3D Multi-Attention Guided Multi-Task Learning Network for Automatic Gastric Tumor Segmentation and Lymph Node Classification

Yongtao Zhang, Haimei Li, Jie Du, Jing Qin, *Member, IEEE*, Tianfu Wang, Yue Chen, Bing Liu, Wenwen Gao, Guolin Ma, and Baiying Lei, *Senior Member, IEEE*

***Abstract*— Automatic gastric tumor segmentation and lymph node (LN) classification not only can assist radiologists in reading images, but also provide image-guided clinical diagnosis and improve diagnosis accuracy. However, due to the inhomogeneous intensity distribution of gastric tumor and LN in CT scans, the ambiguous/missing boundaries, and highly variable shapes of gastric tumor, it is quite challenging to develop an automatic solution. To comprehensively address these challenges, we propose a novel 3D multi-attention guided multi-task learning network for simultaneous gastric tumor segmentation and LN classification, which makes full use of the complementary information extracted from different dimensions, scales, and tasks. Specifically, we tackle task correlation and heterogeneity with the convolutional neural network consisting of scale-aware attention-guided shared feature learning for refined and universal multi-scale features, and task-aware attention-guided feature learning for task-specific discriminative features. This shared feature learning is equipped with two types of scale-aware attention (visual attention and adaptive spatial attention) and two stage-wise deep supervision paths. The task-aware attention-guided feature learning comprises a segmentation-aware attention module and a classification-aware attention module. The proposed 3D multi-task learning network can balance all tasks by combining segmentation and classification loss functions with weight uncertainty. We evaluate our model on an in-house CT images dataset collected from three medical centers. Experimental results demonstrate that our method outperforms the state-of-the-art algorithms, and obtains promising performance for tumor segmentation and LN classification. Moreover, to explore the generalization for other segmentation tasks, we also extend the proposed network to liver tumor segmentation in CT images of the MICCAI 2017 Liver Tumor Segmentation Challenge. Our implementation is released at https://github.com/infinite-tao/MA-MTLN.**

***Index Terms*— Gastric tumor segmentation, lymph node classification, multi-attention, multi-task learning, CT scans.**

Manuscript received January 19, 2021; accepted February 23, 2021. Date of publication March 1, 2021; date of current version June 1, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 81771922, Grant 62071309, Grant 61801305, Grant 62006160, Grant 81971585, and Grant 61871274; in part by the National Natural Science Foundation of Guangdong Province under Grant 2019A1515111205; in part by the Shenzhen Key Basic Research Project under Grant JCYJ20170818094109846, Grant JCYJ20180507184647636, Grant JCYJ20190808155618806, Grant GJHZ20190822095414576, and Grant JCYJ20190808145011259; and in part by the SZU Medical Young Scientists Program under Grant 71201-000001. *(Corresponding authors: Guolin Ma; Baiying Lei.)*

Yongtao Zhang, Jie Du, and Tianfu Wang are with the Health Science Center, School of Biomedical Engineering, Shenzhen University, Shenzhen 518060, China, also with the National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Shenzhen University, Shenzhen 518060, China, and also with the Marshall Laboratory of Biomedical Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: 13707068133@163.com; dujie@szu.edu.cn; tfwang@szu.edu.cn).

Haimei Li is with the Department of Radiology, Fuxing Hospital, Capital Medical University, Beijing 100038, China (e-mail: 1043652709@qq.com).

Jing Qin is with the Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Hong Kong (e-mail: harry.qin@polyu.edu.hk).

Yue Chen, Bing Liu, Wenwen Gao, and Guolin Ma are with the Department of Radiology, China-Japan Friendship Hospital, Beijing 100029, China (e-mail: 71973292@qq.com; liubing0711@163.com; 1196715172@qq.com; maguolin1007@qq.com).

Baiying Lei is with the Health Science Center, School of Biomedical Engineering, Shenzhen University, Shenzhen 518060, China, also with the National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Shenzhen University, Shenzhen 518060, China, also with the Marshall Laboratory of Biomedical Engineering, Shenzhen University, Shenzhen 518060, China, and also with the AI Research Center for Medical Image Analysis and Diagnosis, Shenzhen University, Shenzhen 518060, China (e-mail: leiby@szu.edu.cn).

Digital Object Identifier 10.1109/TMI.2021.3062902

I. Introduction

GASTRIC cancer is a common cancer of the digestive system and the third leading cause of cancer-related deaths in the world [1]. Surgical resection is currently the main treatment for gastric cancer, and preoperative examination is a reliable basis for surgery [2]. Generally, preoperative examination mainly includes the degree of tumor invasion and LN metastasis. Due to the high imaging density resolution, convenient examination, fast speed and non-invasiveness, computed tomography (CT) scan has become a routine imaging modality for gastric cancer [3]. In clinical practice, physicians diagnose and formulate treatment plans are based on CT reports made by radiologists [4]. However, the manual inspection process of slice by slice is quite time-consuming and relies heavily on the experience of radiologists. Therefore, an effective computer-aided diagnosis (CAD) system in CT images is essential for assisting physicians in selecting reasonable surgical approaches and prognosis plans.

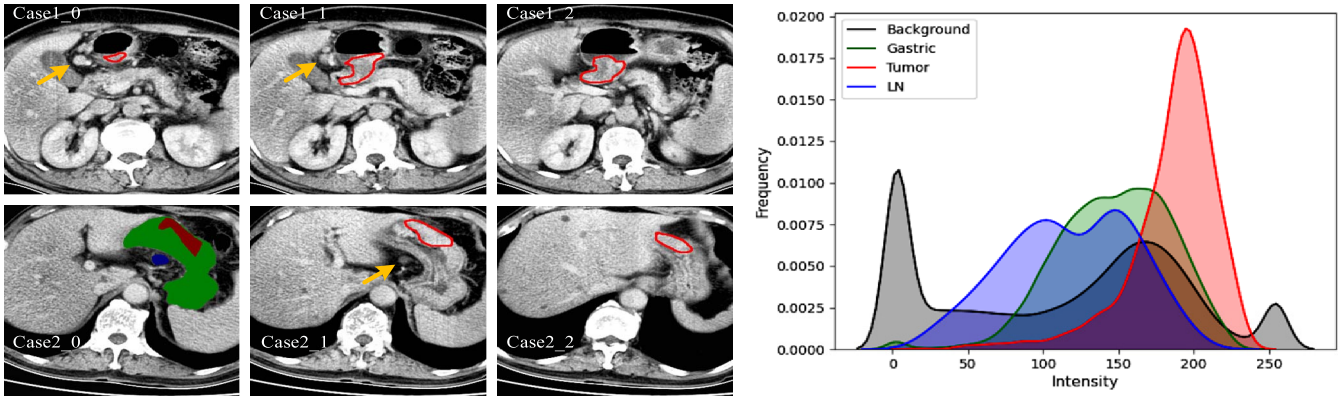

Fig. 1. The major challenges of automatic segmentation of gastric cancer and classification of LN from CT images. The yellow arrow indicates the location of LN metastasis. Case1: LN metastasizes to the pyloric area; Case2: LN metastases to the small curvature of the gastric. In addition, the red line traces the boundary of the tumor. The picture on the right is a deeper expression of Case2_0, showing the intensity distribution of the gastric, tumor, LN and background (adjacent tissue of the gastric). Note: The experimental data does not have label of the gastric, but here is just highlight challenges.

There are two important tasks in building a CAD system of gastric cancer: tumor segmentation and LN classification. The tumor segmentation task is used to achieve tumor location and boundary detection, whereas the LN classification task is used to achieve the diagnosis of negative or positive LN metastasis. LN metastasis is the main route of gastric cancer metastasis and occurs in the drainage direction around the primary tumor [5]. In this regard, tumor segmentation and LN metastasis classification are closely related in clinical practice. Specifically, the location information and morphological characteristics of the tumors can improve the predictive performance of LN metastasis classification [6], [7], whereas the location information of LN metastasis sometimes assists radiologists to locate the primary tumor. Therefore, how to simultaneously boost the performance of automatic segmentation and classification by harnessing the relatedness of these two tasks has also become the focus of our work. Furthermore, accurate tumor segmentation and LN classification in CT images are quite challenging (as shown in Fig. 1) due to the following reasons: (a) there exist many ambiguous/missing boundary caused by low contrast of tumor with gastric and other neighboring tissues, (b) inhomogeneous intensity distribution of tumor and LN in CT images, and (c) tumor and LN have various sizes, shapes and locations in different CT volumes.

In previous studies, many segmentation or classification methods have been proposed for medical image analysis based on traditional machine learning methods, such as graph cuts [8], watershed transform [9], support vector machines [10], bayesian classifer [11], and so on. Most of these methods, however, rely on only hand-crafted features and are incapable of extracting powerful visual representations to address the challenges of our task.

Recently, deep learning with convolutional neural networks (CNNs, e.g., U-Net [12], FPN [13], ResNet [14]) has been attracting widespread attention with respect to medical image computing and has made landmark progress in some abdominal CT image analysis (e.g., liver, kidney, pancreas) [15]–[17]. To solve the ambiguous/missing boundary issue in CT images, Liu *et al.* [18] proposed to construct a graph cut energy function by using the shape information output from an improved U-Net, thereby realizing more accurate boundary location for liver. Yu *et al.* [19] designed a cross network that can simultaneously extract local and global contextual information from both horizontal and vertical directions, to better distinguish the ambiguous boundaries of kidney tumors. To tackle the problem of large tumor size variation, Dou *et al.* [20] employed a 3D fully convolutional network to extract the multi-scale features of the liver, and implemented deep supervision of the multi-scale features to improve the identification ability of the network. Shen *et al.* [21] proposed a multi-scale feature extraction method for LN classification by combining convolutional layers with different kernels and multi-crop pooling layers. In general, deep neural networks (DNNs) can capture multi-scale features containing rich location information and semantic information by convolutional layers with different kernels and pooling layers. However, the multiple down-sampling operation in DNNs inevitably cause the loss of spatial information, which cannot be recovered through the up-sampling operations when generating segmentation results. Moreover, the direct application of simple concatenated multi-scale features at the back end of DNNs tend to include non-target regions, which will have a negative impact on the segmentation results and classification prediction.

To tackle the above issues, we propose a novel 3D multi-attention guided multi-task learning network, which consists of an early-stage scale-aware attention-guided (SA) shared feature learning module to extract the universal features, followed by a task-aware attention-guided (TA) feature learning module to yield task-specific discriminative features. The SA shared feature learning naturally exploits the relatedness between tasks to achieve robust and universal feature representation by a novel bidirectional visual attention (VA) layer and a new adaptive spatial attention (ASA) module. Specifically, the bidirectional VA layer is designed to extract high-level semantic information and rich detailed information with attentive features of different scales. The ASA module can not only use an attention mechanism to focus on useful spatial information at deeper layers, but also selectively leverage

the multi-scale features to refine the single-scale features. These modules are cascaded and can effectively fuse attentive features of different scales by capturing contextual information of targets, which can solve the problem of ambiguous/missing boundary and highly variable shapes of tumor. Meanwhile, two stage-wise deep supervision (SDS) paths are developed to retain rich multi-scale features during model optimization, which is able to increase the sensitivity to target features as well.

The TA feature learning is designed with soft attention masks, which aims at fine-tuning shared features towards the optimal inference of specific tasks. For segmentation task, a segmentation-aware attention module is proposed to further transfer the high-level semantic information of small-scale features to large-scale features. For classification task, a classification-aware attention module is proposed to further transfer the rich detail information of large-scale features to small-scale features. Therefore, these TA modules allow specific tasks to extract corresponding discriminative features from the shared features. Multi-task learning loss is devised to use uncertain weight to balance the performance of segmentation and classification tasks. Experiments on our self-collected CT images dataset demonstrate that our method has achieved a promising performance, outperforming other state-of-the-art methods. In summary, this work has three main contributions:

- We develop a novel end-to-end 3D multi-attention guided multi-task learning network for gastric cancer segmentation and LN classification by integrating global context and spatial information from both small-scale and large-scale features.
- SA shared feature learning is proposed, which first performs weighted cross-scale feature fusion, and then selectively leverages the multi-scale features integrated from different scales to refine the features at each scale.
- TA feature learning is proposed to effectively use the soft attention mask, which combines segmentation and classification losses with uncertain weight for accurate inference of specific tasks.

II. Related Work

A. Attention Mechanism

The attention mechanism [22] effectively solves the problem that the operation of convolution cannot highlight target features and suppress noise features in the hidden layer. Hence, the attention mechanism has become a hot topic. For example, a spatial attention was introduced [23] with the non- local blocks to compute the response at a position as a weighted sum of the features at all positions. A channel attention mechanism [24] was proposed via the squeeze and excitation block to recalibrate the inter-channel dependencies on different channels, which can effectively learn global context features. Later, the concurrent spatial and channel 'squeeze and excitation' was designed for better whole brain segmentation on MRI scans and organ segmentation on CT scans. Meanwhile, a convolutional block attention module (CBAM) [25] was proposed to coordinate the combined effect of channel and spatial attention mechanism. Since then, the spatial and channel attention mechanism is used to recalibrate fully convolutional networks and achieves good performance on multiple medical datasets. Moreover, there are also some works [26], [27] that only use a simple attention model formed by the soft attention mask. In [26], one attention module was proposed to get attentive features for prostate segmentation in 3D transrectal ultrasound, and another attention module [27] was proposed for semantic segmentation and depth estimation in natural images.

B. Multi-Task Learning

Multi-task learning is a paradigm of machine learning and has been successful in many applications [28]–[30]. Since annotated data in medical image analysis is often difficult to collect, it is a good solution when there are multiple related tasks. Therefore, its purpose is to leverage the effective information contained in multiple related tasks to help improve the generalization performance of all tasks. To the best of our knowledge, more and more related studies have shown that multi-task learning can comprehensively analyze medical images. Namburete *et al.* [31] proposed a multi-task FCN architecture to address the problem of 3D fetal brain localization, structural segmentation, and alignment to a referential coordinate system. Chakravarty and Sivswamy [32] employed a multi-task CNN to achieve segmentation and image-level classification for glaucoma by combining image appearance and structural features. Zhou *et al.* [33] designed a multi-task learning network ($C_{MS}VNet_{Iter}$) for joint segmentation and classification of tumors in ABUS images. Chen *et al.* [34] performed three schemes (including three existing feature extraction methods) to extract different types of features for describing the semantic features of the nine lung nodules in CT images. Apart from the architecture design of the network, researchers always pay attention to how to balance the task by improving the multi-task learning loss. Therefore, we introduce uncertain weights [35] at the back end of TA learning to achieve equal learning of all tasks and avoid easier tasks to dominate.

III. Methodology

A. Overview

Fig. 2 shows the proposed 3D multi-attention guided multi-task learning network, which includes three collaborative parts, i.e., the backbone, SA shared feature learning, TA feature learning. The proposed 3D network takes an abdominal CT volume as input and starts with SENet-50 [24] as backbone to acquire coarse multi-scale features at different resolutions. The small-scale feature maps have low-resolution but with high-level semantic information, while the large-scale feature maps have the high resolution but with rich detailed information. Therefore, multi-scale features can effectively deal with tumors and LNs of different sizes. Considering that 3D task consumes lots of computer memory, we have not refined the feature maps at *Scale*0. In addition, based on our dataset characteristics, each volume contains quite few target slices. Therefore, we set down-sampling of *Scale*0,

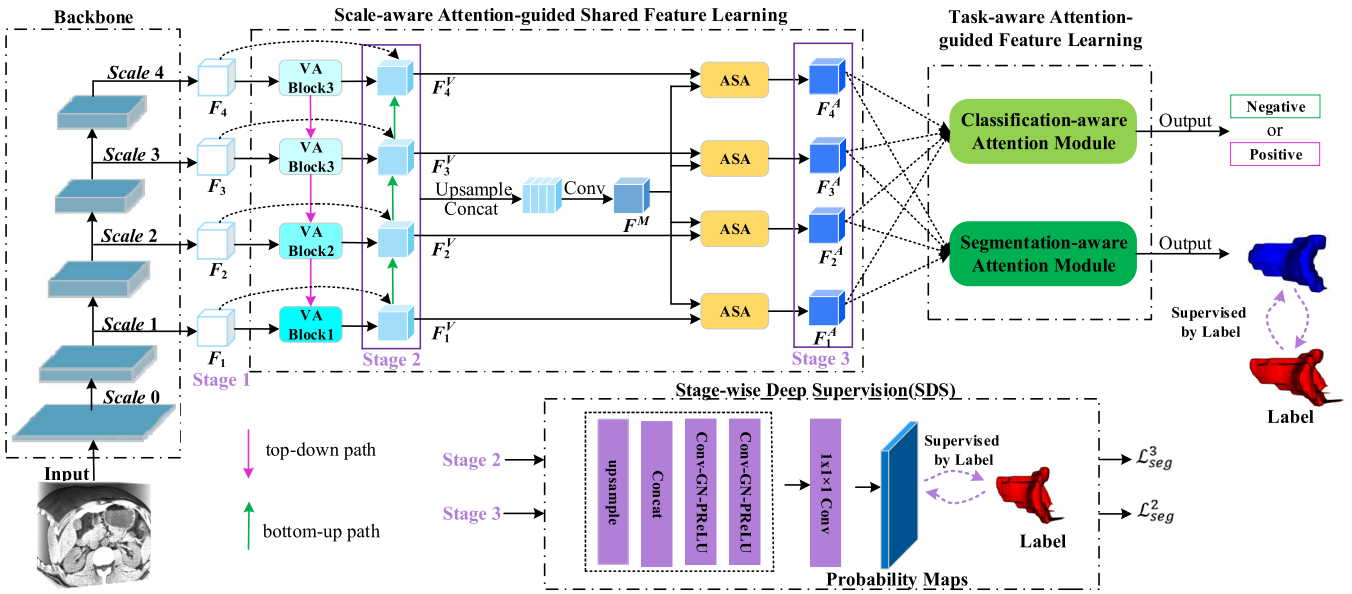

Fig. 2. The framework of the proposed network, which includes scale-aware attention-guided shared feature learning, two SDS paths, and task-aware attention-guided feature learning. VA: visual attention; ASA: adaptive spatial attention. F_1, F_2, F_3 *and* F_4 represent coarse multi-scale features; F_1^V, F_2^V, F_3^V *and* F_4^V mean multi-scale feature with high feature consistency; F_1^A, F_2^A, F_3^A *and* F_4^A indicate refined and universal multi-scale features. Note that the purple and green arrows represent the bidirectional feature fusion path.

*Scale*1 and *Scale*2 by stride (1, 2, 2). Meanwhile, as the network layer deepen, the inconsistency of feature maps at different scales becomes more and more obvious. We use the dilated convolution between *Scale*3 and *Scale*4 to aggregate multi-scale semantic information and get feature maps with the same resolution. Next, the coarse multi-scale features are used as the input of SA shared feature learning, which can retain the detailed information of targets of different sizes and suppress the noise information introduced by the background area. Finally, the refined and attentive multi-scale features are taken as the input of TA task-specific feature learning, which forces the 3D network to learn the discriminative region-wise features for each task.

B. Scale-Aware Attention-Guided Shared Feature Learning

To learn coarse-to-fine features including effective context information, regional semantic and boundary information, we propose the SA shared feature learning with two types of scale-aware attention modules: VA and ASA modules.

The goal of VA blocks is to fully explore the local and global 3D spatial information in different scale features. VA block is inspired by the RFB [36], which is designed to obtain global information through the seamless combination of convolution, dilated convolution. Different from the RFB, we not only introduce attention mechanism, but also set up different branches according to different scales to further promote the visual receptive field. In the VA block, the convolution layers represent the center of vision and the dilated convolution layers are explored to control their eccentricities, which mimic the receptive field structure of the human vision system. Then, these VA blocks further formed the bidirectional VA layer inspired by the theory of human visual perception [37]. Specifically, the human visual perception process is based on an important theory (feature integration theory) and two visual attention mechanisms (bottom-up and top-down) [38]. In the early stage of visual perception, the retina processes various input features in parallel, and there is no visual attention mechanism in this period. In the later stage of visual perception, different features are gradually integrated in the participation of visual attention mechanism, and finally effective features are obtained. In general, we design these VA blocks, which form multi-branch attentive feature pools with different receptive fields. Then, we embed top-down and bottom-up feature fusion strategies to effectively integrate context information at low levels and regional semantic at high levels.

We show VA block1 (see Fig. 3a), which includes all the details of the multi-branch feature pool. Let the coarse multi-scale features generated by the backbone be $\boldsymbol{F}_s$, the refined features be $\boldsymbol{F}_s^V$, $s\epsilon\{1,2,3,4\}$, we formulate the visual perception process of the bidirectional VA layer in different scale features as follows:

$$\boldsymbol{F}_s^{im} = \begin{cases} f_s(\boldsymbol{F}_s), & s=4 \\ f_s\left(\frac{\lambda_1\cdot\boldsymbol{F}_s+\lambda_2\cdot Upsam(\boldsymbol{F}_{s+1}^{im})}{\lambda_1+\lambda_2}\right), & s\neq 4 \end{cases} \tag{1}$$

$$\boldsymbol{F}_s^{V} = \begin{cases} \frac{\lambda_1'\cdot\boldsymbol{F}_s+\lambda_2'\cdot\boldsymbol{F}_s^{im}}{\lambda_1'+\lambda_2'}, & s=1 \\ \frac{\lambda_1'\cdot\boldsymbol{F}_s+\lambda_2'\cdot\boldsymbol{F}_s^{im}+\lambda_3'\cdot Downsam(\boldsymbol{F}_{s-1}^{V})}{\lambda_1'+\lambda_2'+\lambda_3'}, & s\neq 1 \end{cases} \tag{2}$$

where $\boldsymbol{F}_s^{im}$ is the intermediate feature obtained through the top-down attention mechanism at *Scale* s, f_s denotes s-th VA block with soft attention. $\lambda_i \geq 0$ and $\lambda_i' \geq 0$ represent weights obtained by applying a rectified linear unit, $Upsam$ and $Downsam$ represent the up-sampling and down-sampling operation, respectively. All feature fusion processes use the above normalization function instead of the normal softmax layer for the fast feature fusion [39]. In addition, to reduce the

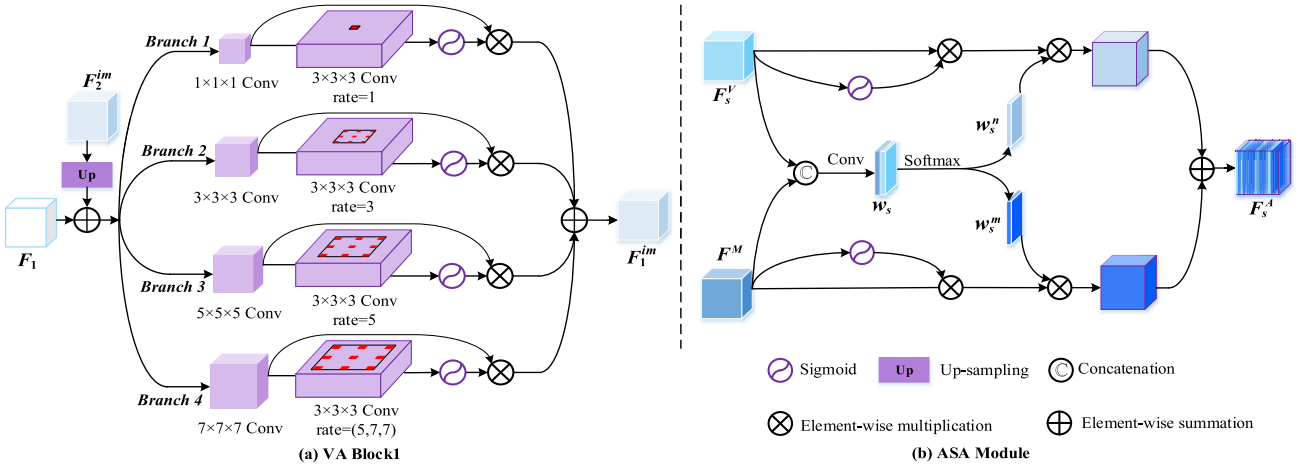

Fig. 3. Two modules of SA shared feature learning. (a) The VA block1 with four branches, which can fully obtain global information. Note that the VA block2 and VA block3 have three branches (*Branch 1*, *Branch 2 and Branch 3*) and two branches (*Branch 1 and Branch 2*), respectively. (b) The ASA module.

parameters of VA block, we choose two and three $3 \times 3 \times 3$ convolutions instead of $5 \times 5 \times 5$ and $7 \times 7 \times 7$ convolution, respectively. Our bidirectional VA layer extracts the refined features with high feature invariance.

Although we have obtained features with high feature consistency, the high-level semantic information in small-scale features and the rich detailed information in large-scale features will inevitably be interfered by noise from non-target regions while transferring each other. Therefore, we design another space-related module to further highlight the feature of tumors and LNs. This novel module is referred as the ASA module as shown in Fig. 3b. Firstly, we enlarge each single scale feature to the same size of *Scale*1's feature map by trilinear interpolation, and then generate the multi-scale features by channel concatenation and convolution operation. Secondly, we use the f_{op} operation to match the feature size of multi-scale feature and each single scale feature, which generates two corresponding soft attention feature maps by applying sigmoid function in parallel. Meanwhile, the module re-fusion multi-scale feature (defined as $\boldsymbol{F}^M$) and each single-scale feature (defined as $\boldsymbol{F}_s^V$) to learn adaptive spatial weight factors $\boldsymbol{w}_s^m$ and $\boldsymbol{w}_s^n$ by using $1 \times 1 \times 1$ convolutional layers and softmax function. Finally, the spatial weight factors and soft attention feature maps are element-wise multiplied to obtain the deep attention-guided coarse-to-fine features. Then the two types of features are element-wise summarized to get the refined features defined as $\boldsymbol{F}_s^A$.

$$\boldsymbol{w}_s = \mathbb{C}\left(c\left(\boldsymbol{F}_s^V\right); c\left(f_{op}\left(\boldsymbol{F}^M\right)\right)\right) \tag{3}$$

$$f_{op} = \begin{cases} \mathcal{I}, & s = 1 \\ Downsam, & s \neq 1 \end{cases} \tag{4}$$

$$\boldsymbol{w}_s^n = \frac{e^{c(\boldsymbol{F}_s^V)}}{e^{c(\boldsymbol{F}_s^V)} + e^{c(f_{op}(\boldsymbol{F}^M))}}, \quad \boldsymbol{w}_s^m = \frac{e^{c(f_{op}(\boldsymbol{F}^M))}}{e^{c(\boldsymbol{F}_s^V)} + e^{c(f_{op}(\boldsymbol{F}^M))}}, \tag{5}$$

$$\boldsymbol{F}_s^A = \boldsymbol{w}_s^n \cdot \boldsymbol{F}_s^V \oplus \sigma\left(\boldsymbol{F}_s^V\right) \oplus \boldsymbol{w}_s^m \cdot f_{op}\left(\boldsymbol{F}^M\right) \oplus \sigma\left(f_{op}\left(\boldsymbol{F}^M\right)\right) \tag{6}$$

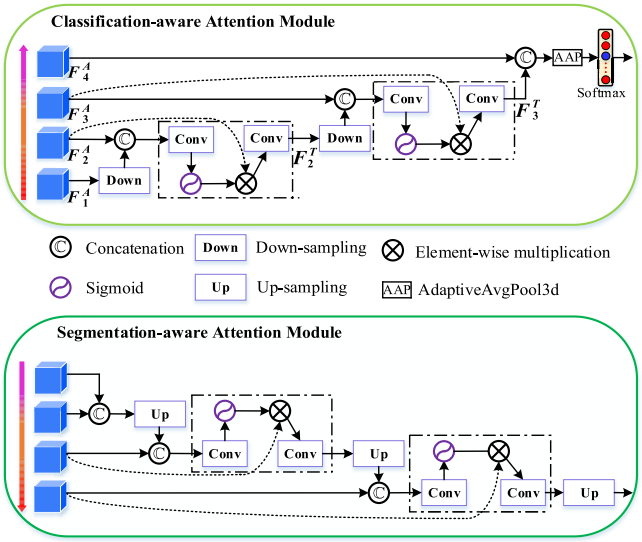

Fig. 4. Two attention modules of task-aware attention-guided learning.

where c and $\mathbb{C}$ denote a $1 \times 1 \times 1$ convolution and channel concatenation operation; $\oplus$ and $\oplus$ mean element-wise product and element-wise summation, $\mathcal{I}$ and σ represent the identity function and sigmoid function, respectively.

Furthermore, two SDS paths are introduced in the network, which guides the acquisition of refined features in the SA shared feature learning. To facilitate implementation, SDS only supervises the segmentation target.

C. Task-Aware Attention-Guided Feature Learning

To make better use of these refined multi-scale features with high feature consistency generated from the SA shared feature learning, we propose the TA feature learning for specific tasks. Unlike other multi-task learning networks [31]–[33] that only extract the features of the middle layer for classification tasks, we use all shared features for segmentation and classification tasks. Fig.4 shows the TA feature learning, which includes two task-aware attention modules in an encoding and decoding path respectively.

1) Classification-Aware Attention Module: The module obtains more accurate classification probabilities by aggregating adjacent scale features in a coding path instead of directly performing multi-branch prediction on refined features in previous works [40]–[42]. Specifically, we first concatenate the features between adjacent scales in the bottom-up direction, and then apply the soft attention mask of the fused features to guide the expression of the LN features at the current scale. Let the refined features be $\boldsymbol{F}_s^A$, $s \in \{1, 2, 3, 4\}$ and the features guided by the attention mechanism $\boldsymbol{F}_s^T$ when $s \in \{2, 3\}$ is expressed as:

$$\boldsymbol{F}_2^T = f_r\left(f_\sigma\left(f_{r'}\left(\mathbb{C}\left(\boldsymbol{F}_2^A; Downsam\left(\boldsymbol{F}_1^A\right)\right)\right)\right) \oplus \boldsymbol{F}_2^A\right) \tag{7}$$

$$\boldsymbol{F}_3^T = f_r\left(f_\sigma\left(f_{r'}\left(\mathbb{C}\left(\boldsymbol{F}_3^A; Downsam\left(\boldsymbol{F}_2^T\right)\right)\right)\right) \oplus \boldsymbol{F}_3^A\right) \tag{8}$$

where $\mathbb{C}$ denotes concatenation operation; $\oplus$ represents the element-wise multiplication; f_r, f_σ and $f_{r'}$ are convolutional layers with group normalization and a non-linear activation function. $f_{r'}$ and f_σ have a convolution with the $1 \times 1 \times 1$ kernel, and obtain the attention mask of the current scale through the sigmoid activation of f_σ. In addition, f_r has a convolution with the $3 \times 3 \times 3$ kernel, which is used to generate features with rich detailed information and pass it to the next attention block until the final prediction of the category.

2) Segmentation-Aware Attention Module: The module obtains more effective prediction of tumor volumes by aggregating adjacent scale features in a decoding path. The only difference between the specific implementation of this module and the classification-aware attention module is the opposite operation path. Specifically, the features between two adjacent scales are concatenated in the top-down direction, and a soft attention mask is used to guide the effective expression of lesion features on different scales. Finally, the high-level semantic information of small-scale features is transferred to large-scale features by the module.

D. Multi-Task Learning Loss

At present, the common multi-task learning method is to optimize the linear weighted loss of all tasks and solve the minimum value. However, this usually cannot find the optimal solution, and the weight parameter setting is cumbersome. In our work, we design the corresponding loss functions for different tasks. On the one hand, a hybrid loss function is designed for challenging segmentation task, which consists of a weighted sum of two functions. The first loss function is the Jaccard loss directly aimed at optimizing the evaluation metric of the segmentation performance, which is defined as:

$$\mathcal{L}_{Jaccard} = 1 - \frac{\sum_{i=1}^{n} q_i \cdot p_i}{\sum_{i=1}^{n} q_i^2 + \sum_{i=1}^{n} p_i^2 - \sum_{i=1}^{n} q_i \cdot p_i} \tag{9}$$

where n is the voxel number of the input CT volume; $p_i \in [0, 1]$ denotes the prediction probability of i-th voxel and $q_i \in \{0, 1\}$ denotes the voxel value of the corresponding ground truth. The second loss function is the Focal loss, which is improved by log loss to solve the problem of positive and negative sample imbalance. It is defined as:

$$\mathcal{L}_{Focal} = -\frac{1}{n}\sum_{i=1}^{n}\left[\alpha q_i \left(1 - p_i\right)^\gamma \log p_i + \left(1 - \alpha\right)\left(1 - q_i\right) p_i^\gamma log\left(1 - p_i\right)\right] \tag{10}$$

where α denotes a balance factor of the Focal loss and is set as 0.2; γ represents a focusing parameter to smoothly adjust the weighting rate and set to 1. In our segmentation task, we use $\mathcal{L}_{Focal}$ to guide the model segmentation of small target regions. Therefore, the segmentation loss function is denoted as:

$$\mathcal{L}_{seg} = \mathcal{L}_{Jaccard} + \eta \cdot \mathcal{L}_{Focal} \tag{11}$$

where η represents the weight factor of Focal loss and η is set as 0.1. Since the two SDS paths act on the segmentation target, the final segmentation loss function is defined as:

$$\mathcal{L}_{SDS} = \sum_{s=2}^{s=3} \mu^s \cdot \mathcal{L}_{seg}^s + \mu^k \cdot \mathcal{L}_{seg}^k \tag{12}$$

where μ^s and $\mathcal{L}_{seg}^s$ represent the weight and loss of s-th stage, respectively, μ^k and $\mathcal{L}_{seg}^k$ are the weight and loss for the segmentation prediction. We empirically set the weights $\left\{\mu^2, \mu^3, \mu^k\right\}$ as $\{0.8, 0.9, 1.0\}$. On the other hand, we choose cross-entropy loss as the loss function ($\mathcal{L}_{class}$) of the classification task.

$$L_{class} = -\frac{1}{N}\sum_{j=1}^{N}\left[q_j \log p_j + \left(1 - q_j\right)\log\left(1 - p_j\right)\right] \tag{13}$$

where N is the number of the input CT volume; $p_j \in [0, 1]$ denotes the prediction probability of j-th volume and $q_j \in \{0, 1\}$ denotes the corresponding ground truth of the volume.

To mitigate the negative impact of segmentation task on classification task, we use uncertainty to weigh losses of segmentation and classification tasks. In fact, this loss design scheme [35] is an application of adaptive strategy. Finally, our multi-task loss ($\mathcal{L}_{MTL}$) function is defined as follows:

$$\mathcal{L}_{MTL} = \frac{1}{2\alpha_1^2} \cdot \mathcal{L}_{SDS} + \frac{1}{2\alpha_2^2} \cdot \mathcal{L}_{class} + \log \alpha_1 \alpha_2 \tag{14}$$

where α_1 and α_2 indicate uncertain weights and are obtained through network learning. In practical applications, the parameters α_1 and α_2 are first initialized as two tensors with value of 1, and then they are iteratively updated during training phase.

IV. Datasets

In this study, we use an in-house images dataset and a publicly available dataset, which are all obtained by CT scans.

A. Our Dataset

The dataset is collected from the three medical centers (Taiyuan People Hospital, China; Xian People Hospital, China; Department of Radiology, China-Japan Friendship Hospital, Beijing, China) by three kinds of medical instruments (Toshiba 320-slice CT, SOMATOM 64-slice CT and Philips 128-slice CT), with a largely varying in-plane resolution from 0.5 mm to

TABLE I
ABLATION ANALYSIS OF DIFFERENT COMPONENTS IN THE PROPOSED 3D NETWORK (MEAN±SD. BASELINE:3D SENET-50)

Method	Segmentation		Classification					
	DSC(%)	JI(%)	Acc(%)	AUC(%)	Pre(%)	Recall(%)	Spec(%)	F1(%)
Baseline	60.1±3.2	43.1±3.2	75.6±3.9	79.4±4.8	77.9±4.7	88.2±6.1	47.5±8.6	83.2±3.2
Baseline+VA	61.0±2.7	43.9±2.7	80.1±6.1	81.9±8.6	81.8±3.5	91.0±8.1	57.0±10.0	**86.5±4.5**
Baseline+ASA	60.8±3.0	43.8±3.1	78.9±6.7	80.5±7.5	80.2±5.4	91.1±7.4	52.5±11.0	85.5±4.6
Baseline+TA	60.4±2.5	43.3±2.5	79.4±5.5	77.0±10.8	78.8±6.6	**95.6±3.4**	44.2±15.1	85.8±3.0
Basline+VA+ASA	61.7±2.1	44.6±2.2	80.8±5.6	83.2±8.7	**82.5±3.4**	88.7±7.6	60.2±18.2	86.2±4.2
Basline+VA+ASA+TA	62.1±1.9	45.1±2.0	**81.2±4.2**	82.7±7.6	82.3±2.8	88.9±12.0	60.1±18.7	86.1±5.4
Ours	**62.7±2.5**	**45.5±2.0**	80.5±5.3	**86.0±9.2**	82.1±5.2	88.8±4.2	**61.6±19.1**	86.2±2.8

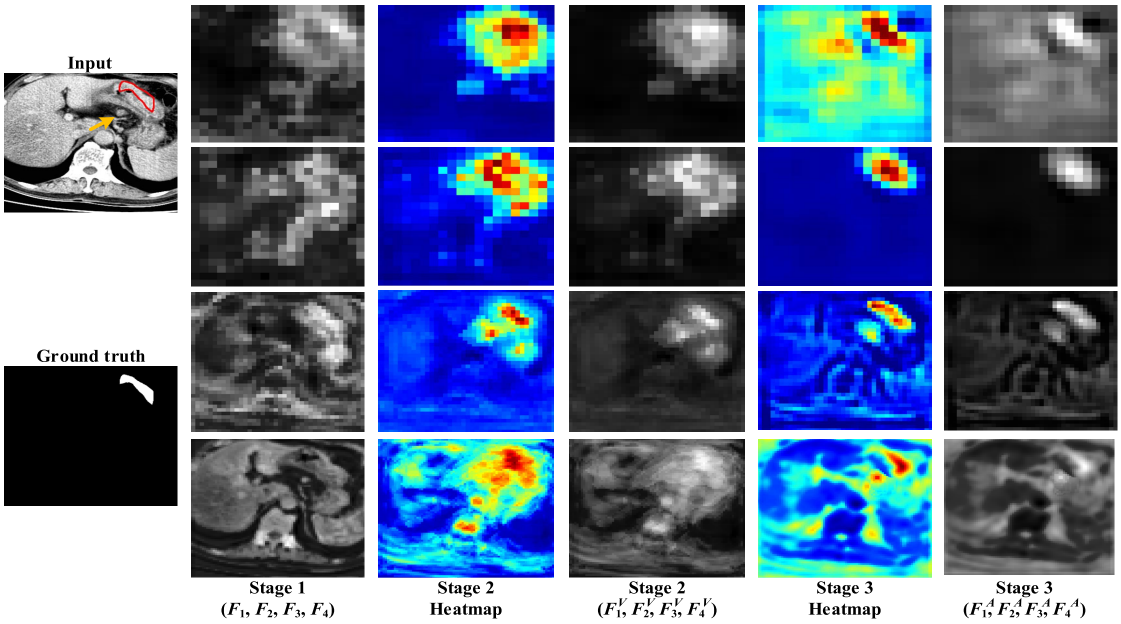

Fig. 5. 2D axial views of feature maps from different stages of the proposed network.

1.0 mm and slice spacing from 5.0 mm to 8.0 mm. Our dataset contains 160 CT image samples (160 ordinary CT volumes and 63 enhanced CT volumes), which are collected from 2015 to 2018. The ground truth of segmentation is annotated by three experienced radiologists using the software ITK-SNAP based on surgical pathology. The classification label comes from the statistical data of clinical information provided by physicians, of which 52 samples were negative LN metastases, while the other 108 samples were positive LN metastases. Note that the dataset used has passed the ethical review of the relevant hospitals and obtained the informed consent of the patients.

B. LiTS Dataset

The MICCAI 2017 Liver Tumor Segmentation (LiTS) Challenge [43] dataset totally has 201 enhanced abdominal CT scans, which is further split to a training set with 131 scans and a test set with 70 scans. The dataset was collected from six different clinical sites by different scanners and protocols, and the organizer only discloses the annotations of the training data and keeps the annotations of the test data confidential.

V. EXPERIMENTS AND RESULTS

A. Implementation Details

1) Data Preprocessing: Because the tumor region is smaller than the background area and to cope with the limitation of 3D data on computer memory consumption, we cut and resample each volume to patches include $24 \times 256 \times 256$ voxels with a voxel size of $5.0 \times 0.741 \times 0.741$ mm^3 or $8.0 \times 0.741 \times 0.741$ mm^3 on our dataset. Moreover, we cut each volume to patches with size $128 \times 128 \times 128$ on the LiTS dataset. For training, we not only use data augmentation (e.g., flipping, rotation, translation), but also perform CT image normalization (automatic clipping operation from 0.5% to 99.5% intensity value of all foreground voxels) and voxel space resampling (with third order spline interpolation).

2) Training Process: The proposed model is implemented on the PyTorch platform and is trained with 1× NVIDIA GeForce TITAN Xp GPU (12GB) using a five-fold group cross- validation strategy. The proposed model and other models utilize an Adam optimizer with an initial learning rate of 2×10^{-4} and a weight decay of 2×10^{-5}. Besides, we set batch size as 2, and training is stopped when the learning rate drops below 10^{-9} or 1000 epochs are exceeded.

B. Evaluation Metrics

In our multi-task learning, we employed six performance metrics to quantitatively evaluate the obtained tumor segmentation results, which include the Dice similarity coefficient (DSC), Jaccard index (JI), Precision (Pre), Recall, Average surface distance (ASD, in voxel) and 95% Hausdorff distance (95HD, in voxel). Note that we only used the first

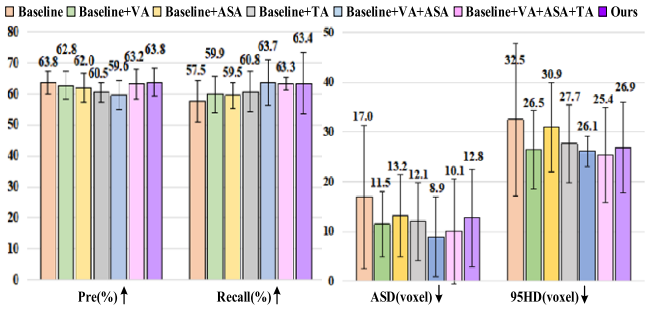

Fig. 6. Automatic segmentation results of different components in the proposed 3D network.

four performance metrics to quantify the results obtained by 2D segmentation models. We employed six performance metrics to quantitatively evaluate the obtained LN classification results, which include Accuracy (Acc), Area under curve (AUC), Precision (Pre), Recall, Specificity (Spec) and F1-score (F1).

We employed two performance metrics to quantitatively evaluate the obtained liver and tumor segmentation results, which include DSC and average symmetric surface distance (ASSD) suggested by the LiTS challenge.

C. Experimental Methods

To demonstrate the effectiveness of the proposed method for simultaneous gastric cancer segmentation and LN classification, we conducted several groups of ablation experiments.

1) Effectiveness of the Different Components: For fair comparison, our baseline consists of the backbone, an encoding branch and a decoding branch. Unlike TA modules, these two branches do not use the soft attention mask. We separately add bidirectional VA layer, ASA module, TA modules and SDS to the baseline, and aggregate different attention modules onto the baseline.

2) Influence of Parameters: In the designed segmentation loss, there are three important hyper-parameters, i.e., γ, α and η. To investigate the impact of their settings on the multi-task learning, we explored to set γ to 0.2, 0.5, 1.0 and 2.0, and set α to 0.2, 0.5 and 0.75, and set η to 0.01 and 0.1.

3) Effectiveness of Multi-Task Learning: To explore the mutual boost between the specific tasks, we not only conducted experiments in a single task, but also compared weight uncertainty and equal weight loss strategy in multi-task learning.

To further verify the effectiveness of the proposed method, we compared it with other state-of-the-art multi-task learning networks (see Section II.*B*) including multi-task FCN [31], multi-task CNN [32] and $C_{MS}VNet_{Iter}$ [33]. Specifically, multi-task FCN implemented an encoding path to extract shared features of correlated tasks, and then branching out into task-specific output streams. Multi-task CNN employed a U-Net architecture for segmentation, and down-sampled the segmentation map to the same size as the feature map of the last down-sampling layer in the original encoding path to extract features for classification. $C_{MS}VNet_{Iter}$ performed V-Net [44] as the backbone network, which implemented encoding-decoding paths for segmentation by stacking convolutional layers at each stage, and the high-level features obtained in the middle three stages are used for classification. For fair comparison, we adopted the basic blocks from SENet-50 in all comparison methods. Moreover, we extended all 2D operations in Multi-task CNN to 3D operations, and changed multi-class classification in multi-task FCN to two-class classification.

In addition to the above comparison methods, we also compared our multi-task network with other single segmentation or classification networks. For segmentation task, we compare the proposed method with other five state-of-the-art segmentation networks: 3D U-Net [12], AnatomyNet [45], V-Net [44], 3D FPN [13], and DAF3D [26]. Most of the existing segmentation methods are based on the U-Net, which is the first network used in biomedical image segmentation. The AnatomyNet is a research work on small target segmentation, and it has achieved good results in the segmentation of multiple small organs. The V-Net is designed to solve the 3D volume segmentation, and has been applied to many CT image data segmentation work [46], [47]. Moreover, FPN and DAF3D are the work of origin and improvement of multi-scale features, respectively. At the same time, to further explore the effect of slice thickness on the segmentation results, we also conducted a set of 2D segmentation experiments. We replace the 3D operations in the proposed 3D segmentation model with 2D operations, and compare the proposed 2D segmentation model with other five state-of-the-art segmentation networks: 2D U-Net [12], SegNet [48], FCN [49], 2D FPN [13] and DeeplabV3+ [50]. The SegNet employed an encoder-decoder, which encourage decoder use the max-pooling indices received from the encoder to perform non-linear up-sampling of their input feature maps. The FCN is the pioneering work of pixel-by-pixel prediction. The DeeplabV3+ implemented the fusion of spatial pyramid pooling (SPP) and encoder-decoder, and applying the atrous separable convolution to both the SPP and decoder modules.

For classification task, we compare the proposed 3D multi-task learning network with other six state-of-the-art classification networks: 3D VGG16 [51], 3D ResNet [14], 3D InceptionV4 [52], 3D ResNeXt [53], 3D SENet [24], 3D DenseNet [54]. The above models have received extensive attention in classification tasks, and more and more medical image processing works choose them as baselines. In addition, we perform comparative experiments on the depth of the network, such as 3D SENet-50 and 3D SENet-101.

To make a fair comparison, we not only use the same data preprocessing for all experiments, but also retrain all compared models using unified implementation and adjust training parameters to obtain the best performance.

D. Ablation Study

1) Effectiveness of the Different Components: As shown in Table I and Fig. 6, we can observe that the designed three modules greatly improve the performance of multi-task learning network compared with the baseline. In the segmentation task, compared with the baseline, the proposed bidirectional

TABLE II
ABLATION ANALYSIS OF SOME MULTI-TASK LOSS PARAMETER SETTINGS IN THE PROPOSED 3D NETWORK (MEAN±SD)

Parameters			Segmentation				Classification					
γ	α	η	DSC(%)	JI(%)	ASD(voxel)	95HD(voxel)	Acc(%)	AUC(%)	Pre(%)	Recall(%)	Spec(%)	F1(%)
0.2	0.75	0.01	54.0±3.5	37.0±3.2	19.7±15.4	34.1±16. 0	79.8±2.9	80.2±8.6	82.1±2.6	89.2±6.6	56.8±13.4	86.0±2.3
0.5	0.5	0.01	59.2±3.6	42.1±3.6	20.4±15.4	34.1±16.5	78.9±6.6	84.3±11.7	81.1±4.2	88.4±13.2	58.0±12.6	84.8±5.8
1.0	0.2	0.01	62.3±3.5	45.3±3.6	13.0±10.5	26.2±11.8	79.8±7.9	85.1±10.5	82.9±5.5	87.2±8.9	63.3±15.4	85.6±5.1
2.0	0.2	0.01	61.5±2.6	44.4±2.7	11.9±6.7	25.3±6.8	80.7±5.2	84.2±8.5	81.6±5.3	91.7±7.7	56.3±15.1	86.6±3.9
0.2	0.75	0.1	54.7±2.6	37.7±2.5	29.8±13.6	46.2±12.5	82.1±3.7	85.7±7.1	82.4±3.4	93.0±4.4	57.3±10.6	87.8±2.2
0.5	0.5	0.1	58.8±3.3	41.7±3.3	23.0±18.2	37.0±19.3	76.1±3.6	83.4±10.8	80.1±3.7	84.4±9.3	57.8±11.7	82.8±2.7
1.0	0.2	0.1	62.7±2.5	45.5±2.0	12.8±9.8	26.9±9.1	80.5±5.3	86.0±9.2	82.1±5.2	88.8±4.2	61.6±19.1	86.2±2.8
2.0	0.2	0.1	61.3±2.8	44.2±2.9	12.4±7.8	26.4±8.9	79.8±4.3	82.0±8.6	81.9±2.6	88.2±9.2	59.5±16.6	85.6±3.5

TABLE III
ABLATION ANALYSIS OF SINGLE-TASK AND MULTI-TASK IN THE PROPOSED 3D NETWORK (MEAN±SD)

Method	Segmentation				Classification					
	DSC(%)	JI(%)	ASD(voxel)	95HD(voxel)	Acc(%)	AUC(%)	Pre(%)	Recall(%)	Spec(%)	F1(%)
Single-task	62.6±2.0	45.6±2.1	**12.5±9.5**	27.9±9.7	-	-	-	-	-	-
Single-task	-	-	-	-	81.2±2.9	75.5±13.4	79.5±1.9	**97.3±2.9**	40.5±27.2	87.7±1.3
Multi-task(Equal weights)	62.7±2.5	45.5±2.0	12.8±9.8	**26.9±9.1**	80.5±5.3	86.0±9.2	82.1±5.2	88.8±4.2	**61.6±19.1**	86.2±2.8
Multi-task(Uncertain weights)	**63.0±2.7**	**45.9±2.9**	14.5±16.5	28.5±14.9	**82.3±4.5**	**86.5±9.9**	**83.0±4.1**	92.2±4.7	59.5±16.7	**87.8±2.2**

VA layer, ASA module, TA modules and SDS can achieve the numerical gain using all evaluation metrics. The comparison experiments by adding different module combinations show that the designed different modules have improved segmentation and classification performance, and two SDS paths also further improve segmentation performance. Finally, the proposed 3D multi-task learning network achieves quite promising performance (DSC: 62.7%, JI: 45.5%, ASD: 12.8 voxels, 95HD: 26.9 voxels) on the CT dataset, and outperforms the baseline by 4.2%, 5.8%, 26.5% and 17.2% in DSC, JI, ASD and 95HD, respectively. This shows that these modules can enhance tumor feature learning, and they interact better about the lesion. In the classification task, the proposed VA, ASA and TA modules also greatly improve performance. Finally, our method yields the mean Acc value of 80.5%, AUC of 86.0%, Pre of 82.1%, Spec of 88.8%, and F1 of 86.2%, respectively. Our network quipped with all components achieves the best performance in most metrics, which demonstrates the ability of these proposed modules to learn the features of LNs. To analyze the feature learning ability of different modules in the proposed network, we visualize and resize 2D axial view feature maps to the original input size from different stages in the proposed network, as shown in Fig. 5. We can observe that we proposed scale-aware attention modules (VA and ASA) completes better feature representation by aggregating detailed information in large-scale features and semantic information in small-scale features. In general, these proposed modules can work simultaneously in segmentation and classification tasks.

2) Effects of Multi-Task Loss Parameter Settings: As shown in Table II, we explored some important multi-task loss parameter (e.g., γ, α, η) settings in detail. To investigate the impact of their settings on the multi-task, we attempted to set γ to 0.2, 0.5, 1.0, and 2.0, and set α to 0.2, 0.5, and 0.75, and set η to 0.01, and 0.1. In Table II, it is clearly that the proposed method achieves the highest DSC on segmentation task when setting γ is set to 1.0, α is set to 0.2, and η is set to 0.1. Although the proposed method achieves the highest ACC on classification task when setting γ is set to 0.2, α is set to 0.75, and η is set to 0.1, the performance of the experiment on segmentation task is very poor. Hence, we empirically set γ to 1.0, α to 0.2, and η to 0.1 for the proposed method.

3) Effectiveness of Multi-Task Learning Network: As shown in Table III, we confirmed the superiority of our multi-task network compared to our single-task network. Meanwhile, we compare the results of the proposed multi-task learning model using the uncertain weights against the equal weights method. Results show that the proposed model can further improve the performance of our classification and segmentation tasks, and the uncertain weights help to balance the segmentation and classification tasks, which avoids the segmentation task being dominated the training of the proposed method. In general, the proposed method has a high performance gain in classification task. We believe that it is mainly due to the characteristics and positional correlation between tumors and LNs, which enhances the classification of negative LNs in the proposed multi-task model. However, compared with the single-task segmentation network, the multi-task learning network does not gain a large performance gain in the segmentation task. The reason may be that it is difficult to promote this pixel-level classification task only based on the positional correlation between tumors and the LNs.

E. Comparison With Other Multi-Task Methods

We compared the proposed method with other three multi-task methods, which include multi-task FCN, multi-task CNN and $C_{MS}VNet_{Iter}$. Table IV displays the results of our method and other three multi-task methods on our dataset. It shows that our model achieves the best segmentation and classification performance metrics, which demonstrate the effectiveness of our multi-task network.

F. Comparison With State-of-the-Art Segmentation Methods

To cope with this extremely challenging task, we use multi-attention guided feature representation and adaptive

TABLE IV
ABLATION ANALYSIS OF THE PROPOSED 3D NETWORK AND THREE MULTI-TASK METHODS (MEAN±SD)

Method	Segmentation				Classification					
	DSC(%)	JI(%)	ASD(voxel)	95HD(voxel)	Acc(%)	AUC(%)	Pre(%)	Recall(%)	Spec(%)	F1(%)
Multi-task FCN [31]	57.8±3.2	40.9±3.2	25.6±12.8	39.8±18.6	78.5±4.1	80.3±9.8	79.5±5.1	85.2±6.7	53.5±14.6	81.2±3.5
Multi-task CNN [32]	56.5±3.5	39.5±3.6	28.2±14.5	42.5±20.6	76.7±5.2	78.2±6.7	77.6±6.9	86.5±4.5	52.3±15.1	80.6±3.8
$C_{MS}VNet_{Iter}$ [33]	60.1±3.2	43.1±3.3	17.6±13.0	32.6±15.4	73.0±9.1	73.6±14.6	76.0±8.0	88.4±9.9	40.1±16.6	81.7±6.4
Ours	62.7±2.5	45.5±2.0	12.8±9.8	26.9±9.1	80.5±5.3	86.0±9.2	82.1±5.2	88.8±4.2	61.6±19.1	86.2±2.8

TABLE V
3D AND 2D SEGMENTATION PERFORMANCE OF DIFFERENT MODELS (MEAN±SD)

Method	3D Segmentation				Method	2D Segmentation			
	DSC(%)	JI(%)	ASD(voxel)	95HD(voxel)		DSC(%)	JI(%)	Pre(%)	Recall(%)
3D U-Net [12]	60.2±3.5	43.2±4.1	18.3±19.1	35.7±22.2	2D U-Net [12]	56.8±4.9	39.8±4.7	68.3±5.9	49.3±7.3
AnatomyNet [45]	59.8±3.2	42.7±3.2	17.0±12.6	37.5±8.3	SegNet [48]	**58.1±3.5**	**41.0±3.5**	**71.0±4.9**	49.6±6.0
V-Net [44]	56.3±2.3	39.2±2.2	15.1±13.0	41.1±9.8	FCN [49]	55.9±1.9	38.8±1.8	69.9±5.1	46.7±3.2
3D FPN [13]	59.3±3.6	42.2±3.7	17.2±15.7	34.6±15.2	2D FPN [13]	55.1±4.1	38.1±3.9	68.1±5.6	46.6±5.4
DAF3D[26]	60.8±4.2	43.8±4.2	14.7±12.5	29.1±10.6	DeeplabV3+ [50]	57.7±3.3	40.6±3.3	70.6±6.1	49.0±3.8
Ours(3D)	**62.7±2.5**	**45.5±2.0**	**12.8±9.8**	**26.9±9.1**	Ours (2D)	**58.1±3.7**	40.8±3.9	61.5±3.7	**54.6±4.2**

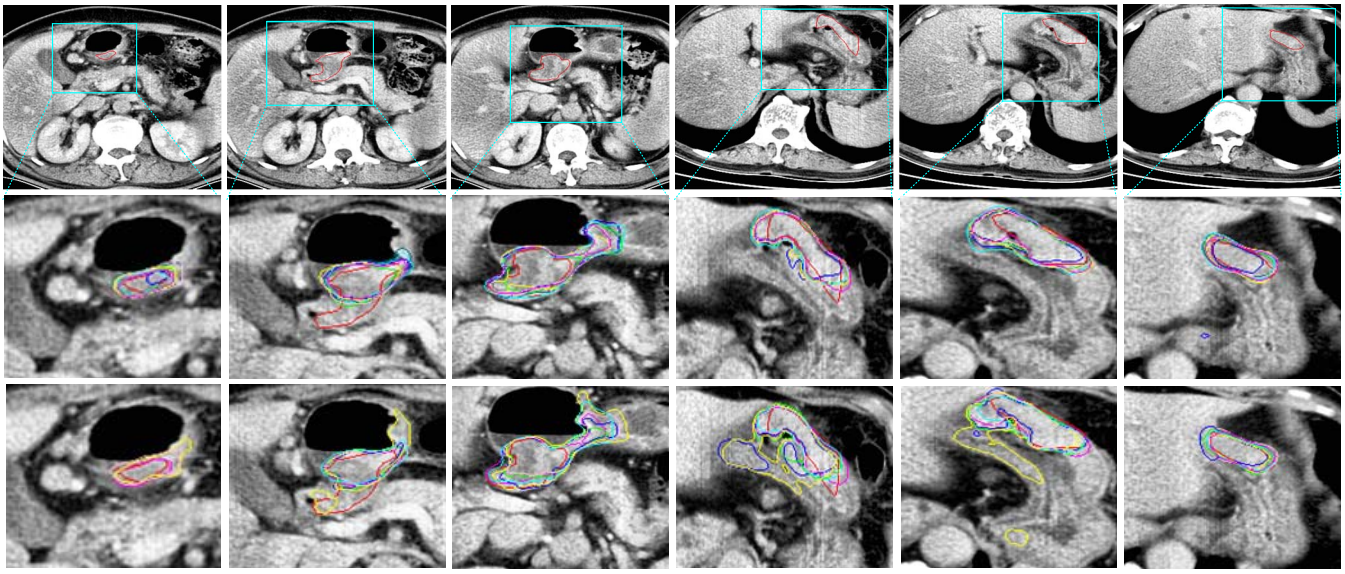

Fig. 7. 2D visual segmentation boundary comparison of different component combinations and competing models. Top row: CT slicers with red lines to indicate the tumor areas; Middle row: ground truth (red) and corresponding tumor boundaries using baseline (blue), baseline+VA (deep blue), baseline+ASA(green), baseline+TA(yellow), baseline+VA+ASA(cyan), baseline+VA+ASA+TA(pink), and the proposed method (purple). Bottom row: ground truth (red) and corresponding tumor boundaries using the proposed method (purple), 3D U-Net (blue), AnatomyNet (green),V-Net(yellow), 3D FPN (cyan) and DAF3D (pink).

feature fusion to predict the volume of lesions of different sizes. The segmentation performance of the proposed method and other segmentation methods are listed in Table V. Experimental results demonstrate that our 3D network outperforms other segmentation networks in gastric tumor segmentation. Specifically, our method yields the mean DSC value of 62.7%, JI of 45.5%, Pre of 67.2%, Recall of 62.5%, ASD of 12.8 voxels and 95HD of 26.9 voxels. Among all segmentation comparison methods, all metric values obtained by our method are the best. In comparison with the results of 3D FPN without refinement modules, the mean DSC, JI, Pre and Recall increased by approximately 5.7 %, 7.8%, 3.5% and 10.1%, respectively. In the 2D segmentation task, the proposed method and the SegNet yield the best DSC score. Besides, the proposed method achieves the best Recall score. In general, the results of 3D segmentation models are better than those of 2D segmentation models. It is worth noting that, compared with 2D U-Net and 2D FPN, both 3D U-Net and 3D FPN achieve higher DSC scores. We believe that almost all 3D segmentation models perform better than all 2D segmentation models because the 3D volume contains more sufficient spatial information.

Figs.7 and 8 visualize the 2D and 3D segmentation results, respectively. Fig. 7 shows the segmented boundaries by all 3D segmentation methods in 2D CT slices. Our method has the most similar segmented boundaries (purple) to the ground truths (red) inmost slices. These comparisons show that our model is quite effective in locating boundaries of lesions. Fig. 8 depicts three cases of 3D visualization surface distances between segmented surface and ground truth. When the green

TABLE VI
CLASSIFICATION PERFORMANCE OF DIFFERENT CLASSIFICATION MODELS (MEAN±SD)

Metric	VGG-16 [51]	ResNet-50 [14]	InceptionV4 [52]	ResNeXt-50 [53]	SENet-50 [24]	DenseNet-100 [54]	SENet-101 [24]	Ours
ACC (%)	68.9±6.9	76.2±4.8	67.5±7.7	78.0±3.2	78.4±3.6	69.3±4.0	78.0±1.9	**80.5±5.3**
AUC (%)	54.5±13.4	78.1±5.4	41.6±6.7	79.0±14.3	76.5±8.9	56.3±12.4	70.9±20.6	**86.0±9.2**
Pre (%)	70.5±8.1	77.7±6.4	70.3±8.4	78.6±4.6	77.7±5.0	73.0±5.2	77.1±1.5	**82.1±5.2**
Recall (%)	94.1±1.3	89.3±5.9	90.5±2.8	91.3±5.9	**95.5±3.7**	85.9±7.0	95.2±4.0	88.8±4.2
Spec (%)	13.8±5.4	45.9±16.8	16.8±4.1	46.6±15.1	40.1±9.4	29.1±23.5	35.1±23.5	**61.6±19.1**
F1 (%)	80.4±5.1	83.6±4.1	79.1±5.8	84.9±3.2	85.8±3.0	79.2±3.6	85.5±2.0	**86.2±2.8**

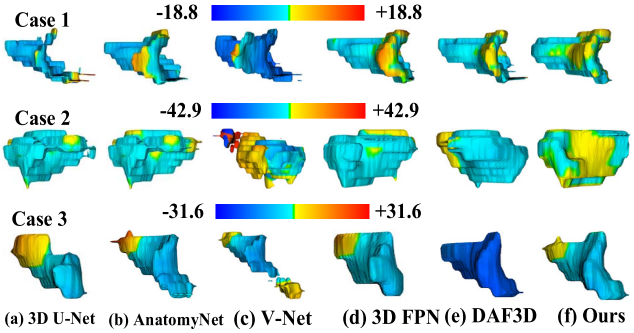

Fig. 8. 3D visualization of the surface distance (in voxel) between segmented surface and ground truth.

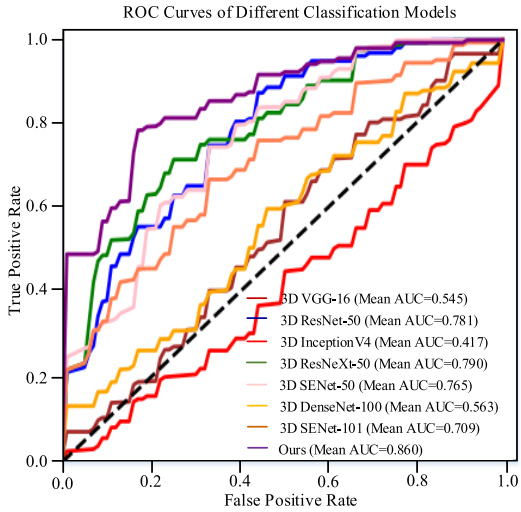

Fig. 9. Visualize the ROC curves of the proposed 3D network and other methods, and the purple line means the ROC curve of our proposed network.

area is larger, it means that the segmentation result is closer to the ground truth.

Based on the above quantitative analysis, we can see that these modules guided by multi-attention mechanisms are helpful for the refinement and fusion of complementary information between multi-scale features, and can achieve even better gastric tumor segmentation results.

G. Comparison With State-of-the-Art Classification Methods

Table VI illustrates that our proposed method obtains quite remarkable classification performance in the LN classification task, which are substantially better than the performance of other state-of-the-art methods. The performance improvement is mainly attributed to the use of multi-attention mechanisms,

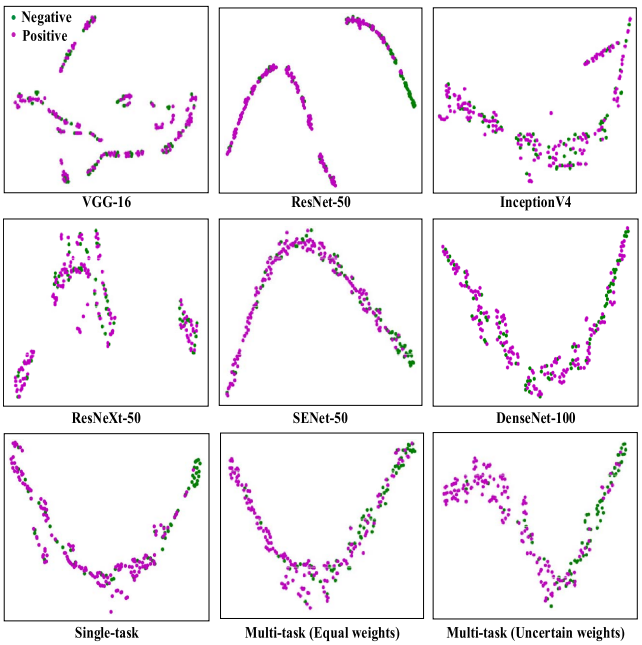

Fig. 10. The t-SNE visualization of different classification methods.

which enables a deep convolutional network to focus more on location information of LNs and thus strengthens the network's power to learn discriminative representation.

Table VI lists the numerical results of the evaluation metrics of all classification comparison methods. It can be observed that our method achieves good results using almost all evaluation metrics. Specifically, our method gets the mean Acc value of 80.5%, AUC of 86.0%, Pre of 82.1%, Recall of 88.8%, Spec of 61.6% and F1 of 87.8%. In comparison with the results of 3D SENet-50, the mean Acc, AUC, Pre, Spec and F1 increases by 5.6%, 10.1%, 5.7%, 34% and 3.1%, respectively. It is worth noting that with the deepening of the model, most of the evaluation metrics are generally declined. In comparison with the results of 3D SENet-101, the mean Acc, AUC, Pre, Spec and F1 increases by 3.2%, 21.3%, 6.5%, 75.5% and 0.8%, respectively. Particularly, in comparison with the results of 3D DenseNet-100, the mean Acc, AUC, Pre, Recall, Spec and F1 increases by 16.2%, 52.8%, 12.5%, 3.4%, 111.7% and 8.8%, respectively.

Fig. 9 shows the ROC curves comparison of competing models and the proposed model, and Fig. 10 visualizes the t-SNE of different classification methods. These two pictures intuitively confirm that our multi-task learning network can accurately identify negative LNs, thereby improving the specificity of the network to LNs. In Fig. 10, we can observe that

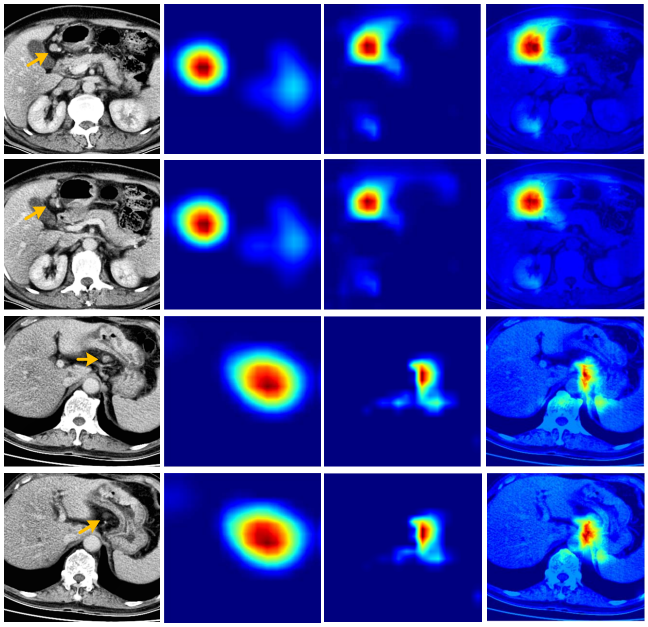

Fig. 11. Visualization of 2D CT images (left column), Overlay (right column) and the corresponding Grad-CAMs (middle two columns) obtained by single-task method and the proposed method.

TABLE VII
AUTOMATIC SEGMENTATION RESULTS OF DIFFERENT METHODS IN LITS CHALLENGE

Method	Liver Segmentation		Tumor Segmentation	
	Dice (%)	ASSD (voxel)	Dice (%)	ASSD (voxel)
H-DenseUNet [15]	96.1	1.69	72.2	1.07
3D DenseUNet [15]	93.6	-	59.4	-
AH-Net [55]	96.3	1.10	65.7	1.15
Med3D [56]	94.6	1.90	-	-
V-Net [44]	93.9	2.20	-	-
Our	93.1	5.67	67.0	1.28

Note: - represents the measurement was not evaluated.

the clusters of negative and positive LNs in the last row are separated more clearly than that in the top and middle rows. And the clustering results of multi-task learning networks are better than single-task learning network.

Furthermore, we visualized the gradient-weighted class activation mapping (Grad-CAM) obtained by our network, as shown in Fig. 11. It shows how much attention our model pays to the LNs, and we can clearly see that the LNs are always in the network's attention area. We can observe that the modules constructed by our multi-attention guided network are helpful for the fusion and refinement of complementary information between multi-scale features, and can achieve better LN classification results by harnessing the positional relatedness between tumors and LNs.

H. Robustness to LiTS Challenge

To further verify the effectiveness of this proposed model in tumor segmentation tasks, we also apply our method to the LiTS challenge. Also, we report the DSC and ASSD of our method and the other five methods: H-DenseUNet [15], 3D DenseUNet [15], AH-Net [55], Med3D [56] and V-Net. The H-DenseUNet not only explored 2D and 3D hybrid features by combining dense connection paths and a connection manner similar to U-Net, but also employed a transfer learning strategy. The 3D DenseUNet only explored 3D features by combining dense connection paths and connection manner (similar to U-Net). The AH-Net first extracts feature in 2D images, and then transfers the features to 3D volumes. The Med3D implemented the transfer learning strategy in eight medical imaging datasets. The comparison results are listed in Table VII. The proposed method achieves 93.1% and 67.0% Dice scores in liver segmentation and tumor segmentation, respectively. Our method only explores 3D spatial information, and does not use transfer learning technology, the results are close to other networks (H-DenseUNet, Med3D) using ensemble techniques. We also compare the Dice score with other networks, such as 3D DenseUNet and AH-Net, and show that our method outperforms those approaches in tumor segmentation. The above results further justify the generality of the proposed method.

VI. DISCUSSION

With the development of medical imaging equipment and deep learning algorithms, more and more neural networks are proposed for automated analysis of various cancers in various imaging modes. In this work, a 3D multi-attention guided multi-task learning network which yields universal multi-scale features and task-specific discriminative features for simultaneous gastric tumor segmentation and LN classification in CT images. Though deep learning methods have achieved good performance on different medical image segmentation and classification tasks, accurate gastric tumor segmentation and LN classification are quite challenging due to the inhomogeneous intensity distribution of gastric tumor and LN in CT scans, the ambiguous/missing boundaries and highly variable shapes of gastric tumor. Recently, since the attention-guided neural networks have demonstrated to be very powerful to learn features for effective fine-grained segmentation and object classification, we are inspired to design a multi-attention guided multi-task learning network to comprehensively tackle the above issues. To the best of our knowledge, there is no existing work to utilize deep learning for automatic analysis of gastric cancer in CT images.

We have also demonstrated that the proposed method improve tumor segmentation and LN classification results by progressive focusing on the tumor region and LN location. As shown in Figs. 5, 7, 8, 10 and 11, it demonstrates that these proposed modules can effectively learn shared feature and task-specific discriminative features for gastric tumor segmentation and LN classification. Our bidirectional VA layer imitates the theory of human visual perception, and focus the tumor region and LN location information through the seamless combination of receptive field and soft attention. In addition, the ASA module achieves cross-scale feature fusion through adaptively weighted single-scale attention feature maps and multi-scale attention feature maps. On the one hand, it can suppress invalid information and adaptively refine

tumor boundaries. On the other hand, it can further improve the specificity of LN classification by fully exploiting the spatial location information between tumors and LNs. As a flexible and effective multi-scale features fusion algorithm, these proposed feature interaction modules are potentially useful to become the plug-and-play components in other 3D segmentation and classification networks, and can improve the performance of other tasks by fully exploring 3D spatial features. In addition, our method can be directly applied to other 3D medical image dataset, and can achieve single task or multi-task. For example, our method also achieves good results on LiTS dataset.

Compared with the single-task learning network, the proposed multi-task learning network has a huge gain in the classification outcomes, but the gain in the segmentation outcomes is mediocre. From a technical perspective, there may be a main reason. In detail, the tumor segmentation task is dedicated to achieve the pixel-level classification of the tumor area, whereas the LN classification task is dedicated to achieve the volume-level classification of the LN metastasis. Therefore, the location and morphological characteristics of the tumor are strongly related to the LN classification task, while the location information of the LN is weakly related to the tumor segmentation task. It is worth noting that there were no distant LN metastases in the 160 samples we collected. Besides, although our method is superior to other state-of-the-art methods in the segmentation and classification tasks, there is still room for improvement regarding segmentation task. In fact, most of the samples in our dataset are ordinary CT volumes, and the slice thicknesses of 5 mm and 8 mm limit the segmentation performance. According to the three experienced radiologists, the layer thickness of CT data is crucial for visualization. When the layer thickness is large, the contrast between the lesion and the organ is weak, so the data carries very little discriminative information. Therefore, our future research work will not only focus on the design of multi-task models, but also collect a wide range of data to make the model more generalized.

VII. Conclusion

In this paper, we propose a multi-attention guided multi-task learning network, which generates scale-aware and task-aware attentive features for automatic gastric tumor segmentation and LN classification. The proposed network can be trained in an end-to-end manner to achieve better segmentation and classification results, simultaneously. Our key idea is to select useful complementary information from multi-scale features by multi-attention mechanisms, and fully exploiting the location information between tumors and LNs. These algorithms in the proposed method mimic the theory of human visual perception and attention fine-tuning, which allow the network to autonomously learn refined multi-scale features with high feature consistency. Compared with other state-of-the-art methods, the proposed method has achieved promising performance results for tumor segmentation and LN classification using our self-collected CT dataset. Meanwhile, we also verified the generality of our method on the LiTS dataset.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, A Cancer J. Clinicians*, vol. 68, no. 6, pp. 394–424, Nov. 2018.

[2] E. Van Cutsem, X. Sagaert, B. Topal, K. Haustermans, and H. Prenen, "Gastric cancer," *Lancet*, vol. 388, no. 10060, pp. 2654–2664, 2016.

[3] R. M. Kwee and T. C. Kwee, "Imaging in local staging of gastric cancer: A systematic review," *J. Clin. Oncol.*, vol. 25, no. 15, pp. 2107–2116, May 2007.

[4] N. Coburn *et al.*, "Staging and surgical approaches in gastric cancer: A systematic review," *Cancer Treatment Rev.*, vol. 63, pp. 104–115, Feb. 2018.

[5] T. Fukuya *et al.*, "Lymph-node metastases: Efficacy for detection with helical CT in patients with gastric cancer.," *Radiology*, vol. 197, no. 3, pp. 705–711, Dec. 1995.

[6] Y. Wang *et al.*, "CT radiomics nomogram for the preoperative prediction of lymph node metastasis in gastric cancer," *Eur. Radiol.*, vol. 30, no. 2, pp. 976–986, Feb. 2020.

[7] Q.-X. Feng *et al.*, "An intelligent clinical decision support system for preoperative prediction of lymph node metastasis in gastric cancer," *J. Amer. College Radiol.*, vol. 16, no. 7, pp. 952–960, Jul. 2019.

[8] X. Chen, J. K. Udupa, U. Bagci, Y. Zhuge, and J. Yao, "Medical image segmentation by combining graph cuts and oriented active appearance models," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2035–2046, Apr. 2012.

[9] V. Grau, A. U. J. Mewes, M. Alcaniz, R. Kikinis, and S. K. Warfield, "Improved watershed transform for medical image segmentation using prior information," *IEEE Trans. Med. Imag.*, vol. 23, no. 4, pp. 447–458, Apr. 2004.

[10] C.-J. Hsieh, K.-W. Chang, C.-J. Lin, S. S. Keerthi, and S. Sundararajan, "A dual coordinate descent method for large-scale linear SVM," in *Proc. 25th Int. Conf. Mach. Learn. ICML*, 2008, pp. 408–415.

[11] D. Grossman and P. Domingos, "Learning Bayesian network classifiers by maximizing conditional likelihood," in *Proc. 21st Int. Conf. Mach. Learn. ICML*, 2004, pp. 46–53.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assisted Intervent*, vol. 2015, pp. 234–241.

[13] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

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

[15] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. 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.

[16] F. Khalifa *et al.*, "3D kidney segmentation from CT images using a level set approach guided by a novel stochastic speed function," in *Proc. MICCAI*, Berlin, Germany, 2011, pp. 587–594.

[17] Y. Zhou, L. Xie, W. Shen, Y. Wang, E. K. Fishman, and A. L. Yuille, "A fixed-point model for pancreas segmentation in abdominal CT scans," in *Proc. MICCAI*, 2017, pp. 693–701.

[18] Z. Liu *et al.*, "Liver CT sequence segmentation based with improved U-Net and graph cut," *Expert Syst. Appl.*, vol. 126, pp. 54–63, Jul. 2019.

[19] Q. Yu, Y. Shi, J. Sun, Y. Gao, J. Zhu, and Y. Dai, "Crossbar-Net: A novel convolutional neural network for kidney tumor segmentation in CT images," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 4060–4074, Aug. 2019.

[20] Q. Dou, H. Chen, Y. Jin, L. Yu, J. Qin, and P.-A. Heng, "3D deeply supervised network for automatic liver segmentation from CT volumes," in *Proc. MICCAI*, Athens, Greece, 2016, pp. 149–157.

[21] W. Shen *et al.*, "Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification," *Pattern Recognit.*, vol. 61, pp. 663–673, Jan. 2017.

[22] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. ICML*, Lille, France, 2015, pp. 2048–2057.

[23] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, Salt Lake City, UT, USA, Jun. 2018, pp. 7132–7141.

[25] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proc. ECCV*, Munich, Germany, Sep. 2018, pp. 3–19.
[26] Y. Wang *et al.*, "Deep attentive features for prostate segmentation in 3D transrectal ultrasound," *IEEE Trans. Med. Imag.*, vol. 38, no. 12, pp. 2768–2778, Dec. 2019.
[27] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1871–1880.
[28] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *Proc. NIPS*, 2007, pp. 41–48.
[29] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Proc. ECCV*, Zurich, Switzerland, 2014, pp. 94–108.
[30] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Robust visual tracking via multi-task sparse learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2042–2049.
[31] A. I. L. Namburete, W. Xie, M. Yaqub, A. Zisserman, and J. A. Noble, "Fully-automated alignment of 3D fetal brain ultrasound to a canonical reference space using multi-task learning," *Med. Image Anal.*, vol. 46, pp. 1–14, May 2018.
[32] A. Chakravarty and J. Sivswamy, "A deep learning based joint segmentation and classification framework for glaucoma assesment in retinal color fundus images," 2018, *arXiv:1808.01355*. [Online]. Available: http://arxiv.org/abs/1808.01355
[33] Y. Zhou *et al.*, "Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images," *Med. Image Anal.*, Nov. 2020, Art. no. 101918.
[34] S. Chen *et al.*, "Automatic scoring of multiple semantic attributes with multi-task feature leverage: A study on pulmonary nodules in CT images," *IEEE Trans. Med. Imag.*, vol. 36, no. 3, pp. 802–814, Mar. 2017.
[35] R. Cipolla, Y. Gal, and A. Kendall, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7482–7491.
[36] S. Liu and D. Huang, "Receptive field block net for accurate and fast object detection," in *Proc. ECCV*, Munich, Germany, Sep. 2018, pp. 385–400.
[37] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Human Neurobiol.*, vol. 4, pp. 219–227, 1985.
[38] S. K. A. L. G. Ungerleider, "Mechanisms of visual attention in the human cortex," *Annu. Rev. Neurosci.*, vol. 23, no. 1, pp. 315–341, Mar. 2000.
[39] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10781–10790.
[40] Q. Zhao *et al.*, "M2det: A single-shot object detector based on multi-level feature pyramid network," in *Proc. AAAI Conf. Artific. Intel.*, 2019, pp. 9259–9266.
[41] S. Liu, D. Huang, and Y. Wang, "Learning spatial fusion for single-shot object detection," 2019, *arXiv:1911.09516*. [Online]. Available: http://arxiv.org/abs/1911.09516
[42] H. Xu and J. Zhang, "AANet: Adaptive aggregation network for efficient stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1959–1968.
[43] *Liver Tumor Segmentation Challenge*. Accessed: Jan. 3, 2021. [Online]. Available: https://competitions.codalab.org/competitions/17094#results
[44] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-NBet: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.
[45] W. Zhu *et al.*, "AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy," *Med. Phys.*, vol. 46, no. 2, pp. 576–589, Feb. 2019.
[46] L. Zhang *et al.*, "Block level skip connections across cascaded V-Net for multi-organ segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 9, pp. 2782–2793, Sep. 2020.
[47] T. Wang *et al.*, "Learning-based automatic segmentation of arteriovenous malformations on contrast CT images in brain stereotactic radiosurgery," *Med. Phys.*, vol. 46, no. 7, pp. 3133–3141, May 2019.
[48] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[49] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
[50] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. ECCV*, Sep. 2018, pp. 801–818.
[51] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[52] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," 2016, *arXiv:1602.07261*. [Online]. Available: http://arxiv.org/abs/1602.07261
[53] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1492–1500.
[54] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[55] S. Liu *et al.*, "3D anisotropic hybrid network: Transferring convolutional features from 2D images to 3D anisotropic volumes," in *Proc. MICCAI*, Granada, Spain, 2018, pp. 851–858.
[56] S. Chen, K. Ma, and Y. Zheng, "Med3D: Transfer learning for 3D medical image analysis," 2019, *arXiv:1904.00625*. [Online]. Available: http://arxiv.org/abs/1904.00625