

The bold entries show that our method obtains the best performance comparing with other methods

Res2-UNeXt w/ 'Demons' achieved a significant performance gaining over the model above, yielding an average improvement of 0.3 and 2.6 points in IoU. Figure 5 showed a qualitative segmentation performance between Res2-UNeXt, FCN [24], CE-Net [11], UNet++ [44], LadderNet [45] and DFANet [22], and Table 4 compared the segmentation accuracy(IoU) of these models. As they were described, U-Net family architectures performed much better than non-U-Net architectures in this cell image segmentation task and our method performed best in the details of the picture, especially at the edge of the cell. In addition, our method could filter some small impurities in cell images. These impurities would trick other models to do a mistake segmentation. But our model could overcome this problem. These advantages were due to our multi-scales architecture and efficient data augmentation. Multi-scales architecture could improve the ability of the model to extract fine-grained and coarse-grained information. It was a benefit to segment better at the edge of the cell and filter the small impurities. What is more, our data augmentation expanded the quantity of training data. It was a helpful for training our deep learning model. Thus, our method could perform better than others.

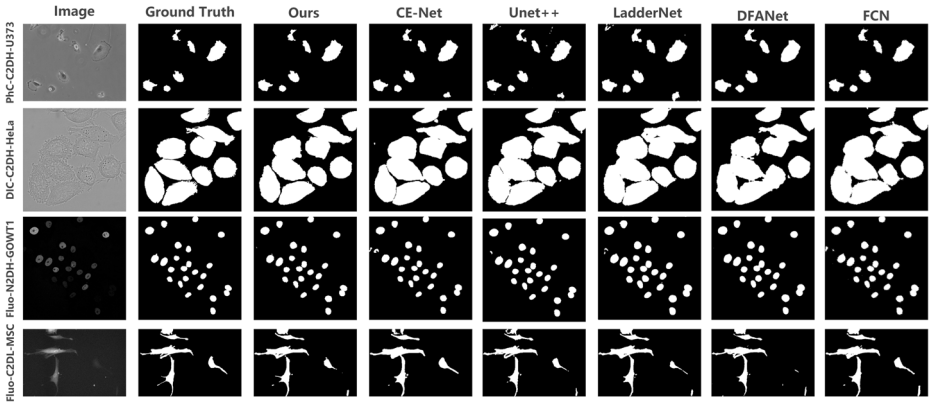

Fig. 5 Qualitative comparison between Res2-UNeXt(Ours), CE-Net [11], UNet++ [44], LadderNet [45], DFANet [22] and FCN [24], showing segmentation results for PhC-C2DH-U373, DIC-C2DH-HeLa, Fluo-N2DH-GOWT1 and Fluo-C2DL-MSC

Table 4 Segmentation results (IoU: %) for Res2-UNeXt(Ours), FCN [24], UNet++ [44], CE-Net [11], LadderNet [45] and DFANet [22]

Architecture	Methods	Dataset			
		DIC-C2DH-HeLa	Fluo-N2DH-GOWT1	PhC-C2DH-U373	Fluo-C2DL-MSC
Non U-Net family	FCN [24](2015)	86.39	85.96	84.93	74.85
	DFANet [22](2019)	84.35	88.86	80.45	72.17
U-Net family	UNet++ [44](2018)	88.02	93.83	86.76	78.17
	CE-Net [11](2018)	87.14	93.25	87.40	77.69
	LadderNet [45](2019)	86.53	94.17	88.81	75.61
	Res2-UNeXt w/ 'Demons'(Ours)	**88.43**	**94.29**	**88.90**	**79.03**

The bold entries show that our method obtains the best performance comparing with other methods

6 Conclusion

In this paper, we proposed Res2-UNeXt to address the multi-scales problem and the few-shot problem of medical image segmentation. The suggested architecture aimed at acquiring both coarse-grained and fine-grained information better. 'Demons' algorithm was applied for expanding our training data. The experiments demonstrated that our data augmentation method is effective as well as proved the effectiveness of Res2-UNeXt. Our algorithm achieved more excellent performance than FCN, CE-Net, UNet++, LadderNet and DFANet.

Acknowledgements This work is partially supported by National Natural Science Foundation of China under Grant No. U1908210, 61906168, Zhejiang Public Welfare Technology Research Plan / Social Development Project No.LGF21F020015.

References

1. Burghelea T, Segre E, Steinberg V (2005) Validity of the taylor hypothesis in a random spatially smooth flow. Physics of Fluids 17(10):103101
2. Caelles S, Maninis KK, Pont-Tuset J, Leal-Taixé L, Cremers D, Van Gool L (2017) One-shot video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 221–230
3. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2017) Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans Pattern Analysis Machine Intell 40(4):834–848
4. Cubuk ED, Zoph B, Mane D, Vasudevan V, Le QV (2019) Autoaugment: learning augmentation strategies from data. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 113–123
5. Dolz J, Gopinath K, Yuan J, Lombaert H, Desrosiers C, Ayed IB (2018) Hyperdense-net: a hyper-densely connected cnn for multi-modal image segmentation. IEEE Trans Medical Imag 38(5):1116–1126
6. Dong N, Xing E (2018) Few-shot semantic segmentation with prototype learning. In: BMVC
7. Dosovitskiy A, Fischer P, Springenberg JT, Riedmiller M, Brox T (2015) Discriminative unsupervised feature learning with exemplar convolutional neural networks. IEEE Trans Pattern Analysis Mach Intell 38(9):1734–1747

8. Fahrig R, Holdsworth D (2000) Three-dimensional computed tomographic reconstruction using a c-arm mounted xrii: image-based correction of gantry motion nonidealities. Med Phys 27(1):30–38
9. Fan H, Su H, Guibas LJ (2017) A point set generation network for 3d object reconstruction from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 605–613
10. Gao S-H, Cheng M-M, Zhao K, Zhang X-Y, Yang M-H, Torr P (2021) Res2net: a new multi-scale backbone architecture. IEEE Trans Pattern Anal Mach Intell 43(2):652–662. https://doi.org/10.1109/TPAMI.2019.2938758
11. Gu Z, Cheng J, Fu H, Zhou K, Hao H, Zhao Y, Zhang T, Gao S, Liu J (2019) Ce-net: context encoder network for 2d medical image segmentation. IEEE Trans Med Imag 38(10):2281–2292
12. Guan S, Khan AA, Sikdar S, Chitnis PV (2019) Fully dense unet for 2d sparse photoacoustic tomography artifact removal. IEEE J Biomed Health Inform 24(2):568–576. IEEE
13. Hauberg S, Freifeld O, Larsen ABL, Fisher J, Hansen L (2016) Dreaming more data: class-dependent distributions over diffeomorphisms for learned data augmentation. In: Artificial intelligence and statistics, pp 342–350
14. Havaei M, Davy A, Warde-Farley D, Biard A, Courville A, Bengio Y, Pal C, Jodoin PM, Larochelle H (2017) Brain tumor segmentation with deep neural networks. Medical Image Analysis 35:18–31
15. Iglovikov V, Shvets A (2018) Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. arXiv:1801.05746
16. Kamnitsas K, Ledig C, Newcombe VF, Simpson JP, Kane AD, Menon DK, Rueckert D, Glocker B (2017) Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. Medical Image Analysis 36:61–78
17. Khan MA, Ashraf I, Alhaisoni M, Damaševičius R, Scherer R, Rehman A, Bukhari SAC (2020) Multimodal brain tumor classification using deep learning and robust feature selection: a machine learning application for radiologists. Diagnostics 10(8):565
18. Khan MA, Qasim M, Lodhi HMJ, Nazir M, Javed K, Rubab S, Din A, Habib U (2020) Automated design for recognition of blood cells diseases from hematopathology using classical features selection and elm. Microscopy Research and Technique. Wiley Online Library
19. Khan MA, Sharif M, Akram T, Bukhari SAC, Nayak RS (2020) Developed newton-raphson based deep features selection framework for skin lesion recognition. Pattern Recogn Lett 129:293–303. https://doi.org/10.1016/j.patrec.2019.11.034
20. Khan MUG, Gotoh Y, Nida N (2017) Medical image colorization for better visualization and segmentation. In: Annual conference on medical image understanding and analysis, pp 571–580. Springer
21. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv:1412.6980
22. Li H, Xiong P, Fan H, Sun J (2019) Dfanet: deep feature aggregation for real-time semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 9522–9531
23. Lin F, Wu Q, Liu J, Wang D, Kong X (2020) Path aggregation u-net model for brain tumor segmentation. Multimedia Tools and Applications: 1–14. Springer
24. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3431–3440
25. McAuliffe MJ, Lalonde FM, McGarry D, Gandler W, Csaky K, Trus BL (2001) Medical image processing, analysis and visualization in clinical research. In: Proceedings 14th IEEE symposium on computer-based medical systems. CBMS 2001, pp 381–386. IEEE
26. Milletari F, Navab N, Ahmadi SA (2016) V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth international conference on 3d vision (3DV), pp 565–571. IEEE
27. Pereira S, Pinto A, Alves V, Silva CA (2016) Brain tumor segmentation using convolutional neural networks in mri images. IEEE Trans Med Imag 35(5):1240–1251
28. Rakelly K, Shelhamer E, Darrell T, Efros AA, Levine S (2018) Few-shot segmentation propagation with guided networks. arXiv:1806.07373
29. Ratner AJ, Ehrenberg H, Hussain Z, Dunnmon J, Ré C (2017) Learning to compose domain-specific transformations for data augmentation. In: Advances in neural information processing systems, pp 3236–3246
30. Rehman A, Khan MA, Mehmood Z, Saba T, Sardaraz M, Rashid M (2020) Microscopic melanoma detection and classification: a framework of pixel-based fusion and multilevel features reduction. Microsc Res Tech 83(4):410–423
31. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention, pp 234–241. Springer

32. Roth HR, Lee CT, Shin HC, Seff A, Kim L, Yao J, Lu L, Summers RM (2015) Anatomy-specific classification of medical images using deep convolutional nets. In: 2015 IEEE 12th international symposium on biomedical imaging (ISBI), pp 101–104. IEEE
33. Roth HR, Shen C, Oda H, Oda M, Hayashi Y, Misawa K, Mori K (2018) Deep learning and its application to medical image segmentation. Medical Imaging Technology 36(2):63–71
34. Seo H, Khuzani MB, Vasudevan V, Huang C, Ren H, Xiao R, Jia X, Xing L (2019) Machine learning techniques for biomedical image segmentation: an overview of technical aspects and introduction to state-of-art applications. arXiv:1911.02521
35. Shaban A, Bansal S, Liu Z, Essa I, Boots B (2017) One-shot learning for semantic segmentation. arXiv:1709.03410
36. Taghanaki SA, Abhishek K, Cohen JP, Cohen-Adad J, Hamarneh G (2019) Deep semantic segmentation of natural and medical images: a review. arXiv:1910.07655
37. Thirion JP (1998) Image matching as a diffusion process: an analogy with maxwell's demons. Med Image Anal 2(3):243–260
38. Wang H, Dong L, O'Daniel J, Mohan R, Garden AS, Ang KK, Kuban DA, Bonnen M, Chang JY, Cheung R (2005) Validation of an accelerated 'demons' algorithm for deformable image registration in radiation therapy. Physics in Medicine & Biology 50(12):2887
39. Xiao X, Lian S, Luo Z, Li S (2018) Weighted res-unet for high-quality retina vessel segmentation. In: 2018 9th international conference on information technology in medicine and education (ITME), pp 327–331. IEEE
40. Xie S, Girshick R, Dollár P, Tu Z, He K (2017) Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1492–1500
41. Zhang W, Li R, Deng H, Wang L, Lin W, Ji S, Shen D (2015) Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. NeuroImage 108:214–224
42. Zhao A, Balakrishnan G, Durand F, Guttag JV, Dalca AV (2019) Data augmentation using learned transformations for one-shot medical image segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8543–8553
43. Zhao X, Wu Y, Song G, Li Z, Zhang Y, Fan Y (2018) A deep learning model integrating fcnns and crfs for brain tumor segmentation. Medical image analysis 43:98–111
44. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J (2019) Unet++: redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans Med Imaging 39(6):1856–1867. IEEE
45. Zhuang J (2018) Laddernet: multi-path networks based on u-net for medical image segmentation. arXiv:1810.07810