SE-U-Net: Contextual Segmentation by Loosely Coupled Deep Networks for Medical Imaging Industry

Lin-Yi Jiang[1(✉)], Cheng-Ju Kuo[1(✉)], O. Tang-Hsuan[1], Min-Hsiung Hung[2], and Chao-Chun Chen[1]

[1] IMIS/CSIE, NCKU, Tainan, Taiwan
{P96094189,P96084087,chaochun}@mail.ncku.edu.tw
[2] CSIE, PCCU, Taipei, Taiwan
hmx4@faculty.pccu.edu.tw

Abstract. We proposed a context segmentation method for medical images via two deep networks, aiming at providing segmentation contexts and achieving better segmentation quality. The context in this work means the object labels for segmentation. The key idea of our proposed scheme is to develop mechanisms to elegantly transform object detection labels into the segmentation network structure, so that two deep networks can collaboratively operate with loosely-coupled manner. For achieving this, the scalable data transformation mechanisms between two deep networks need to be invented, including representation of contexts obtained from the first deep network and context importion to the second one. The experimental results reveal that the proposed scheme indeed performs superior segmentation quality.

Keywords: Deep learning · Medical imaging segmentation · Computed tomography · Transpose convolution · Contextual computing

1 Introduction

Due to the advance of image processing and computer techniques, the medical imaging equipment become popular in hospitals for rapidly and precisely diagnosing various internal medicine symptoms. One frequently used medical image form is the computed tomography (CT), where the computer system controls the movement of the X-ray source to produce the image for further diagnosis.

This work was supported by Ministry of Science and Technology (MOST) of Taiwan under Grants MOST 109-2221-E-006-199, 108-2221-E-034-015-MY2, and 109-2218-E-006-007. This work was financially supported by the "Intelligent Manufacturing Research Center" (iMRC) in NCKU from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education in Taiwan.

N. T. Nguyen et al. (Eds.): ACIIDS 2021, LNAI 12672, pp. 678–691, 2021.
https://doi.org/10.1007/978-3-030-73280-6_54

The medical imaging diagnosis technologies have been widely studied in past decades [9], and provide symptom identifying assistance. Thus, medical imaging industries are eager to add the intelligent diagnosis functions to medical systems to increase doctors' diagnosis performance [2].

Traditional medical imaging diagnosis applied image processing techniques to identify symptoms based on their appearant characteristics. These methods worked well for some symptoms, and are hard to be applied to other symptoms, unless the algorithms are modified according to the new symptoms. The advance of artificial intelligence technologies greatly improves medical image processing capacities. These works use convolutional neural network (CNN), fully-connected neural network (FCN), artificial neural network (ANN) to accomplish missions of object detection, classification, and segmentation. Recently, some works [4,12] provide the medical image diagnosis with the deep learning technologies. Notably, the popular works in deep learning techniques [4] performs detection or segmentation for images of general situations, while the medical images have properties of low color differences between foreground and background, making computer systems hard to distinguish segmentation targets. None of them provide solutions for improving quality of detection to medical images. These existing methods might give inaccurate results so that doctors may be with low confidence to the diagnosis support systems and still need complete effort to diagnose symptoms by themselves. On the other hand literature [1,2,5–8] shows that contextual information processing - especially with neural networks - is very effective in increasing quality of classification of data vectors.

In this work, we proposed a contextual segmentation method for medical images via two deep networks, aiming at providing segmentation semantics and achieving better segmentation quality. The semantics in this paper means the object labels for contextual blocks. Previous works achieve each of these separately. In this work, we solve the two issues in a uniform framework, where the object detection deep network (ODDN) is used to extract semantics, and the developed Semantic U-Net (SE-U-Net) is used to perform medical image segmentation. The kernel idea of our proposed scheme is to develop two mechanisms to elegantly transform object detection labels with associated bounding boxes into the deep segmentation network structure. The first is the scalable data exchange interface between ODDNs and SE-U-Net need to be clearly defined, where a context matrix is invented to connect two networks. The second is that the U-Net structure needs to be modified to fit the additional semantic information. We conduct experiments to validate our proposed scheme on the medical images for identifying symptoms of the coronavirus disease 2019 (COVID-19). The experimental results from the COVID-19 dataset reveal that the proposed scheme indeed performs superior segmentation quality.

2 Backgrounds

2.1 Object Detection Deep Networks

The object detection deep networks (ODDNs) have been studied in the last decade, and many famous methods, e.g., YOLOv3 [10], are proposed and widely recognized. Particularly, they are validated in crucial simulation tests and then open-sourced, so that the area of ODDNs rapidly grows up and become one of most important applications in deep network technologies. The main underlying techniques of ODDNs is the CNNs and the FCNs, as the formal ones extract object features automatically and the later ones map the generated features to specified object classes. Many ODDNs are further developed for fitting different application domains. Nowadays, the deep networks are well trained with data sets via graphic processing unit (GPU) for efficiency.

2.2 Segmentation Deep Networks for Medical Images

Segmentation for medical images has a requirement: the processed images need to preserve most properties of original ones. Thus, the deep networks used in medical imaging industries need to bring portions of the original images to outputs for ensuring high similarity. The popular segmentation deep network for medical images is U-Net [11]. The key design of the U-Net is that extracted features are directly copied and moved to the generation stage, so that many principle properties are incrementally added to the output to preserve the essential elements of original images. Many variants of the U-Net structure are also proposed in recent year, such as ResUNet [3], and they all reserved the copy-and-move components.

3 Proposed Contextual Segmentation Scheme

3.1 Overall Architecture

Figure 1 shows the architecture of our proposed contextual segmentation scheme, which includes two deep networks and they are loosely coupled in the architecture with merely data exchange. The first deep network is used to perform object detection for providing contexts, and the second one is used to perform segmentation over medical images. The advantage of the architecture is two-folded: one is easy to watch the effect of each deep network, the other is avoid complicated model creation procedure. The kernel idea of the scheme includes the loose-coupled of two deep networks and the mechanisms of integrating them. The technical details of the whole architecture, model creation, and network integration will be presented in the following subsections.

Assume a medical image, denoted $\mathbf{X}^{[i]}$, retrieved from the medical image database via the DICOM (Digital Imaging and Communications in Medicine) protocol has $w \times h$ pixels with ch channels (shortly, it denotes $w \times h \times ch$). For example, the medical images used in this study are of $416 \times 416 \times 3$. The

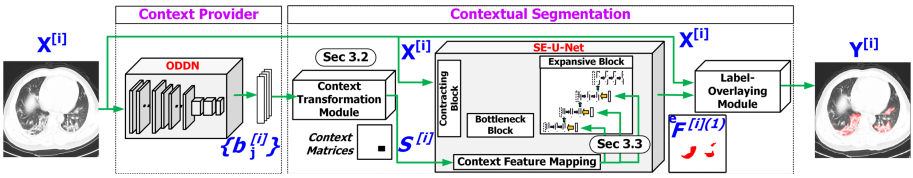


Fig. 1. The proposed contextual segmentation architecture for medical image processing.

output image of the SE-U-Net denotes $\mathbf{Y}^{[i]}$, whose dimension is the same to the input image, i.e., $dim(\mathbf{X}^{[i]}) = dim(\mathbf{Y}^{[i]})$, implying the image size remains after our developed deep network proceeds it. The first network shown in the left-hand side of the figure is an object-detection deep network (ODDN), which generates object detection results in a bounding-box set, denoted BBS, with the format $\{\hat{\mathbf{b}}_j^{[i]} = (\hat{x}_j^{[i]}, \hat{y}_j^{[i]}, \hat{w}_j^{[i]}, \hat{h}_j^{[i]}, \hat{c}_j^{[i]}) | j = 1, \cdots, \hat{k}^{[i]}\}$, where $(\hat{x}_j^{[i]}, \hat{y}_j^{[i]})$ indicates the center of the bounding box (shortly, BBox), $\hat{w}_j^{[i]}$ and $\hat{h}_j^{[i]}$ indicate the width and height, $\hat{c}_j^{[i]}$ is the classification of the BBox, and $\hat{k}^{[i]}$ is the number of BBoxes in $\mathbf{X}^{[i]}$. Since the output format is clearly defined in the architecture, developing deep-network components is flexible and many existing works, such as YOLO, Fast-RCNN, SSD, etc. are considerable candidates. For example, YOLO is used in the experimental study. Without loss of generality, we present the model operation of the ODDN that identifying $\hat{k}^{[i]}$ objects (represented in the form of BBS) for the input $\mathbf{X}^{[i]}$ as

$$M_{ODDN}(\mathbf{X}^{[i]}|\theta_{OD}) = \{\hat{\mathbf{b}}_j^{[i]} | j = 1, \ldots, \hat{k}^{[i]}\} \tag{1}$$

where θ_{OD} is the hyperparameters of the ODDN. For finding out θ_{OD}, the loss function used in the optimizer of the ODDN in this work can be expressed as

$$\begin{aligned} L_{ODDN}(\mathbf{b}^{[i]}, \hat{\mathbf{b}}^{[i]}) &= \sum_{j}^{k}\sum_{\hat{j}}^{\hat{k}} \mathbb{1}^{obj}(\mathbf{b}_j^{[i]}, \hat{\mathbf{b}}_{\hat{j}}^{[i]}) \times \left(\left(x^{[i]_j} - \hat{x}_{\hat{j}}^{[i]}\right)^2 + \left(y_j^{[i]} - \hat{y}_{\hat{j}}^{[i]}\right)^2\right) \\ &+ \sum_{j}^{k}\sum_{\hat{j}}^{\hat{k}} \mathbb{1}^{obj}(\mathbf{b}_j^{[i]}, \hat{\mathbf{b}}_{\hat{j}}^{[i]}) \times \left(\left(\sqrt{w_j^{[i]}} - \sqrt{\hat{w}_{\hat{j}}^{[i]}}\right)^2 + \left(\sqrt{h_j^{[i]}} - \sqrt{\hat{h}_{\hat{j}}^{[i]}}\right)^2\right) \\ &+ \sum_{j}^{k}\sum_{\hat{j}}^{\hat{k}} \mathbb{1}^{obj}(\mathbf{b}_j^{[i]}, \hat{\mathbf{b}}_{\hat{j}}^{[i]}) \times (c_j^{[i]} - \hat{c}_{\hat{j}}^{[i]})^2 + \sum_{j}^{k} \mathbb{1}^{noobj}(\hat{\mathbf{b}}^{[i]}, \mathbf{b}_j^{[i]}) + \sum_{\hat{j}}^{\hat{k}} \mathbb{1}^{noobj}(\mathbf{b}^{[i]}, \hat{\mathbf{b}}_{\hat{j}}^{[i]}) \end{aligned} \tag{2}$$

where $\mathbb{1}^{obj}(\mathbf{b}_j^{[i]}, \hat{\mathbf{b}}_{\hat{j}}^{[i]})$ returns 1 if BBoxes of two parameters predict the same object, 0 otherwise; $\mathbb{1}^{noobj}(\mathbf{b}_1, \mathbf{b}_2)$ returns 1 if the object in $\mathbf{b}_2$ is not inside $\mathbf{b}_1$, 0 otherwise.

The second deep segmentation network, shown in the right-hand side of the figure, is called Semantic U-Net (denoted SE-U-Net), which is a variant of the U-Net structure [11]. Figure 2 illustrates the core SE-U-Net structure, consisting of the constructing block for constructing features of medical images, the

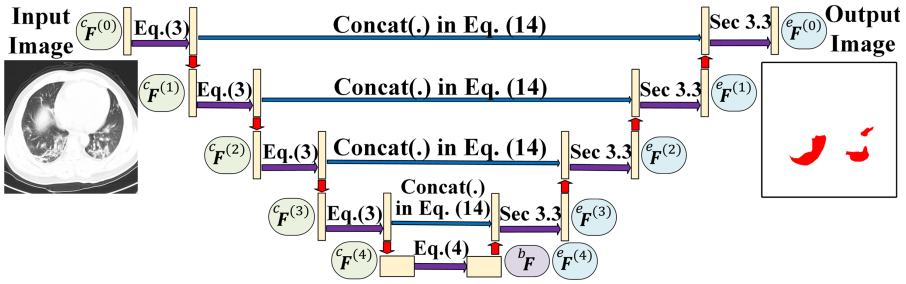


Fig. 2. Illustration of the Core SE-U-Net structure of four levels ($v = 4$) for extracting context from CT images (the feature ${}^{e}\mathbf{F}^{(0)}$ contains generated symptom regions.)

bottlenecking block for transforming features between different feature spaces, and the expensive block for expansive features to medical images. The expensive block will also accommodate the context obtain from the ODDN, which will be formally presented in Sect. 3.3. The key design of the SE-U-Net is that each layer of the constructing block and the expansive block are of the same size and the feature of the former block are directly copy and move to the later block, so that features in the construction block are incrementally added to the output to preserve the essential elements of original images.

The constructing block has v levels and each level performs two convolution operations for feature extraction with one maximum pooling operation for dimension reduction to extracted features. Thus, a feature ${}^{c}\mathbf{F}^{[i](i)}$ of $\mathbf{X}^{[i]}$ generated from level i-th of the constructing block can be expressed as

$$ {}^{c}\mathbf{F}^{[i](i)} = MaxPool(Conv(Conv({}^{c}\mathbf{F}^{[i](i-1)}))), \text{ where } i = 1, \ldots, v. \tag{3} $$

where $MaxPool(.)$ and $Conv(.)$ are the maximum pooling operation and the convolution operation, respectively, and ${}^{c}\mathbf{F}^{[i](0)}$ is the inputted image $\mathbf{X}_i$. The bottleneck block performs two convolution operations, and a feature ${}^{b}\mathbf{F}^{[i]}$ generated in the constructing block for $\mathbf{X}^{[i]}$ can be expressed as

$$ {}^{b}\mathbf{F}^{[i]} = Conv(Conv({}^{c}\mathbf{F}^{[i](v)})) \tag{4} $$

Note that the constructing block and the bottleneck block play as an encoder to extract key features by downsampling procedures via a couple of convolution and max pooling operations. The expansive block has the same number of levels to the constructing block (i.e., v levels) and each level performs two convolution operations and one transposed convolution operation for expanding features obtained from the previous level. A feature ${}^{e}\mathbf{F}^{(i)}$ generated from level i-th of the constructing block is determined by using the feature ${}^{e}\mathbf{F}^{(i+1)}$ and the context provided by the ODDN. Comparing to the previous two blocks, the expansion block plays as a decoder to generate an image from a feature by upsampling procedures. The technical details of ${}^{e}\mathbf{F}^{(i)}$ will be discussed in Sect. 3.3. The specification of the SE-U-Net used in this work are shown in Table 1. The loss

function L_{SEU} is defined based on the idea of Dice coefficient for measuring similarity between pixel sets from a CT image and the labeling image, which can represented as the following equation:

$$L_{SEU} = \sum_{k=1}^{N} 1 - \frac{2 \times (\Sigma_{i=1}^{w} \Sigma_{j=1}^{h} \mathbf{Y}_{i,j}^{(k)} \times \hat{\mathbf{Y}}_{i,j}^{SEU(k)})}{(\Sigma_{i=1}^{w} \Sigma_{j=1}^{h} \mathbf{Y}_{i,j}^{(k)} \times \mathbf{Y}_{i,j}^{(k)}) + (\Sigma_{i=1}^{w} \Sigma_{j=1}^{h} \hat{\mathbf{Y}}_{i,j}^{SEU(k)} \times \hat{\mathbf{Y}}_{i,j}^{SEU(k)})} \tag{5}$$

After obtaining the output ${}^{e}\mathbf{F}^{[i](1)}$ of $\mathbf{X}^{[i]}$ from the SE-U-Net, the label-overlaying module would overlay the refined context on the original medical images, and it can be represented in the following equation:

$$\mathbf{Y}_{j,k,l}^{[i]} = \begin{cases} \mathbf{X}_{j,k,l}^{[i]} + {}^{e}\mathbf{F}_{j,k,l}^{[i](1)}, c \in \text{channel red} \\ \mathbf{X}_{j,k,l}^{[i]}, \text{otherwise} \end{cases} \tag{6}$$

For ease of study, all contexts are highlighted in the red channel, and the overlay function can be modified according to different application needs.

Table 1. The layer specification of SE-U-Net studied in this paper, where the input size is $416 \times 416 \times 3$ and the output size is $512 \times 512 \times 1$.

Contacting block	Kernel size	Depth	Strides	Padding
Convolution	3×3	64	[1, 1]	SAME
Convolution	3×3	64	[1, 1]	SAME
Max Pooling	2×2		[2, 2]	VALID
Convolution	3×3	128	[1, 1]	SAME
Convolution	3×3	128	[1, 1]	SAME
Max Pooling	2×2		[2, 2]	VALID
Convolution	3×3	256	[1, 1]	SAME
Convolution	3×3	256	[1, 1]	SAME
Max Pooling	2×2		[2, 2]	VALID
Convolution	3×3	512	[1, 1]	SAME
Convolution	3×3	512	[1, 1]	SAME
Max Pooling	2×2		[2, 2]	VALID
output size	$26 \times 26 \times 512$			
Convolution	3×3	1024	[1, 1]	SAME
Convolution	3×3	1024	[1, 1]	SAME
Convolution	3×3	1024	[1, 1]	SAME
Convolution	3×3	1024	[1, 1]	SAME
output size	$26 \times 26 \times 1024$			

Expansive block	Kernel size	Depth	Strides	Padding
Transpose convolution	2×2	512	[2, 2]	VALID
Convolution	3×3	512	[1, 1]	SAME
Convolution	3×3	512	[1, 1]	SAME
Transpose convolution	2×2	256	[2, 2]	VALID
Convolution	3×3	256	[1, 1]	SAME
Convolution	3×3	256	[1, 1]	SAME
Transpose convolution	2×2	128	[2, 2]	VALID
Convolution	3×3	128	[1, 1]	SAME
Convolution	3×3	128	[1, 1]	SAME
Transpose convolution	2×2	64	[2, 2]	VALID
Convolution	3×3	64	[1, 1]	SAME
Convolution	3×3	64	[1, 1]	SAME
Convolution	3×3	1	[1, 1]	SAME

Notice that two deep networks in our architecture are loosely coupled with merely data delivery via the context transformation module. Such design has two advantages. The first is that the two deep networks can be trained individually with less training data, and spend less model creation time compared to a single concated deep network. The second is that context extraction and segmentation are both performed with satisfied high quality in the separated deep networks. In case the single deep network solution is adopted, developers

do not know quality of context extraction and segmentation separately as they are all mixtured encoded in the hyperparameters of the single deep network.

Operational Workflows of the Proposed Scheme:
Two operational workflows, model creation and medical image diagnosis, for the proposed context segmentation scheme are presented below.

Model Creation Workflow
Assume a set of medical images $\mathbf{X}^{[i]}, i = 1, \ldots, N$ are given and associated labels of $\mathbf{X}^{[i]}$ are collected from domain experts in the form: $\{\mathbf{Y}^{[i]}, \mathbf{c}^{[i]}\}$, where $\mathbf{Y}^{[i]}$ is the associated segmented image, $\mathbf{c}^{[i]}$ is the associated class label set.

Step 1. Create an optimal ODDN model (θ^*_{OD}).
Assume $BBox(\mathbf{Y}^{[i]}, \mathbf{c}^{[i]}) = \{\mathbf{b}^{[i](j)} = (x_b^{[i](j)}, y_b^{[i](j)}, w_b^{[i](j)}, h_b^{[i](j)}, c_b^{[i](j)}) | j = 1, \ldots, k^{[i]}\}$ is an extraction function that returns a set of minimal bounding boxes $\mathbf{b}^{[i](j)}$ covering irregular labeled segmentation shapes in $\mathbf{Y}^{[i]}$, including the center point, width, and height, where $k^{(i)}$ is the number of collected labels in $\mathbf{X}^{[i]}$. Let $M_{OD}(\mathbf{X}^{[i]}|\theta_{OD}) = \{\hat{\mathbf{b}}^{[i](j)} | j = 1, \ldots, \hat{k}^{[i]}\}$ be the output of the ODDN with hyperparameter θ_{OD} to the medical image $\mathbf{X}^{[i]}$. Given labeled dataset $D_{OD} = \{\mathbf{X}^{[i]}, BBox(\mathbf{Y}^{[i]}, \mathbf{c}^{[i]}) | i = 1, \ldots\}$, in this step, We use the training/testing processes in deep learning to find out the optimal hyperparameters for the ODDN network (i.e., θ^*_{OD}), so that the loss function L_{OD} is minimum. That is,

$$\theta^*_{OD} = \arg\min_{\theta; D_{OD}} L_{OD}(M_{OD}(\mathbf{X}^{[i]}|\theta), BBox(\mathbf{Y}^{[i]}, \mathbf{c}^{[i]}) = \{\mathbf{b}^{[i](j)}\}) \quad (7)$$

Note that the definition of L_{OD} can refer to Eq. (2).

Step 2. Create an optimal SE-U-Net model (θ^*_{SEU}).
Let $M_{SEU}(\mathbf{X}^{[i]}, CM(\mathbf{Y}^{[i]})|\theta_{SEU}) = \hat{\mathbf{Y}}^{SEU[i]}$ be the output of the SE-U-Net with hyperparameter θ_{SEU} to the medical image $\mathbf{X}^{[i]}$, where $CM(\mathbf{Y}^{[i]})$ return the context matrices of $\mathbf{Y}^{[i]}$. Given $D_{SEU} = \{\mathbf{X}^{[i]}, BBox(\mathbf{Y}^{[i]}), \mathbf{Y}^{[i]} | i = 1, \ldots\}$, in this step, we use the training/testing processes to find out the optimal hyperparameters for the SE-U-Net (i.e., θ^*_{SEU}), so that the loss function L_{SEU} of the SE-U-Net, defined in Eq. (5) is minimum. That is,

$$\theta^*_{SEU} = \arg\min_{\theta} L_{SEU}(M_{SEU}(\mathbf{X}^{[i]}, BBox(\mathbf{Y}^{[i]})|\theta), \mathbf{Y}^{[i]}) \quad (8)$$

The pair $(\theta^*_{OD}, \theta^*_{SEU})$ are the models used in our proposed scheme.

Medical Image Diagnosis Workflow
The dashed arrows in Fig. 1 indicate the medical image diagnosis workflow. Due to length limit, we ignore the detailed steps here.

3.2 Context Matrix Transformation

In this subsection, we present a key mechanism, Context Matrix Transformation (CMT), used in the context transformation module for the scalable data

exchange interface between ODDNs and SE-U-Net. The context matrix is scalable as it is adaptive to the medical image size and also accommodate multiple labels in an image. Given a bounding box set $\{\hat{\mathbf{b}}_j^{[i]}\}$ obtained from an ODDN for a medical image, the CMT generates a context matrix, denoted $\mathbf{S}^{[i]}$, whose dimension is the same to the original image $\mathbf{X}^{[i]}$.

Figure 3 illustrates the context matrix transformation, which contains two steps: the matrix-rendering step and the label filling step. The matrix-rendering step is to map the bounding boxes to a matrix with size $w \times h$. Each bounding box $\hat{\mathbf{b}}_j^{[i]}$ of $\mathbf{X}^{[i]}$ has attributes $(\hat{x}_j^{[i]}, \hat{y}_j^{[i]}, \hat{w}_j^{[i]}, \hat{h}_j^{[i]})$, which indicates the corresponding sub-matrix, as shadowed ones in the figure. Each sub-matrix can be indicated by two coordinates: the top-left (tl) point and the bottom-right (br) point, and is represented as

$$tl_j^{[i]} = (tl_j^{[i]}.x, tl_j^{[i]}.y) = ((\hat{x}_j^{[i]} - \hat{w}_j^{[i]}/2) \times w, (\hat{y}_j^{[i]} - \hat{h}_j^{[i]}/2) \times h) \tag{9}$$

$$br_j^{[i]} = (br_j^{[i]}.x, br_j^{[i]}.y) = ((\hat{x}_j^{[i]} + \hat{w}_j^{[i]}/2) \times w, (\hat{y}_j^{[i]} + \hat{h}_j^{[i]}/2)) \times h) \tag{10}$$

Note that for medical imaging applications, each element only belong to one lable at most, that is, it is either marked or unmarked in the context matrix. The tl-br representation is easy to ensure such property, compared to the bounding-box representation. The label filling step is to filling elements in the sub-matrix $(\hat{x}_j^{[i]}, \hat{y}_j^{[i]}, \hat{w}_j^{[i]}, \hat{h}_j^{[i]})$ with the label value $\hat{c}_j^{[i]}$. Other matrix elements are filled with zero, meaning they are unlabeled. After the two steps, all labels of a medical image $\mathbf{X}^{[i]}$ are encoded into a context matrix $\mathbf{S}^{[i]}$ as follows

$$\mathbf{S}_{r,t}^{[i]} = \begin{cases} \hat{c}_j^{[i]}, & \text{if } r \in [tl_j^{[i]}.x, br_j^{[i]}.x] \text{ and } t \in [tl_j^{[i]}.y, br_j^{[i]}.y], \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

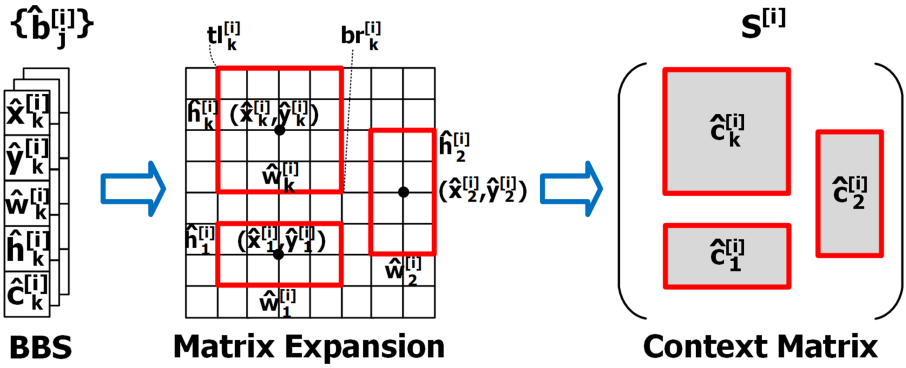


Fig. 3. Illustration of transforming a bounding-box set to a context matrix.

3.3 Computing Expansive Features with Context Matrice

We present another key mechanism in SE-U-Net, Context Feature Mapping (CFM), which is used to refine expansive features in this subsection. In other words, the CFM mechanism is to fuse a context matrix into the core SE-U-Net structure, The inputs of the CFM mechanism include the context matrix $\mathbf{S}^{[i]}$ and the generated features ${}^{c}\mathbf{F}^{[i](k)}$ of the construction block (referring to Fig. 2.)

Let $\mathbf{G}^{[i](k)}$ be the augmented context matrix of the expensive block in the k-th stage and ${}^{g}\mathbf{F}^{[i](k)}$ be the augmented context feature considering $\mathbf{G}^{[i](k)}$. The procedure include two steps. The first step is to produce the augmented context matrices $\mathbf{G}^{[i](k)}$ for $\mathbf{X}^{[i]}$. The purpose of producing $\mathbf{G}^{[i](k)}$ is to scale the context matrix $\mathbf{S}^{[i]}$ to fit the different feature sizes in the k levels of the expensive block. The matrix element $\mathbf{G}^{[i](k)}_{x,y}$ can be computed as the following equation.

$$\mathbf{G}^{[i](k)}_{x,y} = \begin{cases} \hat{c}^{[i]}_{j}, & \text{if } (x \times \frac{w}{w^{(k)}}) \in [tl^{[i]}_{j}.x, br^{[i]}_{j}.x] \text{ and } (y \times \frac{h}{h^{(k)}}) \in [tl^{[i]}_{j}.y, br^{[i]}_{j}.y], \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

The second step is to fuse the augmented context matrix to the expansive block of the core SE-U-Net structure. For level k, the feature ${}^{e}\mathbf{F}^{[i](k+1)}$ obtained in the $(k+1)$-th level is upsampling (via the transpose convolution operation), which is next fused with and the augmented context matrix $\mathbf{G}^{[i](k)}$. The fusion result is then concatenated with the feature ${}^{c}\mathbf{F}^{[i](k)}$, previously defined in Eq. 3, which are then performed convolutions for smoothing the extracted symptoms. The technical details can be represented in the following two formulae.

$${}^{g}\mathbf{F}^{[i](k)} = TransConv({}^{e}\mathbf{F}^{[i](k+1)}) \otimes \mathbf{G}^{[i](k)}, \ k = v-1, \ldots, 0. \tag{13}$$

$${}^{e}\mathbf{F}^{[i](k)} = Conv(Conv(Concat({}^{g}\mathbf{F}^{[i](k)}, {}^{c}\mathbf{F}^{[i](k)}))), \ k = v-1, \ldots, 0. \tag{14}$$

where $\otimes$ is the element-wise production, ${}^{e}\mathbf{F}^{[i](v)} = {}^{b}\mathbf{F}^{[i]}$ for computing ${}^{g}\mathbf{F}^{[i](k)}$ of level $v-1$, and $Concat(.)$ is to concatenate two features. Note that ${}^{c}\mathbf{F}^{[i](k)}$ from the construction block is concatenated to the intermediate feature, which implements the copy-and-move property inspired by the widely used U-Net.

3.4 Discussions: What Does SE-U-Net Do?

The proposed SE-U-Net not only embeds extracted contexts to the image, but also refines the segmentation quality. Thus, the proposed method is worth promotion, compared to existing ones. Figure 4 illustrates conceptual working effects of the proposed SE-U-Net, and explains why SE-U-Net can refine the symptom label with context information. Traditional segmentation methods (e.g., U-Net) perform segment the possible symptom region with features extracted in construction levels, and due to the privacy issue concerned in most hospitals, the number of medical images used for creating models is limited. Thus, U-Net obtains less accurate symptom region, as shown in the middle of the figure.

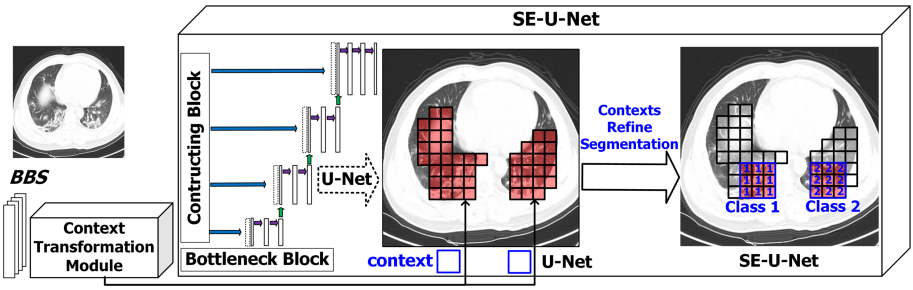


Fig. 4. Illustration of working effects of the proposed SE-U-Net.

For tackling the less quality of a segmentation model, SE-U-Net uses context obtained from another deep network (i.e., ODDN) to further refines the symptom region in the expansive block by means of feature fusion operations (i.e., Eqs. (13) and (14).) In this way, SE-U-Net improves quality of possible symptom regions, as shown in the right of the figure. The conceptual illustration also sustains design of the proposed loose-coupled architecture.

4 Case Study

4.1 System Deployment and Experimental Settings

We deploy the proposed SE-U-Net system on the personal computer with two graphic processing cards of Nvidia GTX 1080 Ti. The personal computer is with Intel i9-7920x processor and 128 GB memory. The system prototype is developed with programming tools mostly used in the deep-learning research, including Python 3.5, TensorFlow 1.6.0, CUDA v9.0.176, and cuDNN 70005. The COVID-CT-Dataset obtained from https://github.com/UCSD-AI4H/COVID-CT has 349 CT images of size $416 \times 416 \times 3$ containing clinical findings of coronavirus disease 2019 (COVID-19) from 216 patients.

4.2 Expr. 1: Visualization Effects of Context Segmentation

Figure 5 shows the visualization effects of the SE-U-Net, where the red is the SE-U-Net, the blue is the human labels, and the pink is the intersection of both the SE-U-Net and the human experts. Three processed CT images are selected from both training and testing datasets for visual validation. We also show the associated object detection results, i.e., contexts provided by the ODDN, for understanding the where the SE-U-Net focuses and how it performs segmentation. From results of the training dataset, we can see most symptoms are marked (i.e., pink color), which indicate the created SE-U-Net model is well trained and performed sufficiently acceptable.

In results of the testing dataset with the created model, the first two cases are randomly selected and the last one is a relative worse instance. The first two

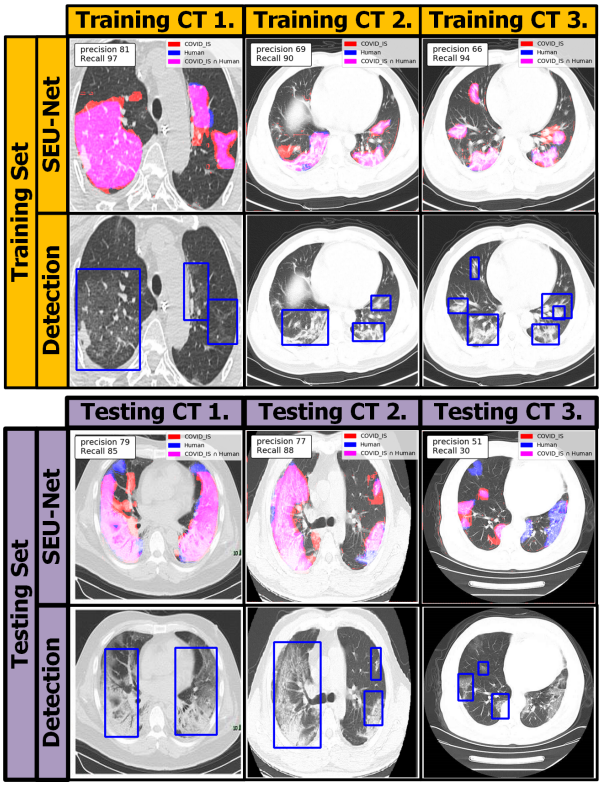


Fig. 5. Visualization effects of the SE-U-Net, where blue boxes are contexts, pink regions are true positives, and blue regions are false negatives. (Color figure online)

shows most marked parts are also recognized by the human experts (i.e., the pink shapes), indicating that our SE-U-Net quite successfully inspects the unseen CT images. In the third one, symptoms in the right lung are missed so that the inspection performance decreases. Note that the ODDN does not provide any context to the SE-U-Net in this instance. In this situation, the segmentation job is performed merely depending on the SE-U-Net model, and its performance in the case is similar to that of the traditional U-Net.

4.3 Expr. 2: Comparisons of Segmentation Methods

This experiment studies the effects of our SE-U-Net and existing U-Net and the results are shown in Fig. 6, where the same CT images adopted in the first experiment are used for fair comparisons. The two models are well trained as possible. From the results of the training dataset, we can see that the SE-U-Net and the U-Net mark most human labeling regions (i.e., high recall value), meaning that both models have sufficient capacity to inspect symptoms. By further analysis based on the precision value, we can see that the SE-U-Net

has lower false positives than the U-Net, indicating that the SE-U-Net provides more accurate information to doctors. Results of the testing dataset have similar phenomena, which supports the above segmentation effects of the two methods.

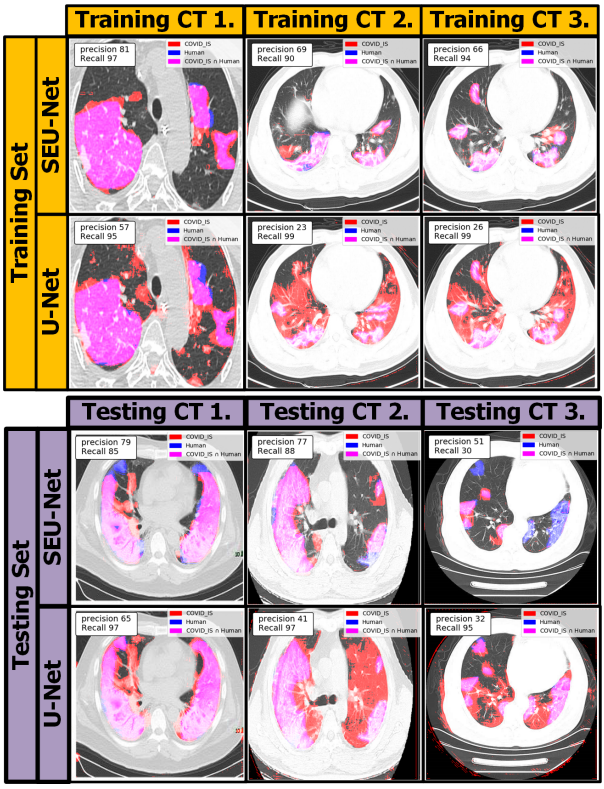


Fig. 6. Visualization comparisons of the SE-U-Net and the U-Net, where pink regions are true positives, red regions are false positives, and blue regions are false negatives. (Color figure online)

4.4 Expr. 3: Quantitative Comparisons

We present quality of the created model used in experiments and comparisons of deep networks in this experiment. Figure 7 shows the loss-function value in various epoches during creating the SE-U-Net model. We can see that the loss-function values of the training dataset are close to zero after 2000 epoches, meaning that the created model has capacity to identify symptoms. Values of the validation dataset are also quite low, verifying that the model is qualified to perform context segmentation tasks.

Figures 8 shows comparisons of our SE-U-Net and the existing U-Net via the receiver operating characteristic (ROC) curve and the precision-recall distribu-

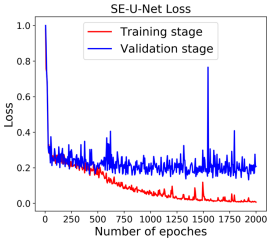
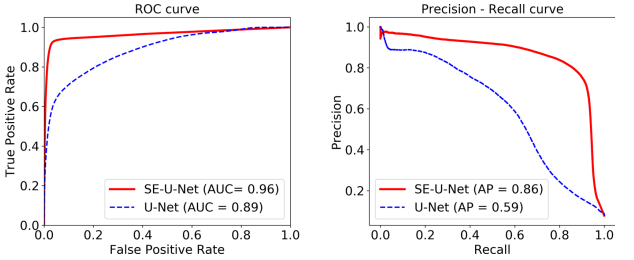


Fig. 7. Loss values during creating the SE-U-Net model.



(a) ROC curves (b) Precision-Recall curves

Fig. 8. Performance comparisons of SE-U-Net and U-Net.

tion, which both are widely used to measure performance of deep learning models. From both plots, we clearly see that the SE-U-Net performs more superior than the U-Net. The experimental results confirm that the fore ODDN indeed assists the SE-U-Net to concentrate on the localized region, so that the SE-U-Net focuses on generating exquisite contextual segmentation outcomes shown previously. The results also verify that our developed mechanisms highly effectively accomplish the purpose of loosely coupling two deep networks.

5 Conclusions and Future Work

In this paper, we proposed SE-U-Net to provide contexts and high-quality segmentation to CT images for assisting doctors to diagnose symptoms. Traditional segmentation only consider symptom identification with a single deep networks, which could be less accurate. Our proposed SE-U-Net can employ object detection deep networks for acquiring contexts with bounding boxes, and then loosely combine those contexts to the attentioned U-Net for further refining the segmentation quality. We also give analysis to explain reasons that the SE-U-Net can refine the segmentation quality. We developed the SE-U-Net prototype and conducted experiments to test its performance. The experimental results revealed that the proposed SE-U-Net indeed performs superior than existing methods in metrics concerned by hospital experts. Our future work will extend the SE-U-Net to diagnose other organs, such as the pancreas, which may be less apparent

in medical images. Certain mechanisms sensitive to such unapparent symptoms need to be additionally invented for the SE-U-Net.

References

1. Bovolo, F., Bruzzone, L.: A context-sensitive technique based on support vector machines for image classification. In: Pal, S.K., Bandyopadhyay, S., Biswas, S. (eds.) PReMI 2005. LNCS, vol. 3776, pp. 260–265. Springer, Heidelberg (2005). https://doi.org/10.1007/11590316_36
2. Cai, G., et al.: One stage lesion detection based on 3D context convolutional neural networks. Comput. Electr. Eng. **79**, 106449 (2019)
3. Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C.: ResUNet-a: a deep learning framework for semantic segmentation of remotely sensed data (2019)
4. Ghamdi, M.A., Abdel-Mottaleb, M., Collado-Mesa, F.: DU-Net: convolutional network for the detection of arterial calcifications in mammograms. IEEE Trans. Med. Imaging **39**, 3240–3249 (2020)
5. Ghimire, B., Rogan, J., Miller, J.: Contextual land-cover classification: incorporating spatial dependence in land-cover classification models using random forests and the Getis statistic. Remote Sens. Lett. **1**, 45–54 (2010)
6. Huk, M.: Non-uniform initialization of inputs groupings in contextual neural networks. In: Nguyen, N.T., Gaol, F.L., Hong, T.-P., Trawiński, B. (eds.) ACIIDS 2019. LNCS (LNAI), vol. 11432, pp. 420–428. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-14802-7_36
7. Huk, M., Mizera-Pietraszko, J.: Context-related data processing in artificial neural networks for higher reliability of telerehabilitation systems. In: 2015 17th International Conference on e-health Networking, Application & Services (HealthCom), pp. 217–221 (2015)
8. Kamara, A.F., Chen, E., Liu, Q., Pan, Z.: Combining contextual neural networks for time series classification. Neurocomputing **384**, 57–66 (2020)
9. Litjens, G., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
10. Redmon, J., Divvala, S., Girshick, R.B., Farhadi, A.: You only look once: unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788 (2016)
11. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: MICCAI (2015)
12. Wang, X., et al.: A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT. IEEE Trans. Med. Imaging **38**, 2615–2625 (2020)