GREnet: Gradually REcurrent Network With Curriculum Learning for 2-D Medical Image Segmentation

Jinting Wang, Yujiao Tang, Yang Xiao, Joey Tianyi Zhou, *Senior Member, IEEE*, Zhiwen Fang, and Feng Yang

***Abstract*— Medical image segmentation is a vital stage in medical image analysis. Numerous deep-learning methods are booming to improve the performance of 2-D medical image segmentation, owing to the fast growth of the convolutional neural network. Generally, the manually defined ground truth is utilized directly to supervise models in the training phase. However, direct supervision of the ground truth often results in ambiguity and distractors as complex challenges appear simultaneously. To alleviate this issue, we propose a gradually recurrent network with curriculum learning, which is supervised by gradual information of the ground truth. The whole model is composed of two independent networks. One is the segmentation network denoted as GREnet, which formulates 2-D medical image segmentation as a temporal task supervised by pixel-level gradual curricula in the training phase. The other is a curriculum-mining network. To a certain degree, the curriculum-mining network provides curricula with an increasing difficulty in the ground truth of the training set by progressively uncovering hard-to-segmentation pixels via a data-driven manner. Given that segmentation is a pixel-level dense-prediction challenge, to the best of our knowledge, this is the first work to function 2-D medical image segmentation as a temporal task with pixel-level curriculum learning. In GREnet, the naive UNet is adopted as the backbone, while ConvLSTM is used to establish the temporal link between gradual curricula. In the curriculum-mining network, UNet++ supplemented by transformer is designed to deliver curricula through the outputs of the modified UNet++ at different layers. Experimental results have demonstrated the effectiveness of GREnet on seven datasets, i.e., three lesion segmentation datasets in dermoscopic images, an optic disc and cup segmentation dataset and a blood vessel segmentation dataset in retinal images, a breast lesion segmentation dataset in ultrasound images, and a lung segmentation dataset in computed tomography (CT).**

***Index Terms*— Curriculum learning, data-driven curriculum, gradually recurrent network, medical image segmentation.**

I. Introduction

MEDICAL image segmentation is a critical step for medical analysis, including skin lesion segmentation for early detection of melanomas in dermoscopic images [1], [2], optic disc and cup segmentation and blood vessel segmentation for discovering structural information in retinal images [3], [4], breast lesion segmentation for auxiliary diagnosis in ultrasound images [5], [6], and lung segmentation for locating organs in computed tomography (CT) [7], [8]. In general, the aforementioned medical image segmentation can be performed manually by skilled physicians. Manual examination, on the other hand, is time-consuming, subjective, and labor-demanding owing to the complexity of medical images, such as *blurry boundaries, confusing regions, and shadow artifacts*. There is a great need for accurate and dependable techniques of computer-aided segmentation [9].

To meet the requirement of automated segmentation, researchers design segmentation models with care. Previous methods using handcraft features have often relied on edge detection [12], [13] and temple matching [14], [15]. Currently, owing to the strong representation ability of the convolutional neural network [16], [17], numerous approaches are continually refining the deep-learning-based models depending on the characteristics of different tasks [18], [19]. However, due to the complexity of 2-D medical image segmentation, it is difficult to obtain satisfactory performance while addressing different issues concurrently. Aiming to alleviate this problem, cascaded-based techniques with contextual guidance, which is inherited from the previous segmentation results, have been frequently employed [2], [20] to polish the outputs incrementally. Nevertheless, because of the inaccuracies in the preceding segmentation outputs, particularly those errors that occur beyond the expected segmentation region, blindly contextual guidance would always result in distractors. It implies that the

Manuscript received 22 April 2022; revised 16 October 2022 and 24 November 2022; accepted 16 January 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 61771233, Grant 61702182, and Grant 62271221; in part by the Science and Technology Program of Guangzhou under Grant 202201011672; in part by the Science and Engineering Research Council (SERC) Central Research Fund (Use-inspired Basic Research); and in part by the Singapore Government's Research, and Innovation and Enterprise 2020 Plan (Advanced Manufacturing and Engineering Domain) under Grant A18A1b0045. *(Corresponding authors: Zhiwen Fang; Feng Yang.)*

Jinting Wang, Yujiao Tang, Zhiwen Fang, and Feng Yang are with the School of Biomedical Engineering, Southern Medical University, Guangzhou 510515, China, also with the Guangdong Provincial Key Laboratory of Medical Image Processing, Southern Medical University, Guangzhou 510515, China, and also with the Guangdong Province Engineering Laboratory for Medical Imaging and Diagnostic Technology, Southern Medical University, Guangzhou 510515, China (e-mail: 3168010128@i.smu.edu.cn; yujiao.tang.smu@gmail.com; fzw310@smu.edu.cn; yangf@smu.edu.cn).

Yang Xiao is with the National Key Laboratory of Science and Technology on Multispectral Information Processing, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: yangxiao@hust.edu.cn).

Joey Tianyi Zhou is with the Centre for Frontier AI Research (CFAR), Research Agency for Science, Technology, and Research (A*STAR), Singapore 138632 (e-mail: zhouty@ihpc.a-star.edu.sg).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2023.3238381.

Digital Object Identifier 10.1109/TNNLS.2023.3238381

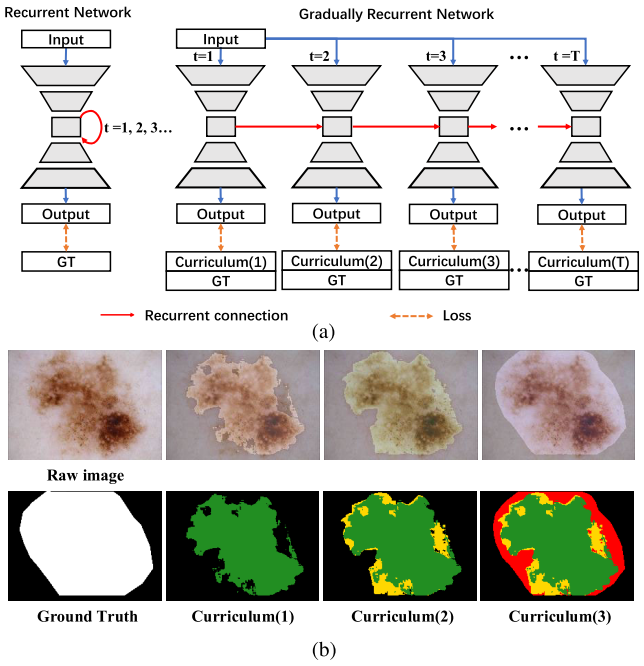


Fig. 1. Illustration of the gradually recurrent network with curriculum learning. The relationship among curricula is curriculum(1) $\subseteq$ curriculum(2) $\subseteq$ curriculum(3) $\subseteq \cdots \subseteq$ curriculum(T). (a) Frameworks of the recurrent network and the gradually recurrent network. (b) Examples of the gradual curricula of skin lesion segmentation on the ISBI2017 dataset. Curriculum(1): green region; Curriculum(2): green and yellow regions; Curriculum(3): green, yellow, and red regions.

subsequent segmentation model must correct the errors while identifying the missing portion of the preceding segmentation outputs. This issue will cause the segmentation model to get confused.

In this work, we innovatively function 2-D medical image segmentation as a temporal task with curriculum learning, which can purposefully provide progressive curriculum guidance shown as Fig. 1. A cognitive theory acknowledges that humans learn much better when different concepts are not randomly presented but organized at different times in a meaningful order, which illustrates gradually more concepts, and gradually more complex ones [21], [22]. Inspired by this theory, Bengio et al. [23] propose curriculum learning. It demonstrates that a curriculum with a "starting small" strategy can benefit machine learning algorithms. In 2-D medical image segmentation, scarce methods use curriculum learning. Two facts may be the primary challenge: 1) how to provide the curriculum at different time steps, i.e., a temporal task and 2) how to provide gradual concepts as pixel-level curricula since segmentation is a pixel-level dense-prediction work. Tang et al. [1], and Li et al. [24] make an attempt to provide sample-level curricula. However, because each sample contains both easy and hard pixels, the sample-level curriculum cannot discriminate gradual pixel-level concepts. Additionally, rather than performing a temporal task, their methods are learned using different curricula in independent training phases. But the gradual concepts in the inference phase are ignored, which will limit the generalization ability of curriculum learning when meeting new samples in the inference phase.

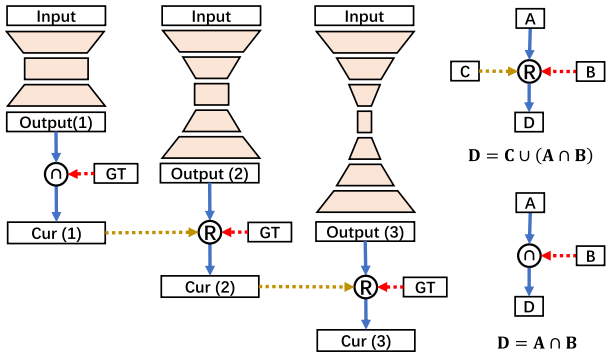


Fig. 2. Illustration of developing curricula via a data-driven strategy. For simplicity, curriculum(T) is denoted as Cur(T). The curriculum-mining network is distinct from the gradually recurrent network. Concretely, UNet++ [10] enhanced by Transformer [11] is adopted to offer the curricula through the outputs of different layers.

To address the two issues, we propose a gradually recurrent network to construct the temporal task for 2-D medical image segmentation and develop data-driven pixel-level curricula. As shown in Fig. 1(a), in contrast to the traditional recurrent networks, the gradually recurrent network will be supervised by gradual curricula. As seen in Fig. 1(b), the curriculum region for 2-D medical image segmentation progressive increases in size, i.e., curriculum(1) $\subseteq$ curriculum(2) $\subseteq$ curriculum(3) $\subseteq \cdots \subseteq$ curriculum(T). The final curriculum is almost the ground truth. To deliver these curricula, a data-driven strategy is introduced using deep-learning models of varying depths, shown in Fig. 2. To a certain extent, shallow models prefer to uncover easy-to-segment regions using low-level features, while deeper models may mine more hard-to-segment regions using high-level features. Visually shown as Fig. 1(b), the third curriculum includes more hard pixels than the first. Furthermore, statistical analyses on five published segmentation datasets for different medical images are given in Fig. 3. It can be observed that gradual curriculum increments of varying degrees occur in a variety of applications. Intuitively, three curricula are established for blood vessel segmentation in Fig. 3(a), skin lesion segmentation in Fig. 3(b), and breast lesion segmentation in Fig. 3(c), but two curricula including Cur(1) and Cur(3) are sufficient for Fig. 3(d) and (e). The number of curricula is determined by the incremental amplitude in statistical analyses, which will further be discussed in the experimental section.

The main contributions can be summarized as follows:

1) Given that segmentation is a pixel-level dense-prediction challenge, to the best of our knowledge, this is the first work to function 2-D medical image segmentation as a temporal task with pixel-level curriculum learning, resulting in generalizability across several applications.
2) We design a gradually recurrent network to learn progressive curricula via a temporal manner, which can purposefully provide progressive curriculum guidance.
3) Rather than manually define curricula, we introduce a data-driven strategy based on modified UNet++ to mine diverse curricula from the training set on different

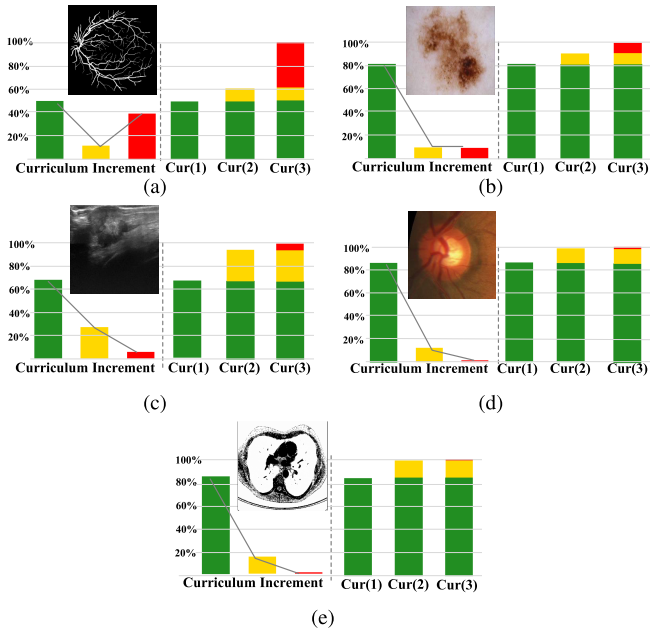


Fig. 3. Statistical analyses of the curriculum increment on five datasets. For simplicity, curriculum(T) is denoted as Cur(T). (a) Curriculum increment on the DRIVE dataset for blood vessel segmentation in retinal images. (b) Curriculum increment on the ISBI 2017 dataset for skin lesion segmentation in dermoscopic images. (c) Curriculum increment on the BUSI dataset for breast lesion segmentation in ultrasound images. (d) Curriculum increment on the REFUGE dataset for optic cup segmentation in retinal images. (e) Curriculum increment on the LUNA dataset for lung segmentation in CT.

datasets. The data-driven strategy will decrease the parameter sensitivity.

The source code of this work is published online.[1]

The remainder of this article is organized as follows. The related network is reviewed in Section II. Then the details of the gradually recurrent network are illustrated in Section III. Section IV illustrates the objective function. The experimental results and discussions are conducted in Section V. Section VI concludes this article.

II. Related Network

Great achievements in 2-D medical image segmentation have been made over the past decades. In this section, we will review four solutions related to our work: 1) segmentation methods based on re-weighting strategy; 2) segmentation methods based on cascading strategy; 3) segmentation methods based on knowledge distillation strategy; and 4) segmentation methods based on curriculum learning strategy.

A. Segmentation Methods Based on Re-Weighting Strategy

Generally, current re-weighting solutions modify popular loss functions, such as the cross entropy loss and Dice loss, for handling the issue of the coexistence of segmentation complexity variations. Focal loss [25], [26] is introduced to balance the distribution weights of different pixels, respectively. Chang et al. [27] employ a multi-class focal loss to emphasize misclassified voxels in magnetic resonance images. The Tversky index [28], [29], which is derived from the Dice similarity coefficient, weights false negatives and positives differently for lesion segmentation. In contrast to the preceding techniques that focus on the issue of data imbalance problems, the curriculum-based learning method stresses progressive learning, which results in a decreased sensitivity of weighted parameters for various applications.

B. Segmentation Methods Based on Cascading Strategy

Cascading strategy [2], [20], [30], [31], [32] initially conducts coarse segmentation and then use the coarse findings as prior information to guide the network to achieve fine segmentation. In [2] and [31], a succession of fully convolutional networks (FCN) is trained sequentially, with the original image concatenated with a degraded probability given by the previous classifier. Tang et al. [32] additionally explore the context included in coarse segmentation at the multi-scale level to improve the accuracy of the final results. In contrast, the studies in [30] and 36 transfer the coarse segmentation produced at the first stage to target localization information maps for further precise target segmentation at the next stage. Due to the coarse-to-fine technique, these cascaded methods generally outperform single-stage methods owing to the coarse-to-fine strategy. Nevertheless, because of the flaws in the preceding segmentation output, especially those errors that occur outside of the expected segmentation region, blind guidance would always result in distractors. This problem will confound the segmentation model.

C. Segmentation Methods Based on Knowledge Distillation Strategy

Knowledge distillation is a strategy for transferring knowledge from a teacher model to a student model by utilizing the outputs of the teacher model as the soft labels of the students [33]. Numerous studies [34], [35], [36], [37] have shown the utility of knowledge distillation in medical image segmentation. Eytan et al. [34] present a method that distills anatomical knowledge for improving lesion segmentation. The knowledge distillation technique is employed to exploit multi-modal image features in [35]. Dual teacher networks are applied to exploit intra- and inter-domain knowledge for cardiac segmentation in [36]. The purpose of knowledge distillation is to get the performance of the student network close to that of the teacher network. Nonetheless, the student network is not as good as the teacher network [37]. Different from the teacher network in knowledge distillation, the curriculum-mining network in our model is responsible for delivering curricula at different times in a meaningful order. The performance of the curriculum-mining network is not the upper bound on the performance of the gradually recurrent network.

D. Segmentation Methods Based on Curriculum Learning Strategy

Curriculum learning is proposed by Bengio [23], which is inspired by human learning patterns. The curriculum strategy suggests that the model training should begin with simple data and progress to harder data, hence improving the generalization performance. Inspired by the successful use of curriculum

[1]Code is available at http://github.com/beria-moon/GREnet

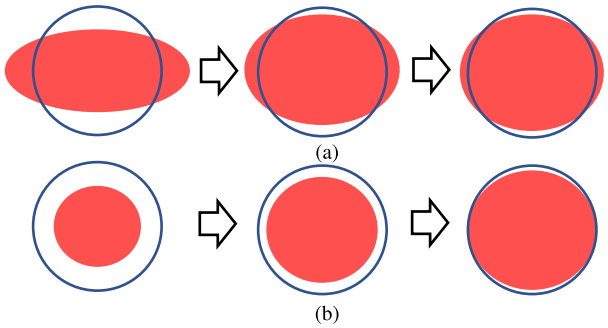


Fig. 4. Diagram of the evolution process of segmentation results. The blue circle represents the target segmentation region, and the red areas are the predicted results at different stages or time steps. (a) Cascaded methods. (b) Curriculum-learning-based methods.

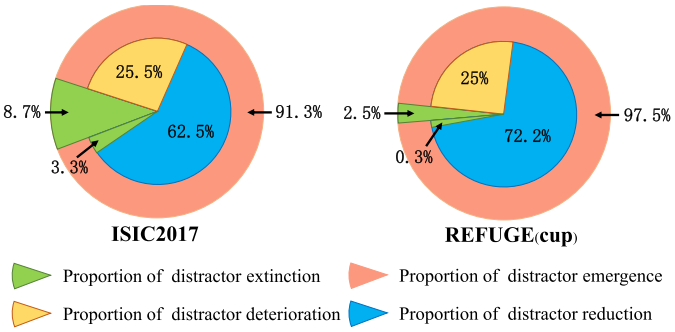


Fig. 5. Statistical analyses of the distractor risk of the traditional cascaded mechanism on the ISIC 2017 and REFUGE(cup) datasets. A cascaded model including two UNet [40] is adopted to illustrate the statistical analyses. The output of the first UNet is utilized as the contextual guidance input of the second UNet. The outer pie chart shows the proportion of the outputs of the first UNet, and the inner pie chart is for the second UNet. To analyze the distractor risk, the inner pie chart only covers the part of distractor emergence of the outer one. Distractor extinction and emergence represent the outputs with and without the distractors beyond the target regions, respectively. Distractor deterioration and reduction mean that the second UNet's output contains more and fewer distractors than that of the first one, separately.

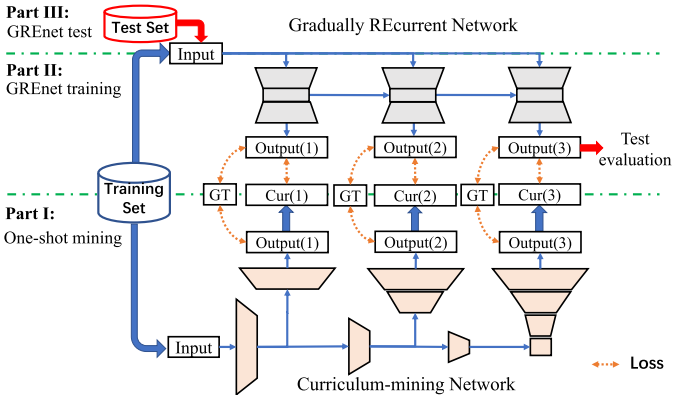


Fig. 6. Pipeline of the proposed method. First, one-shot curriculum mining; Second, GREnet training; Third, GREnet test. GREnet and the curriculum-mining network are trained independently rather than jointly. The curriculum mining technique is one-shot mining, which means that the pixel-level curricula are only mined once for different datasets. Thus, in the training and inference stages of GREnet, no extra time is necessary.

learning in classification [38] and mathematical calculation tasks [39], some researchers try to apply curriculum learning for medical image segmentation [1], [24]. A sample-level curriculum learning strategy is applied in [1] to address the sample-unbalance issue. They develop the curricula in accordance with the complexity of the samples included in the datasets. Li et al. [24] proposes a three-stage curriculum learning strategy for training deep networks to address the issue of class imbalance. These curriculum learning-style techniques concentrate on the issues that arise as a result of imbalance at the sample and class levels. However, they overlook the fact that segmentation is a pixel-level dense prediction task. It implies that easy-to-segmentation and difficult-to-segmentation pixels coexist in each 2-D medical image. Therefore, in this work, we offer a solution based on the pixel-level curriculum learning strategy for tackling the challenge of 2-D medical image segmentation.

III. Gradually Recurrent Network With Curriculum Learning

To accurately predict pixel-level values, it is critical for 2-D medical image segmentation methods to mine patterns from the target areas while avoiding being confused by the distractors beyond the target regions. To illustrate the potential risk caused by the aforementioned distractors, two examples and statistical analyses are provided in Figs. 4 and 5. It can be observed that: 1) more than 90% of the first outputs include distractors; 2) less than 5% of the first outputs can be entirely repaired; and 3) more than 25% are even worse in the second outputs. Thus, without progressive guidance, traditional cascaded methods [2], [20], [30] have a high risk of delivering distractors beyond the target segmentation region as the erroneous guidance to the next stage. Unfortunately, single-stage methods with the re-weighting and knowledge distillation strategies [25], [26], [34], [35] also lack efficient ways to mitigate this risk because internal hard-to-segmentation pixels and external distractors often reside around the boundary of the target segmentation regions.

To reduce this risk, curriculum learning [23] via providing gradual curricula shown as Fig. 4(b) is an optional strategy due to the gradually incremental guidance from the inside out. Inspired by the advantage of curriculum learning, we provide a gradually recurrent network with data-driven curricula to segment 2-D medical images. For simplicity, the network is denoted as GREnet. It functions 2-D medical image segmentation as a temporal curriculum-learning task via a ConvLSTM module and provides a data-driven strategy based on modified UNet++ to effectively offer progressive curricula for 2-D medical image segmentation.

The pipeline of the whole model is illustrated in Fig. 6. The curriculum mining technique is one-shot mining, which means that the pixel-level curricula are only mined once for different datasets. Then, the mined curricula are used to supervise the training phase of GREnet. The curriculum learning will be briefly summarized in Section III-A. The details of GREnet are introduced in Section III-B, and the data-driven curriculum is illustrated in Section III-C.

A. Curriculum Learning

The paradigm of curriculum learning is driven by the learning process of human beings, which advocates that a model starts with easier aspects of the task and then gradually

TABLE I
ENTROPY $E(\text{Cur}(t))$ OF THE tTH CURRICULUM ON FIVE DATASETS

Dataset	$E(Cur(1))$	$E(Cur(2))$	$E(Cur(3))$
DRIVE	0.08	0.19	0.35
ISIC 2017	0.07	0.24	0.34
BUSI	0.11	0.27	0.41
REFUGE(Cup)	0.06	0.17	0.20
LUNA	0.07	0.19	0.22

increases the difficulty level [23]. In curriculum learning, the first step is to measure the difficulty, defining what constitutes "easy" and "hard." The next step is to develop a customized curriculum according to the difficulty assessment. Curricula, in their simplest form, are a sequence of distributions Q_w, which is generated by adjusting the distribution with a different set of weights w. Take these distributions in order of increasing difficulty, w starts from $w = 0$ and ends at $w = 1$. Moreover, these sequential distributions must satisfy the requirement that their entropy increase as (1) [23]. The last step in curriculum learning is to provide the established curricula in sequential time steps to direct the learning process of models

$$E(Q_w) < E(Q_{w+\epsilon}) \quad \forall \epsilon > 0. \tag{1}$$

In the task of 2-D medical image segmentation, the evolution process of the curricular distribution from $Q_{w=0}$ to $Q_{w=1}$ could be understood as a process from an empty set to the ground truth. It is envisaged that this process would convey the curricula from easy-to-segment pixels to hard-to-segment pixels. Intuitively, the difficulty of pixels can be simply determined by their visual properties, including color, saliency, and fuzziness. However, these attributes are not always relevant to a variety of applications. Furthermore, visual priors are one-sided since the contextual relationship between pixels might also convey information about the degree of difficulty. Therefore, based on the outputs of different depth networks, a universal data-driven strategy is developed to provide curricula for various applications. To reduce redundancy in the curricula, we construct a few discrete curricula $\text{Cur}(t), t \in \{1, 2, \ldots, T\}$ rather than the continuous distribution of $Q_w, w \in [0, 1]$. The details about the data-driven curriculum will be introduced in Section III-C.

Given the data-driven curriculum depicted in Fig. 3, the entropies of these curricula on five datasets are listed in Table I. Because the tth curriculum $\text{Cur}(t)$ consists of $\text{Cur}(t-1)$ and the increments at the tth time step, the entropies $E(\text{Cur}(t))$ steadily rise with increasing t, satisfying the criteria of (1). $E(\text{Cur}(t))$ is calculated as

$$E(\text{Cur}(t)) = -\sum_{i=1}^{n} p\left(x_i^t\right) \log\left(p\left(x_i^t\right)\right) \tag{2}$$

where x_i^t is the content of the ith curriculum increment within $\text{Cur}(t)$, $n = t$ is the number of curriculum increment in $\text{Cur}(t)$, and $p(x_i)$ is the probability of x_i^t. For example, shown as Fig. 3, Cur(2) contains two parts $n = 2$ of the curriculum content, i.e., the green part $x_1^{t=2}$ and the yellow one $x_2^{t=2}$.

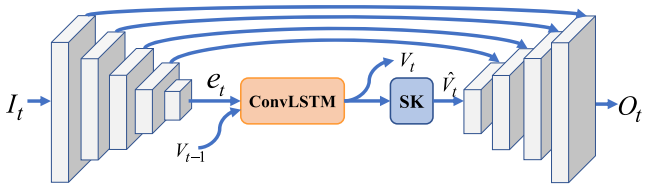


Fig. 7. Illustration of GREnet at the tth time step. The input I_t is the raw image, and the output O_t is the predicted segmentation results jointly supervised by the ground truth and the curriculum $\text{Cur}(t)$ at the tth time step.

$p(x_i^t)$ is computed as

$$p\left(x_i^t\right) = \frac{|x_i^t|}{|\text{GT}|} \tag{3}$$

where $|*|$ indicates the cardinality of $*$ and GT represents the ground truth.

Finally, the gradual curricula are delivered into the training process in sequential time steps to provide incremental guidance.

B. Gradually Recurrent Network

Due to the fact that the curricula are a sequence of supervisions, we design a gradually recurrent network to progressively learn the curriculum context. Taking into account the balance between generalization, lightweightness, and efficacy, we select widely used and simple models UNet [40] and ConvLSTM [41] as the backbone of GREnet. The features supervised by progressive curricula are encoded using UNet, and ConvLSTM is utilized to transform temporal features during the progressive segmentation process. Additionally, ConvLSTM excels in capturing spatiotemporal correlation because it incorporates convolution operations in both the input-to-state and state-to-state transitions [41]. Shown as Fig. 1(b), the size of $\text{Cur}(t)$ steadily increases as t increases. To achieve effective feature representation, we further apply an SK module [42] to adaptively alter the receptive field of the feature extraction.

The visualization of the backbone is given in Fig. 7. Given a medical image as input I_t, UNet can obtain the deep feature representation e_t at the tth time step, which can be defined as

$$e_t = \text{En}\left(I_t; \mathbf{W}_e^t\right) \tag{4}$$

where En is the encoder of UNet, and W_e^t is the parameters that are learned in the encoding process at the tth time step. The naïve UNet has stacked blocks, including two convolution layers and a max-pooling layer. The channels of convolution layers in each block are 16, 32, 64, 128, and 256, respectively. Following the encoder, the ConvLSTM layer can be defined as

$$V_t = \textbf{ConvLSTM}(e_t, V_{t-1}) \tag{5}$$

where V_t and V_{t-1} mean the outputs of ConvLSTM at the tth and $(t-1)$th time step, respectively. The ConvLSTM layer is not used at the 1st time step.

Following the information fusion of ConvLSTM, an SK module is utilized to adaptively optimize the receptive field. Shown in Fig. 8, the output V_t of ConvLSTM will be divided

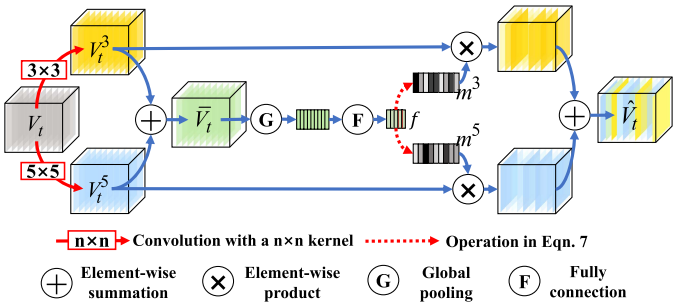


Fig. 8. Architecture of the SK module.

into two paths, i.e., $V_t^3 \in \mathbb{R}^{H\times W\times C}$ and $V_t^5 \in \mathbb{R}^{H\times W\times C}$, using the convolution operations with 3×3 and 5×5 kernel sizes, respectively. V_t^3 and V_t^5 are further fused to provide integrated information $\overline{V}_t \in \mathbb{R}^{H\times W\times C}$ using an elementwise summation as

$$\overline{V}_t = \sum_{k\in\{3,5\}} V_t^k. \tag{6}$$

Then, via global pooling and fully connected operation, $\overline{V}_t$ is converted to a compact feature $f \in \mathbb{R}^{d\times 1}$, where $d = \max(C/16, 32)$ [42]. Driven by the compact feature f, two attention vectors, i.e., m^3 and m^5, are employed to adaptively select different respective fields. The attention vectors $m^k, k \in \{3, 5\}$ are produced as follows:

$$m^k = e^{W_k f} \Big/ \sum_{k\in\{3,5\}} e^{W_k f} \tag{7}$$

where $W_k \in \mathbb{R}^{C\times d}$ is the trainable parameters. By aggregating the feature maps on the channel dimension, the final feature vector $\hat{V}_t$ is generated as

$$\hat{V}_t^c = \sum_{k\in\{3,5\}} m^{k,c} V_t^{k,c} \tag{8}$$

where $\hat{V}_t^c \in \mathbb{R}^{H\times W}$ and $V_t^{k,c} \in \mathbb{R}^{H\times W}$ represent the feature of $\hat{V}_t$ and V_t^k at the cth channel, respectively; $m^{k,c}$ is the cth element of m^k.

Following the SK module, the image scale is recovered by the use of multiple decoding blocks, each of which contains an UpSampling layer and two convolution layers for dense pixel-level prediction. The decoder can be thought of as

$$O_t = \mathrm{De}\left(\hat{V}_t; W_d^t\right) \tag{9}$$

where W_d^t denotes the parameters which are learned in decoding process De, and O_t is the output at the tth time step.

C. Data-Driven Curriculum

This section will discuss the specifics of mining curricula. Intuitively, curricula can be defined by skilled physicians. However, owing to the dense prediction of 2-D medical image segmentation, manually defined curricula are time expensive and subjective, often resulting in an issue of experience bias. Additionally, manual curricula are not uniform for diverse applications. Therefore, we propose a data-driven method to effectively mine curricula from the training set via deep-learning networks.

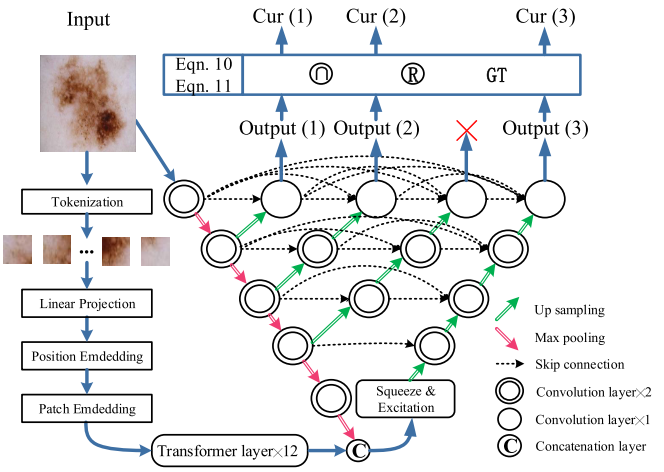


Fig. 9. Illustration of data-driven curriculum-mining. In most cases, the output marked as a red × and output(3) are typically similar experimentally. Therefore, the output marked as a red × is not utilized to develop curricula.

The curricula with progressive contents mainly involve two facts: 1) roughly identifying easy-to-segmentation and hard-to-segmentation regions based on capabilities of various networks with differing depths and 2) carefully integrating the regions as curricula that fulfill the entropy increment criteria specified in (1). Generally, deep-learning networks with different depths will pay attention to different-level information included in the input image [40], [43]. Shallow networks prefer low-level representation information and predict outcomes according to low-level features such as color and edge. In comparison, deep networks excel in modeling high-level semantic information and making predictions based on high-level features as context.

To create distinct networks with varying degrees of depth, numerous independent networks can be designed for the purpose of developing curricula. However, designing many separate networks entails a significant amount of labor. Furthermore, independent networks exhibit distinct model biases. It implies that different networks prefer to mine different types of semantic information according to their unique architecture. Model biases would cause unneeded disruption in the second fact.

To handle this issue, we adopt UNet++ [10] as the backbone of the curriculum-mining network. UNet++ is an integration of CNNs, which has different depths. It means that the outputs of its sub-networks can present different-level segmentation in medical images. Aiming to enhance the curriculum quality, we further introduce transformer to modify the naïve UNet++. The advantage of a transformer is to capture long-range dependencies, which have been demonstrated in numerous fields of image processing [11].

Shown as Fig. 9, transformer and UNet++ are merged in the data-driven curriculum-mining network. In detail, because transformer could optimize the feature representation based on the global information, the features from the transformer and UNet++ encoders are concatenated at the deepest layer. The rough segmentation regions are provided by the outputs of the

curriculum-mining network at different depths. They denoted as Output(t), $t \in \{1, 2, 3\}$, which will be used to produce the curricula Cur(t), $t \in \{1, 2, 3\}$ of the proposed GREnet. The operation of providing Cur(t) is defined as

$$\begin{aligned} \text{Cur}(t) &= \text{(R)}(\text{Output}(t), \text{Cur}(t-1), \text{GT}) \\ &= (\text{Output}(t) \cap \text{GT}) \cup \text{Cur}(t-1) \end{aligned} \tag{10}$$

where Cur(1) is calculated as

$$Cur(1) = \text{(n)}(\text{Output}(1), \text{GT}) = \text{Output}(1) \cap \text{GT} \tag{11}$$

where GT denotes the ground truth, and the visualization examples of (R) and (n) are given in Fig. 2.

Next, we will introduce how to train the curriculum-mining network and deliver the curricula extracted from the training set. It is envisaged that the curricula will be determined solely by the capabilities of the curriculum-mining networks with different depths, avoiding interference from disparate data distributions. Therefore, unlike typical training strategies, which divide the training set into a partial training set and a validation set, the curriculum-mining network trains on the whole training set and subsequently delivers the curricula on the entire training set as well. The primary reasons are as follows: 1) our curriculum-mining network is designed for curriculum-mining rather than general performance; 2) it avoids the data loss situation caused by splitting the training set; and 3) training the network and delivering the curricula both on the same training set forces the final curriculum to approach the ground truth. It may obviate the need for a manually specified curriculum, i.e., the ground truth, which often includes label noise in the task of 2-D medical image segmentation [44].

Algorithm 1 Curriculum-Mining

Input: Training set $\{x_n, y_n\}$, n $\in \{1, 2, \ldots, N\}$;
Curriculum-mining network $\mathcal{C}$; Loss $\mathbb{L}_C$ in (18) and (19).
Output: Number T of curricula;
Curricula $Cur_n(t)$, $t \in \{1, \ldots, T\}$, n $\in \{1, 2, \ldots, N\}$.
Start
Initialize the parameters W_C of the curriculum-mining network
# Train the network on the training set
while not converged **do**
  Update $W_C = argmin_{W_C^*} \sum_{n=1}^{N} \mathbb{L}_C(y_n; \mathcal{C}(W_C^*, x_n))$
end
# Obtain 3-level outputs on the training set
for n $= 1, 2, \ldots, N$ **do**
  $\{O_n(1), O_n(2), O_n(3)\} = \mathcal{C}(W_C, x_n)$
end
# Convert the outputs to curricula
for n $= 1, 2, \ldots, N$ **do**
  for t= 1,2,3 **do**
    if t = 1 **then** $Cur_n(t) = O_n(t) \cap y_n$
    else $Cur_n(t) = (O_n(t) \cap y_n) \cup Cur_n(t-1)$
  end
end
Selection of the number of curricula
if $\frac{1}{N} \sum_{n=1}^{N} (E(Cur_n(3)) - E(Cur_n(2)) < 0.05$
  for n $= 1, 2, \ldots, N$ **do** $Cur_n(2) = Cur_n(3)$
  return $\{T = 2; Cur_n(t), t \in \{1, \ldots, T\}, n \in \{1, 2, \ldots, N\}\}$
else
  return $\{T = 3; Cur_n(t), t \in \{1, \ldots, T\}, n \in \{1, 2, \ldots, N\}\}$
end

* x_n and y_n represent the n^{th} sample and its ground truth, respectively.

IV. Objective Function

In this section, because GREnet and the curriculum-mining network are trained independently rather than jointly, we will detail the objective functions of GREnet and the curriculum-mining network, respectively. The pseudocodes of the curriculum mining and the training procedure of GREnet are provided in Algorithms 1 and 2, separately.

A. GREnet

The objective function $\mathbb{L}_G$ is comprised of the terms $L^c(t)$ from the supervision of the curricula and the terms L^g from the supervision of the ground truth. It is defined as

$$\mathbb{L}_G = \sum_{t=1}^{T} \left(L^c(t) + L^g\right) \tag{12}$$

where T denotes the final recurrent time. $L^c(t)$ and L^g are represented as follows:

$$L^c(t) = L^c_{\text{JA}}(t) + L^c_{\text{BCE}}(t) + L^c_{l1}(t) \tag{13}$$

$$L^g = L^g_{\text{JA}} + L^g_{\text{BCE}} + L^g_{l1} \tag{14}$$

where the subscripts JA, BCE and $l1$ indicate the jaccard loss, binary cross-entropy loss and smooth L_1 loss, respectively. For $L^c(t)$, the sign t means that the tth curriculum is adopted as the supervisor at the tth time step. Because the terms in $L^c(t)$ and L^g are consistent in form, we abbreviate them as L_{JA}, L_{BCE} and L_{l1} for simplification.

The jaccard loss can be written as:

$$L_{\text{JA}} = 1 - \frac{\sum_{(i,j)} \mathcal{Y}(i,j)\widehat{\mathcal{Y}}(i,j)}{\sum_{(i,j)} \mathcal{Y}(i,j)^2 + \sum_{(i,j)} \widehat{\mathcal{Y}}(i,j)^2 - \sum_{(i,j)} \mathcal{Y}(i,j)\widehat{\mathcal{Y}}(i,j)} \tag{15}$$

where (i, j) represents the index of pixels; $\widehat{\mathcal{Y}}_t(i, j)$ is the predicted value of the (i, j)th pixel, and $\mathcal{Y}(i, j)$ is the value of the (i, j)th pixel of the supervisor. For $L^c_{\text{JA}}(t)$, the supervisor is the tth curriculum, and the ground truth for L^g_{JA}. Additionally, we also use binary cross-entropy loss function as a component of the loss function

$$\begin{aligned} L_{\text{BCE}} = -\sum_{(i,j)} &(\mathcal{Y}(i,j))\log\left(\widehat{\mathcal{Y}}(i,j)\right) \\ &+ (1 - \mathcal{Y}(i,j))\log\left(1 - \widehat{\mathcal{Y}}(i,j)\right) \end{aligned} \tag{16}$$

where $\widehat{\mathcal{Y}}_t(i, j)$ and $\mathcal{Y}(i, j)$ have the same meaning as that in (15). The smooth L_1 loss [45] is defined as follows:

$$L_{l1} = \begin{cases} 0.5\ x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases} \tag{17}$$

Algorithm 2 Training GREnet Using the Mined Curricula

Input: Training set $\{x_n, y_n\}$, n $\in \{1, 2, \ldots, N\}$;
Number T of curricula;
Curricula $Cur_n(t)$, t $\in \{1, \ldots, T\}$, n $\in \{1, 2, \ldots, N\}$;
Gradually recurrent network $\mathcal{G}$; Loss $\mathbb{L}_G$ in (12–(17)
Output: Final parameters W_G of GREnet
Start
1. Initialize the parameters W_G of GREnet
Train the network on the training set
2. **while** not converged **do**
3. Update $W_G = argmin_{W_G^*} \sum_{n=1}^{N} \mathbb{L}_G(y_n, \{Cur_n(1), \ldots, Cur_n(T)\}; \mathcal{G}(W_G^*, x_n))$
$= argmin_{W_G^*} \sum_{n=1}^{N} \sum_{t=1}^{T} \mathbb{L}_G(y_n, Cur_n(t); O_n(t))$,
#Obtain gradual outputs
where $\{O_n(1), \ldots, O_n(T)\} = \mathcal{G}(W_G^*, x_n)$
4. **end**
5. **return** W_G

* x_n and y_n represent the n^{th} sample and its ground truth, respectively.

where

$$x = \mathcal{Y}(i, j) - \widehat{\mathcal{Y}}(i, j).$$

B. Curriculum-Mining Network

The objective function $\mathbb{L}_C$ consists of multiple dice coefficient losses $L_D(t), t \in \{1, 2, \ldots T\}$ at the outputs of the curriculum-mining network. The constraints can be defined as

$$\mathbb{L}_C = \sum_{t=1}^{T} L_D(t). \tag{18}$$

Due to the consistent form of $L_D(t)$, $L_D(t)$ is abbreviated as L_D, which can be formulated as

$$L_D = 1 - \frac{2 \sum_{(i,j)} \mathcal{Y}(i, j) \widehat{\mathcal{Y}}(i, j)}{\sum_{(i,j)} \mathcal{Y}(i, j)^2 + \sum_{(i,j)} \widehat{\mathcal{Y}}(i, j)^2} \tag{19}$$

where (i, j) represents the index of pixels; $\widehat{\mathcal{Y}}_t(i, j)$ is the predicted value of the (i, j)th pixel, and $\mathcal{Y}(i, j)$ is the value of the (i, j)th pixel of the Ground Truth.

V. Experiments

We conduct extensive experiments on seven public medical segmentation datasets to verify the effectiveness of GREnet. Three skin lesion segmentation datasets in dermoscopic images are used to show the generalization in the same category task with different data distributions. Except for the skin lesion datasets, an optic disc and cup segmentation dataset and a retinal blood vessel segmentation dataset in retinal images, a breast lesion segmentation dataset in ultrasound images, and a lung segmentation dataset in CT is utilized to analyze the generalization in different tasks.

Section V is organized as follows. The evaluation metric and the implementation details are introduced in Sections V-A and V-B, respectively. The comparison experiments on the skin lesion segmentation datasets, the optic disc and cup segmentation dataset, the retinal blood vessel segmentation dataset, the breast lesion segmentation dataset and the lung segmentation dataset are given in Sections V-C–V-G, separately. Then, the ablation studies of architectures and losses are illustrated in Sections V-H and V-I. We give the analyses of curriculum volume in Section V-J. Section V-K provides the analyses of the time and space complexity. The qualitative evaluation is provided in Section V-L. We give the learning curves in Section V-M. Section V-N will discuss failure cases.

A. Evaluation Metric

To evaluate the segmentation capability of GREnet, we introduce five widely used metrics, including Jaccard index (JA), dice coefficient (DI), accuracy (ACC), sensitivity (SE), and specificity (SP). They are defined as follows:

$$\text{JA} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{20}$$

$$\text{DI} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \tag{21}$$

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{22}$$

$$\text{SE} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{23}$$

$$\text{SP} = \frac{\text{TN}}{\text{FP} + \text{TN}} \tag{24}$$

where true positive (TP) and true negative (TN) represent the number of foreground pixels and background pixels correctly segmented; false positive (FP) denotes the background pixels that are incorrectly labeled as the foreground pixels; false negative (FN) denotes the foreground pixels that are incorrectly predicted as the background pixels.

In the comparison experiments with the state-of-the-arts, we follow the latest works to select the main evaluation metrics of different applications.

B. Implementation Details

Given a 2-D medical image, we select the **X** and **Z** channels in **XYZ** color space as extra channels in our model training [46], [47]. The RGB datasets, which include four skin lesion segmentation datasets, the REFUGE dataset, and the DRIVE dataset, use the input with **RGB** and **XZ** channels, whereas the BUSI and LUNA datasets, which contain gray images, use the input channel unmodified.

All experiments are implemented by Keras on an NVIDIA GeForce RTX 3090 GPU. In the training phase, an Adam optimizer with an initial learning rate of 0.0001 is used to minimize the objective function. All networks are trained for 40 epochs.

The division of the training set, validation set, and test set is consistent with comparison methods.

C. Skin Lesion Segmentation

In this section, we evaluate the performance of GREnet on three skin lesion segmentation datasets, including ISIC

2017 [48], ISIC 2018 [49], and PH2 [50]. The details of the three datasets are as follows.

1) *ISIC 2017:* the ISIC 2017 dataset consists of 2000 training images, 150 validation images, and 600 testing images.
2) *ISIC 2018:* without published testing images, the ISIC 2018 dataset contains 2594 training images. Following the latest methods [2], [32], [51], we divide the training data into 80% training (2076 images) and 20% validation (518 images). To guarantee the robustness of the model against different parts of the available data, five-fold cross-validation is performed.
3) *PH2:* the PH2 dataset is a small dataset that only contains 200 images.

In our experiments, the resolution of all training, validation, and testing images are uniformly resized to 224 × 160. Tables II–IV show the comparison with the state-of-the-art methods on ISIC 2017, PH2, and ISIC 2018 datasets, respectively. It can be seen that as follows.

1) Compared with re-weighting segmentation methods [25], [29], GREnet achieves performance improvement. For example, GREnet outperforms Wang et al. [25] by 3.2% in ACC and 2.7% in SP on the ISIC 2017 dataset. In comparison to Abraham et al. [29], our method improves the DI and SE performance by 6.3% and 6.3% on the ISIC 2018 dataset.
2) In comparison to cascading segmentation methods [2], [32], GREnet obtains the best segmentation performance on all evaluation metrics. Compared with FrCN [2], the JA values of GREnet are enhanced by 3.7% on the ISIC 2017 datasets. Compared with iMSCGnet [32], the proposed method increases the JA value by 3.0% on the ISIC 2017 datasets.
3) When compared to the curriculum-learning-based segmentation method [1], GREnet increases the DI value by 0.7% and the ACC value by 2.6% on the ISIC 2017 datasets.
4) Compared with the approaches developed especially for skin lesion segmentation, such as DAGAN [52], FAT-Net [11], MsRED [53], and DCL-PSI [54], the proposed GREnet achieves performance gains on the four skin lesion segmentation datasets.
5) Given the tiny size of PH2 (only 200 images), some approaches separate it into the training and test sets and evaluate the performance using the five-fold cross-validation. Others utilize it as a test set to assess the generalization capabilities of the model trained on the ISIC 2017 dataset. Shown in Table III, the results in the first part are obtained following the former, and the second part is for the latter to evaluate the generalization ability. GREnet generally outperforms the state-of-the-art methods.

TABLE II
PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE ISIC 2017 DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD

Methods	DI	JA	ACC	SE	SP
Wang *et al.* [25]	0.879	0.804	0.942	0.900	0.958
FrCN [2]	0.871	0.771	0.940	–	–
iMSCGnet [32]	0.858	0.778	0.936	–	–
AFLN-DGCL-FUSION [1]	0.881	0.807	0.948	0.891	0.964
DAGAN [52]	0.859	0.771	0.935	0.835	0.976
FAT-Net [11]	0.850	0.765	0.933	0.839	0.973
MsRED [53]	0.865	0.786	0.941	–	–
DCL-PSI [54]	0.857	0.778	0.941	0.862	0.967
DSNet [55]	–	0.775	–	0.875	0.955
CSARM-CNN [56]	0.846	0.734	0.959	0.802	**0.994**
CSFAG-CNN [57]	0.870	0.789	0.950	0.870	**0.994**
GREnet (ours)	**0.888**	**0.808**	**0.974**	**0.910**	0.985

TABLE III
PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE PH2 DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD

Methods	DI	JA	ACC	SE	SP
FAT-Net [11]	0.944	0.896	0.970	0.944	0.974
MsRED [53]	0.947	0.901	0.968	–	–
GREnet(ours)	**0.963**	**0.921**	**0.989**	**0.965**	**0.980**
FrCN [2]	0.918	0.848	0.951	–	–
iMSCGnet [32]	0.934	**0.882**	0.957	–	–
DSNet [55]	–	0.870	–	0.929	0.969
CSARM-CNN [56]	0.883	0.791	0.952	0.885	**0.995**
CSFAG-CNN [57]	0.910	0.834	0.959	0.898	0.977
GREnet(ours)	**0.935**	0.877	**0.961**	**0.958**	0.942

TABLE IV
PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE ISIC 2018 DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD

Methods	DI	JA	ACC	SE	SP
Abraham *et al.* [29]	0.856	–	–	0.897	–
MsRED [53]	0.900	0.835	0.962	–	–
CPF-Net [51]	0.899	0.829	0.963	–	–
Qamar *et al.* [58]	0.900	0.833	0.970	**0.965**	0.970
SwinPA-Net [18]	0.911	**0.854**	-	0.946	-
GREnet (ours)	**0.919**	0.851	**0.970**	0.960	**0.976**

D. Optic Disc and Cup Segmentation

In this section, we evaluate GREnet on a fundus image dataset, which is acquired from the MICCAI 2018 Retinal Fundus Glaucoma Challenge (REFUGE). Manual annotations of the optic disc and cup were obtained by seven independent ophthalmologists from the Zhongshan Ophthalmic Center, Sun Yat-sen University, China. The REFUGE dataset consists of 1200 images, including 120 glaucoma and 1080 nonglaucoma images, and provides predefined partitions into training and test. Experiments are conducted on the predefined partitions of REFUGE challenge, in which 400, 400, and 400 are for training, validation, and testing, respectively. To segment the optic disc and cup in the retinal fundus images based on their original resolution, a 512 × 512 region around the brightest point is cropped [59]. We adopt the JA and DI metrics from (20) and (21), respectively. The comparison results are listed in Table V. It can be observed that as follows.

1) Segtran [60] and ET-Net [61] are two common segmentation methods, which have provided the segmentation results on the REFUGE dataset. The proposed GREnet outperforms Segtran [60] and ET-Net [61] with increases of $\{0.8\%, 1.6\%\}$ and $\{2.1\%, 0.2\%\}$ in the DI value for optic disc and cup, respectively.

TABLE V
PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART METHODS ON REFUGE DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD

Methods	DI_{disc}	DI_{cup}	JA_{disc}	JA_{cup}
DSFCN [4]	0.873	0.868	0.783	0.771
Bian *et al.* [3]	0.933	0.880	0.876	0.791
ET-Net [61]	0.953	0.891	–	–
TAU [62]	0.960	0.889	0.929	**0.823**
Hervella *et al.* [63]	0.959	0.883	–	–
Segtran [60]	0.961	0.872	–	–
GREnet(ours)	**0.969**	**0.893**	**0.933**	0.820

TABLE VI
PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE DRIVE DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD

Methods	SE	ACC
Yeung *et al.* [64]	0.818	–
SCS-Net [65]	0.829	0.970
Pyramid u-net [66]	0.821	0.962
Bridge-Net [67]	0.800	0.967
Zhou *et al.* [68]	0.829	0.956
Chala *et al.* [69]	0.821	**0.972**
Sine-Net [70]	0.799	0.969
GREnet (ours)	**0.838**	0.968

2) In comparison to methods based on the cascading strategy, such as DSFCN [4] and Bian et al. [3], which are designed for the optic disc and cup segmentation, GREnet obtains gains in DI values of {9.6%, 2.5%} and {3.6%, 1.3%}, separately.
3) GREnet generally outperforms other methods, demonstrating once again that GREnet based on curriculum learning has the benefit of 2-D medical image segmentation.

E. Retinal Blood Vessel Segmentation

The DRIVE dataset is used to evaluate GREnet in fundus images. It consists of 40 fundus images, where 20 images are selected for training, and the remaining are utilized in the testing phase [67]. All images are obtained by a Canon camera at 45° with a resolution of 584 × 565. We resize the images to 512 × 512. As defined in (22) and (23), the SE and ACC metrics are applied to quantitatively analyze the experimental results [65], [67].

Table VI provides the comparison results between GREnet and the state-of-the-art methods. Compared with Yeung et al. [64], which is a re-weighting method, GREnet has a 2.0% increase in SE value. In comparison to the other approaches, which are meticulously designed for retinal vessel segmentation, GREnet rates top at the SE metric, which measures the accuracy of blood vessel segmentation, and achieves comparable ACC performance.

F. Breast Lesion Segmentation

In this section, a breast lesion segmentation dataset, which is denoted as BUSI, is employed to evaluate GREnet in ultrasound images. The BUSI dataset is collected from 600 female patients at Baheya, Cairo, and Egypt. There are 780 breast ultrasound images in the BUSI dataset, which has 647 abnormal samples (210 malignant and 437 benign masses) with annotations and 133 normal samples. Since this task focus on lesion segmentation, only abnormal samples are used for training and testing [6]. All images are scaled to 224 × 224. For a fair comparison, five-fold cross-fold validation is applied [6]. DI, JA, ACC, SE, and SP metrics are employed to evaluate the results. It can be seen from Table VII that GREnet also gets comparable performance with the state-of-the-art methods.

TABLE VII
PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE BUSI DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD

Methods	DI	JA	ACC	SE	SP
UNet [40]	0.706	0.613	0.954	0.739	0.979
SK-U-Net [71]	0.709	–	0.956	–	–
LAEDNet [5]	0.750	0.676	–	–	–
MCRNet [6]	**0.823**	0.699	0.968	0.817	–
FCP-Net [72]	0.791	0.654	0.967	–	**0.988**
GREnet(ours)	0.799	**0.707**	**0.970**	**0.820**	0.974

TABLE VIII
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS ON THE LUNA DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD

Methods	ACC	SE	SP
GC-Net [74]	0.990	0.988	–
GREnet(ours)	**0.995**	**0.991**	**0.994**
ET-Net [61]	0.987	–	–
CE-Net [73]	0.990	0.980	–
R2u-net [7]	0.992	0.983	0.994
PyDiNet [8]	0.996	0.991	–
SA-Net [75]	0.986	0.988	–
GREnet(ours)	**0.998**	**0.993**	**0.995**

G. Lung Segmentation

Next, we employ GREnet to segment CT images on the lung nodule analysis (LUNA) dataset. The dataset contains 534 2-D samples with correspondingly labeled images for lung segmentation. Following [61] and [73], we adopt a five-fold cross-validation strategy. The original image size is 512 × 512 pixels and has been resized to 256 × 256 pixels. In accordance with the latest methods [7], the evaluation metrics including ACC, SE and SP from (22), (23), and (24) are used. Given a ratio of 7:3 (training vs. test) in GC-Net [74] is different from others 8:2. We conduct experiments with a ratio of 7:3 and five-fold cross-validation. The first part in Table VIII is the results with a ratio of 7:3, and the second is the comparisons with a ratio of 8:2.

GREnet is compared to the following methods: ET-Net [61], GC-Net [74], CE-Net [73], R2u-net [7], PyDiNet [8], and SA-Net [75]. The comparison results are listed in Table VIII. It can be observed that the range of ACC improvement is from 0.2% to 1.2%. In particular, R2u-net [7] is an approach based on a recurrent convolutional neural network. R2u-net makes use of recurrent layers but does not apply gradual curricula. The proposed GREnet improves ACC and SE values by 0.6% and 1.0%, respectively. It means that the gradual curriculum supervision with curriculum learning is useful for Lung segmentation.

TABLE IX
ABLATION STUDIES ON THE ISIC 2017 AND REFUGE DATASETS. THE BEST VALUES ARE HIGHLIGHTED IN BOLD

Methods	ISBI 2017					REFUGE			
	DI	JA	ACC	SE	SP	DI_{disc}	DI_{cup}	JA_{disc}	JA_{cup}
UNet	0.843	0.757	0.931	0.858	0.968	0.931	0.854	0.898	0.781
TransUNet++	0.857	0.767	0.943	0.859	0.967	0.961	0.883	0.908	0.803
Cascaded UNet	0.855	0.770	0.941	0.855	0.968	0.956	0.878	0.916	0.788
REnet w/o SK	0.868	0.774	0.943	0.886	0.971	0.955	0.884	0.916	0.791
REnet	0.874	0.785	0.952	0.882	0.966	0.958	0.881	0.921	0.799
GREnet w/o SK	0.876	0.786	0.956	0.884	0.963	0.966	0.886	0.922	0.794
GCnet	0.870	0.786	0.954	0.868	0.973	0.958	0.889	0.920	0.809
GREnet*	0.867	0.781	0.966	0.892	0.968	0.954	0.876	0.919	0.797
GREnet(SPL)	0.875	0.797	0.958	0.893	0.966	0.960	0.882	0.920	0.805
GREnet	**0.888**	**0.808**	**0.974**	**0.910**	**0.985**	**0.968**	**0.893**	**0.930**	**0.819**

H. Ablation Studies of Architectures

To demonstrate the effectiveness of different components in the proposed method, we conduct a step-by-step ablation experiment on the ISIC 2017 and REFUGE datasets, respectively. In summary, the following models are compared.

1) *UNet:* we choose UNet as the base model.
2) *TransUNet++:* TransUNet++ is the proposed curriculum-mining network. In this section, TransUNet++ is utilized as another base model for comparison.
3) *Cascaded UNet:* Cascaded UNet is the model based on UNet with the cascaded strategy. The output of the previous stage and the images are concatenated as the input of the next network. Three UNet are applied in the Cascaded UNet.
4) *REnet:* REnet is the version of GREnet without gradual curriculum supervision.
5) *GCnet:* GCnet is the version of GREnet without ConvLSTM, which is replaced by a concatenation operation.
6) *GREnet*:* GREnet* is the version of GREnet with a different recurrent model [76] based on LSTM.
7) *GREnet(SPL):* GREnet(SPL) is the version of GREnet using the self-paced learning [77].
8) *GREnet:* GREnet is the proposed network.
9) *w/o SK:* it represents the model without the SK module.

The results are listed in Table IX. It can be seen that as follows.

1) GREnet beats the basic models, namely UNet and TransUNet++, by a large margin. This indicates that the gradual curricula can effectively improve the performance of the base model UNet. Additionally, The performance of GREnet is not limited by that of its curriculum-mining network. This is in contrast to knowledge-distillation-based methods, where the performance of the teacher model is the upper bound of the student.
2) Compared with Cascaded UNet, GREnet yields an improvement on the two datasets. It demonstrates that curriculum learning with gradual curricula is superior to the cascaded strategy in the task of 2-D medical image segmentation because of the progressive guidance. As the analyses motioned before, the cascaded strategy would cause the segmentation model to get confused because of the errors in the previous coarse results.
3) Benefiting from the adaptively receptive field of the SK module, the performance of REnet and GREnet is substantially enhanced. It implies that the SK module may assist the model with effectively representing features while adhering to gradual curricula.
4) Due to the gradual curricula, GREnet performs better than REnet. The main reason is that the progressive curricula would provide gradually incremental guidance from easy-to-segmentation regions to hard-to-segmentation parts.
5) Thanks to the recurrent mechanism using ConvLSTM, GREnet outperforms GCnet on the two datasets. It demonstrates the effectiveness of converting 2-D medical image segmentation to a temporal progression task. Additionally, GREnet with ConvLSTM performs better than GREnet*. The main reason is that ConvLSTM can extract spatial features better than LSTM.
6) In comparison to REnet, GREnet(SPL) using sample-level self-paced learning and GREnet using pixel-level curriculum learning can both enhance performance. Thus, future study would focus on the integration of pixel-level curricular learning and self-paced learning.

I. Ablation Studies of Losses

Table X illustrates the impact of different loss functions of GREnet. Combing L_{JA} with L_{BCE} results in an increase of 1.0% in the JA value, and adding the L_{l1} results in an additional increase of 1.2% in the JA value. Using smooth L_1 loss, the sensitivity to outliers is reduced. The number of false positives is reduced using binary cross-entropy loss. This investigation demonstrates that the combination of L_{JA}, L_{BCE}, and L_{l1} improves segmentation results the most.

Additionally, inspired from the adaptive loss function in [78], the L_{l1} term is replaced by an adaptive loss term as $L_{\mathrm{ad}} = \sum_{(i,j)}(((1+\delta)(y(i,j)-\hat{y}(i.j))^2)/(|y(i,j)-\hat{y}(i,j)|+\delta))$, where $\delta = 0.1$ [78]. As can be observed, GREnet can perform as well as GREnet employing an adaptive loss when using the L_{l1} loss. Currently, GREnet pays more attention to general curriculum learning. Thus, we adopt the naïve L_{l1} loss in the present version. We will conduct additional research on the loss design in the future.

TABLE X

COMPARISON OF GRENET WITH DIFFERENT LOSS FUNCTIONS ON THE ISIC 2017 DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD

L_{JA}	L_{l1}	L_{ad}	L_{BCE}	DI	JA	ACC	SE	SP
√				0.868	0.786	0.952	0.874	0.968
	√			0.862	0.777	0.949	0.882	0.976
			√	0.862	0.778	0.952	0.875	0.969
√			√	0.879	0.796	0.963	0.884	0.978
√	√		√	**0.888**	**0.808**	**0.974**	**0.910**	**0.985**
√		√	√	0.890	0.809	0.974	0.913	0.986

TABLE XI

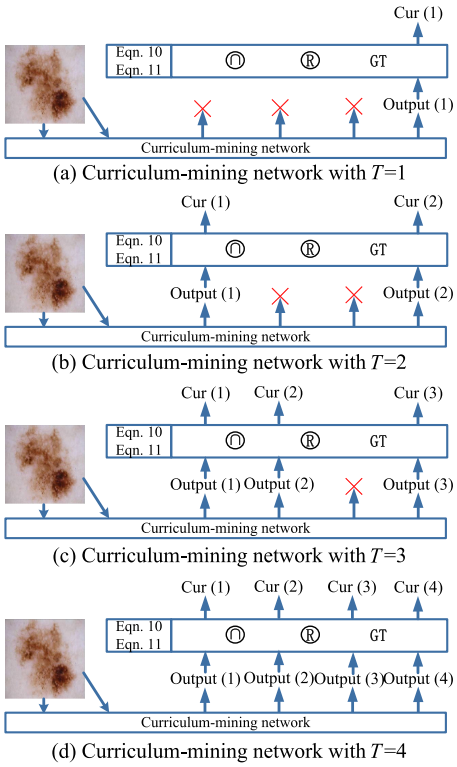
COMPARISON ON FIVE DATASETS WITH DIFFERENT CURRICULUM VOLUME $T \in \{1, 2, 3, 4\}$. $\triangle E$ REPRESENTS THE DIFFERENCE OF ENTROPY BETWEEN THE LAST TWO CURRICULA, I.E., $E(\text{Cur}(2)) - E(\text{Cur}(1))$ AT $T = 2$, $E(\text{Cur}(3)) - E(\text{Cur}(2))$ AT $T = 3$, AND $E(\text{Cur}(4)) - E(\text{Cur}(3))$ AT $T = 4$

Dataset	$\triangle E$	T	DI	JA	ACC	SE	SP
DRIVE	–	1	–	–	0.908	0.678	–
	0.270	2	–	–	0.944	0.796	–
	0.160	3	–	–	**0.968**	**0.838**	–
	0.007	4	–	–	0.967	0.835	–
ISIC 2017	–	1	0.851	0.768	0.932	0.867	0.959
	0.270	2	0.875	0.791	0.955	0.896	0.982
	0.100	3	0.888	**0.808**	**0.974**	**0.910**	**0.985**
	0.011	4	**0.889**	0.806	0.972	0.908	0.984
BUSI	–	1	0.752	0.668	0.974	0.784	0.956
	0.300	2	0.780	0.690	0.968	0.800	0.964
	0.140	3	**0.799**	0.707	**0.974**	**0.820**	**0.974**
	0.010	4	0.797	**0.708**	0.968	0.820	0.970
REFUGE(cup)	–	1	0.875	0.783	–	–	–
	0.140	2	**0.893**	**0.820**	–	–	–
	0.030	3	0.893	0.819	–	–	–
	≈ 0	4	0.892	0.818	–	–	–
LUNA	–	1	–	–	0.986	0.974	0.985
	0.150	2	–	–	**0.998**	**0.993**	0.995
	0.030	3	–	–	0.998	0.991	**0.996**
	≈ 0	4	–	–	0.998	0.990	0.996

TABLE XII

COMPLEXITY ANALYSIS OF GRENET. GRENET(t)* REPRESENTS THE tTH STEP GRENET(t) OF GRENET W/O THE SK MODULE

Parameters (M)			FLOPs (G)	Speed (FPS)
SK module	GRENet(t)*	GRENet(t)	GRENet(t)	GRENet(t)
2.43	6.68	9.11	5.93	500



(a) Curriculum-mining network with T=1

(b) Curriculum-mining network with T=2

(c) Curriculum-mining network with T=3

(d) Curriculum-mining network with T=4

Fig. 10. Curriculum-mining network with various curriculum volumes $T \in \{1, 2, 3, 4\}$.

J. Analysis of Curriculum Volume

In this section, we will give the analysis of curriculum volume, i.e., how to decide the number T of curricula. We do the analyses on five datasets. The comparison results are listed in Table XI. The visualization of the curriculum-mining network with $T \in \{1, 2, 3, 4\}$ can be found in Fig. 10. Here, we only list the difference of entropy $\triangle E$ between the last two curricula in Table XI.

On the DRIVE, ISIC 2017, and BUSI datasets, the results of GREnet with $T = 3$ are generally superior to those of GREnet with $T \in \{1, 2, 4\}$. On the other two datasets, the results of GREnet with $T \in \{2, 3, 4\}$ are almost consistent. As a result of the large curriculum increase, the 3rd curriculum is required for the tasks on the first three datasets. Correspondingly, $T = 2$ is enough for the tasks on the REFUGE and LUNA datasets.

In light of the above observations, we propose an empirical method to determine the number T of curricula. In different tasks, an extra curriculum is needed as $\triangle E > 0.05$ and vice versa. This empirical method can also be used in future research to determine how to select curricula when the total number of curricula is more than 4.

K. Complexity Analysis

Next, we provide the space and time complexity of GREnet in Table XII. Due to the difference of the curriculum number in different applications, we list the complexity analyses of a single time step in GREnet, which is denoted as GREnet(t). GREnet(t) costs about $9.11M$ parameters at the tth time step.

The speed of the curriculum-mining network is 67 frames/s. Because GREnet and the curriculum mining network are trained independently rather than jointly, no extra time is necessary for the GREnet training and inference stages. Thus, GREnet has the capacity to perform on large-scale datasets.

L. Qualitative Evaluation

The qualitative evaluation on the ISIC 2017, REFUGE, and BUSI datasets is illustrated in Fig. 11. It can be observed that as follows.

1) GREnet outperforms UNet and TransUNet++ due to the effectiveness of curriculum learning.
2) In comparison to cascaded UNet, GREnet produces more precise boundaries of segmentation regions. The main reason is that GREnet employs the gradually incremental guidance from the inside out. This mechanism could assist GREnet in avoiding being confused by the distractors located outside of the target regions.
3) In contrast to the results of GREnet(1), which is the 1$^{\text{rt}}$ time step of GREnet, GREnet obtains progressive segmentation outcomes. It means that, from the 1$^{\text{rt}}$ to Tth time step, GREnet can gradually improve the segmentation performance.

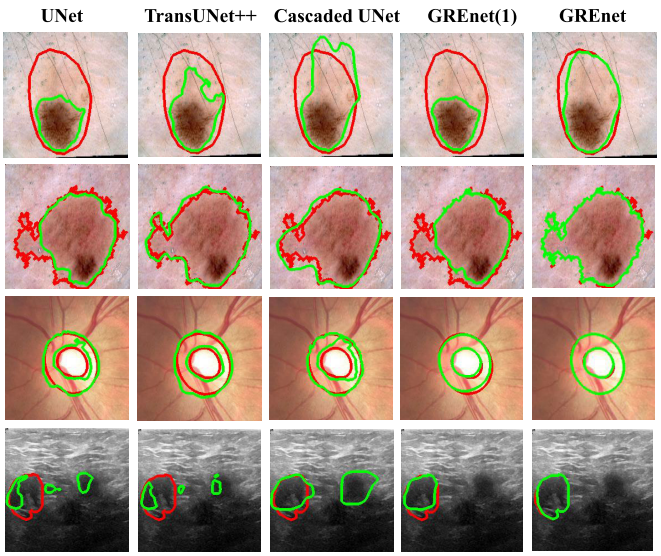


Fig. 11. Qualitative evaluation on the BUSI, REFUGE, and ISIC 2017 datasets. Red and green curves represent the boundary of the ground truth and the segmented results, respectively. GREnet(1) means GREnet at the first time step.

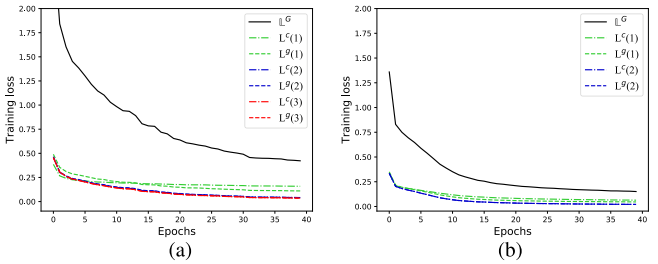


Fig. 12. Learning curves of (12) on the ISIC 2017 and REFUGE(cup) datasets. For clarity, L^g is modified as $L^g(t)$ at the tth, $t \in \{1, 2, \ldots, T\}$ time step. $T = 3$ is used on the ISIC 2017 dataset, and $T = 2$ is used for the REFUGE(cup) dataset. (a) ISIC 2017. (b) REFUGE(cup).

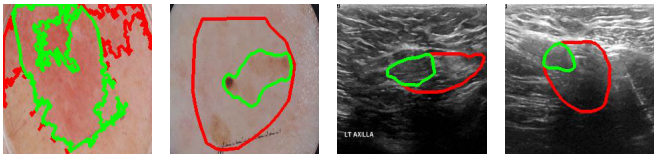


Fig. 13. Failure cases of GREnet on the ISIC 2017 and BUSI datasets. Red and green curves represent the boundary of the ground truth and the segmented results, respectively.

M. Learning Curve

In this section, the learning curves of (12) is provided in Fig. 12. It can be seen that the training losses reduce gradually. The global convergence of the learning curves demonstrates the effectiveness of the training technique. Based on the experimental experience, we train the model for 40 epochs in all situations.

N. Discussion About Failure Cases

We will discuss two scenarios in which the performance of GREnet will be weakened: 1) low-contrast regions and 2) shadow artifacts. Shown in Fig. 13, GREnet is easily disturbed by the low-contrast regions and shadow artifacts, which occupy the bulk of lesion areas. In these failure cases, a more robust base model is required. In the future, to deal with this issue, self-attention modules like Transformer will be investigated in the gradual segmentation framework.

VI. Conclusion

Aiming to improve the performance of 2-D medical image segmentation, we model it as a temporal task with curriculum learning and propose GREnet to gradually learn patterns under the supervision of progressive curricula. To deliver the curricula effectively, we further develop a data-mining approach to adaptively mine the curricula for different applications. The experimental results demonstrate the potential performance of GREnet on seven public datasets and provide guidance on how to determine the number of curricula. In the future, we will do further research about 3-D medical image segmentation using curriculum learning.

References

[1] P. Tang, X. Yan, Q. Liang, and D. Zhang, "AFLN-DGCL: Adaptive feature learning network with difficulty-guided curriculum learning for skin lesion segmentation," *Appl. Soft Comput.*, vol. 110, Oct. 2021, Art. no. 107656.

[2] L. Bi, J. Kim, E. Ahn, A. Kumar, M. Fulham, and D. Feng, "Dermoscopic image segmentation via multistage fully convolutional networks," *IEEE Trans. Biomed. Eng*, vol. 64, no. 9, pp. 2065–2074, Sep. 2017.

[3] X. Bian, X. Luo, C. Wang, W. Liu, and X. Lin, "Optic disc and optic cup segmentation based on anatomy guided cascade network," *Comput. Methods Programs Biomed.*, vol. 197, Dec. 2020, Art. no. 105717.

[4] L. Bi, Y. Guo, Q. Wang, D. Feng, M. Fulham, and J. Kim, "Automated segmentation of the optic disk and cup using dual-stage fully convolutional networks," 2019, *arXiv:1902.04713*.

[5] Q. Zhou, Q. Wang, Y. Bao, L. Kong, X. Jin, and W. Ou, "LAED-Net: A lightweight attention encoder–decoder network for ultrasound medical image segmentation," *Comput. Elect. Eng.*, vol. 99, Apr. 2022, Art. no. 107777.

[6] M. Lou, J. Meng, Y. Qi, X. Li, and Y. Ma, "MCRNet: Multi-level context refinement network for semantic segmentation in breast ultrasound imaging," *Neurocomputing*, vol. 470, pp. 154–169, Jan. 2022.

[7] M. Zahangir Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation," 2018, *arXiv:1802.06955*.

[8] M. Gridach, "PyDiNet: Pyramid dilated network for medical image segmentation," *Neural Netw.*, vol. 140, pp. 274–281, Aug. 2021.

[9] M. A. Al-masni, M. A. Al-Antari, M.-T. Choi, S.-M. Han, and T.-S. Kim, "Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks," *Comput. Methods Programs Biomed.*, vol. 162, pp. 221–231, Aug. 2018.

[10] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.

[11] H. Wu, S. Chen, G. Chen, W. Wang, B. Lei, and Z. Wen, "FAT-Net: Feature adaptive transformers for automated skin lesion segmentation," *Med. Image Anal.*, vol. 76, Feb. 2022, Art. no. 102327.

[12] B. Saha Tchinda, D. Tchiotsop, M. Noubom, V. Louis-Dorr, and D. Wolf, "Retinal blood vessels segmentation using classical edge detection filters and the neural network," *Informat. Med. Unlocked*, vol. 23, 2021, Art. no. 100521.

[13] K. Wang, X. Zhang, X. Zhang, Y. Lu, S. Huang, and D. Yang, "EANet: Iterative edge attention network for medical image segmentation," *Pattern Recognit.*, vol. 127, Jul. 2022, Art. no. 108636.

[14] A. Mihaylova and V. Georgieva, "Spleen segmentation in MRI sequence images using template matching and active contours," *Proc. Comput. Sci.*, vol. 131, pp. 15–22, Dec. 2018.

[15] L. Sun, H. Sun, J. Wang, S. Wu, Y. Zhao, and Y. Xu, "Breast mass detection in mammography based on image template matching and CNN," *Sensors*, vol. 21, no. 8, p. 2855, Apr. 2021.

[16] C. Zhou, C. Ding, X. Wang, Z. Lu, and D. Tao, "One-pass multi-task networks with cross-task guided attention for brain tumor segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 4516–4529, 2020.

[17] M. Havaei et al., "Brain tumor segmentation with deep neural networks," *Med. Image Anal.*, vol. 35, pp. 18–31, Jan. 2017.

[18] H. Du, J. Wang, M. Liu, Y. Wang, and E. Meijering, "SwinPA-Net: Swin transformer-based multiscale feature pyramid aggregation network for medical image segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–12, 2022.

[19] Q. Kang, Q. Lao, and T. Fevens, "Nuclei segmentation in histopathological images using two-stage learning," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*. Cham, Switzerland: Springer, 2019, pp. 703–711.

[20] Y. Zhao et al., "TSASNet: Tooth segmentation on dental panoramic X-ray images by two-stage attention segmentation network," *Knowl.-Based Syst.*, vol. 206, Oct. 2020, Art. no. 106338.

[21] B. F. Skinner, "Reinforcement today," *Amer. Psychologist*, vol. 13, no. 3, p. 94, 1958.

[22] K. A. Krueger and P. Dayan, "Flexible shaping: How learning in small steps helps," *Cognition*, vol. 110, no. 3, pp. 380–394, Mar. 2009.

[23] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 41–48.

[24] H. Li, X. Liu, S. Boumaraf, W. Liu, X. Gong, and X. Ma, "A new three-stage curriculum learning approach for deep network based liver tumor segmentation," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–6.

[25] L. Wang, L. Zhang, and X. Shu, "Focal rank loss function with encoder–decoder network for skin lesion segmentation," in *Proc. J. Phys., Conf.*, 2021, vol. 2010, no. 1, Art. no. 012049.

[26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[27] J. Chang et al., "Brain tumor segmentation based on 3D UNet with multi-class focal loss," in *Proc. 11th Int. Congr. Image Signal Process., Biomed. Eng. Inform. (CISP-BMEI)*, Oct. 2018, pp. 1–5.

[28] S. R. Hashemi, S. S. M. Salehi, D. Erdogmus, S. P. Prabhu, S. K. Warfield, and A. Gholipour, "Asymmetric loss functions and deep densely-connected networks for highly-imbalanced medical image segmentation: Application to multiple sclerosis lesion detection," *IEEE Access*, vol. 7, pp. 1721–1735, 2018.

[29] N. Abraham and N. M. Khan, "A novel focal Tversky loss function with improved attention U-Net for lesion segmentation," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 683–687.

[30] Y. Xie, J. Zhang, Y. Xia, and C. Shen, "A mutual bootstrapping model for automated skin lesion segmentation and classification," *IEEE Trans. Med. Imag.*, vol. 39, no. 7, pp. 2482–2493, Dec. 2020.

[31] Z. Mirikharaji, S. Izadi, J. Kawahara, and G. Hamarneh, "Deep auto-context fully convolutional neural network for skin lesion segmentation," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 877–880.

[32] Y. Tang et al., "iMSCGnet: Iterative multi-scale context-guided segmentation of skin lesion in dermoscopic images," *IEEE Access*, vol. 8, pp. 39700–39712, 2020.

[33] C. Tan and J. Liu, "Improving knowledge distillation with a customized teacher," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jul. 25, 2022, doi: 10.1109/TNNLS.2022.3189680.

[34] E. Kats, J. Goldberger, and H. Greenspan, "Soft labeling by distilling anatomical knowledge for improved MS lesion segmentation," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 1563–1566.

[35] M. Hu et al., "Knowledge distillation from multi-modal to mono-modal segmentation networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*. Cham, Switzerland: Springer, 2020, pp. 772–781.

[36] K. Li, S. Wang, L. Yu, and P.-A. Heng, "Dual-teacher: Integrating intra-domain and inter-domain teachers for annotation-efficient cardiac segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*. Cham, Switzerland: Springer, 2020, pp. 418–427.

[37] D. Qin et al., "Efficient medical image segmentation based on knowledge distillation," *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3820–3831, Dec. 2021.

[38] B. Han, I. W. Tsang, L. Chen, P. Y. Celina, and S.-F. Fung, "Progressive stochastic learning for noisy labels," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 5136–5148, Oct. 2019.

[39] T. Matiisen, A. Oliver, T. Cohen, and J. Schulman, "Teacher–student curriculum learning," *IEEE Trans. neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3732–3740, Sep. 2019.

[40] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[41] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.

[42] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[44] J. Shi and J. Wu, "Distilling effective supervision for robust medical image segmentation with noisy labels," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*. Cham, Switzerland: Springer, 2021, pp. 668–677.

[45] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Deep direct regression for multi-oriented scene text detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 745–753.

[46] R. Ramadan and S. Aly, "CU-Net: A new improved multi-input color U-Net model for skin lesion semantic segmentation," *IEEE Access*, vol. 10, pp. 15539–15564, 2022.

[47] H. Wu, J. Pan, Z. Li, Z. Wen, and J. Qin, "Automated skin lesion segmentation via an adaptive dual attention module," *IEEE Trans. Med. Imag.*, vol. 40, no. 1, pp. 357–370, Jan. 2021.

[48] N. C. Codella et al., "Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 168–172.

[49] N. Codella et al., "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)," 2019, *arXiv:1902.03368*.

[50] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. Marcal, and J. Rozeira, "PH^2—A dermoscopic image database for research and benchmarking," in *Proc. 35th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2013, pp. 5437–5440.

[51] S. Feng et al., "CPFNet: Context pyramid fusion network for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 10, pp. 3008–3018, Oct. 2020.

[52] B. Lei et al., "Skin lesion segmentation via generative adversarial networks with dual discriminators," *Med. Image Anal.*, vol. 64, Aug. 2020, Art. no. 101716.

[53] D. Dai et al., "Ms RED: A novel multi-scale residual encoding and decoding network for skin lesion segmentation," *Med. Image Anal.*, vol. 75, Jan. 2022, Art. no. 102293.

[54] L. Bi, J. Kim, E. Ahn, A. Kumar, D. Feng, and M. Fulham, "Step-wise integration of deep class-specific learning for dermoscopic image segmentation," *Pattern Recognit.*, vol. 85, pp. 78–89, Jan. 2019.

[55] M. K. Hasan, L. Dahal, P. N. Samarakoon, F. I. Tushar, and R. Martí, "DSNet: Automatic dermoscopic skin lesion segmentation," *Comput. Biol. Med.*, vol. 120, May 2020, Art. no. 103738.

[56] Y. Jiang, S. Cao, S. Tao, and H. Zhang, "Skin lesion segmentation based on multi-scale attention convolutional neural network," *IEEE Access*, vol. 8, pp. 122811–122825, 2020.

[57] S. Tao, Y. Jiang, S. Cao, C. Wu, and Z. Ma, "Attention-guided network with densely connected convolution for skin lesion segmentation," *Sensors*, vol. 21, no. 10, p. 3462, May 2021.

[58] S. Qamar, P. Ahmad, and L. Shen, "Dense encoder–decoder–based architecture for skin lesion segmentation," *Cogn. Comput.*, vol. 13, no. 2, pp. 583–594, 2021.

[59] Z. Zhang et al., "Optic disc region of interest localization in fundus image for Glaucoma detection in ARGALI," in *Proc. 5th IEEE Conf. Ind. Electron. Appl.*, Jun. 2010, pp. 1686–1689.

[60] S. Li, X. Sui, X. Luo, X. Xu, Y. Liu, and R. Goh, "Medical image segmentation using squeeze-and-expansion transformers," 2021, *arXiv:2105.09511*.

[61] Z. Zhang, H. Fu, H. Dai, J. Shen, Y. Pang, and L. Shao, "ET-Net: A generic edge-attention guidance network for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*. Cham, Switzerland: Springer, 2019, pp. 442–450.

[62] Y. Zhang, X. Cai, Y. Zhang, H. Kang, X. Ji, and X. Yuan, "TAU: Transferable attention U-Net for optic disc and cup segmentation," *Knowl.-Based Syst.*, vol. 213, Feb. 2021, Art. no. 106668.

[63] Á. S. Hervella, J. Rouco, J. Novo, and M. Ortega, "End-to-end multi-task learning for simultaneous optic disc and cup segmentation and glaucoma classification in eye fundus images," *Appl. Soft Comput.*, vol. 116, Feb. 2022, Art. no. 108347.
[64] M. Yeung, E. Sala, C.-B. Schönlieb, and L. Rundo, "Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation," *Computerized Med. Imag. Graph.*, vol. 95, Jan. 2022, Art. no. 102026.
[65] H. Wu, W. Wang, J. Zhong, B. Lei, Z. Wen, and J. Qin, "SCS-Net: A scale and context sensitive network for retinal vessel segmentation," *Med. Image Anal.*, vol. 70, May 2021, Art. no. 102025.
[66] J. Zhang, Y. Zhang, and X. Xu, "Pyramid U-Net for retinal vessel segmentation," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 1125–1129.
[67] Y. Zhang, M. He, Z. Chen, K. Hu, X. Li, and X. Gao, "Bridge-Net: Context-involved U-Net with patch-based loss weight mapping for retinal blood vessel segmentation," *Expert Syst. Appl.*, vol. 195, Jun. 2022, Art. no. 116526.
[68] Y. Zhou, Z. Chen, H. Shen, X. Zheng, R. Zhao, and X. Duan, "A refined equilibrium generative adversarial network for retinal vessel segmentation," *Neurocomputing*, vol. 437, pp. 118–130, May 2021.
[69] M. Chala, B. Nsiri, M. H. El Yousfi Alaoui, A. Soulaymani, A. Mokhtari, and B. Benaji, "An automatic retinal vessel segmentation approach based on convolutional neural networks," *Expert Syst. Appl.*, vol. 184, Dec. 2021, Art. no. 115459.
[70] I. Atli and O. S. Gedik, "Sine-Net: A fully convolutional deep learning architecture for retinal blood vessel segmentation," *Eng. Sci. Technol., Int. J.*, vol. 24, no. 2, pp. 271–283, Apr. 2021.
[71] M. Byra et al., "Breast mass segmentation in ultrasound with selective kernel U-Net convolutional neural network," *Biomed. Signal Process. Control*, vol. 61, Aug. 2020, Art. no. 102027.
[72] Y. Liu et al., "FCP-Net: A feature-compression-pyramid network guided by game-theoretic interactions for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 6, pp. 1482–1496, Jun. 2022.
[73] Z. Gu et al., "Ce-Net: Context encoder network for 2D medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2281–2292, Oct. 2019.
[74] J. Ni, J. Wu, J. Tong, Z. Chen, and J. Zhao, "GC-Net: Global context network for medical image segmentation," *Comput. Methods Programs Biomed.*, vol. 190, Jul. 2020, Art. no. 105121.
[75] J. Hu, H. Wang, J. Wang, Y. Wang, F. He, and J. Zhang, "SA-Net: A scale-attention network for medical image segmentation," *PLoS ONE*, vol. 16, no. 4, Apr. 2021, Art. no. e0247388.
[76] K. Chen, L. Yao, D. Zhang, X. Wang, X. Chang, and F. Nie, "A semisupervised recurrent convolutional attention model for human activity recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1747–1756, May 2020.
[77] M. Luo, X. Chang, Z. Li, L. Nie, A. G. Hauptmann, and Q. Zheng, "Simple to complex cross-modal learning to rank," *Comput. Vis. Image Understand.*, vol. 163, pp. 67–77, Oct. 2017.
[78] M. Luo, X. Chang, L. Nie, Y. Yang, A. G. Hauptmann, and Q. Zheng, "An adaptive semisupervised feature analysis for video semantic recognition," *IEEE Trans. Cybern.*, vol. 48, no. 2, pp. 648–660, Feb. 2018.

Jinting Wang received the B.S. degree from Southern Medical University, Guangzhou, China, in 2020, where she is currently pursuing the M.S. degree with the School of Biomedical Engineering.

Her research interests include medical image processing and deep learning.

Yujiao Tang received the B.S. and M.S. degrees from Southern Medical University, Guangzhou, China, in 2018 and 2021, respectively.

Her research interests include medical image analysis and machine learning.

Yang Xiao received the B.S., M.S., and Ph.D. degrees from the Huazhong University of Science and Technology, Wuhan, China, in 2004, 2007, and 2011, respectively.

He was a Research Fellow with the School of Computer Engineering and Institute of Media Innovation, Nanyang Technological University, Singapore. He is currently an Associate Professor with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology. His research interests involve computer vision, image processing, and machine learning.

Dr. Xiao was a recipient of the IEEE Innovation Spotlight Research Paper Award 2020, the EurAgEng Outstanding Paper Award 2018, and the Best Paper Award at ICIRA 2018. He also serves as the Associate Editor for *IET Image Processing*.

Joey Tianyi Zhou (Senior Member, IEEE) received the Ph.D. degree in computer science from Nanyang Technological University (NTU), Singapore, in 2015.

He was a Senior Research Engineer with the SONY US Research Center, San Jose, CA, USA. He was with IHPC. He is currently a Senior Scientist, PI, and Group Manager with the A*STAR Centre for Frontier AI Research (CFAR), Singapore. He is also holding an adjunct faculty position with the National University of Singapore (NUS), Singapore. His current interests mainly focus on machine learning with limited resources and their applications to natural language processing and computer vision tasks.

Dr. Zhou organized ICDCS'20-21 Workshop on Efficient AI meets Edge Computing, the ACML'16 Workshop on Learning on Big Data Workshop, and the IJCAI'19 Workshop on Multi-output Learning. He received the NeurIPS Best Reviewer Award in 2017. He is serving as an Associate Editor for the IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE (TETCI), the IEEE ACCESS, and *IET Image Processing*.

Zhiwen Fang received the B.S. and M.S. degrees from the Automation School, Beihang University, and the Ph.D. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2004, 2008, and 2017, respectively.

He was a Research Fellow with the Institute of Media Innovation, Nanyang Technological University, and a Research Scientist with the Institute of High Performance Computing, Research Agency for Science, Technology, and Research, Singapore. He is currently an Associate Professor with the School of Biomedical Engineering, Southern Medical University, Guangzhou, China. His research interests include medical image analysis, object detection, anomaly detection and machine learning.

Dr. Fang also serves as an Associate Editor for *IET Image Processing*.

Feng Yang received the M.S. degree in biomedical signal and image processing from Sun Yat-Sen University, China, in 1993, and the Ph.D. degree in communication and electronic systems from the South China University of Technology, China, in 1998.

He joined the Division of Image Processing (LKEB), Leiden University Medical Center, The Netherlands, from April 2010 to April 2011, as a Visiting Scholar. He is currently with the School of Biomedical Engineering, Southern Medical University, Guangzhou, China, as a Professor and the Director of the Department of Electronic Technology. His research interests include wavelet analysis, medical image processing, and pattern recognition.