Received February 13, 2019, accepted March 1, 2019, date of publication March 11, 2019, date of current version March 29, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2904094

Robust Boundary Segmentation in Medical Images Using a Consecutive Deep Encoder-Decoder Network

NGOC-QUANG NGUYEN AND SANG-WOONG LEE, (Member, IEEE)
Pattern Recognition and Machine Learning Laboratory, Gachon University, Seongnam 13120, South Korea

Corresponding author: Sang-Woong Lee (slee@gachon.ac.kr)

This work was supported by the Gyeonggi-do Regional Research Center (GRRC) Program of Gyeonggi Province (Analysis of behavior based on senior life log) under Grant GRRC-Gachon2017(B01).

ABSTRACT Image segmentation is typically used to locate objects and boundaries. It is essential in many clinical applications, such as the pathological diagnosis of hepatic diseases, surgical planning, and postoperative assessment. The segmentation task is hampered by fuzzy boundaries, complex backgrounds, and appearances of objects of interest, which vary considerably. The success of the procedure is still highly dependent on the operator's skills and the level of hand–eye coordination. Thus, this paper was strongly motivated by the necessity to obtain an early and accurate diagnosis of a detected object in medical images. In this paper, we propose a new polyp segmentation method based on the architecture of a multiple deep encoder–decoder networks combination called CDED-net. The architecture can not only hold multi-level contextual information by extracting discriminative features at different effective fields-of-view and multiple image scales but also learn rich information features from missing pixels in the training phase. Moreover, the network is also able to capture object boundaries by using multiscale effective decoders. We also propose a novel strategy for improving the method's segmentation performance based on a combination of a boundary-emphasization data augmentation method and a new effective dice loss function. The goal of this strategy is to make our deep learning network available with poorly defined object boundaries, which are caused by the non-specular transition zone between the background and foreground regions. To provide a general view of the proposed method, our network was trained and evaluated on three well-known polyp datasets, CVC-ColonDB, CVC-ClinicDB, and ETIS-Larib PolypDB. Furthermore, we also used the Pedro Hispano Hospital (PH^2), ISBI 2016 skin lesion segmentation dataset, and CT healthy abdominal organ segmentation dataset to depict our network's ability. Our results reveal that the CDED-net significantly surpasses the state-of-the-art methods.

INDEX TERMS Image segmentation, medical image segmentation, encoder-decoder network, boundary segmentation, continuous network, deep convolutional neural network.

I. INTRODUCTION

Currently, most of the medical object screening systems are manually operated by clinicians. Owing to the limitation of human vision and the low sensitivity and specificity of the systems, physicians can, therefore, miss the target object during the checking phase. Furthermore, undetected objects often have a diameter smaller than 9 mm, which cannot be observed and localized clearly by the clinicians. Besides, some objects are not detected because they are located in a dangerous area or are even hidden by an intestine fold. They may also be too flat and blurred in appearance to allow them to be seen visually. The high missing rate of clinicians can put patients' lives at risk. For example, in terms of colorectal cancer, according to the report of the American Cancer Society [7], the number of newly diagnosed cancer cases in the United State was approximately 97,220 cases of colon cancer and 43,030 cases of rectal cancer in 2018, and this number is increasing rapidly every year. Unfortunately, 50,630 deaths from colorectal cancer

The associate editor coordinating the review of this manuscript and approving it for publication was Dong Wang.

occurred in 2018. Besides, skin lesions are also a hot medical topic, especially as Melanoma is the most aggressive type of skin cancer and is responsible for a majority of skin cancer deaths [48]. Furthermore, in the US, according to a publication of the American Cancer Society, the estimated number of new cases of melanoma and deaths due to melanoma were 91,270 and 9,320, respectively. With regards to liver cancer, the estimated number of new cases of live cancer was 42,220 (including intrahepatic bile duct cancers) in 2018. Strikingly, liver cancer is about three times more common in men than in women. An estimated 30,200 liver cancer deaths occurred in 2018 [48]. The mortality trends of liver cancer have more than doubled, from 2.8 (per 100,000) in 1980 to 6.6 in 2015, with an increase of 2.5% per year from 2006 to 2015.

Early diagnosis of cancer can greatly reduce its associated mortality; for example, if melanoma is diagnosed in its early stages, it can be cured with prompt excision [1], [2]. The medical image analysis community has taken notice of these pivotal developments. However, the transition from systems that require manual manipulation to systems that learn features from the data has been increased gradually. To help clinicians make faster and more accurate decisions, automatic medical image segmentation approaches have been introduced, and for the last two decades, they have been the most successful methods for medical image analysis. Computer aided segmentation systems can significantly reduce the missing rates of medical objects and help clinicians identify regions of interest despite the complexity of the case. By using apriori knowledge models that contains feature information about the expected shape and appearance of the objects of interest, model-based segmentation methods strive to interpret this knowledge using smart algorithms that have prior knowledge about the object structures. Owing to the information in the dataset, a model-based segmentation method is more stable and more accurate compared to traditional methods, whose performances are sensitive to local image artifacts and perturbations. However, common networks are usually generic models developed for commercial and industrial applications, and if peculiar biological objects with unique attributes need to be detected, their performances are not high enough to allow clinicians to make correct decisions [3]. Therefore, many researchers have developed and investigated fast and precise medical object segmentation algorithms for providing early indications of the diagnosis. Nevertheless, because of restricted clinical requirements, their performances have not convinced doctors, especially as objects of interest always have unpredictable shapes and a large variety of sizes and aspects. Furthermore, in some cases, the shapes of wrinkles and folds are similar to those of tumors and objective cells. Moreover, the transition zone between the object and its surrounding area usually does not exhibit a significant change in texture or color that would enable clinicians to distinguish it from all other normal regions. In order to deal with these main problems, we primarily focused on building a deep convolutional neural network to generate discriminative features that are focused on object boundary regions and the tiny structures of objects.

In medical image segmentation, the pixels in a video or image are classified as object pixels or non-object pixels. Thus, the area to be considered for tracking or recognition is reduced from the entire image to several much smaller blocks. Traditional segmentation methods usually attempt to determine a suitable color space and build a model to classify each pixel individually. There are four main types of segmentation algorithms: explicit skin classifiers, non-parametric classifiers, parametric classifiers, and dynamic classifiers. Explicit skin classifiers, such as RBG, HSV and YCbCr classifiers [4], attempt to segment object points by defining decision boundaries in a color space. To overcome the problems encountered in previous segmentation methods that are triggered by different ethnicities of patients and varying illumination conditions, Long *et al.* [5] introduced deep fully convolutional neural networks (FCNNs), which have recently led to dramatic developments in semantic segmentation research. They can be adopted to recognize and understand images at the pixel level. In the medical field, because of their computational efficiency for discriminative feature extraction, FCNNs are used by many researchers for various purposes and have been shown to be effective when applied to many challenging datasets. Thus, they are have received attention from researchers who are studying approaches for improving medical image segmentation. Ronneberger *et al.* [8] improved FCNNs by introducing a new deep network called U-Net. The architecture, which is considered to be first encoder-decoder architecture, consists of a contracting path that captures context and a symmetric expanding path that enables precise localization. Due to the performance of this network in segmenting biomedical images, it has been widely used in the biomedical field. Besides, after recognizing that the hardest part of segmentation is the object contour, Chen *et al.* [9] proposed DCAN, which pays more attention to contour information by taking multilevel contextual features. Multi-level contextual features from the hierarchical architecture are explored with auxiliary supervision for accurate gland segmentation. Hence, the segmentation performance of DCAN is dramatically proved. These effective networks motivated us to develop a novel deep encoder-decoder network that can also focus on segmenting the boundary of an object.

As mentioned previously, there are two main existing problems in medical image segmentation fields: recognizing the transition space between the object region and non-object region and the large variety of segmenting object shapes.

When it comes to the first issue, unlike the contour of a common object, which clearly distinguishes the object of interest from the background, the boundary of a medical object is hardly defined. There are two causes of this problem. The first is the quality of the camera. Owing to the nature of the medical task, the camera usually goes inside the patient to take images of internal organs and tissue. Therefore, its size should be small, but unfortunately, the images

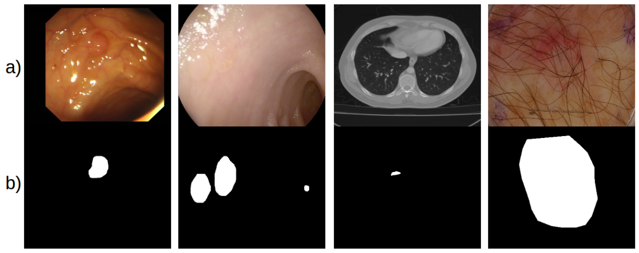

FIGURE 1. Examples for weak contract between the interest object and the surrounded background. a) Medical images. b) Corresponding labels.

taken by small cameras have distorted resolutions. Besides, for computed tomography (CT) images and other types of images, the ability to differentiate materials depends on the images' respective linear attenuation coefficients. Practically speaking, the quality of a CT image is strongly dependent on material properties such as density and atomic composition, the machine parameters of the X-ray spectrum utilized, and the signal-to-noise ratio. Endoscopists have claimed that even when good quality tools are used to take images, the object of interest is still difficult to find because they cannot distinguish between its boundary and the normal area. Hence, Chen *et al.* [9] modified the U-Net network for gland segmentation and inspired us to find a better segmentation method based on prior knowledge. Nevertheless, when applied in medical fields, the DCAN gives poor boundary segmentation, owing to its weak contrast and is therefore less effective than commonly used methods. To overcome this issue, rather than masking polyp boundaries by using fixed contours to explore complementary information [9], we randomly mask object boundary thickness and the objects' neighboring regions that do not differ widely from the object by casually changing pixel values inside the interest area. These masked regions are considered as new labels that provide richer information than previous augmentation methods in the medical segmentation task. Moreover, not only does our proposed augmentation technique enable the model to avoid the overfitting that occurs when the model learns about the object in a detailed manner, but it also enables the network to focus on extracting boundary patterns from each training polyp image. Therefore, instead of extracting only one contour in each training image, as presented in [9], our deep learning network can focus on extracting the most important features in various different parts of the object. Besides, we also propose a new loss function, which effectively calculates the difference between the prediction and the ground truth.

The second issue concerns medical objects' wide variety of appearances, such as their sizes, shapes, and structures. The medical segmented object size directly has an impact on the miss rates in object examinations, because doctors usually cannot easily evaluate small adenomas, which are tiny and difficult to see, yet they can later naturally become cancer tumors. Moreover, the physical size of the medical object is always unpredictable, and it can also be missed by the medical camera that is described above. This is because the distance between the camera and the object which is extremely unpredictable. Furthermore, in terms of the computer-aided detection system, the performance of the system highly depends on the training method, where a lot of important patterns could be missed or insufficiently trained. To the best of our knowledge, these differences render computeraided detection algorithms considerably less effective in real medical environments. Thus, CDED-Net is created such that its inputs are multi-resolution images, and it can also learn completely from training images. It is a combination of a cascade architecture of dilated convolutions and includes an effective decoder module. Our network architecture was inspired by the DeeplabV3+ network for the segmenting task [11] and Mix-nets, which was proposed by Davies and Moore [12]. The cascade architecture of dilated convolutions is used at the end of our network to extract multi-scale context information in local regions and does not require an increased number. This architecture can also effectively learn important information and recover parts related to object boundaries that are lost when the data passes through many convolution and pooling layers because the second network always learns patterns that were missed in the first network training phase. Our proposed method enlarges the perceived ability size without missing important information. Besides, by combining these techniques with our loss function, which will be discussed below, we found that the continuous encoder-decoder network can achieve a considerably better intersection over union (IoU) and give a better prediction.

We compared the performance of our proposed algorithm and its competitors using challenging datasets that were mostly provided by Grand-Challenges[1]. The results of extensive experiments revealed that our algorithms significantly outperform state-of-the-art algorithms. The main contributions of our work can be summarized as follows:

- We propose a new continuous multiple deep encoder-decoder network, CDED-Net, to extract the most useful features from images and learn completely from multi-scale image inputs.
- We introduce a boundary-emphasization augmentation method for making a high number of object boundary patterns from each image in a training set. The novel augmentation method enhances and boosts the segmentation performance of CDED-net.
- In our CDED-net, instated of using constraint dilated convolution, we use different both of strides and rates for each component network to capture contextual information at multiple scales input.
- We present a new Dicoss-loss function, which is a measure of overlap widely used to assess segmentation performance of a network. The combination of the loss function and our CDED-net results in a better performance.

[1] https://grand-challenge.org

The structure of our paper is as follows. First, in Section II, we briefly present the related state-of-the-art algorithms for polyp segmentation that highly motivated our research. Then, in Section III, we discuss our proposed approach to boundary-focused data augmentation and the processing of CDED-Net. Moreover, in this section, we also discuss a novel loss function. Subsequently, the experimental results on challenging databases are presented in Section IV. We summarize our work and describe our future work in Section V.

II. RELATED WORKS

In this section, we are going to briefly discuss the state-of-the-art related algorithms of medical object segmentation.

At the end of the 1990s, supervised techniques were used to develop systems for classification and object detection, which later rapidly became the methodologies of choice for analyzing medical images. Computer-aided diagnosis (CAD) has been used to assist doctors in diagnosing patients faster and more accurately at many hospitals. In particular, in the case of tumor/ lesion detection, the segmentation task plays an important role in medical object localization. It not only gives an output as a coordinate, but also can visualize the object's appearance. These functions of segmentation are extremely helpful when doctors want to diagnose cancer more accurately. Therefore, CAD segmentation applications are being used to precisely segment organs, cancer tumors, and polyps, which is a challenging task medical diagnosis. In the medical diagnostic process, CAD can wisely provide biopsy recommendations and decrease the failure prediction of doctors.

To use deep learning methods for medical image diagnosis, the computer must learn the features that represent the input data for the problem at hand. This concept is based on the basis of many deep learning algorithms: models (networks) composed of many layers that transform input data such as images and videos to outputs while learning increasingly higher level features [13]. The most successful and popular models for image analysis currently are convolutional neural networks (CNNs), which contain many layers that transform their training images using small convolution filters into a matrix called the features matrix. Lo *et al.* [14] applied CNNs to detect a lung nodule in what has been established as the first application of deep learning in the medical field. No sooner was the success of CNNs published than many researchers strove to develop and create new networks for several medical tasks. CNNs can be applied to classify each pixel in the image individually by presenting it with patches extracted around the particular pixel. A disadvantage of CNNs is that input patches from neighboring regions overlap considerably, increasing the number of features and computation time unnecessarily. To solve this problem, Long *et al.* [5] proposed fully convolutional networks (FCNNs) by rewriting the fully connected layers as convolutions so that the network could train with larger images. Moreover, instead of an output for a single pixel, the network is available to give a likelihood map as the result. Most state-of-the-art methods for semantic image segmentation using FCNNs are based on the idea of adding convolutional layers at the end of networks instead of using any fullyconnected layers. Ronneberger *et al.* [8] presented U-net architecture that comprises an FCNN and a decoder path, an upsampling part in which deconvolutions are used to increase the image size.

The segmentation of organs and other substructures in medical images allows quantitative analysis of clinical parameters related to volume and shape such as in polyp, liver, or skin image analysis.

First, in the polyp detection field, Wickstrøm *et al.* [10] enhanced fully convolutional networks (FCNs) for semantic segmentation by adding batch normalization [15] after each layer. Furthermore, the author proposed ESegnet, which is an improvement on SegNet [16] in which the encoder extracts useful features from an image and maps them to a low resolution representation and the decoder maps the low resolution representation back into the same resolution as the input image. Wickstrøm *et al.* [10], inspired by Kendall *et al.* [18], also included Dropout [17] after the three central encoders and decoders. Dropout was used to randomly set units in a layer to zero and can be interpreted as an ensemble of several networks. The addition of Dropout regularizes the model and also enables estimation of uncertainty in the model's prediction. Akbari *et al.* [6] proposed a novel polyp segmentation method that is strongly based on cascading of the network. The authors used a smart patch selection method in the training phase of the network to enhance the performance of the model. Moreover, after using the modified FCN, namely FCN-8S, for segmentation of polyp regions in colonoscopy images, the authors used the Otsu thresholding method to change the probability map output by FCN-8S into a binary image and then found the largest connected component. By using the post-processing method, the number of false positive pixels was slightly decreased. Zhang *et al.* [19] presented a hybrid classification-based method for fully automated polyp segmentation. More specifically, they applied two initial steps: region proposal generation and region area finement. The hierarchical features of polyps were learned by the FCN, while the context information related to the polyp boundaries were modeled by texton patch representation. After the FCN provided a pixel-wise prediction and the initial polyp region candidates, the latter was refined by patch-wise classification using texton based spatial features by using the random forest method. By combining three well-known convolutional networks, which are AlexNet [20], GoogLeNet [21] and VGG [22], Brandao *et al.* [23] refined FCN architectures to recognize specific structures in colonoscopy images. This FCN learned end-to-end, whereby density predictions were obtained by deconvolution layers. Following the success of the FCN in a polyp segmentation computer-aided system, Li *et al.* [24] proposed a new end-to-end FCN structure that was inspired by U-net [8] and FCN [5]. The method can directly give a prediction map that has the same size as the original testing image of the input network without any post-processing methods. The beginning stage is the feature extraction stage and the last stage is the prediction map reconstruction stage. The extraction function extracts low-level features from the input picture and the feature map is permutated and later combined with the continuous convolution operation to produce an abstract, high-level feature map with semantic information.

Second, with respect to the automatic skin lesions segmentation, Vasconcelos *et al.* [25] proposed a morphological geodesic active contour segmentation (MGAC) method with automatic initialization using a mathematical morphology that is known as a great partial differential equation approximation. The method is automatic, with a lower computational cost and no stability problems. More specifically, by using only the blue channel of the dermoscopic image, the method adapts to the contours efficiently owing to the information storage of the channel about the lesion. To speed up the processing time, Vasconcelos *et al.* [25] initialized the geodesic active contour automatically; it gave both a starting point for the beginning of the GAC and a contour that is very close to the lesion. Furthermore, because the training process was not needed, this method was faster than the deep learning method. In addition, this method was also able to manage the noise in images, such as oil bubbles and hairs. In terms of improving the performance of FCN in skin lesions segmentation, Bagher Salimi *et al.* [26] proposed DermoNet, which is an FCN achieved by transforming DenseNets. To take advantage of the capability of high-level feature representations, Bagher Salimi *et al.* [26] used densely connected convolutional blocks and skip connections. With this approach, network layers can, again and again, use the information that is output by their preceding layers. This allows for a loss function to penalize the multi-scale feature maps from different layers. Their proposed network uses fewer parameters; hence the model can quickly achieve a high training accuracy. Furthermore, because the architecture adopts multiple dense blocks in the encoder process, the DermoNet can allow for multi-scale feature maps to be penalized by a loss function [26].

Third, regarding segmentation of liver CT images, traditional methods, such as clustering [27]–[29], and morphological operators [30] used to help clinicians classify texture in feature space. These intensity-based methods are usually efficient and can give excellent results when the liver's intensity is sufficient. The main problem with these traditional methods is that there is no formal shape; thus, boundaries of the liver cannot be defined and sometimes important parts are missed. To address these drawbacks, Yan *et al.* [31] proposed an atlas-based method in the setting of hepatic fat fraction assessment. The fat-fraction map is calculated by using the chemical shift-based method in the delineated region of the liver. Besides, Chartrand *et al.* [32] proposed a method that consists of three main phases: initialization, optimization, and correction. In the initialization step, an initial shape is interpolated from generated contours. This shape is later optimized to converge toward the liver boundary. Lastly, the 3-D surface mesh can be interactively manipulated to obtain a high precision. Lu *et al.* [33] developed a fully automatic liver segmentation framework by combining deep learning and the graph cut approach. Firstly, the author used 3D CNN to simultaneously locate and segment the raw liver surface. Then the graph cut method was used to refine the initial segmentation result.

To surpass these state-of-the-art methods, we propose a method that is effectively used for medical object segmentation for both CT, MR, and common format images. More specifically, after we applied our novel data augmentation method, which we named boundary-emphasization, the augmented data was carried to the CDED-net. In this paper, we also present the new loss function to bias towards the background image rather than the medical interest object.

III. PROPOSED METHOD

In this section, we describe the methodology on which the proposed method is based, including the set of sample images used for evaluation, as well as detailing the theoretical basis of the proposed medical object segmentation. First, we present the novel medical segmentation data augmentation method. Second, we incorporate all the augmented datasets into the CDED-net to teach the model to discriminate the background and foreground. In the last step, we propose a Dicoss-cost function that can effectively boost the segmentation performance of our network.

A. BOUNDARY-EMPHASIZATION DATA AUGMENTATION

Data augmentation is the creation of altered copies of each instance within a training dataset to increase the number of images. A growing challenge of researchers is how to avoid the over-fitting problem that can mislead the deep convolutional neural networks. Researchers are striving to solve this problem and achieve better results by modifying the network's architecture, developing new learning algorithms, and acquiring the data. The most common problem is the lack of good-quality data or uneven class balance within the datasets. Currently, the most effective segmentation networks are very large, hence requiring a large amount of data, which is difficult to obtain [34]. Therefore, data augmentation is an essential step for improving segmentation performance. Recent studies have demonstrated the robustness of data augmentation by generating additional data using the original limited training dataset [35]. It brings these training images into the larger featured space where they can fulfill all their variances. In this work, we used geometric augmentation techniques including reflection, random cropping, translation, and rotation. In particular, we applied elastic distortion which was introduced by Wong *et al.* [36]. The elastic deformation was performed by defining a normalized random displacement field $u(x, y)$ that for each pixel coordinate (x, y) in the image denotes a unit displacement vector, such that $R_w = R_o + \alpha u$, where R_w and R_o denotes the location of pixels in the original and warped images respectively [36]. The strength of the displacement in pixels is given by α.

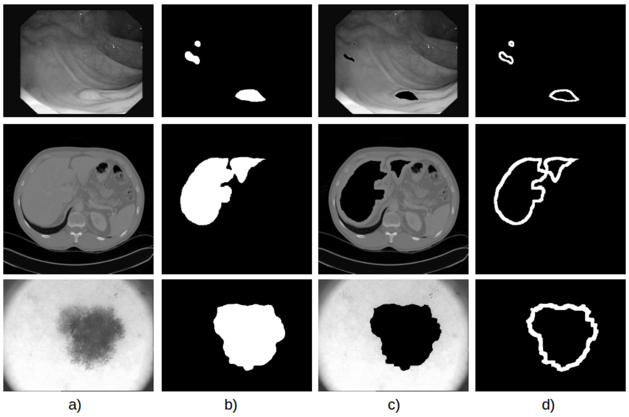

FIGURE 2. Examples of Boundary-emphasization augmentation method. a) Medical images. b) Corresponding labels. c) Processed medical images. d) Processed corresponding labels.

In our experiments we gave $\alpha = 1.12$. The parameter σ, which is the standard deviation of the Gaussian, is convolved with matrices of uniformly distributed random values that form the x and y dimensions of the displacement field u. In medical segmentation, the color of images significantly varies across laboratories as a result of technicians' varying technical skills; therefore, we adopted an effective color constancy method, namely, gray world, which assumes the scene in an image, on average, is a neutral gray, and the source of average reflected color is the color of the light. This technique also enhances the contrast between the object of interest and surrounding areas. We used the following the formula to transform all our experimental datasets into the grayscale format:

$$U_{(x,y)} = 0.2989 * R_o + 0.5870 * G_o + 0.1140 * B_o \qquad (1)$$

where R_o, G_o, B_o represent the red, green, and blue values of the pixel at position (x, y) in the image, respectively. While the $U_{(x,y)}$ denotes a new value in the gray world.

Owing to its medical characteristics, the non-specular transition zone between the medical object and its surrounding area is not easy to discriminate with conventional segmentation methods. This area does not differ dramatically from other areas. Furthermore, especially in the endoscopic field, not only are the folds and wrinkles shapes of the zone similar to those of the tumors, it can partly hide and sometimes overlap the object of interest. To artificially locate the medical object boundary and improve the performance of our CDED-net, we present a new boundary-emphazization augmentation method that can be combined with most existing deep convolutional neural networks to boost the learning ability of the network. After detecting the coordinate of the object in the ground truth images, we apply the erosion method to remove the inner part of the object. Subsequently, we subtract the part which we produced in the previous step from the original images to create the boundary label. In other words, we just delete the inside part of the object of interest to create a foreground with object boundaries. To enlarge the perception capacity of the model, we set the contour thickness arbitrarily. In Figure 2, we introduce some typical examples in three dataset from the top to the bottom, alternately: CVC-ClinicDB [37], liver segmentation dataset [57], PH^2 [38].

Algorithm 1 Boundary-Emphazization Augmentation Algorithm

```
Input: Input image $I$, corresponding label $I_L$;
       Object region $S$;
       Structuring element (erosion kernel size) $C$;
       Erosion represented by $\ominus$;
       Euclidean space $E$;
       Erosion label $I_L^E$;
       $C_z$ is the translation of $C$ by the vector $z$, $\forall z \in E$;
       Processed label $I_L^*$;
Output: Processed image $I^*$, corresponding processed label $I_L^*$
while True do
    $I_L^E = I_L \ominus C = \{z \in E \mid C_z \subseteq I_L\}$;
    $I_L^* = I_L$ - $I_L^E$
    $i(x_I, y_I) \subset (I)$;
    if $i^*(x_I, y_I) \subset I_L$ and $i^*(x_I, y_I) \subset S_{I_L^E}$ then:
        $i = 0$;
    else:
        pass;
result $I^* \leftarrow I, I_L^* \leftarrow I_L$;
```

The entire process of the boundary-emphazization augmentation method is in Algorithm 1. This method is typically applied to binary images whose label format is known. The basic effect of the operator on a binary image is to erode away the boundaries of regions of foreground pixels in object region S (i.e. white pixels, typically) by structuring element C_z. In the first step, the set of Euclidean S coordinates corresponding to the input binary image, namely C_z (also known as a kernel), is the set of coordinates for the structuring element. The erosion of S by C_z can be understood as the locus of points reached by the center of C_z when C_z moves inside S. Then, no sooner does this erosion process provide smaller interest objects, namely erosion labels, from the original image than we deduct the original label L_L to erosion label I_L^E to take the new label. Finally, we turn every pixel $i(x, y)$ in both images and equate the corresponding labels where it has the coordinate (x, y) inside I_L^E into zero. We named this procedure the boundaryemphazization process. Strikingly, when we set the thickness of the boundary randomly, we achieve better results than when we use a constant thickness.

B. CONSECUTIVE DEEP ENCODER-DECODER NETWORK

The encoder-decoder networks have been successfully applied to many computer vision tasks, including semantic segmentation [5], [8], [16], [39]. Recently, the encoder-decoder network has become one of the most effective

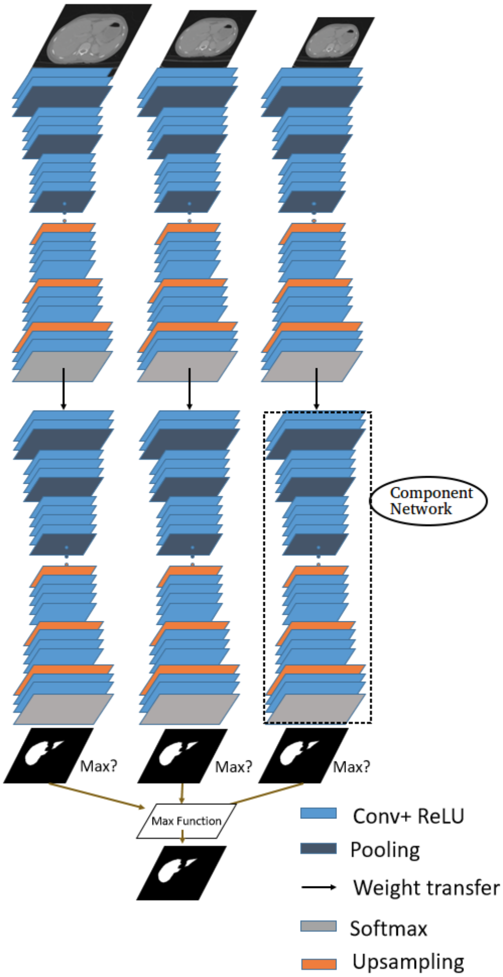

FIGURE 3. The entire architecture of proposed CDED-net.

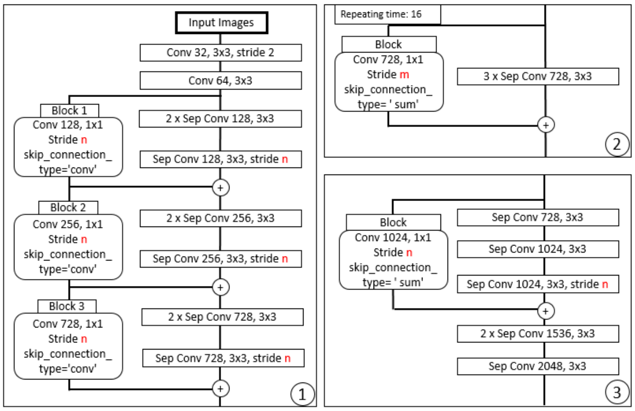

FIGURE 4. The detail of the component proposed CDED-net with Deeplab V3+ [11] as the backbone. 1) Entry flow. 2) Middle flow. 3) Exit flow. We modified the convolution stride to adapt with the resolution of dataset for extracting rich information features.

structures in the segmentation task. Hence, the advantages and disadvantage of DeepLabV3+ [11] inspired us to develop the proposed network. The objective of this study is to build an ensemble of deep encoder-decoder networks to train and obtain rich contextual information for the medical object segmentation task that is shown in Figure 3. Each DEDN does a part of the job of the main model. In other words, a single DEDN is employed to deal with its problem. By comparing the proposed approach with the previous approach, we found that our ensemble network had better segmentation performance than a single network. This may be because our network can not only take discriminative features from the first three networks, but also learn information from missing patterns by using the three last DEDNs.

To capture contextual information at multiple scales, we used deep encoder-decoder networks, namely DeepLab V3X [11], which has several parallel atrous convolutions with different rates, but we also put three types of resolution training images into the network to enlarge the perceived ability of the network to better cover global features. First, DeepLab V3+ [11] is considered as a component network because it has shown an outstanding performance in the PASCAL VOC 2012 challenge [40]. Besides, by using this backbone, we take advantage of the downsampling path that basically contains convolutional and max-pooling layers, where these layers are extensively used in the convolutional neural networks for image classification tasks [20], [41]. Furthermore, the upsampling path contains convolutional and deconvolutional layers that are also known as backwards stride convolution layers [42]. To recover the output score masks and feature maps in their original sizes, we used deconvolutional layers. We did this because the downsampling path aims to extract useful abstraction information, while the upsampling path gives the prediction in the score masks. Moreover, we expand Chen *et al.*'s [11] network further by harnessing multi-level contextual feature representations, which include various levels of contextual information, i.e., intensities appearing in different sizes of perceived ability.

In our proposed CDED-net, instead of using the constant dilation stride we adopt different stride for each component network for denser feature extraction. For instance, the network that is trained with small resolution images we apply dilation stride $m = 1$ and dilation stride $n = 2$ to the last two blocks respectively, which are shown in Figure 4. However, the network is trained by traditional size images we use dilation stride $m = 2$ and dilation stride $n = 4$, while the remaining one we apply dilation stride $m = 3$ and dilation stride $n = 6$. This is because the small image is, the rich information it contains. Besides, we also adopt different atrous rates for each member network to enlarge the perspective field optimally. Meanwhile, the global features are most needed to extract from large image. By using this strategy, out network can effectively extracts most discriminative features. Furthermore, our architecture is highly inspired by the fact that dilated convolutions significantly support exponentially expanding receptive fields without losing coverage [50]. Let $G_0, G_1, G_2 \ldots, G_{n-1} : \mathbb{Z}^2 \rightarrow \mathbb{R}$ be discrete functions and $f_0, f_1, f_2 \ldots f_{n-2}$: $\Omega_1 \rightarrow \mathbb{R}$ be discrete 3×3 kernels. The receptive

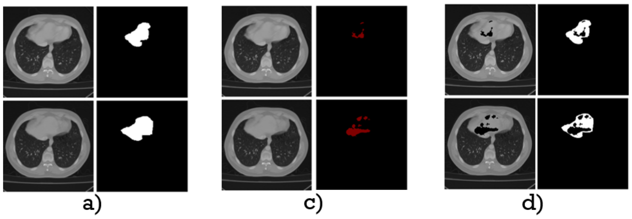

FIGURE 5. Proposed training images for the proposed approach on [57] dataset. a) Images and corresponding labels in three head networks. b) Results in first validation step. c) Images and corresponding labels in three tail networks.

field is showed in Equation 2:

$$G_{i+1} = G_{i*2^i} f_i \quad \text{for } i = 0, 1, 2, \ldots, n-2. \tag{2}$$

Second, unlike in previous research studies, where pre-trained deep learning models were used to extract discriminative features, in this work we trained three first single DEDNs using their pre-trained models and our augmented training dataset. After these training processes, our models can extract considerably better features than their pre-trained models. The training steps are explained in detail below:

1) In the first step, we train our network with our augmented dataset to make the model pay attention to the medical object of interest. In this phase, we separately trained the first group of networks with three different resolution images, for instance, with the liver dataset, the image sizes are 640 × 640, 512 × 512 and 384 × 384 pixels. We also used off the shelf model that is provided by the COCO dataset [43] for the training process as a pre-trained model.
2) Second, after finishing the first training phase, all weights are stored and later used for validation purposes. In the validation stage, we validate the performance of these first networks with the original training images I; we then perform the subtraction stage in which we applied this method $\ominus$ to seek the unknown patterns I_e. Here, we basically subtract the former from the prediction images I_v of the first model. Subsequently, these patterns $I_e = I$ - I_v aare considered as the dataset for the second group of networks. Our training dataset in different phase has been shown in Figure 5. The models that are known as the products of the first training step are added to the network to further strengthen the training process, as shown in Figure 3. This can effectively decrease the problem of vanishing gradients with auxiliary supervision. In this step, we can not only check the missing patterns, but also correct the misunderstanding part from the images. Figure 6 shows the effectiveness of your method with respect to the respective field of the model. Through the last validation, it was shown that the last models do not make a mistake in the same objects by comparing with previous models.

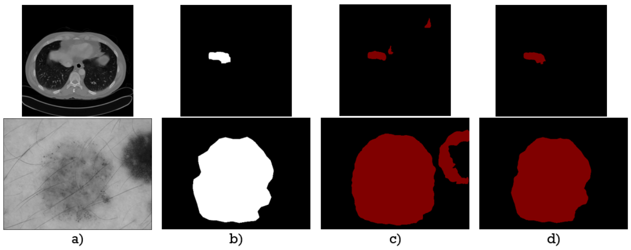

FIGURE 6. Examples from original training images in first validation step (in weight transfer phase) and final validation in dataset [38], [57]. a) Training images. b) Training labels. c) Results in first validation. d) Results in final validation.

3) Finally, after processing, a soft-max layer outputs the probability that each voxel belongs to the foreground and to the background. In particular, in the medical field, such as in skin segmentation, the anatomy of interest usually occupies only a very small region of the scan. This often results in the minima of the loss function affecting the learning process, which yields a network that gives predictions that are mainly biased towards the background. Therefore, our network with multi-level contextual features extracted from input I can be trained by minimizing the overall Dicoss-loss L_c between the predicted results and ground truth annotation, which we discuss in the next section. Besides, regarding a larger object, the model which is trained with large scale images could give better results than the remaining model. However, by training the model with small resolution images, the small-model produced exceptional results with small objects.

C. DICOSS LOSS FUNCTION

To provide better segmentation results, we proposed a novel simple loss function that is a combination of two well-known cost functions and hyperparameters to perform the segmentation. Since Ronneberger *et al.* [8] described the use of the pixel-wise cross entropy loss for the task of image segmentation, it has been adopted widely. This loss simply verified each pixel individually, comparing the class predictions that are defined as depth-wise pixel vector to the the target vector. Because this loss function asserts every single pixel, this may create a problem if various classes are represented in the image. However, medical images usually have a low surface area. Consequently, the segmentation network trained with a cross-entropy loss function is biased towards the background image rather than the object itself. Furthermore, as the foreground region is often missing or only partially detected, it is not easy for the model to see the object. Hence, we combined the function with a dice loss function to reduce the negative aspects of the former. This is because this function can strongly measure the overlap between two objects, one is a prediction and the remaining is ground truth.

$$L_c = -(1-\gamma) * \sum_{i,j} \hat{y}_{i,j} * log(y_{i,j}) + \gamma \frac{2 \sum_{i,j} \hat{y}_{i,j} * y_{i,j}}{\sum_{i,j} \hat{y}_{i,j}^2 + \sum_{i,j}^{N} y_{i,j}^2} \tag{3}$$

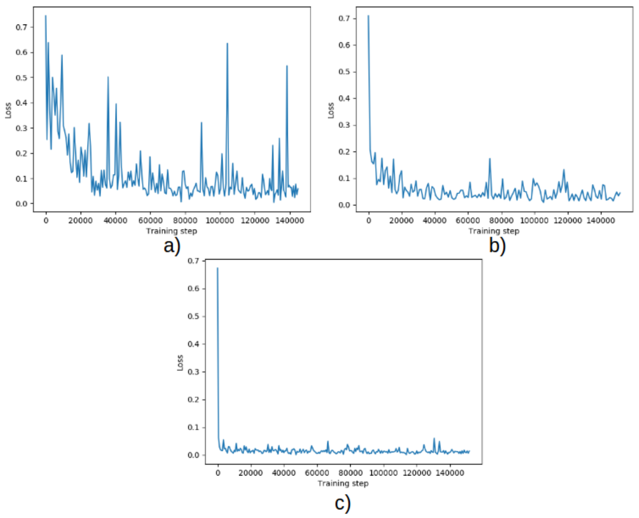

FIGURE 7. The effect of proposed loss function to network learning progress on the same dataset by comparing to two fundamental loss functions. a) Basic cross entropy loss function. b) Basic Dice loss function. c) Proposed loss function.

where $\hat{y}_{i,j}$ the predicted binary segmentation volume and $y_{i,j}$ stands for the ground truth at image pixel (i, j), while hyperparameter γ is used for balancing. To identify which pixel is background and foreground we follow the condition:

$$\begin{cases} T_{i,j} \leq L_c \leq 1 & \text{if } z_{i,j} \in E \\ L_c < T_{i,j} & \text{if } z_{i,j} \notin E \end{cases} \tag{4}$$

where the E_i denotes the interesting area, while $T_{i,j}$ represents for the threshold at pixel coordinate (i, j).

This loss function is able to give a smoother segmentation prediction. Nevertheless, in order to prevent a network from being biased toward the negative class and to clearly predict all zero pixels, we added the γ hyperparameter. Our experimental results prove that this loss function is more robust compared to the classical cross-entropy loss function and basic dice loss function. Figure 7 describes the comparison between our proposed loss function and two fundamental loss functions. Moreover, it is properly suited to the imbalanced classes of the foreground and background. In this Figure we used TensorBoard Visualization to export the loss parameter during the training process and then later draw in Python.

IV. EXPERIMENTS RESULTS AND ANALYSIS

In this section, we study the performance of our proposed segmentation approach. We used six databases to demonstrate our methods and compared our results with those of state-of-the-art algorithms.

A. DATASETS

To evaluate the proposed segmentation method and compare it with the other competitor methods, we used well-known datasets from the MICCAI 2015 polyp detection challenge [44] in colorectal segmentation. Moreover, in terms of skin lesion segmentation and liver segmentation, we report the results highlighted in [38], [56], and [57].

The datasets are briefly described in the following paragraphs.

- CVC-ClinicDB [37] contains 612 images, where all images show at least one polyp. The segmentation labels obtained from 31 colorectal video sequences were acquired from 23 patients.
- CVC-ColonDB [46] ontains 379 frames from 15 different colonoscopy sequences, where each sequence shows at least one polyp each.
- ETIS-LaribPolypDB [45] contains 196 images, where all images show at least one polyp.
- PH^2 [38] contains 200 dermoscopic images with a resolution of 768 × 560 pixels that were acquired at Dermatology Service of Hospital Pedro Hispano, Matosinhos, Portugal Mendonça with Tuebinger Mole Analyzer system, this dataset includes 80 common nevus images, 80 atyp-ical nevus images and 40 melanoma image
- ISBI 2016 [56] contains 900 training imageswith the ground truth provided by experts. The image sizes vary from 1022 × 767 to 4288 × 2848 pixel. This dataset was provided at the 2016 International Symposium on Biomedical Imaging (ISBI 2016).
- CHAOS 2019 [57] contains 980 liver CT images with re resolution is 512 × 512 pixel in DICOM format. This dataset was provided at the IEEE International Symposium on Biomedical Imaging (ISBI) on April 8-11, 2019.

B. CALCULATION METRICS

We used the Jaccard index, also known as the intersection over union (IoU), as the main metric to evaluate the proposed approach's performance. Furthermore, in order to provide a general view of the effectiveness of our method, we also employed Dice score, sensitivity (Sen), specificity (Spec) and accuracy metrics to describe our results. We used these metrics to compare our prediction results (PR) with the ground truth (GT). The former based on the confusion matrix includes true positives (TP), which are the correctly predicted pixels, false negatives (FN) values, which are object pixels that are identified as the background, false positives (FP), which are background pixels classified as objects, and true negatives (TN), which are background pixels that are correctly segmented.

We calculated the mean IoU parameter. Each per-class IoU was computed over a validation/test set according to Equation 5. This is used to calculate the similarity between the GT and the PR proposed by the method.

$$IoU = \frac{PR \cap GT}{PR \cup GT} \tag{5}$$

where ∩ denotes a set of an intersection and ∪ a union set between PR and GT.

Notably, the Dice coefficient is a statistic used for comparing the similarity of prediction images and label images, it is shown in Equation 6. Interestingly, this is also called Dice similarity coefficient and is slightly different with IoU because it can effectively measure the similarity in more heterogeneous datasets while still retaining its sensitivity.

$$Dice = \frac{2PR \cap GT}{PR \cup GT} \tag{6}$$

n addition, we used specificity (Spec) to represents the proportion of the negative values produced by the segmentation method and the values that are real negatives belonged to GT in the Equation 7.

$$Spec = \frac{TN}{TN + FP} \tag{7}$$

Sensitivity (Sen) which is also known as Recall, is the metric which basically measures the proportion of the positive values considered by the segmentation method and the right positives values given by the GT, and it is presented by Equation 8.

$$Sen = \frac{TP}{TP + FN} \tag{8}$$

We also used accuracy (Acc) as one of the main metrics to show how big the gap between the GT and PR of the methods is and the relation between their hits and errors in Equation 9. More specially, the higher the Acc, the better the segmentation methods are. A high Acc shows that most of the pixels were classified correctly.

$$Acc = \frac{TN + TP}{TN + TP + FN + FP} \tag{9}$$

We applied the common metrics to show our results in comparison which are presented in Equation 10. This metric is the ratio of correctly predicted positive observations to the total predicted positive observations.

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

C. COMPARISON WITH OTHER STATE-OF-THE-ART APPROACHES

We designed several experiments, the results of which showed that using grayscale images for both the training and the testing phase is always better than using RGB images. Thus, first of all, we converted the formats of all the images in the training dataset to grayscale. Then, these images were upsampled and downsampled to feed into three different training phases as we mentioned above. Besides, not only did we apply our proposed augmentation method, but in the training phase, we also used other effective data augmentation methods. For instance, a cropping method was used; that is, we cropped from the center to remove the black parts generated by the camera that exist at image corners. The rotation method was applied using random degrees between 0° to 360°. Therefore, we proceeded to use all the images in the datasets. We formatted the datasets in the TFrecords format to optimize the learning ability of the model. Subsequently, after 150000 first training steps with the augmented dataset, the trained models were tested on the original training sets to identify the missing and the incorrect recognition patterns in the prediction images before they were used as input for the next step. As soon as we finished the validation step, these prediction images were subtracted from the original training set to determine the missing parts and to recorrect the failed classification parts. In the following stage, the products of this step were selected for augmentation again and trained with three tail networks. Finally, we tested the performance of the models with three resolution images and used the map score to select the best result. Furthermore, all training processes were executed on TensorFow with a GeForce GTX TITAN X graphics card.

TABLE 1. Comparison of proposed method and three fully convolutional neural networks in terms of mean pixel precision and recall for the ETIS-Larib dataset [45].

Networks	Mean pixel precision	Mean pixel recall
FCN-AlexNet [20]	0.2789	0.3554
FCN-GoogLeNet [21]	0.2583	0.3782
FCN-VGG [22]	0.7023	0.5420
Proposed	**0.9293**	**0.9087**

TABLE 2. Comparison of proposed method with FCN-8S combined with post-processing and a combination of fully convolutional neural network and textons on CVC-ColonDB dataset [46].

Networks	Accuracy	Specificity	Dice	Sensitivity
Zhang et al. [19]	0.975	0.988	0.701	0.757
Akbari et al. [6]	0.977	**0.993**	0.810	0.748
Proposed	**0.980**	0.991	**0.896**	**0.792**

1) RESULTS ON THE POLYP SEGMENTATION DATASET

With regard to polyp segmentation, to compare our method with that in [47], instead of using the combination of images in the MICCAI-challenge datasets, which include 19514 frames, we trained our proposed network with only one dataset [46], which has 379 frames images for the training process. Our proposed method was used to segment polyps contained in the ETIS-Larib dataset of the MICCAI challenge. In the beginning, we mutated all images in the dataset to grayscale images. Then, the images in the dataset [46] were resized from 384 × 288 pixels to 500 × 400 pixels and 300 × 200 pixels. Our results in terms of polyp segmentation are presented in Table 1. The table shows that our proposed model achieved both the highest precision and the highest recall among the models. The experimental results also reveal that the fused method outperformed all the other approaches because of its ability to aggregate multi-scale contextual information.

Moreover, we also evaluated our network's performance on the well-known dataset CVC-ColonDB, as shown in the Table 2. In our approach, the post-processing method was not adopted for fine-tuning the predictions. However,

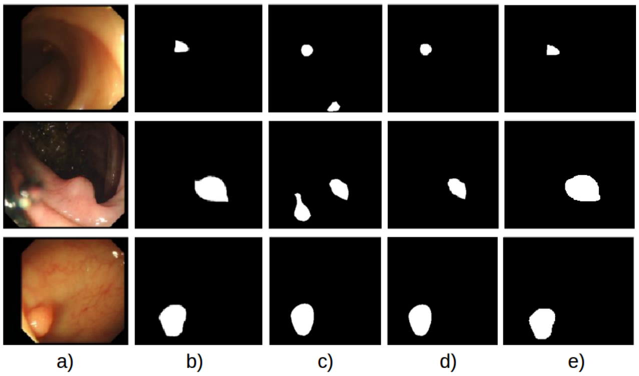

FIGURE 8. Comparison of proposed method with [6]. a) Input images. b) Ground truth. c) FCN-8S with Otsu threshold. d) FCN-8S final result. e) Proposed method.

TABLE 3. Comparison of proposed method on CVC-ClinicDB dataset [37].

Networks	Accuracy	Specificity	Sensitivity	Dice	Precision
Li et al [24]	0.97	0.77	**0.99**	0.83	0.90
Proposed	**0.987**	**0.942**	0.962	**0.891**	**0.950**

Akbari *et al.* [6] applied the erosion method to make predictions smoother. This proves once again that this model can exploit the deep network architecture of multi-resolution input images to aggregate multiscale contextual information, and this attribute can be used to fit it to single models. This table indicates that the performance of our approach is better than that of the second competitor 0.86 in Dice coefficient metrics. The Figure 8 shows that our model can recognize the tumor boundary optimally and achieve a result that the other models can not achieve.

Furthermore, we also evaluated the learning ability of our CDED-net through experiments with one of the most challenging datasets in the polyp segmentation field, CVC-ClinicDB, which owes its reputation not only to its resolution, but also to the wide variety of polyp images that it contains. From this dataset we arbitrarily selected 430 images randomly for training and the remaining 182 images were used as the test set, as was done by Li *et al.* [24]. Therefore, there is no intersection between the training set and the test set. Table 3 demonstrates that the our proposed deep learning network significantly outperforms that proposed by Li *et al.* [24]. This gap in performance can be explained by the weight transfer step, as the presented method of Li *et al.* [24] still could not fully detect interesting objects while our model can segment the object contour entirely.

2) RESULTS ON THE SKIN LESION SEGMENTATION DATASET

To compare the performance of our method with state-of-the-art methods in the skin lesion segmentation field, we also analyzed the performance of the proposed segmentation networks with DermoNet presented in [26] superpixel-based saliency detection approaches that were presented in [48], and several well-known networks on the PH^2 dataset.

TABLE 4. Comparison of proposed method on PH^2 dataset [38].

Networks	IoU	Dice	Sensitivity	Specificity	Accuracy
MGAC [25]	87.03	92.79	93.59	97.81	**96.18**
FCN [5]	82.59	90.46	95.35	94.09	94.44
U-Net [8]	81.63	89.88	86.68	97.63	86.68
SegNet [16]	84.03	91.32	91.57	96.57	95.19
FrCN [49]	84.13	91.38	94.48	95.46	95.20
MSCA [52]	76.88	85.52	85.78	96.33	93.86
MFCN [53]	84.15	90.77	95.64	95.12	95.61
DCL-PSI [51]	86.05	92.26	**97.11**	95.85	96.61
Tong et al. [48]	60.0	75.0	-	-	-
DermoNet [26]	85.3	91.5	-	-	-
Proposed	**88.78**	**94.10**	96.23	**97.84**	95.40

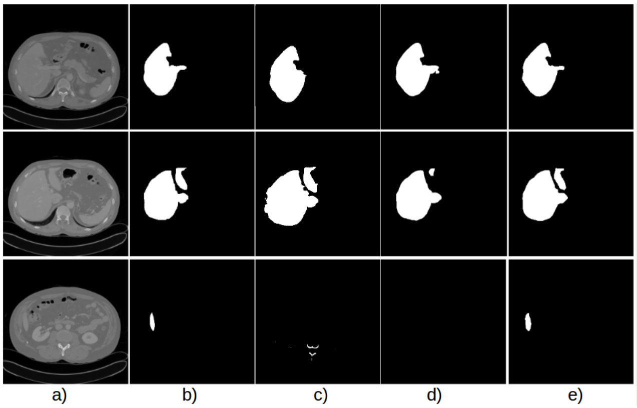

FIGURE 9. Comparison of proposed method on CHAOS dataset [57]. a) Input images. b) Ground truth. c) U-net [8]. d) Deeplab V3+ [11]. e) Proposed method.

Furthermore, we also compared our approach to a powerful method proposed by Vasconcelos *et al.* [25]. In order to evaluate the general segmentation performance of the proposed networks, it was also compared with methods that use machine learning techniques and deep learning. These methods, which include FrCN [49], FCN [5], MFCN [53], MSCA [52] and DCL-PSI [51], achieved exceptional results in skin lesion segmentation. In this comparison, we used the training set that was provided by the International Skin Imaging Collaboration (ISIC) for the ISBI challenge titled "Skin Lesion Analysis Towards Melanoma Detection" and tested the proposed network on 200 skin images from the.The training and test strategy are the same as those presented in [25] and [26]. The comparative results are listed in Table 5. The results reveal that the proposed method significantly outperforms most stateof-the-art methods in two important criteria. The results also reveal that our fused model achieved a Dice score and IoU of 1.75% and 1.31%, respectively, which were better than those of the most effective image processing technique. This is because the proposed method uses a weight transfer step that reminds the network to entirely focus on missing patterns inside the foreground region.

3) RESULTS ON THE CT LIVER SEGMENTATION DATASET

In this section, we adopted the dataset of the CHAOS-Combined (CT-MR) Healthy Abdominal Organ

TABLE 5. Comparison of proposed method on CHAOS dataset [57].

Networks	IoU	Dice	Recall	Precision	Accuracy
U-net [8]	91.52	95.57	93.59	89.51	97.32
Deeplab V3+ [11]	93.36	96.69	93.50	93.46	99.04
Proposed	**96.70**	**98.37**	**94.13**	**96.80**	**99.46**

Segmentation challenge that will be held at The IEEE International Symposium on Biomedical Imaging (ISBI). In computed tomography (CT) images usually acquired for liver diagnosis and monitoring purposes, the intensities of adjacent organs and tissue are extremely similar to those of liver tissue itself. This is often the case for the boundaries of the stomach and heart, but also for the boundary of the subcostal fat of the rib cage [55]. In these problematic regions, automatic segmentation of the liver based on grayscale values alone is very challenging. However, the proposed CDED-net once again proves that it can strongly distinguish the liver features using a similar pattern. In this comparison, we used 800 images for training and 180 images for testing purposes. We apply same augmentation methods with the proposed networks and both two competitive networks. Nevertheless, our network also can correctly recognize very small objects, as shown in 9. Moreover, Table 5 indicates that our proposed method dramatically outperforms the most recent DEDN networks. Ronneberger *et al.* [8] and Chen *et al.* [11] also achieved considerably performance with this dataset, but still could not fully detect the liver.

V. CONCLUSION

In this paper, we proposed a novel approach that uses an ensemble of multimodel deep encoder-decoder networks, called CDED-Net, for medical object segmentation. We also presented a new data augmentation method called boundary-emphasization that can be easily applied in most of the segmentation approaches in the medical field; it can strongly help the network to focus on the object contour. Furthermore, we demonstrated the use of a Dicoss-loss function to boost the performance of the model. Our method outperformed the state-of-the-art polyp segmentation methods on various datasets. The key advantage of the proposed method over existing methods is that it employs an ensemble of encoder-decoder networks trained to extract visual features from multi-scale images that are used in the second training step to re-learn the missing features

We presented an ensemble of DEDNs to extract multi-context information from multiscale training images. This ensemble of DEDNs was able to extract both global and local features, the combination could greatly enhance the segmentation performance of various approaches used in the medical imaging field. Furthermore, in the first training phase, we applied a novel data augmentation method that solved both the limited number of training dataset and over fitting problem. Then, we proposed a new training strategy called weight transfer so that the networks could look at the new dataset taken from the validation subtraction step. Finally, the Dicoss-loss function was used to effectively contribute to boosting the performance of the model in the training stage and later in the test period. The experimental results on challenging datasets demonstrated that this algorithm significantly outperformed state-of-the-art methods.

Experimental results demonstrated the superiority of the proposed method over stateof-the-art medical segmentation approaches. However, our method is still flawed; for example, the training phase takes a considerable amount of time. Thus, in the future development of our methods, we intend to improve the performance of incremental boosting convolution networks by adopting other novel effective methods such as using the advantages of a neural architecture search (NAS) [54] algorithm that can support the network, allowing it to focus on searching the repeatable cell structure, while handdesigning the outer network structure that controls the spatial resolution changes. This can enable the network to segment difficult objects like cells and tissues.

Finally, the proposed approach can be used to improve segmentation performance. To improve segmentation performance, we plan to extract training features from the most important convolution layers of our DCNNs and later check for missing and wrong patterns. This strategy is beneficial for building a strong segmentation model because combining our networks and the novel data augmentation with the Dicoss-Loss function allows the network to strongly focus on the medical object boundary.

REFERENCES

[1] L. Bi, J. Kim, E. Ahn, and D. Feng. (Mar. 12, 2017). "Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks." [Online]. Available: https://arxiv.org/abs/1703.04197

[2] H. Xu, R. Berendt, N. Jha, and M. Mandal, "Automatic measurement of melanoma depth of invasion in skin histopathological images," *Micron*, vol. 97, pp. 56–67, Jun. 2017.

[3] T. Heimann and H.-P. Meinzer, "Statistical shape models for 3D medical image segmentation: A review," *Med. Image Anal.*, vol. 13, no. 4, pp. 543–563, 2009.

[4] M. S. Iraji and A. Yavari, "Skin color segmentation in fuzzy YCBCR color space with the mamdani inference," *Amer. J. Sci. Res.*, vol. 7, pp. 131–137, Jul. 2011.

[5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[6] M. Akbari *et al.* (Feb. 1, 2018). "Polyp segmentation in colonoscopy images using fully convolutional network." [Online]. Available: https://arxiv.org/abs/1802.00368

[7] *American Cancer Society*, Cancer Facts & Figures 2018, 2018.

[8] 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*, 2015, pp. 234–241.

[9] H. Chen, X. Qi, L. Yu, and P. A. Heng, "DCAN: Deep contour-aware networks for accurate gland segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2487–2496.

[10] K. Wickstrøm, M. Kampffmeyer, and R. Jenssen, "Uncertainty Modeling and Interpretability in Convolutional Neural Networks for Polyp Segmentation," in *Proc. IEEE 28th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2018, pp. 1–6.

[11] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. (Feb. 7, 2018). "Encoder-decoder with atrous separable convolution for semantic image segmentation." [Online]. Available: https://arxiv.org/abs/1802.02611

[12] S. Davies and A. Moore, "Mix-nets: Factored mixtures of gaussians in Bayesian networks with mixed continuous and discrete variables," in *Proc. 16th Conf. Uncertainty Artif. Intell.*, 2000, pp. 168–175.

[13] G. Litjens *et al.*, "anchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[14] S. C. B. Lo, S. L. A. Lou, J.-S. Lin, M. T. Freedman, M. V. Chien, and S. K. Mun, "Artificial convolution neural network techniques and applications for lung nodule detection," *IEEE Trans. Med. Imag.*, vol. 14, no. 4, pp. 711–718, Dec. 1995.

[15] S. Ioffe and C. Szegedy. (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift." [Online]. Available: https://arxiv.org/abs/1502.03167

[16] V. Badrinarayanan, A. Kendall, and R. Cipolla. (Nov. 2, 2015) "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." [Online]. Available: https://arxiv.org/abs/1511.00561

[17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[18] A. Kendall, V. Badrinarayanan, and R. Cipolla. (Nov. 9, 2015). "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding." [Online]. Available: https://arxiv.org/abs/1511.02680

[19] L. Zhang, S. Dolwani, and X. Ye, "Automated polyp segmentation in colonoscopy frames using fully convolutional neural network and textons," in *Proc. Annu. Conf. Med. Image Understand. Anal.*, 2017, pp. 707–717.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[21] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[22] K. Simonyan and A. Zisserman. (Sep. 4, 2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[23] P. Brandao *et al.*, "Fully convolutional neural networks for polyp segmentation in colonoscopy," in *Proc. SPIE*, vol. 10134, Mar. 2017, Art. no. 101340F.

[24] Q. Li *et al.*, "Colorectal polyp segmentation using a fully convolutional neural network," in *Proc. 10th Int. Congr. Image Signal Process., BioMed. Eng. Inform. (CISP-BMEI)*, Oct. 2017, pp. 1–5.

[25] F. F. X. Vasconcelos, A. G. Medeiros, S. A. Peixoto, and P. P. R. Filho, "Automatic skin lesions segmentation based on a new morphological approach via geodesic active contour," *Cogn. Syst. Res.*, vol. 55, pp. 44–59, Jun. 2018.

[26] S. B. Salimi, S. Bozorgtabar, P. Schmid-Saugeon, H. K. Ekenel, M. S. Rad, and J.-P. Thiran, "DermoNet: Densely linked convolutional neural network for efficient skin lesion segmentation," Ecole Polytechnique Federale Lausanne, Lausanne, Switzerland, Tech. Rep., Nov. 2018.

[27] M. A. Selver, A. Kocaoğlu, G. K. Demir, H. Doğan, O. Dicle, and C. Güzeliş, "Patient oriented and robust automatic liver segmentation for pre-evaluation of liver transplantation," *Comput. Biol. Med.*, vol. 38, no. 7, pp. 765–784, 2008.

[28] A. H. Foruzan, R. A. Zoroofi, M. Hori, and Y. Sato, "Liver segmentation by intensity analysis and anatomical information in multi-slice CT images," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 4, no. 3, pp. 287–297, 2009.

[29] M. Freiman, O. Eliassaf, Y. Taieb, L. Joskowicz, and J. Sosna, "A bayesian approach for liver analysis: Algorithm and validation study," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, 2008, pp. 85–92.

[30] S.-J. Lim, Y.-Y. Jeong, and Y.-S. Ho, "Automatic liver segmentation for volume measurement in CT Images," *J. Vis. Commun. Image Represent.*, vol.17, no. 4, pp. 860–875, 2006.

[31] Z. Yan *et al.*, "Atlas-based liver segmentation and hepatic fat-fraction assessment for clinical trials," *Computerized Med. Imag. Graph.*, vol. 41, pp. 80–92, Apr. 2015.

[32] G. Chartrand, T. Cresson, R. Chav, A. Gotra, A. Tang, and J. A. De Guise, "Liver segmentation on CT and MR using Laplacian mesh optimization," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 9, pp. 2110–2121, Sep. 2017.

[33] F. Lu, F. Wu, P. Hu, Z. Peng, and D. Kong, "Automatic 3D liver location and segmentation via convolutional neural network and graph cut," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 12, no. 2, pp. 171–182, 2017.

[34] A. Mikołajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," in *Proc. Int. Interdiscipl. PhD Workshop (IIPhDW)*, May 2018, pp. 117–122.

[35] L. Perez and J. Wang. (Dec. 13, 2017). "The effectiveness of data augmentation in image classification using deep learning." [Online]. Available: https://arxiv.org/abs/1712.04621

[36] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell. (Sep. 28, 2016). "Understanding data augmentation for classification: When to warp?" [Online]. Available: https://arxiv.org/abs/1609.08764

[37] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez and F. Vilariño, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Med. Imag. Graph.*, vol. 43, pp. 99–111, Jul. 2015.

[38] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. S. 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.

[39] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1520–1528.

[40] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2009.

[41] D. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 2843–2851.

[42] L. Sifre and S. Mallat, "Rigid-motion scattering for image classification," M.S. thesis, Dept. Mathématiques Appliquées, Ecole Polytechnique, Citeseer, Palaiseau, France, 2014.

[43] T. Y. Lin *et al.*, "ar and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[44] J. Bernal *et al.*, "Comparative validation of polyp detection methods in video colonoscopy: Results from the MICCAI 2015 endoscopic vision challenge," *IEEE Trans. Med. Imag.*, vol. 36, no. 6, pp. 1231–1249, Jun. 2017.

[45] J. Bernal, J. Sánchez and F. Vilariño, "Towards automatic polyp detection with a polyp appearance model," *Pattern Recognit.*, vol. 45, no. 9, pp. 3166–3182, 2012.

[46] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 9, no. 2, pp. 283–293, 2014.

[47] P. Brandao *et al.*, "Fully convolutional neural networks for polyp segmentation in colonoscopy," in *Proc. SPIE*, vol. 10134, Mar. 2017, Art. no. 101340F.

[48] N. Tong, H. Lu, X. Ruan, and M. H. Yang, "Salient object detection via bootstrap learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1884–1892.

[49] 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.

[50] F. Yu and V. Koltun. (Nov. 23, 2015). "Multi-scale context aggregation by dilated convolutions." [Online]. Available: https://arxiv.org/abs/1511.07122

[51] 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.

[52] L. Bi, J. Kim, E. Ahn, D. Feng, and M. Fulham, "Automated skin lesion segmentation via image-wise supervised learning and multi-scale superpixel based cellular automata," in *Proc. IEEE 13th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2016, pp. 1059–1062.

[53] 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.

[54] C. Liu *et al.* (Jan. 10, 2019). "Auto-DeepLab: Hierarchical neural architecture search for semantic image segmentation." [Online]. Available: https://arxiv.org/abs/1901.02985

[55] T. Heimann *et al.*, "Comparison and evaluation of methods for liver segmentation from CT datasets," *IEEE Trans. Med. Imag.*, vol. 28, no. 8, pp. 1251–1265, Aug. 2009.

[56] D. Gutman *et al.* (May 4, 2016). "Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC)." [Online]. Available: https://arxiv.org/abs/1605.01397

[57] (2019). *CHAOS-Combined (CT-MR) Healthy Abdominal Organ Segmentation*. [Online]. Available: https://chaos.grand-challenge.org/Combined_Healthy_Abdominal_Organ_Segmentation/

NGOC-QUANG NGUYEN received the B.S. degree from the University of Engineering and Technology, Vietnam National University, Hanoi, in 2017. He is currently pursuing the master's degree with the Pattern Recognition and Machine Learning Laboratory, Gachon University, Seongnam, South Korea. His current research interests include face recognition, object segmentation, and medical image analysis.

SANG-WOONG LEE received the B.S. degree in electronics and computer engineering and the M.S. and Ph.D. degrees in computer science and engineering from Korea University, Seoul, South Korea, in 1996, 2001, and 2006, respectively. From 2006 to 2007, he was a Visiting Scholar with the Robotics Institute, Carnegie Mellon University. From 2007 to 2017, he was a Professor with the Department of Computer Engineering, Chosun University, Gwangju, South Korea. He is currently an Associate Professor with the Department of Software, Gachon University. His current research interests include face recognition, computational esthetics, machine learning, bioinformatics, and medical imaging analysis.