Quality Quantification in Deep Convolutional Neural Networks for Skin Lesion Segmentation using Fuzzy Uncertainty Measurement

Qiao Lin*, Xin Chen‡, Chao Chen§, Jonathan M. Garibaldi†
School of Computer Science
University of Nottingham
Email:{*qiao.lin, ‡xin.chen, §chao.chen, †jon.garibaldi,}@nottingham.ac.uk

***Abstract*—Deep convolutional neural networks (DCNN)-based methods have achieved promising performance in semantic image segmentation. However, in practical applications, it is important not only to produce the segmentation result but also to inform the segmentation quality (e.g. confidence of the segmentation result). In this paper, we propose to utilize fuzzy sets for estimating segmentation uncertainty, therefore to infer the quality of segmentation produced by a DCNN model. The proposed method combines test-time augmentation and fuzzy sets to estimate an image-level uncertainty. Six different fuzziness measures are implemented and compared, in order to select the best fuzzy uncertainty metric for the proposed method. A public skin lesion dataset is used to evaluate the method. The results show a strong correlation (Pearson correlation coefficient of 0.736) between our proposed uncertainty measure and image segmentation quality measured by Dice coefficient.**

***Index Terms*—fuzzy sets, image segmentation, quality quantification, uncertainty, skin lesion**

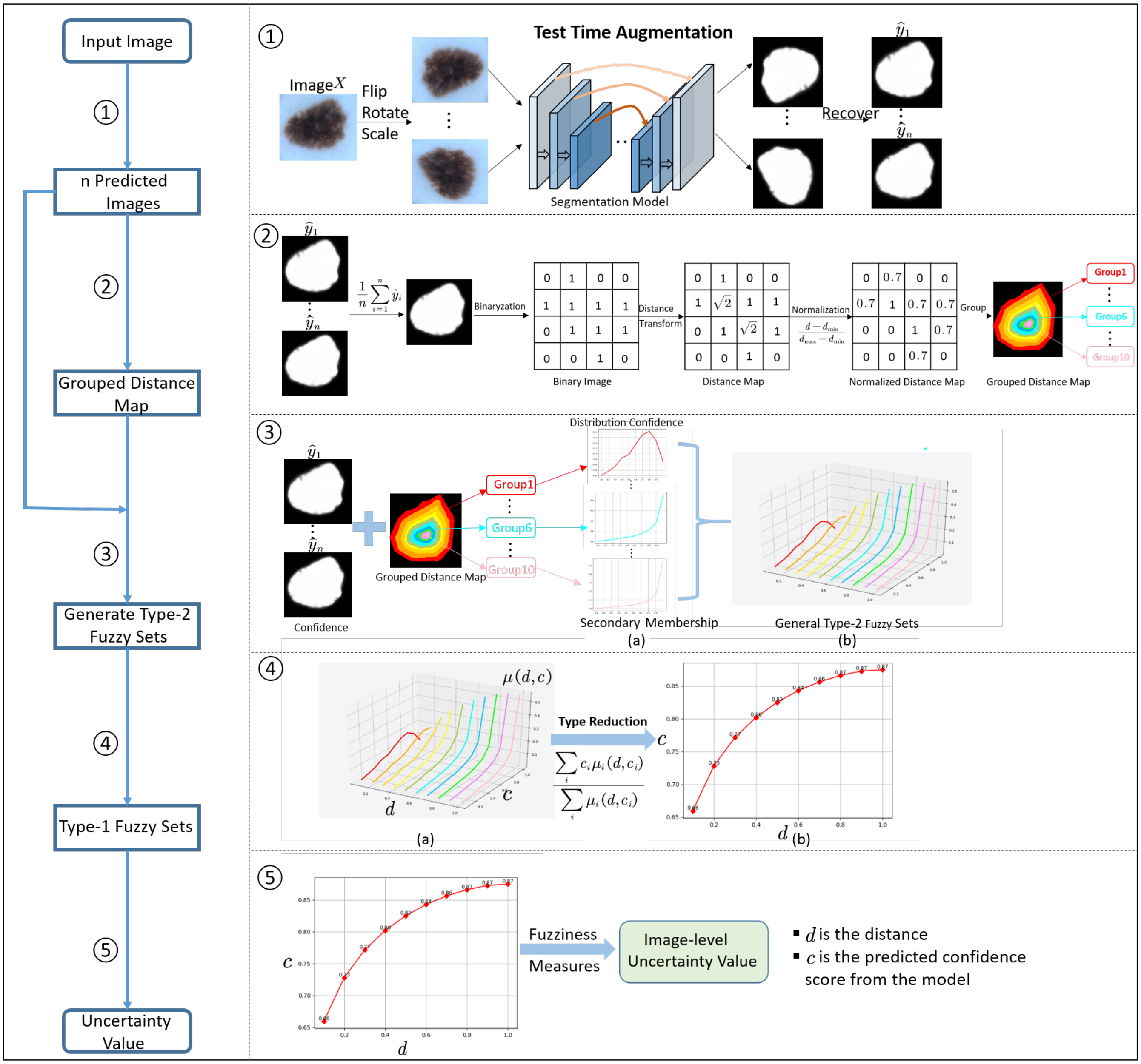
I. Introduction

Medical image segmentation has a pivotal role in medical image analysis. Big data and parallel computing promote the rapid development of medical image segmentation. In recent years, numerous deep convolutional neural networks (DCNN)-based segmentation models including FCN [1], UNet [2], DeepLab [3], have been designed to segment variety of medical image modalities (e.g. CT, MRI, etc.). For some specific diseases (e.g. skin lesion [4], lung tumor [5]), the segmentation performance of these DCNN-based models can be on a par with human experts.

Although modern image segmentation models have achieved high accuracy in numerous public datasets [6], their applications in the real world are still limited due to the fact that no reliable indication of the segmentation quality can be provided. Current image segmentation models have no ability to indicate the success/failure or the level of trustworthiness of the segmentation result. Instead, these models only provide a segmentation result without segmentation quality information, which limits the widespread application of image segmentation models especially in clinical settings. Note that the pixel-wise conference scores provided by the segmentation models are different from the uncertainty or trustworthiness of the segmentation results.

Therefore, it would be of great importance to design a quality quantification algorithm for the image segmentation models. The quality quantification algorithm should be capable of indicating whether the segmentation result has poor or good quality without knowing the ground truth segmentation. Based on the literature, there are many different types of methods to enable quality quantification. Robinson et al. [7] and DeVries et al. [8] constructed a DCNNs-based regression model, which utilized the raw image and the predicted segmentation image to directly predict the segmentation quality measured by Dice coefficient. This method needs to train a new DCNN-based regression model which is time-consuming, and the performance is highly dependent on the training dataset. Roy et al. [9] and Hoebel et al. [10] utilized the segmentation uncertainty information captured by MCdropout [11] to assess the segmentation quality. As a high uncertainty value generally indicates an incorrect prediction, the segmentation quality has a negative relationship with the segmentation uncertainty [12]. However, these methods only considered the pixel-wise uncertainty and their performance can be easily affected by the output of a given segmentation model (see the corresponding formula explanation in [9] and [10]). Different from the above mentioned methods, we propose a novel fuzzy uncertainty estimation method for image-level quality quantification, so that it is useful in selecting high quality segmentation results automatically in practical applications.

Fuzzy sets, proposed by Zadeh in 1965 [13], can efficiently handle ambiguity and vagueness in many fields. Kwak et al. [14] utilized fuzzy sets to manage the uncertainty in face images, which was beneficial for the improvement of face recognition performance. De et al. [15] studied the uncertainty in medical diagnosis based on the intuitionistic fuzzy set theory. Wang et al. [16] adopted fuzzy sets and fuzzy logic to deal with the uncertainty in text and generate interpretable results of the public sentiments analysis model. However, no research investigates the application of fuzzy sets in the evaluation of image segmentation uncertainty. In this paper, fuzzy sets are applied to represent the segmentation results and then the segmentation uncertainty is calculated by evaluating the fuzziness (fuzzy uncertainty) of the given fuzzy sets. A high uncertainty value indicates poor-quality segmentation and



0	1	0	0
1	1	1	1
0	1	1	1
0	0	1	0

0	1	0	0
1	$\sqrt{2}$	1	1
0	1	$\sqrt{2}$	1
0	0	1	0

0	0.7	0	0
0.7	1	0.7	0.7
0	0	1	0.7
0	0	0.7	0

Fig. 1. The left part is the flow chart of our proposed fuzzy-based quality quantification algorithm. The right part is the detailed procedure for each step shown in the left flow chart. First, test-time augmentation is applied to generate n predicted images. The n predicted images are then combined with the grouped distance map calculated by a distance transform operator and the average predicted image, to obtain general type-2 fuzzy sets. Next a type reduction operator is applied to convert type-2 fuzzy sets to type-1 fuzzy sets. Finally, the fuzziness measures of type-1 fuzzy sets are used to evaluate the image-level uncertainty value.

vice versa. Experiments with six different fuzziness measures are implemented to verify the effectiveness of the proposed fuzzy-based quality quantification algorithm on a public skin lesion dataset.

The main contributions are as follows.

- This is the first time that fuzzy sets are applied to design a quality quantification algorithm in DCNN models, which can estimate the segmentation quality without using the ground truth segmentation result.
- Based on the combination of test-time augmentation and distance transform, the segmentation uncertainty can be represented by fuzzy sets and calculated by the fuzziness of the given fuzzy set.
- Six different fuzziness measures are adopted and compared to investigate the relationship between the proposed uncertainty measure and the segmentation quality measured by Dice coefficient. .

The structure of this paper is as follows. Section II introduces the background information for the quality quantification. The detailed procedure for the proposed fuzzy-based quality quantification algorithm is given in Section III. Experimental results and discussion are drawn in Section IV. Finally, conclusions are given in Section V.

II. Background

Image segmentation models can generate the segmentation results but have no ability to provide related information about segmentation quality without the help of ground truth images. To address this limitation of the segmentation model, many researchers devote themselves to the investigation of the quality quantification algorithms. The quality quantification is a auxiliary algorithm, the function of which is to evaluate the segmentation quality without the ground truth images.

Based on the literature, there are two popular classes of methods to design the quality quantification algorithm. The first class of methods is to use learning-based method (e.g. DCNN-based regression model) to measure the segmentation quality. The input of the model is the raw image and the predicted image and the output is a value that indicates the segmentation quality (e.g. Dice coefficient) [7], [8]. The second one is to utilize the segmentation uncertainty to assess the segmentation performance. As the segmentation uncertainty has a negative relationship with the segmentation quality and the calculation procedure of uncertainty is independent of the ground truth images, a low uncertainty refers to a good quality segmentation. After capturing the segmentation uncertainty by MCdropout [11], Hoebel et al. [10] adopted the overlap between pairwise predicted images to measure the uncertainty; Roy et al. [9] utilized the intersection over overlap (IOU) to calculate the uncertainty; and Roy et al. [17] proposed region-wise variation coefficient (VC) to measure the uncertainty. In this paper, our proposed method belongs to the second class of methods, which utilizes fuzzy sets to estimate image-level uncertainty for segmentation quality quantification.

III. fuzzy-based quality quantification algorithm

An overview of the proposed method is shown in Fig. 1. It is assumed that a DCNN-based segmentation model is trained (e.g. Unet [2]), which is capable of performing segmentation on a given input image. Our method works as a computational block in the model inference process, which consists of the following steps. (1) test-time augmentation [18] is firstly applied to generate n predicted segmentation outputs. (2) An average predicted segmentation image is calculated based on the *n* predictions. A distance map is generated using distance transform [19]. The distances are then normalized and discretized into groups, results in a grouped distance map. (3) The *n* predicted segmentation images are used to generate a set of confidence distributions for each of the distance groups, which is then formalized as general type-2 fuzzy sets. (4) Type reduction is applied to convert the type-2 fuzzy sets to type-1 fuzzy sets. (5) Fuzziness measures are subsequently applied to the type-1 fuzzy sets to calculate an image-level uncertainty value, which is used as a surrogate for segmentation quality quantification. The detailed process of each step is given as follows.

A. Segmentation Model

In this paper, a widely used DCNN-based semantic segmentation model (i.e. Unet [2]) is used to perform the task of image segmentation. Compared with other segmentation models, Unet is more suitable for medical image segmentation since it utilizes learned features in a multi-resolution manner. This model is comprised of image/feature map downsampling and upsampling processes. The aim of the downsampling process is to capture multi-resolution features, while the upsampling process (also called deconvolution) is primarily applied to resize the feature maps in order to make the predicted image and the ground truth image have consistent sizes. Skip-connection is the special characteristic of this model, which is applied to transmit learned deep features from the downsampling process to the upsampling process to enable a more effective feature learning. The detailed model structure is given in section IV-B. Note that our method can be applied to other DCNN based models, as long as test-time augmentation (section III-B) can be applied. .

B. Test-time Augmentation

Several segmentation uncertainty estimation methods can be applied to the Unet model. Herein, we focus on estimating one of most acknowledged types of uncertainty in DCNN models, known as data uncertainty. Data uncertainty is generally caused by the process of image acquisition and cannot be eliminated with the increasing amount of data. Inspired by Wang's research [18], test-time augmentation (TTA) shown in Fig. 1–① is adopted to handle the data uncertainty.

Given the true image X_0 and the observed image X, the image acquisition model is represented as: $X = \Theta_\alpha(X_0)$, where Θ is a transformation operator (e.g. translation, rotation, scaling, and flipping), and α depicts the parameters of the given transformation operator.

The segmentation output Y can be $f(X,\omega)$, where $f(.)$ refers to the learned mapping function by DCNN models, and ω is the model parameters including weights, bias and other hyper-parameters.

Then we have

$$Y = \Theta_\alpha(Y_0) = \Theta_\alpha(f(X_0,\omega)) = \Theta_\alpha\left(f\left(\Theta_\alpha^{-1}(X),\omega\right)\right) \tag{1}$$

Y and Y_0 are the segmentation images for X and X_0 respectively. Intuitively, equation (1) indicates that to obtain the segmentation output Y of the input image X, the X can be firstly inversely transformed to the X_0, then input to a segmentation model for producing a segmentation output Y_0, followed by a transformation from Y_0 to Y.

Next the posterior probability of Y given X is calculated as $p(Y|X) = p\left(\Theta_\alpha\left(f\left(\Theta_\alpha^{-1}(X),\omega\right)\right)\right)$, where the precise value of α is unknown but their prior distribution is informed that is $\alpha \sim \mathbb{U}(\alpha)$. Therefore, the final predicted value is

$$\hat{Y} = E(Y|X) = \int \Theta_\alpha\left(f\left(\Theta_\alpha^{-1}(X),\omega\right)\right)\mathbb{U}(\alpha)\,d\alpha. \tag{2}$$

The integral computation is time consuming and extremely complicated. In practical application, Monte Carlo simulation

TABLE I
FUZZINESS MEASURES

Measures	Formulas
F_{Ya} [22]	$1-\frac{\left[\sum_{i=1}^{N}\left\|2\mu_A(x_i)-1\right\|^2\right]^{\frac{1}{2}}}{N^{\frac{1}{2}}}$
F_{Dt} [23]	$-\frac{1}{N}\sum_{i=1}^{N}\mu_A(x_i)Log_2(\mu_A(x_i))+(1-\mu_A(x_i))Log_2(1-\mu_A(x_i))$
F_{Be} [24]	$1-\left(\frac{1}{N}\sum_{i=1}^{N}\mu_A(x_i)^2\right)$
F_{Ka} [25]	$2\sqrt{\frac{\sum_{i=1}^{N}\left\|\mu_A(x_i)-\mu_{A_{near}}(x_i)\right\|^2}{N}}$, where $\mu_{A_{near}}(x)=\begin{cases}1,\mu_A\geqslant 0.5\\0,\mu_A<0.5\end{cases}$
F_{Ko} [26]	$\sqrt{\frac{\sum_{i=1}^{N}\left\|\mu_A(x_i)-\mu_{A_{near}}(x_i)\right\|^2}{\sum_{i=1}^{N}\left\|\mu_A(x_i)-\mu_{A_{far}}(x_i)\right\|^2}}$ where $\mu_{A_{near}}(x)=\begin{cases}1,\mu_A\geqslant 0.5\\0,\mu_A<0.5\end{cases}$ and $\mu_{A_{far}}(x)=\begin{cases}0,\mu_A\geqslant 0.5\\1,\mu_A<0.5\end{cases}$
F_{Bp} [28]	$-\frac{1}{N}\sum_{i=1}^{N}Log_2\left(\mu_A(x_i)^2+(1-\mu_A(x_i))^2\right)$

method is utilized to assess $E(Y|X)$. We obtain the set $[\alpha_0,\alpha_1,\alpha_2,\cdots,\alpha_n]$ which is sampled from the prior distribution. For each parameter α_i, the semantic segmentation model is able to generate one predicted image $\hat{y}_i$ based on equation (2). Hence, the predicted image set $[\hat{y}_0,\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_n]$is equal to the values sampled from $p(Y|X)$. In another word, $\hat{Y}$ is estimated by aggregating the segmentation outputs of n transformed versions of X. Then the n predicted images $[\hat{y}_1,\hat{y}_1,\hat{y}_3,\cdots,\hat{y}_n]$ are applied to evaluate the segmentation uncertainty in the next step of our method.

C. Calculation of the Grouped Distance Map

TTA captures the segmentation uncertainty by generating n predicted images $[\hat{y}_1,\hat{y}_1,\hat{y}_3,\cdots,\hat{y}_n]$. Then, a grouped distance map (shown in Fig. 1–②) is calculated to help divide the segmented object region into sub-groups according to their distance to the object boundary. The reason of choosing the pixel distance from the segmentation boundary to group the pixels is that the segmentation results normally have higher uncertainty for the pixels closer to the boundary and vice versa. Hence, it is sensible to represent the predicted pixel-wise confidence values as a function of their distance to the object boundary. The detailed calculation procedure of the grouped distance map is given in the following:

(1) Given an input image X, the predicted images are $[\hat{y}_1,\hat{y}_2,\hat{y}_3,\cdots,\hat{y}_n]$, where n is the number of TTA. The average of the predicted images is calculated as $\hat{Y}=\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i$.

(2) The binary image Υ of $\hat{Y}$ is then calculated using the function $B(p)=\begin{cases}1,p\geqslant 0.5\\0,p<0.5\end{cases}$, where p is the pixel value of $\hat{Y}$.

(3) The Euclidean distance transform algorithm [19] is then applied to the binary image Υ, resulting in a distance map $\mathscr{D}$. Each pixel value in the distance map means the minimum distance to the segmented object boundary.

(4) To ensure consistent measurements of small and large objects, the distance values in the distance map is normalized to the interval of $[0,1]$ using Min-Max scaling algorithm $d_{norm}=\frac{d-d_{\min}}{d_{\max}-d_{\min}}$, where d is the pixel value in the distance map $\mathscr{D}$.

(5) Finally, a grouped distance map is obtained by dividing the distance into 10 evenly distributed sub-groups, as shown in Fig. 1–②.

D. Fuzzy-Sets-based Uncertainty Estimation

Having the grouped distance map generated, the next step is to calculate the image-level uncertainty to quantify the segmentation quality. It is known that the degree of uncertainty can be calculated based on the fuzziness of given type-1 fuzzy sets [21]. In this section, a novel method is proposed to generate fuzzy sets based on predicted images and the grouped distance map obtained in the previous sections. The method is described in detail below and the flow-chart can be seen in Fig. 1.

It is observed that the predicted pixel-wise confidence values of a segmentation model are positively correlated to their distance to the predicted object boundary. In other words, the further away a pixel is from the object boundary the more confident the prediction is, as illustrated in Fig. 1–⑤. This curve can be considered as a type-1 fuzzy set.

On the other hand, as we have obtained the predicted images and the grouped distance map, it is possible to get such a type-1 fuzzy set described above by following the steps below.

As illustrated in Step ③ of Fig. 1, all pixels in $[\hat{y}_1,\hat{y}_1,\hat{y}_3,\cdots,\hat{y}_n]$ are divided into 10 groups based on the obtained grouped distance map (section III-C). It is known that the predicted pixel value indicates the confidence level of the pixel belonging to the target class. Therefore, for each distance group, distribution of confidence values (see Fig. 1–③ (a)) for all pixels in that group is obtained. Hence, by combining the distributions of all ten groups, a 3D distribution is then generated as shown in Fig. 1–③ (b). We treat this 3D distribution as a type-2 fuzzy set. The primary variable (x-axis) is the distance from the pixel to the segmentation boundary, and the secondary variable (y-axis) is the confidence value to represent whether this pixel belongs to the target segmentation class. The distribution of confidence values for each group can be considered as the secondary membership function.

In the next step, to get the type-1 fuzzy set, a type reduction method is applied to the type-2 fuzzy set. An efficient method, known as the centroid method (i.e. weighted average in formula (3)), is used to convert the distribution (the secondary membership function) of each group to a single confidence level (see Fig. 1–④).

$$\frac{\sum_i c_i \mu_i(d, c_i)}{\sum_i \mu_i(d, c_i)} \tag{3}$$

Note that this is not a standard type reduction method for type-2 fuzzy sets. However, it can be considered as an extension (or a variation) of the Nie-Tan type reduction operator [20], which is for interval type-2 fuzzy sets. When the centroid method is applied to an interval type-2 fuzzy set (where all secondary membership degrees are equal to 1) in the way we used above, the type-reduced results will be the same as that based on the Nie-Tan operator. Hence, in this paper, we call this centroid method as a type reduction method. After applying the type reduction, a type-1 fuzzy set (as illustrated in Fig. 1–④ (b)) is obtained.

The next step is to calculate the uncertainty level based on the fuzziness of the type-1 fuzzy set (Step ⑤ in Fig. 1). There are many studies about the fuzziness measures [22]–[28]. In this paper, six commonly used fuzziness measures are chosen and summarised in Table I. F_{Ya}, F_{Ka} and F_{Ko} take the distance from the given membership value to a chosen threshold value into consideration, while F_{Dt}, F_{Bp} and F_{Be} only consider the membership value itself. It is noted that the threshold value in F_{Ya}, F_{Ka} and F_{Ko} is generally specified by the highest fuzziness level [21].

Finally, an image-level uncertainty score is obtained from the fuzziness measure, which is used as a surrogate for segmentation quality quantification. The performance of these six fuzziness measures in our proposed method is investigated in next section.

IV. Evaluation

In this section, the performance of the six aforementioned fuzziness measures in our proposed method is discussed. Firstly the experiments of evaluating the correlation between the segmentation uncertainty and the segmentation quality with various fuzziness metrics were conducted. Then by setting a threshold, the ability of detecting good-quality segmentation images for the six different fuzziness measures were investigated. The detail for the dataset, experimental methods and experimental results are presented in the following subsections.

A. Dataset

A public skin lesion dataset is used to evaluate the performance of our proposed fuzzy-based quality quantification method. This dataset is from a grand challenge (ISIC2018), which includes 2594 dermoscopic lesion images with corresponding ground truth segmentation labels. Compared to other medical image segmentation datasets, the skin lesion dataset has a wide variety of shapes, colours, sizes, textures, and lesion boundaries. Therefore, it was selected to evaluate our method.

B. Experimental Methods

In the experiments, we compared the performance of six commonly used fuzziness measures as discussed in III-D (Table I) for segmentation uncertainty estimation. Pearson correlation coefficient [30] is used to measure the relationship between the estimated uncertainty value and the image segmentation quality measured by Dice coefficient. The Pearson correlation coefficient measures the linear correlation of two variables and the value is between -1 and 1. When the value is close to 1 or -1, it means the given two variables have a strong positive relationship or a strong negative relationship. When the value is close to 0, it means the given two variables are almost unrelated. The formula of Pearson correlation coefficient is given as

$$\rho_{X,Y} = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sqrt{\mathbb{E}[X^2] - (\mathbb{E}[X])^2}\sqrt{\mathbb{E}[Y^2] - (\mathbb{E}[Y])^2}}, \tag{4}$$

where X and Y refers to uncertainty values and Dice values respectively, $\mathbb{E}$ means the expectation. It is noted that the proposed fuzzy uncertainty measure has a negative relationship with the segmentation quality. The Dice coefficient is a widely used metric to measure the segmentation quality, and its formula is $Dice = \frac{2|A \cap B|}{|A|+|B|}$, where A and B refer to the foreground areas of the predicted image and the ground truth image respectively. Therefore, when the Pearson correlation coefficient between the fuzzy uncertainty and Dice is closer to -1, it means the given fuzzy uncertainty algorithm has a better ability to evaluate the segmentation performance and the corresponding fuzziness measure is better.

Five-fold cross validation is applied to obtain the final experimental results. The training data was used to train the Unet semantic segmentation model. This Unet model is an encoder-decoder framework. The encoding stage consists of 5 layers. Each layer contains two convolutional blocks with 3 × 3 convolutional filters. Max-pooling operator is used to down-sampling the feature maps between each layer. The decoding stage consists of 4 layers. Each layer also contains two convolutional blocks with 3 × 3 convolutional filters. Deconvolutional operator with 2 × 2 deconvolutional filter is used to up-sampling the feature maps between each layer. During the training process, cross entropy loss and adam optimization algorithm were applied to update the parameters in the Unet model. The initial learning rate was 10^{-4}. After obtaining the pre-trained Unet model, TTA was utilized to capture the segmentation uncertaint by using rotation, scaling, and flipping transformation operators. In this way, 24 predicted images $[\hat{y}_1, \hat{y}_2, \hat{y}_3, \cdots, \hat{y}_{24}]$ for one given input image were generated. Next type-1 fuzzy set was generated to describe the 24 predicted images based on the procedure given in section III-D. Finally, the image-level segmentation uncertainty was calculated by the fuzziness measure of the type-1 fuzzy sets shown in Table I.

Furthermore, we investigated the feasibility of using the estimated uncertainty value for image segmentation quality control. In practice, the ground truth segmentation label is not

TABLE II

EXPERIMENTAL RESULTS FOR OUR PROPOSED FUZZY-BASED QUALITY QUANTIFICATION ALGORITHM USING SIX DIFFERENT FUZZINESS MEASURES ON SKIN LESION DATASET. MEAN AND STANDARD DEVIATION VALUES ARE REPORTED FOR PEARSON CORRELATION COEFFICIENT MEASURED IN 5-FOLD CROSS VALIDATION. $*$ MEANS THE GIVEN METHOD AND F_{Ya} ARE SIGNIFICANTLY DIFFERENT WITH P VALUE <0.01

Method	Mean	Standard Deviation
F_{Ya}	-0.736	0.027
$F_{Dt}*$	-0.689	0.021
$F_{Be}*$	-0.721	0.025
$F_{Ka}*$	-0.716	0.029
$F_{Ko}*$	-0.677	0.026
$F_{Bp}*$	-0.719	0.024

provided for a given unseen image to be segmented. In this case, it is desirable to inform the user, if the predicted segmentation result is in good quality or not. In this experiment, we explored if a threshold can be applied to the uncertainty value, so that the predicted segmentation can be classified into good and poor quality.

For this lesion segmentation dataset, based on the results reported in the literature [29], Dice score of greater than 0.8 was considered to be good quality and vice versa. Hence by setting a threshold for the uncertainty value, we compared the number of good quality images and poor quality images produced by the six fuzziness measures.

All the experiments were implemented using PyTorch and trained on a workstation with NVIDIA GeForce GTX1080Ti GPU and i7-3820 CPU.

C. Experimental Results

Table II shows the experimental results for the Pearson correlation coefficient between Dice and our proposed uncertainty measure. F_{Ya}, F_{Dt}, F_{Be}, F_{Ka}, F_{Ko}, F_{Bp} means six different fuzziness measures and their formulas are given in Table I. In Table II, the Pearson correlation coefficient between the F_{Ya} measure and Dice is -0.736, which performs the best compared with other fuzziness measures. The result means that F_{Ya} metric has the strongest linear relationship with the segmentation quality. Wilcoxon sign rank test results show that there is a statistically significant difference between F_{Ya} fuzziness measured values and other fuzziness measured values with $P < 0.01$.

In order to use the measured uncertainty values to infer segmentation quality (good or poor) using the proposed thresholding method (described in section IV-B), a threshold needs to be determined first. Mathematically, all six methods should produce the uncertainty value in the range of [0,1]. Hence we use all the data points from all six methods to determine the threshold. The Dice and uncertainty values for the test images using all six fuzziness measures are plotted in Fig. 2. Yellow dots means that their Dice values are greater than 0.8 while the green dots means that their Dice values are less than 0.8. Red dotted lines are different uncertainty threshold values. In Fig. 2, each red dotted line and the sky-blue dotted line divide all dots into four groups. By maximising the number of yellow dots in group ① and minimising the number of green dots in group ②, the optimal uncertainty threshold is determined. In this application, the uncertainty threshold is set as 0.2.

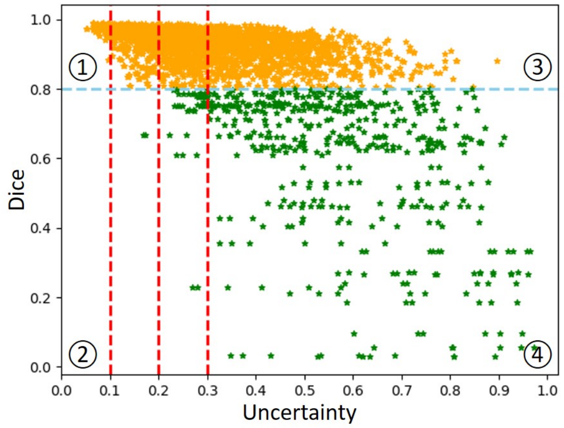


Fig. 2. Choosing the optimal uncertainty threshold: red dotted lines mean different uncertainty threshold values. Yellow dots mean their Dice> 0.8, while green dots mean their Dice≤ 0.8.

Next, the performance of six fuzziness measures with the given uncertainty threshold 0.2 is explored. Fig. 3 shows the distribution of the Dice for six different fuzziness metrics. In Fig. 3 (a), it is seen that almost all the test images with a fuzzy uncertainty value less than 0.2 have good-quality segmentation (Dice>0.8). Moreover, F_{Ya} and F_{Be} fuzziness metrics are capable of detecting more good-quality segmentation images compared with other fuzzy uncertainty metrics. In Fig. 3 (b), when the fuzzy uncertainty is greater than or equal to 0.2, there are still many more good-quality segmentation images (Dice>0.8) for F_{Dt}, F_{Ka}, F_{Ko}, F_{Bp} measures than the F_{Ya} and F_{Be} measures, which also indicates F_{Ya} and F_{Be} are the best two methods. This is consistent with the conclusion drawn from Table II that F_{Ya} and F_{Be} measures have the strongest linear correlations with Dice values.

Furthermore, to visualize the relationship between the segmentation performance index and the six uncertainty measures, the scatter plot of individual method is shown in Fig. 4. The x-axis refers to the specific fuzziness metric and the y-axis represents the Dice values. Each point in the scatter plot represents one test image. Yellow dots mean their uncertainty values are less than 0.2. Green dots mean their uncertainty values are greater than or equal to 0.2. The red dotted line indicates the threshold of 0.2. The grey line is the best fitted straight line (BFSL) for all test images. All dots in Fig. 4 (a) are evenly distributed on both sides of the BFSL, which means that there's a reasonably high linear correlation between the F_{Ya} measured uncertainty and the segmentation performance index (Dice). Moreover, when the uncertainty value is less than the threshold 0.2, there are more yellow dots in Fig. 4 (a) and Fig. 4 (c) than other four scatter plots, and almost all yellow dots have a high Dice value. It further verifies that

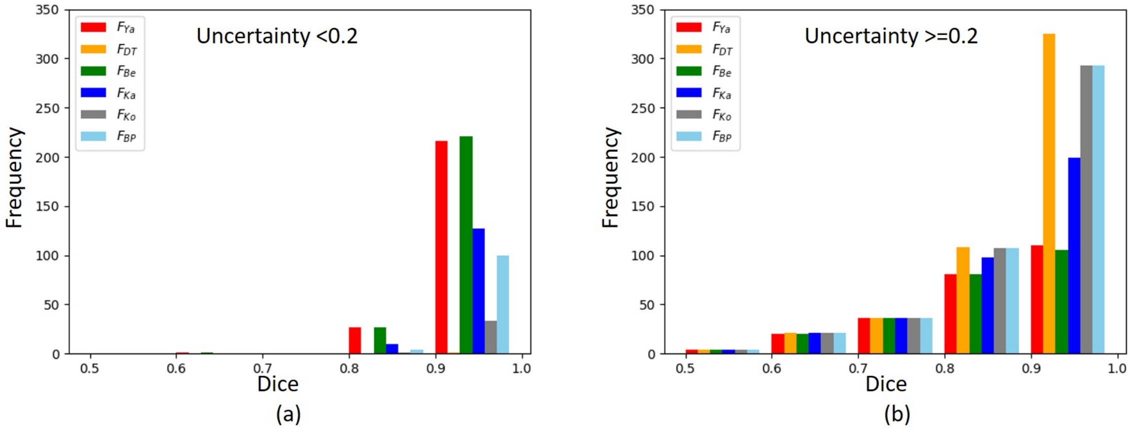


Fig. 3. (a) The distribution of Dice when the fuzzy uncertainty is less than 0.2. (b) The distribution of Dice when the fuzzy uncertainty is greater than or equal to 0.2.

the abilities of detecting good quality segmentation images using F_{Ya} and F_{Be} fuzziness metrics are better than other four fuzziness metrics.

Therefore, based on the result in Table II and Fig. 4, we conclude that F_{Ya} is the best fuzziness measure for our proposed fuzzy-based quality quantification algorithm compared with other fuzziness metrics.

V. Conclusions

In this paper, we have proposed a novel quality quantification method based on the TTA and fuzzy sets. TTA is implemented to capture the segmentation uncertainty by generating n predicted images. Then the distance transform algorithm is applied to capture each pixel's distance to the closest boundary. Next the n predicted images are represented by fuzzy sets with the distance information. Finally, six different fuzzy uncertainty metrics are utilized to calculate the fuzziness of the fuzzy set. The fuzziness value is used to quantify the image-level uncertainty of the predicted segmentation result, and therefore to infer the segmentation performance when there are no ground truth labels. Experimental results show that our proposed method is capable of estimating the skin lesion segmentation quality, and F_{Ya} is the best fuzziness measure compared with other five fuzziness metrics.

In future work, we will compare our fuzzy based method to other quality quantification methods on more datasets. Furthermore, the number of groups in grouped distance map and the number of predicted images generated by TTA are fixed, and the influence of these parameters will be investigated in our future work.

References

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proceedings of the IEEE conference oncomputer vision and pattern recognition, 2015, pp. 3431–3440.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networksfor biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.

[3] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," arXiv preprint arXiv:1412.7062, 2014

[4] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau,and S. Thrun, "Dermatologist-level classification of skin lesion with deep neural networks," nature, vol. 542, no. 7639, pp. 115–118, 2017.

[5] J. Jiang, Y. C. Hu, C. J. Liu, D. Halpenny, M. D. Hellmann, J. O. Deasy,G. Mageras, and H. Veeraraghavan, "Multiple resolution residually connected feature streams for automatic lung tumor segmentation from ct images," IEEE transactions on medical imaging, vol. 38, no. 1, pp. 134–144, 2018

[6] S. Hao, Y. Zhou, and Y. Guo, "A brief survey on semantic segmentation with deep learning," Neurocomputing, vol. 406, pp. 302–321, 2020.

[7] R. Robinson, O. Oktay, W. Bai, V. V. Valindria, M. M. Sanghvi, N. Aung,J. M. Paiva, F. Zemrak, K. Fung, E. Lukaschuk, et al., "Realtime prediction of segmentation quality," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 578–585, 2018.

[8] T. DeVries and G. W. Taylor, "Leveraging uncertainty estimates for predicting segmentation quality,"arXiv preprint arXiv:1807.00502, 2018.

[9] A. G. Roy, S. Conjeti, N. Navab, C. Wachinger, A. D. N. Initiative, et al., "Bayesian quicknat: Model uncertainty in deep whole-brain segmentation for structure-wise quality control," NeuroImage, vol. 195 ,pp. 11–22, 2019.

[10] K. Hoebel, V. Andrearczyk, A. Beers, J. Patel, K. Chang, A. Depeursinge, H. M̈uller, and J. Kalpathy-Cramer, "An exploration of uncertainty information for segmentation quality assessment," in Medical Imaging 2020: Image Processing, vol. 11313. International Society for Optics and Photonics, 2020.

[11] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation:Representing model uncertainty in deep learning," in international conference on machine learning. PMLR, pp. 1050–1059, 2016.

[12] A. Mehrtash, W. M. Wells, C. M. Tempany, P. Abolmaesumi, and T. Kapur, "Confidence calibration and predictive uncertainty estimation for deep medical image segmentation," IEEE transactions on medical imaging, vol. 39, no. 12, pp. 3868–3878, 2020.

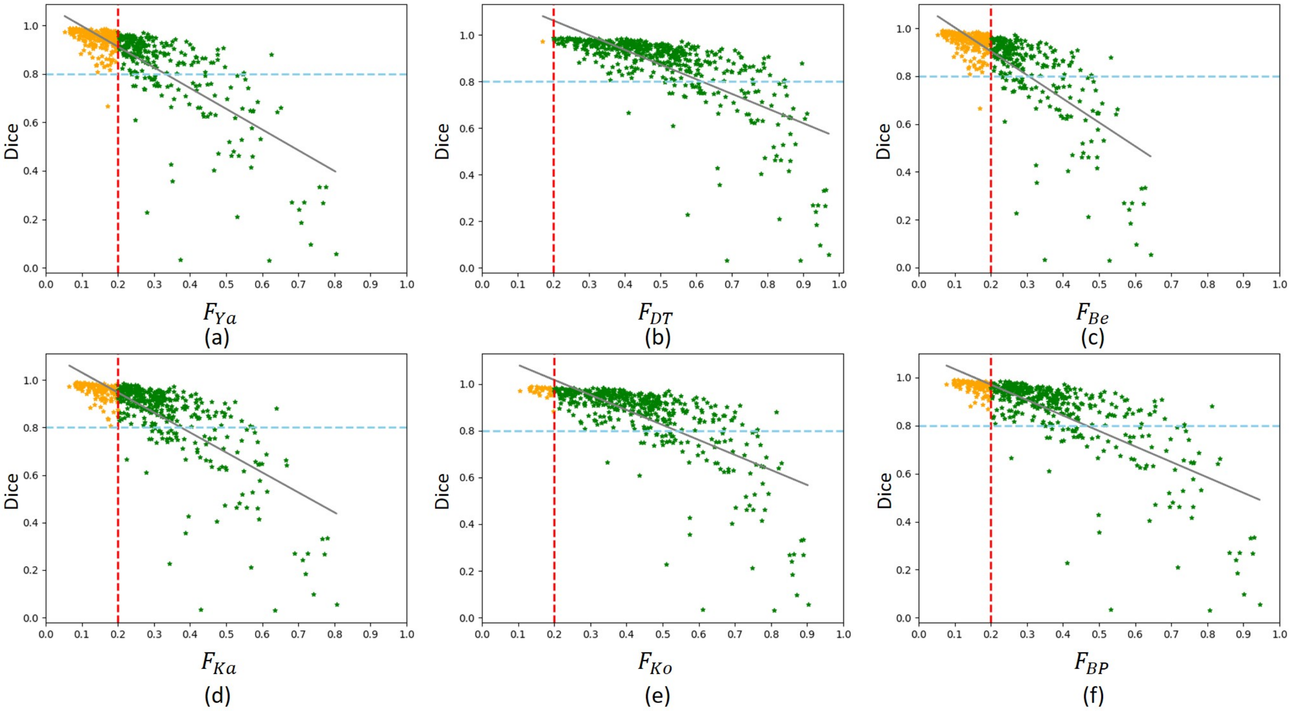
[13] L. A. Zadeh, "Fuzzy sets,"Information and control, vol. 8, no. 3, pp. 338–353, 1965.

[14] K. C. Kwak and W. Pedrycz, "Face recognition using a fuzzy fisher face classifier," Pattern recognition, vol. 38, no. 10, pp. 1717–1732, 2005

[15] S. K. De, R. Biswas, and A. R. Roy, "An application of intuitionistic fuzzy sets in medical diagnosis," Fuzzy sets and Systems, vol. 117, no. 2, pp. 209–213, 2001.

[16] X. Wang, H. Zhang, and Z. Xu, "Public sentiments analysis based on fuzzy logic for text," International Journal of Software Engineering and Knowledge Engineering, vol. 26, no. 09n10, pp. 1341–1360, 2016

[17] A. G. Roy, S. Conjeti, N. Navab, and C. Wachinger, "Inherent brain segmentation quality control from fully convnet monte carlo sampling,"



(a) (b) (c) (d) (e) (f)

Fig. 4. Scatter plot for the relationship between the segmentation performance index and our proposed quality quantification method with various fuzzy uncertainty metrics. The x-axis refers to the specific fuzzy uncertainty metric, and the y-axis means the Dice. Each point in the scatter plot represents one test image. Yellow dots mean their uncertainty values are less than 0.2. Green dots mean their uncertainty values are greater than or equal to 0.2. The red dotted line is the threshold of 0.2. The grey line is the best fitting straight line for all test images.

in International Conference on Medical Image Computing and Computer Assisted Intervention. Springer, pp. 664–672, 2018.

[18] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren, "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks," Neurocomputing, vol. 338, pp. 34–45, 2019.

[19] P. F. Felzenszwalb and D. P. Huttenlocher, "Distance transforms of sampled functions," Theory of computing, vol. 8, no. 1, pp. 415–428, 2012.

[20] M. Nie and W. W. Tan, "Towards an efficient type-reduction method for interval type-2 fuzzy logic systems," in 2008 IEEE International Conference on Fuzzy Systems (IEEE World Congress on Computational Intelligence). IEEE, 2008, pp. 1425–1432.

[21] N. R. Pal and J. C. Bezdek, "Measuring fuzzy uncertainty," IEEE Transactions on Fuzzy Systems, vol. 2, no. 2, pp. 107–118, 1994

[22] R. R. Yager, "A measurement-informational discussion of fuzzy union and intersection," International Journal of Man-Machine Studies, vol. 11, no. 2, pp. 189–200, 1979.

[23] A. De Luca and S. Termini, "A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory," Information and control, vol. 20, no. 4, pp. 301–312, 1972.

[24] J. C. Bezdek, FUZZY-MATHEMATICS IN PATTERN CLASSIFICATION. Cornell University, 1973.

[25] A. Kaufmann, Introduction to the theory of fuzzy subsets. Academic press, 1975.

[26] B. Kosko, "Fuzzy entropy and conditioning," Information sciences, vol. 40, no. 2, pp. 165–174, 1986.

[27] N. R. Pal and S. K. Pal, "Object-background segmentation using new definitions of entropy," IEE Proceedings E-Computers and Digital Techniques, vol. 136, no. 4, pp. 284–295, 1989.

[28] D. Bhandari and N. R. Pal, "Some new information measures for fuzzy sets," Information Sciences, vol. 67, no. 3, pp. 209–228, 1993.

[29] Y. Yuan, M. Chao, and Y. C. Lo, "Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance," IEEE transactions on medical imaging, vol. 36, no. 9, pp. 1876–1886, 2017.

[30] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in Noise reduction in speech processing. Springer, 2009, pp. 1–4.