KNOWLEDGE FUSION IN ALGORITHMS FOR MEDICAL IMAGE ANALYSIS

by

Fengze Liu

A dissertation submitted to Johns Hopkins University
in conformity with the requirements for the degree of
Doctor of Philosophy

Baltimore, Maryland
October, 2022

Abstract

Medical imaging is one of the primary modalities used for clinical diagnosis and treatment planning. Building up a reliable automatic system to assist clinicians read the enormous amount of images benefits the efficiency and accuracy in general clinical trail. Recently deep learning techniques have been widely applied on medical images, but for applications in real clinical scenario, the accuracy, robustness, interpretability of those algorithms requires further validation.

In this dissertation, we introduce different strategies of knowledge fusion for improving current approaches in various tasks in medical image analysis. (i) To improve the robustness of segmentation algorithm, we propose to learn the shape prior for organ segmentation and apply it for automatic quality assessment. (ii) To detect pancreatic lesion with patient-level label only, we propose to extract shape and texture information from CT scans and combine them with a fusion network. (iii) In image registration, semantic information is important yet hard to obtain. We propose two methods for introducing semantic knowledge without the need of segmentation label. The first one designs a joint framework for registration synthesis and segmentation to share knowledge between different tasks. The second one introduces unsupervised semantic embedding to improve regular registration framework. (iv) To reduce the false positives in tumor detection task, we propose a hybrid feature engineering system extracting features of the tumor candidates from various perspectives and merging them in the decision stage.

Thesis Readers

Dr. Alan L. Yuille (Primary Advisor)
Bloomberg Distinguished Professor
Department of Computer Science
Johns Hopkins University

Dr. Wei Shen
Associate Professor
Artificial Intelligence Institute
Shanghai Jiao Tong University

Dr. Le Lu
Head of Medical AI RD
Alibaba Group
IEEE Fellow

Acknowledgments

First and foremost, I want to express thanks to my advisor Prof. Alan L. Yuille for his guidance and support during my journey as a Ph.D. student. I still remember the very first question I asked him when I came as a summer intern, and he pointed me at several related papers, which then introduced me to the world of computer vision. I appreciate that he can always inspire me during the discussion on research directions and also give me freedom to explore various projects. This encourages me to grow as an independent researcher.

Next, I would like to thanks all my collaborators and mentors. I want to thank Le Lu for organizing so many meaningful internship projects and thank Dakai Jin, Adam Harrison, Daguang Xu, Dong Yang, Ke Yan for insightful discussions during my internship at Nvidia and PAII. Most of my research works are completed during the internship. This dissertation would have been impossible without them. Besides, I want to thanks Prof. Rama Chellappa, Vishal Patel, Elliot Fishman, Linda Chu, Wei Shen, Alex Szalay for serving on my GBO committee and providing valuable suggestions. I would also like to thank Zachary Burwell, MaDonna Perry, and Kim Franklin for scheduling my GBO and help on completing Ph.D. requirements.

Last, I want to thank all members in CCVL, for numerous discussions and collaboration, including Lingxi, Jun, Vittal, Ehsan, Wei, Yan, Adam, Yongyi, Weichao, Zhuotun, Chenxi, Zhishuai, Siyuan, Cihang, Yuyin, Qing, Zhe, Huiyu, Chenxu, Qi, Yi, Yingda, Hongru, Jieru, Qihang, Yingwei, Yixiao, Zhuowan, Zihao, Chenglin, Yutong, Angtian, Chen, Jieneng.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

With advanced medical imaging techniques, an enormous amount of medical images has been generated every year. These medical images are crucial in lesion detection, treatment planning and medical intervention to assist doctors making clinical decisions. Missed or inaccurate diagnosis will lead to improper treatment, which causes the patients debilitation or even death. However, precise analysis of medical images requires years of experience from qualified radiologists, and it is limited by the speed, fatigue, and time cost. Ideally, developing an automatic system that can well read and understand the medical images surely improves the accuracy and efficiency of the diagnosis.

Computer-aided detection and diagnosis (CAD) is the system that assists doctors interpreting medical images. The goal of the system is to understand the anatomical structure inside the medical images and highlight the suspicious area for lesion detection as well as providing quantitative analysis of the lesion. So it is usually equipped with machine learning and compute vision algorithm to achieve that goal. As a result, the performance of the system is limited by the algorithms. Even for an expert system which has the same performance with experienced radiologists, it is not clear whether the diagnosis quality can be improved after using CAD [156]. There are also conditions that the usage of CAD causes drop of the diagnosis accuracy [30] when the radiologists get misled by the CAD system. To ensure the positive effect of CAD system, more robust and accurate algorithms are always pursued.

In recent years, with the success application of deep learning in the natural image domain, huge progresses have been made on various vision tasks, including image classification[66, 121, 40], object detection [105, 82, 102, 103, 104, 16] and semantic segmentation [84, 18]. Soon after that the researchers discover that the ability of deep learning to learn representation can also be applied to medical images. Then the success of deep learning models continues on tasks in CAD, including organ segmentation[112, 54, 166], lesion detection[147, 170], image registration[7, 155].

Despite the initial achievements, there remain challenges within medical image analysis community. Firstly getting annotation on medical images requires massive human efforts from experts, especially for accurate pixel-wise annotation, which prevents researchers from training or evaluating models on large scale data. Secondly, due to various image modalities, there usually is a domain gap when applying a trained model to the data from other institutes. Even when the data are from the same modality, the performance is not guaranteed because of the different parameter settings when producing the medical images. As a result, the actual performance of the models in the real application scenario is hard to predict. On facing these challenges, considerable efforts have been made to develop more robust models, including semi-supervised learning [85, 32, 71] and transfer learning [133, 59, 28]. However, for most of the methods, the principle idea is still similar with approaches that are designed for natural images. We argue that medical images possess unique prior information which is not shared with natural images and combining such knowledges into the AI system benefits the robustness of algorithms [29].

1.1 Priors in Medical Images

Medical images contain several types of prior information which are not shared with natural images.

Content Prior The objects inside a medical images is stable within a certain body range. For example, the radiologists are expecting to find a pancreas, a liver etc in every abdominal CT of a healthy patient. But in a natural image, there is no guarantee whether a certain object, like a car will appear in this image or not.

Positional Prior When taking a medical images, the patient is usually required to keep a straight pose until finish. For modalities like CT and MRI, which are able to reflect the anatomical structure in 3D, the relative position between organs is stable. As shown in [33], such prior information and can be utilized for improving model robustness across datasets.

Shape Prior Firstly the shape of organs and tissues obeys a certain distribution, which is shared in human species. Although due to genetic variation, the size and shape of organs change for each specific person, the variation still stays in a range [106]. Secondly, 2D natural images are produced by projecting a 3D scene from a certain view into a surface. The shape of an object will change according to the view and it also causes ambiguity due to the information loss at projecting step. However, there is no such ambiguity brought by view change in 3D medical images. Even for 2D medical images, the images are also produced under a preset view. [134, 152, 101] have proposed additional shape related module to guide the training process, but how to model the complex shape variation and further incorporate such shape prior constraints in the decision level of algorithms remains an open question.

Intensity Prior Due to the basic principles of producing medical images [14], the intensity value in each pixel of medical image corresponds to the physical properties inside the body. Although the intensity value is related with parameter setting during the imaging process, modality, usage of contrast media, the intensity value still has a physical meaning which can be shared across all the hospitals and institutes. For natural images, the appearance for an object is more arbitrary and is affected by lighting, occlusion.

1.2 Outlines

In this dissertation we study how to improve robustness and efficiency by incorporating additional knowledge into the AI system in medical image analysis.

1.2.1 Automatic Quality Assessment

We aim at building a quality assessment system for segmentation task. Quality assessment module is used to monitor the output from the algorithm and tries to give confidence and quality evaluation of the output. It is important in large-scale data related applications to ensure the input and output of system have expected quality and can save a lot of human efforts. It is usually hard for a learning system to predict correctly on rare events that never occur in the training data. Meanwhile, manual inspection of each case to locate the failures becomes infeasible due to the trend of large data scale and limited human resource, especially for the segmentation task, which requires pixel-level annotation first in order to evaluate.

In chapter 2 we propose an alarm system [76] that will set off alerts when the segmentation result is possibly unsatisfactory, assuming no corresponding ground truth mask is provided. One plausible solution is to project the segmentation results into a low dimensional feature space; then learn classifiers/regressors to predict their qualities. Motivated by this, we learn a feature space using the shape information which is a strong prior shared among different datasets and robust to the appearance variation of input data. The shape feature is captured using a Variational Auto-Encoder [63] (VAE) network that trained with only the ground truth masks. During testing, the segmentation results with bad shapes shall not fit the shape prior well, resulting in large loss values. Thus, the VAE is able to evaluate the quality of segmentation result on unseen data, without using ground truth. Finally, we learn a regressor in the one-dimensional feature space to predict the qualities of segmentation results. Our alarm system is evaluated on several recent state-of-art segmentation algorithms [91, 168, 19, 68] for 3D medical segmentation tasks across three public datasets. Compared with other standard quality assessment methods, our system consistently provides more reliable prediction on the qualities of segmentation results.

1.2.2 Lesion Detection

We focus on early detection of pancreatic tumors (PDAC) in CT scans. Compared with segmentation framework [164] which requires accurate voxel-wise annotation of tumors from radiologists, we treat the detection as a 3D volume classification task so that we only need patient-level annotation.

Considering that the tumor inside pancreas will cause the shape change of pancreas. In chapter 3 we propose to obtain the representation of the shape of pancreas and then use it for classifying anomaly [77]. A two-stage framework is developed, which first segments the pancreas into a binary mask, then compresses the mask into a shape vector and performs abnormality classification. Shape representation and classification are performed in a *joint* manner, both to exploit the knowledge that PDAC often changes the shape of the pancreas and to prevent over-fitting. Experiments are performed on normal scans and PDAC cases in JHU PDAC dataset. The system achieves accurate result with only the shape information of pancreas and shows promise for clinical applications.

Although shape prior is important itself, the texture information is also crucial in detecting pancreatic tumors. It is sub-optimal for the system to only consider shape of pancreas. In chapter 4, we propose to fuse the cues from both shape and texture by designing a system to extract the shape and texture feature at the same time for detecting PDAC [80]. A two-stage method is used for this 3D classification task. First, we segment the pancreas into a binary mask. Second, a FusionNet is proposed to take both the binary mask and CT image as input and perform a binary classification. The optimal architecture of the FusionNet is obtained by searching a pre-defined functional space. We show that the classification results using either shape or texture information are complementary, and by fusing them with the optimized architecture, the performance improves by a large margin.

1.2.3 Unsupervised Image Registration

Image registration is a challenging but also important clinical task for many real applications and scenarios. As the first step in analysis, deformable registration among different image modalities is often required in order to provide complementary visual information. During registration, semantic information is the key to match homologous points and pixels. Nevertheless, many conventional registration methods are incapable in capturing high-level semantic anatomical dense correspondences.

In chapter 5, we propose a novel multi-task learning system, JSSR [73], based on an end-to-end 3D convolutional neural network that is composed of a generator, a registration and a segmentation component. The system is optimized to satisfy the implicit constraints between different tasks in an unsupervised manner. It first synthesizes the source domain images into the target domain, then an intra-modal registration is applied on the synthesized images and target images. The segmentation module are then applied on the synthesized and target images, providing additional cues based on semantic correspondences. The supervision from another fully-annotated dataset is used to regularize the segmentation. We extensively evaluate JSSR on a large-scale medical image dataset containing 1,485 patient CT imaging studies of four different contrast phases (i.e., 5,940 3D CT scans with pathological livers) on the registration, segmentation and synthesis tasks. The performance is improved after joint training on the registration and segmentation tasks compared to a highly competitive and accurate deep learning baseline. The registration also consistently outperforms conventional state-of-the-art multi-modal registration methods.

In chapter 6, we take advantages of dense anatomical/semantic representation from another work [146] and propose a fast and accurate system [79] for unsupervised 3D medical image registration. The system breaks down image registration into three steps: affine transformation, coarse deformation, and deep deformable registration. Using high level semantic embeddings, we enhance these steps by finding more coherent correspondences, and providing features and a loss function with better semantic guidance. We collect a multi-phase chest computed tomography dataset with 35 annotated organs for each patient and conduct inter-subject registration for quantitative evaluation. Results show that our system outperforms widely-used traditional registration techniques and learning based method for two separate tasks of within-contrast-phase and across-contrast-phase registration, respectively. Our system achieves the comparable performance to the best traditional registration method, DEEDS [46] (from our evaluation), while being orders of magnitude faster.

1.2.4 False Positive Reduction

False positive reduction (FPR) is an important stage in automatic lesion detection framework, especially when early diagnosis is needed on the target lesion. Most recent methods rely on learning discriminative features for FPR via deep networks [130, 27, 119, 127], which however can be biased towards texture. The complementary features, such like shape, are usually better captured by hand-crafted feature engineering. How to effectively combine comprehensive features regarding to different aspects and different engineering processes to boost FPR performance remains an unsolved problem.

In chapter 7, we propose a Hybrid Feature Engineering (HFE) framework for FPR in pancreatic lesion detection. It consists of 1) A massive hybrid feature pool which contains both hand-crafted and learned features to represent each false positive candidate from different aspects; 2) A sequential feature selection which efficiently picks up useful features from the pool; 3) A random forest classifier, which is learned upon the selected hybrid feature pool. Our HFE framework achieves high accuracy on identifying false positives from the lesion detection candidates extracted from JHU CT dataset and reduces the false positive rate by a large margin with merely drop in the sensitivity.

In Chapter 8, we state the conclusion of this thesis and provide potential directions for future work.

1.3 Relevant Publications

The following publications contribute to the main idea of this dissertation.

- Chapter 2 - **Fengze Liu**, Yingda Xia, Dong Yang, Alan L Yuille, Daguang Xu. An Alarm System for Segmentation Algorithm Based on Shape Model, in ICCV 2019.
- Chapter 3 - **Fengze Liu**, Lingxi Xie, Yingda Xia, Elliot Fishman, Alan Yuille. Joint Shape Representation and Classification for Detecting PDAC, in MLMI 2019.
- Chapter 4 - **Fengze Liu**, Yuyin Zhou, Elliot Fishman, Alan Yuille. FusionNet: Incorporating Shape and Texture for Abnormality Detection in 3D Abdominal CT Scans, in MLMI 2019.
- Chapter 5 - **Fengze Liu**, Jinzheng Cai, Yuankai Huo, Chi-Tung Cheng, Ashwin Raju, Dakai Jin, Jing Xiao, Alan Yuille, Le Lu, ChienHung Liao, Adam P Harrison. JSSR: A Joint Synthesis, Segmentation, and Registration System for 3D Multi-modal Image Alignment of Large-Scale Pathological CT Scans, in ECCV 2020.
- Chapter 6 - **Fengze Liu**, Ke Yan, Adam P Harrison, Dazhou Guo, Le Lu, Alan L Yuille, Lingyun Huang, Guotong Xie, Jing Xiao, Xianghua Ye, Dakai Jin. SAME: Deformable Image Registration based on Embeddings, in MICCAI 2021.

- Chapter 7 - **Fengze Liu**, Yongyi Lu, Yingda Xia, Wei Shen, Elliot Fishman, Alan Yuille. Hybrid Feature Engineering for False Positive Reduction in Pancreatic Lesion Detection.

The other related publications provide contexts for this dissertation:

- Yan Wang, Xu Wei, **Fengze Liu**, Jieneng Chen, Yuyin Zhou, Wei Shen, Elliot K Fishman, Alan L Yuille. Deep Distance Transform for Tubular Structure Segmentation in CT Scans, in CVPR 2020.
- Yuan Yao, **Fengze Liu**, Zongwei Zhou, Yan Wang, Wei Shen, Alan Yuille, Yongyi Lu. Unsupervised Domain Adaptation through Shape Modeling for Medical Image Segmentation, in MIDL 2022.
- Yingda Xia, Yi Zhang, **Fengze Liu**, Wei Shen, Alan L Yuille. Synthesize Then Compare: Detecting Failures and Anomalies for Semantic Segmentation, in ECCV 2020.

Chapter 2

An Alarm System for Segmentation Algorithm Based on Shape Model

2.1 Introduction

Segmentation algorithms often fail on rare events, and it is hard to fully avoid such issue. The rare events may occur due to limited number of training data. The most intuitive way to handle this problem is to increase the number of training data. However, the labelled data is usually hard to collect especially in medical domain, e.g., fully annotating a 3D medical CT scan requires professional radiology knowledge and several hours of work. Meanwhile, even large number of labelled data is usually unable to cover all possible cases. Previously, various methods have been proposed to make better use of the training data, like sampling strategies paying more attention to the rare events [135]. But still they may fail on rare events that never occur in the training data. Another direction is to increase the robustness of the segmentation algorithm to rare events. [60] proposed the Bayesian neural network that models the uncertainty as an additional loss to

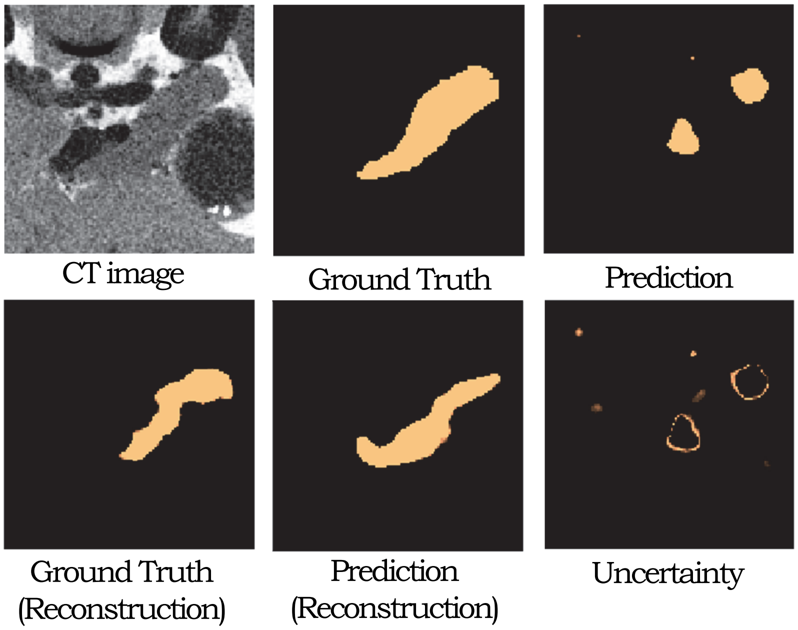


Figure 2.1: The visualize on an NIH CT data for pancreas segmentation. The Dice between GT and Prediction is 47.06 (real Dice) while the Dice between Prediction and Prediction(Reconstruction) from VAE is 47.25 (fake Dice). Our method uses the fake Dice to predict the former real Dice which is usually unknown at inference phase of real applications. This case shows how these two Dice scores are related to each other. In contrast, the uncertainty used in existing approaches (introduced in section 2) mainly distributes on the boundary of predicted mask, which makes it a vague information when detecting the failure cases.

make the algorithm more robust to noisy data. These kinds of methods make the algorithm insensitive to certain types of perturbations, but the algorithms may still fail on other perturbations.

Since it is hard to completely prevent the segmentation algorithm from failure, we consider detecting the failure instead: build up an alarm system cooperating with the segmentation algorithm, which will set off alerts when the system finds that the segmentation result is not good enough. It is assumed that there is no corresponding ground truth mask, which is usually true after the model deployment due to the trend of large data scale and limited human resource. This task is also called as quality assessment. Several works have been proposed in this field. [58] applied Bayesian neural network to capture the uncertainty of the segmentation result and set off alarm based on its value. However, this system also suffers from rare events since the segmentation algorithms often make mistakes confidently on some rare events [143], shown in Figure 2.1. [65] provided an effective way by projecting the segmentation results into a feature space and learn from this low dimension space. They manually designed several heuristic features, e.g., size, intensity, and assumed such features would indicate the quality of the segmentation results. After projecting the segmentation results into a low-dimensional feature space, they learned a classifier to predict its quality which distinguishes good segmentation results from bad ones directly. In a reasonable feature space, the representation of the failure output should be far from that of the ground truth when the segmentation algorithm fails. So the main problems is what these "good" features are and how to capture them. Many features selected in [65] are actually less related to the quality of segmentation results, e.g., size.

In our system, we choose the shape feature which is more representative and robust because the segmented objects (foreground in the volumetric mask) usually have stable shapes among different cases even though their image appearance may vary a lot, especially in 3D. So the shape feature could provide a strong prior information for judging the quality of segmentation results, i.e., bad segmentation

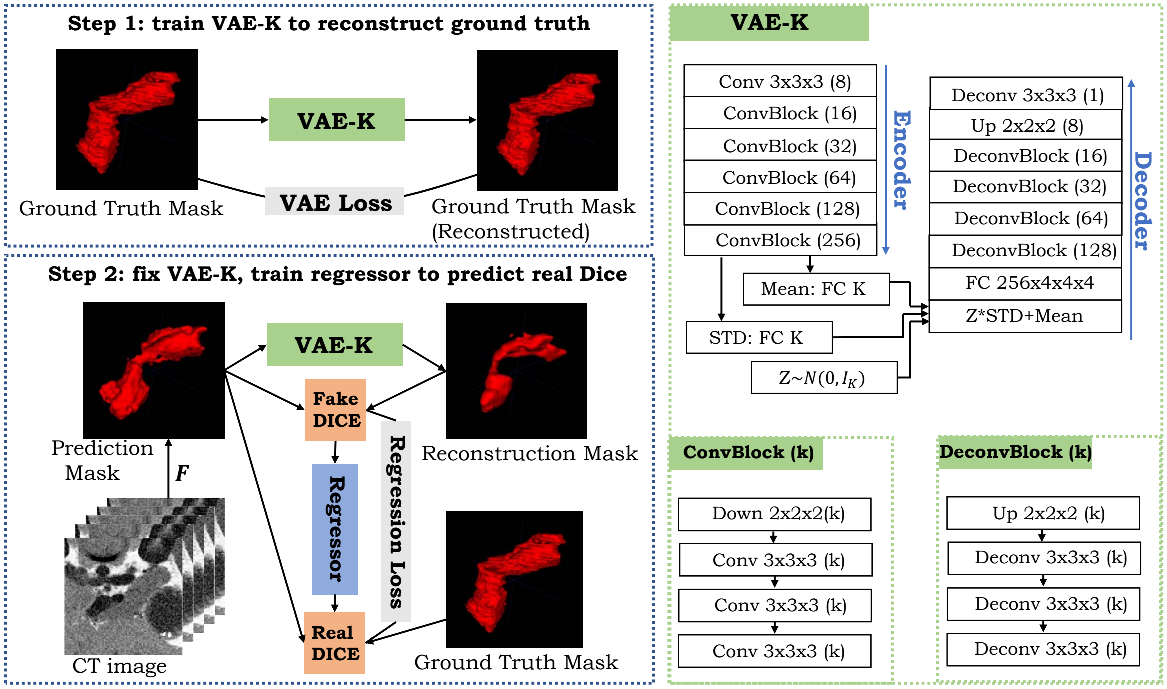


Figure 2.2: The architecture of our alarm system. In train step 1, the VAE is trained to reconstruct the ground truth masks. In train step 2, the parameters of VAE are fixed and a regressor is trained to predict the real Dice score. F represents a preparation segmentation algorithm which is used to generate prediction masks for training the regressor. During testing, F is replaced with the target segmentation algorithm to be evaluated. On the right side we show the structure of VAE used. (**Conv**: convolution layers with stride 1. **Down**: convolution layers with stride 2. **Deconv**: transpose convolution layers with stride 1. **Up**: transpose convolution layers with stride 2. **FC**: fully connected layers. ***k***: convolution kernel numbers.) Further details about the structure are presented in section 4.3.

results tend to have bad shapes and vice versa. Furthermore, modeling the prior from the segmentation mask space is much easier than doing it in the image space. The shape prior can be shared among different datasets while the features like image intensity are affected by many factors. Thus, the shape feature can deal with not only rare events but also different data distributions in the image space, which shows great generalization power and potential in transfer learning. We propose to use the Variational Auto-Encoder(VAE) [63] to capture the shape feature. The VAE is trained on the ground truth masks, and afterwards we define the value of the loss function as the shape feature of a segmentation result when it is tested with VAE network. Intuitively speaking, after the VAE is trained, the bad segmentation results with bad shapes are just rare events to VAE because it is trained using only the ground truth masks, which are under the distribution of normal shapes. Thus they will have larger loss value. In this sense we are utilizing the fact that the learning algorithms will perform badly on the rare events. Formally speaking (detailed in Sec. 3.1), the loss function, known as the variational lower bound, is optimized to approximate the function $\log P(Y)$ during the training process. So after the training, the value of the loss function given a segmentation result $\hat{Y}$ is close to $\log P(\hat{Y})$, thus being a good definition for the shape feature.

In this paper, we proposed a VAE-based alarm system for segmentation algorithms, shown in Figure 4.1. The qualities of the segmentation results can be well predicted using our system. To validate the effectiveness of our alarm system, we test it on multiple segmentation algorithms. These segmentation algorithms are trained on one dataset and tested on several other datasets to simulate when the rare events occur. The performance for the segmentation algorithms on the other datasets (rather than the training dataset) varies a lot but our system can still predict their qualities accurately. We compare our system with several other alarm systems on the above tasks and ours outperforms them by a large margin, which shows the importance of shape feature in the alarm system and the great power of

VAE in capturing the shape feature.

2.2 Related Work

2.2.1 Quality Assessment:

[60] employed Bayesian neural network (BNN) to model the aleatoric and epistemic uncertainty. Afterwards, [68] applied the BNN to calculate the aleatoric and epistemic uncertainty on medical segmentation tasks. [58] utilized the BNN and model another kind of uncertainty-based on the entropy of segmentation results. They calculated a doubt score by summing over weighted pixel-vise uncertainty.

Other methods like [131][108] used registration based approach for quality assessment. It registered the image of testing case with a set of reference image and also transfer the registration to the segmentation mask to find the most matching one. However it can be slow to register with all the reference image especially in 3D. Also the registration based approach can hardly be transferred between datasets or modalities. [17] and [34] used unsupervised methods to estimate the segmentation quality using geometrical and other features. However their application in medical settings is not clear. [65] introduced a feature space of shape and appearance to characterize a segmentation. The shape features in their system contain volume size and surface area, which are not necessarily related with the quality of the segmentation results. Meanwhile, [107] tried a simple method using image-segmentation pairs to directly regress the quality. [12] used the feature from deep network for quality assessment.

2.2.2 Anomaly Detection:

Quality assessment is also related with Out-of-Distribution (OOD) detection. Investigation related research papers can be found in [97]. Previous works in this field [48] [72] made use of the softmax output in the last layer of a classifier to calculate the out-of-distribution level. In our case, however, for a segmentation method, we can only get a voxel-wise out-of-distribution level using these methods. How to calculate the out-of-distribution level for the whole mask as an entity becomes another problem. In addition, the segmentation algorithm can usually predict most of background voxels correctly with a high confidence, making the out-of-distribution level on those voxels less representative.

2.2.3 Auto-Encoder:

Auto-Encoder(AE), as a way of learning representation of data automatically, has been widely used in many areas such as anomaly detection [171], dimension reduction, etc. Unlike [139] which needs to pre-train with RBM, AE can be trained following an end-to-end fashion. [98] learned the shape representation from point cloud form, while we choose the volumetric form as a more natural way to corporate with segmentation task. [95] utilizes AE to evaluate the difference between prediction and ground truth but not in an unsupervised way. [165] explored shape features using AE. [10] utilized the reconstruction error of brain MRI image by AE and [117] used GAN for anomaly detection but it is sometimes hard to generate a realistic image abdominal CT scan. [118] used AE and a one-class SVM to identify anomalous regions in OCT images through unsupervised learning on healthy examples. Variational autoencoder(VAE) [63], compared with AE, adds more constraint on the latent space, which prevents from learning a trivial solution identity mapping. [2] applied VAE for anomaly detection on MNIST and KDD datasets. In this paper we employ VAE to learn the shape representation for the volumetric mask and use that for quality assessment task.

2.3 Our VAE-based Alarm System

We first define our task formally. Denote the datasets as $(\mathcal{X},\mathcal{Y})$, where $\mathcal{Y}$ is the label set of $\mathcal{X}$. We divide $(\mathcal{X},\mathcal{Y})$ into training set $(\mathcal{X}_t,\mathcal{Y}_t)$ and validation set $(\mathcal{X}_v,\mathcal{Y}_v)$. Suppose we have a segmentation algorithm F trained on $\mathcal{X}_t$. Usually we validate the performance of F on $\mathcal{X}_v$ using $\mathcal{Y}_v$. Now we want to do this task without $\mathcal{Y}_v$. Formally, we try to find a function L such that

$$\mathcal{L}(F(X),Y) = L(F,X;\omega) \tag{2.1}$$

where $\mathcal{L}$ is a function used to calculate the similarity of the segmentation result $F(X)$ respect to the ground truth Y, i.e., the quality of $F(X)$. How to design L to take valuable information from F and X, is the main question. Recall that the failure may happen when X is a rare event. But to detect whether an image X is within the distribution of training data is very hard because of the complex structure of image space. In uncertainty-based method [58] and [68], the properties of F are encoded by sampling its parameters and calculating the uncertainty of output. The uncertainty does help predict the quality but the performance strongly relies on F. It requires F to have Bayesian structure, which is not in our assumption.

Also for a well-trained F, the uncertainty will mainly distribute on the boundary of segmentation prediction. So we change the formulation above to

$$\mathcal{L}(F(X), Y) = L(F(X); \omega) \tag{2.2}$$

By adding this constraint, we still take the information from F and X, but not in a direct way. The most intuitive idea to do is directly training a regressor on the segmentation results to predict the quality. But the main problem is that the regression parameters trained with a certain segmentation algorithm F highly relate with the distribution of $F(X)$, which varies from different F.

Following the idea of [65], we develop a two-step method. Firstly we encode the segmentation result $F(X)$ into the feature space, denoting as $S(F(X); \theta)$. Secondly we learn from the feature space to predict the quality of $F(X)$. Finally it changes to

$$\mathcal{L}(F(X), Y) = L(S(F(X); \theta); \omega) \tag{2.3}$$

2.3.1 Shape Feature from Variational Autoencoder

In the first step we learn a feature space of shape from Variational Autoencoder (VAE) trained with the ground masks $Y \in \mathcal{Y}_t$, using $S(Y; \theta)$ to indicate how perfect the shape of Y is. Here we define the shape of the segmentation masks as the distribution of the masks in volumetric form. We assume the normal label Y obeys a certain distribution $P(Y)$. For a predictive mask $\hat{y}$, its quality should be related with $P(Y = \hat{y})$. Our goal is to estimate the function $P(Y)$ using $S(Y; \theta)$. Recall the theory of VAE, we hope to find an estimation function $Q(z)$ minimizing the difference between $Q(z)$ and $P(z|Y)$, where z is the variable of the latent space we want encoding Y into, i.e. optimizing

$$\mathcal{KL}[Q(z)||P(z|Y)] = E_{z\sim Q}[\log Q(z) - \log P(z|Y)] \tag{2.4}$$

$\mathcal{KL}$ is Kullback-Leibler divergence. By replacing $Q(z)$ with $Q(z|Y)$, finally it would be deduced to the core equation of VAE [25].

$$\begin{aligned} &\log P(Y) - \mathcal{KL}[Q(z|Y)||P(z|Y)] \\ &= E_{z\sim Q}[\log P(Y|z)] - \mathcal{KL}[Q(z|Y)||P(z)] \end{aligned} \tag{2.5}$$

where $P(z)$ is the prior distribution we choose for z, usually Gaussian, and $Q(z|Y), P(Y|z)$ correspond to encoder and decoder respectively. Once Y is given, $\log P(Y)$ is a constant. So by optimizing the RHS known as variational lower bound of $\log P(Y)$, we optimize for $\mathcal{KL}[Q(z|Y)||P(z|Y)]$. Here however we are interested in $P(Y)$. By exchanging the second term in LHS with all terms in RHS in equation (5), we rewrite the training process as minimizing

$$\begin{aligned} &E_{Y\sim\mathcal{Y}_t}\, \mathcal{KL}[Q(z|Y)||P(z|Y)] \\ &= E_{Y\sim\mathcal{Y}_t}|\log P(Y) - S(Y;\theta)| \end{aligned} \tag{2.6}$$

We choose $E_{z\sim Q}[\log P(Y|z)] - \mathcal{KL}[Q(z|Y)||P(z)]$ to be $S(Y;\theta)$. $S(Y;\theta)$ is the loss function we use for training VAE and the training process is actually learning the parameters θ to best fit $\log P(Y)$ over the distribution of Y. So after training

VAE, $S(Y;\hat{\theta})$ becomes a natural approximation for $\log P(Y)$ where $\hat{\theta}$ is the learned parameter. So we can just use $S(Y;\hat{\theta})$ as our shape feature. In this method we use Dice Loss [91] when training VAE, which is widely used in medical segmentation task. The final form of S is

$$\begin{aligned} S(Y;\theta) = E_{z\sim\mathcal{N}(\mu(Y),\Sigma(Y))} \frac{2|g(z)\cdot Y|}{|Y|^2 + |g(z)|^2} \\ - \lambda\, \mathcal{KL}[\mathcal{N}(\mu(Y),\Sigma(Y))||\mathcal{N}(0,1)] \end{aligned} \tag{2.7}$$

where encoder μ, Σ and decoder g are controlled by θ, and λ is a coefficient to balance the two terms. The first term is the Dice's coefficient between Y and $g(z)$, ranging from 0 to 1 and equal to 1 if Y and $g(z)$ are equal.

2.3.2 Shape Feature for Predicting Quality

In the second step we regress on the shape feature to predict the quality. We assume that the shape feature is good enough to obtain reliable quality assessment because intuitively thinking, for a segmentation result $F(X)$, the higher $\log P(F(X))$ is, the better shape $F(X)$ is in, thus the higher $\mathcal{L}(F(X), Y)$ is and vice versa. Formally, taking the shape feature in section 3.1, we can predict the quality by learning ω such that

$$\mathcal{L}(F(X), Y) = L(S(F(X);\hat{\theta});\omega) \tag{2.8}$$

Here the parameter $\hat{\theta}$ is learned by training the VAE, using labels in the training data $\mathcal{Y}_t$, and is then fixed during train step two. We choose L to be a simple linear

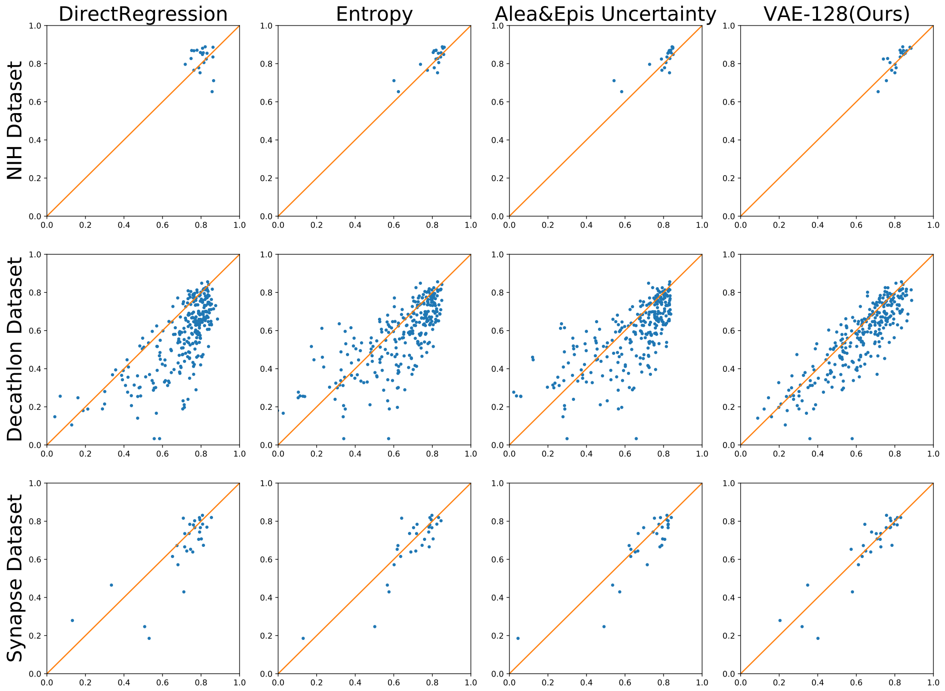


Figure 2.3: This figure shows our predictive Dice score (x axis) vs real Dice score (y axis). For each row, the segmentation algorithm is tested on the left most dataset. The four figures in each row show how the segmentation results are evaluated by 4 different methods.

model, so the energy function we want to optimize is

$$E(S(F(X);\hat{\theta});a,b) = ||aS(F(X);\hat{\theta}) + b - \mathcal{L}(F(X),Y)||^2 \tag{2.9}$$

We only use linear regression model because the experiments show strong linear correlation between the shape features and the qualities of segmentation results. $\mathcal{L}$ is the Dice's coefficient, i.e. $\mathcal{L}(F(X),Y) = \frac{2|F(X)\cdot Y|^2}{|F(X)|^2+|Y|^2}$.

2.3.3 Training Strategy

In step one, the VAE is trained only using labels in training data. Then in step two θ is fixed as $\hat{\theta}$. To learn a, b, the standard way is to optimize the energy function in 3.2 using the segmentation results on the training data, i.e.

$$\arg\min_{a,b} \sum_{(X,Y)\in(\mathcal{X}_t,\mathcal{Y}_t)} ||aS(F(X);\hat{\theta}) + b - \mathcal{L}(F(X),Y)||^2. \tag{2.10}$$

Here the segmentation algorithm F we use to learn a, b is called the preparation algorithm. If F is trained on $\mathcal{X}_t$, the quality of $F(X)$ would be always high, thus providing less information to regress a, b. To overcome this, we use jackknifing training strategy for F on $\mathcal{X}_t$. We first divide $\mathcal{X}_t$ into $\mathcal{X}_t^1$ and $\mathcal{X}_t^2$. Then we train two versions of F on $\mathcal{X}_t \setminus \mathcal{X}_t^1$ and $\mathcal{X}_t \setminus \mathcal{X}_t^2$ respectively, say F_1 and F_2. The optimizing function is then changed to

$$\arg\min_{a,b} \sum_{k=1,2} \sum_{(X,Y)\in(\mathcal{X}_t^k,\mathcal{Y}_t^k)}$$

$$||aS(F_k(X);\hat{\theta}) + b - \mathcal{L}(F_k(X),Y)||^2. \tag{2.11}$$

In this way we solve the problem above by simulating the performance of F on the testing set. The most accurate way is to do leave-one-out training for F, but the time consumption is not acceptable, and two-fold split is effective enough according to experiments. When the training is done, we can test on any segmentation algorithm G and data X to predict the quality $Q = \hat{a}S(G(X);\hat{\theta}) + \hat{b}$ where $\hat{a}$ and $\hat{b}$ are the learned parameters for step 2 using the above strategy.

2.4 Experimental Results

In this section we test our alarm system on several recent algorithms for automatic pancreas segmentation that are trained on a public medical dataset. Our system achieves reliable predictions on the qualities of segmentation results. Furthermore, the alarm system remains effective when the segmentation algorithms are tested on other unseen datasets. We show better quality assessment capability and transferability compared with uncertainty-based methods and direct regression method. The quality assessment results are evaluated using mean absolute error (MAE), standard deviation of residual error (STD), Pearson correlation (P.C.) and Spearman's correlation (S.C.) between the real quality (Dice's coefficient) and predictive quality.

2.4.1 Dataset and Segmentation Algorithm

We adopt three public medical datasets and four recently published segmentation algorithms in total. All datasets consist of 3D abdominal CT images in portal venous phase with pancreas region fully annotated. The CT scans have resolutions of $512 \times 512 \times h$ voxels with varying voxel sizes.

- **NIH Pancreas-CT Dataset (NIH)** The NIH Clinical Center performed 82 abdominal 3D CT scans[113] from 53 male and 27 female subjects. The subjects are selected by radiologists from patients without major abdominal pathologies or pancreatic cancer lesions.

	NIH Dataset				MSD Dataset				SYN Dataset			
	MAE	STD	P.C.	S.C.	MAE	STD	P.C.	S.C.	MAE	STD	P.C.	S.C.
Direct Regression	6.30	7.93	-18.36	-1.50	14.47	12.50	72.26	70.17	8.22	10.82	78.29	71.39
Direct Regression+Image	11.74	13.67	2.13	3.16	21.87	20.83	5.53	9.22	13.80	17.65	36.83	39.80
Jungo *et al.*[58]	3.51	3.98	82.21	61.95	11.86	16.31	71.24	77.71	9.45	20.61	73.32	79.93
Kwon *et al.*[68]	4.07	4.71	**82.41**	75.93	12.68	18.31	70.42	77.77	9.77	22.30	74.80	81.13
VAE-2 (53.93)	5.31	6.45	56.66	57.14	14.86	10.73	81.21	77.63	9.63	11.23	79.66	68.19
VAE-16 (72.46)	4.39	4.84	62.10	76.69	9.83	9.56	84.86	83.93	6.29	8.30	89.57	82.56
VAE-128 (76.00)	**2.89**	**3.60**	81.08	**82.86**	**8.14**	**9.14**	**86.23**	85.02	**4.93**	**7.20**	**90.92**	**86.07**
VAE-1024 (79.65)	3.50	4.15	73.78	80.90	8.42	9.24	85.81	**85.17**	5.71	8.00	88.61	85.98

Table 2.1: Comparison between our method and baseline methods. The target segmentation (BNN) algorithm is evaluated automatically without using ground truth. We have tried different structures for VAE (VAE-128 for 128-dimensional latent space). Of all the methods, VAE-128 achieves the highest performance. The numbers in brackets following the VAE methods are the average Dice score of reconstructing the ground truth masks on validation data. Usually with more accurate reconstruction of ground truth masks, the evaluation result is better but too accurate reconstruction may harm the evaluation capability (thinking of the identity mapping).

- **Medical Segmentation Decathlon (MSD)**[1] The medical decathlon challenge collects 420 (281 Training +139 Testing) abdominal 3D CT scans from Memorial Sloan Kettering Cancer Center. Many subjects have cancer lesions within pancreas region.

- **Synapse Dataset**[2] The multi-atlas labeling challenge provides 50 (30 Training +20 Testing) abdomen CT scans randomly selected from a combination of an ongoing colorectal cancer chemotherapy trial and a retrospective ventral hernia study.

The testing data of the last two datasets is not used in our experiment since we do not have their annotations. The segmentation algorithms we choose are V-Net [91], 3D Coarse2Fine [168], DeepLabv3 [19], and 3D Coarse2Fine with Bayesian structure [68]. The first two algorithms are based on 3D networks while the DeepLab is 2D-based. The 3D Coarse2Fine with Bayesian structure is employed to compare with the uncertainty-based method, and we denote it as Bayesian neural network (BNN) afterwards.

2.4.2 Baseline

Our method is compared with three baseline methods. Two of them are based on uncertainty and the last one directly applies regression network on the prediction mask to regress quality in equation (2):

[1] http://medicaldecathlon.com/index.html

[2] https://www.synapse.org/#!Synapse:syn3193805/wiki/217789

- **Entropy Uncertainty**. [58] calculated the pixel-vise predictive entropy using Bayesian inference. Then, the uncertainty is summed up over the whole image to get the doubt score which would replace the shape feature in (8) to regress the quality. The sum is weighted by the distance to predicted boundary, which somehow alleviates the bias distribution of uncertainty. Their method is done in 2D image and here we just transfer it to 3D image without essential difficulty.

- **Aleatoric and Epistemic Uncertainty**. [68] divided the uncertainty into two terms called aleatoric uncertainty and epistemic uncertainty. We implement both terms and calculate the doubt score in the same way as [58] because the original paper does not provide a way. The two doubt scores are used in predicting the quality.

- **Direct Regression**. A regression neural network is employed to directly learn the quality of predictive mask. It takes a segmentation mask as input and output a scalar for the predictive quality.

2.4.3 Implementation Detail

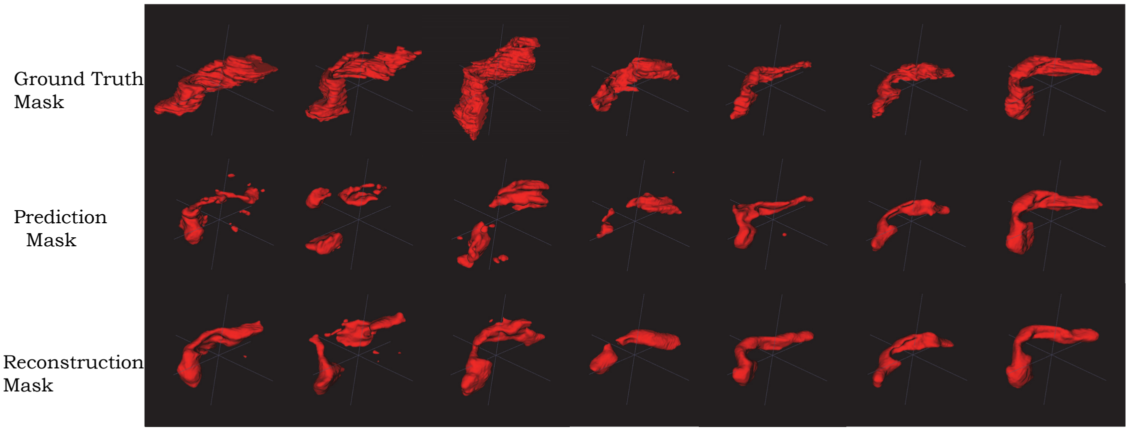
The structure of VAE is shown in Figure 4.1. We apply instance normalization on each convolution layer. The ReLU activation is applied on each layer except for the fully connected layer for mean value and the output layer is activated using the sigmoid function. The structure we use in the direct regression method is the encoder part of the VAE so that they are fair for comparison.

	3D Coarse2Fine					3D VNet				
	MAE	STD	P.C.	S.C.	Dice	MAE	STD	P.C.	S.C.	Dice
NIH	3.46	4.09	89.95	85.41	79.38	2.57	3.24	91.35	84.51	81.21
MSD	10.02	9.45	89.67	87.54	51.88	9.34	9.60	86.52	82.50	55.90
Synapse	6.24	9.00	92.39	84.29	62.10	5.67	7.28	91.65	80.11	64.93
	DeepLabV3					BNN				
	MAE	STD	P.C.	S.C.	Dice	MAE	STD	P.C.	S.C.	Dice
NIH	5.35	5.83	63.34	78.80	81.53	2.89	3.60	81.08	82.86	82.15
MSD	9.34	9.60	86.52	82.50	54.96	8.14	9.14	86.23	85.02	57.10
Synapse	5.67	7.28	91.65	80.11	61.03	4.93	7.20	90.92	86.07	66.36

Table 2.2: Results of different target segmentation algorithms are evaluated by our alarm system on different datasets. The Dice column means the average Dice score for the segmentation algorithm tested with groundtruth on different datasets, provided for reference. Our system achieves comparable performance as in Table 2.1 (see also in the right bottom cell) although the segmentation performance differs a lot between datasets. Without tuning parameters, our alarm system can be directly applied to evaluate other segmentation algorithms

For data pre-processing, since the voxel size varies from case to case, which would affect the shape of pancreas and prediction of segmentation, we first resample the voxel size of all CT scans and annotation mask to $1mm \times 1mm \times 1mm$. For training VAE, we apply simple alignment on the annotation mask. We employ a cube bounding box which is large enough to contain the whole pancreas region, centered at the pancreas centroid, then crop both volume and label mask out and resize it to a fixed size $128 \times 128 \times 128$. We only employ a simple alignment because the human pose is usually fixed when taking CT scans, e.g. stance, so that the organ will not rotate or deform heavily. For a segmentation prediction, we also crop and resize the predictive foreground to $128 \times 128 \times 128$ and feed it into VAE to capture the shape feature.

During the training process, we employ rotation for -10, 0, and 10 degree along x,y,z axes(27 conditions in total) and random translation for smaller than 5 voxel on



caseID	03	14	40	09	41	23	60
Real Dice	0.32	0.44	0.47	0.62	0.73	0.85	0.89
Fake Dice	0.57	0.50	0.47	0.65	0.72	0.85	0.83

Figure 2.4: We visualize the performance of our evaluation system on different qualities of segmentation results. The real Dice score increases from left to right. The fake Dice score is highly correlated with the real Dice so that we can get good prediction of real Dice by applying simple regressor on the fake Dice.

annotation mask as data augmentation. This kind of mild disturbance can enhance the data distribution but keep the alignment property of our annotation mask. We tried different dimension of latent space and finally set it to 128. We found that VAE with latent space of different dimension will have different capability in quality assessment. The hyper parameter λ in object function of VAE is set to 2^{-5} to balance the small value of Dice Loss and large KL Divergence. We trained our network by SGD optimizer. The learning rate for training VAE is fixed to 0.1. Our framework and other baseline models are built using TensorFlow. All the experiments are run on NVIDIA Tesla V100 GPU. The first training step is done in total 20000 iterations and takes about 5 hours.

2.4.4 Primary Results and Discussion

We split NIH data into four folds and three of them are used for training segmentation algorithms and VAE; the remaining one fold, together with all training data from MSD and Synapse datasets forms the validation data to evaluate our evaluation method. First we learn the parameter of VAE using the training label of NIH dataset. Then we choose BNN as the preparation algorithm mentioned in section 3.3. The training strategy in section 3.3 is applied on it to learn the parameters of regression. For all the baseline methods, we employ the same training strategy of jackknifing as in our method and choose the BNN as preparation algorithm for fair comparison. Finally we predict the quality of segmentation mask on the validation data for all the segmentation algorithms. Note that all segmentation algorithms are trained only on the NIH training set.

Table 2.1 compared our method and three baselines by assessing the BNN segmentation result of validation datasets. In general, our method achieves the lowest error and variance on all datasets. In our experiment, the preparation algorithm BNN achieves 82.15, 57.10 and 66.36 average Dice score tested on NIH, MSD and Synapse datasets respectively. The segmentation algorithm trained on NIH will fail on some cases of other datasets, and our alarm system still works well without tuning the parameters of VAE and regressor on other datasets. More detailed result is as shown in Figure 2.3. We can clearly observe that our method provides more accurate quality assessment result. For uncertainty-based methods, as shown in Figure 2.1, the uncertainty often distributes on the boundary of predicted masks but not on the missing parts or false positive parts and the transferability is not

	MSD Dataset Pancreas				MSD Dataset Tumor			
	MAE	STD	P.C.	S.C.	MAE	STD	P.C.	S.C.
Direct Regression	7.48	8.64	56.48	44.49	23.20	29.81	45.50	45.36
Jungo *et al.*[58]	7.24	8.79	54.38	49.29	26.57	29.78	-23.87	-20.23
Kwon *et al.*[68]	6.94	8.54	62.15	**61.20**	26.14	29.24	14.61	14.70
VAE-1024(Ours)	**6.03**	**7.63**	**68.40**	59.65	**20.21**	**23.60**	**60.24**	**63.30**

Table 2.3: Results for evaluating both pancreas and tumor segmentation. The MAE number for pancreas is better than those in Table 2.1 since there are more training samples in the MSD dataset. For tumor evaluation, all the methods are not doing well but our method reveal the strongest correlation between the real quality and the predictive quality. Since detecting tumor itself is a very hard task, the segmentation prediction for tumor is often with more variance. The alarm system needs more careful design to deal with that big variance.

strong since it relies on the segmentation algorithm. For direct regression method, we use the encoder part of VAE-1024 followed by a 2-layer fully connection. The training data of direct regression method is the augmentated testing data of F_1, F_2 on $\mathcal{X}_t^1$, $\mathcal{X}_t^2$ respectively as in section 3.3. So the number of training data for direct regression method is the same as ours but our method shows better capability of predicting the quality.

Table 2.2 shows the quality assessment results of our method for 4 different segmentation algorithms. The result of BNN is better because the preparation algorithm we use for training the regressor is also BNN. Without tuning parameters, our method remains reliable when the segmentation algorithms to be evaluated and the dataset to be tested on are changed, which shows strong transferability.

Why it works: In the experiments we use $S(F(X);\hat{\theta})$ as the input of regressor. However we find the second term of $S(F(X);\hat{\theta})$ is less related with the real Dice (So in Figure 4.1 we only put the fake Dice there, which is the first term of $S(F(X);\hat{\theta})$).

That means VAE can encode masks with bad shape into normal points in the latent space so that the reconstructions are of normal shape, which makes the fake Dice low. We visualize some cases in Figure 2.4 for showing this property of VAE. For bad segmentation predictions, the reconstruction masks from VAE indeed look more like a pancreas.

2.4.5 Ablation Experiments

We also run ablation experiments for different structures of VAE and for evaluating foreground without strong shape prior, tumor region.

2.4.5.1 Different VAE Structures:

Table 2.1 also shows results of VAE with latent space of different dimensions. With bigger latent space, VAE can reconstruct the ground truth masks better which generally indicates stronger evaluation capability. But for VAE-1024, the reconstruction is the best but the prediction result is not as good as VAE-128. We have also tried larger latent space like VAE-10000, and it can reconstruct the ground truth masks almost perfectly. But it is more like an identity mapping, making it impossible for the evaluation task.

2.4.5.2 Combine With Texture:

Since our alarm system only uses the information of segmentation masks, the texture information, which can be important in evaluating the segmentation quality, is missing. We tested it with a very intuitive setting, , for the direct regression method, we concatenate the image and segmentation masks together and use that as input for training the regression network. The result is shown in Table 2.1 “Direct Regression+Image". We see that with the same number of training data, the performance is even worse than only taking the segmentation mask as input. We think it is because the complex structure of image will confuse the regression network for learning the quality. [117] and [10] developed textured based methods on OCT and brain MRI data respectively, while in our experiments, it is hard to generate realistic abdominal CT scans. So how to better combine the texture with the segmentation mask is another direction worth exploring.

2.4.5.3 Evaluate Object With Large Shape Variance:

We also compare baseline methods and our method on evaluating segmentation of object with less stable shape tumor. The MSD dataset also provides voxel-wised label of pancreatic tumor. Instead of only evaluating the tumor prediction (requires accurate localization of tumor bounding boxes which is a hard task already), we evaluate both the tumor and pancreas segmentation at the same time so that we can use the bounding box of pancreas. Since this is a multi-class problem now, we adapt the VAE to take the one-hot encoding segmentation masks as input and change the original Dice loss to multi-class Dice loss. Similarly, we adapt the baseline methods so that they can fit in this multi-class evaluating problem. For direct regression method, it is trained to regress pancreas Dice score and tumor Dice score at the same time. For uncertainty-based method, uncertainty for both pancreas and tumor are calculated. We randomly split the MSD dataset into two parts and one is used for training while the other one for validation. For the training process we still apply the strategy as in section 3.3. We also train a BNN for pancreas and tumor segmentation as the target algorithm to evaluate and it reaches 72.52 and 35.34 average Dice score on pancreas and tumor respectively. The detailed comparison is shown in Table 2.3. For the uncertainty-based method, the tumor segmentation evaluation is quite bad because the segmentation algorithm often wrongly segments the tumor confidently, which also proves the limitation of uncertainty-based method on quality assessment. For the direct regression method, as there are more training data (60 → 140 before augmentation), the number is better than that in Table 2.1, which is common for a learning system. Our method still performs the best although it is not satisfactory, as there are many cases with 0 Dice score on tumor segmentation which are hard to predict the quality only from the segmentation mask. Note that the correlation between the real quality and predictive quality of our method is much stronger, which means even with weak shape prior, our method can still capture some useful information from the segmentation mask.

2.5 Conclusion

In the paper we presented a VAE based alarm system for segmentation algorithms which predicts the qualities of the segmentation results without using ground truth. We claim that the shape feature is useful in predicting the qualities of the segmentation results. To capture the shape feature, we first train a VAE using ground truth masks. We utilize the fact that rare events usually achieve larger loss value, and successfully detect the out-of-distribution shape according to the loss value in the testing time. In the second step we collect the segmentation results of the segmentation algorithm on the training data, and extract the shape feature of them to learn the parameters of regression. By applying jackknifing training on the preparation algorithm we can obtain more accurate regression parameters.

Our proposed method outperforms the standard uncertainty-based methods and direct regression methods, and possesses better transferability to other datasets and other segmentation algorithms. The reliable quality assessment results prove both that the shape feature capturing from VAE is meaningful and that the shape feature is useful for quality assessment in the segmentation task.

Chapter 3

Joint Shape Representation and Classification for Detecting PDAC

3.1 Introduction

Pancreatic cancer is a major killer causing hundreds of thousands of deaths globally every year. It often starts with a small set of localized cells multiplying themselves out of control and invading other parts of the body. The five-year survival rate of the patient can reach 20% [96] if the cancer is detected at an early stage, but quickly drops to 5% if it is discovered late and the cancerous cells have spread to other organs [123]. Therefore, early diagnosis of pancreatic cancer can mean the difference between life and death for the patients.

This paper deals with pancreatic ductal adenocarcinoma (PDAC), the major type of pancreatic cancer accounting for about 85% of the cases [123], and attempts to detect it by checking abdominal CT scans. The pancreas, even in a healthy state, is difficult to segment from a CT volume [114], partly because its 3D shape is irregular [151]. The segmentation, particularly for the cancer lesion area, becomes even more challenging when the pancreas is abnormal, *e.g.*, cystic [160]. In recent years, with the development of deep learning frameworks [66], researchers were able to construct effective deep encoder-decoder networks [84] for organ segmentation [111] or shape representation [13], boosting the accuracy of conventional models for a wide range of medical imaging analysis tasks.

The goal of this paper is to discriminate abnormal pancreases from normal ones[1]. This is a classification task, but directly training a volumetric classifier may suffer from over-fitting due to limited training data. Inspired by the fact that PDAC often changes the pancreas shape, we set shape representation as an intermediate goal, so as to constrain the learning space and regularize the model. Our framework contains two stages. First, we train an encoder-decoder network [149] for voxel-wise pancreas segmentation from CT scans[2]. Second, we use a joint shape representation and classification network to predicts if the patient suffers from PDAC. The weights of the shape representation module are initialized using an auto-encoder [13][49], and then jointly optimized with the classifier. Joint optimization improves classification accuracy at the testing stage.

The radiologists in our team collected and annotated a dataset with 436 CT scans, including 300 normal cases and 136 PDAC cases. Our approach achieves a sensitivity of 80.2% at a specificity of 90.2%, *i.e.*, finding 4/5 of abnormal cases with false alarms on only 1/10 of the normal cases. Some detected PDAC cases contain tiny tumors, which are easily missed by segmentation algorithms and even

[1] Throughout this paper, an *abnormal* pancreas is defined as one suffering from PDAC.

[2] To make our approach generalized, we do not assume the tumors are annotated in the training set, and so we do not perform tumor segmentation.

some professional radiologists. According to the radiologists, our approach can provide auxiliary cues for clinical purposes.

3.2 Detecting PDAC in Abdominal CT Scans

3.2.1 The Overall Framework

A CT-scanned image, $\mathbf{X}$, is a $W \times H \times L$ matrix, where W, H and D are the width, height and length of the cube, respectively. Each element in the cube indicates the Hounsfield unit (HU) at the specified position. Each volume is annotated with a binary pancreas mask $\mathbf{S}^{\star}$ which shares the same dimensionality with $\mathbf{X}$. Our goal is to design a discriminative function $p(\mathbf{X}) \in \{0, 1\}$, with 1 indicating that this person suffers PDAC and 0 otherwise.

Our idea is to decompose the function into two stages. The first stage is a segmentation model $\mathbf{f}(\cdot)$ for voxel-wise pancreas segmentation, *i.e.*, where $\mathbf{S} = \mathbf{f}(\mathbf{X})$. The second stage is a mask classifier $c(\cdot)$ which assigns a binary label to the mask $\mathbf{S}$. To make use of shape information, $c(\cdot)$ is further decomposed into a shape encoder $\mathbf{g}(\cdot)$ which produces a compact vector $\mathbf{v} = \mathbf{g}(\mathbf{S})$ to depict the shape properties of the binary mask $\mathbf{S}$, and a shape classifier $h(\cdot)$ which determines if the shape vector $\mathbf{v}$ corresponds to a pancreas suffering from PDAC.

Therefore, the overall framework, shown in Figure 3.1, can be written as:

$$p(\mathbf{X}) = c \circ \mathbf{f}(\mathbf{X}) = h \circ \mathbf{g} \circ \mathbf{f}(\mathbf{X}). \tag{3.1}$$

We can of course design an alternative function, *e.g.*, a 3D classifier which works on

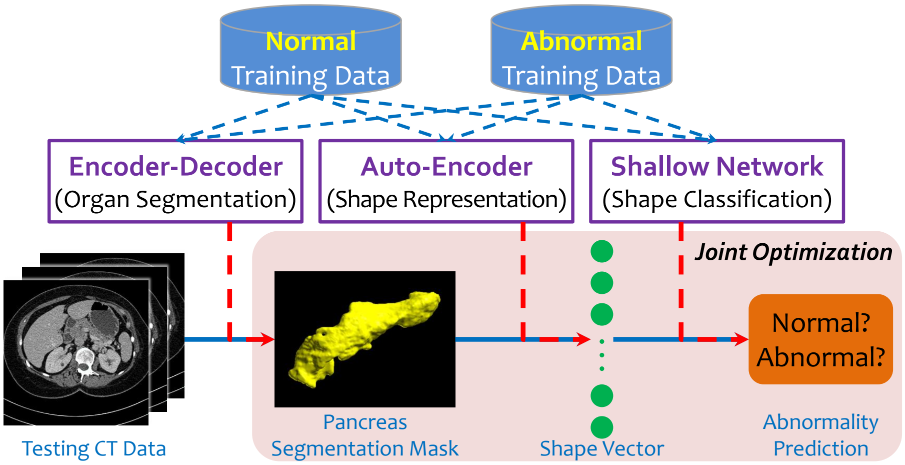


Figure 3.1: The overall framework of our approach (best viewed in color).

CT image data directly, but our stage-wise model makes use of the prior knowledge from the radiologists, *i.e.*, PDAC often changes the shape of the pancreas. This sets up an intermediate goal of optimization and shrinks the search space of our model, which is especially helpful in preventing over-fitting given limited training data. In addition, this also enables us to interpret our prediction. We will show in experiments that, without such prior knowledge, the classifier produces unstable results and less satisfying prediction accuracy.

3.2.2 Pancreas Segmentation by Encoder-Decoder Networks

Our approach starts with an encoder-decoder network for pancreas segmentation. There are typically two choices, which differ from each other in the way of processing volumetric data. The first one applies 2D segmentation networks [111][114] from orthogonal planes, while the other one trains a 3D network directly [90] in a patch-based manner. Either method requires cutting volumetric data into 2D slices or 3D patches at both training and testing stages. As a result, the segmentation function $\mathbf{S} = \mathbf{f}(\mathbf{X})$ cannot be optimized together with the subsequent modules, namely shape representation and classification.

In practice, we apply a recent 2D segmentation approach named STN [149] for pancreas segmentation. It trains three models from the *coronal*, *sagittal* and *axial* planes, respectively. In our own dataset, STN works very well, providing an average DSC of over 87% for normal pancreas segmentation, and over 70% for abnormal pancreas segmentation. We make two comments here. First, the segmentation accuracy of 87% almost reaches the agreement between two individual annotations by different radiologists. Second, the abnormal pancreases are often more difficult to segment, as their appearance and geometry properties can be changed by PDAC. However, as shown later, such imperfections in segmentation only cause little accuracy drop in abnormality classification.

3.2.3 Joint Shape Representation and Classification

Based on pancreas segmentation $\mathbf{S} = \mathbf{f}(\mathbf{X})$, it remains to determine the abnormality of this pancreas. We achieve this by first compressing the segmentation mask into a low-dimensional vector $\mathbf{v} = \mathbf{g}(\mathbf{S})$ to compress $\mathbf{v}$, and then applying a classifier $h(\cdot)$ on top of $\mathbf{v}$.

The shape representation network $\mathbf{g}(\cdot)$ involves down-sampling the segmentation mask gradually. Following [13], this is implemented by a series of 3D convolutional layers. The detailed network configuration is shown in Figure 3.2.

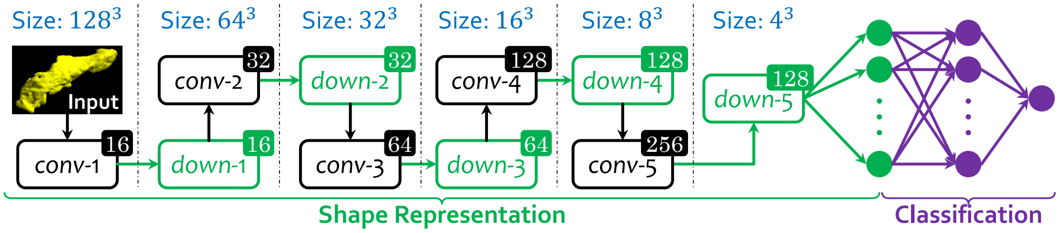


Figure 3.2: Shape representation and classification network (best viewed in color). Each rectangle is a layer, with the number at the upper-right corner indicating the number of channels. Each convolution (*conv*) layer contains a set of $3 \times 3 \times 3$ kernels, and each down-sampling (*down*) layer uses $2 \times 2 \times 2$ convolution with a stride of 2. Batch normalization and ReLU activation are used after all these layers. The last layer in shape representation (the green neurons) is the low-dimensional shape vector, followed by a 2-layer fully-connected network for classification.

Regarding the dimensionality of the shape vectors (*i.e.*, the number of output neurons), a high-dimensional representation carries more information, but also risks over-fitting under limited training data. We analyze this parameter in experiments. Essentially, both segmentation and shape representation networks perform image down-sampling. The former starts with the raw input image and thus requires complicated and expensive computations. The latter, however, is much simpler, with the network much shallower, which processes the entire volume at once. This makes it possible to be optimized together with the classifier.

In the final step, we implement $h(\cdot)$ as a 2-layer fully-connected network. The simplicity of $h(\cdot)$ aligns with our motivation, *i.e.*, the vector $\mathbf{v}$ carries discriminative shape information which is easy to classify. Being a differentiable module, it can the optimized with the shape representation network in a joint manner (details are elaborated below), which brings consistent accuracy gain.

The training process starts by sampling a segmentation mask $\mathbf{S}$ from training data. We first perform slight rotation (0° or $\pm 10^\circ$ along three axes individually, 27 possibilities) as data augmentation, and rescale the region within the minimal bounding box into $128 \times 128 \times 128$. Note that direct optimization on $h \circ \mathbf{g}(\cdot)$ cannot guarantee that $\mathbf{g}(\cdot)$ learns shape information. In addition, direct optimization can lead to over-fitting with limited training data, even after data augmentation (see experiments). Hence, we use a two-step method for gradual optimization.

In the first step, we deal with $\mathbf{g}(\cdot)$ by concatenating this module with a decoder network $\tilde{\mathbf{g}}(\cdot)$, which performs reverse operations (all convolutions are replaced by deconvolutions) to restore the compressed vector into the original image. This framework, named an auto-encoder [13][49], can be trained in a weakly-supervised manner, *i.e.*, given an input mask $\mathbf{S}$, we can minimize the difference between $\mathbf{S}$ and $\tilde{\mathbf{S}} = \tilde{\mathbf{g}} \circ \mathbf{g}(\mathbf{S})$ by minimizing the loss function $\mathcal{L}_\mathrm{S}\big(\mathbf{S}, \tilde{\mathbf{S}}\big)$. This forces the compressed vector $\mathbf{v}$ to store sufficient information in order to restore $\mathbf{S} = \tilde{\mathbf{g}}(\cdot)$. Auto-encoder provides a reasonable initialization for $\mathbf{g}(\cdot)$ in the next step (joint optimization). We use a mini-batch size of 1 and train the auto-encoder for 40,000 iterations with a fix learning rate of 10^{-6}.

The second step optimizes $\mathbf{g}(\cdot)$ and $h(\cdot)$ jointly. We use the cross-entropy loss $\mathcal{L}_\mathrm{C}(y, p) = y \ln p + \eta \cdot (1 - y) \ln(1 - p)$ where y is the ground-truth and $p = h \circ \mathbf{g}(\mathbf{S})$ is the predicted confidence. η performs class-balancing to avoid model bias. The mini-batch size is still set to be 1, and we perform a total of 40,000 iterations. We start with a learning rate of 0.0005, and divide it by 10 after 20,000 and 30,000 iterations. To maximally preserve stability, we freeze all weights of $\mathbf{g}(\cdot)$ in the first 5,000 iterations, so that the 2-layer network $h(\cdot)$, initialized as scratch, is reasonably trained before being optimized together with $\mathbf{g}(\cdot)$.

Last but not least, there is an alternative way of jointly optimizing $\mathbf{g}(\cdot)$ and $h(\cdot)$, *i.e.*, applying a discriminative auto-encoder [110], which preserves the shape restoration loss in the second step and optimizes $\mathcal{L}_{\mathrm{S}}(\mathbf{S}, \tilde{\mathbf{S}}) + \lambda \cdot \mathcal{L}_{\mathrm{C}}(y, p)$. We do not use this strategy because our ultimate goal is classification – shape representation is an important cue, but we do not hope the constraints in shape restoration harms classification accuracy. In experiments, we find that a discriminative auto-encoder produces less stable classification accuracy.

3.3 Experiments

3.3.1 Dataset and Settings

To the best of our knowledge, there are no publicly available datasets for PDAC diagnosis. We collect a dataset with the help of the radiologists in our team. There are 300 normal CT scans and 136 biopsy-proven abnormal (PDAC) cases, and all of them were scanned by the same machine. The pancreas annotation was done by four expert in abdominal anatomy and each case was checked by a experienced board certified Abdominal Radiologist. The spatial resolution of our data is relatively high, *i.e.*, the physical distance between the neighboring voxels is 0.5mm in the long axis, and varies from 0.5mm to 1.0mm in the other two axes. We do not use data scanned from other types of machines (*e.g.*, the NIH dataset [114]) to avoid dataset bias, *i.e.*, the classifier works by simply checking the spatial resolution or other meta-information of the scan.

Dimension	SVM		2LN (I)		2LN (J)	
	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.
128	73.4 ± 3.1	87.8 ± 2.9	77.5 ± 2.2	87.6 ± 1.5	79.3 ± 1.0	89.9 ± 1.0
256	75.0 ± 1.9	87.6 ± 3.2	78.2 ± 1.6	89.1 ± 1.2	79.0 ± 0.4	90.5 ± 0.8
512	78.1 ± 1.9	89.5 ± 1.0	80.7 ± 1.5	88.3 ± 1.0	79.0 ± 0.8	**90.9** ± 0.9
1,024	75.0 ± 0.0	89.0 ± 0.0	78.8 ± 0.7	90.5 ± 0.6	**80.2** ± 0.5	90.2 ± 0.2

Table 3.1: The sensitivity (sens., %) and specificity (spec., %) reported by different approaches and dimensionalities of shape. We denote the models optimized individually and jointly by (I) and (J), respecitively. All these numbers are the average over 5 individual runs. 2LN (J) with 1,024-dimensional vectors has the best average performance.

We use 100 normal cases for training the STN [149] and auto-encoder [13] for pancreas segmentation and shape representation, respectively. The remaining 200 normal and 136 abnormal scans are first segmented using the STN then compressed by the auto-encoder. These examples are randomly split into 4 folds, each of which has 50 normal and 34 abnormal cases. We perform cross-validation, *i.e.*, training a classifier on three folds and testing it on the remaining one. We report the sensitivity and specificity of different models.

3.3.2 Quantitative Results

Results are summarized in Table 3.1. To compare with the joint training strategy, we provide two other competitors, namely a support vector machine (SVM) and the individually-optimized 2-layer network (equivalent to freezing the parameters in the auto-encoder throughout the entire training process). We observe consistent accuracy gains brought by the proposed approach over both competitors, in particular the 2-layer network optimized individually. This stresses the importance and effectiveness of joint optimization. Regarding other options, we find that the classification accuracy of our approach either drops or becomes unstable if we (i) train the entire network from scratch; (ii) preserve the shape restoration loss with classification loss; or (iii) do not freeze the weights of the auto-encoder in the early training sections.

In clinics, an important issue to consider is the tradeoff between sensitivity and specificity. A higher sensitivity implies that more abnormal cases are detected, but also brings the price of a lower specificity. Our approach, by simply tuning the classification threshold, can satisfy different requirements. The ROC curves of different models are shown in Figure 3.3. Using our best model (1,024-dimensional shape vector with joint optimization), we can achieve a sensitivity of 95% at a specificity of 53.8%, or a specificity of 95% at a sensitivity of 67.9%.

3.3.3 Qualitative Analysis

We first investigate the relationship between pancreas segmentation quality and classification accuracy. Trained on a standalone set of 100 normal cases, STN reports average DSCs of 86.66% and 71.45% on the 200 testing normal and 136 abnormal cases, respectively. The radiologists randomly checked around 20 cases, and verified that our segmentation results, especially on the normal pancreases, have achieved the level of being used for diagnosis. We also use the ground-truth segmentation masks of these $200 + 136$ pancreases in classification. With 1,024-dimensional shape vectors, the sensitivity and specificity of the SVM classifier are improved by by 9.0% and 2.0%, and these numbers for the 2-layer network are 5.6% and 0.6%, respectively. This indicates that the imperfection of abnormal pancreas segmentation mainly causes drops in sensitivity. But, built on top of automatic segmentation, our framework can be applied to a wide range of scenarios where the manual annotation is not available.

Next, we consider the accuracy of shape representation, or more specifically, the similarity between the restored segmentation mask and the original one. It is obvious that a higher dimension in shape representation stores richer information and thus produces more accurate restoration. However, as shown in Table 3.1, we do not observe significant gain brought by high dimensionalities. This verifies our assumption, *i.e.*, the classifier does not require accurate shape reconstruction. This also explains the advantage of joint optimization, in which the classifier can capture discriminative information from shape representation, and the shape model can also adjust itself to help classification.

We visualize several successful and failure examples in Figure 3.3. Our approach is able to detect some cases with tiny tumors which are easily missed even by the radiologists[3]. On the other hand, our approach is likely to fail when the pancreas segmentation is less accurate, leading to a strange pancreas shape which is not seen in training data and thus confuses the classifier. One false-negative and one false-positive cases are shown in Figure 3.3.

Finally, we point out that there is an alternative to our approach, which directly trains segmentation/detection networks to find the tumors in these PDAC cases. In comparison, our approach has two advantages. First, we do not require the

[3] The early diagnosis of PDAC is difficult and can be uncertain from CT scans. In our case, the radiologists proved these PDAC cases with biopsy checks. They can easily miss some of these cases if they were not told their abnormality beforehand.

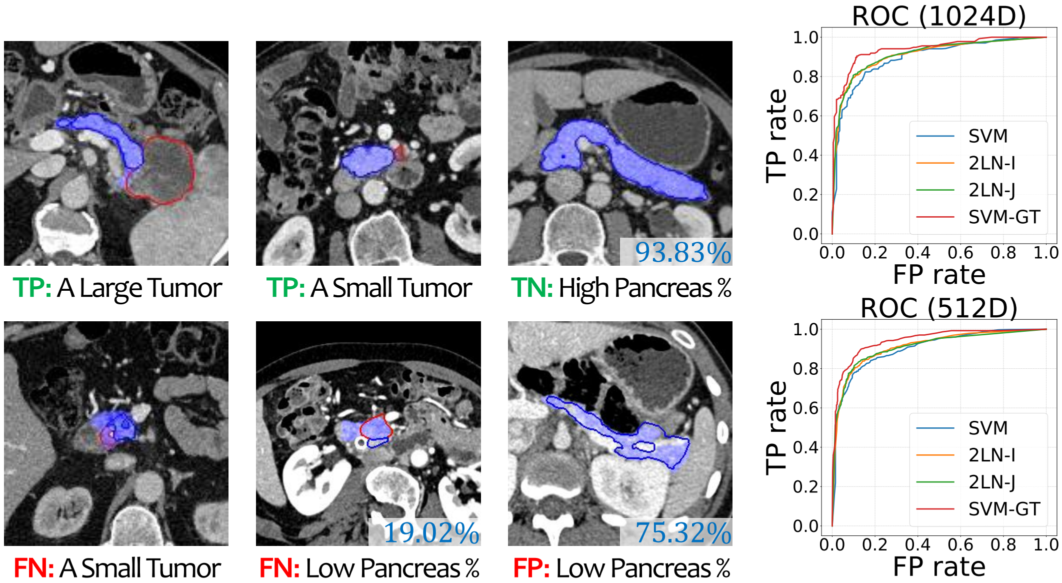


Figure 3.3: Left: classification results by our approach. Right: the ROC curves. Red and blue contours mark the labeled pancreas and tumor, and blue regions indicate the predicted pancreas. TP, TN, FP, FN: for {true,false}-{positive,negative}, respectively.

tumors to be annotated in the training data, which is an extremely challenging task. Second, our approach can detect some PDAC cases with very small tumors (which largely changed the shape of the pancreas) that are missed by segmentation. We train a tumor segmentation network individually, and find that more than half of the false negative can be recovered by our approach. This suggests that shape representation serves as an auxiliary cue. However, a clear drawback of our approach is not being able to find the exact position of the lesion area. In all, our approach provides an important cue (shape), and it can be integrated with other cues in the future towards more accurate diagnosis, *e.g.*, when voxel-wise tumor annotations are available, we can incorporate pancreas/tumor segmentation into our joint optimization framework.

3.4 Conclusions

Our approach is motivated by knowledge from surgical morphology, which claims that the pancreatic ductal adenocarcinoma (PDAC) can be discovered by observing the shape change of the pancreas. We first use an encoder-decoder network to obtain pancreas segmentation, and design a joint framework for shape representation and classification. We initialize shape representation using an auto-encoder, and optimize it with the classifier in a joint manner.

In experiments, our approach achieved a sensitivity of 80.2% with a specificity of 90.2%. It even detected several challenging cases which are easily missed by the radiologists. Given a larger amount of training data, we can expect even higher performance. Our future research directions also involve adding other cues (*e.g.*, tumor segmentation) and training the entire framework in a joint manner.

Chapter 4

FusionNet: Incorporating Shape and Texture for Abnormality Detection in 3D Abdominal CT Scans

4.1 Introduction

Pancreatic cancer is one of the most dangerous type of cancer. In 2019, about 56770 people will be diagnosed with pancreatic cancer, and pancreatic cancer accounts for about 3% of all cancers in the US and about 7% of all cancer deaths [1]. The 5-year relative survival rate for all stages of pancreatic cancer is only about 9%, while it can rise to 34% if the cancer is detected in an early stage. However, even experienced doctors may miss an early stage cancer because it is small and hard to observe. So developing an reliable automatic system to assist doctors to diagnosis can help decrease the missing rate of patients with early stage of cancer.

This paper is aimed at discriminating normal cases from cases with pancreatic ductal adenocarcinoma (PDAC), the major type of pancreatic cancer accounting for about 85% of the cases, by checking into the abdominal 3D CT scans. With the development of deep learning in recent years [67], researchers have made significant progress in automatically segmenting organs like pancreas from CT scans [114, 162, 167], which is already a hard task due to the irregular shape of pancreas [151]. Even though, segmenting the lesion region is an even more challenging task due to the large variation in shape, size and location of the lesion [160]. And the full annotation for the lesion region requires more expertise and time to obtain. So instead of directly segmenting the lesion region, detecting the patients with PDAC can already help the diagnosis and, more importantly, is more feasible when the annotation is limited.

We choose to utilize the segmentation mask and CT image for pancreatic abnormality detection, since the segmentation mask can represent the shape while the CT image represents the texture, which are both important for abnormality detection. However we find that the classification results of using only shape and only texture information are quite complementary, which motivates us to combine them in a unified system and thereby can improve the classification outcome. In the natural image domain, how to effectively combine different information has been explored in several different works. [39] proposes a fusion network incorporating depth to improve the segmentation. [138] calculates the normal, depth and silhouette from a single image for better 3D reconstruction. Other works like [140, 159] build different networks for different views of the same data and present co-training strategy to enable the models to incorporate different views.

In this paper we develop a two-stage method for this problem. Firstly, a recent state-of-the-art segmentation network [150] is used to segment the pancreas and then tested on all the data to get the prediction mask for pancreas. Secondly, the CT image is fed into a deep discriminator together with the prediction mask. The discriminator is employed to extract information from both the image and segmentation mask for abnormality classification. We optimize the architecture of the discriminator by searching from a functional space, which includes functions with different fusion strategies. Unlike [169] that needs full annotation for the lesion, our method only requires annotation masks for the pancreas region on cases without PDAC in the first stage, and image-level labels indicating abnormality in the second stage. Other works like [20, 78] make use of the information from either the prediction mask or CT image for classification. We show in the experiments that these two kinds of information are complementary to each other and the combination can improve the classification result by a large margin.

We test our framework on 200 normal and 136 abnormal (with PDAC) CT scans. We report a 92% sensitivity and 97% specificity, *i.e.* missing 11 out of 136 abnormal cases with 6 false alarms out of 200 normal cases. Compared with using only single branch, our method improves the result by more than 5% in specificity and 10% in sensitivity.

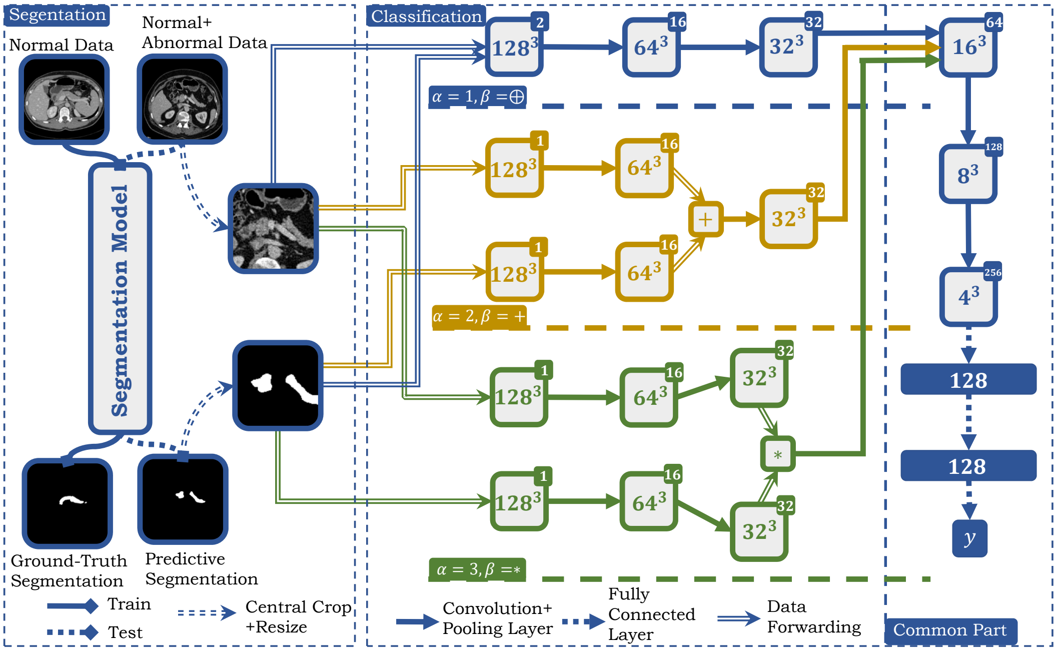


Figure 4.1: The pipeline of our framework. In stage 1, a segmentation network is trained using the normal data. Then the segmentation network is tested on both normal and abnormal data. The 3D mask and image are cropped and scaled as the input of second stage. At the right side, we show the examples of fusion model using different α, β. Note that these three models share the same architecture after layer 3 because $\alpha <= 3$ in the examples but they do not share the weights. Each convolution layer uses a set of $3 \times 3 \times 3$ kernels, and each pooling layer uses $2 \times 2 \times 2$ kernels with a stride of 2. Batch normalization and ReLU activation are used after all these layers.

4.2 Fusion Network for detecting PDAC

4.2.1 The Overall Framework

The CT scan $\mathbf{X} \in \mathcal{X}$ is a volume of size $L \times W \times H$, where L,W,H represents the length, width and height of the volume respectively. Typically, a CT scan is of size $512 \times 512 \times H$, where H is the number of slices along the axial axis. Each element in the volume indicates the Hounsfield Unit (HU) at a certain position. Our goal is to learn a discriminative function $f(\mathbf{X}) \in \{\mathbf{0}, \mathbf{1}\}$, where 1 indicates PDAC and 0 otherwise.

Directly learning the function $f(\cdot)$ is feasible but not optimal. Because the high dimensionality and rich texture information in the CT image can easily make the model overfit, especially when the number of training data is limited. [78] introduces a constraint by segmenting the pancreas first and learn $f(g(\mathbf{X}))$, where $g(\cdot)$ is a segmentation function to get a binary mask of pancreas $\mathbf{S}$. However this will result in loss of texture information since $g(\mathbf{X})$ is only a binary mask. In order to fully exploit both shape and texture information we consider learning

$$f(g(\mathbf{X}), \mathbf{X}),$$

which takes both the segmentation mask and image as input. The major problem here is how to design the function $f(\cdot)$ so that it can well extract shape information from $g(\mathbf{X})$ and texture information from $\mathbf{X}$ and combine them for the classification task. Our idea is to define a functional space representing a set of different fusion strategies and the optimal architecture is obtained by searching that functional space. Given a normal CT dataset $\mathcal{X}_1 = \{(\mathbf{X}, \mathbf{Y})\}$, where the annotation for pancreas $\mathbf{Y}$ is available, and $\mathcal{X}_2 = \{(\mathbf{X}, \mathbf{z})\}$ which contains both normal and abnormal cases with only image-level label $\mathbf{z}$ indicating the abnormality, we split our framework into two stages. First we train a segmentation function $g(\cdot)$ on $\mathcal{X}_1$ and test it on $\mathcal{X}_2$, then the prediction masks together with CT images on $\mathcal{X}_2$ become the input for the second stage to train a classification function $f(\cdot)$. We will introduce each stage in detail in the following sections.

4.2.2 The Segmentation Stage

This stage is necessary in the framework for getting the segmentation mask which will provide shape information in the second stage. Since the focus in this paper is how to combine $g(\mathbf{X})$ and $\mathbf{X}$ in $f(\cdot)$, and also the two stages are executed separately, so the form of $g(\cdot)$ is out of range of this study and will be investigated in the future. In this paper we choose a recent stat-of-the-art segmentation framework [150] for g. Since $g(\cdot)$ is a 2D-based method so we need to concatenate the output of different slices to reconstruct the 3D volume like in [141]. We train the segmentation algorithm on $\mathcal{X}_1$ and test it on $\mathcal{X}_2$. After that, we crop out the region-of-interest(ROI) from both CT image and prediction mask, defined as the cube bounding box covering all foreground voxels in the prediction mask and padded by 20 voxels in each dimension. Then the cropped regions are resampled to $128 \times 128 \times 128$ volumes. We denote the predictive mask after cropping and resampling as $\hat{\mathbf{S}} = \mathbf{g}(\mathbf{X})$.

4.2.3 The Classification Stage

The two branches of the input represent different information. The image domain contains rich texture information, while the binary mask can indicate shape of the target object. Directly concatenating them in the very first layer is an intuitive way but may not be optimal. To explore the optimal fusion strategy, we start from a base model with $L = 6$ convolution layers similar with 3D VNet [92], followed by two fully connected layers, as shown in Figure 4.1. Then a functional space for different architectures is defined as $\{(\alpha, \beta) | \alpha \in \{1, 2, ..., L\}, \beta \in \{+, *, \oplus\}\}$, where α indicates at which layer to fuse and β indicates how to fuse. Here $\oplus$ represents concatenation. See also in Figure 4.1 for specific examples for different combination of α, β. We formulate each fusion function in the functional space $f_{\alpha\beta}(\cdot)$ as following.

$$f_{\alpha\beta}(\mathbf{S}, \mathbf{X}; w) = f_{\alpha:L}(\beta(f_{1:\alpha}(\mathbf{S}; w^1_{1:\alpha}), f_{1:\alpha}(\mathbf{X}; w^2_{1:\alpha})); w_{\alpha:L}).$$

Here $f_{1:\alpha}(\cdot)$ is the first α convolution layers of the base model while $f_{\alpha:L}(\cdot)$ is the remaining layers. $w = \{w^1_{1:\alpha}, w^2_{1:\alpha}, w_{\alpha:L}\}$ is the parameters to learn. The feature maps of two branches after the first α layers are fused using operation $\beta(\cdot)$ as $\beta(f_{1:\alpha}(\mathbf{S}; w^1_{1:\alpha}), f_{1:\alpha}(\mathbf{X}; w^2_{1:\alpha}))$, and then fed into $f_{\alpha:L}(\cdot)$. The idea of this design is to alleviate the effect of changing the model structure but only focus on finding the best way to combine two different input.

Once given α, β, we learn w by optimizing a weighted cross-entropy loss

$$L = -\lambda \log p^z - (1 - \lambda) \log(1 - p)^{1-z},$$

where $p = f_{\alpha\beta}(\hat{S}, X; w)$. The output of $f_{\alpha\beta}(\cdot)$ is activated by a sigmoid function so that $p \in [0, 1]$. $z \in \{0, 1\}$ is the label for a CT scan indicating whether this study suffers from PDAC. We set $\lambda = 0.7$ for balancing the class difference during training.

4.3 Experiments

In this section we test our two-stage framework on our dataset containing 3D abdominal CT scans with both patients with and without PDAC. We compare our method with other method using single source input and also show the result of different fusion architectures. We report the sensitivity(SEN), specificity(SPEC), ROC AUC Score(AUC) and F1 Score(F1) to evaluate the classification model.

4.3.1 Dataset and Settings

We collect the dataset with the help of the radiologists. There are 300 normal cases and 136 biopsy proven PDAC cases. 100 out of 300 normal cases have voxel-wise annotations for pancreas (denoted as set $\mathcal{X}_1$), and the remaining 200 normal cases as well as the 136 PDAC cases only have image-level labels, *i.e.*, abnormal/normal (denoted as set $\mathcal{X}_2$). In the first stage, we train the segmentation network on $\mathcal{X}_1$ and test it on $\mathcal{X}_2$. In the second stage, $\mathcal{X}_2$ is randomly split into four folds for cross-validation, where each fold contains 50 normal and 34 abnormal cases and the fusion network is trained on three of the folds and tested on the remaining one.

For the first stage, we follow the instruction of [150] to train a segmentation network. For the second stage, we apply grid search on α and β, *i.e.* we choose for every pair of $(\alpha, \beta) \in \{(\alpha, \beta) | \alpha \in \{1, 2, ...L\}, \beta \in \{+, *, \oplus\}\}$. In our case, $L = 6$, so there are $3L = 18$ different architectures in total in the search space. After setting α and β, for training $f_{\alpha\beta}(\cdot)$, we use stochastic gradient descent(SGD) with batch size of 4. The learning rate is set to 0.01 with exponential decay rate 0.9997. We

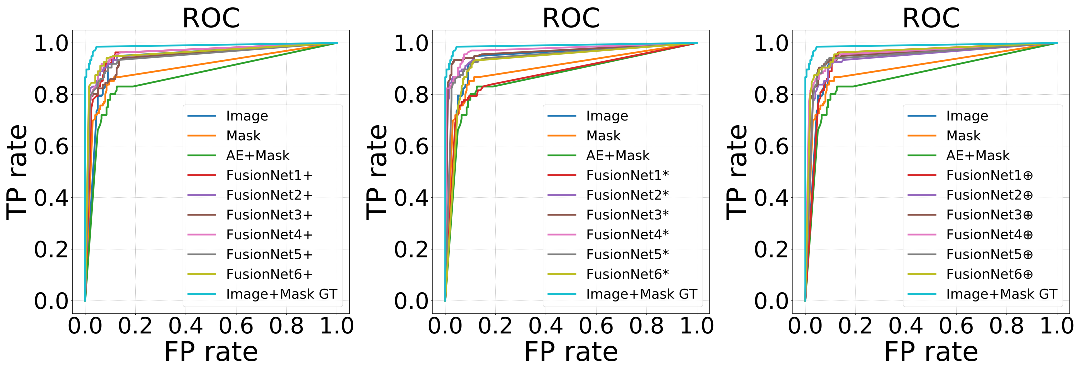


Figure 4.2: ROC curves for comparison of different fusion strategies. **Left**: fused by $+$. **Mid**: fused by $*$. **Right**: fused by $\oplus$. The Image, Mask and AE+Mask are the baseline methods without fusing. The Image+Mask GT is the pseudo upper bound of the fusing.

also perform data augmentation on both the CT image and prediction mask by slightly rotating $0°$, $\pm 10°$ along three axes individually (27 possibilities) to prevent from overfitting, since the number of training data is very limited. For each pair of α, β, the model is trained for 10,000 iterations, which takes about 1.5 hours on a NVIDIA TITAN RTX(24GB) GPU.

4.3.2 Primary Results

We compare our method with [78] which utilizes the feature from pre-trained auto-encoder for classification (AE+Mask). We also compare with the base model using either the CT image (Image) or prediction mask (Mask) as input. Our fusion model has the same network structure with the base model after the fusing point for fair comparison. The best result is achieved when fusing the two branches in third layer with multiplication operation. The result is summarized in Table 4.1. The ROC curves of different models are shown in Figure 4.2. Image+Mask GT indicates

	SEN	SPEC	AUC	F1
AE+Mask [78]	77.94	91.00	89.04	81.54
Mask	82.35	91.50	92.94	84.53
Image	83.09	92.00	95.95	85.28
Naive Fusion	83.09	95.50	97.17	87.60
FusionNet3*(Ours)	**92.65**	**97.00**	**97.72**	**94.03**
Mask+Image GT	94.12	97.50	99.53	95.17

Table 4.1: Comparison between our method and baseline methods on the sensitivity(SEN), specificity(SPEC), area under the curve(AUC) and F1 score(F1). FusionNet3* achieves the best result, indicating the best way to fuse is to multiply two branches in the third layer.

the strategy that if either one of the two methods (Image and Mask) correctly classifies the case, then we treat this case as correctly classified. This can be the upper bound of merging because it fuses the result based on the ground-truth label. The large improvement in the upper bound result shows that the information provided by the CT image and prediction mask for abnormality detection are quite complementary to each other, which proves the necessity of combining them together. Naive Fusion is done by taking the average of output from Mask and Image and only shows limited improvement. This fact further validates the efficacy of our proposed FusionNet.

4.3.3 Analysis and Discussion

4.3.3.1 Single Branch Comparison:

From Table 4.1 we can see the comparison among Image, Mask and AE+Mask which all use only one branch of information. Using only the image works the best, which indicates the importance of texture for detecting PDAC. For the other two

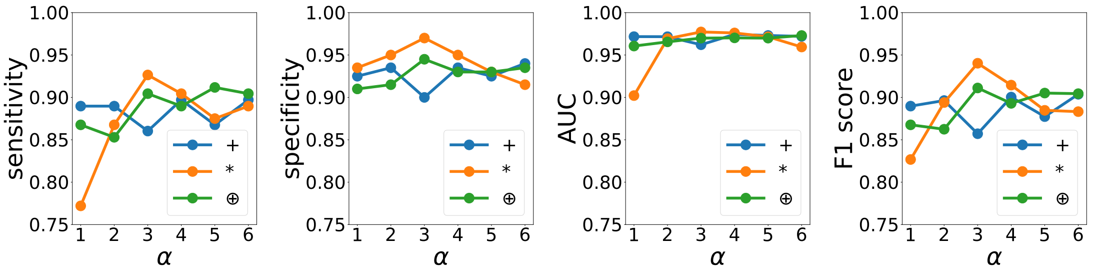


Figure 4.3: Comparison on the sensitivity, specificity, AUC and F1 score between different fusion architectures.

methods using only shape information, directly training a discriminator achieves better results, showing that the constraint of auto-encoder can harm the classification performance.

4.3.3.2 Fusion Comparison:

The result of fusing at different layers with different operations is as shown in Table 4.2 and Figure 4.3. First of all, almost all the fusion models can perform better than the single branch model, which proves the advantages of fusing shape and texture. Table 4.2 shows the number of parameters and floating-point operations for each model to show how the size of model affects the classification result. We can see as α increases, the size of model increases, but the classification result does not always improve correspondingly. For the $+$ fusion operation, the performance is better when fusing at the earlier or later layers of the network. For the $*$ and $\oplus$ operation, however, fusing at the middle layer of the network shows better performance. The best result is obtained when fusing at the third layer with $*$ operation.

α		1	2	3	4	5	6
$\beta = +$	F1	88.97	89.63	85.71	90.04	87.73	90.37
	# Para	3.99M	3.99M	4.01M	4.10M	4.45M	5.86M
	FLOPs	7.99M	7.99M	8.04M	8.21M	8.92M	11.74M
$\beta = *$	F1	82.68	89.39	94.03	91.45	88.48	88.32
	# Para	3.99M	3.99M	4.01M	4.10M	4.45M	5.86M
	FLOPs	7.99M	7.99M	8.04M	8.21M	8.92M	11.74M
$\beta = \oplus$	F1	86.76	86.25	91.11	89.30	90.51	90.44
	# Para	3.99M	4.01M	4.07M	4.32M	5.34M	7.96M
	FLOPs	7.99M	8.02M	8.15M	8.66M	10.69M	15.94M

Table 4.2: Comparison between different fusion architectures on the F1 scores, number of parameters and floating-point operations(FLOPs). As α increases, the two branches fuse at more latter layer, so the parameters number also increases. However the best performance is achieved when $\alpha = 3$.

4.4 Conclusion

In this paper we propose a FusionNet which combines shape and texture information from the segmentation mask and CT image for detecting PDAC. Compared with using only single source of information, using both shape and texture information improves the performance by a large margin. We also explore the best network structure for fusing these two branches together by searching from a functional space, which is to multiply the feature map of two branches in the middle of the network. We report a 92% sensitivity and a 97% specificity by doing 4-fold cross-validation on 200 normal patients and 138 patients with PDAC.

Chapter 5

JSSR: A Joint Synthesis, Segmentation, and Registration System for 3D Multi-Modal Image Alignment of Large-scale Pathological CT Scans

5.1 Introduction

Image registration attempts to discover a spatial transformation between a pair of images that registers the points in one of the images to the homologous points in the other image [137]. Within medical imaging, registration often focuses on inter-patient/inter-study mono-modal alignment. Another important and (if not more) frequent focal point is multi-channel imaging, dynamic-contrast computed tomography (CT), multi-parametric magnetic resonance imaging (MRI), or positron emission tomography (PET) combined with CT/MRI. In this setting, the needs of intra-patient multi-modal registration are paramount, given the unavoidable patient movements or displacements between subsequent imaging scans. For scenarios where deformable misalignments are present, the abdomen, correspondences can be highly complex. Because different modalities provide complementary visual/diagnosis information, proper and precise anatomical alignment benefits human reader's radiological observation and is crucial for any downstream computerized analyses. However, finding correspondences between homologous points is usually not trivial because of the complex appearance changes across modalities, which may be conditioned on anatomy, pathology, or other complicated interactions.

Unfortunately, multi-modal registration remains a challenging task, particularly since ground-truth deformations are hard or impossible to obtain. Methods must instead learn transformations or losses that allow for easier correspondences between images. Unsupervised registration methods, like [8, 44], often use a local modality invariant feature to measure similarity. However these low-level features may not be universally applicable and cannot always capture high level semantic information. Other approaches use generative models to reduce the domain shift between modalities, and then apply registration based on direct intensity similarity [128]. A different strategy learns registrations that maximize the overlap in segmentation labels [8, 51]. This latter approach is promising, as it treats the registration process similarly to a segmentation task, aligning images based on their semantic category. Yet, these approaches rely on having supervised segmentation labels in the first place for every deployment scenario.

Both the synthesis and segmentation approaches are promising, but they are each limited when used alone, especially when fully-supervised training data is not available, no paired multi-modal images and segmentation labels, respectively. As Fig. 5.1 elaborates, the synthesis, segmentation, and registration tasks are linked together and define implicit constraints between each other. That motivates us to develop a joint synthesis, segmentation, and registration (JSSR) system which satisfies these implicit constraints. JSSR is composed of a generator, a segmentation, and a registration component that performs all three tasks simultaneously. Given a fixed image and moving image from different modalities for registration, the generator can synthesize the moving image to the same modality of the fixed image, conditioned on the fixed image to better reduce the domain gap. Then the registration component accepts the synthesized image from the generator and the fixed image to estimate a deformation field. Lastly, the segmentation module estimates the segmentation map for the moving image, fixed image and synthesized image. During the training procedure, we optimize several consistency losses including (1) the similarity between the fixed image and the warped synthesized image; (2) the similarity between the segmentation maps of the warped moving image and the fixed image; (3) an adversarial loss for generating high fidelity images; and (4) a smoothness loss to regularize the deformation field. To stop the segmentation module from providing meaningless segmentation maps, we regularize the segmentation by training it on fully supervised data obtained from a different source than the target data, e.g., public data. We evaluate our system on a large-scale clinical liver CT image dataset containing four phases per patient, for unpaired image synthesis, multi-modal image registration, and multi-modal image segmentation tasks. Our system outperforms the state-of-the-art conventional multi-modal registration methods and significantly improves the baseline model we used for the fother two tasks, validating the effectiveness of joint learning.

We summarize our main contributions as follows:

- We propose a novel joint learning approach for multi-modal image registration that incorporates the tasks of synthesis, registration and segmentation. Each task connects to the other two tasks during training, providing mutually reinforcing supervisory signals.
- We evaluate and validate the performance improvement of baseline methods for synthesis and segmentation after joint training by our system, demonstrating the effectiveness of our joint training setup and revealing the possibility of obtaining a better overall system by building upon and enhancing the baseline models.
- Our system consistently and significantly outperforms state-of-the-art conventional multi-modal registration approaches based on a large-scale multi-phase CT imaging dataset of 1,485 patients (each patient under four different intravenous contrast phases, i.e., 5,940 3D CT scans with various liver tumors).
- While we use supervised data from *single-phase* public data to regularize our segmentation, our method does not use or rely upon any manual segmentation labels from the target *multi-phase* target CT imaging dataset. Compared

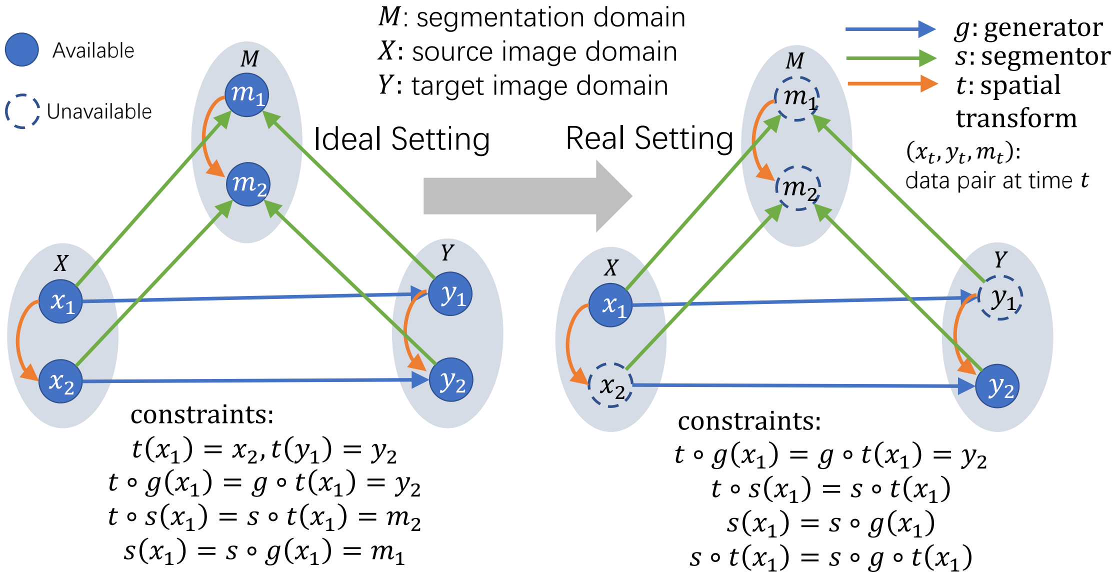

Available
Unavailable
M: segmentation domain
X: source image domain
Y: target image domain
g: generator
s: segmentor
t: spatial transform
(x_t, y_t, m_t): data pair at time t
Ideal Setting
Real Setting
M
X
Y
m_1
m_2
x_1
x_2
y_1
y_2
constraints:
$t(x_1) = x_2, t(y_1) = y_2$
$t \circ g(x_1) = g \circ t(x_1) = y_2$
$t \circ s(x_1) = s \circ t(x_1) = m_2$
$s(x_1) = s \circ g(x_1) = m_1$
constraints:
$t \circ g(x_1) = g \circ t(x_1) = y_2$
$t \circ s(x_1) = s \circ t(x_1)$
$s(x_1) = s \circ g(x_1)$
$s \circ t(x_1) = s \circ g \circ t(x_1)$

Figure 5.1: The relationship between the synthesis, segmentation and registration tasks. In the ideal setting, spatially transformed examples from each domain, and their segmentation labels, are fully available. In more realistic settings, only one example is available from each domain, each under a different spatial transform. Moreover, segmentation labels are not available. Should segmentation, synthesis, and spatial transform *mappings* be available, the constraints in the ideal case can be mapped to analogous constraints in the real case.

to approaches expecting target segmentation labels, JSSR enjoys better scalability and generalizability for varied clinical applications.

5.2 Related Work

5.2.1 Multi-modal Image Registration

Multi-modal image registration has been widely studied and applied in medical imaging. Existing registration methods can be based on additional information, e.g., landmarks [109, 126] or a surface [148], or they can operate directly on voxel intensity values without any additional constraints introduced by the user or segmentation [88]. For voxel-based methods, there are two typical strategies. One is to transform each image using self-similarity measurements that are invariant across modalities. These include local self-similarities [120] or the modality independent neighbourhood descriptor [43]. Notably the DEEDS algorithm [44, 45, 41] employed a discrete dense displacement sampling for deformable registration using self-similarity context (SSC) [47]. The other common strategy is to map both modalities into a shared space and measure the mono-modal difference. Prominent examples include mutual information [86] and normalized mutual information [124] similarity measures that can be applied directly on cross-modal images. However, such methods can suffer from low convergence rates and loss of spatial information. [11] employed a convolutional neural network (CNN) to learn modality invariant features using a small amount of supervision data. [15] used Haar-like features from paired multi-modality images to fit a random forest regression model for bi-directional image synthesis, and [128, 87] applied CycleGANs to reduce the gap between modalities for better alignment. Recently [3] developed a joint synthesis and registration framework on natural 2D images.

Recently a variety of deep learning-based registration methods have been proposed. Because ground truth deformation fields are hard to obtain, unsupervised methods, like [24, 70, 23, 8], are popular. These all rely on a CNN with a spatial transformation function [56]. These unsupervised methods mainly focus on monomodal image registration. Some methods make use of correspondences between labelled anatomical structures to help the registration process [51]. [8] also showed how the segmentation map can help registration. However, in many cases the segmentation map is not available, which motivates us to combine the registration and segmentation components together.

5.2.2 Multi-task Learning Methods

As the registration, synthesis, segmentation tasks are all related with each other, there are already several works that explore combining them together. [128, 87, 136] used CycleGANs to synthesize multi-modal images into one modality, allowing the application of mono-modal registration methods. [61] projected multi-modal images into a shared feature space and registered based on the features. [100] made use of a generative model to disentangle the appearance space from the shape space. [69, 145, 99] combined a segmentation model with a registration model to let them benefit each other, but the focus was on mono-modal registration. [158] performed supervised multi-phase segmentation based on paired multi-phase images but did not jointly train the registration and segmentation. [153, 157, 53] used a generative model to help guide the segmentation model. In contrast, our work combines *all three* of the tasks together to tackle multi-modal registration problem in the most general setting where the deformation ground truth, paired multi-modal images and segmentation maps are *all unavailable*.

5.3 Methodology

Given a moving image $x \in \mathcal{X}$ and fixed image $y \in \mathcal{Y}$ from different modalities, *but from the same patient*, we aim to find a spatial transformation function τ that

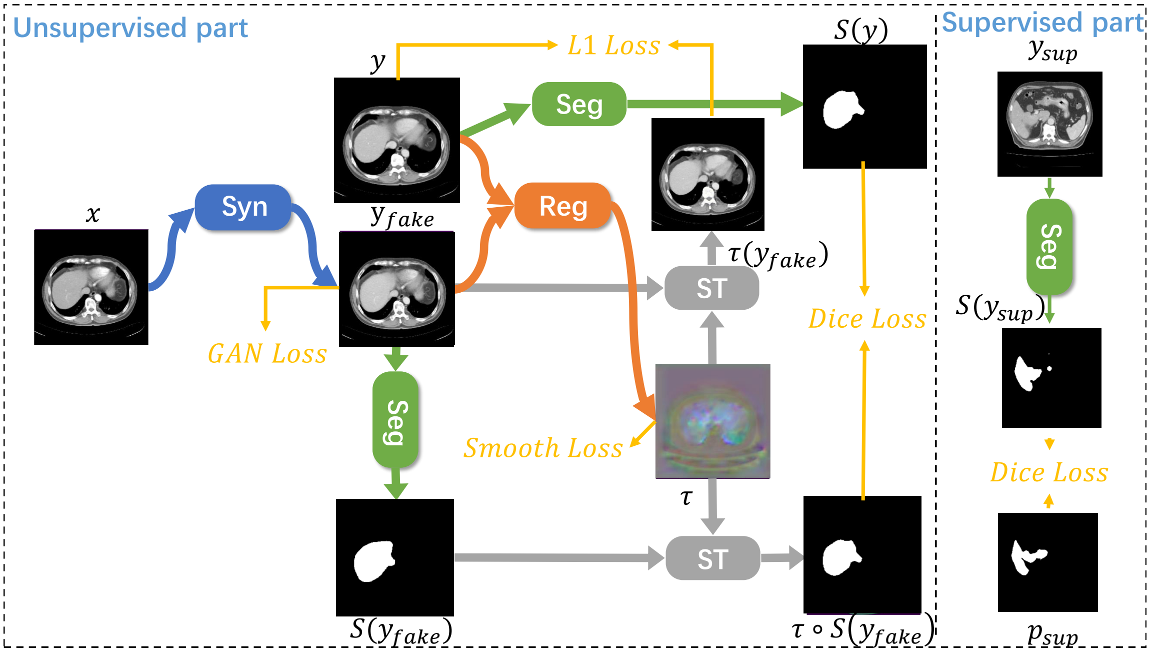


Figure 5.2: The JSSR system. We denote the generator, segmentation, registration module and spatial transform as Syn, Seg, Reg and ST respectively.

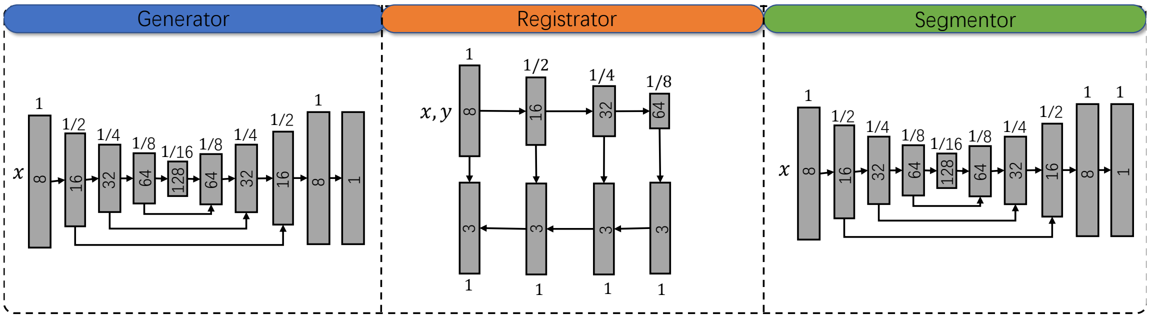


Figure 5.3: The model structure for each component. We use a 3D PHNN [37] for registration and 3D VNet [90] for segmentation and the generator.

corrects for any misalignments between the two. We tackle this multi-modal image registration problem in a fully unsupervised way to meet common applications settings, where none of the ground truth deformation fields, segmentation maps, or paired multi-modal images are available. As Fig. 5.1 depicts, image synthesis, segmentation and registration can be related together via a set of constraints. Motivated by this, we develop a system consisting of three parts: a generator G, a registration module Φ and a segmentation module S. By satisfying the constraints in Fig. 5.1, we can satisfy the conditions for a correct registration, segmentation and image synthesis. During optimization, these three tasks will benefit from each other. Refer to Fig. 7.1 for the overall framework of our system.

5.3.1 Unpaired Image Synthesis

Although good unpaired image synthesis works exist, e.g., [52], they may generate a variety of different target domain images based on the random sampling. However, for registration, the synthesized images should have identical anatomical and other pertinent modality-invariant properties. Thus, a conditional synthesis is a natural choice. Similar to [55], but without random noise, we use a GAN with a dual-input generative model G which learns a mapping from x, y to $\tau^{-1}(y)$, $G : \{x, y\} \to \tau^{-1}(y)$. Here τ is the true deformation from x to y, meaning the generator attempts to generate a version of x that looks like y, but removing any spatial transformation between the two. In reality, τ itself must be estimated, which we will outline in Sec. 5.3.2. A discriminator D is also equipped to detect the fake images from the generator.

The objective of the conditional GAN is

$$\mathcal{L}_{GAN}(G, D) = E_y \log D(y) - E_{x,y} \log D(G(x, y)). \tag{5.1}$$

In a classical paired GAN setup, we would use $E_y \log D(\tau^{-1}(y))$, but this is not available, so use unpaired synthesis, based on the assumption that spatial transform τ does not alter the likelihood of any one sample. We also add another appearance-based loss to benefit the GAN objective:

$$\mathcal{L}_{L1}^{syn}(G) = E_{x,y}||\tau^{-1}(y) - G(x,y)||_1. \tag{5.2}$$

The final objective for the synthesis part is

$$G^* = \arg\min_G \max_D \mathcal{L}_{L1}^{syn}(G) + \lambda_{syn}\mathcal{L}_{GAN}(G,D). \tag{5.3}$$

5.3.2 Multi-Modal Image Registration

For two images x and y, the registration module learns a function $\Phi : x, y \rightarrow \tau$ where τ is a spatial transformation function [56], also called the deformation field. For mono-modal registration, the L_1 loss can be used to estimate a deformation field that directly matches the intensities between the fixed image and warped image. Here we are registering two images from different modalities. [8] proposed to use a cross-modal similarity measure like cross-correlation [5]. Instead, if we assume a generative model is available to transform x into the $\mathcal{Y}$ domain, then we can use a simple mono-modal similarity measure:

$$\mathcal{L}_{L1}^{reg}(\Phi) = E_{x,y}||\tau(G(x,y)) - y||_1, \tag{5.4}$$

where $\tau = \Phi(G(x,y), y)$, and G is the generator that synthesizes images from $\mathcal{X}$ to $\mathcal{Y}$. Another smoothness term is added to prevent non-realistic deformation:

$$\mathcal{L}_{smooth}(\Phi) = E_{x,y} \sum_{v \in \Omega} ||\nabla \tau_v||^2, \tag{5.5}$$

where v represents the voxel location and $\nabla \tau_v$ calculates the differences between neighboring voxels of v. We use the same implementation for the smoothness term as in [8]. The final objective is:

$$\Phi^* = \arg\min_{\Phi} \mathcal{L}_{L1}^{reg}(\Phi) + \lambda_{reg}\mathcal{L}_{smooth}(\Phi). \tag{5.6}$$

Of course, we cannot optimize this objective without a G. However, to get a good G, we need a good Φ as discussed in Sec. 5.3.1, which makes this problem a chicken-and-egg conundrum. One way is to optimize the two objectives from the synthesis and registration modules together, which leads to

$$\begin{aligned}\Phi^*, G^* &= \arg\min_{\Phi,G} \mathbb{F}(\Phi, G) \\ &= \arg\min_{\Phi,G}\max_{D} \mathcal{L}_{L1}^{reg}(\Phi, G) + \mathcal{L}_{L1}^{syn}(\Phi, G) \\ &+ \lambda_{reg}\mathcal{L}_{smooth}(\Phi, G) + \lambda_{syn}\mathcal{L}_{GAN}(G, D) \\ &\approx \arg\min_{\Phi,G}\max_{D}\ 2\mathcal{L}_{L1}^{reg}(\Phi, G) \\ &+ \lambda_{reg}\mathcal{L}_{smooth}(\Phi, G) + \lambda_{syn}\mathcal{L}_{GAN}(G, D).\end{aligned} \tag{5.7}$$

However, there is no guarantee that we can get the optimal solution by minimizing $\mathbb{F}(\Phi, G)$. Actually there is a trivial solution that minimizes $\mathbb{F}(\Phi, G)$, which is when $G(x, y) = y$ and $\Phi(G(x, y), y) = \Phi(y, y) = I$, i.e., the identity transform. To mitigate this, we add skip connections from the source domain to keep the spatial information in the structure of generator, as shown in Fig. 5.3.

5.3.3 Multi-Modal Image Segmentation

We enforce segmentation-based constraints for two reasons. Firstly, as noted in [8], the additional information of segmentation maps can help guide the registration process. However, [8] assumes the segmentation maps are available for the target dataset, which we do not assume. Secondly, as noted by others [69, 145, 99, 153, 157], synthesis and registration can benefit segmentation, which can help develop better segmentation models on datasets without annotation.

We denote the segmentation model as a function $S : x \rightarrow p$, where $p \in \mathcal{P}$ represents the segmentation map domain. Based on the constraint between synthesis, registration and segmentation tasks, we define the objective as:

$$\mathcal{L}_{dice}^{reg}(S, \Phi, G) = E_{x,y} 1 - Dice[\tau(S(G(x, y))), S(y)], \tag{5.8}$$

where $\tau = \Phi(G(x, y), y)$ and $Dice(x, y) = \frac{2x^T y}{x^T x + y^T y}$ is the widely used measurement for the similarity between two binary volumes. This loss term connects three components together and in the experiments afterwards we show this crucial toward the whole system's performance.

To make (5.8) work properly, we need the segmentation to be as accurate as possible. However only with the consistency loss, the segmentation module is not able to learn meaningful semantic information. For instance, a segmentation module that predicts all background can trivially minimize (5.8). To avoid this, we use fully supervised data, e.g., from public sources, to regularize the segmentation. Importantly, because (5.8) is only applied on the $\mathcal{Y}$ domain, we need only use supervised data from one modality, e.g., if we are registering dynamic contrast CT data, we need only fully-supervised segmentation maps from the more ubiquitous venous-phase CTs found in public data. Thus, the supervision loss is defined as

$$\mathcal{L}_{dice}^{sup}(S) = E_{y_{sup}} 1 - Dice[S(y_{sup}), p_{sup})], \tag{5.9}$$

where $y_{sup} \in \mathcal{Y}$ is in the same modality with $y \in \mathcal{Y}$, but the two datasets do not overlap. $p_{sup} \in \mathcal{P}_{sup}$ is the corresponding annotation. The total loss provided by the segmentation module is

$$\mathbb{H}(S, \Phi, G) = \mathcal{L}_{dice}^{reg}(S, \Phi, G) + \mathcal{L}_{dice}^{sup}(S). \tag{5.10}$$

5.3.4 Joint Optimization Strategy

Based on previous sections, the final objective for our whole system is

$$\Phi^*, G^*, S^* = \arg\min_{\Phi, G, S} \mathbb{F}(\Phi, G) + \lambda_{seg} \mathbb{H}(S, \Phi, G). \tag{5.11}$$

In order to provide all the components with a good initial point, we first train S on the fully-supervised data, $\{y_{sup}, p_{sup}\}$ and also train Φ and G using (5.7) on the unsupervised data. Finally, we jointly optimize all modules by (5.11). When optimizing (5.7) and (5.11), we use the classic alternating strategy for training GAN models, which alternately fixes Φ, G, S and optimizes for D and then fixes D and optimizes for the others.

5.4 Experiments

Datasets. We conduct our main experiments on a large-scale dataset of 3D dynamic contrast multi-phase liver CT scans, extracted from the archives of the Chang Gung Memorial Hospital (CGMH) in Taiwan. The dataset is composed of 1485 patient studies and each studies consists of CT volumes of four different intravenous contrast phases: venous, arterial, delay, and non-contrast. The studied population is composed of patients with liver tumors who underwent CT imaging examinations prior to an interventional biopsy, liver resection, or liver transplant. Our end goal is to develop a computer-aided diagnosis system to identify the pathological subtype of any given liver tumor. Whether the analysis is conducted by human readers or computers, all phases need to be precisely pre-registered to facilitate downstream analysis, which will observe the dynamic contrast changes within liver tumor tissues across the sequential order of non-contrast, arterial, venous and delay CTs.

The different phases are obtained from the CT scanner at different time points after the contrast media injection and will display different information according to the distribution of contrast media in the human body. The intensity value of each voxel in the CT image, measured by the Hounsfield Unit (HU), is an integer ranging from -1000HU to 1000HU, which will also be affected by the density of contrast media. The volume size of the CT image is $512 \times 512 \times L$, where L can vary based on how the image was acquired. The *z*-resolution is 5mm in our dataset. Since the venous phase is one of the most informative for diagnosis, and is also ubiquitous in public data, we choose it as the anchor phase and register images from other three phases to it. Consequently, we also synthesize the other three phases images to the venous phase. We divide the dataset into 1350/45/90 patients for training, validation and testing, respectively, and we manaully annotate the liver masks on the validation and testing sets for evaluation. *Note that there are in total* $1485 \times 4 = 5940$ *3D CT scans (all containing pathological livers) used in our work. To the best of our knowledge, this is the largest clinically realistic study of this kind to-date.* For the supervised part, we choose a public dataset, i.e., **MSD** [122], that contains 131 CT images of venous phase with voxel-wise annotations of the liver and divide it into 100/31 for training and validation. We evaluate the performance of all three registration, synthesis and segmentation tasks to measure the impact of joint training.

5.4.1 Baseline

We compare with several strong baselines for all three tasks:

- For image synthesis, we choose **Pix2Pix** [55]. We approximately treat the multi-phase CT scans from the same patient as paired data, so that we can better compare to see how incorporating registration can benefit the synthesis module when there is no paired data.
- For image registration, we first compare with **Deeds** [44], one of the best

	Dice ↑			HD95 ↓		
	Arterial	Delay	Non-Contrast	Arterial	Delay	Non-Contrast
Initial State	90.94 (7.52)	90.52 (8.08)	90.08 (6.74)	7.54 (4.89)	7.86 (5.83)	7.87 (4.37)
Affine [89]	92.01 (6.57)	91.69 (6.80)	91.52 (5.48)	6.81 (4.83)	6.95 (5.32)	6.73 (3.63)
Deeds [44]	94.73 (2.10)	94.70 (1.91)	94.73 (1.90)	4.74 (1.96)	4.76 (1.69)	4.62 (1.05)
VoxelMorph [8]	94.28 (2.53)	94.23 (3.15)	93.93 (2.58)	5.29 (2.33)	5.42 (3.25)	5.40 (2.48)
JSynR-Reg	94.81 (2.35)	94.71 (2.62)	94.57 (2.52)	4.93 (2.14)	5.07 (3.06)	4.87 (2.30)
JSegR-Reg	95.52 (1.76)	95.39 (2.14)	95.37 (1.80)	4.47 (2.21)	4.70 (3.24)	4.45 (1.85)
JSSR-Reg	**95.56**(1.70)	**95.42**(2.00)	**95.41**(1.72)	**4.44**(2.19)	**4.65**(3.14)	**4.35**(1.60)
	ASD ↓			Time ↓		
	Arterial	Delay	Non-Contrast	Arterial	Delay	Non-Contrast
Initial State	2.12 (1.86)	2.27 (2.19)	2.37 (1.77)	-/-	-/-	-/-
Affine [89]	1.74 (1.58)	1.86 (1.89)	1.87 (1.41)	-/7.77	-/7.77	-/7.77
Deeds [44]	1.01 (0.44)	1.01 (0.39)	0.99 (0.36)	-/41.51	-/41.51	-/41.51
VoxelMorph [8]	1.10 (0.53)	1.12 (0.87)	1.20 (0.67)	1.71/1.76	1.71/1.76	1.71/1.76
JSynR-Reg	0.95 (0.45)	0.98 (0.72)	0.98 (0.56)	3.14/1.76	3.14/1.76	3.14/1.76
JSegR-Reg	0.80 (0.37)	0.83 (0.59)	0.83 (0.40)	3.14/1.76	3.14/1.76	3.14/1.76
JSSR-Reg	**0.79**(0.36)	**0.83**(0.56)	**0.82**(0.37)	1.71/1.76	1.71/1.76	1.71/1.76

Table 5.1: Evaluation for the registration task on the CGMH liver dataset in terms of Dice score, HD (mm), ASD (mm), and GPU/CPU running time (s). Standard deviations are in parentheses.

registration methods to date for abdominal CT [144]. The advantage of learning-based methods compared with conventional ones is often on the speed of inference, but we can also show performance improvement. We also compare with the learning-based **VoxelMorph** [8] with local cross-correlation to handle multi-modal image registration.

- For the segmentation task, we compare with **VNet** [90], which is a popular framework in medical image segmentation.

5.4.2 Implementation Details

We conduct several preprocessing procedures. First, since the CT images from different phases, even for the same patient, have different volume sizes, we crop the maximum intersection of all four phases based on the physical coordinates to make their size the same. Second, we apply rigid registration using [89] between the four phases, using the venous phase as the anchor. Third, we window the intensity values to -200HU to 200HU and normalize to -1 to 1, and then we resize the CT volume to $256 \times 256 \times L$ to fit into GPU memory. For the public dataset, we sample along the axial axis to make the resolution also 5mm, and then apply the same intensity preprocessing.

The structure of each component is shown in Fig. 5.3. We choose 3D V-Net [90] for the generator and segmentation module and 3D PHNN [37] for the registration. To optimize the objectives, we use the Adam solver [62] for all the modules, setting the hyper parameters to $\lambda_{seg} = \lambda_{reg} = 1,\ \lambda_{syn} = 0.02$. We choose different learning rates for different modules in order to better balance the training:. 0.0001, 0.001, 0.1, and 0.1 for the generator, registration module, segmentation module, and discriminator, respectively. Another way to balance the training is to adjust the loss term weights. However, there are loss terms that relate with multiple modules, which makes it more complex to control each component separately. We train on the Nvidia Quadro RTX 6000 GPU with 24 GB memory, with instance normalization and batch size 1. The training process takes about 1.4 GPU days.

5.4.3 Main Results

5.4.3.1 Multi-modal image registration

We summarize the results of registration task in Table ??. We use the manual annotations of the test set and evaluate the similarity between those of fixed image,

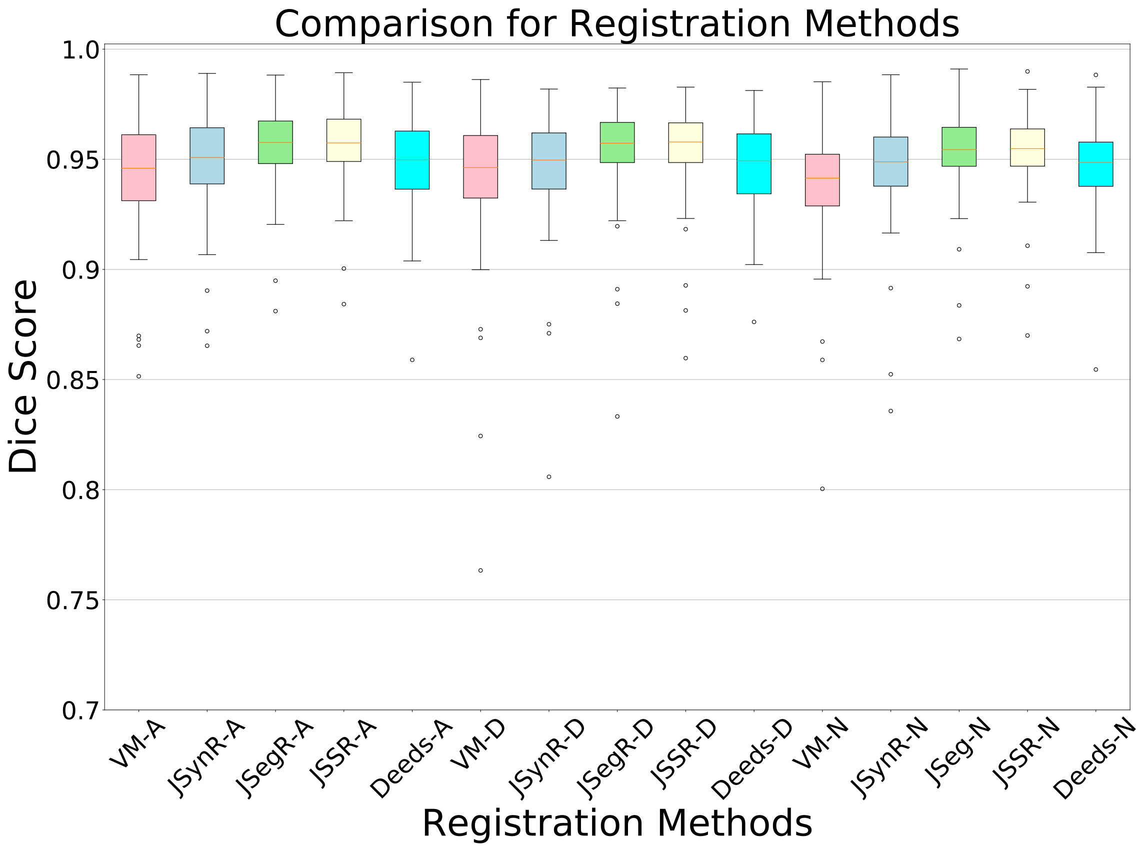


Figure 5.4: Box-plots for the registration results (DSC). Suffixes indicate the moving phases (A, D, N for arterial, delay, non-contrast). VM stands for VoxelMorph.

which is always in the venous phase here, and the warped labels of the moving images chosen from arterial, delay and non-contrast. The similarity is measured using the Dice score, 95 percent hausdorff distance (HD), and the average surface distance (ASD). We also report the consumed time on GPU/CPU in sec for each method. We use the term "Initial State" to refer to the result before applying any registration and "Affine" to the result after rigid registration. We denote our joint system as JSSR and JSSR-Reg is only the registration part of JSSR. We also compare

	VNet [90]			
Dice ↑	Venous	Arterial	Delay	Non-Contrast
No-Synthesis	90.47 (6.23)	89.47 (7.05)	89.88 (6.38)	89.38 (6.38)
Pix2Pix [55]	90.47 (6.23)	76.50 (17.77)	79.60 (13.13)	67.48 (15.97)
JSynR-Syn	90.47 (6.23)	89.69 (7.09)	90.01 (6.27)	90.15 (6.21)
JSSR-Syn	90.47 (6.23)	89.44 (7.15)	89.76 (6.34)	89.31 (7.57)

	JSegR-Seg			
Dice ↑	Venous	Arterial	Delay	Non-Contrast
No-Synthesis	91.88 (4.84)	90.91 (5.06)	91.18 (4.68)	91.12 (4.72)
Pix2Pix [55]	91.88 (4.84)	89.59 (5.51)	87.78 (5.78)	89.59 (5.51)
JSynR-Syn	91.88 (4.84)	91.15 (4.93)	91.37 (4.56)	91.36 (4.54)
JSSR-Syn	91.88 (4.84)	91.12 (4.99)	91.30 (4.63)	91.39 (4.53)

	JSSR-Seg			
Dice ↑	Venous	Arterial	Delay	Non-Contrast
No-Synthesis	92.24 (3.88)	91.25 (4.10)	91.34 (3.76)	91.37 (3.81)
Pix2Pix [55]	92.24 (3.88)	85.30 (7.11)	84.68 (9.29)	79.89 (8.49)
JSynR-Syn	92.24 (3.88)	91.42 (4.06)	91.58 (3.64)	91.67 (3.67)
JSSR-Syn	92.24 (3.88)	91.39 (4.10)	91.51 (3.72)	91.60 (3.69)

Table 5.2: Evaluation for the synthesis and segmentation tasks on the CGMH liver dataset in terms of average Dice score

two ablations of JSSR. JSynR, which only contains the generator and registration module, is optimized using (5.7). JSegR has the segmentation and registration module instead. More details will be discussed in Section 5. As can be seen, our JSSR method outperforms Deeds by 0.83% by average Dice, while executing much faster in terms of inference. Also by taking advantage of the joint training, JSSR achieves significantly higher results than VoxelMorph (exceeded by 1.28%) with comparable inference time. We can observe gradual improvements from VoxelMorph to JSynR to JSSR, which demonstrates the successive contributions of joint training. Fig. 5.4 depicts a box plot of these results.

5.4.3.2 Multi-modal image segmentation and synthesis

Table ?? presents the synthesis and segmentation evaluations. Following the practice of [55], we evaluate the synthesis model by applying the segmentation model on the synthesized image. The intuition is that the better the synthesized image is, the better the segmentation map can be estimated. We evaluate with three segmentation models. The VNet baseline is trained on the MSD dataset with full supervision. JSegR-Seg is the segmentation part of JSegR as described in Section 5. JSSR-Seg is the segmentation module of our JSSR system. For each segmentation model, we test it on different synthesis model, thus comparing all possible synthesis/segmentation combinations. For "No-Synthesis", we directly apply the segmentation model on original images. For the three synthesis models, we test the segmentation model on the original venous image and also on the "fake" venous images synthesized from arterial, delay, non-contrast phases. From the No-Synthesis lines we can observe a clear performance drop when directly applying the segmentation model to arterial, delay and non-contrast phases, since the supervised data is all from the venous phase. For Pix2Pix, the performance goes through different levels of reduction among different segmentation algorithms and is not as high as the Non-Synthesis. That may be caused by artifacts introduced by the GAN model and the L1 term is providing less constraint since there is no paired data. Comparing the JSynR-Syn and JSSR-Syn generators, the performance is improved by creating true paired data via the registration process, but even so, it is just comparable to No-Synthesis. For JSynR-Syn, the JSynR is not jointly learned with a segmentation process, so the performance for synthesized images does not necessarily go up. For JSSR-Syn, however, it means the constraints we are using for optimizing the system does not bring enough communication between the generator and segmentor to improve the former. Even so, we can improvements from VNet to JSegR-Seg to JSSR-Seg on both the No-Synthesis and various synthesis options, indicating that the segmentation process can still benefit from a joint system, which includes the synthesis module. Please refer to Fig. 5.7 for qualitative examples of JSSR registration, synthesis and segmentation results.

5.5 Ablation and Discussion

5.5.1 JSegR vs JSSR

We implement JSegR as another ablation. The purpose is to explore the importance of the synthesis module for the JSSR system. Since JSegR does not have a generator, the registration module takes images from different phases directly as input. The segmentation consistency term in (5.8) is then replaced with

$$\mathcal{L}^{reg}_{dice}(S, \Phi) = E_{x,y} 1 - Dice[\tau(S(x)), S(y)], \tag{5.12}$$

where $\tau = \Phi(x, y)$. This framework is similar to [145], which jointly learned the registration and the segmentation module . In our case, though, x, y are in a different domain and the annotations are unavailable. This method is expected to struggle, since x, y are in different phases. However, as shown in Table **??**, the performance drop across phases is not too severe even for the baseline VNet. Correspondingly, JSegR can achieve a higher result on registration than JSynR and

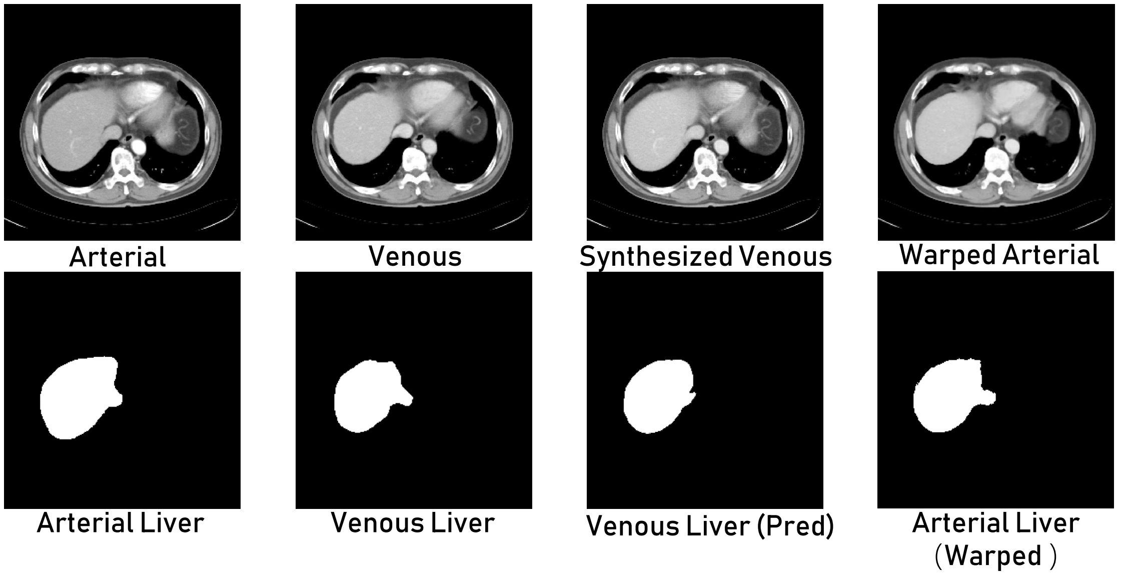


Figure 5.5: Results on the arterial CT phase.

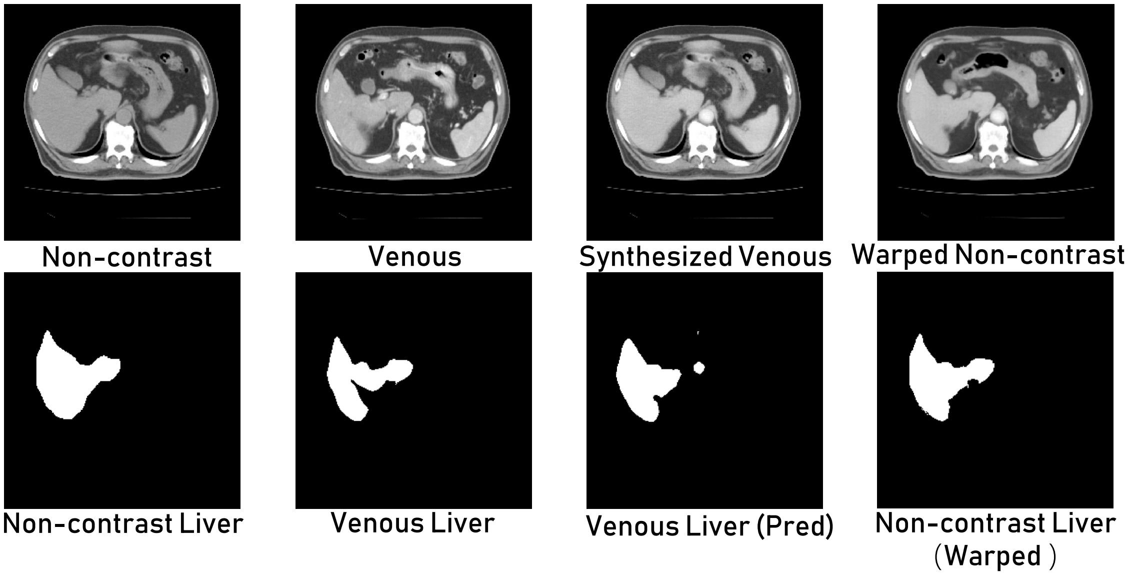


Figure 5.6: Results on the non-contrast CT phase.

Figure 5.7: Qualitative examples of JSSR synthesis, segmentation and registration.

performs close to JSSR, which demonstrates the great importance of incorporating semantic information into the registration.

5.5.2 Extra constraints

The constraints detailed in Fig. 5.1 are not the only possible constraints. For instance, constraints can be added to ensure consistency between "register first" vs "register last" pipelines:

$$\mathcal{L}_{L1}^{reg}(\Phi, G) = E_{x,y}||G(\tau(x), y) - \tau(G(x, y))||_1. \tag{5.13}$$

However, each constraint introduces additional complexity. Future work should explore whether (5.13), or other constraints, can boost performance further.

5.6 Conclusion

In this paper, we propose a novel JSSR system for multi-modal image registration. Our system takes advantages of joint learning based on the intrinsic connections between the synthesis, segmentation and registration tasks. The optimization can be conducted end-to-end with several unsupervised consistency loss and each component benefits from the joint training process. We evaluate the JSSR system on a large-scale multi-phase clinically realistic CT image dataset without any segmentation annotations. After joint training, the performance of registration and segmentation increases by 0.91% and 1.86% respectively on the average Dice score for all the phases. Our system outperforms the recent VoxelMorph algorithm [8] by 1.28%, and the state-of-the-art conventional multi-modal registration method [44] by 0.83%, but has considerably faster inference time.

5.7 Appendix

5.7.1 More Visualization

We visualize more examples in detail here. The denotations follow Figure 7.1. We can see some potential improvement for JSSR system.

Firstly, the generator part is conditioned on both x and y, which brings both benefits and shortcomings. In the column $y \rightarrow \tau(x_{fake})$ we can see the synthesized image can well capture the intensity change between y and x since the checkerboard image only shows mild difference between y and $\tau(x_{fake})$. However, in the $x \rightarrow x_{fake}$ column, like Figure 5.8 Arterial row and Figure 5.12 Arterial row, the generator also introduces additional boundary information from y, which will affect the register.

Secondly, as in Figure 5.11, the segmentor part produces bad segmentation but the overlap in $\tau \circ S(x_{fake}) \rightarrow S(y)$ is still large, meaning that the consistency is well satisfied but will provide wrong supervision to the register. A better consistency term may help this condition.

This system is now only tested on multi-phase CT images. However, equipped with a generator and a segmentor, the system can be applied to many application scenes like the registration from CT to MRI, or the domain adaptation for segmentation between CT and MRI, or it can help the tumor detection by combining multi-modality information if we extend the segmentor to segment both normal

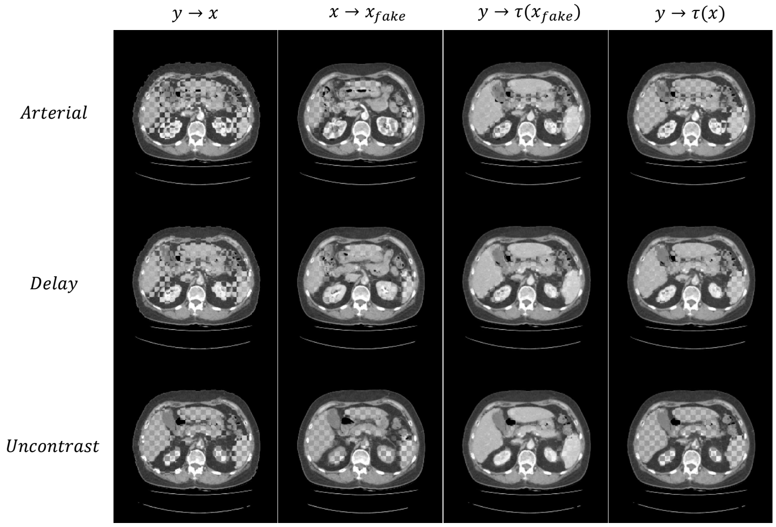


Figure 5.8: An Example for for evaluation of JSSR system. Each picture shows the difference in checkerboard style between the two inputs indicated on the top (on the left and right of $\rightarrow$). y $\in$Venous and x $\in$Arterial, Delay, Uncontrast for each row.

organ and tumor region.

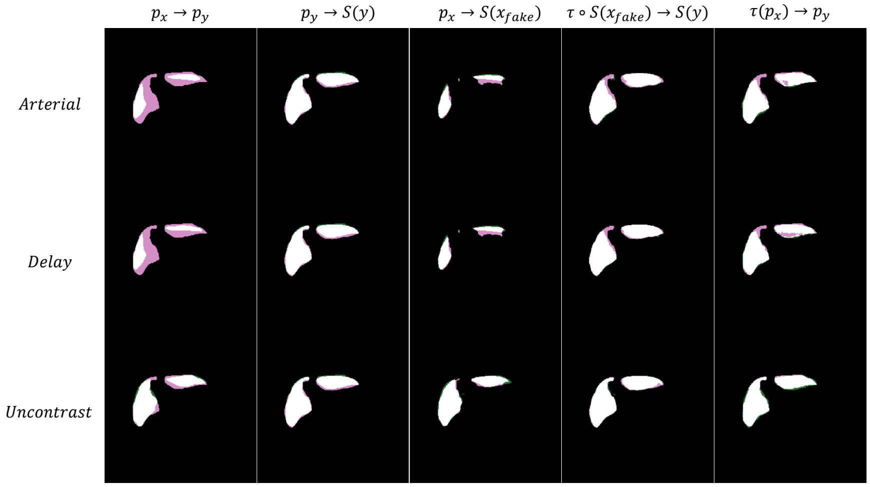


Figure 5.9: Visualization for the segmentation part of JSSR on the same example of Figure 5.9. In each picture, the pink part belongs to the input on the right of → and green part belongs to the left input and white part is the overlap.

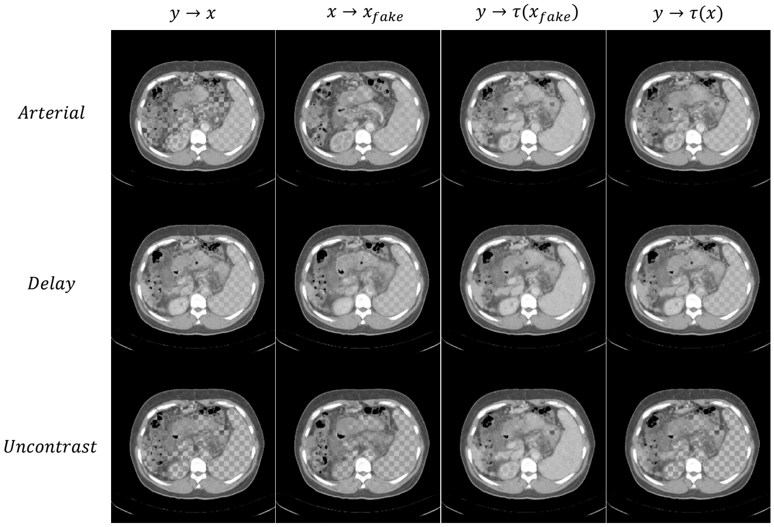


Figure 5.10: Another example

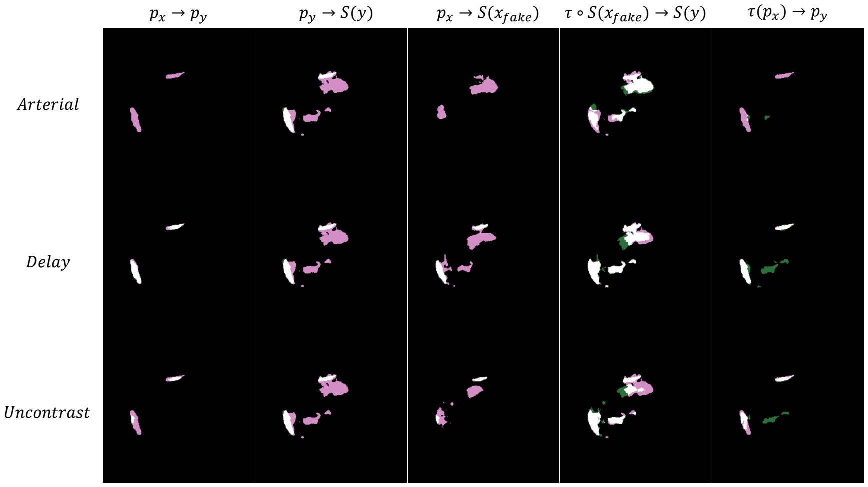


Figure 5.11: Same example as Figure 5.10

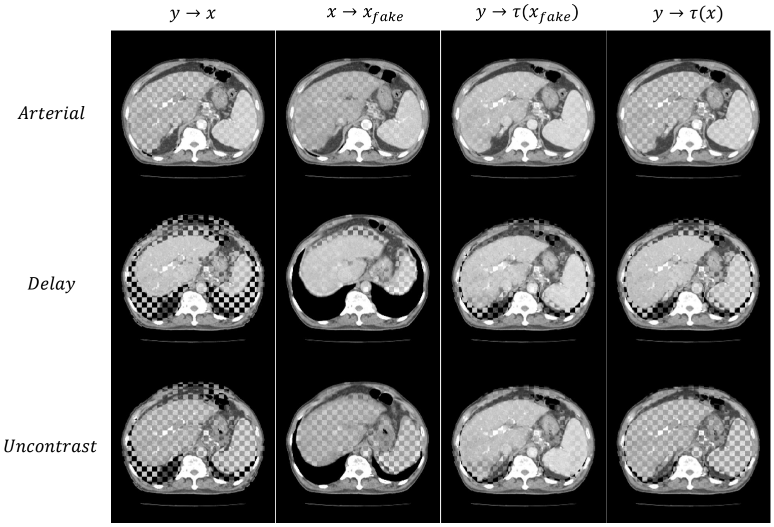


Figure 5.12: Another example

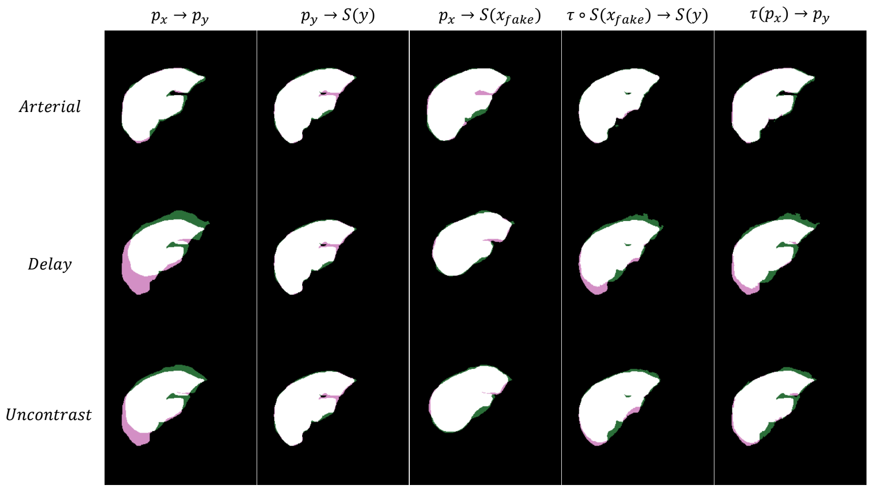


Figure 5.13: Same example as Figure 5.12

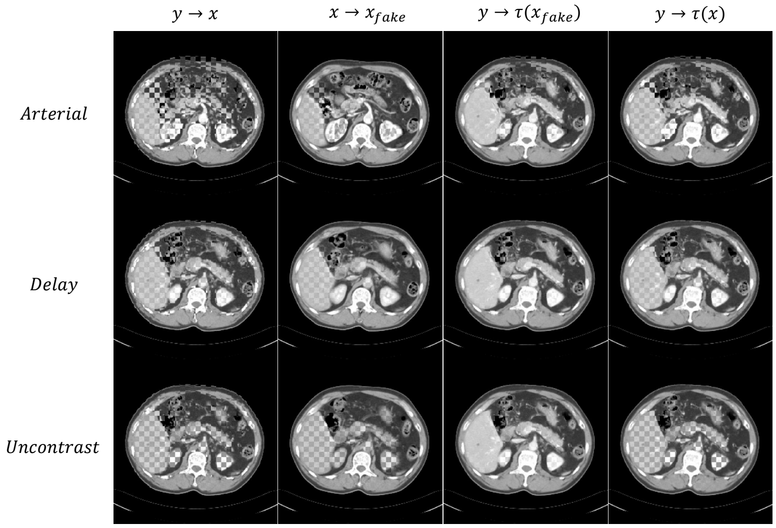


Figure 5.14: Another example

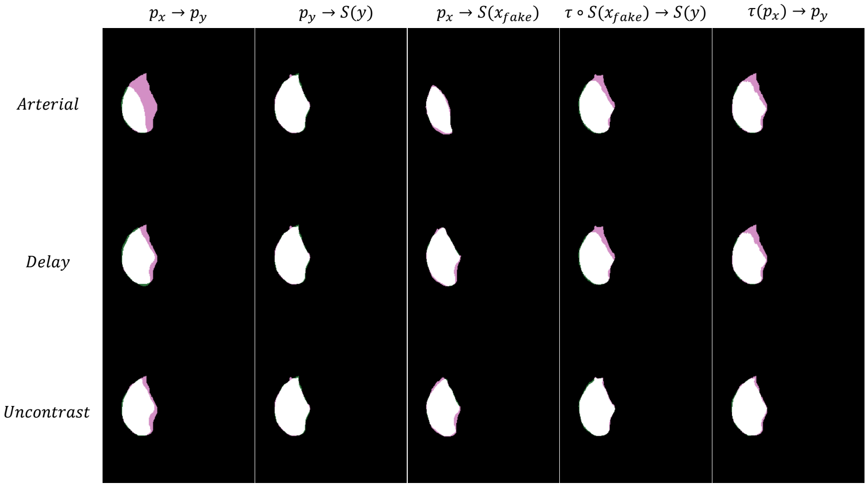

$p_x \rightarrow p_y$
$p_y \rightarrow S(y)$
$p_x \rightarrow S(x_{fake})$
$\tau \circ S(x_{fake}) \rightarrow S(y)$
$\tau(p_x) \rightarrow p_y$
Arterial
Delay
Uncontrast

Figure 5.15: Same example as Figure 5.14

5.7.2 Proof

We made an approximation in the paper that $\mathcal{L}_{L1}^{syn}(\Phi, G) \leq k\mathcal{L}_{L1}^{reg}(\Phi, G)$ for some constant k when the τ generated by Φ is smooth enough. Here we give the prove.

$$\begin{aligned}
||\tau^{-1}(y) - G(x,y)||_1 &= \int_\Omega |\tau^{-1}(y)_i - G(x,y)_i| di \\
&= \int_\Omega |y_{\tau(i)} - G(x,y)_i| di \\
&= \int_\Omega |y_j - G(x,y)_{\tau^{-1}(j)}| d\tau^{-1}(j) \\
&= \int_\Omega |(y)_j - \tau(G(x,y))_j| \tau^{-1\prime}(j) dj \\
&\leq k \int_\Omega |(y)_j - \tau(G(x,y))_j| dj \\
&= k||\tau(G(x,y)) - y||_1
\end{aligned}$$

using the smoothness assumption that $|\tau^{-1\prime}(j)| \leq k \; \forall j, x, y$ and the identity transform $\tau(y)_i = y_{\tau^{-1}(i)}$. Then we have

$$\mathcal{L}_{L1}^{syn}(\Phi, G) = E_{x,y}||\tau^{-1}(y) - G(x,y)||_1 \leq kE_{x,y}||\tau(G(x,y)) - y||_1 = k\mathcal{L}_{L1}^{reg}(\Phi, G).$$

Chapter 6

SAME: Deformable Image Registration based on Embeddings

6.1 Introduction

Deformable image registration is a fundamental task in medical image analysis [115]. Traditional registration methods solve an optimization problem and iteratively minimize a preset similarity measure to align a pair of images. Recently, learning-based deformable registration, using deep networks, have been investigated [9, 50, 154, 93, 74]. Compared with their conventional counterparts, learning-based methods can incorporate more flexible losses, integrate other computing modules and are much faster in inference. VoxelMorph was a representative work [9] that learns a parameterized registration function using a convolutional neural network (CNN). Many recent methods focus on designing more sophisticated networks using pyramid [93] or cascaded structures [50, 154], or connecting registration to pipelines that include synthesis and segmentation [74]. Ideally, registration should focus on aligning semantically similar/coherent voxels, e.g., the same anatomical locations. This semantic information can come in the form of extra manual annotations (e.g. organ masks) [9], but requiring prohibitive labor costs from professionals. Existing unsupervised methods instead optimize similarity measures describing local intensities as a proxy of the semantic information, such as the mean squared error (MSE) or normalized cross correlation (NCC). However, these are less reliable in settings with large deformations, complex anatomical differences, or cross-modality/cross-phase imagery.

In this paper, we exploit incorporating a novel form of semantic information in registration. SAM is a recent work as a means to produce pixel-wise embeddings in radiological images by encoding anatomical semantic information [146]. It requires no annotations in training. SAM can match corresponding points between two images, which is exactly the fundamental goal of image registration. The most simple and straightforward way to register two images with SAM is to extract SAM embeddings from both fixed and moving images, match each moving pixel to the closest fixed pixel in SAM space, and calculate the corresponding coordinate offsets to generate a deformation field. However, this approach is highly inefficient, as there are millions of pixels in a typical 3D CT scan. Besides, SAM would not incorporate spatial smoothness constraints [9], which is useful when the correspondences predicted by SAM contain noises.

We propose SAME to address these issues. SAME is comprised of three consecutive steps. (1) **SAM-affine**, which uses correspondence points generated from SAM on a sparse grid to compute the affine transformation matrix. Affine registration [64] has been widely used either alone or as an initialization of deformable methods [9, 46]. (2) **SAM-coarse**, which uses a coarse correspondence grid to directly produce a coarse-level deformation field. These first two steps are efficient, require no additional training, and can provide a good initialization for the final step. (3) Lastly, **SAM-VM** enhances the deep learning-based VoxelMorph registration method [9], using SAM-based correlation features [26] and a newly formulated SAM similarity loss. SAME is evaluated on a multi-phase chest CT dataset for inter-subject registration with 35 thoracic organs annotated. Quantitative experimental results show that SAM-affine significantly outperforms traditional optimization-based affine registration in both accuracy and speed. The complete SAME consistently outperforms traditional approaches [116, 6] and VoxelMorph [9] in both within-contrast-phase and across-contrast-phase tasks by average Dice scores of 4.7% and 2.7%, respectively. SAME matches DEEDS [46], as the state-of-the-art in CT registration [144], while being orders of magnitude faster (1.2 sec vs. 45 sec).

6.2 Method

In this section, we present the details of the proposed SAME for deformable registration and describe how SAM is integrated in each of the three steps.

6.2.1 SAM

SAM is recently proposed by [146], as a novel pixel-level contrastive learning framework with a coarse-to-fine network and a hard-and-diverse negative sampling strategy. In an unsupervised manner, it predicts a global and a local embedding vector with semantic meanings per pixel in a CT volume—the same anatomical location in different images expressing similar embeddings. SAM is readily used to find correspondences between images, providing a means to solve the registration problem from a new perspective. Let $X_f, X_m \in \mathbb{R}^{D\times H\times W}$ be the fixed and moving images to be registered. For each image, we extract the global and local SAM embedding volumes and concatenate them in the channel dimension, resulting in $S_f, S_m \in \mathbb{R}^{C\times D\times H\times W}$ (C is the concatenated channel dimension). Given a point $p_f = (x, y, z)$ in X_f, we take its embedding vector $S_f(:, z, y, x)$ and convolve it with S_m to get a similarity heatmap volume. The point with the highest similarity score becomes the matched point in the moving image. Results show that matching for a single point only consumes 0.2 sec on a common chest CT scan [146].

6.2.2 SAM-affine and SAM-coarse

Matched SAM correspondences can be directly employed to estimate an affine transformation matrix [64, 46, 9]. First, we select a set of points on X_f for matching. Intuitively, evenly distributed points on the image may lead to a better estimation. Therefore, we use the points on a regular grid on X_f, see Fig. 6.1. It would be more precise to run point matching on every pixel (instead of a coarse grid) and directly generate a fine deformation field, but that would consume 0.5h for a CT with 200 slices. To balance accuracy and speed, we use a grid with stride 8. Since SAM is only designed for points inside the body, we segment the body mask of X_f using intensity thresholding and morphological post processing, and then remove grid points outside the mask. When doing point matching, we downsample S_m

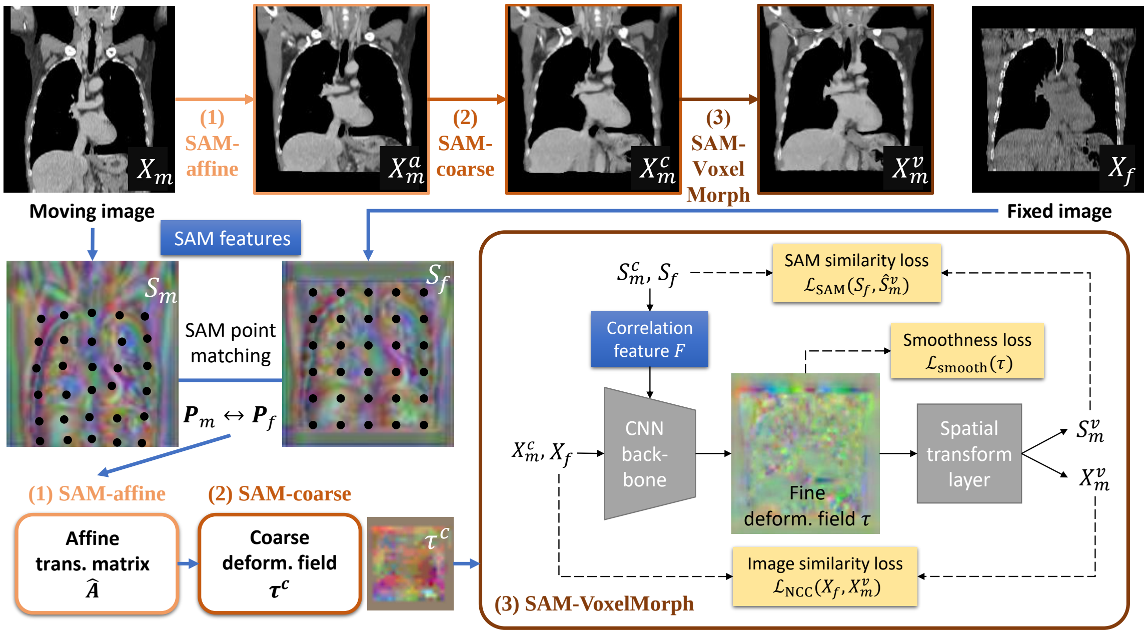


Figure 6.1: SAME framework. The moving image is warped by three consecutive steps: SAM-affine, SAM-coarse, SAM-VM, gradually approaching the fixed image. Variables X, S, and P denote the image, SAM embedding, and point coordinates, respectively. Subscripts m, f stand for moving or fixed, respectively. Superscripts a, c and v indicate the variable is generated after each of the three steps (affine, coarse deform, or VoxelMorph).

with spatial stride of 4 to reduce computation. After the corresponding points in X_m are located, we need to filter out low-quality matches. We examine their similarity scores and discard those lower than a threshold θ. After that, we can get k matched points in X_f, X_m, which can be represented by $3 \times k$ matrices: $\mathbf{P}_f$ and $\mathbf{P}_m$, respectively. We pad them with 1s to create homogeneous versions of the matched points coordinates, $\tilde{\mathbf{P}}_f, \tilde{\mathbf{P}}_m \in \mathbb{R}^{4\times k}$, and estimate the affine matrix $\hat{\mathbf{A}} \in \mathbb{R}^{4\times 4}$ by a simple least squares fitting:

$$\hat{\mathbf{A}} =_{\mathbf{A}} \|\mathbf{A}\tilde{\mathbf{P}}_m - \tilde{\mathbf{P}}_f\|_F^2. \tag{6.1}$$

Next, we transform X_m with $\hat{\mathbf{A}}$ to obtain X_m^a and extract new SAM embeddings S_m^a from it. Then, points in $\mathbf{P}_f$ are matched again on X_m^a to get $\mathbf{P}_m^a$. $\mathbf{P}_m^a$ and $\mathbf{P}_f$ actually represent a mapping from X_m^a to X_f on k sparse points. We can compute their difference $\Delta = \mathbf{P}_f - \mathbf{P}_m^a$, and map each point in Δ back to the original coordinates of the image to get $\tau^c \in \mathbb{R}^{3\times D\times H\times W}$. Note, there are only k deformation in Δ that are not necessarily uniformly spaced. Thus values in τ^c are filled in using linear interpolation. This gives us the final coarsely estimated deformation map, which is applied to warp (X_m^a, S_m^a) to (X_m^c, S_m^c). Although coarsely estimated (on only k points), τ_c can effectively reduce the difference between the moving and the fixed images. Compared to a global affine alignment, this provides local warps that can serve as a better initialization for a final learning-based deformable registration step. One question is that whether we could omit SAM-affine and compute τ^c directly. We observed that before affine registration, the two images may have significant offsets, so τ^c is potentially large in magnitude, which will magnify the noises in the matched points. Thus, we first perform affine registration to reduce the magnitude of deformations.

6.2.3 SAM-VM

The objective of the final step is to predict a fine deformation map $\tau \in \mathbb{R}^{3\times D\times H\times W}$, which is a spatial transformation function that can warp the moving image to best match the fixed one. Following the framework of VoxelMorph [9], we learn a function $\Phi : (X_f, X_m^c) \rightarrow \tau$ with a CNN. The original VoxelMorph uses pure pixel intensity-based features and similarity losses. We improve them by leveraging the semantic information contained in SAM embeddings using SAM correlation features and a SAM loss (see Fig. 6.1).

The loss function in VoxelMorph and follow-up works includes two parts, an image similarity loss and a smoothness loss. We use the local normalized cross-correlation (NCC) loss [9] for the former, while the latter is defined as

$$\mathcal{L}_{smooth}(\tau) = \frac{1}{|\Omega|} \sum_{\mathbf{u} \in \Omega} ||\nabla \tau_{\mathbf{u}}||^2, \tag{6.2}$$

where Ω is the set of all pixels within the body mask. However, the NCC loss only compares local image intensities, which may not be robust under CT contrast injection, pathological changes, and large or complex deformations in the two images. On the other hand, the SAM embeddings can uncover semantic similarities between two pixels. Thus, we add a proposed SAM loss:

$$\mathcal{L}_{SAM}(S_f, S_m^v) = \frac{1}{|\Omega|} \sum_{\mathbf{u} \in \Omega} \langle S_f(\mathbf{u}), S_m^v(\mathbf{u}) \rangle, \tag{6.3}$$

where the superscript v indicates the feature map has been warped by τ predicted by SAM-VM. The final loss is

$$\mathcal{L} = \mathcal{L}_{NCC}(X_f, X_m^v) + \lambda \mathcal{L}_{SAM}(S_f, S_m^v) + \gamma \mathcal{L}_{smooth}(\tau). \tag{6.4}$$

While the SAM loss is an effective means to more semantically align images, the *features* extracted in standard VoxelMorph still lack semantic information, which may be needed to better guide predictions. The correlation feature was originally proposed in FlowNet [26] to manage this problem for optical flow. It was also used in [42] for registration. Briefly, it computes the similarity of pixel $\mathbf{u}$ on X_f and pixel

Table 6.1: Comparison of different registration methods. We show the average Dice score (%) of two tasks: CE-to-CE and CE-to-NC registration. VM: VoxelMorph. Best and second best performance is shown in bold and gray box, respectively.

Methods	CE-to-CE	CE-to-NC	Inference time (s)	std of $\lvert J_\phi \rvert$
Elastix-affine [64]	28.44	27.96	3.38	-
MIND-affine [43]	28.24	27.91	7.86	-
SAM-affine (SA)	33.80	33.77	0.48	-
SAM-coarse (SC)	44.67	43.68	0.78	-
SA + SC	46.76	45.67	1.05	0.40
SA + VM [9]	48.79	47.35	0.78	0.38
SA + SAM-VM	51.99	49.90	0.84	0.36
SA + SC + VM	54.12	50.64	1.13	0.68
SA + SC + SAM-VM (ours)	**54.42**	50.96	1.16	0.66
SyN [6]	49.75	47.95	74.34	-
FFD [116]	49.36	48.22	93.51	0.51
DEEDS [46]	52.72	**51.15**	45.35	0.40

*Paired t-tests show SAME significantly outperforms all other methods ($p < 10^{-4}$), except for DEEDS in the CE-to-NC setting. SAM-VM significantly outperforms VM ($p < 10^{-7}$).

**The average surface distance (ASD) in CE-to-CE: FFD 4.6mm, SA+VM 4.1mm, DEEDS 4.0mm, SA+SAM-VM 3.9mm, SA + SC + SAM-VM 3.8mm.

$\mathbf{u} + \mathbf{d}$ on X_m, where $\mathbf{d}$ is a small displacement. This similarity is computed for each pixel and for n possible displacement values to generate an n-channel feature map, which is then concatenated to the original feature map at some point in the network. When using SAM, the semantic similarity of two pixels can be simply computed as the inner product of two SAM vectors, $F(\mathbf{u}) = \langle S_f(\mathbf{u}), S_m^c(\mathbf{u} + \mathbf{d}) \rangle$. We empirically find that using 27 displacement values $\mathbf{d} \in \{-2, 0, 2\}^3$ yields good results. Injecting the SAM correlation features provides improved cues to the network when predicting deformations, thus brings further boosts in accuracy.

6.3 Experiments

6.3.1 Dataset and task.

To evaluate SAME, we collected a chest CT dataset containing 94 subjects, each with a contrast-enhanced (CE) and a non-contrast (NC) scan. We randomly split the patients to 74, 10, and 10 for training, validation, and testing. Each image has manually labeled masks of 35 organs (including lung, heart, airway, esophagus, aorta, bones, muscles, arteries and veins) [36]. For the validation and test sets, we construct 90 image pairs for inter-subject registration and calculate an atlas-based segmentation accuracy on the 35 organs. Performances of two tasks are evaluated: intra-phase registration (CE-to-CE) and cross-phase registration (CE-to-NC). Every image is resampled to an isotropic resolution of 2mm and cropped to $208 \times 144 \times 192$ by clipping black borders. The image intensity is normalized to $(-1, 1)$ using a window of $(-800, 400)$ HU.

6.3.2 Implementation details.

Our method was developed using PyTorch 1.5. It was run on a Ubuntu server with 12 CPU cores of 3.60GHz. It requires one NVIDIA Quadro RTX 6000 GPU to train and test. We trained a SAM model using the training set of the chest CT dataset. Its structure is identical with the one in [146], which outputs a 128D global embedding and a 128D local one for each pixel. This model is fixed and applied in all three steps of SAME. In SAM-affine and SAM-coarse, the similarity threshold θ is set to 0.7 to select high-confidence matches. In SAM-VM, we use a 3D progressive holistically-nested network (P-HNN) [38] as the backbone and concatenate the correlation feature before the third convolutional block. We also tried 3D U-Net [22] but observed no significant accuracy gains. The loss weights in Eq. 6.4 are empirically set to $\lambda = 1, \gamma = 0.5$. We train SAM-VM using the Adam optimizer with a learning rate of 0.001 for 10 epochs. Each training batch contains 2 image pairs with random contrast phases (CE or NC). We evaluate the registration results using average Dice score over 35 organ masks. The organ masks are not used during training.

6.3.3 Quantitative results.

From Table 6.1 we can see that **SAM-affine** outperforms the traditional affine registration method in Elastix [64] by 5-6%, meanwhile being 6 times faster. It is also better than affine registration with the MIND [43] robust descriptor. This is because SAM can match corresponding anatomical locations between two images accurately and efficiently. Compared with other methods that iteratively optimizes the affine parameters, SAM-affine directly calculates affine matrix by least squared fitting. **SAM-coarse** surpasses SAM-affine by 10% since it allows for locally deformable warping with more degrees of freedom. Cascading these two steps further boosts the accuracy. VoxelMorph pre-aligned by SAM-affine outperforms SAM-affine + SAM-coarse moderately since the latter can only perform a coarse deformable transformation. However, note that the former is a learning-based dense registration method, while the latter does not require any extra training. It only utilizes the matching result of a pretrained SAM model on grid points. The

Table 6.2: Ablation study for different settings on incorporating SAM to VoxelMorph (VM). The average Dice score (%) is reported. All methods are initialized by SAM-affine without SAM-coarse.

Methods	SAM loss	SAM correlation feature	CE-to-CE	CE-to-NC
VM [9]	×	×	48.79	47.35
SAM-VM	✓	×	50.43	48.24
	×	✓	51.37	48.99
	✓	✓	**51.99**	**49.90**

2% small gap demonstrates the capability of our proposed SAM-coarse.

SAM-affine + SAM-coarse can provide a good initialization to the learning-based VM in the third step, allowing it to better perform. From the 4 rows in the middle block of Table 6.1, we also observe consistent improvement by replacing the original VoxelMorph [9] with **SAM-VM**. The SAM embeddings contain more semantic information than the raw pixel intensities, which is incorporated to SAM-VM by the SAM-based correlation feature and SAM loss. An ablation study of SAM-VM is shown in Table 7.3, where the best result is achieved when both the correlation feature and SAM loss are used. On one hand, explicitly inputting the correlation feature calculated by SAM provides extra guidance for determining the deformation fields. On the other hand, the SAM loss provides a more semantically informed supervisory signal.

In the bottom block of Table 6.1, we evaluate several widely-used non-rigid registration methods including FFD [116], SyN [6], and DEEDS [46]. FFD was implemented using Elastix [10], where parameters matched the best performing FFD method in EMPIRE10 Challenge [94]. The only modification was an extra bending energy term with weight 0.01 to regularize the smoothness. For SyN

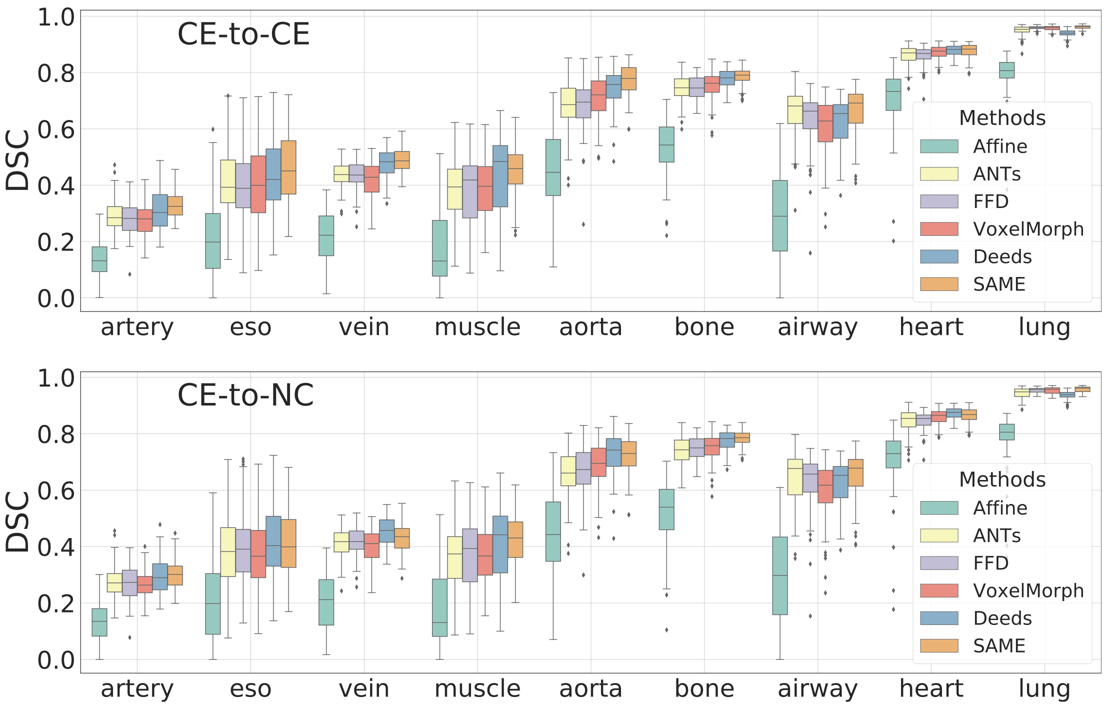


Figure 6.2: Comparison of registration methods on all organ groups. Eso: esophagus.

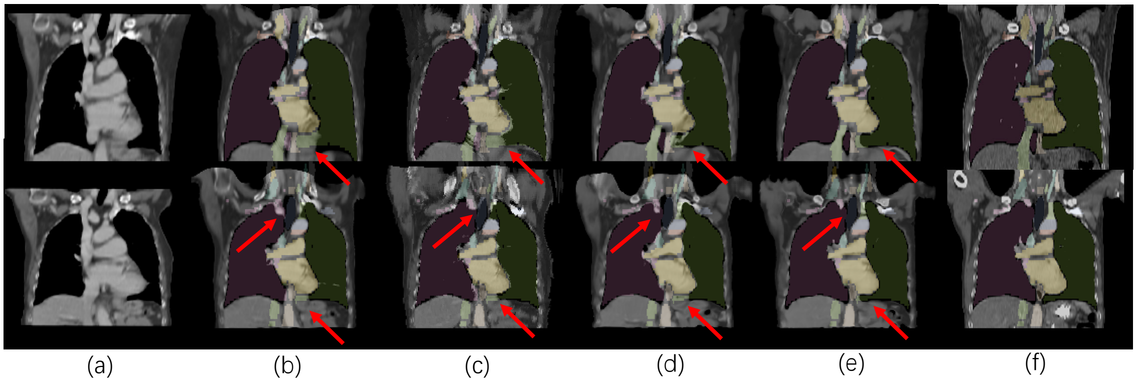


Figure 6.3: Visualization of registration results from different methods. From left to right is (a) the moving image, (b) warped moving image of ANTs, (c) DEEDS, (d) SAM-affine + VoxelMorph, (e) SAME, and (f) the fixed image.

(implemented in ANTS) and DEEDS (implemented by the original author), parameters were set according to those used in [144]. For affine transform, the default implementation in each package was used. The proposed SAME (combination of three steps) achieves markedly better results than SyN and FFD. Compared with the best traditional method (DEEDS), it performs better in the within-phase setting and comparably in the cross-phase setting, meanwhile is 38 times faster. Cross-phase registration is more difficult because the brightness and appearance of contrast-enhanced and non-contrast CTs can be very different (see X_m and X_f in Fig. 6.1), and DEEDS has explicitly designed the modality independent features in its registration. SAME takes a different approach that uses the modality invariant SAM embeddings to align images.

We have computed the standard deviation of Jacobian determinants to measure the smoothness of the deformation field. In Table 6.1, it is observed that SAME achieves the best Dice with a certain degree of sacrifice in smoothness. This is mainly because SAME cascades two deformable methods, SAM-coarse (SC) and SAM-VM. The smoothness of SAM-VM alone is slightly better than the original VM (0.36 vs. 0.38), but SC itself brings more non-smoothness (0.40). SC generates a deformation field by directly differentiating two sets of coordinates without any constraint. This approach gives SC more flexibility to model large deformation but may also produce less smoothed results. We will study on adding constraints to improve the smoothness of SC in the future. On the other hand, if SC is not used, SA + SAM-VM can also achieve competing accuracy (52.0% Dice score) with good smoothness (0.36), where the overall performance is still comparable to DEEDS (52.7%, 0.40) while significantly better than FFD (49.4%, 0.51), and SA+VM (50.8%, 0.38).

Organ-specific results are shown in Fig. 6.2. For the sake of conciseness, we divide the 35 organs in our dataset into 9 groups and calculate the median and inter-quartile range of Dice score within each group. The affine in Fig. 6.2 is from Elastix [64], whereas the VoxelMorph refers to SAM-affine + VM [9] in Table 6.1. The results of SAME surpass DEEDS on 8 out of 9 groups except heart in the within-phase condition. In the cross-phase setting, SAME outperforms DEEDS on the artery, bone, airway and lung organs. In other organs, like esophagus and muscle, SAME shows results with smaller variance and comparable median performance with DEEDS. Organ groups such as artery, esophagus, vein, and muscle display lower Dice scores for all methods because they are typically small and can be confused with surrounding tissues. Qualitative examples are illustrated in Fig. 6.3. Manual organ masks of the fixed images are overlaid to show whether the warped moving images align well with the fixed image. Arrows pointed to regions where SAME works better than other methods.

6.4 Conclusion

In this paper, we propose SAME, a fast and accurate framework for unsupervised medical image registration. We expect SAM-affine and SAM-coarse to be promising alternatives of traditional optimization-based methods for registration initialization. The SAM correlation feature and SAM loss may also be combined with other learning-based algorithms [74, 154] for further accuracy improvement.

Chapter 7

Hybrid Feature Engineering for False Positive Reduction in Pancreatic Lesion Detection

7.1 Introduction

Pancreatic cancer is a low-incident but highly lethal disease. Patients with pancreatic cancer suffer from low survival rate with less than 5% surviving five years and the average life expectancy is less than six months after diagnosed with metastatic pancreatic cancer [123]. However, even expert radiologists struggle to distinguish the real tiny cancer, which causes the missing of best treatment period. To deal with this, automatically detecting the cancer in an early stage is urgently needed.

Recently, deep convolutional neural networks (CNNs) have shown great promise in detecting pancreatic cancer due to their strong ability to learn features automatically in a data-driven manner [142, 161, **?**, 163]. Nevertheless, it is still challenging to achieve a satisfying pancreatic lesion detection result, since there is a trade-off between sensitivity and specificity. False positive reduction (FPR) has been widely used in many lesion detection system in order to maintain a high sensitivity as well as a high specificity. Significant efforts have been dedicated to reducing the number of false positives for pulmonary nodule detection [130, 27, 119, 127]. False positive reduction for pancreatic tumors, however, is under-explored in previous research for at least two reasons. First, the annotated data is extremely scarce on pancreatic cancers compared to other diseases, especially for those containing really small cancers, which impedes the progress of detecting pancreatic tumors. More importantly, small pancreatic tumors cannot be easily identified as they share similar textures with surrounding tissues, while the contrast is larger in other pulmonary nodule cases (*e.g.*, for pulmonary nodules the surrounding environment is mostly air). Pancreatic tumors smaller than 2cm are often inconspicuous on computerized tomography (CT) scans even with radiologists' expertise [4]. This suggests that false positives of pancreatic tumors tend to carry similar appearance to true positives (*e.g.*, false positives correspond to the focal fat infiltration in the pancreas), which makes this task more challenging.

Consequently, there is a large room for improvement in false positive reduction for pancreatic tumors. On one hand, for feature extraction, most existing CNN-based false positive reduction methods focus on how to design powerful networks to extract features from candidate patches, such as by exploiting different levels of contexts [27, 57] and multi-view cues [119]. However, as pointed out by a recent work [35], the features learned by CNNs tend to bias towards texture. We claim that such biased feature representation is less discriminative in the FPR stage since the major discrepancy towards appearance has been filtered out more or less by the first stage detection, where a deep neural network is typically used to provide initial detection results. Also in [130, 21, 81, 83], there is an evidence that other features like shape can provide complementary information for more accurate classification. On the other hand, for feature fusion, although there are previous studies to fuse CNNs features with hand-crafted features [130, 125] by concatenating them to different level of CNN outputs, we argue that it is suboptimal since CNNs features and hand-crafted features are different, where for hand-crafted features each dimension can represent different information while for CNNs features information is mixed up in a high dimensional space. Thus, a robust feature engineering is highly required for exploiting the information in these different feature spaces to facilitate false positive reduction.

To address above issues, we propose a Hybrid Feature Engineering framework (HFE) to tackle challenges in the false positive reduction for pancreatic lesion detection. We first extract features that represent different information, *i.e.* texture, shape and uncertainty, including both hand-crafted and CNNs features to construct a hybrid feature pool. Then a combination of features from the pool is picked using a sequential feature selection technique. Finally we train a random forest to distinguish false positives from true positives. The random forest is chosen for its strong interpretability and capability in dealing with hand-crafted data. To better apply feature selection and tree-based classification to the learning based features, we apply principal component analysis (PCA) on CNNs features before adding it into the feature pool. Our preliminary study shows that the picked combination

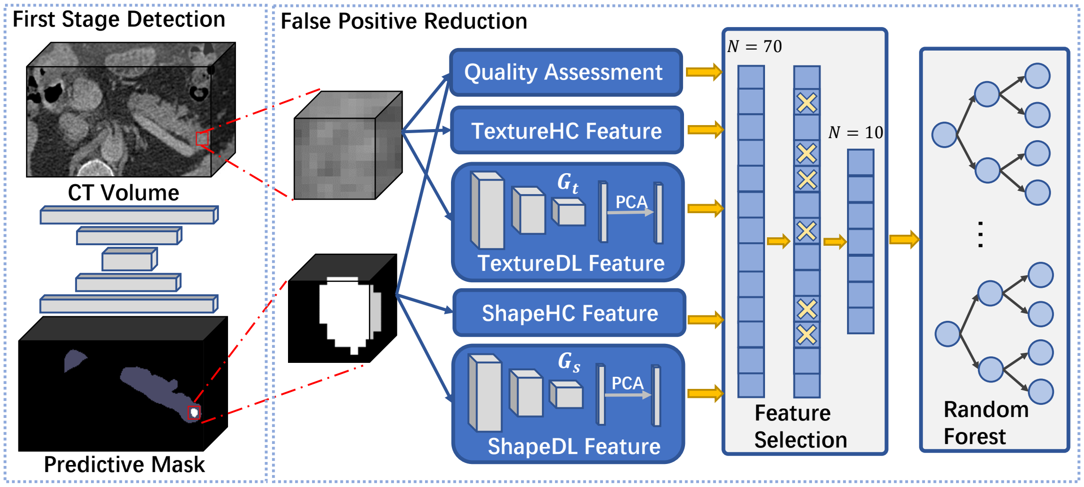


Figure 7.1: The framework of our proposed Hybrid Feature Engineering framework.

of features boost the performance of FPR, and PCA on CNNs features can help both feature selection and random forest based method. We report 80% accuracy in classifying 475 tumor candidates generated from 281 3D CT scans and reduce the false positive rate by a relatively 60.3% from 0.63/scan to 0.25/scan with an only 3.6% sensitivity drop.

7.2 Method

In this section, we describe details of the proposed false positive reduction method, including (i) the initial detection algorithm to provide lesion candidates, and hybrid feature engineering for (ii) comprehensive feature extraction by both hand-crafted and learning based processes and (iii) feature selection and classifier learning for false positive reduction.

7.2.1 Pancreatic Lesion Detection

Given that pancreatic lesion detection is a challenging task, we need the sensitivity of the detection stage to be as high as possible so that the false positive reduction after that makes more sense. Our focus of this paper is to better reject false positives while keeping the sensitivity comparable as before. In this stage we choose a state-of-art segmentation framework for pancreatic cancer detection [142] to generate the initial lesion candidates, which reported 97% sensitivity on a large CT dataset containing subjects with or without pancreatic tumors.

Formally, let $X \in \mathcal{X}$ be the CT scan volume, F be the segmentation algorithm and Y be the voxel-wise annotation of class $N = \{\text{background}, \text{pancreas}, \text{lesion}\}$. We learn the parameters of F under the supervision of Y, following [142]. After training, the estimation $\hat{Y}$ is obtained by testing F on X for a separate split. Then we extract patches centered on the connected components of lesion prediction of $\hat{Y}$ from X, Y and $\hat{Y}$ correspondingly and assign each patch with a binary label indicating whether $\hat{Y}$ and Y overlaps in this patch area on the lesion class. We denote the patch data and the binary label as $X^p, Y^p, \hat{Y}^p, l$, respectively.

7.2.2 Comprehensive Feature Extraction

For each candidate patch obtained in section 2.1, we extract its features from three different perspectives, namely quality assessment (QA), shape and texture. The QA feature is usually targeted at anomaly detection. In FPR we treat the properties within tumor region as target distribution so that the false positives, which do not correspond to tumor region become anomalies. Shape and texture can represent orthogonal properties of pancreatic lesions so that they provide complementary cues when combined together.

7.2.2.1 Quality Assessment Feature

It is shown in [58] that the entropy based uncertainty helps assess the quality of segmentation. Here we use it to distinguish between false positives and true positives, since segmentation with bad quality are more likely to be a false positive. We calculate the uncertainty in a way by accumulating the entropy on the voxel that is predicted as lesion in $\hat{Y}^p$. Specifically, we have

$$f_{\text{entropy}} = -\frac{1}{|\Omega|} \sum_{i \in \Omega} \sum_{c \in N} P(\hat{Y}_i^p = c) \log P(\hat{Y}_i^p = c),$$

where $\Omega = \{i | \arg\max_{c \in N} P(\hat{Y}_i^p = c) = \text{lesion}\}$.

We combine another feature proposed in [75], where a variational auto-encoder (VAE) is learned to reconstruct the ground-truth segmentation and then the reconstruction error of the predictive segmentation is used to evaluate the segmentation quality. Specifically,

$$f_{\text{vae}} = DSC(\hat{Y}^p, \text{VAE}(\hat{Y}_i^p)),$$

where DSC represents dice coefficient score.

7.2.2.2 Shape Feature

The shape feature is important in the organ segmentation since human organs usually has template 3D shapes. In [21] and [129], hand-crafted features on shape are used in FPR of liver tumor and pulmonary nodules respectively. We extract 15 hand-crafted shape features as f_{shapeHC} using the radiomics library [132].

Apart from hand-crafted feature, we also extract learning based shape feature by training a neural network G_s to classify the false positives. We learn the parameter of G_s by minimizing the cross entropy loss, *i.e.*,

$$G_s^* = \arg\min_{G_s} \frac{1}{|\hat{\mathcal{Y}}^p|} \sum_{y \in \hat{\mathcal{Y}}^p} -l_y \log\left[G_s(y)\right] - (1 - l_y) \log\left[1 - G_s(y)\right],$$

where $\hat{\mathcal{Y}}^p$ contains all the predicted patch in section 2.1. In order to fit the input size of neural network, we resample the input patches into the same dimension first. As in [81] we force the neural network to learn the shape feature by feeding it with a binary mask to make a decision. After converging, we extract the feature from the output of second last fully connected layer and apply Principal Component Analysis (PCA) on it. The PCA is to make the feature of neural network independent on each dimension as well as apply dimension reduction. We show that this benefits when applying feature selection and tree-based classification on it. We denote this learning based feature after PCA as f_{shapeDL}.

7.2.2.3 Texture Feature

The texture is usually the crucial feature in many computer vision task. However in traditional FPR, the candidates are usually obtained by a learned neural network, which can bias to texture properties [35], which makes the texture feature less discriminative in the FPR stage. To compare the effect with other types of feature, we also extract texture based features. For hand-crafted texture features, we extract gray level co-occurrence matrix (GLCM) features and first order statistics using [132] and form a 29 dimension feature $f_{\text{textureHC}}$.

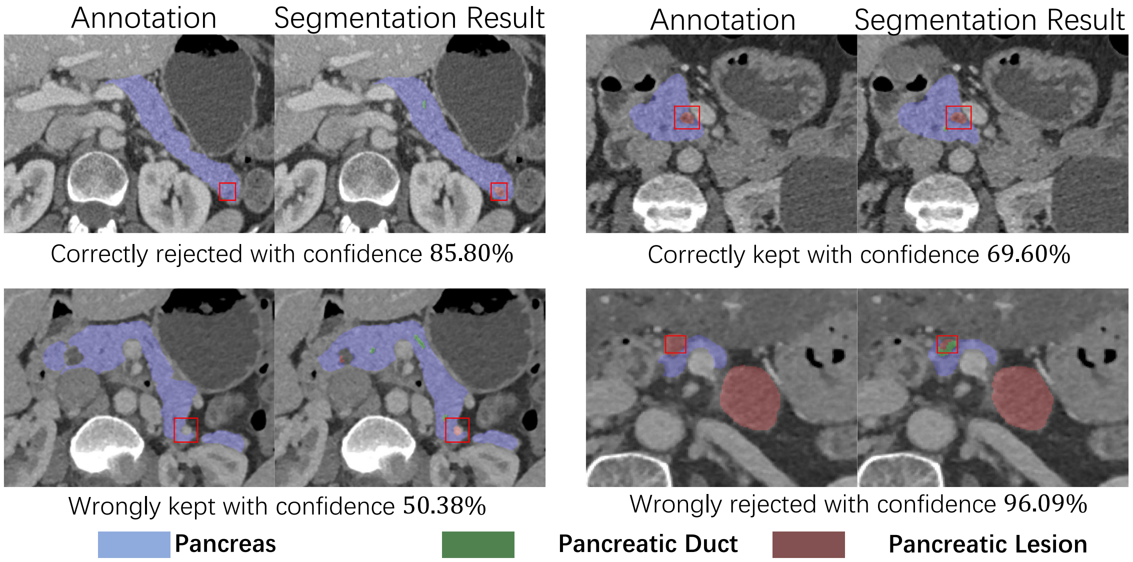
Similarly we extract learning based texture feature $f_{\text{textureDL}}$ by training a neural network G_t of the same structure with G_s. The difference is that we only input X^p instead of $\hat{Y}^p$ to let it bias to the texture properties. The feature is extracted from the same fully connected layer and applied PCA afterwards.

7.2.3 Feature Selection and Classifier Learning

We first apply sequential feature selection [31] on the hybrid feature pool. Specifically, starting from an empty set, we pick one feature at one time from the remaining feature pool that minimize a validation loss, until the number of picked features becomes m. At each loop we train a random forest to learn false positive label l for each candidate based on the picked features and test it on a validation set to calculate error. After feature selection, we finally train a new random forest under the same parameter setting as during feature selection

Table 7.1: Results for pancreatic lesion detection on testing set before and after FPR. SEN per correspondence means sensitivity in terms of total number of lesions

	SEN per	SEN per	SPE per	FPs per
	case(%)	correspondence(%)	case(%)	scan
Init. Detection	93.3	85.3	82.5	0.63
Init. Detection+FPR	89.7	77.3	91.2	0.25



Correctly rejected with confidence **85.80%**

Correctly kept with confidence **69.60%**

Wrongly kept with confidence **50.38%**

Wrongly rejected with confidence **96.09%**

Figure 7.2: Visualization of FPR examples. Tumor candidates are within the red box. Our method can distinguish false positives even with small size.

7.3 Experiments

In this section, we evaluate the proposed Hybrid Feature Engineering (HFE) on a large scale 3D CT dataset. We compare different combinations of feature sets with our feature selection result and also different classifiers to prove the effectiveness of our proposed HFE.

7.3.0.1 Datasets

The experiments are conducted on a large scale 3D CT dataset containing 2285 subjects. The CT volume is of size 512x512xL,where the resolution for sagittal and coronal axis is 0.6-0.9mm and for axial is 0.5mm. In the dataset there are 1734 biopsy-proven abnormal cases, containing at least one of the three types of pancreatic lesion, namely pancreatic ductal adenocarcinoma (PDAC), pancreatic cystic lesions (Cyst) and pancreatic neuroendocrine tumors (PNET). This dataset is collected under the help of radiologists and voxel-wise annotation for the lesion is available. The size of lesion ranges from 5mm to 120mm in diameter. The variance of size causes extra difficulty in lesion detection, since detecting smaller lesion usually means more false positives, and that motivates us to develop this specialized FPR stage.

7.3.0.2 Experimental Settings

We first split the dataset into two separate parts $\mathcal{X}_1, \mathcal{X}_2$, where $\mathcal{X}_1$ contains 1062 abnormal cases and 400 normal cases and $\mathcal{X}_2$ contains 672 abnormal cases and 151 normal cases (823 in total). The first part $\mathcal{X}_1$ is used to train a detection model that provides the initial candidates as mentioned in section 2.1. We then test that model on $\mathcal{X}_2$ and conduct the FPR experiments on it. In our experiments, the segmentaion model generate 1359 candidates in total on $\mathcal{X}_2$, where 823 of them are true positives, *i.e.*, have overlap with the lesion annotation, and 536 are false positives. Initial detection result is as shown in Tab.7.1. We treat the FPR as a binary classification problem (1 for keeping the candidate and 0 for rejecting it). We

Table 7.2: Results for FPR using different feature sets. N is the number of used feature

Feature	N	ACC(%)	SEN(%)	SPE(%)	AUC(%)	F-score(%)
textureHC	29	76.89	81.51	69.71	81.89	81.10
textureDL	12	73.50	79.14	64.73	79.15	78.42
shapeHC	15	75.99	76.71	74.87	82.06	79.53
shapeDL	12	75.44	78.37	70.88	84.44	79.52
QA	2	76.71	81.18	69.77	84.42	80.91
All	70	76.63	79.62	71.98	83.01	80.56
All+FS (Ours)	10	**80.20**	**83.07**	**75.75**	**85.88**	**83.62**

further equally partition $\mathcal{X}_2$ into training, validation and testing set,which contains 434, 450, 475 candidates respectively. The model for generating learning based features and random forest for classifying false positives are trained on the training set. The random forest consists of 20 decision trees with maximum depth 3 and minimum leaf size 30. The validation loss during feature selection is calculated on the validation set. Detailed structure of models used in FPR can be found in appendix. The final result is reported on the testing set. For the learning based feature, the original dimension is 64, we extract the first 12 principle components, which explain 80% of the data. We evaluate different methods using standard metrics including accuracy (ACC), sensitivity (SEN), specificity (SPE) and $F1$-score. We also report the area under the receiver operating characteristic (ROC) curve, AUC to measure the performance. To eliminate randomness, each experiment is repeated for 50 times and report the average result.

7.3.0.3 Results and Discussion

We summarize the results in this section,including the performance using different feature sets (Tab.7.2), different number of selected features (Fig. 7.4), ablations of

Table 7.3: Comparison results for FPR with different settings PCA and Feature Selection(FS). N for number of used features

FS	PCA	N	ACC(%)	SEN(%)	SPE(%)	AUC(%)	F-score(%)
×	×	174	76.19	81.58	67.81	83.14	80.65
×	✓	70	76.63	79.62	71.98	83.01	80.56
✓	×	10	78.13	81.56	72.82	85.30	81.94
✓	✓	10	**80.20**	**83.07**	**75.75**	**85.88**	**83.62**

PCA and feature selection (Tab.7.3).

In Tab.7.2, we compare our method with simply input one group of feature defined in section 2.2 or input all the features. The result shows that the both shape feature and QA feature perform better than learning based texture feature, which proves the idea that the texture based features are less discriminative in the FPR stage. The QA feature achieves the highest performance with only 2 features, showing that quality assessment task can provide crucial cues in FPR. In our experiments, using all the features do not yield to better result since the limited number of training data will lead to overfitting can cause performance drop on the test set. Our best result is achieved after feature selection. We select 10 features from the hybrid feature pool as described in section 2.3 and observe that those 10 features are made up by features from all the feature set we have mentioned, which proves the importance of building up the hybrid feature pool. Details of those 10 selected features can be found in the appendix.

To illustrate the effect of feature selection and PCA, in Fig.7.4 we vary the number of selected features and report the result on both validation and testing set. We can see that on the validation set, as the number of picked features increases, the

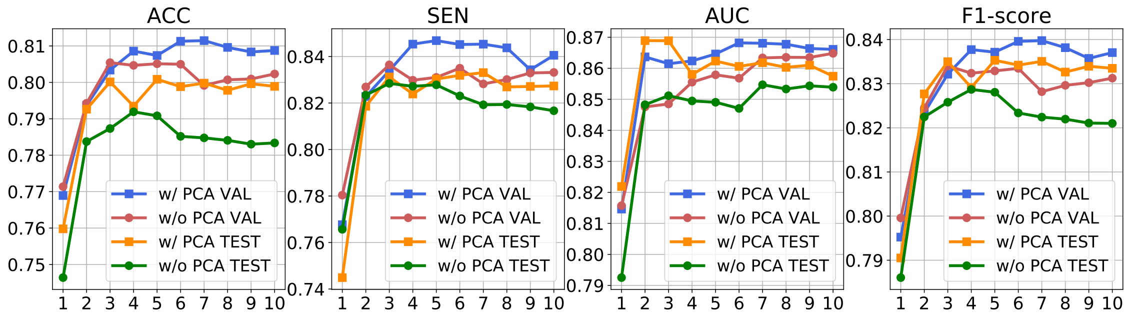


Figure 7.3: Visualization of how the number of picked features affect the result. X-axis is for the number of selected features.

result tends to get better. That is because we are using the error on the validation set to guide the selection. However on the test set, we observe more vibration on the result, indicating that the feature selection method may have chance to overfit the validation set. But when comparing selecting feature with and without PCA, the result after PCA is consistently better than before PCA. In Tab.7.3 we conduct throughout ablation to show how the feature selection (FS) and PCA affect the results. The feature selection procedure boosts the performance significantly (2% w/o PCA and 3.6% w/PCA) and applying PCA both decreases the feature dimension and makes learning based feature orthogonal so that it will further benefit the feature selection.

We visualize several typical conditions as in Fig.7.2. The statistic shows that around 70% of false positives in our experiments are smaller than 1cm in diameter and our method can correctly reject 52% of them while keeping 93% of the true positives. Our method is not able to handle cases where FPs are outside the pancreas or mixed with pancreatic duct prediction, since such information is still not exploited in the feature engineering framework, which could be a promising direction for the future work.

7.4 Conclusion

We propose a Hybrid Feature Engineering method for false positive reduction in pancreatic lesion detection. We show that CNNs features can be better merged with hand-crafted features with PCA and random forest classifier. In our experiments, hand-crafted shape feature combined with learning based texture feature and QA feature achieves the best result. With this optimal combination, we finally reduce the false positive rate from 0.62/scan to 0.25/scan, while sensitivity only changes from 93.3% to 89.7% on a large CT dataset containing 281 volumes. Given that FPR in pancreatic lesion detection is still under-explored in most current research, the proposed method already achieves promising result in the first attempt.

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

7.5 Appendix

7.5.1 Feature List

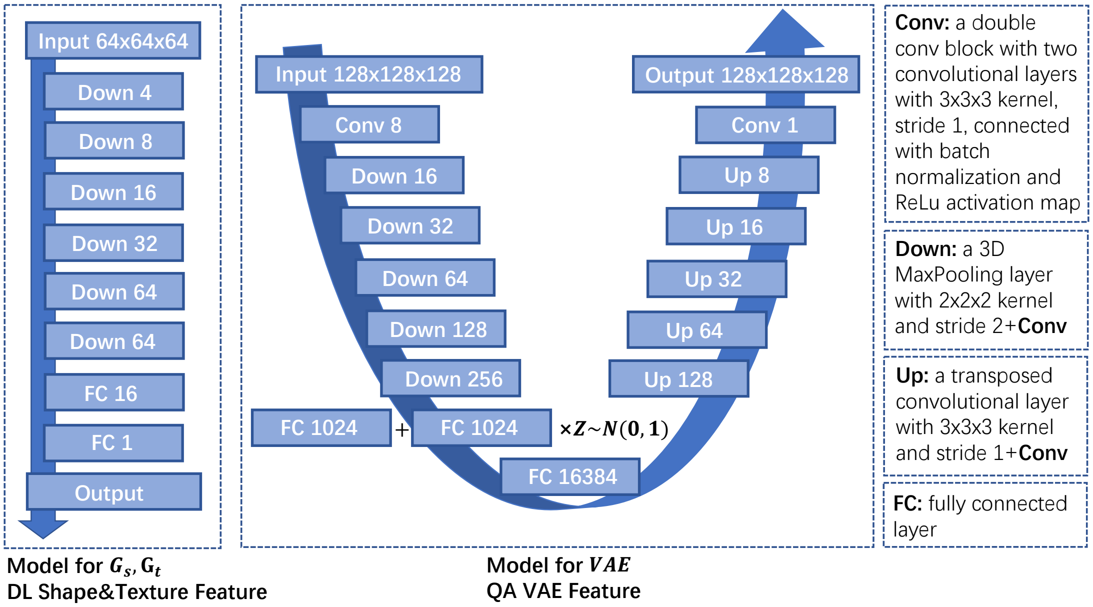
We include all the features used in our HFE system in Tab.7.4 and the selected features after feature selection in Tab.7.5.

Table 7.4: All features used in HFE system

HC Shape Features (HCS)	HC Texture Features (HCT)	
Radius	Autocorrelation	JointAverage
Elongation	ClusterProminence	JointEnergy
Flatness	ClusterShade	JointEntropy
LeastAxisLength	ClusterTendency	MCC
MajorAxisLength	Contrast	MaximumProbability
Maximum2DDiameterColumn	Correlation	SumAverage
Maximum2DDiameterRow	DifferenceAverage	SumSquares
Maximum2DDiameterSlice	DifferenceEntropy	SumEntropy
Maximum2DDiameterRow	DifferenceVariance	Kurtosis
Maximum3DDiameter	InverseDifference	MeanAbsoluteDeviation
MeshVolume	IDM	Range
MinorAxisLength	IDMN	Skewness
Sphericity	IDN	Uniformity
SurfaceArea	IMC1	
SurfaceVolumeRatio	IMC2	
VoxelVolume	InverseVariance	
QA Features (QA)	DL Shape Features (DLS)	DL Texture Features (DLT)
ENTROPY	Shape-PC-1 to	Texture-PC-1 to
VAE	Shape-PC-12	Texture-PC-12

Table 7.5: Feature selection(FS) result by the sequential feature selection method. Note that without applying PCA, the deep learning features are 64 dimension instead of 12 when applying PCA.

FS w/ PCA	Feature Group	FS w/o PCA	Feature Group
SurfaceVolumeRatio	HCS	Shape-DL-40	DLS
ENTROPY	QA	ENTROPY	QA
Shape-PC-12	DLS	Shape-DL-23	DLS
Texture-PC-8	DLT	Shape-DL-32	DLS
Range	HCT	Texture-DL-26	DLT
VAE	QA	Shape-DL-39	DLS
Texture-PC-11	DLT	Shape-DL-64	DLS
SumEntropy	HCT	Sphericity	HCS
MaximumProbability	HCT	Shape-DL-25	DLS
IDN	HCT	JointAverage	HCT


Input 64x64x64
Down 4
Down 8
Down 16
Down 32
Down 64
Down 64
FC 16
FC 1
Output
Input 128x128x128
Conv 8
Down 16
Down 32
Down 64
Down 128
Down 256
FC 1024
+
FC 1024
×$Z \sim N(0,1)$
FC 16384
Output 128x128x128
Conv 1
Up 8
Up 16
Up 32
Up 64
Up 128
Conv: a double conv block with two convolutional layers with 3x3x3 kernel, stride 1, connected with batch normalization and ReLu activation map
Down: a 3D MaxPooling layer with 2x2x2 kernel and stride 2+**Conv**
Up: a transposed convolutional layer with 3x3x3 kernel and stride 1+**Conv**
FC: fully connected layer

Model for G_s, G_t
DL Shape&Texture Feature

Model for VAE
QA VAE Feature

Figure 7.4: The model structure used for extracting learning based shape and texture features and QA-VAE feature. Both model are trained with SGD optimizer and learning rate 0.01. Random rotation and scaling are applied for data augmentation.

7.5.2 Model Structure

We include the model structure used for extracting deep shape feature, deep texture feature and QA VAE feature in Fig.7.4

Chapter 8

Conclusion

8.1 Summary

In this dissertation, we focus on topics on knowledge fusion and the applications in fundamental tasks in medical image analysis. In chapter 2, we study the shape prior knowledge of organ segmentation with a variational auto-encoder and apply it for automatic quality assessment. Then we focus on detecting pancreatic cancer. In chapter 3, we propose to obtain the shape representation with a learned auto-encoder from the predictive pancreas segmentation and then use it for classifying abnormal patients. Further in chapter 4, we design a fusion framework which combines information from both shape and texture extracted from predictive segmentation and CT scans. Next we aim at improving multi-modal image registration. Considering the relationship between image segmentation, synthesis and registration, in chapter 5, we propose a joint multi-task framework in order to benefit each individual task by incorporating additional knowledge via joint learning. In chapter 6, we improve current deep image registration methods by introducing high-level semantic representation of images into both feature level and loss function of basic model. Finally in chapter 7, we design a pipeline for false positive removal in pancreatic lesion detection. It first extracts features from different perspectives including shape, texture, uncertainty, quality. Then the features are selected and used to train a classification model.

8.2 Future Work

Although with the development of deep learning, the performance on the public benchmarks shows consistently improvement, there still exist challenges in order to really benefit the efficiency and accuracy when deploying AI system into clinical procedures. Firstly, poor robustness of model limits the performance in the scenario of large-scale testing, where the data comes from various sources. A reliable quality assessment system and the ability to adapt to different source of data is necessary. Secondly, the ability to reasoning is crucial for both doctors and AI system in the diagnosis procedure. The interpretation of the result from AI system will help doctors understand the condition of patients and give more accurate clinical decisions. Thirdly, current algorithms are usually designed for a specific organ or lesion. It is worth exploring whether transferring knowledge from multiple tasks benefits the model robustness.

Bibliography

[1] Seer cancer statistics review 1975-2015. *National Cancer Institute. Bethesda, MD*.

[2] Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probability. 2015.

[3] Moab Arar, Yiftach Ginger, Dov Danon, Amit H Bermano, and Daniel Cohen-Or. Unsupervised multi-modal image registration via geometry preserving image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13410–13419, 2020.

[4] J. C. Ardengh, G. A. de Paulo, and A. P. Ferrari. Pancreatic carcinomas smaller than 3.0 cm: endosonography (EUS) in diagnosis, staging and prediction of resectability. *HPB (Oxford)*, 5(4):226–230, 2003.

[5] Brian B Avants, Charles L Epstein, Murray Grossman, and James C Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, 12(1):26–41, 2008.

[6] B B Avants, C L Epstein, M Grossman, and J C Gee. Symmetric diffeomorphic

image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Med. Image Anal.*, 12(1):26–41, 2008.

[7] Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. Voxelmorph: A learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, 2019.

[8] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800, 2019.

[9] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. VoxelMorph: A Learning Framework for Deformable Medical Image Registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, 2019.

[10] Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Deep autoencoding models for unsupervised anomaly segmentation in brain MR images. *CoRR*, abs/1804.04488, 2018.

[11] Max Blendowski and Mattias P Heinrich. Learning interpretable multi-modal features for alignment with supervised iterative descent. In *International Conference on Medical Imaging with Deep Learning*, pages 73–83, 2019.

[12] S. Bosse, D. Maniry, K. Müller, T. Wiegand, and W. Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27(1):206–219, Jan 2018.

[13] A. Brock, T. Lim, J. M. Ritchie, and N. Weston. Generative and discriminative voxel modeling with convolutional neural networks. *arXiv:1608.04236*, 2016.

[14] K S Caldemeyer and K A Buckwalter. The basic principles of computed tomography and magnetic resonance imaging. *J Am Acad Dermatol*, 41(5 Pt 1):768–771, Nov. 1999.

[15] Xiaohuan Cao, Jianhua Yang, Yaozong Gao, Yanrong Guo, Guorong Wu, and Dinggang Shen. Dual-core steered non-rigid registration for multi-modal images via bi-directional image synthesis. *Medical image analysis*, 41:18–31, 2017.

[16] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[17] Sebastien Chabrier, Bruno Emile, Christophe Rosenberger, and Helene Laurent. Unsupervised performance evaluation of image segmentation. *EURASIP Journal on Applied Signal Processing*, 2006:217–217, 2006.

[18] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[19] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

[20] Xuan Chen, Yufei Chen, Chao Ma, Xianhui Liu, and Xin Tang. Classification of pancreatic tumors based on mri images using 3d convolutional neural networks. In *ISICDM*, 2018.

[21] G. Chlebus, A. Schenk, J. H. Moltz, B. van Ginneken, H. K. Hahn, and H. Meine. Automatic liver tumor segmentation in CT with fully convolutional neural networks and object-based postprocessing. *Sci Rep*, 8(1):15497, 10 2018.

[22] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-net: Learning dense volumetric segmentation from sparse annotation. In *MICCAI*, volume 9901 LNCS, pages 424–432, 2016.

[23] Bob D de Vos, Floris F Berendsen, Max A Viergever, Hessam Sokooti, Marius Staring, and Ivana Išgum. A deep learning framework for unsupervised affine and deformable image registration. *Medical image analysis*, 52:128–143, 2019.

[24] Bob D de Vos, Floris F Berendsen, Max A Viergever, Marius Staring, and Ivana Išgum. End-to-end unsupervised deformable image registration with a convolutional neural network. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 204–212. Springer, 2017.

[25] CARL DOERSCH. Tutorial on variational autoencoders. *stat*, 1050:13, 2016.

[26] Alexey Dosovitskiy, Philipp Fischery, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas

Brox. FlowNet: Learning optical flow with convolutional networks. In *ICCV*, volume 2015 Inter, pages 2758–2766, 2015.

[27] Q. Dou, H. Chen, L. Yu, J. Qin, and P. Heng. Multilevel contextual 3-d cnns for false positive reduction in pulmonary nodule detection. *IEEE Transactions on Biomedical Engineering*, 64(7):1558–1567, 2017.

[28] Qi Dou, Cheng Ouyang, Cheng Chen, Hao Chen, and Pheng-Ann Heng. Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss. In *IJCAI International Joint Conference on Artificial Intelligence*, page 691, 2018.

[29] Rosana El Jurdi, Caroline Petitjean, Paul Honeine, Veronika Cheplygina, and Fahed Abdallah. High-level prior-based loss functions for medical image segmentation: A survey. *Computer Vision and Image Understanding*, 210:103248, 2021.

[30] Joshua J. Fenton, Stephen H. Taplin, Patricia A. Carney, Linn Abraham, Edward A. Sickles, Carl D'Orsi, Eric A. Berns, Gary Cutter, R. Edward Hendrick, William E. Barlow, and Joann G. Elmore. Influence of computer-aided detection on performance of screening mammography. *New England Journal of Medicine*, 356(14):1399–1409, 2007. PMID: 17409321.

[31] F.J. Ferri, P. Pudil, M. Hatef, and J. Kittler. Comparative study of techniques for large-scale feature selection. volume 16 of *Machine Intelligence and Pattern Recognition*. North-Holland, 1994.

[32] Roman Filipovych, Christos Davatzikos, and Alzheimer's Disease Neuroimaging Initiative. Semi-supervised pattern classification of medical images: application to mild cognitive impairment (MCI). *Neuroimage*, 55(3):1109–1119, Dec. 2010.

[33] Shuhao Fu, Yongyi Lu, Yan Wang, Yuyin Zhou, Wei Shen, Elliot Fishman, and Alan Yuille. Domain adaptive relational reasoning for 3D multi-organ segmentation. In Anne L Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A Zuluaga, S Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 656–666, Cham, 2020. Springer International Publishing.

[34] Han Gao, Yunwei Tang, Linhai Jing, Hui Li, and Haifeng Ding. A novel unsupervised segmentation quality evaluation method for remote sensing images. *Sensors*, 17(10):2427, 2017.

[35] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

[36] Dazhou Guo, Xianghua Ye, Jia Ge, Xing Di, Le Lu, Lingyun Huang, Guotong Xie, Jing Xiao, Zhongjie Lu, Ling Peng, Senxiang Yan, and Dakai Jin. DeepStationing: Thoracic Lymph Node Station Parsing in CT Scans using Anatomical Context Encoding and Key Organ Auto-Search . In *MICCAI*, volume LNCS, 2021.

[37] Adam P Harrison, Ziyue Xu, Kevin George, Le Lu, Ronald M Summers, and Daniel J Mollura. Progressive and multi-path holistically nested neural networks for pathological lung segmentation from ct images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 621–629. Springer, 2017.

[38] Adam P Harrison, Ziyue Xu, Kevin George, Le Lu, Ronald M Summers, and Daniel J Mollura. Progressive and multi-path holistically nested neural networks for pathological lung segmentation from CT images. In *MICCAI*, volume 10435 LNCS, 2017.

[39] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *ACCV*, 2017.

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[41] Mattias Heinrich, Oskar Maier, and Heinz Handels. Multi-modal multi-atlas segmentation using discrete optimisation and self-similarities. *CEUR Workshop Proceedings*, 2015.

[42] Mattias P Heinrich and Lasse Hansen. Highly accurate and memory efficient unsupervised learning-based discrete CT registration using 2.5 D displacement search. In *MICCAI*, 2020.

[43] Mattias P Heinrich, Mark Jenkinson, Manav Bhushan, Tahreema Matin, Fergus V Gleeson, Sir Michael Brady, and Julia A Schnabel. MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration. *Medical Image Analysis*, 16(7):1423–1435, 2012.

[44] Mattias P Heinrich, Mark Jenkinson, Michael Brady, and Julia A Schnabel. Globally optimal deformable registration on a minimum spanning tree using dense displacement sampling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 115–122. Springer, 2012.

[45] Mattias P Heinrich, Mark Jenkinson, Michael Brady, and Julia A Schnabel. Mrf-based deformable registration and ventilation estimation of lung ct. *IEEE transactions on medical imaging*, 32(7):1239–1248, 2013.

[46] Mattias P Heinrich, Mark Jenkinson, Sir Michael Brady, and Julia A Schnabel. Globally optimal deformable registration on a minimum spanning tree using dense displacement sampling. In *MICCAI*, volume 7512 LNCS, pages 115–122, 2012.

[47] Mattias Paul Heinrich, Mark Jenkinson, Bartlomiej W Papież, Michael Brady, and Julia A Schnabel. Towards realtime multimodal fusion for image-guided interventions using self-similarities. In *International conference on medical image computing and computer-assisted intervention*, pages 187–194. Springer, 2013.

[48] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[49] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[50] Xiaojun Hu, Miao Kang, Weilin Huang, Matthew R. Scott, Roland Wiest, and Mauricio Reyes. Dual-stream pyramid registration network. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Springer International Publishing, 2019.

[51] Yipeng Hu, Marc Modat, Eli Gibson, Wenqi Li, Nooshin Ghavami, Ester Bonmati, Guotai Wang, Steven Bandula, Caroline M Moore, Mark Emberton, et al. Weakly-supervised convolutional neural networks for multimodal image registration. *Medical image analysis*, 2018.

[52] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018.

[53] Yuankai Huo, Zhoubing Xu, Hyeonsoo Moon, Shunxing Bao, Albert Assad, Tamara K Moyo, Michael R Savona, Richard G Abramson, and Bennett A Landman. Synseg-net: Synthetic segmentation without target modality ground truth. *IEEE transactions on medical imaging*, 2018.

[54] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, Feb 2021.

[55] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[56] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[57] H. Jin, Z. Li, R. Tong, and L. Lin. A deep 3D residual CNN for false-positive reduction in pulmonary nodule detection. *Med Phys*, 45(5):2097–2107, May 2018.

[58] Alain Jungo, Raphael Meier, Ekin Ermis, Evelyn Herrmann, and Mauricio Reyes. Uncertainty-driven sanity check: Application to postoperative brain tumor cavity segmentation. *CoRR*, abs/1806.03106, 2018.

[59] Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia Newcombe, Joanna Simpson, Andrew Kane, David Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert, and Ben Glocker. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In Marc Niethammer, Martin Styner, Stephen Aylward, Hongtu Zhu, Ipek Oguz, Pew-Thian Yap, and Dinggang Shen, editors, *Information Processing in Medical Imaging*, pages 597–609, Cham, 2017. Springer International Publishing.

[60] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017.

[61] Michael D Ketcha, Tharindu S De Silva, Runze Han, Ali Uneri, Sebastian Vogt, Gerhard Kleinszig, and Jeffrey H Siewerdsen. Learning-based deformable image registration: effect of statistical mismatch between train and test images. *Journal of Medical Imaging*, 2019.

[62] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[63] Diederik P Kingma and Max Welling. Auto-encoding variational bayes [j]. 2013.

[64] Stefan Klein, Marius Staring, Keelin Murphy, Max A. Viergever, and Josien P.W. Pluim. Elastix: A toolbox for intensity-based medical image registration. *IEEE Transactions on Medical Imaging*, 29(1):196–205, 2010.

[65] Timo Kohlberger, Vivek Singh, Chris Alvino, Claus Bahlmann, and Leo Grady. Evaluating segmentation error without ground truth. In *MICCAI*, pages 528–536. Springer, 2012.

[66] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[67] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*. 2012.

[68] Yongchan Kwon, Joong-Ho Won, Beom Joon Kim, and Myunghee Cho Paik. Uncertainty quantification using bayesian neural networks in classification:

Application to ischemic stroke lesion segmentation. 2018.

[69] Bo Li, Wiro J Niessen, Stefan Klein, Marius de Groot, M Arfan Ikram, Meike W Vernooij, and Esther E Bron. A hybrid deep learning framework for integrated segmentation and registration: evaluation on longitudinal white matter tract changes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 645–653. Springer, 2019.

[70] Hongming Li and Yong Fan. Non-rigid image registration using fully convolutional networks with deep self-supervision. *arXiv preprint arXiv:1709.00799*, 2017.

[71] Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, Lei Xing, and Pheng-Ann Heng. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):523–534, 2021.

[72] Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

[73] Fengze Liu, Jinzheng Cai, Yuankai Huo, Chi-Tung Cheng, Ashwin Raju, Dakai Jin, Jing Xiao, Alan Yuille, Le Lu, ChienHung Liao, et al. Jssr: A joint synthesis, segmentation, and registration system for 3d multi-modal image alignment of large-scale pathological ct scans. In *European Conference on Computer Vision*, pages 257–274. Springer, 2020.

[74] Fengze Liu, Jinzheng Cai, Yuankai Huo, Chi-Tung Cheng, Ashwin Raju,

Dakai Jin, Jing Xiao, Alan Yuille, Le Lu, ChienHung Liao, and Adam P. Harrison. Jssr: A joint synthesis, segmentation, and registration system for 3d multi-modal image alignment of large-scale pathological ct scans. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 257–274, Cham, 2020. Springer International Publishing.

[75] F. Liu, Y. Xia, D. Yang, A. Yuille, and D. Xu. An alarm system for segmentation algorithm based on shape model. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10651–10660, 2019.

[76] Fengze Liu, Yingda Xia, Dong Yang, Alan L Yuille, and Daguang Xu. An alarm system for segmentation algorithm based on shape model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10652–10661, 2019.

[77] Fengze Liu, Lingxi Xie, Yingda Xia, Elliot Fishman, and Alan Yuille. Joint shape representation and classification for detecting pdac. In *International Workshop on Machine Learning in Medical Imaging*, pages 212–220. Springer, 2019.

[78] Fengze Liu, Lingxi Xie, Yingda Xia, Elliot K. Fishman, and Alan L. Yuille. Joint shape representation and classification for detecting pdac. *ArXiv*, 2018.

[79] Fengze Liu, Ke Yan, Adam P Harrison, Dazhou Guo, Le Lu, Alan L Yuille,

Lingyun Huang, Guotong Xie, Jing Xiao, Xianghua Ye, et al. Same: Deformable image registration based on self-supervised anatomical embeddings. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 87–97. Springer, 2021.

[80] Fengze Liu, Yuyin Zhou, Elliot Fishman, and Alan Yuille. Fusionnet: Incorporating shape and texture for abnormality detection in 3d abdominal ct scans. In *International Workshop on Machine Learning in Medical Imaging*, pages 221–229. Springer, 2019.

[81] Fengze Liu, Yuyin Zhou, Elliot Fishman, and Alan Yuille. Fusionnet: Incorporating shape and texture for abnormality detection in 3d abdominal ct scans. In *Machine Learning in Medical Imaging*, 2019.

[82] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[83] Xiaoming Liu and Zhigang Zeng. A new automatic mass detection method for breast cancer with false positive reduction. *Neurocomputing*, 152:388–402, 2015.

[84] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[85] Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. *Proceedings of*

the AAAI Conference on Artificial Intelligence, 35:8801–8809, 05 2021.

[86] Frederik Maes, Andre Collignon, Dirk Vandermeulen, Guy Marchal, and Paul Suetens. Multimodality image registration by maximization of mutual information. *IEEE transactions on Medical Imaging*, 16(2):187–198, 1997.

[87] Dwarikanath Mahapatra, Bhavna Antony, Suman Sedai, and Rahil Garnavi. Deformable medical image registration using generative adversarial networks. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1449–1453. IEEE, 2018.

[88] JB Antoine Maintz and Max A Viergever. An overview of medical image registration methods. In *Symposium of the Belgian hospital physicists association (SBPH/BVZF)*, volume 12, pages 1–22. Citeseer, 1996.

[89] Kasper Marstal, Floris Berendsen, Marius Staring, and Stefan Klein. Simpleelastix: A user-friendly, multi-lingual library for medical image registration. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2016.

[90] F. Milletari, N. Navab, and S. A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 2016.

[91] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 565–571. IEEE, 2016.

[92] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. pages 565–571, 10 2016.

[93] Tony C. W. Mok and Albert C. S. Chung. Large deformation image registration with anatomy-aware laplacian pyramid networks. *Segmentation, Classification, and Registration of Multi-modality Medical Imaging Data: MICCAI 2020 Challenges, ABCs 2020, L2R 2020, TN-SCUI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Proceedings*, 12587:61–67, Feb 2021.

[94] Keelin Murphy, Bram Van Ginneken, Joseph M Reinhardt, Sven Kabus, Kai Ding, Xiang Deng, Kunlin Cao, Kaifang Du, Gary E Christensen, Vincent Garcia, et al. Evaluation of registration methods on thoracic ct: the empire10 challenge. *IEEE transactions on medical imaging*, 30(11):1901–1920, 2011.

[95] Ozan Oktay, Enzo Ferrante, Konstantinos Kamnitsas, Mattias Heinrich, Wenjia Bai, Jose Caballero, Stuart A Cook, Antonio de Marvao, Timothy Dawes, Declan P O'Regan, et al. Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation. *IEEE transactions on medical imaging*, 37(2):384–395, 2018.

[96] PDQ Adult Treatment Editorial Board. Pancreatic cancer treatment (PDQ®). 2017.

[97] M.A. Pimentel, D.A. Clifton, L. Clifton, and L. Tarassenko. A review of novelty detection. *Signal Processing 99 (2014) 215–249.*

[98] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep

learning on point sets for 3d classification and segmentation. *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 1(2):4, 2017.

[99] Chen Qin, Wenjia Bai, Jo Schlemper, Steffen E Petersen, Stefan K Piechnik, Stefan Neubauer, and Daniel Rueckert. Joint learning of motion estimation and segmentation for cardiac mr image sequences. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 472–480. Springer, 2018.

[100] Chen Qin, Bibo Shi, Rui Liao, Tommaso Mansi, Daniel Rueckert, and Ali Kamen. Unsupervised deformable registration for multi-modal images via disentangled representations. In *International Conference on Information Processing in Medical Imaging*, pages 249–261. Springer, 2019.

[101] Ashwin Raju, Shun Miao, Dakai Jin, Le Lu, Junzhou Huang, and Adam P Harrison. Deep implicit statistical shape models for 3d medical image delineation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2135–2143, 2022.

[102] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[103] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[104] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[105] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

[106] Mauricio Reyes, Miguel A Gonzalez Ballester, Zhixi Li, Nina Kozic, See Chin, Ronald M Summers, and Marius George Linguraru. ANATOMICAL VARIABILITY OF ORGANS VIA PRINCIPAL FACTOR ANALYSIS FROM THE CONSTRUCTION OF AN ABDOMINAL PROBABILISTIC ATLAS. *Proc IEEE Int Symp Biomed Imaging*, 2009:682–685, 2009.

[107] Robert Robinson, Ozan Oktay, Wenjia Bai, Vanya Valindria, Mihir Sanghvi, Nay Aung, José Paiva, Filip Zemrak, Kenneth Fung, Elena Lukaschuk, et al. Real-time prediction of segmentation quality. *arXiv preprint arXiv:1806.06244*, 2018.

[108] Robert Robinson, Vanya V. Valindria, Wenjia Bai, Hideaki Suzuki, Paul M. Matthews, Chris Page, Daniel Rueckert, and Ben Glocker. Automatic quality control of cardiac mri segmentation in large-scale population imaging. In *MICCAI*, 2017.

[109] K. Rohr, H. S. Stiehl, R. Sprengel, T. M. Buzug, J. Weese, and M. H. Kuhn. Landmark-based elastic registration using approximating thin-plate splines. *IEEE Transactions on Medical Imaging*, 2001.

[110] J. T. Rolfe and Y. LeCun. Discriminative recurrent sparse auto-encoders. *arXiv:1301.3775*, 2013.

[111] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[112] Holger Roth, Amal Farag, Le Lu, Baris Turkbey, and Ronald Summers. Deep convolutional networks for pancreas segmentation in ct imaging. 9413, 04 2015.

[113] Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *MICCAI*, pages 556–564. Springer, 2015.

[114] H. R. Roth, L. Lu, A. Farag, H. C. Shin, J. Liu, E. B. Turkbey, and R. M. Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *MICCAI*, 2015.

[115] Daniel Rueckert and Julia A. Schnabel. *Medical Image Registration*, pages 131–154. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[116] D Rueckert, L I Sonoda, C Hayes, D L G Hill, M O Leach, and D J Hawkes. Nonrigid Registration Using Free-Form Deformations: Application to Breast MR Images. *IEEE Trans. Med. Imaging*, 18(8), 1999.

[117] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *Information Processing in*

Medical Imaging, pages 146–157, Cham, 2017. Springer International Publishing.

[118] Philipp Seeböck, Sebastian M. Waldstein, Sophie Klimscha, Bianca S. Gerendas, René Donner, Thomas Schlegl, Ursula Schmidt-Erfurth, and Georg Langs. Identifying and categorizing anomalies in retinal imaging data. *CoRR*, abs/1612.00686, 2016.

[119] A. A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. J. van Riel, M. M. W. Wille, M. Naqibullah, C. I. Sánchez, and B. van Ginneken. Pulmonary nodule detection in ct images: False positive reduction using multi-view convolutional networks. *IEEE Transactions on Medical Imaging*, 35(5):1160–1169, 2016.

[120] Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[121] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

[122] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.

[123] B. W. K. P. Stewart, C. P. Wild, et al. World cancer report 2014. *Health*, 2017.

[124] Colin Studholme, Derek LG Hill, and David J Hawkes. An overlap invariant entropy measure of 3d medical image alignment. *Pattern recognition*, 32(1):71–86, 1999.

[125] Ran Su, Tianling Liu, Changming Sun, Qiangguo Jin, Rachid Jennane, and Leyi Wei. Fusing convolutional neural network features with hand-crafted features for osteoporosis diagnoses. *Neurocomputing*, 2020.

[126] Sharmin Sultana, Daniel Y. Song, and Junghoon Lee. A deformable multimodal image registration using PET/CT and TRUS for intraoperative focal prostate brachytherapy. In Baowei Fei and Cristian A. Linte, editors, *Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 10951, pages 383 – 388. International Society for Optics and Photonics, SPIE, 2019.

[127] Hao Tang, Chupeng Zhang, and Xiaohui Xie. Nodulenet: Decoupled false positive reduction for pulmonary nodule detection and segmentation. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, 2019.

[128] Christine Tanner, Firat Ozdemir, Romy Profanter, Valeriy Vishnevsky, Ender Konukoglu, and Orcun Goksel. Generative adversarial networks for mr-ct deformable image registration. *arXiv preprint arXiv:1807.07349*, 2018.

[129] A. Teramoto, H. Fujita, K. Takahashi, O. Yamamuro, T. Tamaki, M. Nishio, and T. Kobayashi. Hybrid method for the detection of pulmonary nodules

using positron emission tomography/computed tomography: a preliminary study. *Int J Comput Assist Radiol Surg*, 9(1):59–69, Jan 2014.

[130] A. Teramoto, H. Fujita, O. Yamamuro, and T. Tamaki. Automated detection of pulmonary nodules in PET/CT images: Ensemble false-positive reduction using a convolutional neural network technique. *Med Phys*, 43(6):2821–2827, Jun 2016.

[131] Vanya V Valindria, Ioannis Lavdas, Wenjia Bai, Konstantinos Kamnitsas, Eric O Aboagye, Andrea G Rockall, Daniel Rueckert, and Ben Glocker. Reverse classification accuracy: predicting segmentation performance in the absence of ground truth. *IEEE transactions on medical imaging*, 36(8):1597–1606, 2017.

[132] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J. C. Fillion-Robin, S. Pieper, and H. J. W. L. Aerts. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res*, 77(21):e104–e107, 11 2017.

[133] Annegreet van Opbroek, M Arfan Ikram, Meike W Vernooij, and Marleen de Bruijne. Transfer learning improves supervised image segmentation across imaging protocols. *IEEE Trans Med Imaging*, 34(5):1018–1030, Nov. 2014.

[134] Yan Wang, Xu Wei, Fengze Liu, Jieneng Chen, Yuyin Zhou, Wei Shen, Elliot K Fishman, and Alan L Yuille. Deep distance transform for tubular structure segmentation in ct scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3833–3842, 2020.

[135] Yan Wang, Yuyin Zhou, Peng Tang, Wei Shen, Elliot K Fishman, and Alan L Yuille. Training multi-organ segmentation networks with sample selection by relaxed upper confident bound. *MICCAI*, 2018.

[136] Dongming Wei, Sahar Ahmad, Jiayu Huo, Wen Peng, Yunhao Ge, Zhong Xue, Pew-Thian Yap, Wentao Li, Dinggang Shen, and Qian Wang. Synthesis and inpainting-based mr-ct registration for image-guided thermal ablation of liver tumors. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 512–520. Springer, 2019.

[137] Roger P. Woods. Handbook of medical image processing and analysis. 2009.

[138] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, William T Freeman, and Joshua B Tenenbaum. MarrNet: 3D Shape Reconstruction via 2.5D Sketches. In *NeurIPS*, 2017.

[139] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.

[140] Yingda Xia, Fengze Liu, Dong Yang, Jinzheng Cai, Lequan Yu, Zhuotun Zhu, Daguang Xu, Alan L. Yuille, and Holger Roth. 3d semi-supervised learning with uncertainty-aware multi-view co-training. *arXiv*.

[141] Yingda Xia, Lingxi Xie, Fengze Liu, Zhuotun Zhu, Elliot K. Fishman, and Alan L. Yuille. Bridging the gap between 2d and 3d organ segmentation with volumetric fusion net. In *MICCAI*.

[142] Yingda Xia, Qihang Yu, Wei Shen, Yuyin Zhou, Elliot K. Fishman, and Alan L. Yuille. Detecting pancreatic ductal adenocarcinoma in multi-phase ct scans via alignment ensemble. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 285–295, 2020.

[143] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection, 10 2017.

[144] Z. Xu, C. P. Lee, M. P. Heinrich, M. Modat, D. Rueckert, S. Ourselin, R. G. Abramson, and B. A. Landman. Evaluation of six registration methods for the human abdomen on clinically acquired ct. *IEEE Transactions on Biomedical Engineering*, 63(8):1563–1572, 2016.

[145] Zhenlin Xu and Marc Niethammer. Deepatlas: Joint semi-supervised learning of image registration and segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 420–429. Springer, 2019.

[146] Ke Yan, Jinzheng Cai, Dakai Jin, Shun Miao, Adam P Harrison, Dazhou Guo, Youbao Tang, Jing Xiao, Jingjing Lu, and Le Lu. Self-supervised learning of pixel-wise anatomical embeddings in radiological images, 2020.

[147] Ke Yan, Xiaosong Wang, Le Lu, and Ronald M Summers. DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *J Med Imaging (Bellingham)*, 5(3):036501, July 2018.

[148] Xiaofeng Yang, Hamed Akbari, Luma Halig, and Baowei Fei. 3d non-rigid registration using surface and local salient features for transrectal ultrasound image-guided prostate biopsy. *Proceedings of SPIE–the International Society for Optical Engineering*, 7964:79642V–79642V, 2011.

[149] Q. Yu, L. Xie, Y. Wang, Y. Zhou, E. K. Fishman, and A. L. Yuille. Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation. *arXiv:1709.04518*, 2017.

[150] Qihang Yu, Lingxi Xie, Yan Wang, Yuyin Zhou, Elliot K. Fishman, and Alan L. Yuille. Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation. *CVPR*, 2018.

[151] L. Zhang, L. Lu, R.M. Summers, E. Kebebew, and J. Yao. Personalized pancreatic tumor growth prediction via group learning. In *MICCAI*, 2017.

[152] Shaoting Zhang, Yiqiang Zhan, Maneesh Dewan, Junzhou Huang, Dimitris N Metaxas, and Xiang Sean Zhou. Towards robust and effective shape modeling: sparse shape composition. *Med Image Anal*, 16(1):265–277, Sept. 2011.

[153] Zizhao Zhang, Lin Yang, and Yefeng Zheng. Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern Recognition*, 2018.

[154] S. Zhao, Y. Dong, E. Chang, and Y. Xu. Recursive cascaded networks for unsupervised medical image registration. In *2019 IEEE/CVF International*

Conference on Computer Vision (ICCV), pages 10599–10609, 2019.

[155] Shengyu Zhao, Yue Dong, Eric I-Chao Chang, and Yan Xu. Recursive cascaded networks for unsupervised medical image registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[156] Wan-Jun Zhao, Lin-Ru Fu, Zhi-Mian Huang, Jing-Qiang Zhu, and Bu-Yun Ma. Effectiveness evaluation of computer-aided diagnosis system for the diagnosis of thyroid nodules on ultrasound: A systematic review and meta-analysis. *Medicine*, 98(32), 2019.

[157] Huangjie Zheng, Lingxi Xie, Tianwei Ni, Ya Zhang, Yan-Feng Wang, Qi Tian, Elliot K. Fishman, and Alan L. Yuille. Phase collaborative network for multi-phase medical imaging segmentation. *ArXiv*, abs/1811.11814, 2018.

[158] Yuyin Zhou, Yingwei Li, Zhishuai Zhang, Yan Wang, Angtian Wang, Elliot K Fishman, Alan L Yuille, and Seyoun Park. Hyper-pairing network for multi-phase pancreatic ductal adenocarcinoma segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 155–163. Springer, 2019.

[159] Y. Zhou, Y. Wang, P. Tang, S. Bai, W. Shen, E. Fishman, and A. Yuille. Semi-supervised 3d abdominal multi-organ segmentation via deep multi-planar co-training. In *WACV*, 2019.

[160] Y. Zhou, L. Xie, E. K. Fishman, and A. L. Yuille. Deep supervision for pancreatic cyst segmentation in abdominal ct scans. In *MICCAI*, 2017.

[161] Yuyin Zhou, Lingxi Xie, Elliot K. Fishman, and Alan L. Yuille. Deep supervision for pancreatic cyst segmentation in abdominal ct scans. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*, 2017.

[162] Y. Zhou, L. Xie, W. Shen, Y. Wang, E. K. Fishman, and A. L. Yuille. A fixed-point model for pancreas segmentation in abdominal ct scans. In *MICCAI*, 2017.

[163] Zhuotun Zhu, Yongyi Lu, Wei Shen, Elliot K. Fishman, and Alan L. Yuille. Segmentation for classification of screening pancreatic neuroendocrine tumors, 2020.

[164] Zhuotun Zhu, Yongyi Lu, Wei Shen, Elliot K Fishman, and Alan L Yuille. Segmentation for classification of screening pancreatic neuroendocrine tumors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3402–3408, 2021.

[165] Zhuotun Zhu, Xinggang Wang, Song Bai, Cong Yao, and Xiang Bai. Deep learning representation using autoencoder for 3d shape retrieval. *Proceedings 2014 IEEE International Conference on Security, Pattern Analysis, and Cybernetics, SPAC 2014*, 09 2014.

[166] Zhuotun Zhu, Yingda Xia, Wei Shen, Elliot Fishman, and Alan Yuille. A 3d coarse-to-fine framework for volumetric medical image segmentation. In *2018 International conference on 3D vision (3DV)*, pages 682–690. IEEE, 2018.

[167] Z. Zhu, Y. Xia, W. Shen, E. Fishman, and A. Yuille. A 3d coarse-to-fine framework for volumetric medical image segmentation. In *2018 International*

Conference on 3D Vision (3DV), 2018.

[168] Zhuotun Zhu, Yingda Xia, Wei Shen, Elliot K. Fishman, and Alan L. Yuille. A 3d coarse-to-fine framework for volumetric medical image segmentation. In *3DV*, 2018.

[169] Zhuotun Zhu, Yingda Xia, Lingxi Xie, Elliot K. Fishman, and Alan L. Yuille. Multi-scale coarse-to-fine segmentation for screening pancreatic ductal adenocarcinoma. *ArXiv*, 2018.

[170] Zhuotun Zhu, Yingda Xia, Lingxi Xie, Elliot K Fishman, and Alan L Yuille. Multi-scale coarse-to-fine segmentation for screening pancreatic ductal adenocarcinoma. In *International conference on medical image computing and computer-assisted intervention*, pages 3–12. Springer, 2019.

[171] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. 2018.

Vita

Fengze Liu is completing his Ph.D. degree of Computer Science at the Johns Hopkins University, under the supervision of Bloomberg Distinguished Professor Alan L. Yuille. Fengze received his B.S. degree in mathematics and applied mathematics from Tsinghua University in 2017. Fengze's research interests lie in the fields of deep learning and computer vision, with focus on medical image analysis.