IEEE Journal of Translational Engineering in Health and Medicine

Received 6 December 2022; revised 14 March 2023, 28 April 2023, and 5 June 2023; accepted 8 June 2023.
Date of publication 15 June 2023; date of current version 26 June 2023.

Digital Object Identifier 10.1109/JTEHM.2023.3286423

Multiple Field-of-View Based Attention Driven Network for Weakly Supervised Common Bile Duct Stone Detection

YA-HAN CHANG[1], MENG-YING LIN[2], MING-TSUNG HSIEH[2], MING-CHING OU[3], CHUN-RONG HUANG[1,4], (Senior Member, IEEE), AND BOR-SHYANG SHEU[2]

[1]Department of Computer Science and Engineering, National Chung Hsing University, Taichung 402202, Taiwan
[2]Department of Internal Medicine, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan 701401, Taiwan
[3]Department of Medical Image, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan 701401, Taiwan
[4]Cross College Elite Program, and Academy of Innovative Semiconductor and Sustainable Manufacturing, National Cheng Kung University, Tainan 701401, Taiwan

(Ya-Han Chang and Meng-Ying Lin are co-first authors.) CORRESPONDING AUTHOR: C.-R. HUANG (crhuang@gs.ncku.edu.tw)

This work was supported in part by the National Science and Technology Council of Taiwan under Grant NSTC 111-2634-F-006-012, Grant NSTC 111-2628-E-006-011-MY3, and Grant NSTC 112-2327-B-006-008.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board, National Cheng Kung University Hospital, under Application No. B-ER-111-186.

ABSTRACT Objective: Common bile duct (CBD) stones caused diseases are life-threatening. Because CBD stones locate in the distal part of the CBD and have relatively small sizes, detecting CBD stones from CT scans is a challenging issue in the medical domain. Methods and procedures: We propose a deep learning based weakly-supervised method called multiple field-of-view based attention driven network (MFADNet) to detect CBD stones from CT scans based on image-level labels. Three dominant modules including a multiple field-of-view encoder, an attention driven decoder and a classification network are collaborated in the network. The encoder learns the feature of multi-scale contextual information while the decoder with the classification network is applied to locate the CBD stones based on spatial-channel attentions. To drive the learning of the whole network in a weakly-supervised and end-to-end trainable manner, four losses including the foreground loss, background loss, consistency loss and classification loss are proposed. Results: Compared with state-of-the-art weakly-supervised methods in the experiments, the proposed method can accurately classify and locate CBD stones based on the quantitative and qualitative results. Conclusion: We propose a novel multiple field-of-view based attention driven network for a new medical application of CBD stone detection from CT scans while only image-levels are required to reduce the burdens of labeling and help physicians automatically diagnose CBD stones. The source code is available at https://github.com/nchucvml/MFADNet after acceptance. Clinical impact: Our deep learning method can help physicians localize relatively small CBD stones for effectively diagnosing CBD stone caused diseases.

INDEX TERMS Common bile duct (CBD) stone detection, choledocholithiasis, weakly-supervised learning, deep learning, object detection.

I. INTRODUCTION

The presence of gallstones in the common bile duct refers as the common bile duct (CBD) stone which is also known as choledocholithiasis. Most cases of choledocholithiasis result from gallstones stuck in the common bile duct [1]. As shown in [2], up to 20% of gallbladder stone cases are associated with CBD stones. CBD stones caused acute suppurative cholangitis and acute biliary pancreatitis [3] are life-threatening. These diseases should be diagnosed and treated immediately even in asymptomatic ones [4].

The gold standard treatment in managing CBD stones nowadays is endoscopic retrograde cholangiopancreatography (ERCP). However, ERCP is an invasive procedure and may result in about 6.9% to 12% of adverse events even performed by experienced endoscopists [5], [6]. Some adverse events are lethal and needed to be prevented. To preventing

IEEE Journal of Translational Engineering in Health and Medicine

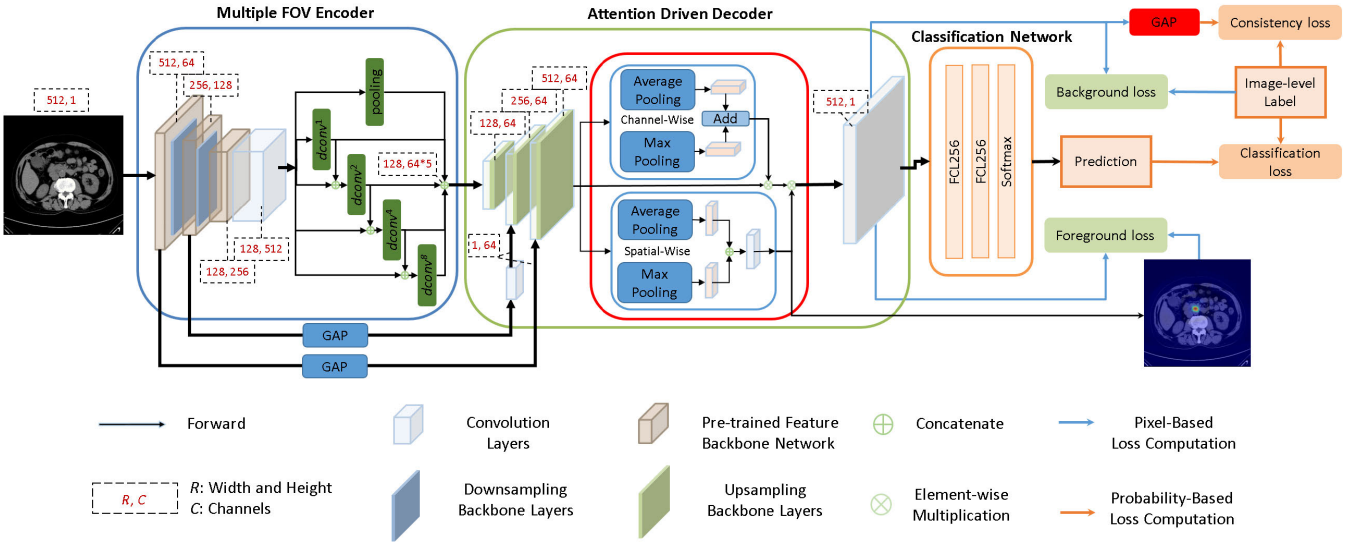

FIGURE 1. **The network architecture of the proposed method. The multiple field-of-view encoder aims to generate deep features of different resolutions to represent CBD stones of different sizes. The attention driven decoder generates the channel attention map, spatial attention map and probability map. The channel attention map enhances the decoder features via effective channel information. The spatial attention map aims to help locate the CBD stones based on image-level labels. The probability map reduces the false detection of CBD stones based on the background loss and serves as the feature for the classification network. The consistency loss ensures that the dominant feature response of the probability map is consistent with the image-level label. Finally, the classification loss drives the predictions of the classification network to classify the CT scan. The dash rectangles show the widths, heights and channels of the features maps.**

ERCP [7], patients suspected of having choledocholithiasis can also be diagnosed by history taking, blood test, physical examination, and ultrasound scanning. However, the positive prediction rates of these tests are ranged from 59% to 64% [8], [9].

Compared with these methods, diagnosing CBD stones from computed tomography (CT) scans achieves better diagnostic accuracy [10], [11]. However, the results are interpreter dependent and the process is time-consuming. Compared with larger gallstones, CBD stones locate in the distal part of the CBD and have relatively small sizes. Thus, they may not be clearly captured by CT scans. Automatically and effectively detecting CBD stones from CT scans becomes a novel and emerging issue in the medical domain. A novel technical solution is expected to address this clinical need in the interdisciplinary field and improve the quality of patient care efficiently.

Recently, supervised convolutional neural networks (CNNs) are widely utilized to solve medical image processing problems [12], [13], [14]. In general, a large number of training data is required for CNNs to learn representative models for object detection. Moreover, object-level labels such as bounding boxes are required for supervised CNNs. These labels bring the time-consuming burdens for physicians and also heavily rely on physicians' experience.

In this paper, we would like to propose a novel technical method to solve the clinic CBD stone detection problem from CT scans in a weakly-supervised manner, i.e. only the image-level labels are given without the requirement of the ground truth bounding boxes of the CBD stones. To achieve the goal, we propose a novel multiple field-of-view based attention driven network (MFADNet). As shown in Fig. 1, the proposed network is composed of a multiple field-of-view encoder, an attention driven decoder and a classification network. We apply the multiple field-of-view encoder to extract the encoder feature based on the dilated convolutions [15] of different dilated rates. The encoder feature represents the multi-scale contextual information of the CT scan. Through the decoder, the encoder feature is further upsampled to obtain the decoder feature of the higher resolution for CBD stone detection. Via the spatial-channel attention scheme [16] of the decoder, the spatial attention map and channel attention map are generated to represent salient responses of the decoder feature for localizing the CBD stones. By integrating the decoder feature, spatial attention map and channel attention map, the probability map is generated to distinguish CBD stones from normal tissues.

To drive the learning of the network via image-level labels, four losses including the foreground loss, background loss, consistency loss and classification loss are proposed. Here, the foreground loss aims to learn the locations of the CBD stones. It indicates the correlations of the spatial attention map and the probability map based on the image-level labels with CBD stones. The background loss aims to avoid miss-detection of CBD stones for normal CT scans. It drives the probability map to represent normal regions based on the image-level labels without CBD stones. The consistency loss aims to ensure that the learned dominant feature response of the probability map needs to be consistent with the image-level labels. Finally, the classification loss is computed based

on the prediction of the classification network to achieve accurate image-level classification. With the combination of these losses, we can train the proposed network with weak labels in an end-to-end training manner. During the inference, the bounding boxes of the detected CBD stones are located by searching the regions of the dominant responses in the spatial attention map. As shown in the experiments, the proposed method can successfully achieve CBD stone detection compared with the state-of-the-art weakly-supervised methods.

There are three main contributions in this paper. First, this is the first deep learning based weakly-supervised method for CBD stone detection from CT scans based on our best knowledge. Our method not only proposes a novel weakly-supervised network for a novel application in the medical domain, but also reduces the burdens of annotations. Second, the attention driven decoder with the proposed foreground loss, background loss and consistency loss helps accurately locate CBD stones by using image-level labels. These losses also drive the end-to-end training of the network. Third, while the class activation map (CAM) based methods [17], [18] aim to obtain salient features based on the learned networks, the proposed method utilizes the aforementioned losses to effectively drive the learning of the features and locations of the CBD stones. Thus, the proposed method achieves outstanding performance compared with state-of-the-art weakly-supervised methods.

The remaining parts of the paper are organized as follows. Sec. II introduces the related weakly-supervised methods. Sec. III presents the proposed method and implementation details. The collected dataset and experimental results are shown in Sec. IV. Finally, the conclusions are given in Sec. V.

II. RELATED WORK

Although supervised deep learning methods [19], [20], [21], [22], [23], [24] have achieved amazing performance for lesion classification, detection or segmentation in the medical domain, time-consuming annotated labels are required. To reduce the labeling burdens of the physicians, weakly-supervised methods are proposed to achieve computer-aided diagnosis for CT scans based on weakly annotated labels. Many weakly-supervised methods [17], [18], [25], [26], [27] are proposed in the computer vision domain. In the following, we focus on the reviews of the weakly-supervised methods in the medical domain.

Wang et al. [28] proposed a label denoising network (LDnet) to segment male pelvic organs from CT scans with 3-D bounding box labels. Li and Xia [29] proposed a deep reinforcement learning-based method for lymph node segmentation by using two cross line-annotations on the lymph node. They used GrabCut [30] to generate pseudo ground truths for U-Net [19]. Tang et al. [31] proposed an attention enhanced model with a regional level set loss to achieve lesion segmentation.

Due to the COVID-19 issues, many state-of-the-art weakly-supervised methods focus on the detection and segmentation of lesions from chest CT scans. Wang et al. [12] proposed a 3-D deep convolutional neural network for COVID-19 detection from CT scans. They segmented lung regions by using a pre-trained U-Net. Then, a 3-D deep neural network and a class activation map (CAM) [17] were applied to predict the COVID-19 infection and localize the regions. Yang et al. [32] proposed a generative adversarial network (GAN) [33] based weakly-supervised method for COVID-19 lesion localization. They subtracted the output image from the input image to localize lesions. Similar GAN based weakly-supervised method can also be found in [34]. Qian et al. [35] presented a multi-task multi-slice deep learning method with two networks. Their method diagnosed diseases for each single CT scan and generated localization maps of abnormalities, while the patient-level classification network provided the prediction based on the features of the network. Liu et al. [36] proposed using scribble-level annotations for segmenting COVID-19 lung infections. Uncertainty-aware and transformation-consistent schemes were considered to make consistent segmentation results with respect to different perturbations of the CT scan.

Besides CT scans, Madooei et al. [37] proposed using a multiple instance learning (MIL) framework for identifying blue-white structure from dermoscopy images based on image-level labels. Chamanzar and Nie [38] proposed a deep learning method to achieve cell segmentation and detection based on point labels. van Sloun and Demi [39] proposed using a fully convolutional neural network to locate B-lines from ultrasound scans. Ma et al. [40] proposed a multi-scale class activation map (MS-CAM) to solve the weakly-supervised geographic atrophy lesion segmentation problem from spectral-domain optical coherence tomography images. The geographic atrophy lesion is retrieved based on the projection of the segmentation of the MS-CAM. Meng et al. [41] proposed a complementary heatmap-based method to achieve multi-retinal disease detection from fundus images with weakly-supervised labels. Qi et al. [42] proposed a graph-regularized embedding network to model the cross-region and cross-image relationships on chest X-ray images for weakly-supervised disease localization.

Compared with the aforementioned weakly-supervised methods, the proposed method extracts the spatial-channel attention feature from the multiple field-of-view feature to detect CBD stones of different sizes. Then, based on the proposed losses and image-level labels, the proposed method can successfully classify the CT scans and locate the CBD stones in an end-to-end trainable manner. In addition, the proposed method is the first weakly-supervised method to detect CBD stones and has been shown to achieve the state-of-the art performance compared with competing weakly-supervised methods from different research domains.

III. PROPOSED METHOD

In this section, an overview of the proposed method is first presented. Then, the multiple field-of-view encoder, the

IEEE Journal of Translational Engineering in Health and Medicine

attention driven decoder and the classification network are described. The losses are introduced to address how can the proposed method effectively locate the CBD stones based on image-level labels. Finally, we provide the implementation details.

A. OVERVIEW

Let a weakly annotated set of CT scans with image-level labels be $D = \{I_n, y_n\}_{n=1}^{N}$, where I_n denotes the nth CT scan and $y_n \in \{1, 0\}$ is its image label to indicate if I_n contains the CBD stone or not, and N is the number of CT scans. The width and height of each CT scan are denoted as W and H. With the weakly annotated dataset D, we aim to derive a detection model to locate CBD stones in CT scans. Fig. 1 illustrates the proposed network architecture.

The proposed network contains a multiple field-of-view encoder, an attention driven decoder and a classification network. The encoder aims to represent CBD stones by using the multiple field-of-view scheme. Then, the feature of the encoder is refined by the attention driven decoder with the proposed losses to locate CBD stones. The classification network provides the image-level predictions of the CT scans.

During training, we consider a CT scan I and its label y, where I serves as the input of the multiple field-of-view encoder. The encoder contains a feature backbone network and a multiple field-of-view network to produce the encoder feature $\boldsymbol{f}_E$ of I. Then, the attention driven decoder with an upsampling backbone network and a spatial-channel attention network [16] is proposed to locate the CBD stones based on the proposed losses. The feature $\boldsymbol{f}_U$ produced by the upsampling backbone network is the input of the spatial-channel attention network. The spatial-channel attention network generates the channel attention map $\boldsymbol{m}_c \in \mathbb{R}^{1\times 64}$ and spatial attention map $\boldsymbol{m}_s \in \mathbb{R}^{WH\times 1}$. By fusing $\boldsymbol{f}_U$, $\boldsymbol{m}_c$ and $\boldsymbol{m}_s$ with a 1×1 convolutional layer, a probability map $\boldsymbol{m}_p \in \mathbb{R}^{WH\times 1}$ is generated to distinguish CBD stones from normal regions.

To enforce the learned spatial attention map $\boldsymbol{m}_s$ to locate CBD stones based on the weakly annotated image-level label, $\boldsymbol{m}_s$ is applied to compute the foreground loss ℓ_{fg} with respect to $\boldsymbol{m}_p$. When the training image does not contain the CBD stone, the background loss ℓ_{bg} is computed based on $\boldsymbol{m}_p$ to avoid the miss-detection of CBD stones of the normal training image. Besides ℓ_{bg}, to ensure that the spatial prediction of $\boldsymbol{m}_p$ can be consistent with the weakly annotated image-level label, a global average pooing layer is applied to $\boldsymbol{m}_p$ to compute the consistency loss ℓ_{con}. Finally, $\boldsymbol{m}_p$ is passed to the classification network to obtain the image-level prediction. A classification loss ℓ_{cls} is computed based on the prediction and the image-level label y. In summary, the whole network is optimized in a weakly-supervised manner by using the following loss function:

$$\ell = \omega_{fg}\ell_{fg} + \omega_{bg}\ell_{bg} + \omega_{con}\ell_{con} + \omega_{cls}\ell_{cls}, \tag{1}$$

where ω_{fg}, ω_{bg}, ω_{con} and ω_{cls} are the weights of the losses. The network will be described in details in the following.

B. MULTIPLE FIELD-OF-VIEW ENCODER

The multiple field-of-view encoder contains a feature backbone network and a multiple field-of-view network. The feature backbone network is used to extract the deep feature $\boldsymbol{f}_b$ to represent I. It is a pre-trained convolutional neural network based on the ImageNet dataset [43]. The downsampling layers of the feature backbone network are achieved by 2×2 max pooling. Because the sizes and spatial context relationship of CBD stones are variant, $\boldsymbol{f}_b$ is hard to provide representative deep features to handle the scale problem of CBD stone detection. Thus, we enhance $\boldsymbol{f}_b$ by using the multiple field-of-view network to extract features of different resolutions as follows.

As shown in Fig. 1, the multiple field-of-view network contains 4 parallel dilated convolutional layers [15] and a maximum pooling layer to represent features in different resolutions. The first dilated feature $\boldsymbol{f}_d^1$ is obtained by a 3×3 dilated convolutional layer of the dilated rate 1. To further extend the discriminability of following learned dilated feature, $\boldsymbol{f}_d^1$ is concatenated with $\boldsymbol{f}_b$ and passed to a 3×3 dilated convolutional layer of the dilated rate 2 to obtain the second dilated feature $\boldsymbol{f}_d^2$. Similarly, the third dilated feature $\boldsymbol{f}_d^3$ and the fourth dilated feature $\boldsymbol{f}_d^4$ of the dilated rates 4 and 8 are obtained based on $\boldsymbol{f}_b$ and the dilated features of previous dilated rates, respectively. The kth dilated feature is defined as follows:

$$\boldsymbol{f}_d^k = \begin{cases} dconv^{2^{k-1}}(\boldsymbol{f}_b), & k = 1 \\ dconv^{2^{k-1}}(\boldsymbol{f}_b \oplus \boldsymbol{f}_d^{k-1}), & k > 1 \end{cases}, \tag{2}$$

where $dconv^{2^{k-1}}(\cdot)$ is a 3×3 dilated convolutional layer of the dilated rate 2^{k-1} and $\oplus$ is the concatenation operator of the backbone feature and the previous dilated feature.

Finally, $\boldsymbol{f}_b$ is passed to a 2×2 max pooling layer to obtain the pooling feature $\boldsymbol{f}_p$. By concatenating the dilated features and the pooling feature of the multiple field-of-view network, the obtained encoder feature $\boldsymbol{f}_E$ can represent the multi-scale contextual information of I and is defined as:

$$\boldsymbol{f}_E = \boldsymbol{f}_d^1 \oplus \boldsymbol{f}_d^2 \oplus \boldsymbol{f}_d^3 \oplus \boldsymbol{f}_d^4 \oplus \boldsymbol{f}_p, \tag{3}$$

where $\oplus$ is the concatenation operator. The encoder feature then serves as the input of the attention driven decoder for CBD stone detection.

C. ATTENTION DRIVEN DECODER

The attention driven decoder contains an upsampling backbone network and a spatial-channel attention network to locate the CBD stones based on the proposed losses. The upsampling backbone network upsamples the encoder feature $\boldsymbol{f}_E$ to obtain the decoder feature $\boldsymbol{f}_U$ which has the same spatial resolution of the input image. Two forward connections from the feature backbone network of the encoder provide features of different resolutions to help obtain better decoder features during upsampling. The encoder features of the first block and second block of the feature backbone network are passed to spatial-wise global average pooling layers to

generate the features. To ensure the consistency of the feature dimension, a 1×1 convolutional layer with 64 channels is used to modify the dimension of the feature of the second block after the global average pooling. These features then serve as the weights to multiply the decoder features of the corresponding blocks of the decoder as shown in Fig. 1.

Instead of considering the feature responses of the low-resolution feature maps, we propose applying the spatial-channel attention network based on the output of the upsampling backbone network to compute the channel attention map $\boldsymbol{m}_c$ and spatial attention map $\boldsymbol{m}_s$. In this way, the computed attention maps can better represent the locations of the CBD stones in the original resolution. The spatial-channel attention network based on the convolutional block attention module [16] is consisted of a channel attention module and a spatial attention module for feature and loss computation. In our method, the channel attention module aims to extract the channel attention map $\boldsymbol{m}_c$ which contains representative information of different channels of $\boldsymbol{f}_U$. $\boldsymbol{m}_c$ is computed by the addition of the channel features after a spatial-wise max pooling layer and a spatial-wise global average pooling layer. Because effective channel information will be reserved, $\boldsymbol{m}_c$ is used as the weight map to enhance important features in $\boldsymbol{f}_U$.

The spatial attention module aims to obtain spatial attention map $\boldsymbol{m}_s$ which is used to locate the CBD stones. $\boldsymbol{f}_U$ is passed to a channel-wise max pooling layer and a channel-wise average pooling layer, respectively, to obtain spatial attention features. These features are concatenated and passed to a 7×7 convolutional layer $conv^7(\cdot)$ to obtain the spatial attention map $\boldsymbol{m}_s$ as follows:

$$\boldsymbol{m}_s = conv^7(pool_c^{max}(\boldsymbol{f}_U) \oplus pool_c^{ave}(\boldsymbol{f}_U)), \tag{4}$$

where $pool_c^{max}(\cdot)$ and $pool_c^{ave}(\cdot)$ are the channel-wise max pooling and channel-wise average pooling functions. If I contains CBD stones, these two pooling functions help emphasize the responses of the CBD stones.

The channel attention map $\boldsymbol{m}_c$ aims to extract representative channel features of $\boldsymbol{f}_U$ and is defined as follows:

$$\boldsymbol{m}_c = pool_s^{max}(\boldsymbol{f}_U) + pool_s^{ave}(\boldsymbol{f}_U), \tag{5}$$

where $pool_s^{max}(\cdot)$ and $pool_s^{ave}(\cdot)$ are the spatial-wise max pooling and spatial-wise average pooling functions. These two spatial-wise pooling functions help find representative features for CBD stones from $\boldsymbol{f}_U$.

Finally, a probability map $\boldsymbol{m}_p$ is computed as follows:

$$\boldsymbol{m}_p = conv^1(\boldsymbol{m}_s \otimes (\boldsymbol{m}_c \otimes \boldsymbol{f}_U)), \tag{6}$$

where $conv^1(\cdot)$ is a 1×1 convolutional layer followed by a sigmoid function, and $\otimes$ is the element-wise multiplication. $\boldsymbol{m}_p$ aims to help distinguish CBD stones from backgrounds.

To locate CBD stones, we cooperate $\boldsymbol{m}_p$ with $\boldsymbol{m}_s$ to compute the foreground loss ℓ_{fg} which enforces the spatial attention map to learn the locations of the CBD stones. When $\boldsymbol{m}_p$ indicates the low probability of the CBD stones for certain pixels, the spatial attention map should also have low attention responses for these pixels. When a training image contains the CBD stone, the network needs to learn the locations of CBD stones based on the pixel-wise feature responses of $\boldsymbol{m}_s$. In other words, the pixels with high feature responses of CBD stones in the spatial attention map $\boldsymbol{m}_s$ should have low feature responses of $\boldsymbol{m}_p$.

By enforcing the learning of $\boldsymbol{m}_s$ to locate the CBD stones with respect to $\boldsymbol{m}_p$, the foreground loss ℓ_{fg} is defined based on pixel-wise feature responses between $\boldsymbol{m}_s$ and $\boldsymbol{m}_p$ as follows:

$$\ell_{fg} = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} (1 - \boldsymbol{m}_s^{(i,j)}) \times \boldsymbol{m}_p^{(i,j)}, \tag{7}$$

where $\boldsymbol{m}_s^{(i,j)}$ is the feature response of the pixel at position (i, j) of $\boldsymbol{m}_s$, $\boldsymbol{m}_p^{(i,j)}$ is the feature response of the pixel at position (i, j) of $\boldsymbol{m}_p$. The foreground loss offers an supervisory signal to identify possible locations of CBD stones in the weakly-supervised training process of the network.

Besides the foreground loss to locate CBD stones, it is also important to identify normal CT scans without CBD stones based on image-level labels. To distinguish CBD stones from normal regions, we propose the probability map $\boldsymbol{m}_p$. To drive the learning of $\boldsymbol{m}_p$ to learn normal regions without CBD stones, we propose the background loss ℓ_{bg}. The background loss aims to reduce the feature responses of $\boldsymbol{m}_p$ when inputting a normal training CT scan. Because we only have the image-level label y of the training CT scan, we define the label $y^{(i,j)}$ of the pixel at position (i, j) of the training CT scan as $y^{(i,j)} = y$. The background loss is then defined as follows:

$$\ell_{bg} = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} (1 - y^{(i,j)}) \times \boldsymbol{m}_p^{(i,j)}. \tag{8}$$

If the input training CT scan contains the CBD stone, i.e. $y = 1$, it should not be used to learn $\boldsymbol{m}_p$ to avoid the incorrect learning of CBD stones as backgrounds. Thus, only the normal training CT scans affect the computation of $\boldsymbol{m}_p$. By minimizing ℓ_{bg}, the network can effectively represent normal training CT scans. When the feature responses of $\boldsymbol{m}_p$ are small for normal CT scans, the false detection of CBD stones can be avoided.

Besides the pixel-based foreground loss and background loss, we also propose the consistency loss and classification loss which are probability-based loss functions computed by using the dominant feature responses and the predictions to drive the learning of the network based on global image-level labels. While the probability map $\boldsymbol{m}_p$ indicates possible normal background regions and CBD stones of the input CT scans, we would like to ensure that the dominant feature response of $\boldsymbol{m}_p$ is consistent with the image-level labels of the input CT scans. If the CT scans are normal, the learned $\boldsymbol{m}_p$ should contain feature responses with respect to normal image-level labels. Similarly, when the CT scans contain CBD stones, their dominant feature responses should also

IEEE Journal of Translational Engineering in Health and Medicine

be consistent with the image-level labels. To address the dominant feature responses for both cases, we propose the consistency loss ℓ_{con} as follows:

$$\ell_{con} = -y\log(GAP(\boldsymbol{m}_p)) + (1-y)\log(1-GAP(\boldsymbol{m}_p)), \tag{9}$$

where $GAP(\cdot)$ is the global average pooling layer which represents the dominant feature response of $\boldsymbol{m}_p$. By using the consistency loss, each training CT scan produces an extra dominant feature response which needs to be consistent with the ground truth image-level label to optimize the whole network.

Finally, $\boldsymbol{m}_p$ is used as the input of the classification network consisting of two fully connected layers followed by a softmax layer. The classification network aims to figure out if CBD stones exist in I or not. To guide the model learning, the classification loss ℓ_{cls} is defined as follows:

$$\ell_{cls} = -y\log(\hat{y}) + (1-y)\log(1-\hat{y}), \tag{10}$$

where $\hat{y}$ is the prediction of the classification network. Based on Eq. (10), we can ensure that the predictions of the classification network are consistent with the ground truth image-level labels. Moreover, it will also drive the whole network to learn proper features.

D. IMPLEMENTATION DETAILS

The feature backbone network [44] is modified from a pre-trained VGG-16 network [45] by adding dropout layers to VGG blocks to avoid overfitting. The upsampling backbone network is composed of three 3×3 convolutional layers and one 1×1 convolutional layer. Each 3×3 convolutional layer is followed by an instance normalization layer. Finally, the classification network consists of two fully connected layers with 256 neurons and a softmax layer for image-level prediction.

The Keras framework on an Intel Core i7 computer with a 3.7 GHz CPU and RTX 2080 GPU is used to implement the proposed method. We use the RMSProp optimizer to train the model. The parameters of ρ and ϵ of the RMSProp optimizer are set to 0.9 and 10^{-8}. The batch size is set to 3. The initial learning rate is set to 10^{-4}. When the loss of the validation stops improving for 5 epochs, the learning rate will be decreased by a factor of 10. The maximal training epoch is set to 100, and if the loss of the validation stops improving in 10 epochs, the training will end early. The weights of the losses are set to $\omega_{fg} = 1$, $\omega_{bg} = 0.5$, $\omega_{con} = 1$ and $\omega_{cls} = 1$, because we empirically found that suppressing the background loss helps increase the detection rate of the CBD stones. To locate the CBD stones in the CT scans, we first extract the feature map from the spatial-channel attention network. To extract salient regions which reflect locations of CBD stones, we apply the channel-wise average pooling to obtain important channel information of the feature map through the channel dimensions. To avoid false detection of the CBD stone of normal CT scans, we apply the probability map as the weight map. In this way, we can obtain the CBD stone attention map $\boldsymbol{m}_a$ as follows:

$$\boldsymbol{m}_a = pool_c^{ave}(\boldsymbol{m}_s \otimes (\boldsymbol{m}_c \otimes \boldsymbol{f}_U)) \otimes (1-\boldsymbol{m}_p). \tag{11}$$

Here, $\boldsymbol{m}_a$ is normalized by the maximal value of $\boldsymbol{m}_a$. When the normalized values of pixels are larger than the threshold th $(= 0.6)$ and the classification network predicts that the CBD stone exists in the CT scan, these pixels are considered as pixels with CBD stones. Finally, a bounding box is used to extract detection results based on the largest connected component region which is the same as CAM [17].

IV. EXPERIMENTAL RESULTS

A. EXPERIMENTAL SETTINGS

1) DATASET

From January 2018 to August 2019, patients who were clinically suspicious of CBD stones and met 2010 American society for gastrointestinal endoscopy (ASGE) high probability criteria for CBD stones in the National Cheng Kung University Hospital were included. All patients were presented with one of the following conditions: (a) total bilirubin more than 4 mg/dL, (b) total bilirubin level ranged from $1.8-3.9$ mg/dL with a dilated (diameter $>$ 6mm) common bile duct on images, (c) presented ascending cholangitis clinically, and (d) abdominal ultrasound revealed CBD stone. After the enrollment, patients with known malignancy or medical implants inside the biliary system that causes obstruction of the CBD were excluded. Patients without pre-treatment abdominal CT scans, adolescent and pregnant women were also excluded from the initial cohort. The dataset is approved by Institutional Review Board, National Cheng Kung University Hospital under B-ER-111-186.

In the experiments, abdomen CT scans near the gallbladder regions of 252 patients were collected. 428 CT scans with CBD stones were selected by the physicians. Because the number of the normal CT scans are significantly more than that of the CT scans with CBD stones, the physicians limited the number of the normal CT scans to be twice the number of CT scans with CBD stones to reduce the data imbalance problem and retain the fact that the number of the normal CT scans is more than that of CT scans with CBD stones. To provide more diversity, the number of selected normal CT scans for each patient was randomly decided by the physicians and thus 856 normal CT scans were randomly selected. As a result, each patient has $\sim$ 5 selected CT scans on average. The CT scans of the dataset is divided into 7:3 for the training and testing as the setting in [46], [47], and [48]. The training dataset contains 298 CT scans with CBD stones and 596 normal CT scans. The testing dataset contains 130 CT scans with CBD stones and 260 normal CT scans. Only the image-level labels were applied for a weakly-supervised training manner. The resolution of the CT scans is 512×512.

We apply accuracy, sensitivity, specificity, F1-score metrics to evaluate the classification performance of the proposed method and state-of-the-art methods. Moreover, to evaluate the weakly-supervised detection performance of CBD

TABLE 1. Ablation study.

Methods	Accuracy	Sensitivity	Specificity	F1-score
w/o M-FOV	0.7821	0.7154	0.8154	0.6836
w/o $\boldsymbol{m}_c$	0.8769	0.7462	**0.9423**	0.8017
w/o $\boldsymbol{m}_s$	0.8692	0.7308	0.9385	0.7884
w/o ℓ_{bg}	0.8744	0.8000	0.9115	0.8093
w/o ℓ_{con}	0.8718	0.8000	0.9077	0.8062
Ours	**0.9051**	**0.8692**	0.9231	**0.8593**

TABLE 2. Classification results compared with competing methods.

Methods	Accuracy	Sensitivity	Specificity	F1-score
CAM	0.8179	0.8077	0.8231	0.7473
Grad-CAM	0.8128	0.7308	0.8538	0.7224
MS-CAM	0.8154	0.8462	0.8000	0.7534
SPA	0.8846	0.8077	**0.9231**	0.8235
Ours	**0.9051**	**0.8692**	**0.9231**	**0.8593**

stones for each method, the mean intersection over union (mIoU) and the average precision (AP) [49] values were employed, and the ground truth bounding boxes of the testing images were manually labelled by an experienced physician. In addition, we used the same procedure shown in CAM [17] to draw bounding boxes for all of the competing methods.

2) COMPARATIVE BASELINES

Based on our best knowledge, the proposed method is the first weakly-supervised method for CBD stone detection from CT scans. Thus, we compared our method with four state-of-the-art weakly-supervised learning methods from the computer vision domain and medical domain for the evaluations of CBD stone classification and detection. The first competing method is the class activation map (CAM) [17] which applies the global average pooling on the convolutional feature maps before a fully-connected layer to identify the importance of the image regions for object detection. To provide more general explanations of activation maps in convolutional neural networks, Grad-CAM [18] is proposed by using the gradient information back-propagated to the convolutional layer of interest. To extract activation features by using multiple scale information, MS-CAM [40] is proposed for geographic atrophy lesion detection. While the localization of the CAM based methods is easily affected by salient feature responses, structure-preserving activation (SPA) [50] is proposed to extract object structural information for object detection.

B. ABLATION STUDY

The results of the ablation study are shown in Table 1. The first row shows the results without (w/o) the multiple field-of-view (M-FOV) network. When the multiple field-of-view information is not considered, the sensitivity significantly drops. Such results show the importance of the multiple field-of-view network to help extract representative deep features. The second row shows the results without the channel attention map $\boldsymbol{m}_c$ which represents effective channel information of different channels of the decoder feature. When the decoder feature cannot be enhanced by the learned channel features, the classification accuracy of the network degrades. Moreover, the learned features are hard to represent CBD stones and thus the sensitivity also degrades. Compared with $\boldsymbol{m}_c$, the spatial attention map $\boldsymbol{m}_s$ aims to learn the locations of the CBD stones based on image-level labels. It is also used to compute the foreground loss ℓ_{fg}. As shown in the third row of Table 1, without $\boldsymbol{m}_s$, the sensitivity significantly drops which indicates the importance of $\boldsymbol{m}_s$ and ℓ_{fg} to identify CBD stones. Please note that results of the ablation study without $\boldsymbol{m}_c$ and $\boldsymbol{m}_s$ were obtained by ignoring corresponding terms in Eqs. (4) and (5), respectively. The background loss ℓ_{bg} aims to ensure that $\boldsymbol{m}_p$ can indicate normal regions when the training CT scan does not contain CBD stones. Without ℓ_{bg}, the specificity drops compared with the proposed method in the fourth row of Table 1. This result shows that ℓ_{bg} helps reduce false detection of CBD stones for normal CT scans. The fifth row shows the results without the consistency loss ℓ_{con}. Because ℓ_{con} aims to ensure that the dominant feature response of $\boldsymbol{m}_p$ is consistent with the image-level labels, both sensitivity and specificity of the method without ℓ_{con} drop compared with the proposed method. The ablation study shows that the proposed network structure and losses are effective.

C. QUANTITATIVE RESULTS

Two kinds of the quantitative performance are compared. The first one is the image-based classification performance shown in Table 2. Compared with the CAM based methods, SPA considers a restricted activation module during object localization to avoid the affections of local extremely high responses in CAM. Thus, SPA achieves a better specificity compared with CAM based methods. The proposed method further considers the foreground loss, background loss and the consistency loss. Thus, it can achieve high sensitivity compared with SPA.

The second quantitative performance is to evaluate the weakly-supervised detection results of each method. When the methods can correctly locate CBD stones in the weakly-supervised manner, the detected regions should overlap the ground truth regions. In the CAM based methods, they do not consider to learn the locations of CBD stones but only consider the local extremely high responses of the feature maps. SPA contains a self-correlation map generating module which can improve the attention map based on the structural information to better locate target objects. Because CBD stones are usually inconspicuous and may not contain self-correlations in CT scans, these competing methods fail to locate CBD stones based on their salient feature responses in the weakly-supervised manner. Thus, the area of intersection between the detected regions of these methods and the ground truth regions is very small which leads to significantly low mIoU and AP values of these methods as shown in Table 3. The visualization results in Sec. IV-D also indicate

IEEE Journal of Translational Engineering in Health and Medicine

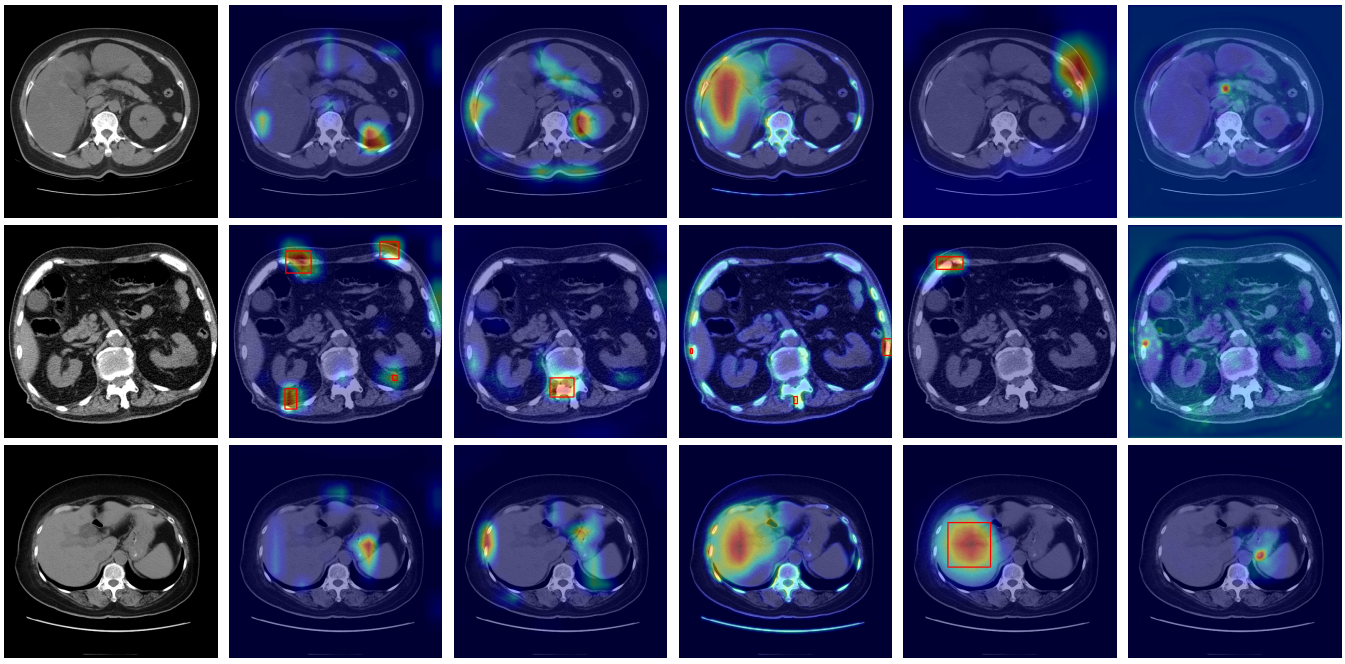

FIGURE 2. The CBD stone detection results with attention maps for normal CT scans. (a) Ground truth, (b) CAM, (c) Grad-CAM, (d) MS-CAM, (e) SPA, and (f) the proposed method. The red rectangles indicate the false detection results of the CBD stones.

TABLE 3. Detection results compared with competing methods.

Methods	mIoU	AP
CAM	0.0206	0.0263
Grad-CAM	0.0011	0.0002
MS-CAM	0.0127	0.0108
SPA	0.0015	0.0017
Ours	**0.5309**	**0.5707**

TABLE 4. The confusion matrix of the proposed method.

	Predict-Positive	Predict-Negative
Real-Positive	113	17
Real-Negative	20	240

that the detected regions of these methods fail to overlap the ground truth regions. In contrast, the proposed method can achieve better mIoU and AP values under the guidance of the spatial-channel attention network with the proposed losses.

Table 4 shows the confusion matrix of the proposed method. Most CT scans are correctly classified to show that the proposed method can distinguish the CBD stone cases from normal cases. In addition, the average inference time of the proposed method is 0.067 seconds to show the potential of the real-time usage of the proposed method.

D. QUALITATIVE RESULTS

Fig. 2 shows the qualitative results of the state-of-the-art methods and the proposed method for normal CT scans of three patients. The attention maps of each method are also shown in Fig. 2. The red rectangles show the CBD stone detection results when the method incorrectly classifies the normal CT scans as the abnormal CT scans with CBD stones. Fig. 2(a) shows the ground truth of normal CT scans. The results of CAM are shown in Fig. 2(b). Because CAM considers the feature responses after global average pooling, the learned features then easily focus on bone and angiosteosis regions which are salient compared with other organs in the CT scans. Similar results can also be observed in Grad-CAM as shown in Fig. 2(c). As a result, the false detection and false classification results of CAM and Grad-CAM are observed from the CT scan of the second patient.

Compared with CAM and Grad-CAM, MS-CAM considers the scale information with the attention mechanism of the fully connected operations. It can then observe liver regions as shown in the CT scans of the first and third patients of Fig. 2(d). Nevertheless, MS-CAM still incorrectly detects CBD stones for the second patient due to the salient features of bones. As indicated by SPA, these CAM based methods easily miss object structure information because extremely high feature responses are considered. To solve the problem, a restricted activation module is proposed in SPA. Nevertheless, extremely high feature responses such as bone and liver regions still affect the classification results of SPA for the second and third patients as shown in Fig. 2(e). Thus, SPA also incorrectly locates the bone and liver regions as CBD stones. Fig. 2(f) shows the results of the proposed method. With the foreground loss, background loss and consistency loss, the learned attention maps focus on the regions which can distinguish CBD stones from ambiguous regions such as bone regions of the second patient and angiosteosis regions of

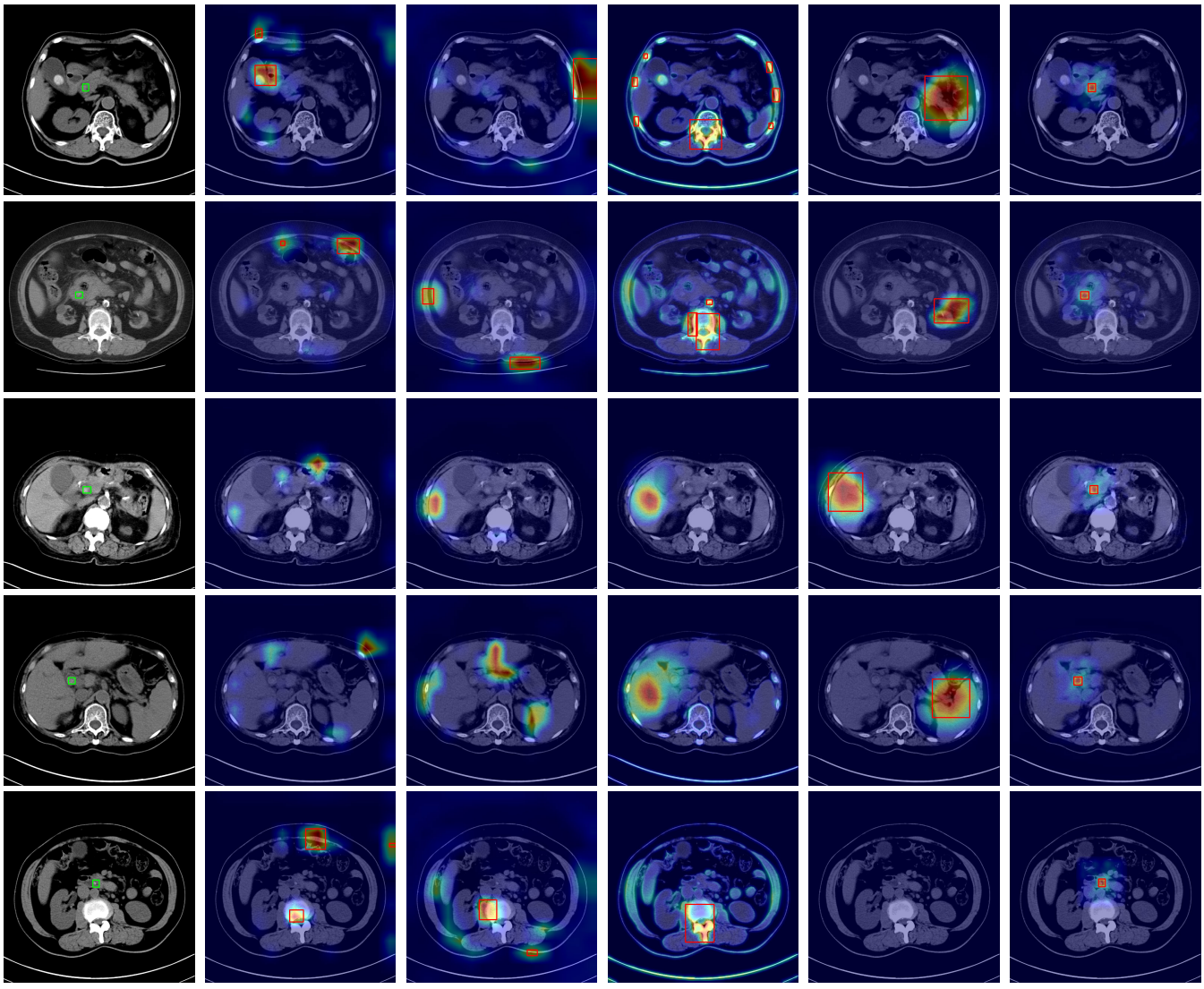

FIGURE 3. The CBD stone detection results with attention maps for CT scans with CBD stones. (a) Ground truth, (b) CAM, (c) Grad-CAM, (d) MS-CAM, (e) SPA, and (f) the proposed method. The red rectangles indicate the false detection results of the CBD stones.

the third patient. Thus, the proposed method can successfully classify these CT scans as normal CT scans based on the learned features.

Fig. 3(a) shows the ground truth regions of CBD stones in CT scans by using green rectangles. The results of CAM, Grad-CAM, MS-CAM, SPA and the proposed method are shown in Figs. 3(b), (c), (d), (e), and (f), respectively. The same as the observations in Fig. 2(b), the salient feature responses of CAM are affected by bone regions for the second and fifth patients, and the gallstones of the first patient. Although CAM successfully classifies the CT scans of the first, second, and fifth patients as CT scans with CBD stones, the locations of detected CBD stones are incorrect compared with the ground truth. Because the CBD stones of the third and fourth patients are not clearly captured by CT scans, the learned salient features of CAM cannot represent the CBD stones for classification. Thus, miss-detection of CBD stones occurs for these two patients. Similar results can also be observed for Grad-CAM and MS-CAM in Fig. 3(c) and (d).

Compared with CAM based methods, SPA considers the restricted activation module to learn local object structure and self-correlation to refine localization maps. Thus, the salient features of SPA can focus on non-bone regions. Nevertheless, SPA still fails to correctly locate CBD stones compared with the ground truth for the first four patients as shown in Fig. 3(e). Because CBD stones are relatively small and are not clearly captured in CT scans, the modules of SPA are hard to learn local object structure of CBD stones from only image-level labels. As a result, the learned features of SPA cannot represent CBD stones for the fifth patient.

As shown in Table 3, the competing methods have low mIoU and AP values. Such results can be visually explained

IEEE Journal of Translational Engineering in Health and Medicine

by the salient feature responses of these methods shown in Fig. 3(b), (c), (d) and (e), respectively. These competing methods focus on regions without CBD stones compared with the ground truth regions. Thus, the mIoU values of these methods are naturally low because the detected regions do not overlap the ground truth regions. In contrast, the visualizations shown in Fig. 3(f) reveal that the proposed method provides an interpretable AI model which truly focuses on CBD stones. The cooperation of losses and the whole network provides a novel weakly-supervised learning way to learn salient features to represent CBD stones. Thus, the mIoU and AP values of the proposed method are significantly better than those of the competing methods.

V. CONCLUSION

In summary, we propose a novel multiple field-of-view based attention driven network for a new medical application of CBD stone detection from CT scans. Different from CAM based methods, the proposed method is composed of a multiple field-of-view encoder, an attention-driven decoder and a classification network. While the encoder learns representative multiple field-of-view features from CT scans, the decoder learns the locations of CBD stones based on the spatial-channel attention network with the proposed foreground loss, background loss and consistency loss from image-level labels. The classification network provides the image-level prediction. By the guidance of the proposed losses, the network is end-to-end trainable in a weakly-supervised manner. Also shown in the experimental results, CBD stones can be accurately detected and located compared with the competing methods. To address the clinic problem of CBD stone detection, we develop the unique engineering solution with the collaboration between physicians and engineers to achieve the CBD stone diagnosis. In the future, we will apply the proposed method to different interdisciplinary fields of biomedical engineering such as [48], [51], and [52] to evaluate the generalization capability of the proposed method.

ACKNOWLEDGMENT

The authors thank the National Center for High-performance Computing (NCHC) for providing computational and storage resources.

REFERENCES

[1] C. Molvar and B. Glaenzer, "Choledocholithiasis: Evaluation, treatment, and outcomes," in *Proc. Seminars Interventional Radiol.*, 2016, vol. 33, no. 4, pp. 268–276.

[2] R. Costi, A. Gnocchi, F. Di Mario, and L. Sarli, "Diagnosis and management of choledocholithiasis in the golden age of imaging, endoscopy and laparoscopy," *World J. Gastroenterol.*, vol. 20, no. 37, p. 13382, 2014.

[3] A. Ahmed, R. C. Cheung, and E. B. Keeffe, "Management of gallstones and their complications," *Amer. Family Physician*, vol. 61, no. 6, pp. 1673–1680, 2000.

[4] S. Tazuma et al., "Evidence-based clinical practice guidelines for cholelithiasis 2016," *J. Gastroenterol.*, vol. 52, no. 3, pp. 276–300, 2017.

[5] A. Andriulli et al., "Incidence rates of post-ERCP complications: A systematic survey of prospective studies," *Amer. J. Gastroenterol.*, vol. 102, no. 8, pp. 1781–1788, Aug. 2007.

[6] T. Glomsaker, G. Hoff, J. T. Kvaløy, K. Søreide, L. Aabakken, and J. A. Søreide, "Patterns and predictive factors of complications after endoscopic retrograde cholangiopancreatography," *Brit. J. Surg.*, vol. 100, no. 3, pp. 373–380, Jan. 2013.

[7] U. B. Kuzu et al., "Management of suspected common bile duct stone: Diagnostic yield of current guidelines," *HPB*, vol. 19, no. 2, pp. 126–132, Feb. 2017.

[8] R. M. Narvaez-Rivera et al., "Accuracy of ASGE criteria for the prediction of choledocholithiasis," *Revista Espanola de Enfermedades Digestivas*, vol. 108, no. 6, pp. 309–314, 2016.

[9] H. He et al., "Accuracy of ASGE high-risk criteria in evaluation of patients with suspected common bile duct stones," *Gastrointestinal Endoscopy*, vol. 86, no. 3, pp. 525–532, Sep. 2017.

[10] S. W. Anderson, B. C. Lucey, J. C. Varghese, and J. A. Soto, "Accuracy of MDCT in the diagnosis of choledocholithiasis," *Amer. J. Roentgenol.*, vol. 187, no. 1, pp. 174–180, Jul. 2006.

[11] C.-W. Tseng, C.-C. Chen, T.-S. Chen, F.-Y. Chang, H.-C. Lin, and S.-D. Lee, "Can computed tomography with coronal reconstruction improve the diagnosis of choledocholithiasis?" *J. Gastroenterol. Hepatol.*, vol. 23, no. 10, pp. 1586–1589, Oct. 2008.

[12] X. Wang et al., "A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2615–2625, Aug. 2020.

[13] T.-H. Lin, J.-Y. Jhang, C.-R. Huang, Y.-C. Tsai, H.-C. Cheng, and B.-S. Sheu, "Deep ensemble feature network for gastric section classification," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 1, pp. 77–87, Jan. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9107452

[14] J.-Y. Jhang, Y.-C. Tsai, T.-C. Hsu, C.-R. Huang, H.-C. Cheng, and B.-S. Sheu, "Gastric section correlation network for gastric precancerous lesion diagnosis," *IEEE Open J. Eng. Med. Biol.*, pp. 1–9, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10128879

[15] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–13.

[16] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Euro. Conf. Comput. Vis.*, 2018, pp. 3–19.

[17] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[19] 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.

[20] X. Li, H. Chen, X. Qi, Q. Dou, C. Fu, and P. Heng, "H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018.

[21] S. Park, H. M. Gach, S. Kim, S. J. Lee, and Y. Motai, "Autoencoder-inspired convolutional network-based super-resolution method in MRI," *IEEE J. Transl. Eng. Health Med.*, vol. 9, pp. 1–13, 2021.

[22] J. Zhang et al., "MLBF-Net: A multi-lead-branch fusion network for multi-class arrhythmia classification using 12-lead ECG," *IEEE J. Transl. Eng. Health Med.*, vol. 9, pp. 1–11, 2021.

[23] M. A. Ottom, H. A. Rahman, and I. D. Dinov, "ZNet: Deep learning approach for 2D MRI brain tumor segmentation," *IEEE J. Transl. Eng. Health Med.*, vol. 10, pp. 1–8, 2022.

[24] S. Huang, C. Yu, Y. Liao, and C. Huang, "Evaluations of deep learning methods for pathology image classification," in *Proc. IEEE Biomed. Circuits Syst. Conf. (BioCAS)*, Oct. 2022, pp. 95–99.

[25] K. K. Singh and Y. J. Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3544–3553.

[26] J. Choe and H. Shim, "Attention-based dropout layer for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2214–2223.

[27] T. Wu et al., "Embedded discriminative attention mechanism for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16760–16769.

[28] S. Wang et al., "Iterative label denoising network: Segmenting male pelvic organs in CT from 3D bounding box annotations," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 10, pp. 2710–2720, Oct. 2020.

[29] Z. Li and Y. Xia, "Deep reinforcement learning for weakly-supervised lymph node segmentation in CT images," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 3, pp. 774–783, Mar. 2021.

[30] C. Rother, V. Kolmogorov, and A. Blake, "'GrabCut': Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004.

[31] Y. Tang et al., "Weakly-supervised universal lesion segmentation with regional level set loss," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2021, pp. 515–525.

[32] Z. Yang, L. Zhao, S. Wu, and C. Y. Chen, "Lung lesion localization of COVID-19 from chest CT image: A novel weakly supervised learning method," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 6, pp. 1864–1872, Jun. 2021.

[33] I. J. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[34] Y. Yang et al., "Towards unbiased COVID-19 lesion localisation and segmentation via weakly supervised learning," in *Proc. IEEE 18th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2021, pp. 1966–1970.

[35] X. Qian et al., "M^3Lung-Sys: A deep learning system for multi-class lung pneumonia screening from CT imaging," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 12, pp. 3539–3550, 2020.

[36] X. Liu et al., "Weakly supervised segmentation of COVID19 infection with scribble annotation on CT images," *Pattern Recognit.*, vol. 122, Feb. 2022, Art. no. 108341.

[37] A. Madooei, M. S. Drew, and H. Hajimirsadeghi, "Learning to detect blue-white structures in dermoscopy images with weak supervision," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 2, pp. 779–786, Mar. 2019.

[38] A. Chamanzar and Y. Nie, "Weakly supervised multi-task learning for cell detection and segmentation," in *Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2020, pp. 513–516.

[39] R. J. G. van Sloun and L. Demi, "Localizing B-lines in lung ultrasonography by weakly supervised deep learning, in-vivo results," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 4, pp. 957–964, Apr. 2020.

[40] X. Ma, Z. Ji, S. Niu, T. Leng, D. L. Rubin, and Q. Chen, "MS-CAM: Multi-scale class activation maps for weakly-supervised segmentation of geographic atrophy lesions in SD-OCT images," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 12, pp. 3443–3455, Dec. 2020.

[41] Q. Meng, L. Liao, and S. Satoh, "Weakly-supervised learning with complementary heatmap for retinal disease detection," *IEEE Trans. Med. Imag.*, vol. 41, no. 8, pp. 2067–2078, Aug. 2022.

[42] B. Qi et al., "GREN: Graph-regularized embedding network for weakly-supervised disease localization in X-ray images," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 10, pp. 5142–5153, Oct. 2022.

[43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[44] L. A. Lim and H. Y. Keles, "Learning multi-scale features for foreground segmentation," *Pattern Anal. Appl.*, vol. 23, no. 3, pp. 1369–1380, Aug. 2020.

[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, May 2015, pp. 1–14.

[46] T. Tan et al., "Optimize transfer learning for lung diseases in bronchoscopy using a new concept: Sequential fine-tuning," *IEEE J. Transl. Eng. Health Med.*, vol. 6, pp. 1–8, 2018.

[47] T. I. Mahmud, S. A. Imran, and C. Shahnaz, "Res-SE-ConvNet: A deep neural network for hypoxemia severity prediction for hospital in-patients using photoplethysmograph signal," *IEEE J. Transl. Eng. Health Med.*, vol. 10, pp. 1–9, 2022.

[48] C.-M. Nien, E.-H. Yang, W.-L. Chang, H.-C. Cheng, and C.-R. Huang, "Criss-cross attention based multi-level fusion network for gastric intestinal metaplasia segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent. Workshop Imag. Syst. GI Endoscopy*, 2022, pp. 13–23.

[49] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Euro. Conf. Comput. Vis.*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. 2014, pp. 740–755.

[50] X. Pan et al., "Unveiling the potential of structure preserving for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11637–11646.

[51] A. Masood et al., "Cloud-based automated clinical decision support system for detection and diagnosis of lung cancer in chest CT," *IEEE J. Transl. Eng. Health Med.*, vol. 8, pp. 1–13, 2020.

[52] A. K. Mondal, A. Bhattacharjee, P. Singla, and A. P. Prathosh, "XViT-COS: Explainable vision transformer based COVID-19 screening using radiography," *IEEE J. Transl. Eng. Health Med.*, vol. 10, pp. 1–10, 2022.