Pattern Recognition 127 (2022) 108625

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

2K-Fold-Net and feature enhanced 4-Fold-Net for medical image segmentation

Yunchu Zhang[a,b], Jianfei Dong[b,*]

[a] School of Biomedical Engineering (Suzhou), Division of Life Sciences and Medicine, University of Science and Technology of China, Suzhou 215163, China
[b] Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, Suzhou 215163, China

ARTICLE INFO

Article history:
Received 10 June 2021
Revised 22 February 2022
Accepted 5 March 2022
Available online 13 March 2022

Keywords:
2K-Fold-Net
EF^3-Net
U-Net
AFE
Image segmentation

ABSTRACT

For segmenting medical images, U-Net has become a popular and effective tool. However, it also has some shortcomings in segmenting fuzzy boundaries and eliminating interferences. Improvements of the original U-Net have been proposed by many authors, resulting in many variants such as MultiResUNet, DoubleU-Net and W-Net. Based on the common characteristics of these structures, we propose in this work a generalized structure by multiplying the folds of a fully convolutional network (FCN) for even more times, and thus name it as “2K-Fold-Net”. The more folds in this structure provide more freedoms to create cross links between the neighboring folds. The influence of the fold-pair number K on its performance is also studied. The realizations with K up to 6 are compared to three other variants of cascaded U-Nets using the CVC-ClinicDB dataset. Then the special case “4-Fold-Net” is further empowered with the feature enhancing functionalities recently seen in the attention-aware feature enhancement method. This new net is hence named as “Enhanced-Feature-4-Fold-Net”, abbreviated as “EF^3-Net”. Finally, 2K-Fold-Net and EF^3-Net have been compared with U-Net, SegNet, DoubleU-Net, MultiResUNet and its variants using four challenging medical image datasets. The results have demonstrated that the proposed nets outperform the other variants of U-Net, even with slightly lower amount of parameters. The code is available on: https://github.com/raik7/EF3-Net.

1. Introduction

Various modalities of medical imaging provide important bases for diagnosis [1], automatic early screening [2], treatment response prediction [3], lesion localization [4], and surgical navigation [5]. Medical images generally suffer from low contrast, blur boundaries, variable tissue characteristics and complex distribution of fine structures, which bring tremendous challenges.

In recent years, the applications of deep learning methods in image processing has been developing rapidly. Via a data-driven approach, a series of nonlinear mode transformations are used to automatically extract multi-layer features from data. Especially for medical imaging, deep learning based techniques can greatly improve the diagnosis efficiency [6].

Since fully convolutional networks (FCNs) [7] were brought into the public, end-to-end image segmentation research has begun to prosper. These networks are usually in the Encoder-Decoder structure [8]. The encoders extract increasingly complex features from the inputs [9], the decoders reconstruct a pixel level segmentation mask based on the extracted features. Among them, U-Net [10] is one of the most popular network. It can be regarded as folding an FCN once and connecting its sub-blocks facing each other across the folded “U” curve, which creates the so-called “skip connections”. Such a connection is the most obvious feature of U-Net [11]. Inspired by the U-Net, we will propose in this work a novel structure by folding an FCN for even more times to get $2K$ folds, and connecting the sub-blocks facing each other between two neighboring sub-U-Nets. This network is thus named as 2K-Fold-Net, and can be regarded as a generalization of the U-Net and its recently emerging variants, such as DoubleU-Net [12] and W-Net [13].

There are two main contributions in this work. Firstly, 2K-Fold-Net is proposed to improve the performance of the U-Net without increasing parameters. The influence of the fold-pair number K on its performance is also studied, by testing and comparing it with three other variants of U-Net using the CVC-ClinicDB dataset consisting of 612 polyp images [14]. Secondly, a specific realization with $K = 2$ is developed, and further empowered with the attention-aware feature enhancement (AFE), channel-wise feature enhancement (CFE) and spatial-wise feature enhancement (SFE) modules recently devised in [15]. This net is hence named as Enhanced-Feature-4-Fold-Net, or in short EF^3-Net. It is then tested

[*] Corresponding author.
E-mail addresses: jfeidong@hotmail.com, dongjf@sibet.ac.cn (J. Dong).

https://doi.org/10.1016/j.patcog.2022.108625
0031-3203/© 2022 Elsevier Ltd. All rights reserved.

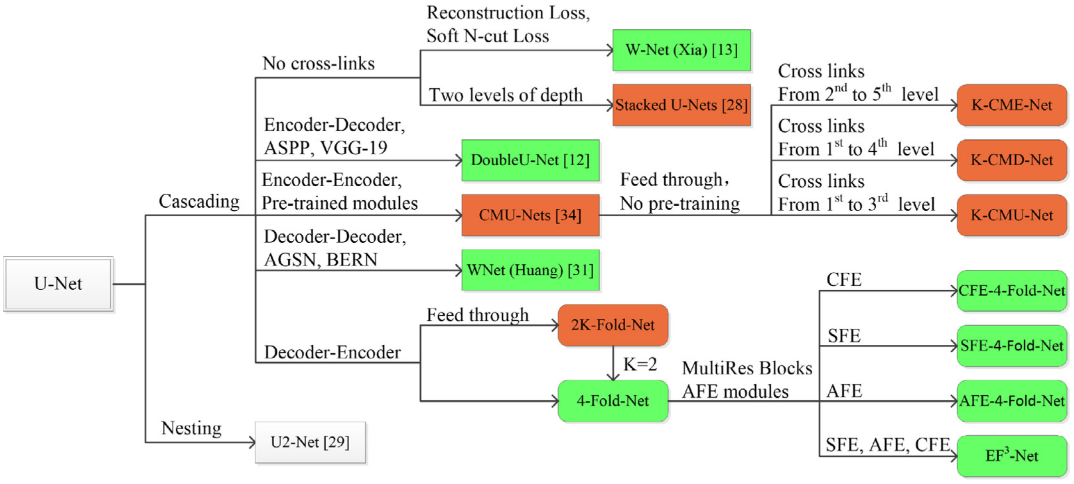


Fig. 1. Structural differences among the variants of U-Net. Green: models with two sub-U-nets; Orange: models with more than two sub-U-nets; Rectangles: the nets proposed by other authors; Rounded rectangles: the variants devised in this work.

and compared with U-Net, SegNet, DoubleU-Net, MultiResUNet and its variants using four challenging medical image segmentation datasets, including CVC-ClinicDB, the ISBI-2012: 2D EM segmentation challenge dataset (ISBI-2012) [16,17] containing 30 section images of the Drosophila first instar larva ventral nerve cord, the ISIC-2018: Lesion Boundary Segmentation challenge dataset (ISIC-2018) consisting of 2594 dermoscopic images of Melanoma [18,19], and the Gland Segmentation in Colon Histology Images Challenge dataset (GlaS) containing 165 images of H&E stained histological sections of colorectal adenocarcinoma [20].

The rest of the paper is organized as follows. The recent progresses based on U-Net and its variants are introduced in Section 2. The 2K-Fold-Net and EF^3-Net are developed in Section 3. In Section 4, the experimental results of the proposed networks are presented and discussed. Section 5 concludes the paper.

2. Related works

2.1. Modifying the U-Net architecture

MultiResUNet [21] is one of the successful modifications of U-Net. It enables extracting different scales of features by stacking a series of 3×3 convolutional layers. A more innovative modification therein is the shortcut connection named “ResPath”, which matches the possible semantic gap between the corresponding encoder and decoder. A deformable U-Net has been proposed to segment retina vessels [22]. By adding trainable offsets to convolutional kernel grid locations, the receptive fields of the filters are able to adapt to the shapes of the retina vessels. Likewise, dense blocks are also used to handle the artifacts in photoacoustic images to improve the image quality [23]. Inception-Res and Dense-Inception modules have also been devised to increase the width and depth of the network, which result in a DENSE-INception U-net [24]. Besides, separable convolutional blocks have been introduced to the U-Net to reduce the computational complexity [25]. Furthermore, the consecutive pooling operations in the U-Net often lead to the loss of boundary features. Incorporating multi-scale input features can improve the sensitivity to the morphological changes [26].

Most of the aforementioned improvements to U-Net are based on feature extraction modules. Whereas, studies on enhancing its performance by further enriching the connections among the sub-blocks are still rare.

2.2. Stacking multiple U-Nets

Stacking multiple U-Net-like networks is another popular approach to improve the performance. These methods usually connect a series of U-Nets to form a K-cascading U-Net, where K denotes the number of sub-U-Nets. For instance, a DoubleU-Net [12] has been proposed with two U-Nets stacked on top of each other, and further improved by Atrous Spatial Pyramid Pooling and pre-trained VGG-19. “Stacked U-Nets” has been proposed for classification and segmentation by stacking u-net blocks. The depth of the u-net module is reduced to two to limit the number of parameters [27]. Moreover, a two-level nested U-structure network has been designed, whose encoders and decoders are also composed of U-structured modules [28]. The network depth can be further increased without significantly increasing the computation.

On the other hand, W-shaped networks have also been developed recently. For instance, a W-Net is proposed by concatenating two U-Nets into an autoencoder [13], and achieve good results in unsupervised image segmentation. In order to handle the problem of crowd counting, an independent decoding branch is added to the expanding stage of a U-Net [29], which also ends up with a W-shaped net. The outputs from the two branches are element-wise multiplied to generate the reinforced map.

All the aforementioned methods link multiple U-Nets in the fashion of encoder-encoder connections, and can thus extract another group of features from the same set of original features. However, the same features may be extracted twice, and can therefore reduce the efficiency of the network. On the other hand, decoder-decoder connections have also been utilized in the W-Net proposed in [30] to deliver the features in atlas-guided segmentation network to a boundary-enhanced refinement network.

The structural differences of the aforementioned networks are illustrated in Fig. 1; while their main limitations are summarized in Table 1.

3. Methods

3.1. 2K-Fold-Net

3.1.1. Motivations

The main objective of stacking two and even more U-Nets is to improve the performance in a simple and efficient way, but at the cost of multiplying the number of parameters. This may lead to unaffordable burdens on the memory and the computing power, and may also cause over-fitting.

Table 1
Drawbacks of various variations of U-Net. First half: the nets proposed by other authors; Second half: the variants devised in this work.

Network	Drawbacks
W-Net [13]	1) Lack of inter-module skip-connection, deterioration of the foreground. 2) Possible negative results for more complex medical images.
Stacked U-Nets [27]	1) Not utilizing cross links may result in a vanishing gradient problem. 2) Shallow depth, low generative capacity.
DoubleU-Net [12]	1) Involvement of many parameters and thus heavy models. 2) Specialized computational requirements [31].
CMU-Nets [32]	Inconvenient pre-training for a specific task.
WNet [30]	Inevitable registration errors [33].
U2-Net [28]	1) Class agnostic [34]. 2) High spatial complexity.
K-CME-Net	Repeated extraction of features.
K-CMD-Net	Repeated extraction of features.
K-CMU-Net	1) Repeated extraction of features. 2) Less focus on the higher level features.
CFE-4-Fold-Net	Not focusing on the spatial information of the feature maps.
SFE-4-Fold-Net	1) Not focusing on the channel information of the feature maps. 2) Insufficient feature maps channels.
AFE-4-Fold-Net	Insufficient feature map channels.
EF^3-Net	Requiring more RAM and training time compared to U-Net.

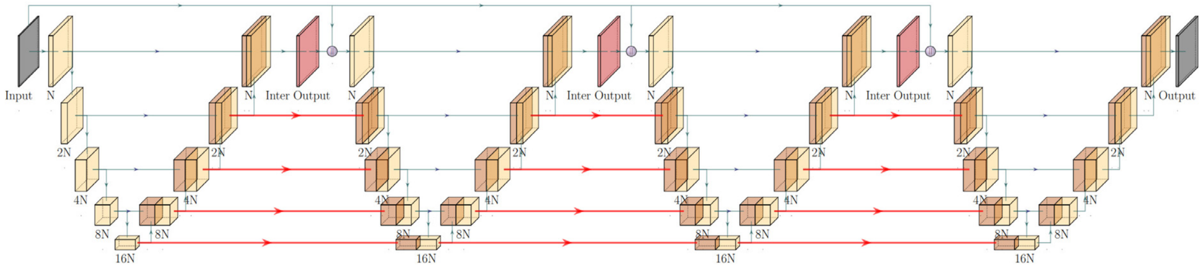


Fig. 2. Architecture of 2K-Fold-Net, with $K = 4$ as an example. "N" represents the number of channels of the first feature map extracted from the input image.

In this work, we attempt to explore the more freedoms in linking the neighboring folds. Specifically, the proposed structure is characterized by folding an FCN for K times, which results in K sub-U-Nets. Within each of these sub-U-Nets, the skip connections are made. Unlike the aforementioned variants of U-Nets, we also connect the blocks facing each other between two neighboring sub-U-Nets. To distinguish from the skip connections within each individual sub-U-Net, these new connections are called "cross links" in what follows. Such formulated networks are therefore called 2K-Fold-Net.

3.1.2. Architecture of 2K-Fold-Net

The architecture of a 2K-Fold-Net with the specific realization of $K = 4$ is depicted in Fig. 2. Inspired by the dense block [35], we concatenate the original input image and the output of each sub-U-Net to generate a new input image for its follower. In this way, all the individual sub-U-Nets have the access to the original image. Consequently, the deeper sub-U-Nets will *n*ot starve from the input information, and will *n*ot depend only on the features extracted by their leaders. The vanishing gradient problem is also avoided. Moreover, in the simulation experiments, feeding through the original input image to each sub-U-Net indeed helped improve the performance.

More specifically, suppose an FCN is folded for K times; and there are four skip connections in each of the K sub-U-Nets by default. Then, the number of the main elements in a 2K-Fold-Net can be counted as:

$$N_{sc} = 4K,\ N_{cl} = 4(K-1),\ N_{fb} = 9K, \qquad (1)$$

where N_{sc}, N_{cl}, N_{fb} respectively stand for the number of the skip connections, cross links and feature extracting blocks in 2K-Fold-Net.

3.1.3. Cascading modular U-Nets

To validate the proposed 2K-Fold-Net, we will compare it with a similar structure called Cascading Modular U-Nets (CMU-Nets) [32].

The original form of CMU-Net also suffers from the aforementioned starvation from the original input image, as K increases. Therefore, for a fair comparison, we have modified the CMU-Net by feeding through the original input to each subsequent sub-U-Net, as shown in Fig. 3(a).

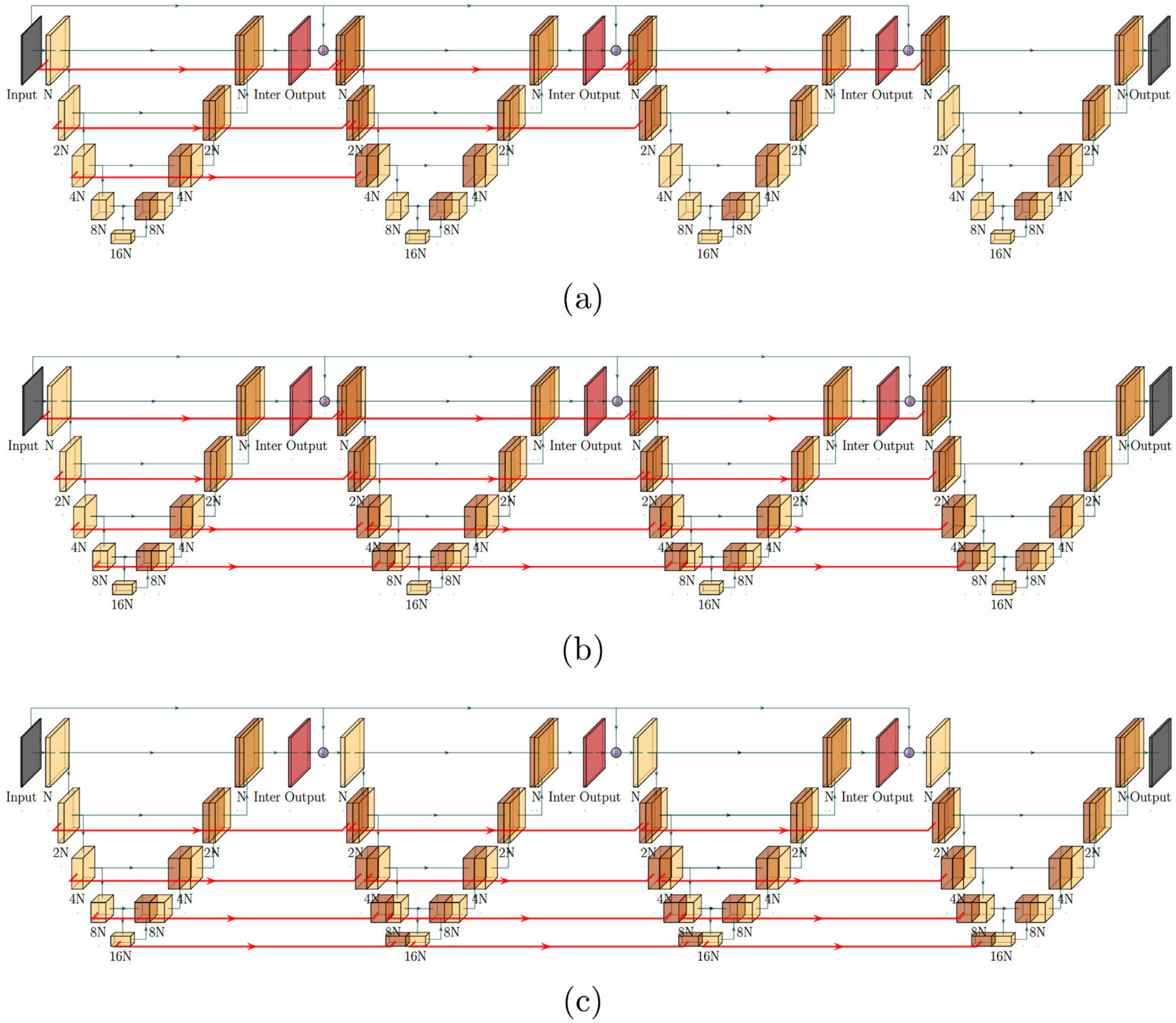
Furthermore, we have also developed a variant of the CMU-Net for better comparisons. Unlike the gradual decrease in the number of cross links between the encoding modules in the original CMU-Net, a network without such a reduction is developed under the name of Cascading Modular U-Net with Dense cross links, abbreviated as K-CMD-Net and depicted in Fig. 3(b).

Besides, to demonstrate the disadvantage of the encoder-encoder connection, we have developed a variant of 2K-Fold-Net, where the cross links are from the encoding modules of the leader to the corresponding encoders of its follower, as depicted in Fig. 3(c). This network is named as Cascading Modular U-Net with Encoder-encoder cross links, abbreviated as K-CME-Net.

Note that in the names, K-CMU-Net, K-CMD-Net and K-CME-Net, the prefix K is added to specify the number of sub-U-Nets contained therein. The major differences between them and 2K-Fold-Net are highlighted by the thick red lines and arrows in Figs. 2 and 3, and summarized in Table 2.

3.2. EF^3-Net as a special realization with K=2

Based on the 2K-Fold-Net architecture above, we further propose a special realization with $K = 2$. Besides, this network is fur-



(a)

(b)

(c)

Fig. 3. Three variants of CMU-Net: (a) K-CMU-Net, (b) K-CMD-Net, (c) K-CME-Net. The differences among the three variants of CMU-Net and 2K-Fold-Net are highlighted in red arrows. "N" represents the number of channels of the first feature map extracted from the input image. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
The differences in cross links between the original and three variants of CMU-Nets and 2K-Fold-Net, where "D-E, E-E" respectively stand for Decoder-Encoder and Encoder-Encoder.

Network	Type	Number	Levels	Feed through
2K-Fold-Net	D-E	4	from 2nd to 5th	Yes
K-CME-Net	E-E	4	from 2nd to 5th	Yes
K-CMD-Net	E-E	4	from 1st to 4th	Yes
K-CMU-Net	E-E	gradual decrease from 3	from 1st to 3rd	Yes
Original CMU-Nets [32]	E-E	gradual decrease from 3	from 1st to 3rd	No

ther equipped with more feature enhancing modules to improve its performance.

3.2.1. Motivations

The MultiRes block, as proposed in MultiResUNet [21] and shown in Fig. 4, can improve the performance effectively over the classic U-Net. Compared to the sequence of two convolutional layers in U-Net, a MultiRes block is comparatively lightweight, and can learn features of larger scales.

The ResPath in MultiResUNet aims at matching the possible disparity of feature maps between the corresponding layers of the contracting stage and expanding stage.

Spatial information is essential for accurate boundaries extraction; while some channels in the deep feature maps are more important to the final labeling process. Giving higher weights to the specific parts of the feature maps can enhance the specific features in the expanding stage.

3.2.2. AFE architectures

The architecture of an AFE module and its sub-modules are depicted in Fig. 5. In the CFE module, the $(W \times H \times C)$-dimensional input features are firstly processed in a global average pooling layer to set the initial weights of each channel, and become a $(1 \times 1 \times C)$-dimensional tensor. The initial weights then go through two fully-connected layers respectively with ReLu and sigmoid activation functions. The obtained weights are multiplied channel-wise with the original input. These procedures can be mathemati-

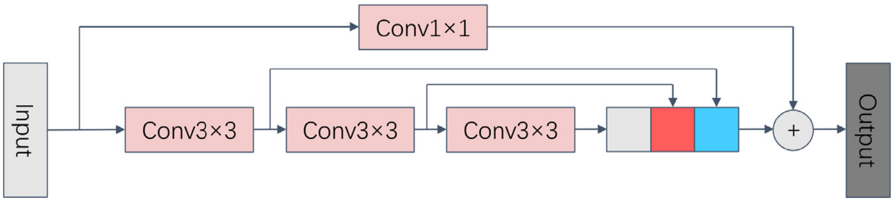


Fig. 4. The MultiRes block architecture.

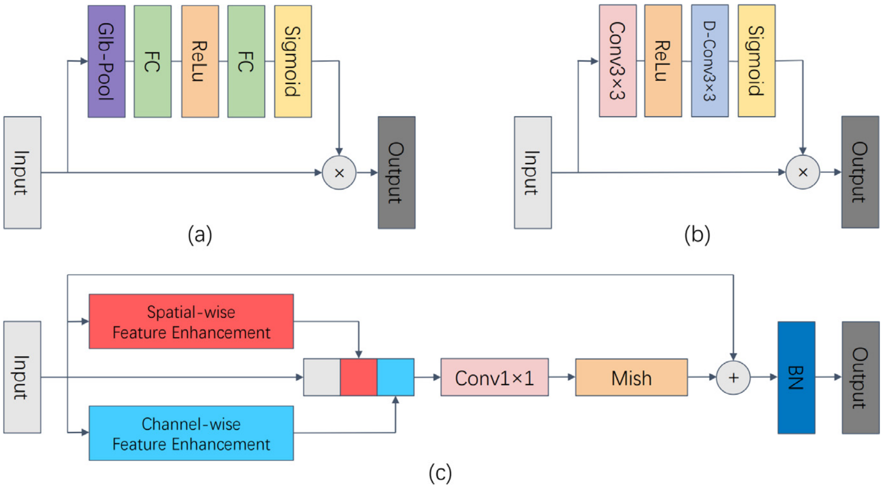


Fig. 5. Architecture of AFE related modules: (a) CFE, (b) SFE and (c) AFE.

cally represented as follows [15].

$$\boldsymbol{Y}_{CFE} = \boldsymbol{U} \otimes \alpha_{sig}\{\mathrm{FC}[\alpha_{ReLu}(\mathrm{FC}(\mathrm{AvePool}(\boldsymbol{U})))]\}. \tag{2}$$

Here, $\boldsymbol{Y}_{CFE}$ and $\boldsymbol{U}$ are respectively the output and input. $\mathrm{AvePool}(\cdot)$ and $\mathrm{FC}(\cdot)$ respectively represent the global average-pooling and fully-connected layer. $\alpha_{sig}(\cdot)$ and $\alpha_{ReLu}(\cdot)$ are respectively the sigmoid and ReLu activation functions. The operator "$\otimes$" stands for channel-wise multiplication.

In the SFE module, the input feature maps firstly go through a 3×3 convolutional layer and a depth-wise convolutional layer respectively with ReLu and sigmoid activation. Then the results are multiplied element-wise with the original input features. Therefore, one can write

$$\boldsymbol{Y}_{SFE} = \boldsymbol{U} \circledast \alpha_{sig}\{\mathrm{DConv}_{3\times3}[\alpha_{ReLu}(\mathrm{Conv}_{3\times3}(\boldsymbol{U}))]\}. \tag{3}$$

Here, $\mathrm{Conv}_{3\times3}(\cdot)$ and $\mathrm{DConv}_{3\times3}(\cdot)$ respectively represent the traditional and depth-wise convolutional layer with the kernel size of 3×3. The operator "$\circledast$" stands for pixel-wise multiplication.

An AFE module consists of both a CFE and a SFE module. The features of the two submodules are concatenated with the original input. Then, a 1×1 convolutional layer with mish activation function is adopted to mix up the information. Residual connections are also applied to prevent gradient vanishing. The AFE module can finally be represented as follows.

$$\boldsymbol{Y}_{AFE} = \mathrm{BN}\{\boldsymbol{U} \oplus \alpha_{Mish}[\mathrm{Conv}_{1\times1}(\mathrm{Conc}(\boldsymbol{U}, f_{SFE}(\boldsymbol{U}), f_{CFE}(\boldsymbol{U})))]\}. \tag{4}$$

Here, $\mathrm{Conc}(\cdot,\cdot,\cdot)$ represents the concatenation operation. $Conv_{1\times1}(\cdot)$ stands for a 1×1 convolutional layer. The operators "BN" and "$\oplus$" are respectively batch normalization and pixel-wise addition. The ReLu activation function in the original AFE module is replaced by the "mish" function [36], since it can preserve small negative outputs and make the optimization more effective.

3.2.3. Architecture of EF^3-Net

For further improvement, we first incorporate the feature enhancement modules into a MultiResUNet in two ways. The first method replaces all the ResPaths in the MultiResUNet with the AFE modules. It is hence named as AFE-MRUNet. The second method is designed to take advantage of a mix of AFE, CFE and SFE modules. It is hence named as Enhanced-Feature-MRUNet, or EF-MRUNet. Its structure is explained in details as follows.

For brevity, we shall call the four skip connections from top to bottom in a MultiResUNet consecutively the first level to the fourth level. The SFE module is installed at the first level due to the fewer channels therein. The whole AFE modules are incorporated into the second and third level because of the relatively more complex spatial and channel features therein. The CFE module is applied at the fourth level for two reasons. Firstly, at this level, both the height and width of the feature maps are reduced to 1/8 of those of the original input image. Enhancing the spatial features is not necessary. Secondly, for a feature map with n channels, adopting CFE, SFE and AFE modules respectively requires $2n^2$, $9n^2 + 11n$ and $14n^2 + 12n$ parameters. Therefore, it is not cost-effective to apply SFE and AFE at this deepest level with 8 times of channels compared to the first level.

Note that EF-MRUNet contains a single U structure. To further improve its performance, we use the trick of 2K-Fold-Net, and clone the single EF-MRUNet to two of them. The architecture actually corresponds to a special realization of 2K-Fold-Net with $K = 2$ and the mixed feature enhancement modules. Therefore, it is named as Enhanced-Feature-4-Fold-Net, or in short EF^3-Net. Here we selecte $K = 2$ to meet a balance of performance and computational complexity, which will be detailed in Section 4.5. This new net also inherits the other characteristics of the generic 2K-Fold-Net. Besides the applied feature enhancement modules in a single sub-EF-MRUNet, they are also installed in the four cross links between the two sub-EF-MRUNets. Specifically, from the top to the bottom, AFE, AFE, CFE and CFE modules are consecutively added. Moreover, the original input image is concatenated with the output of the first sub-EF-MRUNet and fed to the second one. The channel number of the intermediate output (abbreviated as inter output in Fig. 6) is set to equal to that of the original input. The

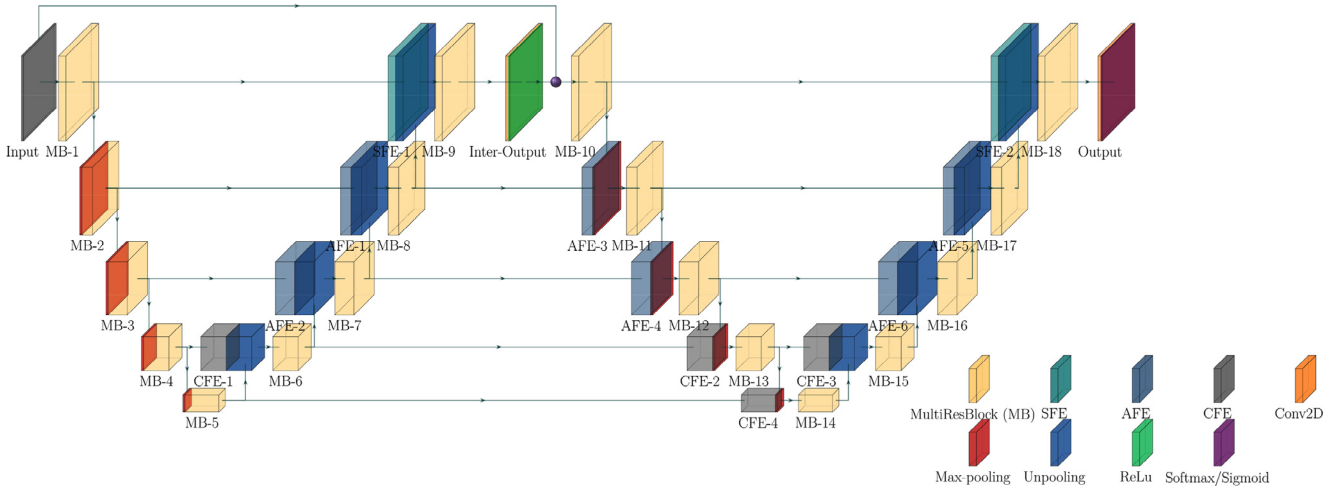


Fig. 6. Architecture of EF^3-Net.

Table 3
Overview of the datasets.

Dataset	Modality	No. of images	No. of classes	Original resolution	Input resolution
ISBI-2012	Electron microscopy	30	2	512 × 512	256 × 256
CVC-ClinicDB	Endoscopy	612	2	384 × 288	256 × 192
ISIC-2018	Dermoscopy	2594	2	Variable	256 × 192
GlaS	Microscopy	165	2	Variable	256 × 192

detailed architecture is finally illustrated in Fig. 6 and described in the Supplementary Materials.

To demonstrate the advantages of using the mixed feature enhancement modules in EF^3-Net, we also create three other variants purely with a single type of respectively the CFE, SFE and AFE modules, which are thus called CFE-4-Fold-Net, SFE-4-Fold-Net and AFE-4-Fold-Net. The differences among the total numbers of parameters in EF^3-Net and its variants are kept within $\pm 0.5\%$ by limiting the extracted feature map channels, since the CFE, SFE and AFE modules actually require different amounts of parameters. More specifically, when compared with EF^3-Net as the benchmark, CFE-4-Fold-Net, SFE-4-Fold-Net and AFE-4-Fold-Net respectively require 103%, 84% and 75% of the total channel numbers of the feature maps fed to the feature enhancement modules.

The relationships among all the aforementioned variants of U-Nets are illustrated in Fig. 1. Moreover, in Section 4.4, we will discuss the pros and cons of the different combinations of the feature enhancement modules. Also, we will compare the proposed EF^3-Net with its three variants mentioned above, 2K-Fold-Net (with $K = 3$ for the reason to be detailed in Section 4.2), and the original forms of U-Net, SegNet and MultiResUNet.

4. Experiments and results

In this section, the testing results and comparisons of the proposed networks and some other variants of U-Nets are presented, based on four challenging medical image datasets, as detailed in Table 3.

4.1. Experimental settings

All the models were implemented in Keras with Tensorflow 2.0 backend, on a desktop with an AMD R7 3700X CPU, 64GB RAM, and a NVIDIA GTX 1080Ti GPU. To train all the networks, we used the Adam optimizer with default settings (learning rate $= 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, decay $= 0$) and the binary cross-entropy loss function to train all the models.

5-fold cross-validation tests were executed to estimate the overall performance of these models. In these tests, the datasets were randomly split into 5 equal or nearly equal subsets. The models were trained and evaluated for 5 rounds by selecting in turn four different subsets as the training set and the rest as the validation set. In each round, the models were trained for 150 epochs with the batch size of 8, and evaluated in terms of six metrics after each epoch. The best results in these epochs were recorded. Finally, the average of every metric in the 5-fold cross-validation tests was calculated. The "He_normal" initializer was used to set the initial random weights.

4.2. Experiments of 2K-Fold-Net

We first tested 2K-Fold-Net and the three variants of cascading modular U-Nets described in Section 3.1.3, i.e., K-CMU-Net, K-CMD-Net and K-CME-Net, on a widely used and challenging dataset, CVC-ClinicDB. To study the effects of the fold-pair/sub-U-Net number K on the performances of these networks, its value was taken from $K = 1$ to $K = 6$. However, since for $K > 4$, there were no more cross links in K-CMU-Net to be further decreased, its simulation stopped at $K = 4$.

It shall be noted that for fair comparison, training K-CMU-Net did not follow the pre-training procedures of each sub-U-Net suggested in the original work of CMU-Nets [32], but followed the same end-to-end network training procedures as those in training the other three networks.

In the experiments, the metric of mean intersection over union (mIoU) was used to evaluate the segmentation accuracy, for its attribute of emphasizing the precision and penalizing the mistakes; i.e., mIoU $= \frac{A \cap B}{A \cup B}$. Here, A, B are respectively the set of ground truth and segmentation results.

Since the objective of this work is *not* to improve the performance at the cost of increased amount of parameters, the amount

Table 4
Comparisons of the four networks using the CVC-ClinicDB dataset in terms of the mIoU resulted from 5-fold cross-validation tests.

	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$
2K-Fold-Net	0.769159	**0.829555**	**0.838101**	**0.834577**	**0.833272**	**0.825715**
K-CME-Net	0.769159	0.805996	0.835040	0.833810	0.826601	0.823741
K-CMD-Net	0.769159	0.809990	0.824484	0.819861	0.817198	0.800350
K-CMU-Net	0.769159	0.817343	0.815860	0.802880	N.A.	N.A.

Table 5
Number of parameters (denoted as "N_p") and the time cost of training and predicition (denoted as "T_t" and "T_p" respectively) with the unit of millisecond per image (abbreviated as ms/img) of the compared networks.

Network	N_p	T_t	T_p	Network	N_p	T_t	T_p
SegNet	29.5M	27.5	9.63	U-Net	7.76M	12.9	3.50
MultiResUNet	7.26M [21]	35.0	9.38	4-Fold-Net	7.74M	15.1	3.75
AFE-MRUNet	7.24M	41.1	9.50	6-Fold-Net	7.77M	18.6	4.50
EF-MRUNet	7.16M	40.9	9.25	8-Fold-Net	7.75M	21.5	5.13
CFE-4-Fold-Net	7.24M	35.6	9.75	10-Fold-Net	7.76M	23.6	5.63
SFE-4-Fold-Net	7.11M	49.5	9.63	12-Fold-Net	7.76M	25.0	6.00
AFE-4-Fold-Net	7.25M	56.5	11.0	EF^3-Net	7.11M	60.9	12.5
DoubleU-Net	29.3M [12]	N.A.	N.A.				

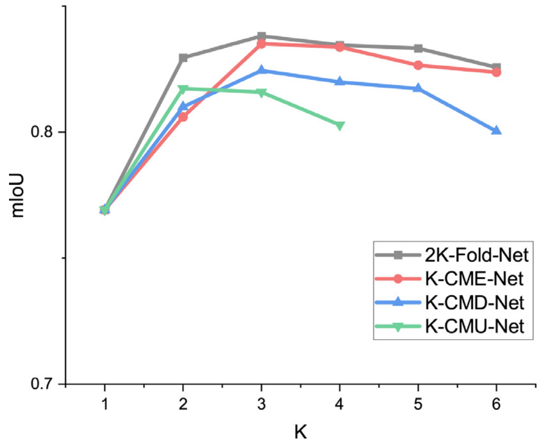


Fig. 7. mIoU v.s. the fold-pair/sub-U-Net number K in the four networks tested on CVC-ClinicDB.

of parameters of 2K-Fold-Net, K-CME-Net, K-CMD-Net and K-CMU-Net were kept in the range of $7.75M \pm 0.5\%$ by limiting the number of filters in each layer when taking different K values.

The 5-fold cross-validation results of the mIoU metric for the four networks using the CVC-ClinicDB images are listed in Table 4 and plotted in Fig. 7. Since the standard deviations are very small and range from 0.017 to 0.027 in all the cases, the error bars are not shown in the figure. From the results, all the four methods were able to significantly improve the segmentation accuracy by increasing K from 1 to 2. For $K = 3$, all but K-CME-Net could still make a slight improvement. However, further increasing K from 3 led to a trend of performance degradation. This clearly indicates that with limited amounts of parameters, the performance of these networks cannot be further improved by simply folding an FCN for more than three times. Among all the four methods, 2K-Fold-Net outperformed the others for all $K > 1$. Actually when $K = 1$, all these four nets boil down to the original U-Net, and thus generate the same results.

The comparison between 2K-Fold-Net and K-CME-Net shows that the decoder-encoder cross links outperform the encoder-encoder links. One reason can be attributed to the repeated extraction of the features. In K-CME-Net, the features extracted by an encoding module in the leading sub-U-Net are both fed into its own decoding module and the encoding module of its follower. In 2K-Fold-Net, the features extracted by an encoding module in the leading sub-U-Net have to be first processed by its own decoding module, and are then fed into the encoding module of its follower.

On the other hand, K-CMU-Net performed worst among the four for $K \geq 2$. This may be due to the gradually decreasing number of cross links, which actually puts more focus on the lower level features that does not help much to classify the labels of each pixel.

4.3. Experiments of EF^3-Net

The settings of the 5-fold cross-validation experiments were consistent with those mentioned in Section 4.1. We limited the number of parameters of the proposed models to be lower than that of the MultiResUNet (7.26M). The numbers of parameters of all the compared networks are listed in Table 5. In addition to the main metric mIoU, several other popular metrics were also used to evaluate the performance, including Dice coefficient (Dice), Accuracy (Accu.), Precision (Prec.), Sensitivity (Sens.), and Specificity (Spec.).

The results tested by the four datasets are respectively presented in Tables 6. Box plots of the mIoU for the four datasets are presented in Fig. 9. The results clearly demonstrated that EF^3-Net outperformed all the other networks in terms of almost all the metrics on the CVC-ClinicDB, ISIC-2018 and GlaS datasets. Judging from the box plots, the overall performance of EF^3-Net in the 5-fold cross-validation tests was also the best, as it resulted in higher median lines and higher upper and lower limits. Moreover, the relatively small boxes reflected the better generalization performance of EF^3-Net. Whereas, 6-Fold-Net performed best on ISBI-2012.

The largest differences in the performance metrics were found between EF^3-Net and U-Net or MultiResUNet on all the four types of medical images. For instance, as tested by the CVC-ClincDB dataset, the difference in the achieved mIoU between EF^3-Net and MultiResUNet was about 6.85%. As tested by the GlaS dataset, the difference in the achieved mIoU between EF^3-Net and U-Net was about 15.40%. It is worth emphasizing again that these improvements were *not* achieved by increasing the number of parameters in EF^3-Net, which was instead even less than those of the others.

To better demonstrate the results, the representative segmented images by EF^3-Net, U-Net and MultiResUNet are illustrated in Fig. 8 for all the testing cases. Obviously, the segmented masks by U-Net

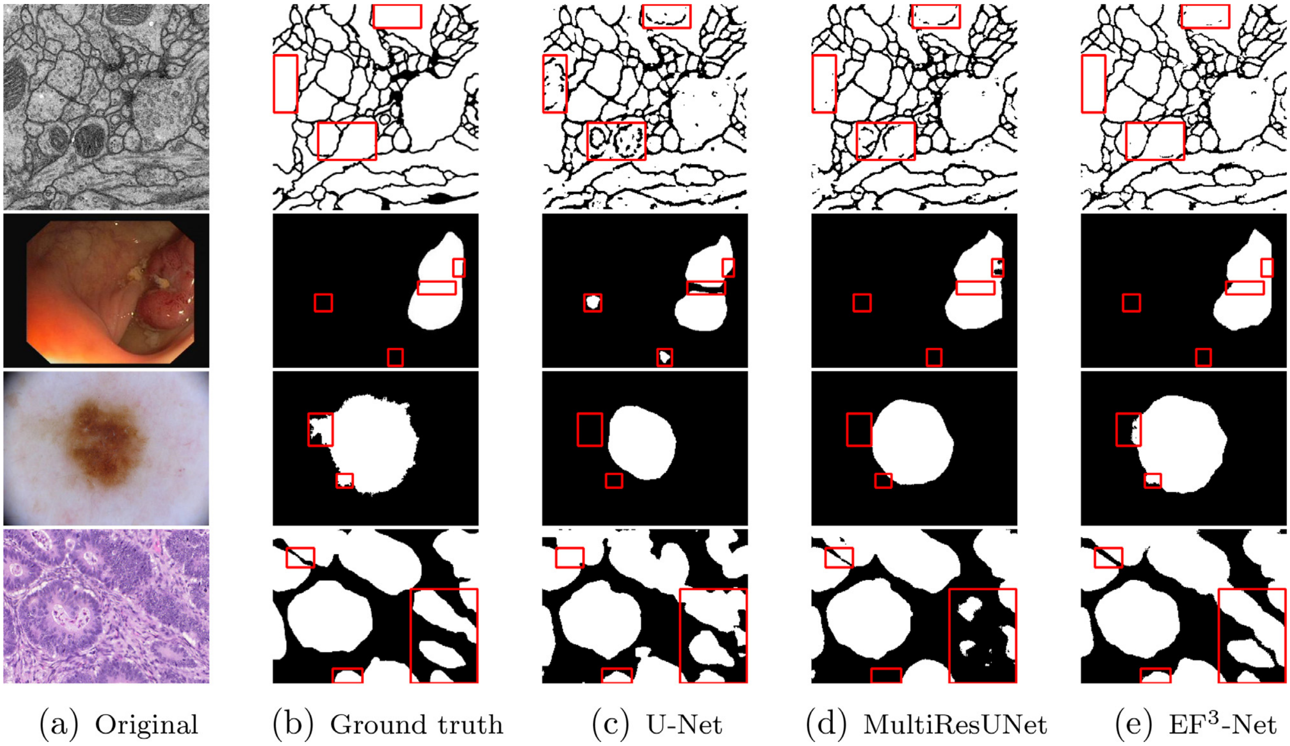

(a) Original (b) Ground truth (c) U-Net (d) MultiResUNet (e) EF^3-Net

Fig. 8. Representative segmented masks by U-Net, MultiResUNet and EF^3-Net of the images in ISBI-2012 (first row), CVC-ClinicDB (second row), ISIC-2018 (third row) and GlaS (fourth row). The major differences are marked in the red boxes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

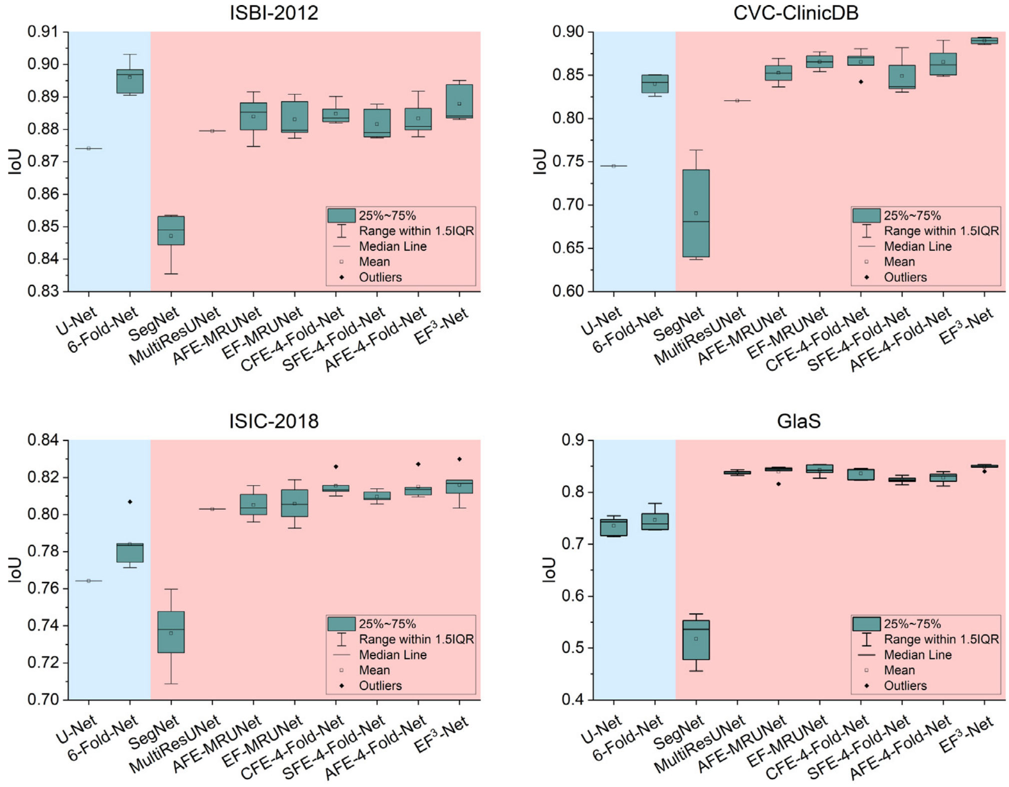


Fig. 9. Box plots of the mIoUs resulted from testing the ten nets on the four datasets.

Table 6
Testing results on the four datasets.

Testing results on the ISBI-2012 Electron Microscopy dataset						
Network	mIoU	Dice	Prec.	Accu.	Sens.	Spec.
SegNet	0.847111	0.917132	0.938991	0.872476	0.903361	0.735385
U-Net [21]	0.874092	N.A.	N.A.	N.A.	N.A.	N.A.
MultiResUNet [21]	0.879477	N.A.	N.A.	N.A.	N.A.	N.A.
AFE-MRUNet	0.883916	0.938371	0.927157	0.899754	0.958919	0.679610
EF-MRUNet	0.883078	0.937900	0.946788	0.899443	0.956052	0.699749
6-Fold-Net	**0.896000**	**0.945141**	**0.956054**	**0.912362**	0.953839	**0.834417**
CFE-4-Fold-Net	0.884834	0.938896	0.928880	0.901304	0.952733	0.713866
SFE-4-Fold-Net	0.881609	0.937074	0.925084	0.898095	0.954291	0.696833
AFE-4-Fold-Net	0.883314	0.938034	0.926900	0.899801	0.953695	0.705291
EF^3-Net	0.887898	0.940612	0.931363	0.903738	**0.959438**	0.703950
	Testing results on the CVC-ClinicDB dataset					
Network	mIoU	Dice	Prec.	Accu.	Sens.	Spec.
SegNet	0.544324	0.702758	0.904227	0.951431	0.831789	0.979960
U-Net [12]	0.7881	0.8781	0.9329	N.A.	0.7865	N.A.
MultiResUNet [21]	0.820574	N.A.	N.A.	N.A.	N.A.	N.A.
DoubleU-Net [12]	0.8611	0.9239	0.9592	N.A.	0.8457	N.A.
AFE-MRUNet	0.847932	0.917647	0.958461	0.985279	0.921063	0.996459
EF-MRUNet	0.857545	0.923214	0.954264	0.986251	0.924554	0.996008
6-Fold-Net	0.840110	0.913080	0.963199	0.984676	0.922316	**0.996912**
CFE-4-Fold-Net	0.865343	0.927760	0.956777	0.986955	0.944898	0.996871
SFE-4-Fold-Net	0.849011	0.918220	0.959332	0.985373	0.921377	0.996532
AFE-4-Fold-Net	0.865370	0.927750	0.952685	0.986914	0.931460	0.995595
EF^3-Net	**0.885437**	**0.939212**	**0.967354**	**0.988921**	**0.950661**	0.996832
	Testing results on the ISIC-2018 dataset					
Network	mIoU	Dice	Prec.	Accu.	Sens.	Spec.
SegNet	0.735965	0.847784	0.895583	0.937889	0.804280	0.973566
U-Net [21]	0.764277	N.A.	N.A.	N.A.	N.A.	N.A.
MultiResUNet [21]	0.802988	N.A.	N.A.	N.A.	N.A.	N.A.
DoubleU-Net [12]	**0.8212**	0.8962	0.9459	N.A.	0.8780	N.A.
AFE-MRUNet	0.805204	0.892074	0.963047	0.954721	0.918278	0.992169
EF-MRUNet	0.805844	0.892454	0.959033	0.955081	0.925962	0.991013
6-Fold-Net	0.784065	0.878910	**0.972083**	0.949356	0.908814	0.987010
CFE-4-Fold-Net	0.815495	0.886697	0.954404	0.957185	0.937583	0.991810
SFE-4-Fold-Net	0.809712	0.882620	0.946981	0.956238	0.931145	0.990971
AFE-4-Fold-Net	0.815151	0.886482	0.945580	0.956799	**0.939552**	0.988039
EF^3-Net	0.816758	**0.898419**	0.965763	**0.957476**	0.930548	**0.993189**
	Testing results on the GlaS dataset					
Network	mIoU	Dice	Prec.	Accu.	Sens.	Spec.
SegNet	0.515874	0.679390	0.693806	0.554987	0.870546	0.838109
U-Net	0.735441	0.847377	0.895587	0.843255	0.885492	0.875746
MultiResUNet	0.838332	0.912052	0.943251	0.912012	**0.942273**	0.949975
AFE-MRUNet	0.839579	0.912749	0.948386	0.912214	0.914908	0.950225
EF-MRUNet	0.843016	0.914792	0.947285	0.914442	0.932348	**0.950449**
6-Fold-Net	0.746711	0.854847	0.892287	0.852270	0.899896	0.893242
CFE-4-Fold-Net	0.836335	0.910840	0.933216	0.909852	0.919013	0.932070
SFE-4-Fold-Net	0.823929	0.903453	0.945538	0.903938	0.906569	0.906569
AFE-4-Fold-Net	0.828067	0.905915	0.942792	0.905594	0.916520	0.943144
EF^3-Net	**0.848710**	**0.918158**	**0.953192**	**0.917694**	0.925299	0.946949

and MultiResUNet contained a lot of under- or over-segmentation, which were absent in those resulted from EF^3-Net. More detailed comparisons are presented as follows.

In the ISBI-2012 task, U-Net and MultiResUNet incorrectly segmented the border of the nucleus into the profiles of neurites. EF^3-Net presented significantly fewer errors in distinguishing these two elements. This can be attributed to the CFE modules that can improve the labeling accuracy.

In the CVC-ClinicDB task, U-Net contained discontinuities and over-segmentation. MultiResUNet was disturbed by some local highlights and dark spots. Whereas, EF^3-Net showed better boundary tracking performance, and was more robust to these artifacts.

In the ISIC-2018 task, U-Net and MultiResUNet failed to segment the subtly different areas. In this case, the SFE modules helped improve the capability of EF^3-Net to distinguish these nuances.

In the GlaS task, U-Net and MultiResUNet respectively showed significant over- and under-segmentation. As for EF^3-Net, less false predictions were made, which demonstrated the capability of the AFE module to enhance the spatial and class information.

4.4. Comparisons of the eleven networks on the four datasets

We further compared EF^3-Net with the other networks listed in Table 5 and 2K-Fold-Net with $K = 3$ as the best choice demonstrated in Section 4.2. The following observations can be made.

Firstly, the comparison between U-Net and 6-Fold-Net verified that the 2K-Fold-Net structure can significantly improve the performance over U-Net in different tasks, because of the new decoder-encoder cross links between the sub-U-Nets and feeding through the original input image to each sub-U-Net.

Secondly, with only a quarter amount of parameters used by DoubleU-Net, EF^3-Net achieved better segmentation results on the CVC-ClinicDB dataset and comparative results on the ISIC-2018 dataset.

Thirdly, it can also be observed from the results of MultiResUNet and its two variants, i.e., AFE-MRUNet and EF-MRUNet, that the performance of the feature-enhanced MultiResUNets was improved to a certain extent, especially on the CVC-ClinicDB dataset, with the mIoU increased from 0.8206 to 0.8661 (5.54%). The overall performance of EF-MRUNet and AFE-MRUNet was not much different. In terms of the mIoU and Dice coefficient, EF-MRUNet showed slightly better performance on the CVC-ClincDB and GlaS dataset, and comparative performance on the ISBI-2012 and ISIC-2018 dataset. AFE-MRUNet showed a bit more advantage in dealing with less uniform images such as those in GlaS. This had actually motivated us to extend EF^3-Net based on the structure of EF-MRUNet.

Fourthly, comparing the results of EF^3-Net with those of EF-MRUNet, the mIoU values were improved respectively by 0.55%, 3.25%, 1.35%, 0.68% on the four datasets. Although the improvement in EF^3-Net was not remarkably big, it is still a meaningful advancement in medical diagnosis. Generally speaking, as the result of segmentation tends to be perfect, it becomes more and more difficult to further improve the performance. As can be seen from the box plots, EF^3-Net exhibited a smaller variance compared to EF-MRUNet, which reflected that extending EF-MRUNet to EF^3-Net improved the generalization performance of the network.

Finally, the performance of the other three "4-Fold-Net" was not as good as EF^3-Net. Among them, SFE-4-Fold-Net performed the worst, while CFE-4-Fold-Net and AFE-4-Fold-Net were not much different.

4.5. Discussion

The results listed above have demonstrated that the proposed 2K-Fold-Net architecture and the feature enhanced EF^3-Net are indeed able to improve the segmentation performance. The main observations are as follows.

Firstly, equipped with the feature enhancement modules, both EF-MRUNet and AFE-MRUNet outperformed the original MultiResUNet, indicating that enhancing specific parts of the feature maps is indeed helpful. On the other hand, comparing AFE-4-Fold-Net with CFE-4-Fold-Net, the presence of the SFE modules in the former net can achieve a comparable mIoU value with the latter, while reducing the channels of the extracted feature maps by 28%. Whereas, comparing AFE-4-Fold-Net with SFE-4-Fold-Net, the presence of the CFE modules can improve the mIoU from 0.19% to 1.93% on the four datasets, with a 9% reduction in the channel numbers.

Secondly, it is better to adopt different functional feature enhancement modules at suitable levels. Due to the structural complexity of the SFE and AFE module, with limited amount of parameters, applying more AFE modules will in turn cause a decrease in the amount of parameters assignable to the feature extraction modules, and thus fewer features can be extracted.

Thirdly, by comparing the results of EF^3-Net with those of EF-MRUNet, the capability of improving the performance by the 2K-Fold-Net structure has been proved, since EF^3-Net are extended from EF-MRUNet.

Fourthly, according to Table 5, every additional two folds in 2K-Fold-Net result in a 6% to 28% increase in training and prediction time. For a more complex architecture, i.e. EF-MRUNet and EF^3-Net, the time cost increases even more (48% and 35% respectively). We therefore selected $K = 2$ rather than $K = 3$, which showed better results according to Table 4, to meet a balance of performance and computational complexity. In fact, extending EF-MRUNet to 6 folds ($K = 3$) in the aforementioned method showed no significant improvement in performance in our pre-experiments.

5. Conclusions

We have proposed in this work a generalized structure of U-Net by folding an FCN for $K \geq 1$ times, which results in a 2K-Fold-Net. While the case of $K = 1$ becomes the original U-Net, bigger K values can further improve the image segmentation performance, as proved by the experimental results on the CVC-ClinicDB dataset. Then in the special case of $K = 2$, the 4-Fold-Net is further empowered with the attention-aware feature enhancement method. Tested by the four medical image datasets, the newly proposed EF^3-Net has demonstrated superior performance than U-Net, SegNet, DoubleU-Net, MultiResUNet and its variants. Especially, when compared to the similar structure of DoubleU-Net with three times more parameters than it, EF^3-Net was still able to perform slightly better in terms of the Dice coefficient and precision metrics.

The limitations of the proposed networks are as follows. Firstly, every increase of K by 1 in the 2K-Fold-Net also leads to 6% to 28% increase in the training and prediction time with similar amount of parameters. On the other hand, the testing results on the CVC-ClinicDB dataset also show that with limited amount of parameters, the performance of 2K-Fold-Net cannot be further improved by folding an FCN for more than three times. Secondly, training an EF^3-Net requires much more RAM than training a U-Net, when both networks contain a comparable number of parameters. As can be seen in Table 5, training EF^3-Net takes 3.7 and 2.3 times more time required to train U-Net and 6-Fold-Net respectively. The reason may be attributed to the fact that EF^3-Net possesses lower degree of parallelism and imposes heavier memory access cost (MAC) than those two nets.

The proposed networks are suitable in the following scenarios. Firstly, for the tasks whose executing time is not critical, the 2K-Fold-Net structure is an effective way to improve the performance of the variants of U-Net, in terms of better image segmentation performance using similar or even slightly lower amount of parameters. Secondly, with the demonstrated performance, EF^3-Net can be applied in the general and even more challenging medical image segmentation tasks where U-Net and its earlier variants have been applied.

Further simplifying the EF^3-Net architecture and improving its computational efficiency will be the most immediately potential extension of the current work. The recently proposed adaptive gradient clipping technique and the novel operator of Involution can be incorporated into the current architectures for further improvement. On the other hand, extending the proposed methods to segment 3D volumes is another task worth exploring. Finally, applying deep reinforcement learning to optimize the combination of cross links and feature enhancing modules into the proposed networks can also be a challenging and fruitful task to fulfill.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

J. Dong would like to thank the sponsorship of National Natural Science Foundation of China (Grant No. 61873263).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.patcog.2022.108625.

References

[1] J. Shi, X. Zheng, J. Wu, B. Gong, Q. Zhang, S. Ying, Quaternion Grassmann average network for learning representation of histopathological image, Pattern Recognit. 89 (2019) 67–76, doi:10.1016/j.patcog.2018.12.013.

[2] G. Yang, J. Cao, Z. Chen, J. Guo, J. Li, Graph-based neural networks for explainable image privacy inference, Pattern Recognit. 105 (2020) 107360, doi:10.1016/j.patcog.2020.107360.

[3] Y. Xu, A. Hosny, R. Zeleznik, C. Parmar, T. Coroller, I. Franco, R.H. Mak, H.J. Aerts, Deep learning predicts lung cancer treatment response from serial medical imaging, Clin. Cancer Res. 25 (11) (2019) 3266–3275, doi:10.1158/1078-0432.CCR-18-2495.

[4] A. Oulefki, S. Agaian, T. Trongtirakul, A.K. Laouar, Automatic COVID-19 lung infected region segmentation and measurement using CT-scans images, Pattern Recognit. 114 (2021) 107747, doi:10.1016/j.patcog.2020.107747.

[5] C. Nafis, V. Jensen, L. Beauregard, P. Anderson, Method for estimating dynamic EM tracking accuracy of surgical navigation tools, in: Medical Imaging 2006: Visualization, Image-Guided Procedures, and Display, vol. 6141, 2006, pp. 152–167, doi:10.1117/12.653448.

[6] N. Sharma, L.M. Aggarwal, Automated medical image segmentation techniques, J. Med. Phys. 35 (1) (2010) 3–14, doi:10.4103/0971-6203.58777.

[7] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

[8] A. Sinha, J. Dolz, Multi-scale self-guided attention for medical image segmentation, IEEE J. Biomed. Health Inform. 25 (1) (2021) 121–130, doi:10.1109/JBHI.2020.2986926.

[9] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, G.-Z. Yang, Deep learning for health informatics, IEEE J. Biomed. Health Inform. 21 (1) (2017) 4–21, doi:10.1109/JBHI.2016.2636665.

[10] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2015, pp. 234–241, doi:10.1007/978-3-319-24574-4_28.

[11] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. Van Der Laak, B. Van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88, doi:10.1016/j.media.2017.07.005.

[12] D. Jha, M.A. Riegler, D. Johansen, P. Halvorsen, H.D. Johansen, DoubleU-Net: a deep convolutional neural network for medical image segmentation, in: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), 2020, pp. 558–564, doi:10.1109/CBMS49503.2020.00111.

[13] X. Xia, B. Kulis, W-Net: a deep model for fully unsupervised image segmentation, arXiv preprint arXiv:1711.08506 (2017).

[14] J. Bernal, F.J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, F. Vilariño, WM-DOVA maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians, Comput. Med. Imaging Graph. 43 (2015) 99–111, doi:10.1016/j.compmedimag.2015.02.007.

[15] J. Ji, B. Zhong, K.-K. Ma, Image interpolation using multi-scale attention-aware inception network, IEEE Trans. Image Process. 29 (2020) 9413–9428, doi:10.1109/TIP.2020.3026632.

[16] A. Cardona, S. Saalfeld, S. Preibisch, B. Schmid, A. Cheng, J. Pulokas, P. Tomancak, V. Hartenstein, An integrated micro-and macroarchitectural analysis of the drosophila brain by computer-assisted serial section electron microscopy, PLoS Biol. 8 (10) (2010) e1000502.

[17] I. Arganda-Carreras, S.C. Turaga, D.R. Berger, D. Cireşan, A. Giusti, L.M. Gambardella, J. Schmidhuber, D. Laptev, S. Dwivedi, J.M. Buhmann, et al., Crowdsourcing the creation of image segmentation algorithms for connectomics, Front. Neuroanat. 9 (2015) 142, doi:10.3389/fnana.2015.00142.

[18] N. Codella, V. Rotemberg, P. Tschandl, M.E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, et al., Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the international skin imaging collaboration, arXiv preprint arXiv:1902.03368 (2019).

[19] P. Tschandl, C. Rosendahl, H. Kittler, The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, Sci. Data 5 (1) (2018) 1–9, doi:10.1038/sdata.2018.161.

[20] K. Sirinukunwattana, J.P.W. Pluim, H. Chen, X. Qi, P.-A. Heng, Y.B. Guo, L.Y. Wang, B.J. Matuszewski, E. Bruni, U. Sanchez, et al., Gland segmentation in colon histology images: the GlaS challenge contest, Med. Image Anal. 35 (2017) 489–502, doi:10.1016/j.media.2016.08.008.

[21] N. Ibtehaz, M.S. Rahman, MultiResUNet: rethinking the U-Net architecture for multimodal biomedical image segmentation, Neural Netw. 121 (2020) 74–87, doi:10.1016/j.neunet.2019.08.025.

[22] Q. Jin, Z. Meng, T.D. Pham, Q. Chen, L. Wei, R. Su, DUNet: a deformable network for retinal vessel segmentation, Knowl. Based Syst. 178 (2019) 149–162, doi:10.1016/j.knosys.2019.04.025.

[23] S. Guan, A.A. Khan, S. Sikdar, P.V. Chitnis, Fully dense UNet for 2-D sparse photoacoustic tomography artifact removal, IEEE J. Biomed. Health Inform. 24 (2) (2020) 568–576, doi:10.1109/JBHI.2019.2912935.

[24] Z. Zhang, C. Wu, S. Coleman, D. Kerr, Dense-inception U-net for medical image segmentation, Comput. Methods Programs Biomed. 192 (2020) 105395, doi:10.1016/j.cmpb.2020.105395.

[25] P. Tang, Q. Liang, X. Yan, S. Xiang, W. Sun, D. Zhang, G. Coppola, Efficient skin lesion segmentation using separable-Unet with stochastic weight averaging, Comput. Methods Programs Biomed. 178 (2019) 289–301, doi:10.1016/j.cmpb.2019.07.005.

[26] L. Wang, J. Gu, Y. Chen, Y. Liang, W. Zhang, J. Pu, H. Chen, Automated segmentation of the optic disc from fundus images using an asymmetric deep learning network, Pattern Recognit. 112 (2021) 107810, doi:10.1016/j.patcog.2020.107810.

[27] S. Shah, P. Ghosh, L.S. Davis, T. Goldstein, Stacked U-Nets: a no-frills approach to natural image segmentation, arXiv preprint arXiv:1804.10343 (2018).

[28] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O.R. Zaiane, M. Jagersand, U2-Net: going deeper with nested u-structure for salient object detection, Pattern Recognit. 106 (2020) 107404, doi:10.1016/j.patcog.2020.107404.

[29] V.K. Valloli, K. Mehta, W-Net: reinforced U-Net for density map estimation, arXiv preprint arXiv:1903.11249 (2019).

[30] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, J. Wu, WNET: an end-to-end atlas-guided and boundary-enhanced network for medical image segmentation, in: 2020 IEEE 17th International Symposium on Biomedical Imaging, 2020, pp. 763–766, doi:10.1109/ISBI45749.2020.9098654.

[31] N. Paluru, A. Dayal, H.B. Jenssen, T. Sakinis, L.R. Cenkeramaddi, J. Prakash, P.K. Yalavarthy, Anam-Net: anamorphic depth embedding-based lightweight CNN for segmentation of anomalies in COVID-19 chest CT images, IEEE Trans. Neural Netw. Learn. Syst. 32 (3) (2021) 932–946, doi:10.1109/TNNLS.2021.3054746.

[32] S. Kang, B.K. Iwana, S. Uchida, Complex image processing with less data-document image binarization by integrating multiple pre-trained U-Net modules, Pattern Recognit. 109 (2021) 107577, doi:10.1016/j.patcog.2020.107577.

[33] O.M. Benkarim, G. Piella, M.A.G. Ballester, G. Sanroma, A.D.N. Initiative, et al., Discriminative confidence estimation for probabilistic multi-atlas label fusion, Med. Image Anal. 42 (2017) 274–287, doi:10.1016/j.media.2017.08.008.

[34] A. Sauer, A. Geiger, Counterfactual generative networks, arXiv preprint arXiv:2101.06046 (2021).

[35] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2261–2269, doi:10.1109/CVPR.2017.243.

[36] D. Misra, Mish: a self regularized non-monotonic activation function, arXiv preprint arXiv:1908.08681 (2019).

Yunchu Zhang received the BEng degree in Optoelectronic Information Science and Engineering from Soochow University, Suzhou, China, in 2018. Since then, he has been a postgraduate student both with University of Science and Technology of China and with Suzhou Institute of Biomedical Engineering and Technology, where he is now pursuing the PhD degree. His main research interests include deep learning, convolutional neural networks, and medical image segmentation.

Jianfei Dong received the PhD degree in systems and control from the Delft University of Technology (TUD), The Netherlands, in 2009. From November 2009 to July 2011, he was a Postdoctoral Researcher with TUD. Then, he worked as a Research Scientist with Philips Research, Eindhoven, The Netherlands. He is currently a Professor with the Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, Suzhou, China. His current research interests focus on data-driven modeling and control methodologies for complex dynamic systems, including biological, optoelectronic and mechatronic systems.