An Alarm System for Segmentation Algorithm Based on Shape Model

Fengze Liu[1], Yingda Xia[1], Dong Yang[2], Alan Yuille[1], Daguang Xu[2]
[1]Johns Hopkins University, [2]NVIDIA Corporation

Abstract

It is usually hard for a learning system to predict correctly on rare events that never occur in the training data, and there is no exception for segmentation algorithms. Meanwhile, manual inspection of each case to locate the failures becomes infeasible due to the trend of large data scale and limited human resource. Therefore, we build an alarm system 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, in this paper, 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 (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 for 3D medical segmentation tasks. Compared with other standard quality assessment methods, our system consistently provides more reliable prediction on the qualities of segmentation results.

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 [25]. 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. [10] proposed the Bayesian neural network that models the uncertainty as an additional loss to 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.

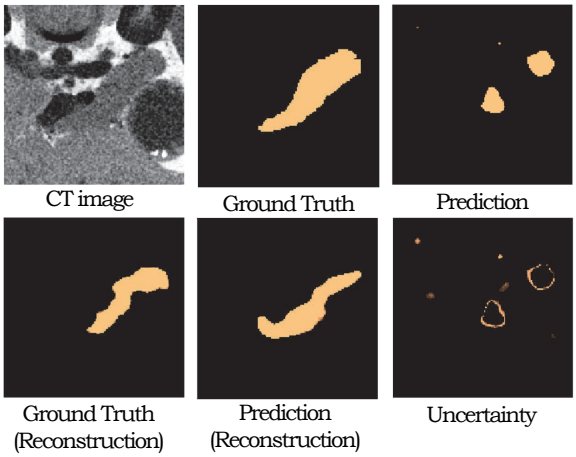


Figure 1. The visualize on an NIH CT data for pancreas segmentation. The Dice between GT and Prediction is 47.06 (real Dice) while the Dice between Prediction and Prediction(Reconstruction) from VAE is 47.25 (fake Dice). Our method uses the fake Dice to predict the former real Dice which is usually unknown 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.

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. [9] 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 [27], shown in Figure 1. [12] 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 [12] 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 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) [11] 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 2. 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. Related Work

Quality Assessment: [10] employed Bayesian neural network (BNN) to model the aleatoric and epistemic uncertainty. Afterwards, [13] applied the BNN to calculate the aleatoric and epistemic uncertainty on medical segmentation tasks. [9] 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 [24][20] 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. [4] and [7] used unsupervised methods to estimate the segmentation quality using geometrical and other features. However their application in medical settings is not clear. [12] 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, [19] tried a simple method using image-segmentation pairs to directly regress the quality. [3] used the feature from deep network for quality assessment.

Anomaly Detection: Quality assessment is also related with Out-of-Distribution (OOD) detection. Investigation related research papers can be found in [17]. Previous works in this field [8] [14] 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 us-

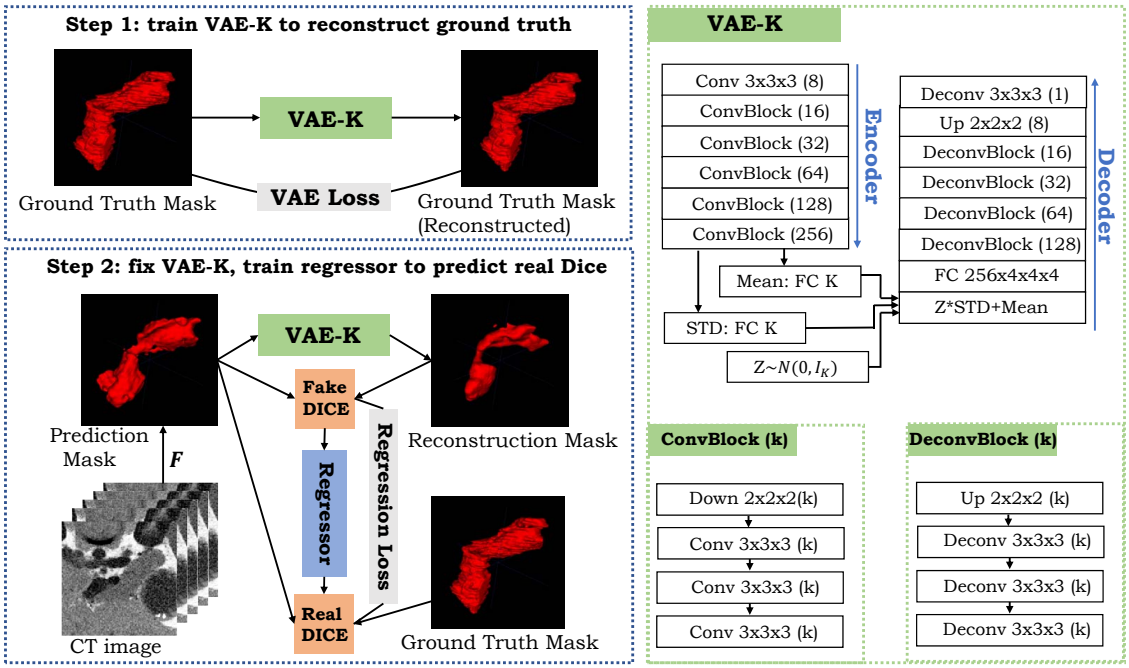


Figure 2. The architecture of our alarm system. In train step 1, the VAE is trained to reconstruct the ground truth masks. In train step 2, the parameters of VAE are fixed and a regressor is trained to predict the real Dice score. F represents a preparation segmentation algorithm which is used to generate 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.

ing 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.

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 [30], dimension reduction, etc. Unlike [26] which needs to pre-train with RBM, AE can be trained following an end-to-end fashion. [18] learned the shape representation from point cloud form, while we choose the volumetric form as a more natural way to corporate with segmentation task. [16] utilizes AE to evaluate the difference between prediction and ground truth but not in an unsupervised way. [28] explored shape features using AE. [2] utilized the reconstruction error of brain MRI image by AE and [22] used GAN for anomaly detection but it is sometimes hard to generate a realistic image *e.g.* abdominal CT scan. [23] used AE and a one-class SVM to identify anomalous regions in OCT images through unsupervised learning on healthy examples. Variational autoencoder(VAE) [11], compared with AE, adds more constraint on the latent space, which prevents from learning a trivial solution *e.g.* identity mapping. [1] 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.

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

$$\mathcal{L}(F(X), Y) = L(F, X; \omega) \tag{1}$$

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 [9] and [13], 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

$$\mathcal{L}(F(X), Y) = L(F(X); \omega) \tag{2}$$

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 [12], 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

$$\mathcal{L}(F(X), Y) = L(S(F(X); \theta); \omega) \tag{3}$$

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 $Y \in \mathcal{Y}_t$, *i.e.* 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

$$\mathcal{KL}[Q(z)||P(z|Y)] = E_{z\sim Q}[\log Q(z) - \log P(z|Y)] \tag{4}$$

$\mathcal{KL}$ is Kullback-Leibler divergence. By replacing $Q(z)$ with $Q(z|Y)$, finally it would be deduced to the core equation of VAE [6].

$$\begin{aligned} &\log P(Y) - \mathcal{KL}[Q(z|Y)||P(z|Y)] \\ &= E_{z\sim Q}[\log P(Y|z)] - \mathcal{KL}[Q(z|Y)||P(z)] \end{aligned} \tag{5}$$

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

$$\begin{aligned} &E_{Y\sim\mathcal{Y}_t}\, \mathcal{KL}[Q(z|Y)||P(z|Y)] \\ &= E_{Y\sim\mathcal{Y}_t} |\log P(Y) - S(Y; \theta)| \end{aligned} \tag{6}$$

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 [15] when training VAE, which is widely used in medical segmentation task. The final form of S is

$$\begin{aligned} S(Y; \theta) = E_{z\sim\mathcal{N}(\mu(Y),\Sigma(Y))} &\frac{2|g(z) \cdot Y|}{|Y|^2 + |g(z)|^2} \\ &- \lambda\, \mathcal{KL}[\mathcal{N}(\mu(Y), \Sigma(Y))||\mathcal{N}(0, 1)] \end{aligned} \tag{7}$$

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.

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

$$\mathcal{L}(F(X), Y) = L(S(F(X); \hat{\theta}); \omega) \tag{8}$$

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 model, so the energy function we want to optimize is

$$E(S(F(X); \hat{\theta}); a, b) = ||aS(F(X); \hat{\theta}) + b - \mathcal{L}(F(X), Y)||^2 \tag{9}$$

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. $\mathcal{L}(F(X), Y) = \frac{2|F(X)\cdot Y|^2}{|F(X)|^2 + |Y|^2}$.

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.

$$\arg\min_{a,b} \sum_{(X,Y)\in(\mathcal{X}_t,\mathcal{Y}_t)} ||aS(F(X); \hat{\theta}) + b - \mathcal{L}(F(X), Y)||^2. \tag{10}$$

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

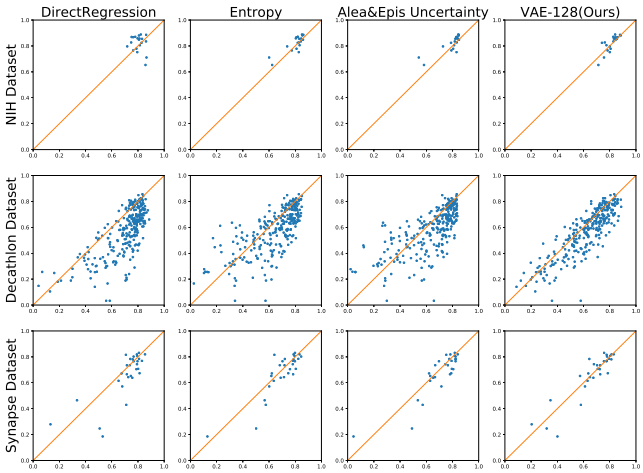


Figure 3. This figure shows our predictive Dice score (x axis) vs real Dice score (y axis). For each row, the segmentation algorithm is tested on the left most dataset. The four figures in each row show how the segmentation results are evaluated by 4 different methods.

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

$$\arg\min_{a,b} \sum_{k=1,2} \sum_{(X,Y)\in(\mathcal{X}_t^k,\mathcal{Y}_t^k)} ||aS(F_k(X);\hat{\theta}) + b - \mathcal{L}(F_k(X), Y)||^2. \quad (11)$$

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.

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.

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 $512\times512\times h$ voxels with varying voxel sizes.

- **NIH Pancreas-CT Dataset (NIH)** The NIH Clinical Center performed 82 abdominal 3D CT scans[21] from 53 male and 27 female subjects. The subjects are selected by radiologists from patients without major abdominal pathologies or pancreatic cancer lesions.
- **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 [15], 3D Coarse2Fine [29], DeepLabv3 [5], and 3D Coarse2Fine with Bayesian structure [13]. 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.

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):

- **Entropy Uncertainty**. [9] 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.

[1] http://medicaldecathlon.com/index.html
[2] https://www.synapse.org/#!Synapse:syn3193805/wiki/217789

	NIH Dataset				MSD Dataset				Synapse 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*. [9]	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*. [13]	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 1. Comparison between our method and baseline methods. The target segmentation (*i.e*. BNN) algorithm is evaluated automatically without using ground truth. We have tried different structures for VAE (*e.g*. 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).

- **Aleatoric and Epistemic Uncertainty**. [13] 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 [9] 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.

4.3. Implementation Detail

The structure of VAE is shown in Figure 2. 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.

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 re-sample 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 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.

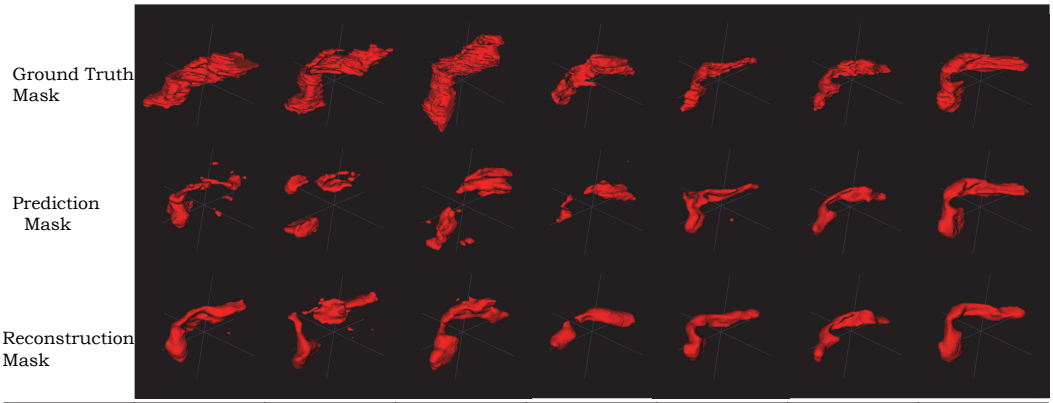
4.4. Primary Results and Discussion

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 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

	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. 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 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



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 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.

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 3. We can clearly observe that our method provides more accurate quality assessment result. For uncertainty-based methods, as shown in Figure 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 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 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 2 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 4 for showing this property of VAE. For bad segmentation predictions, the reconstruction masks from VAE indeed look more like a pancreas.

4.5. Ablation Experiments

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

Different VAE Structures: Table 1 also shows results of VAE with latent space of different dimensions. With bigger latent space, VAE can reconstruct the ground truth masks

	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*. [9]	7.24	8.79	54.38	49.29	26.57	29.78	-23.87	-20.23
Kwon *et al*. [13]	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 3. Results for evaluating both pancreas and tumor segmentation. The MAE number for pancreas is better than those in Table 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.

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.

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, *i.e*., 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 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. [22] and [2] 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.

Evaluate Object With Large Shape Variance: We also compare baseline methods and our method on evaluating segmentation of object with less stable shape *e.g*. 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 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 ($60 \rightarrow 140$ before augmentation), the number is better than that in Table 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.

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.

References

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

[2] 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.

[3] S. Bosse, D. Maniry, K. Mller, 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.

[4] 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.

[5] 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.

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

[7] 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.

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

[9] 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.

[10] 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.

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

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

[13] 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.

[14] 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.

[15] 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.

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

[17] M.A. Pimentel, D.A. Clifton, L. Clifton, and L. Tarassenko. A review of novelty detection. *Signal Processing 99 (2014) 215249*.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

[28] 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.

[29] 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.

[30] 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.