Multi-Scale Coarse-to-Fine Segmentation for Screening Pancreatic Ductal Adenocarcinoma

Zhuotun Zhu[*,1], Yingda Xia[*,1],
Lingxi Xie[1], Elliot K. Fishman[2], Alan L. Yuille[1]

[1]The Johns Hopkins University, Baltimore, MD 21218, USA
{zhuotun,philyingdaxia,198808xc,alan.l.yuille}@gmail.com
[2]The Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA
efishman@jhmi.edu

Abstract. We propose an intuitive approach of detecting pancreatic ductal adenocarcinoma (PDAC), the most common type of pancreatic cancer, by checking abdominal CT scans. Our idea is named **multi-scale segmentation-for-classification**, which classifies volumes by checking if at least a sufficient number of voxels is segmented as tumors, by which we can provide radiologists with tumor locations. In order to deal with tumors with different scales, we train and test our volumetric segmentation networks with multi-scale inputs in a coarse-to-fine flowchart. A post-processing module is used to filter out outliers and reduce false alarms. We collect a new dataset containing 439 CT scans, in which 136 cases were diagnosed with PDAC and 303 cases are normal, which is the largest set for PDAC tumors to the best of our knowledge. To offer the best trade-off between sensitivity and specificity, our proposed framework reports a sensitivity of 94.1% at a specificity of 98.5%, which demonstrates the potential to make a clinical impact.

Keywords: PDAC, Pancreas Segmentation, CT Scan.

1 Introduction

Pancreatic cancer is one of the most dangerous killers to human lives, causing more than 330,000 deaths globally in 2014 [11]. Pancreatic ductal adenocarcinoma (PDAC) is the most common type of pancreatic cancer, accounting for about 85% of cancer cases. In early stages, this disease often has few symptoms and is very difficult to discover. By the time of diagnosis, the cancer has often spread to other parts of the body, leading to a very poor prognosis (*e.g.*, a five-year survival rate of 5% [11]). But, for cases diagnosed early, the survival rate rises to about 20% [7]. Hence, it is very important to study the possibility of detecting PDAC in common examinations, *e.g.*, the abdominal CT scan.

The early diagnosis of pancreatic cancer requires much expertise in reading the scanned images and making decisions, but the increasing number of cases

* The first two authors equally contributed to the work.

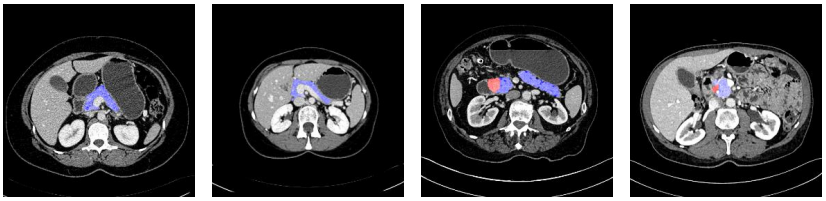

Fig. 1. Examples of normal and abnormal (PDAC) pancreases (best viewed in color). Blue and red region mark the normal pancreas and tumor regions, respectively.

makes it impossible for a limited number of experienced radiologists to check all CT scans manually. Therefore, an artificial intelligence system for this purpose is in need. In particular, the radiologists in our team are interested in a system working on abdominal CT scans, which filters out a large fraction of normal cases, but preserves almost all abnormal cases for further investigation. To the best of our knowledge, there is no existing work on this task.

With the development of deep learning [5], it is possible to construct a system which learns from professional knowledge in data annotation, and apply it to helping doctors in various clinical purposes. The pancreas is one of the most challenging organs in CT segmentation [9]. The difficulty mainly lies in its irregular shape and low contrast around the boundary. Powered by the recent progress in deep learning for 2D [1][8] and 3D [6][12] image segmentation, researchers designed various approaches [10][15] towards accurate pancreas segmentation. In the pathological cases, the morphology of the pancreas can be largely impacted by the difference in the pancreatic cancer stage [13][14].

Our work is aimed to detect PDAC from a mixture of normal and abnormal CT scans. This is not a simple classification since radiologists also want to know the location of PDAC, we suggest a solution named **segmentation-for-classification**, which trains segmentation models and uses their outputs for classification. To deal with tumors of various sizes (Fig. 1), we adopt a segmentation network with multiple input scales, *i.e.*, 64^3, 32^3 and 16^3 volumes. But, voting that small input regions lead to a high false alarm rate, we adopt a **coarse-to-fine** testing strategy, which uses the 64^3 network for a coarse scan, and then the 32^3&16^3 networks inside the bounding box to detect small tumors that are possibly ignored in the previous stage. A non-parameterized post-processing algorithm is designed to remove outliers.

Our contributions are three folds: 1) we voxelwisely annotate an abdominal CT dataset with 439 cases in total, in which 136 cases are diagnosed with PDAC while the remaining 303 cases are normal, which is currently the largest PDAC dataset to the best of our knowledge; 2) we adopt a **multi-scale segmentation-for-classification** framework to conduct an **interpretable** abnormality detection, which provides radiologists with suspicious regions for further diagnosis; 3) our framework achieves a sensitivity of 94.1% at a specificity of 98.5%, which shows a promising direction to make a potential significant clinical impact.

2 The Segmentation-for-Classification Approach

2.1 The Overall Framework

Let a dataset be $\mathbf{S} = \{(\mathbf{X}_1, y_1^\star), \ldots, (\mathbf{X}_N, y_N^\star)\}$, where N is the number of CT scans, $\mathbf{X}_n \in \mathbb{R}^{W_n \times H_n \times L_n}$ is the 3D volume with each element indicating the Hounsfield unit (HU) of a voxel, and $y_n \in \{0, 1\}$ is the label (0 for a normal case, 1 for an abnormal case). Throughout this paper, by *abnormal* we refer to the cases diagnosed as PDAC. The goal is to design a model $\mathbb{M} : y = f(\mathbf{X})$ to predict the label for each testing volume. We evaluate our approach by ranking all volumes by the probability of being a PDAC, computing the sensitivity and specificity at a given threshold, and plotting the ROC curve indicating the relationship between the sensitivity and specificity at different thresholds. For clinical purposes, we shall guarantee a high sensitivity with a reasonable specificity.

Although some previous work suggested to classify CT or MRI volumes directly using 3D networks [3][4], we argue that a better solution is to perform tumor segmentation at the same time of classification. This makes the classification results **interpretable** by segmentation cues, by which radiologists can take a further investigation of the suspicious abnormal regions. In addition, this integrates voxel-wise annotations into the classification model as deep supervision, so that the entire network is better trained [14]. Therefore, we propose a two-stage framework named *segmentation-for-classification*, in which a segmentation stage first extracts voxel-wise cues from the input CT scan, and a classification stage follows to summarize this information into the final prediction. Our multi-scale segmentation strategy is different from [15], which applied another network of the same scale in the fine stage. **Tumor detection requires multiple scales.**

Mathematically, let each training data be augmented by a segmentation mask $\mathbf{M}_n^\star$ of the same dimensionality as $\mathbf{X}$, so that $m_{n,i}^\star \in \{0, 1, 2\}$ indicates the category of the i-th voxel, *i.e.*, in the tumor ($m_{n,i} = 2$), outside the tumor but inside the pancreas ($m_{n,i} = 1$), or outside the pancreas ($m_{n,i} = 0$). Note that the tumor voxel set is a subset of the pancreas voxel set. The segmentation module is a high-dimensional function $\mathbf{M} = \mathbf{s}(\mathbf{X})$, which is implemented by a deep encoder-decoder network. The classification module is a binary function $y = c(\mathbf{M})$. The overall framework is thus written as:

$$y = f(\mathbf{X}) = c \circ \mathbf{s}(\mathbf{X}). \tag{1}$$

2.2 Training: Multi-Scale Deeply-Supervised Segmentation

We start with describing the segmentation stage. The tumor region in a pancreas, as shown in Fig. 1, can vary in scale, appearance and geometric properties. In particular, the largest tumor in our dataset occupies over one million voxels, but the smallest one has only thousands. This motivates us to train multi-scale networks to deal with such a large variation in scale.

In practice, we train three networks, taking input volumes of 64^3, 32^3 and 16^3 voxels, respectively. Each segmentation network follows an encoder-decoder

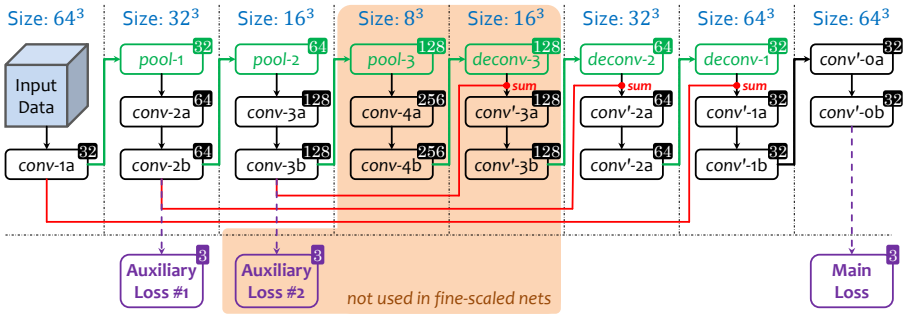

Fig. 2. The architecture of our segmentation backbone (best viewed in color). Each rectangle is a layer, green arrows indicate operations changing spatial resolution, and red arrows mean residual connections. We illustrate the situation when the input volume size is 64^3. If it is 32^3 or 16^3, all volumes are shrunk accordingly (to $1/2^3$ or $1/4^3$ of the displayed size). The number at the upper-right corner of each cube is the number of channels. Each convolution uses $3\times3\times3$ kernels with 1 as stride, each pooling $2\times2\times2$ with 2 as stride, and each deconvolution $4\times4\times4$ with 2 as stride. The weight ratio for auxiliary losses #1, #2 and main loss is $1:2:5$ for the 64^3 network, and $1:3$ for the auxiliary loss #1 and the main loss for the 32^3 and 16^3 networks.

flowchart shown in Fig. 2.2. It has a series of convolutional layers to learn 3D patterns from training data. Down-sampling and up-sampling are implemented by max pooling and deconvolutional layers, respectively. Following [15], we introduce deep supervision in the training process, which is implemented by adding several auxiliary losses to intermediate layers, which delivers better performance for the normal and cystic pancreas segmentation in [15]. Deep supervision is considered as a way of incorporating multi-stage visual cues, which constrains intermediate layers and improves the stability of training deep networks. Multi-scale segmentation is complementary to deep supervision, which aims at capturing visual patterns of various scales. As can be seen in experiments, multi-scale segmentation can take advantage of different scales, *i.e.*, a large network produces a high specificity, and a small network gives a high sensitivity.

The training process starts with sampling patches of a specified size. Since the pancreas and the tumor only occupy a small fraction of the entire volume, a random sampling strategy may lead to that only few patches contain pancreas or tumor voxels, and thus the segmentation models are biased towards the background class. To deal with the issue, we sample lots of foreground patches for training the 32^3 and 16^3 networks. We first compute the region-of-interest (ROI) by padding a 32-voxel margin around the minimal 3D bounding box covering the entire pancreas. Within it, we categorize the randomly sampled patches into three types (*i.e.*, *background*, *tumor* and *pancreas*) according to the fraction of pancreas and tumor voxels, and make the numbers of training patches of these types o be approximately the same. Data augmentation is performed by randomly flipping patches and rotating by 90°, 180° and 270° over three axes.

We use the same configuration for training these networks. The base learning rate is 0.01 and decayed polynomially (the power is 0.9) in a total of 80,000 iterations (the mini-batch size is 16, 32 and 128 for 64^3, 32^3 and 16^3, respectively). The weight decay and momentum are set to be 0.0005 and 0.9, respectively.

2.3 Testing: Coarse-to-Fine Segmentation with Post-Processing

The first goal is to perform the pancreas and tumor segmentation. We first slide a 64^3 window in the entire CT volume. The spatial stride is 20 along three axes, which is chosen to have the average testing time for each case within 11 minutes on a TITAN Xp GPU. Based on the *coarse* segmentation, we compute the ROI, *i.e.*, the smallest box covering all pancreas and tumor voxels padded by 32, and crop the CT image accordingly. Then, we scan the ROI with sliding windows of 32^3 and 16^3 voxels, and the strides are set to be 10 and 5, respectively. We do not run the two small networks on the entire volume because it can easily hallucinate tumors in the background regions. In addition, shrinking the scanning region for the 32^3 and 16^3 networks leads to a significant speedup in the testing process. The predictions of three networks are averaged into final segmentation.

Then, based on the segmentation mask, we classify each volume as normal or abnormal. Advised by the radiologists who desire the classification result to be explainable, we do not formulate the classifier $c(\cdot)$ as another deep network, but use a simple, non-parametrized approach to filter out the outliers. We construct a graph on all voxels predicted as *normal pancreas* or *tumor*. Each voxel is a node, and there exists an edge between the adjacent voxels (each voxel is adjacent to 6 neighbors). We compute all connected component in the graph. A component is preserved if it is larger than 20% of the maximal connected component, otherwise it is removed, *i.e.*, all voxels within this component are predicted as *background*. To obtain our final goal, a volume is predicted as PDAC if at least K voxels are predicted as tumor. In practice, we empirically set $K = 50$.

3 Experiments

3.1 Dataset and Settings

We collected a dataset with 303 normal cases from potential renal donors, as well as 136 biopsy-proven PDAC cases. Four experts in abdominal anatomy annotated the pancreas and tumor voxels on these data using the Varian Velocity software, and each case was checked by an experienced board-certified Abdominal Radiologist. For a radiologist, an average normal case took 20 minutes, and an average abnormal case 40 minutes to segment. Since the abnormal cases are much harder to obtain and annotate than the normal cases, we adopt a 4-fold cross-validation on our 136 PDAC scans to have testing results on every abnormal case while we use a hard split of training and testing on our 303 normal cases. All in all, each training set contains 103 normal and 102 abnormal cases where the normal-to-abnormal ratio is close to 1, and each testing set contains 34 abnormal and 200 normal cases. The average size of CT scans is $512\times512\times667$.

Scale	N. Pancreas	A. Pancreas	Tumor	Misses	Sensitivity	Specificity
64^3	$\mathbf{86.9} \pm 8.6\%$	$\mathbf{81.0} \pm 10.8\%$	$\mathbf{57.3} \pm 28.1\%$	10/136	92.7%	**99.0**%
32^3	$82.0 \pm 12.2\%$	$75.7 \pm 14.9\%$	$53.8 \pm 26.1\%$	7/136	94.9%	96.0%
16^3	$61.5 \pm 20.6\%$	$64.1 \pm 20.2\%$	$42.5 \pm 25.6\%$	**4**/136	**97.1**%	86.5%
Multi	$84.5 \pm 11.1\%$	$78.6 \pm 13.3\%$	$56.5 \pm 27.2\%$	8/136	94.1%	98.5%

Table 1. Comparison of segmentation and classification results by networks of different scales and their combination. From left to right: normal/abnormal pancreas and tumor segmentation accuracy (DSC, %), the number of missing tumors (*i.e.*, DSC is 0%), and the sensitivity (= 1 − miss rate) and specificity.

One goal is to measure the segmentation accuracy by the Dice-Sørensen Coefficient (DSC) between the predicted and the ground-truth tumor sets $\mathcal{Y}$ and $\mathcal{Y}^\star$, *i.e.*, $\text{DSC}(\mathcal{Y}, \mathcal{Y}^\star) = 2 \times |\mathcal{Y} \cap \mathcal{Y}^\star| / (|\mathcal{Y}| + |\mathcal{Y}^\star|)$. Our main goal is the tumor classification, which involves a tradeoff between sensitivity and specificity.

3.2 Segmentation Results

We first summarize the segmentation results in Table 1, which makes the normal v.s. abnormal classification to be interpretable by segmentation cues. The 64^3 network achieves reasonable pancreas and tumor segmentation accuracies. The segmentation result of normal pancreas is as high as 86.9%, which means that the normal pancreases are easier to segment, as there are often unpredicted changes in shape and geometry in the abnormal cases. As a side comment, the lowest DSC of an abnormal pancreas is 38.4%, lower than the number (44.0%) of a normal pancreas. In tumor segmentation, we observe a lower accuracy and a higher standard deviation ($57.3 \pm 28.1\%$). Except for the 10 missing cases, we find 20 more cases with a tumor DSC lower than 30%. All these evidences imply the challenging of finding tumors considering their various size, shape and locations. Note that a recent work on the pancreatic cyst segmentation achieves a DSC of $63.4 \pm 27.7\%$ [14], which is not as hard as the tumor segmentation.

Going to smaller scales, fewer tumors are missed, though segmentation accuracies become lower. This is the tradeoff between sensitivity and specificity: a network with a smaller input region has the ability to detect tiny regions, but without seeing contexts, it can be easily confused by false positives. Thus, combining multi-scale predictions achieves a balance between sensitivity and specificity. Fig. 3 shows two examples that benefit from multi-scale segmentation.

We replace our backbone with 3D UNet [2] and VNet [6] at the 64^3 scale setting and report their results in Table 2 for comparison. We can find that the three backbones perform roughly similar in terms of the segmentation results. However, our backbone achieves the best results for the sensitivity and specificity.

Scale	N. Pancreas	A. Pancreas	Tumor	Misses	Sensitivity	Specificity
Ours	$86.9 \pm 8.6\%$	$81.0 \pm 10.8\%$	$57.3 \pm 28.1\%$	**10/136**	**92.7%**	**99.0%**
UNet	$\mathbf{87.0 \pm 8.4\%}$	$\mathbf{81.6 \pm 10.2\%}$	$57.6 \pm 27.8\%$	11/136	91.9%	99.0%
VNet	$86.7 \pm 8.8\%$	$80.6 \pm 11.4\%$	$\mathbf{58.7 \pm 28.0\%}$	10/136	92.7%	98.0%

Table 2. Comparison of different networks as backbone at the 64^3 setting.

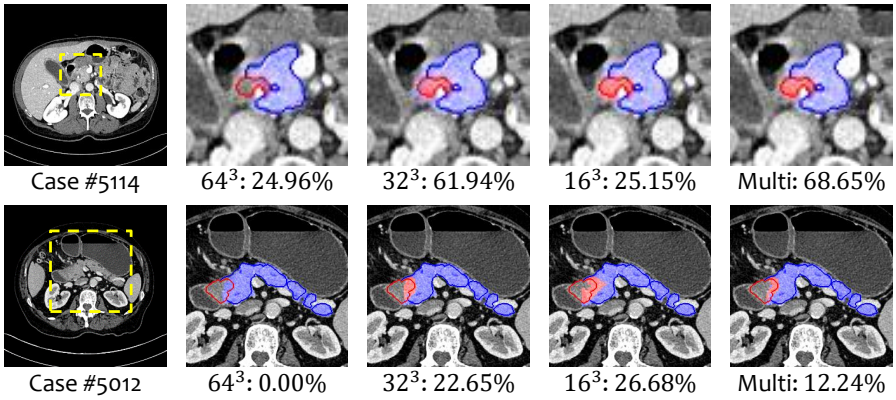

Fig. 3. Multi-scale segmentation examples (best viewed in color). Top: a case that all three scales work well, and multi-scale combines them to achieve a higher DSC. Bottom: a failure case in the 64^3 network, but found by the 32^3 and 16^3 networks. The yellow frames indicate the zoomed-in regions, the blue and red contours mark the annotated pancreas and tumor respectively, and the masked regions mark segmentation results.

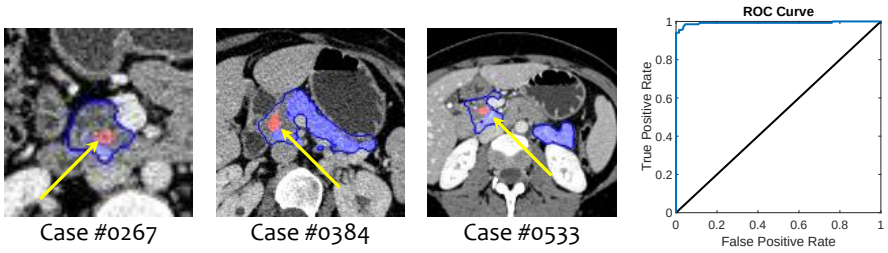

Fig. 4. Left: three false alarm examples, in which the blue contour marks the annotated pancreas, and the blue and red regions mark the predicted pancreas and tumor, respectively. We use yellow arrows to indicate the detected tiny "tumors". Right: the ROC curve of multi-scale classification. This figure is best viewed in color.

3.3 Classification Results

Finally, we summarize classification results in Table 1, which is the crucial goal of making earlier diagnosis possible for doctors. Radiologists care more about a high

sensitivity since they don't want to miss a patient who has an abnormal pancreas, which inspires us to adopt a multi-scale strategy to improve the sensitivity while keeping a reasonable specificity. The model with multi-scale information achieves the best overall performance, *i.e.*, a sensitivity of 94.1% at a specificity of 98.5%. These high scores imply that tumor segmentation provide strong cues for PDAC screening. We show all three false alarms in Fig. 4. The radiologists of our team confirmed that 2 out of these 3 false positives have focal fatty infiltration in the pancreas corresponding to the detected "tumors". Focal fatty infiltration is difficult for radiologists to distinguish from tumor in current clinical practice. In this case, the predicted "false alarm" was not normal in view of our radiologists.

By augmenting our segmentation for classification framework with cues from number of predicted tumor voxels since the more voxels predicted as PDAC the more likely this case is abnormal, we can output a confident score for each case, indicating the possibility that this case suffers PDAC. More specifically, a confidence score is computed by a weighted sum of the volume size and the segmentation probability of predicted tumor voxels. By sorting all testing cases according to their confident scores, we obtain a ROC curve of sensitivity and specificity. From the ROC curve, we can make different emphasis to change the tradeoff between sensitivity and specificity, *e.g.*, we can achieve a sensitivity of 98.5% at a specificity of 95.6%, or a specificity of 99.5% at a sensitivity of 94.1%.

4 Conclusion

In this paper, we study an important and challenging task, *i.e.*, detecting pancreases diagnosed with PDAC in abdominal CT scans. This topic is crucial in saving lives from pancreatic cancer yet few studied before, possibly due to the lack of data. We propose a **segmentation-for-classification** framework which trains a segmentation network and performs **interpretable** abnormality classification by simply checking the existence of tumor voxels in each testing volume. There are two key points to improve classification accuracy, known as **multi-scale** network training and **coarse-to-fine** testing. To offer a best tradeoff between sensitivity and specificity on our own collected dataset containing 303 normal and 136 PDAC cases, we achieve a sensitivity of 94.1% at a specificity of 98.5%. The strong numbers show the promising direction to make a significant impact in clinics for early detection of pancreatic cancer, which would save lives.

Acknowledgements This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research.

References

1. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. In: ICLR (2016)
2. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D u-net: learning dense volumetric segmentation from sparse annotation. MICCAI (2016)

3. Dou, Q., Chen, H., Yu, L., Qin, J., Heng, P.A.: Multilevel contextual 3-d cnns for false positive reduction in pulmonary nodule detection. IEEE TBE (2017)
4. Hussein, S., Chuquicusma, M.M., Kandel, P., Bolan, C.W., Wallace, M.B., Bagci, U.: Supervised and unsupervised tumor characterization in the deep learning era. arXiv:1801.03230 (2018)
5. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
6. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3DV (2016)
7. PDQ Adult Treatment Editorial Board: Pancreatic cancer treatment (PDQ®)
8. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
9. Roth, H.R., Lu, L., Farag, A., Shin, H.C., Liu, J., Turkbey, E.B., Summers, R.M.: Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In: MICCAI (2015)
10. Roth, H.R., Lu, L., Farag, A., Sohn, A., Summers, R.M.: Spatial aggregation of holistically-nested networks for automated pancreas segmentation. MICCAI (2016)
11. Stewart, B.W.K.P., Wild, C.P., et al.: World cancer report 2014. Health (2017)
12. Xia, Y., Xie, L., Liu, F., Zhu, Z., Fishman, E.K., Yuille, A.L.: Bridging the gap between 2d and 3d organ segmentation. MICCAI (2018)
13. Zhang, L., Lu, L., Summers, R.M., Kebebew, E., Yao, J.: Personalized pancreatic tumor growth prediction via group learning. In: MICCAI (2017)
14. Zhou, Y., Xie, L., Fishman, E.K., Yuille, A.L.: Deep supervision for pancreatic cyst segmentation in abdominal ct scans. In: MICCAI (2017)
15. Zhu, Z., Xia, Y., Shen, W., Fishman, E.K., Yuille, A.L.: A 3d coarse-to-fine framework for volumetric medical image segmentation. 3DV (2018)