Learning Inductive Attention Guidance for Partially Supervised Pancreatic Ductal Adenocarcinoma Prediction

Yan Wang, Peng Tang, Yuyin Zhou, Wei Shen, Elliot K. Fishman, and Alan L. Yuille, *Fellow, IEEE*

***Abstract*— Pancreatic ductal adenocarcinoma (PDAC) is the third most common cause of cancer death in the United States. Predicting tumors like PDACs (including both classification and segmentation) from medical images by deep learning is becoming a growing trend, but usually a large number of annotated data are required for training, which is very labor-intensive and time-consuming. In this paper, we consider a partially supervised setting, where cheap image-level annotations are provided for all the training data, and the costly per-voxel annotations are only available for a subset of them. We propose an Inductive Attention Guidance Network (IAG-Net) to jointly learn a global image-level classifier for normal/PDAC classification and a local voxel-level classifier for semi-supervised PDAC segmentation. We instantiate both the global and the local classifiers by multiple instance learning (MIL), where the attention guidance, indicating roughly where the PDAC regions are, is the key to bridging them: For global MIL based normal/PDAC classification, attention serves as a weight for each instance (voxel) during MIL pooling, which eliminates the distraction from the background; For local MIL based semi-supervised PDAC segmentation, the attention guidance is inductive, which not only provides bag-level pseudo-labels to training data without per-voxel annotations for MIL training, but also acts as a proxy of an instance-level classifier. Experimental results show that our IAG-Net boosts PDAC segmentation accuracy by more than 5% compared with the state-of-the-arts.**

***Index Terms*— Attention, multiple instance learning, semi-supervised learning, medical image segmentation.**

I. Introduction

PANCREATIC ductal adenocarcinoma (PDAC) is one of the most deadly diseases, whose prognosis is dismal as more than 50% of patients have evidence of metastatic disease at the time of diagnosis. Currently, detecting or segmenting PDACs through medical imaging at the localized disease stage followed by complete resection can offer the best chance of survival [1]. Computed tomography (CT) screening is the most commonly used imaging modality for the initial evaluation of PDACs. However, finding PDACs in CT images is challenging, even for experienced radiologists. Therefore, to build up computer-aided diagnosis (CAD) systems with the ability to automatically identify suspicious cases and alert radiologists is vital. In clinical environments, this cannot be simply formulated as either a classification or localization/segmentation problem, but a joint problem of these two tasks. Our goal is to address the PDAC cancer prediction problem: given a CT scan of a patient, we need to determine (*i.e.*, classify) whether this patient suffers from PDAC cancer or not, and if yes, to localize where the PDAC region is. The latter is of great importance as it provides a visual interpretation to support the former result.

A growing literature has proposed various techniques based on supervised learning for medical image segmentation/localization [2]–[7]. But these methods require a large amount of per-voxel annotated data. Usually, this high-quality manual contouring process is not only tedious, but also expensive (considering the cost of salaries, segmentation software, and training). This situation is more conspicuous for PDAC segmentation in CT images as PDAC masses in the early stage are usually small and their boundaries are usually weak, which can even confuse radiologists. Thus, we consider addressing the PDAC cancer prediction problem under a (weakly) partially supervised setting: we are given the training data with image-level annotations, *i.e.*, we know whether each CT image has PDAC masses or not, but only a small subset of them have per-voxel annotations, *i.e.*, we know whether each voxel belongs to a PDAC region or not.

Under this partially supervised setting, to address the PDAC cancer prediction problem, *i.e.*, the joint (multi-task) problem of normal/PDAC classification and PDAC segmentation, two types of weakly supervised learning techniques can be used. One is inexact supervised learning (ISL) [8], such as multiple instance learning (MIL), the other is semi-supervised learning (SSL) [9], such as Expectation-Maximization (EM) [10]. MIL based methods build a bag-level (image-level) classifier upon bag representations aggregated from instance (voxel) features over the whole image, with the ability to infer per-voxel labels from image-level annotations [11]–[13]. EM-like methods make use of the large amount of training

Manuscript received January 7, 2021; accepted February 9, 2021. Date of publication February 18, 2021; date of current version September 30, 2021. This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research. *(Corresponding author: Wei Shen.)*

Yan Wang is with the Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai 200241, China, and also with the Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: wyanny.9@gmail.com).

Peng Tang, Yuyin Zhou, and Alan L. Yuille are with the Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: tangpeng723@gmail.com; zhouyuyiner@gmail.com; alan.l.yuille@gmail.com).

Wei Shen is with the MoE Key Laboratory of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: shenwei1231@gmail.com).

Elliot K. Fishman is with the Department of Radiology, The Johns Hopkins University School of Medicine, Baltimore, MD 21287 USA (e-mail: efishman@jhmi.edu).

Digital Object Identifier 10.1109/TMI.2021.3060066

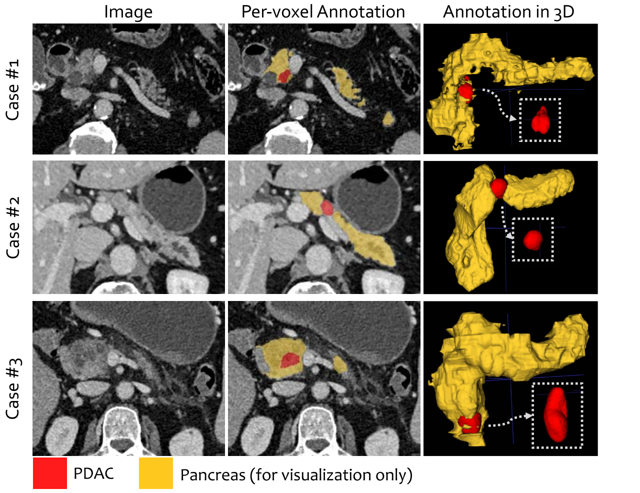

Fig. 1. PDAC examples shown in 2D and rendered in 3D. Per-voxel pancreas labels are also shown as reference. Since part of the PDAC can be inside the pancreas, the whole PDAC is shown in white dashed boxes in 3D.

data without per-voxel annotations in conjunction with the small amount of training data with per-voxel annotations to improve instance-level classifiers for localization/segmentation [14]–[17]. Nevertheless, due to the difficulties in PDAC segmentation in the early stage, such as small sizes, weak boundaries, variety of shapes and diverse locations with a pancreas region (see Fig. 1), current MIL-based classification methods or EM-like segmentation methods or multi-task learning methods cannot achieve satisfactory results. The MIL-based methods, *e.g.*, [18], used a classic MIL operator (*i.e.,* xor) to aggregate all the instance features to form the bag representation, which is vulnerable to the distraction from the background, since a PDAC is usually much smaller than the whole pancreas region; The EM-like methods, *e.g.*, [19] relied on self-training strategies [20]. But the generated per-voxel pseudo-labels are often heavily noisy due to the difficulties in PDAC segmentation; The multi-task learning methods, *e.g.*, [21], essentially conducted these two learning tasks separately, except for shared feature learning, and thus suffer from both above issues.

To cope with the above difficulties, in this paper, we propose an attention-guided framework to jointly learn a global (image-level) classifier for normal/PDAC classification and a local (instance-level) classifier for semi-supervised PDAC segmentation. The attention guidance, indicating roughly where PDAC regions are, is explicitly learned from the training data with per-voxel annotations and inducted on the training data without per-voxel annotations. In this framework, both the global and the local classifiers are instantiated by MIL and the attention guidance is the key to bridging them. For normal/PDAC classification, the attention guidance serves as a weight for each instance, resulting in weighted MIL pooling to suppress the distraction from background when training the global MIL classifier; For semi-supervised PDAC segmentation, the attention guidance separates the PDAC and the background regions on the training data without per-voxel annotations as bag-level pseudo labels for training the local MIL classifier in the EM manner. Since the instances of these training data without per-voxel annotations are "bagged" instead of treated as singletons, the local MIL classifier offers a possible way to mitigating the effects of noises in per-voxel pseudo labels [22], as shown in Fig. 2. Note that, the attention guidance is **inductive** [23], since it can not only provide bag-level pseudo-labels to training data without per-voxel annotations, but also act as a proxy of a local instance-level classifier defined on all the data which have PDAC masses.

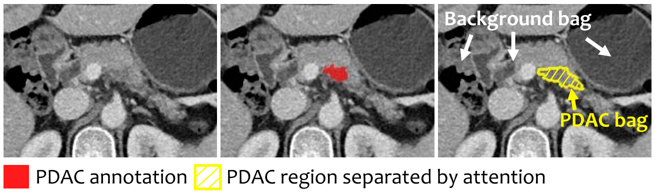

Fig. 2. The PDAC region separated by attention consists of background voxels (middle & right images). Per-voxel pseudo-label is noisy, but from the MIL perspective, the bag-level pseudo-labels assigned according to the separation are correct.

We instantiate our attention-guided framework by a single network with two streams. The backbone of the network, *e.g.*, a VGG-Net [24] or an U-Net [25], generates a convolutional feature map from an input image, in which each feature vector at a spatial location is an instance. Then the attention is learned explicitly on the feature map from training data with per-voxel annotations. The feature map is further branched into two streams: The first stream trains the global MIL classifier by using the attention guidance as a weighting mechanism for instances in MIL pooling; The second stream first separates PDAC regions and background regions on the training data without per-voxel annotations according to the attention guidance, and then trains the local MIL classifier by bagging the instances in the PDAC regions and background regions, respectively. We refer to this network with two streams as **IAG-Net**, for **I**nductive **A**ttention **G**uidance **Net**work.

Given a training set consisting of normal (*i.e.*, healthy cases) and abnormal (*i.e.*, cases diagnosed with PDAC) CT scans under the partially supervised setting, all the streams and the backbone of IAG-Net can be jointly trained. It provides both image-level (*i.e.*, normal/PDAC classification) results and corresponding voxel-level (*i.e.*, PDAC segmentation) visual evidences. Experimental results show such a design performs favorably for PDAC prediction, boosting PDAC segmentation results by a large margin.

The contribution of this paper is four-fold:

1) We propose an attention-guided framework to address classification and segmentation of PDAC under the partially supervised setting, which is of great potential for practical applications.
2) Unlike previous semi-supervised segmentation methods, which train segmentors (instance-level classifiers) on per-voxel pseudo-labels, our framework trains the instance-level classifier by MIL. It takes the pseudo-PDAC and background regions as bags, which addresses the issue of heavy noises in per-voxel pseudo labels.

3) Our framework has active interactions between PDAC/normal classification and PDAC segmentation, bridging by the inductive attention guidance.
4) We instantiate our framework by a single network with two streams, IAG-Net, in which the backbone and the two streams can be trained jointly. IAG-Net achieves a substantial improvement of around 5% DSC for PDAC segmentation.

II. Related Work

Weakly supervised learning is widely used in the field of computer vision [8], [26]–[31]. In this section, we briefly review the related works on weakly-supervised learning for medical image classification and segmentation/detection.

A. Semi-Supervised Learning

Semi-supervised learning (SSL) [9], is also known as incomplete supervised learning [8], where only a small subset of training data are labeled whereas the other data remain unlabeled. The most widely used techniques for SSL are EM-like methods such as self-training [14] and co-training [19]. Other directions such as consistency-based methods [15], [32], [33] are becoming popular recently.

Self-training propagates labels from the labeled to the unlabeled data, and then using the larger, newly labeled set for training. This approach assumes that the method's high confidence predictions are correct. The expectation-maximization procedure alternates between assigning pseudo-labels to the unlabeled data given the labeled data and model parameters, and updating the model parameters given all the data [14], [19]. [14] trained and predicted on a single plane while the work proposed in [19], termed as DMPCT, distilled consensus information from three planes of the 3D volume of a CT scan. DMPCT [19] adopted the idea from co-training [34], where classifiers are trained with independent sets of features, and the classifiers rely on each other for estimating the confidence of their predictions. Many SSL methods heavily rely on the quality of per-voxel pseudo-labels, which is hard to be guaranteed when the segmentation task is challenging, *e.g.*, such as PDAC segmentation. Our method belongs to self-training based methods. But the pseudo-label generated by our method is in bag-level (see Fig. 2), addressed by MIL [35], and it can tolerate voxel-level errors. Another research direction is consistency-based methods [15], [32], [36], which encouraged consistent segmentation/classification of the network-in-training for the same input (on both labeled and unlabeled images) under a given class of transformations. These methods are complementary with ours [37].

B. Inexact Supervised Learning

In inexact supervised learning (ISL) [8], only coarse-grained labels are provided. One particular form of ISL is multiple instance learning (MIL), first proposed for drug activities prediction [35]. In MIL, a training set consists of a number of bags, each of which is assigned with a positive or negative label (or multi-class label). Each bag contains a group of instances, which are not individually labeled like traditional supervised learning setting. The goal of MIL is to learn instance classifiers under MIL constraints (*i.e.*, a bag should be labeled as positive if at least one of its instances is positive and labeled as negative otherwise). Many previous methods in the medical imaging community adopted the multiple instance learning (MIL) pipeline for weakly supervised abnormality classification and detection. These methods have achieved encouraging results on various tasks [11]–[13]. Xu *et al.* [11] proposed a first integrated framework for histopathology cancer image classification, segmentation and clustering. Yao *et al.* [12] embedded MIL into deep learning frameworks for thoracic disease identification and localization. Breast tumor histology classification uses quantile aggregation to predict the class of the cropped image region [13]. But, there is still a big gap between the results of these MIL algorithms with fully supervised ones. Besides, most MIL-based algorithms [11]–[13] do not have any instance label, thus localization results are far from satisfactory.

C. Partially Supervised Learning

There are also some methods considering partially supervised learning, which use a large number of image-level annotated data, and a subset of pixel-level annotated data [18], [21], [38], [39]. Li *et al.* [18] proposed a unified framework for disease identification and localization of abnormalities in chest X-ray images. CIA-Net [38] exploited the highly structured property of chest X-ray images which localized diseases via a pair of aligned positive and negative samples. Collaborative learning [39] jointly improves the performance of disease grading and lesion segmentation on fundus images for diabetic retinopathy. It used attention maps generated by lesion attentive classification module as pseudo-labels for SSL.

Different from [18], [38] which are tailored for detection, our IAG-Net is designed for segmentation. Additionally, our work differs from [18] in that: Li's work did not leverage the localization information from the data without per-voxel annotations. Meanwhile, our differences compared with [38] are 1) our IAG-Net does not rely on aligned images, while the attention map of Liu's CIA-Net relies on a pair of aligned positive and negative images; 2) we assign pseudo-labels to the data without per-voxel annotation while CIA-Net leverages the spatial-wise attention map indicating the possible location of the disease in images without bounding box annotation. The differences of our work compared with Zhou's collaborative learning method [39] are 1) for images without per-voxel annotations, our IAG-Net uses bag-level pseudo-labels, while Zhou's work uses per-voxel pseudo-labels; 2) the global classifier in IAG-Net is designed under popular MIL constraints for solving classification problems where positive instances only exist in positive bags, while Zhou's work is aiming at classifying disease severity gradings, where lesion symptoms may co-exist in different classes.

III. Inductive Attention Guidance Network

Mathematically, let the 3D volume of a CT image[1] denoted by $\mathbf{Z} \in \mathbb{R}^{W \times H \times L}$. The goal of PDAC prediction is to predict 1) the image-level label $\hat{y} \in \{0, 1\}$ of $\mathbf{Z}$, indicating whether it contains PDAC masses ($\hat{y} = 1$) or not ($\hat{y} = 0$), and

[1] 3D image will be termed as *image* for short in the rest of the paper

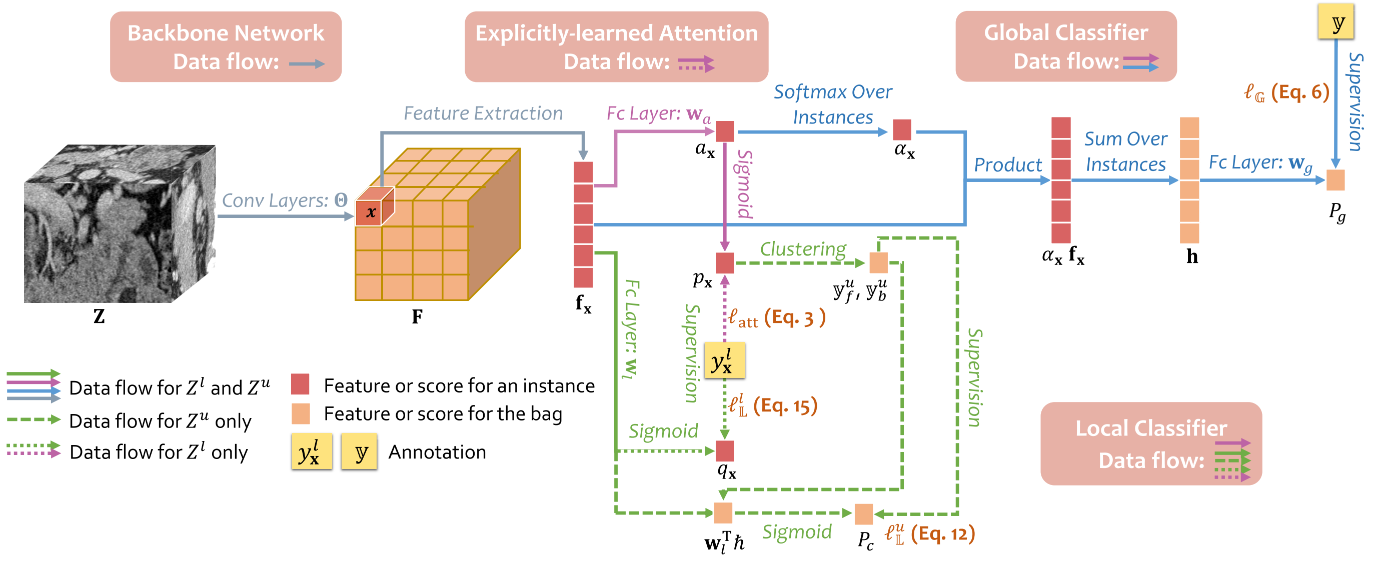

Fig. 3. The architecture of IAG-Net. Our IAG-Net has a backbone network and two streams: 1) Global classifier for normal/PDAC classification (Sec. III-B) and 2) Local classifier (Sec. III-C) for PDAC segmentation. Explicitly-learned attention (Sec. III-A) bridges the two streams, and it is also a proxy of the local classifier. The volume of a CT scan $\mathbf{Z}$ is fed into a backbone network, parameterized by $\boldsymbol{\Theta}$. $\mathbf{F}$ is the feature map. $\mathbf{f}_{\mathbf{x}}$ represents the feature vector at spatial location $\mathbf{x}$ on the feature map. Attention value $a_{\mathbf{x}}$ is obtained through Eq. 1. Attention values are normalized through a softmax function to get $\alpha_{\mathbf{x}}$. Image-level feature $\mathbf{h}$ is obtained through Eq. 5. P_g is the probability that the image belongs to PDAC (see Eq. 7). If $\mathbf{Z} = \mathbf{Z}^l$, the probability-like attention value $p_{\mathbf{x}}$ (Eq. 2) and probability $q_{\mathbf{x}}$ (Eq. 14) are supervised by the annotation of the instance $y_{\mathbf{x}}^l$. If $\mathbf{Z} = \mathbf{Z}^u$, pseudo-labels y_f^u and y_b^u are generated by attention-based separation (Eq. 9), which are used to train a local MIL classifier under the SSL setting (Eq. 12).

2) the per-voxel label map $\hat{\mathbf{Y}} \in \{0,1\}^{W\times H\times L}$, indicating where PDAC masses are in $\mathbf{Z}$. Our training set $\mathcal{D}$ consists of two subsets: $\mathcal{D} = \mathcal{D}^l \bigcup \mathcal{D}^u$, where $\mathcal{D}^l = \{(\mathbf{Z}^l, \mathbf{Y}^l)\}_{l=1}^{L}$ and $\mathcal{D}^u = \{(\mathbf{Z}^u, \mathrm{y}^u)\}_{u=L+1}^{U}$. The training data in $\mathcal{D}^l$ are given by **annotated per-voxel label maps** while those in $\mathcal{D}^u$ are only given by **annotated image-level labels**. Note that, if the per-voxel label map $\mathbf{Y}$ is known, then the image-level label y is also known, but not vice versa if $\mathrm{y} = 1$.

The overall pipeline of the proposed IAG-Net is shown in Fig. 3. An image $\mathbf{Z}$ is first fed into the backbone parameterized by $\boldsymbol{\Theta}$ to produce a feature map $\mathbf{F}(\mathbf{Z}; \boldsymbol{\Theta}) \in \mathbb{R}^{W'\times H'\times L'\times C}$, where C is the number of feature channels. The feature vector $\mathbf{f}_{\mathbf{x}}(\mathbf{Z}; \boldsymbol{\Theta}) \in \mathbb{R}^C$ at each spatial location $\mathbf{x}$ on the feature map $\mathbf{F}(\mathbf{Z}; \boldsymbol{\Theta})$ represents the feature of an instance. Then attention values are learned explicitly on the feature map from the training data with per-voxel annotations which have PDAC masses (Sec. III-A). The feature map is branched into two streams: The first stream performs attention-guided MIL to train a global image-level (bag-level) classifier on all the training data (Sec. III-B); The second stream trains the local MIL classifier by bagging the instances in the PDAC regions and background regions inducted by attention on the training data without per-voxel annotations (Sec. III-C). The overall loss function for training IAG-Net is the sum of loss functions for the two streams. Next, we describe each module in detail. Finally, we show the testing procedure of IAG-Net.

A. Explicitly-Learned Attention

Intuitively, high attention values highlight possible PDAC regions. Since attention acts as a proxy of an instance-level local classifier, explicitly learning attention on the training data with per-voxel annotations is straightforward.

We define the attention value of an instance $\mathbf{f}_{\mathbf{x}}(\mathbf{Z}; \boldsymbol{\Theta})$ as its linear projection:

$$a_{\mathbf{x}}(\mathbf{Z}; \boldsymbol{\Theta}, \mathbf{w}_a) = \mathbf{w}_a^\top \mathbf{f}_{\mathbf{x}}(\mathbf{Z}; \boldsymbol{\Theta}), \tag{1}$$

where $\mathbf{w}_a \in \mathbb{R}^C$. To learn the attention value for each instance, we first apply a sigmoid function to it to obtain a probability-like attention value:

$$p_{\mathbf{x}}(\mathbf{Z}; \Theta, \mathbf{w}_a) = \frac{1}{1 + \exp\big(-a_{\mathbf{x}}(\mathbf{Z}; \boldsymbol{\Theta}, \mathbf{w}_a)\big)}, \tag{2}$$

then minimize the cross-entropy loss on the training data with per-voxel annotations:

$$\begin{aligned}\ell_{\text{att}}(\mathbf{Z}^l, \mathbf{Y}^l; \Theta, \mathbf{w}_a) = -\sum_{\mathbf{x}\in\mathcal{X}} \big[&y_{\mathbf{x}}^l \log p_{\mathbf{x}}^a(\mathbf{Z}^l; \Theta, \mathbf{w}_a) \\ &+ (1 - y_{\mathbf{x}}^l)\log(1 - \big(p_{\mathbf{x}}^a(\mathbf{Z}^l; \Theta, \mathbf{w}_a)\big)\big],\end{aligned} \tag{3}$$

where $y_{\mathbf{x}}^l$ is the annotation of the instance at spatial location $\mathbf{x}$ obtained directly from $\mathbf{Y}^l$ and $\mathcal{X}$ is the whole $H' \times W' \times L'$ location lattice.

B. Global Classifier

We now introduce our attention-guided MIL module for image-level label prediction, *i.e.*, normal/PDAC classification. MIL pooling usually plays an important role to form bag representations from the instance features. Typical MIL pooling algorithms include max pooling and average or sum pooling. But, max pooling can be easily influenced by noises, and average pooling or sum pooling may filter out PDAC signals. To address this problem, inspired by attention-based MIL [40], we use a weighted average pooling method, where the weights are determined by the attention values. First, we normalize the attention values (Eq. 1) over spatial locations to be attention weights by a softmax function:

$$\alpha_{\mathbf{x}}(\mathbf{Z};\Theta,\mathbf{w}_a)=\frac{\exp\left(a_{\mathbf{x}}(\mathbf{Z};\boldsymbol{\Theta},\mathbf{w}_a)\right)}{\sum_{\mathbf{x}\in\mathcal{X}}\exp\left(a_{\mathbf{x}}(\mathbf{Z};\boldsymbol{\Theta},\mathbf{w}_a)\right)}. \tag{4}$$

Note that, Eq. 1 is generic for both $\mathbf{Z}^l$ and $\mathbf{Z}^u$. Then, we obtain the bag vector representation $\mathbf{h}(\mathbf{Z};\Theta,\mathbf{w}_a)\in\mathbb{R}^C$ of image $\mathbf{Z}$ by our attention-guided MIL pooling:

$$\mathbf{h}(\mathbf{Z};\Theta,\mathbf{w}_a)=\sum_{\mathbf{x}\in\mathcal{X}}\alpha_{\mathbf{x}}(\mathbf{Z};\Theta,\mathbf{w}_a)\mathbf{f}_{\mathbf{x}}(\mathbf{Z};\Theta). \tag{5}$$

Based on the bag representation, we train a global image-level classifier for normal/PDAC classification by minimizing the cross-entropy loss:

$$\begin{aligned}\ell_{\mathbb{G}}(\mathbf{Z},\mathrm{y};\Theta,\mathbf{w}_a,\mathbf{w}_g)=&-\big[\mathrm{y}\log P_g(\mathbf{Z};\Theta,\mathbf{w}_a,\mathbf{w}_g)\\&+(1-\mathrm{y})\log(1-P_g\big(\mathbf{Z};\Theta,\mathbf{w}_a,\mathbf{w}_g)\big)\big],\end{aligned} \tag{6}$$

where $\mathbf{w}_g\in\mathbb{R}^C$ is the parameter of the global image-level classifier, and

$$P_g(\mathbf{Z};\Theta,\mathbf{w}_a,\mathbf{w}_g)=\frac{1}{1+\exp\big(-\mathbf{w}_g^\top\mathbf{h}(\mathbf{Z};\Theta,\mathbf{w}_a)\big)}, \tag{7}$$

representing the probability that image $\mathbf{Z}$ contains PDAC masses. Note that, the parameter $\mathbf{w}_a$ for computing attention values is optimized by minimizing both Eq. 3 and Eq. 6. So for both $\mathbf{Z}^l$ and $\mathbf{Z}^u$, their attention values are implicitly learned by minimizing Eq. 6, while the attention values of $\mathbf{Z}^l$ are explicitly learned by minimizing Eq. 3, additionally. The loss function for training the global classifier over the whole trainig set $\mathcal{D}$ is defined by

$$L_{\text{global}}(\mathcal{D};\boldsymbol{\Theta},\mathbf{w}_a,\mathbf{w}_g)=\frac{1}{|\mathcal{D}|}\sum_{(\mathbf{Z},\mathrm{y})\in\mathcal{D}}\ell_{\mathbb{G}}(\mathbf{Z},\mathrm{y};\Theta,\mathbf{w}_a,\mathbf{w}_g). \tag{8}$$

C. Local Classifier

In this section, we describe how to use attention to guide semi-supervised PDAC segmentation. The key to improving segmentation results under the SSL setting is how to generate reliable pseudo-labels on the large amount of training data without per-voxel annotations. The attention values for $\mathbf{Z}^u$ computed by Eq. 1 can be used as the per-voxel pseudo-labels after binarization. But, they are implicitly supervised by image-level annotation y^u. Consequently, they are obtained by searching over all spatial locations, and thus are coarse and noisy. Since many PDACs occupy only a small portion of pancreas regions, this issue becomes more severe.

Our basic strategy is to generate bag-level pseudo-labels to the data without per-voxel annotations, and then learn the local classifier based on these bag-level pseudo-labels by MIL. The bag-level pseudo-labels are obtained by separating the whole location lattice $\mathcal{X}$ into two bags (regions) according to the attention values: the PDAC bag $\mathcal{X}_f$ and the background bag $\mathcal{X}_b$. This strategy leads to a benefit: The correctness of the bag-level pseudo-labels is much easier to be guaranteed than per-voxel pseudo-labels, *e.g.*, a PDAC bag covers at least parts of PDAC masses and most of the instances in a background bag come from the background region (see Fig. 2).

1) Attention Based PDAC and Background Separation: $\mathcal{X}_f$ and $\mathcal{X}_b$ can be obtained by setting an attention threshold t_a: $\mathcal{X}_f=\{\mathbf{x}\in\mathcal{X}|p_{\mathbf{x}}\leq t_a\}$ and $\mathcal{X}_b=\{\mathbf{x}\in\mathcal{X}|p_{\mathbf{x}}>t_a\}$. Here, $p_{\mathbf{x}}=p_{\mathbf{x}}(\mathbf{Z}^u;\Theta,\mathbf{w}_a)$ is computed by Eq. 2. We omit these parameters for notational simplicity. Intuitively, the instances within each bag $\mathcal{X}_f$ or $\mathcal{X}_b$ should have similar attention values, while the instances from the two bags respectively should have different attention values. Thus, we determine the threshold t_a for each volume $\mathbf{Z}^u$ adaptively by minimizing

$$\ell_{\mathbb{C}}=\sum_{c\in\{f,b\}}\sum_{\mathbf{x}\in\mathcal{X}_c}|p_{\mathbf{x}}-\frac{1}{|\mathcal{X}_c|}\sum_{\mathbf{x}\in\mathcal{X}_c}p_{\mathbf{x}}|. \tag{9}$$

We realize this minimization by K-means clustering, since $\ell_{\mathbb{C}}$ is equivalent to a K-means clustering loss when $K=2$. Note that, this minimization is performed in the forward pass, and does not participate in the backward pass.

2) Learning Local Instance-Level Classifier by MIL: We treat both the PDAC region and the background region as small bags: $\mathbf{Z}_f^u$ and $\mathbf{Z}_b^u$, and express the bag vector representation $\hbar(\mathbf{Z}_c^u;\Theta)\in\mathbb{R}^C$ by the aggregation of the instance features:

$$\hbar(\mathbf{Z}_c^u;\Theta,\mathbf{w}_a)=\frac{1}{|\mathcal{X}_c|}\sum_{\mathbf{x}\in\mathcal{X}_c}\mathbf{f}_{\mathbf{x}}(\mathbf{Z}^u;\Theta,\mathbf{w}_a),c\in\{f,b\}. \tag{10}$$

Note that, $\mathcal{X}_c$ depends on parameter $\mathbf{w}_a$. Hence, the left hand side (LHS) of Eq. 10 also has parameter $\mathbf{w}_a$. Here we use average pooling, because we want the PDAC bags and background bags to be compact, and to progressively approach ground-truth segmentation maps.

The probability that a bag is a PDAC region is defined by:

$$P_c(\mathbf{Z}_c^u;\Theta,\mathbf{w}_a,\mathbf{w}_l)=\frac{1}{1+\exp\big(-\mathbf{w}_l^\top\hbar(\mathbf{Z}_c^u;\Theta,\mathbf{w}_a)\big)}, \tag{11}$$

where $\mathbf{w}_l\in\mathbb{R}^C$ is the parameter of the local classifier. Then an MIL loss is defined on $\mathbf{Z}^u$ to learn the local classifier:

$$\begin{aligned}\ell_{\mathbb{L}}^u(\mathbf{Z}^u;\boldsymbol{\Theta},\mathbf{w}_a,\mathbf{w}_l)=&-\big[\mathrm{y}_f^u\log P_c(\mathbf{Z}_f^u;\Theta,\mathbf{w}_a,\mathbf{w}_l)\\&+(1-\mathrm{y}_b^u)\log\big(1-P_c(\mathbf{Z}_b^u;\Theta,\mathbf{w}_a,\mathbf{w}_l)\big)\big].\end{aligned} \tag{12}$$

Note that, since $\mathbf{w}_l$ are linear coefficients, we have:

$$\mathbf{w}_l^\top\hbar(\mathbf{Z}_c^u;\Theta,\mathbf{w}_a)=\frac{1}{|\mathcal{X}_c|}\sum_{\mathbf{x}\in\mathcal{X}_c}\mathbf{w}_l^\top\mathbf{f}_{\mathbf{x}}(\mathbf{Z}^u;\Theta,\mathbf{w}_a), \tag{13}$$

Since the bag-level feature is the average of the instance-level features and the bag-level classifier is linear, applying the linear coefficients to the bag-level feature (the bag score) is equivalent to applying the linear coefficients to each instance-level feature (the instance score) then averaging. This implies that the local classifier can be directly applied to an instance. With $\mathbf{w}_l$, we have a unified function to compute the probability that an instance $\mathbf{f}_{\mathbf{x}}(\mathbf{Z};\Theta)$, no matter from $\mathbf{Z}^l$ or $\mathbf{Z}^u$, belongs to a PDAC region:

$$q_{\mathbf{x}}(\mathbf{Z};\Theta,\mathbf{w}_l)=\frac{1}{1+\exp\big(-\mathbf{w}_l^\top\mathbf{f}_{\mathbf{x}}(\mathbf{Z};\Theta)\big)}. \tag{14}$$

Eq. 11 and Eq. 14 show a unified function to compute the probabilities for both a bag and an instance by using the local classifier $\mathbf{w}_l$. Then, we can also define a loss function on the training data with per-voxel annotations to learn the local classifier:

$$\ell^l_{\mathbb{L}}(\mathbf{Z}^l, \mathbf{Y}^l; \Theta, \mathbf{w}_l) = -\sum_{\mathbf{x}\in\mathcal{X}} \big[y^l_{\mathbf{x}} \log q_{\mathbf{x}}(\mathbf{Z}^l; \Theta, \mathbf{w}_l) + (1 - y^l_{\mathbf{x}}) \log\big(1 - q_{\mathbf{x}}(\mathbf{Z}_l; \Theta, \mathbf{w}_l)\big)\big]. \tag{15}$$

Now, by combing Eq. 12 and Eq. 15, we can write down the loss function over the training set $\mathcal{D}$ to learn the local instance-level classifier:

$$L_{\mathbb{L}}(\mathcal{D}; \boldsymbol{\Theta}, \mathbf{w}_a, \mathbf{w}_l) = \frac{1}{N^l_p} \sum_{(\mathbf{Z}^l,\mathbf{Y}^l)\in\mathcal{D}^l} \mathrm{y}^l \ell^l_{\mathbb{L}}(\mathbf{Z}^l, \mathbf{Y}^l; \Theta, \mathbf{w}_l) + \frac{\lambda}{N^u_p} \sum_{(\mathbf{Z}^u,\mathrm{y}^u)\in\mathcal{D}^u} \mathrm{y}^u \ell^u_{\mathbb{L}}(\mathbf{Z}^u; \boldsymbol{\Theta}, \mathbf{w}_a, \mathbf{w}_l), \tag{16}$$

where $N^l_p = \sum_{(\mathbf{Z}^l,\mathbf{Y}^l)\in\mathcal{D}^l} \mathrm{y}^l$ and $N^u_p = \sum_{(\mathbf{Z}^u,\mathrm{y}^u)\in\mathcal{D}^u} \mathrm{y}^u$ are the numbers of the training data which have PDAC masses, with and without per-voxel annotations, respectively. Since the local classifier is used for PDAC segmentation, we learn it on the training data which have PDAC masses. The training data which have no PDAC mass are excluded for learning local classifier, as shown in Eq. 16. λ is a weight factor to balance the two terms of the right hand side (RHS) of Eq. 16. Intuitively, at the beginning of training, the pseudo PDAC region $\mathcal{X}_f$ and the background region $\mathcal{X}_b$ are very noisy, which makes the second term of the RHS of Eq. 16 unreliable. As the optimization proceeds, it progressively become reliable. Here, we make use of the attention guidance again, to design an adaptive weight factor to automatically reflect the reliability of the second term: $\lambda = \max_{\mathbf{x}\in\mathcal{X}} p_{\mathbf{x}}(\mathbf{Z}_u; \Theta, \mathbf{w}_a)$.

Since the explicitly-learned attention is a proxy of the local classifier, we add its loss function over the training set $\mathcal{D}$ to Eq. 16, which obtains the final loss function to learn the local classifier:

$$L_{\text{local}}(\mathcal{D}; \boldsymbol{\Theta}, \mathbf{w}_a, \mathbf{w}_l) = L_{\text{att}}(\mathcal{D}^l; \boldsymbol{\Theta}, \mathbf{w}_a) + L_{\mathbb{L}}(\mathcal{D}; \boldsymbol{\Theta}, \mathbf{w}_a, \mathbf{w}_l), \tag{17}$$

where

$$L_{\text{att}}(\mathcal{D}^l; \boldsymbol{\Theta}, \mathbf{w}_a) = \frac{1}{N^l_p} \sum_{(\mathbf{Z}^l,\mathbf{Y}^l)\in\mathcal{D}^l} \mathrm{y}^l \ell_{\text{att}}(\mathbf{Z}^l, \mathbf{Y}^l; \Theta, \mathbf{w}_a). \tag{18}$$

3) Overall Loss Function: Finally, we write down the overall loss function for training our IAG-Net:

$$L_{\text{IAG}}(\mathcal{D}; \boldsymbol{\Theta}, \mathbf{w}_a, \mathbf{w}_g, \mathbf{w}_l) = L_{\text{global}}(\mathcal{D}; \boldsymbol{\Theta}, \mathbf{w}_a, \mathbf{w}_g) + \beta L_{\text{local}}(\mathcal{D}; \boldsymbol{\Theta}, \mathbf{w}_a, \mathbf{w}_l), \tag{19}$$

where β is a trade-off parameter which balances the two terms (we set $\beta = 20$ in our implementation, which is not sensitive between [10, 30]). All parameters are jointly optimized during network training. The optimized parameters are obtained by

$$(\boldsymbol{\Theta}, \mathbf{w}_a, \mathbf{w}_g, \mathbf{w}_l)^* = \arg\min_{\boldsymbol{\Theta},\mathbf{w}_a,\mathbf{w}_g,\mathbf{w}_l} L_{\text{IAG}}(\mathcal{D}; \boldsymbol{\Theta}, \mathbf{w}_a, \mathbf{w}_g, \mathbf{w}_l). \tag{20}$$

Algorithm 1: The Training Process of IAG-Net

```
Input : Training set $\mathcal{D} = \mathcal{D}^l \bigcup \mathcal{D}^u$, where $\mathcal{D}^l = \{(\mathbf{Z}^l, \mathbf{Y}^l)\}_{l=1}^{L}$ and $\mathcal{D}^u = \{(\mathbf{Z}^u, \mathrm{y}^u)\}_{u=L+1}^{U}$;
        Max number of iterations $T$;
Output: Parameters $\boldsymbol{\Theta}^*$, $\mathbf{w}_a^*$, $\mathbf{w}_l^*$ and $\mathbf{w}_g^*$;
$t \leftarrow 0$;
Randomly initialize $\boldsymbol{\Theta}$, $\mathbf{w}_a$, $\mathbf{w}_l$ and $\mathbf{w}_g$;
repeat
    $t \leftarrow t + 1$;
    Randomly select a data sample $(\mathbf{Z}, \cdot)$ from $\mathcal{D}$
    if $(\mathbf{Z}, \cdot) \in \mathcal{D}^l$ then
        Compute $\ell_{\text{att}}(\mathbf{Z}^l, \mathbf{Y}^l; \Theta, \mathbf{w}_a)$ by Eq. 3 and
        $\ell^l_{\mathbb{L}}(\mathbf{Z}^l, \mathbf{Y}^l; \boldsymbol{\Theta}, \mathbf{w}_l)$ by Eq. 15;
    else
        Obtain $\mathcal{X}_f$ and $\mathcal{X}_b$ by minimizing Eq. 9
        Compute $\ell^u_{\mathbb{L}}(\mathbf{Z}^u; \boldsymbol{\Theta}, \mathbf{w}_a, \mathbf{w}_l)$ on $\{\mathcal{X}_f, \mathcal{X}_b\}$ by
        Eq. 12;
    end
    Compute $\ell_{\mathbb{G}}(\mathbf{Z}, \mathrm{y}; \Theta, \mathbf{w}_a, \mathbf{w}_g)$ by Eq. 6;
    Compute $L_{\text{IAG}}(\mathcal{D}; \boldsymbol{\Theta}, \mathbf{w}_a, \mathbf{w}_g, \mathbf{w}_l)$ by Eq. 19;
    Update $\boldsymbol{\Theta}$, $\mathbf{w}_a$, $\mathbf{w}_l$ and $\mathbf{w}_g$ by Gradient Descent;
until $t = T$;
Return: $(\boldsymbol{\Theta}, \mathbf{w}_a, \mathbf{w}_l, \mathbf{w}_g)^* \leftarrow (\boldsymbol{\Theta}, \mathbf{w}_a, \mathbf{w}_l, \mathbf{w}_g)$.
```

The overall training procedure is summarized in Algorithm 1. We also show the detailed architecture of our IAG-Net in Fig. 4. The backbone is not illustrated for simplicity.

D. Testing IAG-Net for PDAC Prediction

Given a testing image $\mathbf{Z}$, whether it has PDAC masses is determined by the global image-level classifier (Eq. 7):

$$\hat{\mathrm{y}} = \begin{cases} 1, & \text{if } P_g(\mathbf{Z}; \Theta^*, \mathbf{w}_a^*, \mathbf{w}_g^*) \geqslant 0.5; \\ 0, & \text{otherwise.} \end{cases} \tag{21}$$

If $\hat{\mathrm{y}} = 1$, then we further need to segment the PDAC masses out from $\mathbf{Z}$, *i.e.*, to predict $\hat{y}_{\mathbf{x}}$ at each location $\mathbf{x}$. For this task, we have the result of the local instance-level classifier (Eq. 14): $q_{\mathbf{x}}(\mathbf{Z}; \Theta^*, \mathbf{w}_l^*)$. Recall that, our learned attention guidance is inductive. It acts as a proxy of a local instance-level classifier (Eq. 2): $p_{\mathbf{x}}(\mathbf{Z}; \Theta^*, \mathbf{w}_a^*)$. We simply average these two results to determine the value of $\hat{y}_{\mathbf{x}}$:

$$\hat{y}_{\mathbf{x}} = \begin{cases} 1, & \text{if } p_{\mathbf{x}}(\mathbf{Z}; \Theta^*, \mathbf{w}_a^*) + q_{\mathbf{x}}(\mathbf{Z}; \Theta^*, \mathbf{w}_l^*) \geqslant 1.0; \\ 0, & \text{otherwise.} \end{cases} \tag{22}$$

IV. Experimental Results

In this section we describe the implementation details and compare IAG-Net with other competitors. We first evaluate our approach in a JHMI dataset which consists of both normal and PDAC cases, and then provide some diagnostic experiments for further analysis. After that, we apply our approach to segment pancreatic tumors in a public MSD challenge dataset under the semi-supervised setting.

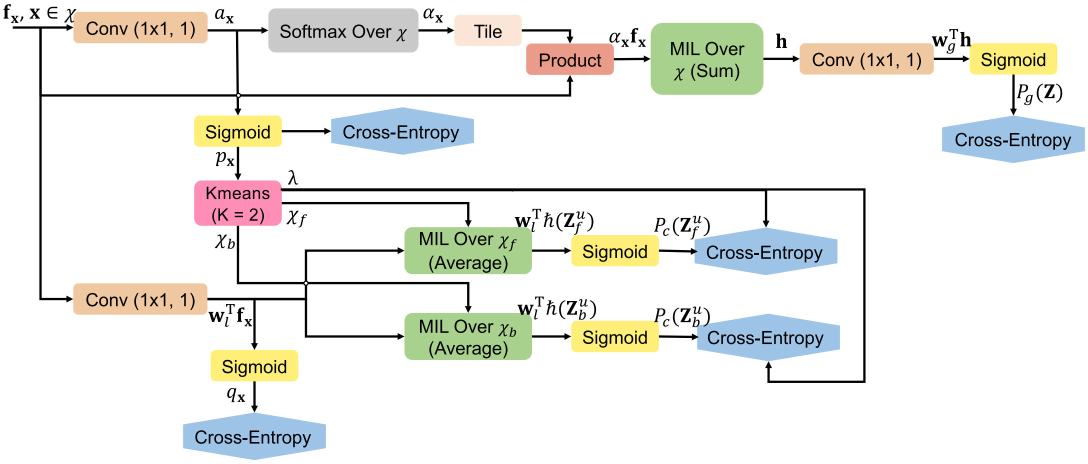

Fig. 4. The detailed architecture shown layer by layer of our IAG-Net. The backbone is not illustrated for simplicity.

A. Dataset

We evaluate IAG-Net in a JHMI dataset, which contains 400 CT images of normal cases and 400 biopsy-proven PDAC cases under an IRB (Institutional Review Board) approved protocol in Johns Hopkins Hospital as a part of the FELIX project for pancreatic cancer research. CT images have voxel spatial resolution of $([0.523 - 0.977] \times [0.523 - 0.977] \times 0.5) mm^3$. All CT scans are pancreas region of interest (RoI) of contrast enhanced images in the Venous phase. Each PDAC case has tumor annotation in biopsy-proven PDAC cases, annotated and verified by an experienced board-certified Abdominal Radiologist. We randomly partition the dataset into 4 equally-sized folds, using three of them for training, and the remaining one for testing. Unless otherwise specified, half of the biopsy-proven PDAC cases are randomly chosen to be given only voxel-level annotations during training. We also evaluate the generality of IAG-Net in a public MSD pancreas tumor dataset [41]. This dataset was comprised of patients undergoing resection of pancreatic masses (intraductal mucinous neoplasms, pancreatic neuroendocrine tumours, or PDAC).

B. Evaluation Metric

For PDAC segmentation, Dice-Sørensen similarity coefficient (DSC) over the whole pancreas RoI is computed, and mean DSC scores, standard deviation, max DSC and median DSC over all testing cases are reported. For normal/PDAC classification, sensitivity and specificity are calculated.

C. Implementation Details

For the data pre-processing, we truncate the raw intensity values within the range [-125, 350] HU and normalize the intensity values of each raw CT case to [0, 1] HU to decrease the data variance caused by the physical processes of medical images [42]. In the experiment, we test our IAG-Net based on different popular CNN backbones: VGG-Net [24] and 2D U-Net [25]. Given an input 3D volume $\mathbf{Z}$, we use the CNN backbone to compute a feature map of each slice, and then concatenate all the feature maps as the feature map $\mathbf{F}$ of the volume $\mathbf{Z}$. We choose the output of the last feature extraction layer in both backbones as the feature map $\mathbf{F}$. For VGG-Net, extracting feature maps from `conv6` will cause *out-of-memory* errors on an NVidia TITAN RTX GPU with 24GB of GDDR6 memory. So we use the output of `conv5` as our feature map $\mathbf{F}$, and $C = 512$. The per-voxel annotation is downsampled to match the dimension of $\mathbf{F}$. The predicted per-voxel label map is upsampled to the original size. For 2D U-Net, we use the output of the layer before the last as our feature map $\mathbf{F}$, and $C = 64$. The dimension of $\mathbf{F}$ matches with the input image.

Noted that feeding 2D slices from different viewing directions into 2D deep networks, and combining segmentation results from multi-view images as the final prediction leads to comparable performance to 3D deep networks in medical image segmentation [43], [44]. In addition, radiologists also view 3D scans slice-by-slice when they do annotation. So we simply adopt 2D deep networks rather than 3D networks. Our IAG-Net can be also built on 3D backbones, but this is out of the scope of this paper.

Given a 3D pancreas RoI, we train IAG-Nets on 2D views based on the normal vector directions of the sagittal (X), coronal (Y) and axial (Z) planes, respectively. When training an IAG-Net for one direction, *e.g.*, Z plane, in order to save memory, instead of feeding all slices in the whole 3D volume into the IAG-Net as one batch, we sample slices with an interval which is a random integer between $[1, 5]$, so that the size of the real input is $\hat{H} \times W \times L$ where $\hat{H} < H$. The same strategy is also applied to other directions. Following [19], [44], the final prediction, including classification and segmentation, is a combination of three 2D views. These strategies are applied for other competing methods, unless otherwise specified.

VGG-Net was pre-trained on ImageNet, and it was trained on the PascalVOC dataset to transfer the learned weights from classification networks to segmentation networks. U-Net was pre-trained on a separate in-house multi-organ segmentation dataset. These pre-trained models are used for the rest of the experiments. We set the initial learning rate to be 10^{-5} for VGG-Net and 10^{-4} for U-Net. Models are trained for $120,000$ iterations. We use exponential learning rate decay

TABLE I

PERFORMANCE COMPARISON (%) ON PDAC SEGMENTATION (DSC, MEAN ± STANDARD DEVIATION, MAX AND MEDIAN OF ALL CASES) AND CLASSIFICATION (SENSITIVITY AND SPECIFICITY). AS A REFERENCE, WE ALSO SHOW PERFORMANCES OF USING ALL THE TRAINING DATA IN A FULLY-SUPERVISED MANNER AS IAG-NET (FULLY-SUPERVISED). **BOLD** DENOTES THE BEST RESULTS OF SEMI-SUPERVISED METHODS. NUMBER OF PARAMETERS OF THE NETWORKS, SIGNIFICANT STATISTICAL IMPROVEMENTS OF IAG-NET VS. FCN-8S [45]/ U-NET [25], LI *et al.* [18], DMPCT [19], COLLABORATIVE LEARNING METHOD [39], AND STATISTICAL IMPROVEMENTS OF IAG-NET (FULLY-SUPERVISED) VS. IAG-NET ON PDAC SEGMENTATION ARE SHOWN

Backbone	Method	Mean DSC (Max, Median)	Sensitivity	Specificity	# Parameters	p-value
VGG-Net	Attention-based MIL [40]	–	99.00	**98.00**	14.78M	–
	FCN-8s [45]	50.02 ± 26.15 (89.75, 56.50)	–	–	134.27M	7.93×10^{-5}
	Li *et al.* [18]	41.31 ± 21.41 (82.80, 41.55)	99.00	96.75	14.72M	3.68×10^{-51}
	DMPCT [19]	49.24 ± 27.04 (90.89, 54.12)	–	–	134.27M	1.87×10^{-7}
	Collaborative [39]	52.52 ± 19.35 (85.88, 55.37)	98.25	96.75	21.02M	1.31×10^{-7}
	IAG-Net (Ours)	54.38 ± 18.77 (87.65, 57.00)	99.25	97.50	14.72M	–
	IAG-Net (fully-supervised)	55.45 ± 19.60 (88.36, 58.29)	99.25	97.75	14.72M	3.90×10^{-3}
U-Net	Attention-based MIL [40]	–	97.00	94.50	30.44M	–
	U-Net [25]	51.87 ± 25.94 (93.63, 56.52)	–	–	30.43M	1.22×10^{-32}
	Li *et al.* [18]	47.91 ± 26.13 (90.84, 51.73)	99.25	93.75	30.43M	7.70×10^{-45}
	DMPCT [19]	52.35 ± 26.38 (92.23, 56.69)	–	–	30.43M	9.61×10^{-30}
	Collaborative [39]	55.24 ± 24.96 (93.88, 60.94)	98.75	95.75	36.74M	3.70×10^{-18}
	IAG-Net (Ours)	**60.29** ± 21.60 (**94.04**, **64.37**)	**99.75**	96.50	30.43M	–
	IAG-Net (fully-supervised)	60.38 ± 23.83 (93.61, 66.93)	98.25	98.00	30.43M	6.16×10^{-1}

with $\gamma = 0.99$. The learning rate, maximum training iterations, and learning rate decay for training are the same for other competing methods unless otherwise specified.

D. Comparison Between IAG-Net and Other Methods

In this section, we conduct comparison between IAG-Net and four competitors: 1) attention-based MIL [40] 2) Li *et al.* [18], which is a unified framework to combine disease identification and localization of abnormalities, 3) a semi-supervised learning method for segmentation, *i.e.*, Deep Multi-Planar Co-Training (DMPCT) [19], and 4) a collaborative learning method for segmentation and classification under the semi-supervised setting [39]. Besides, we also compare with the segmentation backbones, *i.e.*, directly using FCN-8s [45] and U-Net for segmentation. Noted that the backbone of FCN-8s is VGG-Net, but it adopts additional strategies such as skip connections to enhance the segmentation performance.

Attention-based MIL [40] uses only image-level annotation during training. But, it shows the ability to do segmentation *i.e.*, there is a substantial matching between the heat map obtained from attention and the segmentation ground-truth. We use attention-based MIL as a baseline method. Following [40], we treat $a'_{\mathbf{x}} = (a_{\mathbf{x}} - \min(\mathbf{a}))/(\max(\mathbf{a}) - \min(\mathbf{a}))$ as the instance probability to obtain the PDAC segmentation results.

The work of Li *et al.* [18] is designed for disease identification and localization for 2D images, given part of the images with bounding boxes annotations and all images have disease identities. Since our task is to perform segmentation on 3D volumes, we customize Li's method to fit our setting: We replace the MIL xor operation used in Li's method by the commonly used MIL max pooling, since the number of instances of a 3D volume is 50× more than that of a 2D image, which makes MIL based on xor operation difficult to converge, even with the smoothing trick (*i.e.*, used in Li's method, normalize the patch score from $[0, 1]$ to $[0.98, 1]$). We also change the loss function in Li's method to the same cross-entropy loss as ours for PDAC segmentation and replace the backbone utilized in Li's method with our backbones.

In implementing DMPCT [19], We also adopt VGG-Net and U-Net as backbones. When using VGG-Net as its backbone, followed by [19], we also use the strategies in FCN-8s [45], such as skip connections and fusing coarse and fine feature maps [45] to improve its segmentation results [19]. We follow the same settings as reported in [19], *i.e.*, we set the learning rate to be 10^{-9} for FCN-8s. The teacher model and the student model are trained for 80,000 and 160,000 iterations, respectively for both FCN-8s and U-Net. For U-Net, we set the learning rate to be 10^{-8}. Followed by [19], models are trained and tested from multiple planes separately in a slice-by-slice manner. The final prediction is a combination of three 2D views.

To implement the collaborative learning method [39], we replace the lesion segmentation network with our backbones, and adopt the other network components (*e.g.*, lesion attention classification) as illustrated in [39]. Since the number of our class is 2, we slightly modify the network accordingly.

Results are shown in Table I. Attention-based MIL [40] achieves high classification accuracy with VGG-Net, but does not work for PDAC segmentation at all. More specifically, attention values are uniformly distributed over all locations. It is no surprise to observe such a phenomenon, since without any per-voxel supervision, it is hard to learn the attention for our dataset. Li *et al.* [18] achieves comparable classification results with our IAG-Net using VGG-Net as the backbone, but it performs worse in PDAC segmentation compared to ours. We achieve much better segmentation and classification

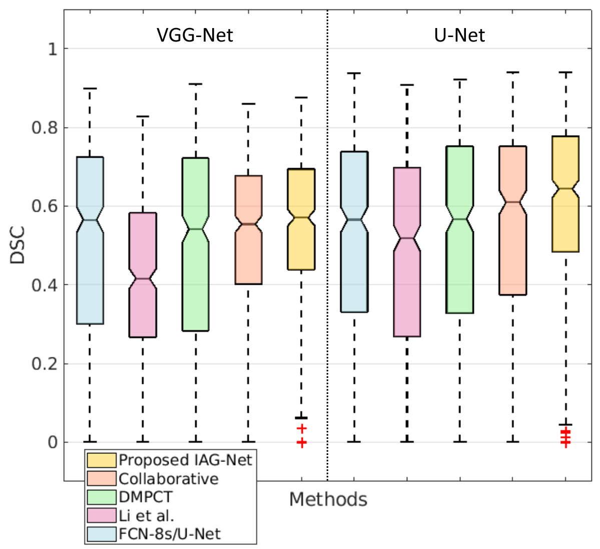

Fig. 5. Performance comparison (DSC) in box plots of PDAC segmentation. Proposed IAG-Net improves the overall mean and median DSC and also reduces the standard deviation.

results than Li's method when using U-Net as the backbone model. DMPCT [19] is designed only for segmentation. We can outperform DMPCT by 5.14% and 7.94% in terms of mean DSC with two backbones, respectively. Note that DMPCT mines consensus information from multiple planes, which is demonstrated to generate more reliable pseudo-labels than single-planar based method [14]. Since our current method learns attention-guided pseudo-labels for each plane separately, more performance gain is expected if consensus information from multiple planes can be extracted. Moreover, DMPCT with VGG-Net adopts additional strategies from FCN-8s [45] to enhance the segmentation results, Our results can be further improved by using these strategies. IAG-Net achieves superior performances than the collaborative learning method [39] in terms of both segmentation and classification tasks. We show the performance of a fully supervised IAG-Net in Table I, termed as IAG-Net (fully-supervised) for reference. The fully supervised IAG-Net is trained under the setting that all the training data are with per-voxel annotations, *i.e.*, $|\mathcal{D}^l| = 300$, $|\mathcal{D}^u| = 0$, and Eq (12) is omitted during training. The results show that the proposed IAG-Net can approach its fully-supervised version.

Fig. 5 shows comparison results of our IAG-Net, two backbone networks, Li *et al.* [18], DMPCT [19] and collaborative learning [39] by box plots. Besides, the p-values for testing significant difference between our IAG-Net and backbones, Li *et al.*, DMPCT and collaborative learning method for PDAC segmentation are shown in the last column of Table I. Our IAG-Net is significantly better than all competitors in PDAC segmentation. We compare parameter sizes among different methods. As shown in Table I, compared with other methods, our IAG-Net is effectively designed for PDAC prediction with a negligible increase in parameter sizes. We also illustrate PDAC segmentation results in Fig. 6 for qualitative comparison. We can see that compared with other methods, IAG-Net can output more accurate segmentation results, which are more robust to the complicated background.

We compare with two state-of-the-art consistency-related methods whose codes are publicly available [32], [46], achieving 57.22 ± 21.68 (89.53, 64.22) and 51.21 ± 24.11 (91.86, 55.24) in terms of DSC for PDAC segmentation, respectively. To run these methods, we use the same training models (backbone: DenseUNet [47] for [32] and ResNet-50 [48] for [46]) and parameters as reported in [32] and [46]. Models are trained and tested from multiple planes separately in a slice-by-slice manner. The final prediction is a combination of three 2D views.

Note that, [46] uses ResNet-50 as the backbone network, which is a stronger backbone than VGG-Net, as shown in [48]. We also test our IAG-Net with ResNet-50, and achieve 56.36 ± 18.37 (91.90, 58.77). This shows that IAG-Net outperforms [46] by a large margin (with 5.15% improvement in terms of mean DSC). Moreover, we adopt the cross-consistency training strategy in our IAG-Net ([32] takes advantages of both transformation consistency and self-ensembling, while [46] designs a stand-alone cross-consistency training strategy which can be directly integrated into our framework), *i.e.*, we add perturbations on the encoder's output, and add the auxiliary global classifier and the local classifier corresponding to each perturbation. More specifically, we use $K = 2$ for Con-Msk and Obj-Msk, $K = 2$ for I-VAT, and $K = 6$ for the rest of the perturbations, as suggested in [46]. With cross-consistency training, our IAG-Net-consistency can obtain 57.44 ± 23.23 (92.92, 63.33), whose performance is better than pure IAG-Net or consistency-based method [46]. This shows that the consistency-based methods are another direction for semi-supervised learning which are also demonstrated to be complementary with self-training methods in [37].

Last but not the least, we briefly show the results of applying 3D U-Net [49] for only PDAC segmentation as a reference. We set the initial learning rate to be 10^{-2}, and use exponential learning rate decay with $\gamma = 0.99$. During training, we randomly sample patches of a specified size (*i.e.*, 64). During testing, we employ the sliding window strategy to obtain the final predictions. The mean DSC (%, max, median) is 50.13 ± 26.75 (93.06, 56.64) with half of the PDAC cases with per-voxel annotation during training. Our IAG-Net performs much better than the 3D U-Net baseline for PDAC segmentation.

E. Ablation Study

We conduct ablation experiments to analyze the influence of different designs and components for IAG-Net.

(1) Ablation study on the joint learning framework.

We first conduct an ablation study on the joint framework, *i.e.*, to explore what if only the global or the local classifier is learned in IAG-Net. To train the global classifier only, we disable the supervision for the local classifier in IAG-Net; To train the local classifiers only, we disable the supervision for the global classifier in IAG-Net. During testing, the predicted image-level label $\hat{y}$ and the predicted instance-level label $\hat{y}_{\mathbf{x}}$ under both of the two ablations are also obtained by Eq. 21 and Eq. 22, respectively.

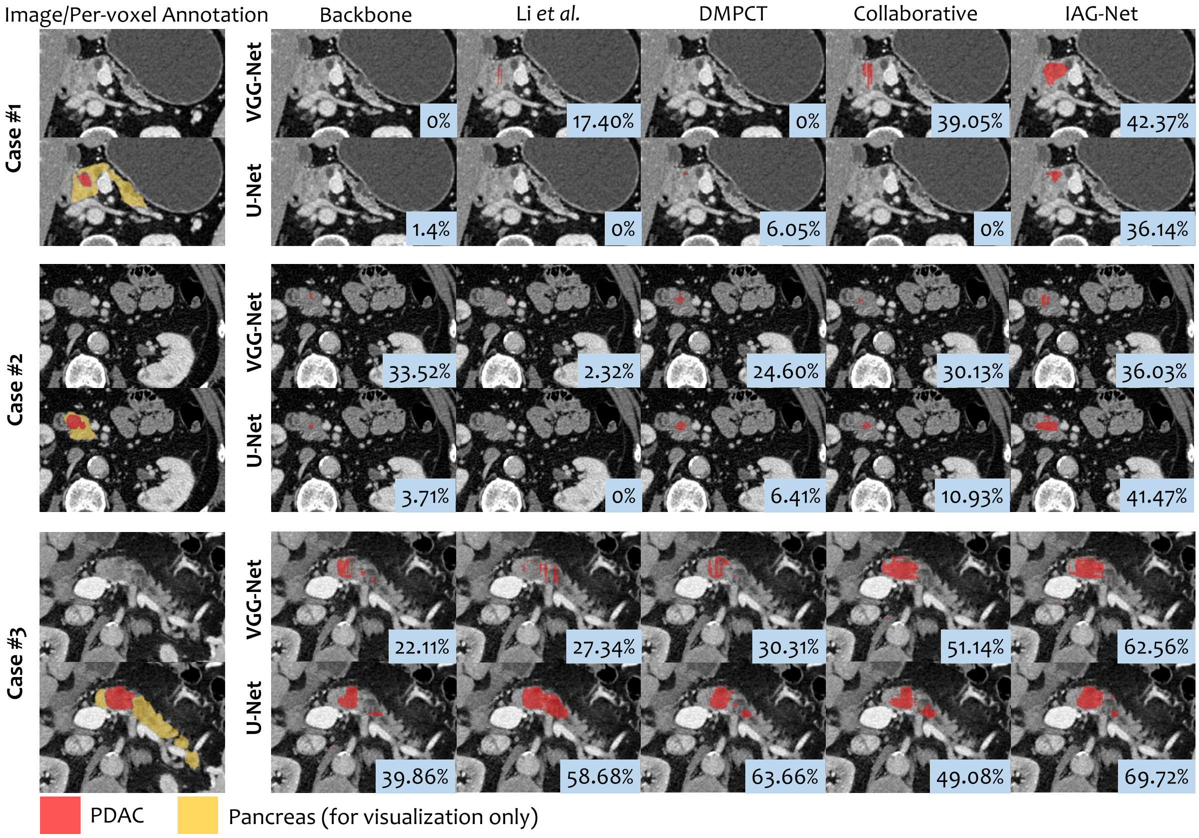

Fig. 6. Qualitative comparisons of PDAC segmentation. Three cases are shown in 2D slices in axial view. For each case, we show the results obtained by different methods (*i.e.*, a backbone network (VGG-Net [24] based FCN-8s [45], or U-Net [25]), Li *et al.* [18], DMPCT [19], collaborative learning method [39] and proposed IAG-Net) with the two backbone models, *i.e.*, VGG-Net (upper) and U-Net (lower). Per-voxel pancreas annotation is also shown as reference. Numbers on the bottom right are segmentation DSCs.

TABLE II
ABLATION ON THE JOINT LEARNING FRAMEWORK

Global classifier	Local classifier	Mean DSC (Max, Median)	Sens.	Spec.
	✓	51.19 ± 18.49 (86.80, 52.95)	98.00	29.50
✓		6.900 ± 6.100 (47.04, 5.210)	98.00	98.75
✓	✓	54.38 ± 18.77 (87.65, 57.00)	99.25	97.50

TABLE III
ABLATION STUDY ON THE GLOBAL CLASSIFIER

MIL method	Mean DSC (Max, Median)	Sens.	Spec.
Max	51.85 ± 20.63 (87.65, 55.21)	98.75	96.50
Average	34.12 ± 19.46 (83.25, 32.91)	86.75	99.75
Attention	54.38 ± 18.77 (87.65, 57.00)	99.25	97.50

Results with VGG-Net are shown in Table II. We observe that jointly training the global and the local classifiers achieves better PDAC segmentation results. Training with only the global classifier leads to comparable classification results to jointly training (*i.e.*, Sensitivity: 98.00%, Specificity: 98.75% vs. Sensitivity: 99.25%, Specificity: 97.50%). But surprisingly, it does not work for PDAC segmentation (instance probabilities are uniformly distributed over all locations). **The reason might be that the existence of PDAC makes the whole pancreas abnormal.** Without the local classifier, the global classifier alone cannot segment PDAC regions. Compared with jointly training, training with only local classifiers is confronted with a slight performance drop (-3.19%) for PDAC segmentation, and a significant specificity drop (-68%) for normal/PDAC classification. This is reasonable, since local classifiers aim at localizing the PDAC masses, which is more prone to identify a case as a PDAC case. These observations verify the importance of our joint learning for PDAC prediction.

(2) Ablation study on the global classifier.

To show the effectiveness of our global classifier by attention-guided MIL for PDAC segmentation, we replace it with max-pooling and average-pooling MIL. More specifically, after acquiring the attention $a_{\mathbf{x}}$ for instance $\mathbf{x}$, on one hand, $a_{\mathbf{x}}$ can be supervised by per-voxel annotation $y_{\mathbf{x}}$, if any. On the other hand, $a_{\mathbf{x}}$ is fed into a max/average pooling operation. The probability that an image $\mathbf{Z}$ contains PDAC masses is

$$P_g(\mathbf{Z}; \boldsymbol{\Theta}, \mathbf{w}_a) = \frac{1}{1+\exp\left(-\mathbf{g}(\mathbf{Z}; \boldsymbol{\Theta}, \mathbf{w}_a)\right)}, \tag{23}$$

where $\mathbf{g}(\mathbf{Z}; \boldsymbol{\Theta}, \mathbf{w}_a) = \max_{\mathbf{x}\in\mathcal{X}} a_{\mathbf{x}}(\mathbf{Z}; \boldsymbol{\Theta}, \mathbf{w}_a)$ or $\mathbf{g}(\mathbf{Z}; \boldsymbol{\Theta}, \mathbf{w}_a) = \frac{\sum_{\mathbf{x}\in\mathcal{X}} a_{\mathbf{x}}(\mathbf{Z}; \boldsymbol{\Theta}, \mathbf{w}_a)}{|\mathcal{X}|}$. Results with VGG-Net are shown in Table III. Attention-guided MIL achieves better prediction results than max/average-pooling MIL.

(3) Ablation study on the local classifier.

(i) What if we train it on per-voxel pseudo-labels? To train the local classifier on per-voxel pseudo-labels, we follow the typical teacher-student model [50] for self-training. We remove

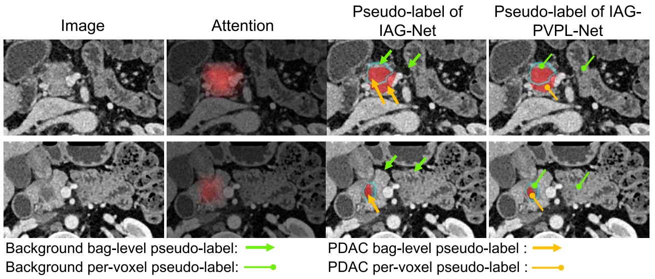
Image
Attention
Pseudo-label of IAG-Net
Pseudo-label of IAG-PVPL-Net
Background bag-level pseudo-label:
Background per-voxel pseudo-label:
PDAC bag-level pseudo-label :
PDAC per-voxel pseudo-label :

Fig. 7. Two examples of the attention and pseudo-labels of unlabeled training cases. Each row shows an input image, attention values of IAG-Net, bag-level pseudo-label of IAG-Net and per-voxel pseudo-label of IAG-PVPL-Net. The blue contours are the boundaries of the annotated PDAC segmentation masks (shown as references, which are not used during training). Since usually the misclassified instances in a bag are minor, the correctness of the bag-level pseudo-labels can be guaranteed after average pooling. But the per-voxel pseudo-labels of these instances are treated as noises for IAG-PVPL-Net.

the second stream of IAG-Net and train the explicitly-learned attention on the training data with per-voxel annotations as the teacher model. Then, we use the teacher model to generate per-voxel pseudo-labels for the training data without per-voxel annotation. After that, We re-train the explicitly-learned attention by including the generated per-voxel pseudo-labels as the student model. The first stream is still supervised by image-level annotations when training both teacher and student models. This method is termed as IAG-PVPL-Net (PVPL: Per-Voxel Pseudo-Labels). The classification and segmentation results are given by P_g and $p_{\mathbf{x}}$, respectively. Training the local classifier on per-voxel pseudo-labels obtains 51.74 ± 20.40 (87.16, 54.96) for mean (% max, median) DSC, 98.75% sensitivity and 97.25% specificity, leading to around 3% DSC drop compared with IAG-Net, while it requires much more training iterations.

To better understand why the bag-level pseudo-labels are better than the per-voxel pseudo-labels, some of the pseudo-labels generated by IAG-Net and IAG-PVPL-Net are shown in Fig. 6. Bag-level pseudo-labels obtained by average pooling impose soft constraints on instances whose scores are low. For example, among 100 instances, 95 instances are scored 1.0, and the rest 5 are scored 0.0. Then the loss w.r.t a bag-level pseudo-label is $-\log(0.95)$. But per-voxel pseudo-labels impose strict constraints on each instance, *i.e.*, the losses of the instances with very low scores w.r.t. per-voxel pseudo-labels are nearly ∞. Consequently, the bag-level pseudo-labels are better for training.

(ii) Is the attention-based weight factor λ better than a constant weight factor? We set the weight factor λ to a default constant: $\lambda = 1.0$. The result is: 52.69± 18.11 Mean, 86.39 max and 54.18 median DSC (%) for PDAC segmentation, and 98.5% sensitivity and 97.75% specificity for normal/PDAC classification. Compared with the attention-based weight factor, it leads to 1.69% mean DSC drop.

(iii) What if we do not train the local classifier on data without per-voxel annotations? We set $|\mathcal{D}^u| = 0$ during training. The PDAC segmentation results based on both VGG-Net and U-Net are shown in Table IV. Training the local classifier on $\mathcal{D}^u$ can always boost the PDAC segmentation performance (+4.55% for VGG-Net and +5.24% for U-Net).

TABLE IV
ABLATION ON THE LOCAL CLASSIFIER

Backbone	$\|\mathcal{D}^l\|$	$\|\mathcal{D}^u\|$	Mean DSC (Max, Median)	Sens.	Spec.
VGG-Net	150	0	49.83 ± 20.59 (86.80, 52.77)	98.75	98.00
	150	150	54.38 ± 18.77 (87.65, 57.00)	99.25	97.50
U-Net	150	0	55.05 ± 25.13 (94.15, 59.81)	98.50	97.25
	150	150	60.29 ± 21.60 (94.04, 64.37)	99.75	96.50

TABLE V
ABLATION ON THE LOCAL CLASSIFIER BY VARYING $|\mathcal{D}^l|$ AND $|\mathcal{D}^u|$ FOR EACH FOLD WITH VGG-NET

$\|\mathcal{D}^l\|$	$\|\mathcal{D}^u\|$	Mean DSC (Max, Median)	Sens.	Spec.
50	0	44.31 ± 21.29 (82.37, 46.46)	98.25	97.75
50	250	49.54 ± 22.87 (85.63, 52.54)	98.50	97.25
150	0	49.83 ± 20.59 (86.80, 52.77)	98.75	98.00
150	150	54.38 ± 18.77 (87.65, 57.00)	99.25	97.50
300	0	55.45 ± 19.60 (88.36, 58.29)	99.25	97.75

(iv) How does the PDAC segmentation performance change by varying the ratio between $|\mathcal{D}^l|$ and $|\mathcal{D}^u|$? Since each fold of our PDAC dataset has 300 PDAC cases, $|\mathcal{D}^l| + |\mathcal{D}^u| \leq 300$. We train our IAG-Net by varying the ratio between $|\mathcal{D}^l|$ and $|\mathcal{D}^u|$. Table V shows the PDAC segmentation performance of our IAG-Net is improved as $|\mathcal{D}^l|$ is increased, which is not surprising. But we also observe that training the local classifier on $\mathcal{D}^u$ always significantly boosts the PDAC segmentation performance, which is even approaching the result obtained by fully-supervised segmentation (the last row).

(v) Does the attention guidance ($p_{\mathbf{x}}$) itself have a good segmentation ability in IAG-Net? Using $p_{\mathbf{x}}$ or $q_{\mathbf{x}}$ alone in Eq. 22 leads to 54.28 ± 19.04 or 54.38 ± 18.53 mean DSC (%), which is comparable to the result by combining $p_{\mathbf{x}}$ and $q_{\mathbf{x}}$ (Eq. 22). This shows that the attention guidance itself has a good segmentation ability.

F. Discussions

For the early stage of a PDAC case, as the PDAC mass is subtle, obtaining an accurate contour for PDAC segmentation is difficult. But, the pancreas may have some abnormal changes, *e.g.*, duct dilation or texture abnormality. PDAC segmentation is a much more challenging task than normal/PDAC classification. This may be the reason that although PDAC segmentation results are not very high by using even state-of-the-art methods, we can still achieve good classification results.

Our IAG-Net can be applied to multi-class segmentation, *i.e.*, to segment the entire pancreas and PDAC simultaneously. Let us define an attention map as a matrix consisting of the attention values of all the instances on the feature map $\mathbf{F}(\mathbf{Z}; \boldsymbol{\Theta})$. The key to the extension is to explicitly learn two attention maps under the supervision of the per-voxel annotations for both the pancreas and the PDAC, respectively. How to explicitly learn an attention map is shown in Sec. III-A. Then the global classifier which targets normal/PDAC classification, is also trained based on the attention map of the PDAC, as the same as Sec. III-B. For the local classifier, we train two

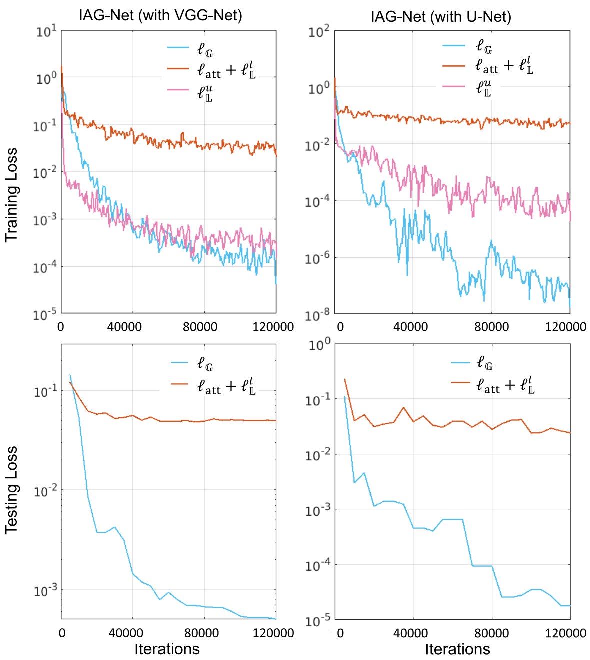

Fig. 8. Training and testing loss curves of IAG-Net with VGG-Net and U-Net as backbones.

"one-vs-all" classifiers for semi-supervised pancreas segmentation and semi-supervised PDAC segmentation, based on the attention maps of the pancreas and the PDAC, respectively. The training process of each of the two "one-vs-all" classifiers is the same as the training process shown in Sec. III-C. The loss function for the local classifier is the sum of the losses of the two "one-vs-all" classifiers.

Finally, Fig. 8 shows training and testing curves of the image-level classification loss $\ell_{\mathbb{G}}$ (Eq. 6), instance-level segmentation loss for training data with per-voxel annotations ℓ_{att} + $\ell^l_{\mathbb{L}}$ (Eq. 3 + Eq. 15), and the instance-level segmentation loss for training data without per-voxel annotations $\ell^u_{\mathbb{L}}$ (Eq. 12). Training losses are obtained via mini-batches every 20 iterations, and testing losses are acquired from the whole testing set every 5000 iterations. There is no $\ell^u_{\mathbb{L}}$ loss for testing. We observe that all losses converge after a certain number of iterations, and the testing losses decrease accordingly with the decrease of the training losses. The Y-axis in the plots are shown in the log space, so the loss values are very small when approaching 120,000 iterations.

G. IAG-Net on Semi-Supervised Tumor Segmentation

To verify the generality of IAG-Net on semi-supervised pancreatic tumor segmentation, we test it on the pancreas tumor segmentation dataset in MSD challenge [41]. There are two targets in the pancreas tumor dataset: pancreas and tumor. Here we only focus on tumor segmentation, which consists of multiple types of pancreatic tumors. 282 labeled data are released for training and validation. Noted that in this dataset, there are no normal cases, thus we exclude the global classifier in our IAG-Net for semi-supervised tumor segmentation only.

We randomly partition the dataset into 4 folds. The numbers of the cases belonging to each fold are 70, 70, 71 and 71. Following the same setting as we used for the JHMI dataset, we randomly choose half of the training data as the data without per-voxel annotation during training, which leaves only 35 cases with per-voxel annotation for each fold. The 2D U-Net is adopted as the backbone network unless otherwise specified. We train and test on the axial view of each case. We compare our IAG-Net with other competitors, and the results are shown in Table VI. IAG-Net outperforms other methods by a large margin. IAG-Net (fully-supervised) in Table VI is also shown as a reference for comparison.

TABLE VI
COMPARISON ON PANCREAS TUMOR DATASET IN MSD CHALLENGE

Method	Mean DSC (Max, Median)
Li *et al.* [18]	19.04 ± 24.48 (85.06, 3.17)
DMPCT [19]	25.60 ± 25.20 (89.32, 23.75)
Collaborative [39]	22.42 ± 24.86 (87.67, 13.17)
Consistency-based [46]	24.63 ± 27.10 (90.37, 15.12)
Consistency-self-ensembling [32]	30.01 ± 27.55 (90.07, 28.40)
IAG-Net (Ours)	32.49 ± 27.82 (94.08, 31.41)
IAG-Net (fully-supervised)	33.91 ± 28.21 (90.83, 34.25)

TABLE VII
COMPARISON ON PANCREAS TUMOR DATASET IN MSD CHALLENGE BY VARYING $|\mathcal{D}^l|$ AND $|\mathcal{D}^u|$. EACH FOLD HAS 70 OR 71 CASES. $|\mathcal{D}^l| + |\mathcal{D}^u| \leq 70/71$

$\lvert\mathcal{D}^l\rvert$	$\lvert\mathcal{D}^u\rvert$	Mean DSC (Max, Median)
35	0	30.65 ± 27.44 (92.37, 29.60)
35	35/36	32.49 ± 27.82 (94.08, 31.41)
70/71	0	33.91 ± 28.21 (90.83, 34.25)

We also vary the ratio between $|\mathcal{D}^l|$ and $|\mathcal{D}^u|$. Results are summarized in Table VII, which show that our IAG-Net can improve the tumor segmentation result by leveraging the training data without per-voxel annotations (around 2% accuracy gain) and even surpasses the fully-supervised method in terms of max DSC.

V. CONCLUSION

This paper addresses the problem of PDAC prediction *i.e.*, normal/PDAC classification and PDAC segmentation under the partially supervised setting. We present an Inductive Attention Guidance (IAG) strategy for learning a global image-level classifier for normal/PDAC segmentation and a local instance-level classifier for semi-supervised PDAC segmentation, which enjoys the advantages of bridging the MIL-based global and local classifiers. We showed empirically on the JHMI dataset the superiority of the proposed IAG-Net for PDAC prediction, which is helpful to computer-assisted clinical diagnoses. Additionally, we verified the generality of IAG-Net on the pancreas tumor segmentation dataset in MSD challenge.

REFERENCES

[1] Y. Zhou *et al.*, "Hyper-pairing network for multi-phase pancreatic ductal adenocarcinoma segmentation," in *Proc. MICCAI*, 2019, pp. 155–163.

[2] H. R. Roth *et al.*, "DeeporGAN: Multi-level deep convolutional networks for automated pancreas segmentation," in *Proc. MICCAI*, 2015, pp. 556–564.

[3] H. Roth, L. Lu, A. Farag, A. Sohn, and R. M. Summers, "Spatial aggregation of holistically-nested networks for automated pancreas segmentation," in *Proc. MICCAI*, 2016, pp. 451–459.

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

[5] A.-A.-Z. Imran, A. Hatamizadeh, S. P. Ananth, X. Ding, N. Tajbakhsh, and D. Terzopoulos, "Fast and automatic segmentation of pulmonary lobes from chest CT using a progressive dense V-network," *Comput. Methods Biomechanics Biomed. Eng., Imag. Visualizat.*, vol. 8, no. 5, pp. 509–518, Sep. 2020.

[6] K. Yan, X. Wang, L. Lu, and R. M. Summers, "Deeplesion: Automated deep mining, categorization and detection of significant radiology image findings using large-scale clinical lesion annotations," *J. Med. Imag.*, vol. 5, no. 3, 2018, Art. no. 036501.

[7] 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*, 2016, pp. 149–157.

[8] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 44–53, Jan. 2018.

[9] X. Zhu, "Semi-supervised learning with graphs," Ph.D. dissertation, School Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, USA, 2005.

[10] A. L. Yuille and C. Liu, "Deep nets: What have they ever done for vision?" *Int. J. Comput. Vis.*, Nov. 2020.

[11] Y. Xu, J.-Y. Zhu, E. Chang, and Z. Tu, "Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 964–971.

[12] L. Yao, J. Prosky, E. Poblenz, B. Covington, and K. Lyman, "Weakly supervised medical diagnosis and localization from multiple resolutions," 2018, *arXiv:1803.07703*. [Online]. Available: http://arxiv.org/abs/1803.07703

[13] H. D. Couture, J. S. Marron, C. M. Perou, M. A. Troester, and M. Niethammer, "Multiple instance learning for heterogeneous images: Training a CNN for histopathology," in *Proc. MICCAI*, 2018, pp. 254–262.

[14] W. Bai *et al.*, "Semi-supervised learning for network-based cardiac MR image segmentation," in *Proc. MICCAI*, 2017, pp. 253–260.

[15] G. Bortsova, F. Dubost, L. Hogeweg, I. Katramados, and M. D. Bruijne, "Semi-supervised medical image segmentation via learning consistency under transformations," in *Proc. MICCAI*, 2019, pp. 810–818.

[16] D. Wang, Y. Zhang, K. Zhang, and L. Wang, "FocalMix: Semi-supervised learning for 3D medical image detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3951–3960.

[17] Y. Zhou *et al.*, "Prior-aware neural network for partially-supervised multi-organ segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10672–10681.

[18] Z. Li *et al.*, "Thoracic disease identification and localization with limited supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 8290–8299.

[19] Y. Zhou *et al.*, "Semi-supervised 3D abdominal multi-organ segmentation via deep multi-planar co-training," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 121–140.

[20] H. Scudder, "Probability of error of some adaptive pattern-recognition machines," *IEEE Trans. Inf. Theory*, vol. 11, no. 3, pp. 363–371, Jul. 1965.

[21] S. Y. Shin, S. Lee, I. D. Yun, S. M. Kim, and K. M. Lee, "Joint weakly and semi-supervised deep learning for localization and classification of masses in breast ultrasound images," *IEEE Trans. Med. Imag.*, vol. 38, no. 3, pp. 762–774, Mar. 2019.

[22] T. Leung, Y. Song, and J. Zhang, "Handling label noise in video classification via multiple instance learning," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 2056–2063.

[23] A. Sarkar and G. Haffari, "Tutorial on inductive semi-supervised learning methods: With applicability to natural language processing," in *Proc. HLT-NAACL*, 2006, pp. 307–308.

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

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.

[26] X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning," *Synth. Lect. Artif. Intell. Mach. Learn.*, vol. 3, no. 1, pp. 1–130, 2009.

[27] P. Tang, X. Wang, X. Bai, and W. Liu, "Multiple instance detection network with online instance classifier refinement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2843–2851.

[28] P. Tang *et al.*, "PCL: Proposal cluster learning for weakly supervised object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 176–191, Jan. 2020.

[29] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille, "Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1742–1750.

[30] R. G. Cinbis, J. Verbeek, and C. Schmid, "Weakly supervised object localization with multi-fold multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 189–203, Jan. 2017.

[31] H. R. Roth *et al.*, "An application of cascaded 3D fully convolutional networks for medical image segmentation," *Computerized Med. Imag. Graph.*, vol. 66, pp. 90–99, Jun. 2018.

[32] X. Li, L. Yu, H. Chen, C. Fu, and P. Heng, "Transformation-consistent self-ensembling model for semisupervised medical image segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 523–534, Feb. 2021.

[33] G. Fotedar, N. Tajbakhsh, S. P. Ananth, and X. Ding, "Extreme consistency: Overcoming annotation scarcity and domain shifts," in *Proc. MICCAI*, 2020, pp. 699–709.

[34] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. 11th Annu. Conf. Comput. Learn. Theory (COLT)*, 1998, pp. 92–100.

[35] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, nos. 1–2, pp. 31–71, Jan. 1997.

[36] Q. Liu, L. Yu, L. Luo, Q. Dou, and P. A. Heng, "Semi-supervised medical image classification with relation-driven self-ensembling model," *IEEE Trans. Med. Imag.*, vol. 39, no. 11, pp. 3429–3440, Nov. 2020.

[37] P. Tang, C. Ramaiah, Y. Wang, R. Xu, and C. Xiong, "Proposal learning for semi-supervised object detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jun. 2021, pp. 2291–2301.

[38] J. Liu, G. Zhao, Y. Fei, M. Zhang, Y. Wang, and Y. Yu, "Align, attend and locate: Chest X-ray diagnosis via contrast induced attention network with limited supervision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10632–10641.

[39] Y. Zhou *et al.*, "Collaborative learning of semi-supervised segmentation and classification for medical images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2079–2088.

[40] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 2127–2136.

[41] A. L. Simpson *et al.*, "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," 2019, *arXiv:1803.07703*. https://arxiv.org/abs/1803.07703

[42] Z. Zhu, Y. Xia, W. Shen, E. K. Fishman, and A. L. Yuille, "A 3D coarse-to-fine framework for volumetric medical image segmentation," in *Proc. 3DV*, 2018, pp. 682–690.

[43] Y. Wang, Y. Zhou, W. Shen, S. Park, E. K. Fishman, and A. L. Yuille, "Abdominal multi-organ segmentation with organ-attention networks and statistical fusion," *Med. Image Anal.*, vol. 55, pp. 88–102, Jul. 2019.

[44] L. Xie, Q. Yu, Y. Zhou, Y. Wang, E. K. Fishman, and A. L. Yuille, "Recurrent saliency transformation network for tiny target segmentation in abdominal CT scans," *IEEE Trans. Med. Imag.*, vol. 39, no. 2, pp. 514–525, Feb. 2020.

[45] 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.

[46] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12674–12684.

[47] 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.

[48] 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.

[49] O. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. MICCAI*, 2016, pp. 424–432.

[50] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. ICML Workshop*, 2013, pp. 1–6.