Restorable Segmentation Synthesis Using Fourier Descriptors

Shuyue Guan[0000−0002−3779−9368] and Weijie Chen[0000−0001−7437−7829]

Division of Imaging, Diagnostics, and Software Reliability
Office of Science and Engineering Laboratories
Center for Devices and Radiologic Health
United States Food and Drug Administration, Silver Spring, Maryland, United States
shuyue.guan@fda.hhs.gov
weijie.chen@fda.hhs.gov

Abstract. To evaluate truthing (also known as label fusion) methods in medical image segmentation, synthetic segmentation contours can be useful especially when the reference standard is established by combining multiple segmentation results, such as those produced by multiple experts. This is because ground-truth segmentation is often unavailable in real medical images but is predefined in synthetic data. For this purpose, we developed the Restorable Segmentation Synthesis (RSS) tool. The RSS tool generates segmentation contours by modifying the Fourier descriptors of a truth contour, which, for realism, can be the contour of an anatomical structure extracted from a real medical image. The tool allows for the creation of contours with various segmentation errors relative to the ground truth. A favorable feature of our segmentation contour synthesis tool for evaluating truthing methods is that the average of a large number of synthetic contours asymptotically converge to the truth contour. This is important because such a dataset can help benchmark and compare the truthing methods. Our RSS tool is developed to have this restorability property, which we validated here through simulation studies. We further show that simulating contours is a promising approach for truthing method analysis and data augmentation for segmentation tasks.
The RSS tool with a GUI is available: https://github.com/DIDSR/RSS-tool

Keywords: Restorable segmentation synthesis · Synthetic segmentation · Medical image segmentation · Restorability · Segmentation truthing · Label fusion · Segmentation data augmentation · Generative model

1 Introduction

Despite the rapid advancements in artificial intelligence and machine learning (AI/ML) model development for medical image segmentation [6], there is a lack of consensus on the evaluation methods. Numerous metrics have been proposed in the literature, but guidelines are still needed to select the most suitable ones

for specific clinical tasks [10,11]. Various truthing (*aka* label fusion) methods, which typically establish a reference standard by combining multiple segmentations from experts [12,13], require further evaluation and comparison studies. Ground truth segmentation of medical images is an important consideration for comparing truthing methods. However, segmentation ground truth is generally not available in real medical images, but can be predefined in synthetic data.

In our previous work, we developed the Medical Image Segmentation Synthesis tool (MISS-tool) [3] to create synthetic segmentation contours based on predefined truth contours (*e.g.*, contours derived from real-world medical images). The MISS-tool allows users to customize segmentation errors through adjustable parameters. These emulated segmentation contours can be used to inform the selection of performance metrics to evaluate image segmentation methods [4]. By setting specific parameters, synthetic segmentation contours are generated from truth masks to simulate various types of possible segmentation errors. Although multiple synthetic contours can be created from a single truth mask, their average does not necessarily converge to the truth contour. A desired property for synthesizing segmentation contours to evaluate truthing methods is the ability for the average of synthetic contours to converge to the truth contour. This is favorable because such a dataset can act as a benchmark to assess truthing methods - for example, to assess if a truthing method is biased, meaning that the reference standard either systematically over- or under-segments the truth on average. We previously developed a method to generate synthetic segmentation contours that converge to a polygon approximation of the truth contour [5]. As a significant improvement, in this study, we propose a method, Restorable Segmentation Synthesis (RSS), to generate synthetic segmentation contours that can directly converge to the pixel-level truth contour, instead of its polygon approximation.

Besides verifying the restorability and variability features of synthetic segmentations generated from our RSS method, as the preliminary study, in Section 3.5 we used these synthetic segmentations to analyze three truthing methods: Majority Vote (MV) [9], Truth Estimate from Self Distances (TESD) [2], and Simultaneous Truth And Performance Level Estimation (STAPLE) [12]. And by including the generative model, the RSS can be potentially applied to augment the training dataset for segmentation tasks. In addition, the results in Section 3.6 indicate our contour augmentation approach, when used to supplement training data, can improve a segmentation model's performance.

2 Methods

2.1 Restorable segmentation synthesis

Considering a truth contour in a 2D digital image that consists of N pixels: $P^k(x_k, y_k), k = 1, 2, .., N$ represented by their Cartesian coordinates, we add independent Gaussian noise to the coordinates of each point. For a point $P^k(x_k, y_k)$, it is adjusted to a new location $P_1^k(x_k + \epsilon_1, y_k + \epsilon_2)$ after adding Gaussian noise of zero-mean: $\epsilon_1, \epsilon_2 \sim \mathcal{N}(0, \sigma^2)$. The zero mean of the Gaussian

distribution ensures that the expectation of the new coordinates is the original coordinate. Obviously, the location of P_1^k follows a 2-D Gaussian distribution with a mean at the original location P^k. However, if we apply the adjustment to all pixels in the contour, it will break the contour into unconnected pieces, as shown in the contour $P + \epsilon$ in Figure 1. To overcome this issue, we consider adjusting the contour in the frequency domain using the Fourier transform.

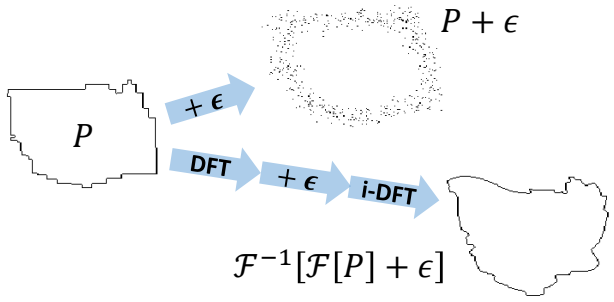

Fig. 1. Adding Gaussian noise ϵ to the pixels' coordinates of a contour in the frequency domain instead of the spatial domain improves the continuity of the contour. DFT is the Discrete Fourier Transform, and i-DFT is the inverse-DFT.

A point in spatial domain with coordinates $P^k(x_k, y_k)$ can be represented by a complex number:

$$P^k = x_k + y_k i,$$

where $k = 0, 1, 2, \cdots, N-1$, and N is the number of points (pixels) on the contour. By applying the Discrete Fourier Transform (DFT):

$$F^u = \frac{1}{N} \sum_{k=0}^{N-1} P^k e^{-i\frac{2\pi}{N}uk},$$

where F^u is a Fourier descriptor (FD) of the contour, and $u = 0, 1, 2, \cdots, N-1$, the complex number of points P^k in the spatial domain can be converted to a F^u in the frequency domain, and P^k can be retrieved from F^u by applying inverse-DFT:

$$P^k \underset{\text{i-DFT}}{\overset{\text{DFT}}{\rightleftharpoons}} F^u = \alpha_u + \beta_u i. \tag{1}$$

F^u is a complex number with a real part α_u and an imaginary part β_u. An important fact is that the Fourier transform (including DFT and inverse-DFT) of a zero-mean Gaussian function is also a zero-mean Gaussian function. Specifically, for $x \sim \mathcal{N}(0, \sigma^2)$, its Gaussian function is:

$$f(x, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}},$$

and, its Fourier transform is:

$$\mathscr{F}\{f(x,\sigma)\} = \frac{1}{\sqrt{\sigma}} f\left(\omega, \frac{1}{\sigma}\right).$$

It follows that the Gaussian noise added to F^u in the frequency domain can be asymptotically eliminated by taking the average in the spatial domain after the inverse-DFT is applied. And unlike adding Gaussian noise to P^k in the spatial domain, adding Gaussian noise to F^u in the frequency domain generally keep the continuity of contour in the spatial domain, as shown in Figure 1. In summary, as shown in Equation 1, all the points on a contour are transformed to FDs by the DFT; then the two coefficients, the real part α_u and the imaginary part β_u of FDs are updated by adding zero-mean Gaussian noise $\epsilon_1, \epsilon_2 \sim \mathcal{N}(0, \sigma^2)$:

$$\begin{cases} \alpha'_u = \alpha_u(1+\epsilon_1) \\ \beta'_u = \beta_u(1+\epsilon_2) \end{cases} \tag{2}$$

Applying the inverse-DFT to the modified FDs results in the contour change, as shown in the lower-right contour in Figure 1. Each coordinate of a point on the changed contour is still subject to an additive independent zero-mean Gaussian noise, and such noise can be asymptotically eliminated by averaging. Thus, the original contour is restorable from a larger number of changed contours.

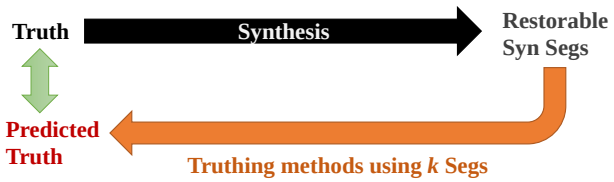

Fig. 2. Our process of truthing methods evaluation compares a truth mask with the predicted truth mask from a truthing method using a number of restorable synthetic segmentations (Syn Segs) generated from that truth mask.

2.2 Assessment of truthing methods

In our study of truthing methods, we applied three truthing methods: Majority Vote (MV) [9], Truth Estimate from Self Distances (TESD) [2], and Simultaneous Truth And Performance Level Estimation (STAPLE) [12] to a set of our restorable synthetic segmentations generated from a truth mask. We then compared the fusion results from the truthing methods with the truth mask.

The assessment process is shown in Figure 2. Truth contours are defined as a set of segmentation contours from real medical images. Restorable synthetic segmentation contours are then generated from the truth contours using our RSS

tool. These synthetic contours are combined with a truthing method into the "predicted truth", which simulates a situation where segmentations by multiple observers are combined into a reference standard. The "predicted truth" (*i.e.*, simulated reference standard) is then assessed by comparing with the truth contours. We assess the quality of the truthing method using the following criteria:

1. The similarity between the original truth mask and the restored mask, as measured by segmentation performance metrics such as the Dice-Sørensen coefficient (DSC).
2. The convergence rate reflected by the number of synthetic masks needed to reach a predefined similarity level. The fewer masks needed, the better.

2.3 Data augmentation for segmentation tasks

Besides the assessment of truthing methods, our restorable segmentation synthesis (RSS) method is also applied to augment the training dataset for segmentation tasks. The dataset in supervised segmentation tasks contains images and their corresponding segmentation masks. In brief, data augmentation using the RSS method for segmentation tasks has three steps:

1. Train a generative model, such as the image-to-image translation model: pix2pix [8] by using pairs of real images and masks.
2. Apply the RSS method to real segmentation masks to generate additional synthetic masks.
3. Transform the synthetic masks to synthetic images by using the synthetic masks as input to a pre-trained generative model (pix2pix).

Figure 3 shows the pipeline to augment data for segmentation tasks using the RSS method. In the following section, we show that augmenting the training dataset for a segmentation task directly benefits the training process of the segmentor and improves its performance.

3 Experiments and Results

3.1 Parameters

Suppose there are N points (pixels) on the contour. After applying the Discrete Fourier Transform (DFT), in the frequency domain, there are N Fourier descriptors (FDs): $\{F^0, F^1, \cdots, F^u, \cdots, F^{(N-1)}\}$. The $F^{(N/2)}$ (or $F^{[(N\pm1)/2]}$) has the maximum frequency; the sequence $\{F^1, F^2, \cdots, F^{(N/2)}\}$ is from low to high frequencies, and the sequence $\{F^{(N/2)}, F^{(N/2)+1}, \cdots, F^{(N-1)}\}$ is from high to low frequencies. Due to this mirror nature of the frequency domain signal, the DFT of a time-domain signal is often shifted such that the first half of its spectrum is in positive frequencies and the second half is in negative frequencies [1], with

[1] https://www.mathworks.com/help/matlab/ref/fft.html

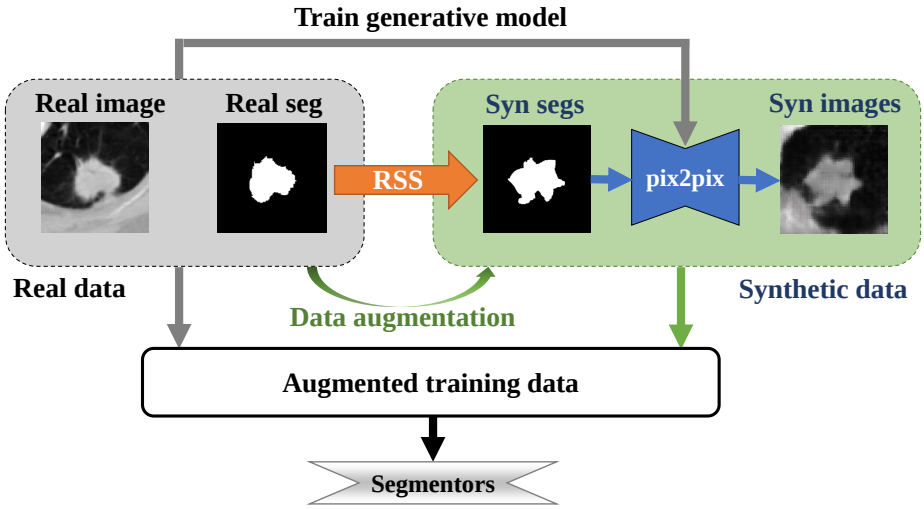

Fig. 3. The process of how our RSS method is applied in data augmentation for segmentation tasks.

the first element F^0 reserved for the zero frequency in the middle, or named the *Direct Current* (DC)-component.

In our method, we do not change the DC-component F^0 because the DC-component controls the location of the contour, and it does not affect the contour's shape. We also do not change the very high frequency components because these only result in tiny differences in the contour's shape. Thus, to synthesize a new contour, we perturb FDs in a middle band of the frequency domain. Formally, the user specifies the **range** $\{l, h\}$ $(l \leq h)$ for FDs $\{F^l, F^{(l+1)}, \cdots, F^{(h-1)}, F^h\}$ and $\{F^{(N-h)}, F^{(N-h+1)}, \cdots, F^{(N-l-1)}, F^{(N-l)}\}$ to add noise, where l refers to the lower-end frequency and h for the higher end. By definition, $l \geq 1$ and $h \leq \lceil N/2 \rceil - 1$. For example, for the range $\{l = 2, h = 5\}$ (suppose $N = 20$), the FDs to add noise are:

$$\{F^2, F^3, F^4, F^5\} \text{ and } \{F^{15}, F^{16}, F^{17}, F^{18}\}.$$

The other parameter to specify is the **standard deviation** σ for the zero-mean Gaussian noise added to these middle-band FDs (Equation 2).

3.2 Materials

In this study, the original input (truth mask) is the lung nodule segmentation from the LIDC-IDRI dataset [1]. The LIDC-IDRI dataset consists of diagnostic and lung cancer screening thoracic computed tomography (CT) scans with radiologist-annotated lesions. Figure 4 displays an example of creating synthetic

segmentations using our RSS method from a lung nodule segmentation and the convergence to the original (truth mask) input. The truth mask in this example is a STAPLE-fused mask combining four annotations for the image *LIDC-IDRI-0078/nod_0/slice_3*.

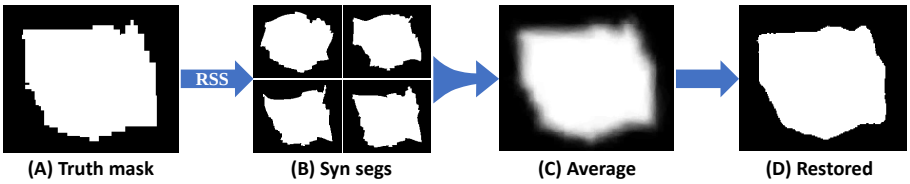

Fig. 4. Synthetic segmentations from a lung nodule segmentation and the convergence to the truth mask (original input). In (A), the truth mask is a fused mask of annotations by radiologists on the CT image. (B) shows four examples of synthetic segmentations ($l = 2, h = 10, \sigma = 1$) generated from the truth mask by our RSS method. (C) is the average image through 100 synthetic segmentations. (D) is the restored binary segmentation from (C) by applying a threshold of 0.5 to the average image. The DSC score between (D) and (A) is 0.98.

3.3 Restorability

To validate the restorability property of our RSS method, *i.e.*, the ability to asymptotically converge to the truth mask by averaging the synthetic segmentation masks, we applied the RSS method with parameters: $l = 2, h = 10, \sigma = 1$ on the truth mask shown in Figure 4A to generate six groups of synthetic segmentations including 50, 100, 500, 1000 (1k), 2k, and 5k masks, respectively (Figure 4B).

As shown in the fourth (*Restored_DSC*) column of Table 1, **the DSC scores between the average of the synthetic masks (using a threshold of 0.5) and the truth mask** are greater than 0.98 for groups including 100 synthetic segmentations and more. Figure 4 (C) and (D) show the result for group 2 with 100 synthetic masks. Columns *mean_DSC* and *std_DSC* in Table 1 show the mean and standard deviation of DSC scores between synthetic segmentations and the truth mask. Results in the column *Restored_DSC* show our method can provide a high degree of restorability as we were able to achieve a relatively high DSC score.

3.4 Variability

It is useful that the synthetic segmentations have a variability to mimic the real-world. Here we examine how the parameters **range** ($\{l, h\}$) and **standard deviation** (σ) affect the variability of the synthetic contours. We generated eight groups of synthetic segmentation by setting $\{l = 2, h = 10\}$ and $\{l = 4, h = 12\}$,

Table 1. The second column is the mean of DSC scores when comparing individual synthetic segmentation masks with the truth mask. The third column is the standard deviation (std) of these DSC scores. The fourth column is the DSC scores between the truth mask and the reproduced mask by averaging the synthetic segmentations.

#	mean_DSC	std_DSC	**Restored_DSC**
50	0.9248	0.0175	**0.9778**
100	0.9237	0.0180	**0.9804**
500	0.9228	0.0187	**0.9819**
1k	0.9240	0.0171	**0.9827**
2k	0.9225	0.0182	**0.9829**
5k	0.9222	0.0193	**0.9830**

and $\sigma = 1, 2, 3, 4$ from the same truth mask shown in Figure 4A. Each group includes 1000 synthetic segmentations. Figure 5 (left) shows the DSC between synthetic segmentation and the truth mask, and Figure 5 (right) shows the histogram of the 1000 DSC values for the group: $\{l = 2, h = 10, \sigma = 1\}$. Their DSC ranges from 0.8474 to 0.9658, and the histogram's bin width is 0.006.

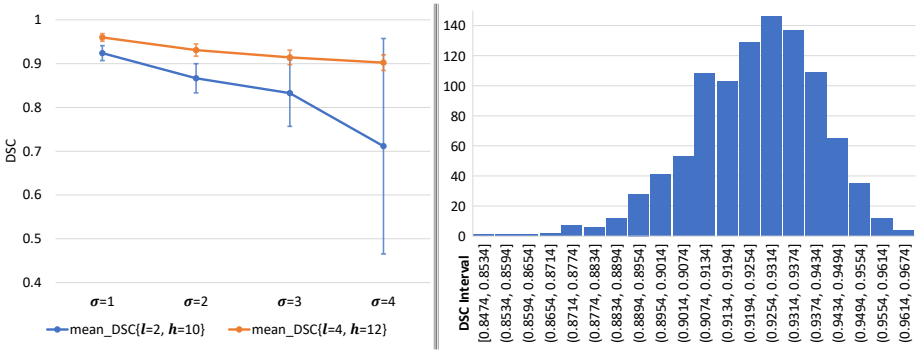

Fig. 5. *Left*: DSC between the truth mask and synthetic segmentations generated by different range ($\{l, h\}$) and standard deviation (σ) parameters for RSS. Bars are the standard deviation (std) of these DSC scores for 1000 synthetic segmentations. *Right*: the histogram of DSC for 1000 synthetic segmentations in the group: $\{l = 2, h = 10, \sigma = 1\}$. The x-axis represents the intervals of DSC, and the y-axis represents the counts of DSC for each interval.

The results in Figure 5 (left) show that a larger standard deviation (σ) results in a smaller mean DSC and a wider range (variability) of synthetic segmentations. Changes in the Fourier descriptors (FDs) frequencies $\{l, h\}$ also control the variability. Specifically, changing the lower frequency FDs results in a smaller mean DSC and a wider range (variability) for the synthetic segmentations. Thus,

the variability of the generated synthetic segmentation is managed by these parameters for a given truth mask.

3.5 Performance of truthing methods

As described in Methods (Section 2.2), we generated k synthetic segmentation contours with our RSS method (Section 2.1) with parameters: $l = 2, h = 10, \sigma = 1$ from the truth mask shown in Figure 4A. Second, we applied the three truthing methods: MV, TESD, and STAPLE, to the k synthetic masks to obtain a fused (predicted) mask as the reference standard. Then, we compared the reference standard with the original truth mask based on the DSC. For each truthing method and number of synthetic segmentations k, the process was repeated 10 times to obtain 10 DSC scores. Finally, we plotted the mean and standard deviation (std) of these DSC scores for k from 2 to 100 in Figure 6 and show selected mean DSC scores in Table 2.

Table 2. Selected mean DSC scores between the truth mask and the reference standard masks fused by truthing methods from k synthetic masks.

k	2	3	4	5	6	7	8	9	10	20	30	40	50	60	70	80	90	100
MV	0.9227	0.9463	0.9511	0.9572	0.9594	0.9629	0.9651	0.9661	0.9676	0.9733	0.9758	0.9769	0.9787	0.9796	0.9801	0.9805	0.9804	0.9804
TESD	0.9390	0.9497	0.9538	0.9597	0.9618	0.9644	0.9649	0.9678	0.9680	0.9732	0.9758	0.9768	0.9788	0.9796	0.9800	0.9804	0.9804	0.9804
STAPLE	0.9227	0.9463	0.9439	0.9396	0.9543	0.9506	0.9439	0.9448	0.9512	0.9461	0.9464	0.9454	0.9456	0.9463	0.9458	0.9456	0.9456	0.9454

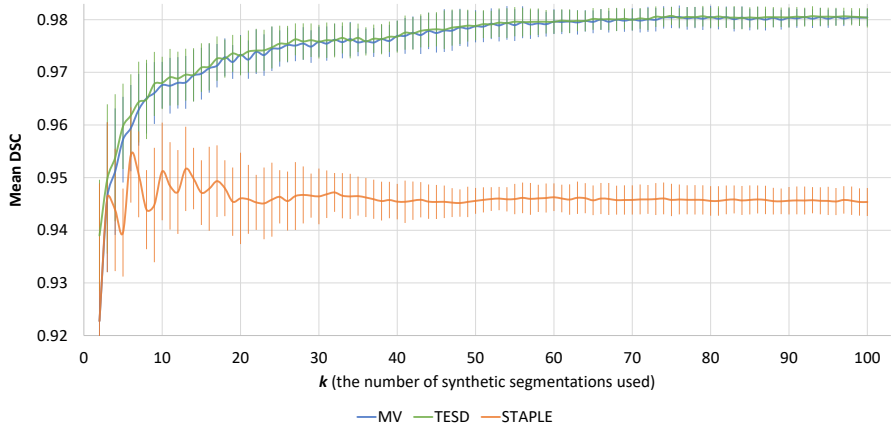

Fig. 6. Mean DSC scores between the truth mask and the reference standard masks fused using truthing methods, MV (blue), TESD (green), and STAPLE (orange), for different numbers of synthetic segmentations (parameters: $l = 2, h = 10, \sigma = 1$). The x-axis is the number of synthetic segmentations k, and the y-axis is the mean DSC score. The error bar show one std above and below.

We observed that adding more images does not help STAPLE converge to the truth mask, while MV and TESD, performed similarly, converging to higher DSC scores than STAPLE with more samples. Interestingly, for $k = 2$ and 3, MV and STAPLE were identically in DSC scores. For small k (2 to 7), TESD's DSC scores are greater, and STAPLE's scores are lower for almost all k. By looking at the comparisons of fusion results for k =100 and the truth mask in Figure 7, STAPLE tends to create larger areas with False positive (FP) [2] areas but very few False negative (FN) areas. MV and TESD perform very similarly to each other, and their results have fewer FP but more FN areas than STAPLE.

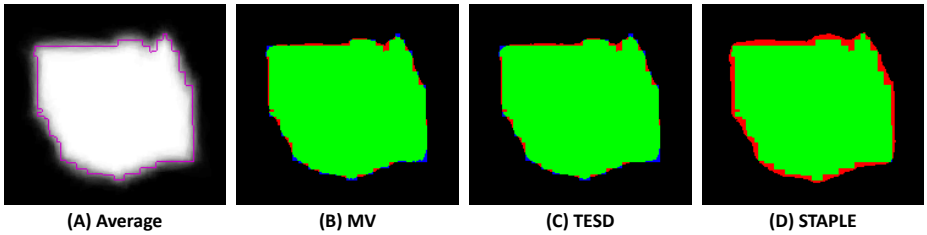

Fig. 7. Comparisons of fusion/average results and the truth mask/contour. In (A), the truth contour is in *Purple* and the gray-scale image is the average of 100 synthetic masks. (B), (C), and (D) show the fusion results when applying the different truthing methods to 100 synthetic masks that overlap the truth mask. The True positive (TP) areas are in *Green*, False positive (FP) in *Red*, and False negative (FN) in *Blue*.

3.6 Segmentation data augmentation

To study the power of our RSS method in segmentation data augmentation, we collected 2579 pairs of real lung nodule images and truth masks from the LIDC-IDRI dataset (Section 3.2). The truth masks are MV-fused masks of annotations from the central slice of nodules. About 70% of the image pairs (1805) are divided (by patients or cases) to serve as the real data to train the generative model (pix2pix) and segmentation models, and to generate synthetic masks using our RSS method. The remaining 30% of the real image pairs (774) are kept as test data. The test data are only used to evaluate the performance of the trained segmentation models and are never used in model training or synthetic masks generation. We applied the processes described in Section 2.3 to augment the segmentation data using only the training data (1805 image pairs):

1. Use the 1805 image pairs to train the pix2pix [8] model (generator architecture: unet_128). The image-to-image translation model can transform the input binary masks to (fake) lung nodule images. The saved model is trained for 200 epochs.

[2] https://en.wikipedia.org/wiki/Confusion_matrix

2. Apply the RSS method (parameters: $l = 2, h = 20, \sigma = 3$) to the 1805 training truth masks. For each truth mask, K synthetic masks are generated. There are $1805 \cdot K$ synthetic masks in total.
3. Input the $1805 \cdot K$ synthetic masks into the trained the pix2pix model and generate $1805 \cdot K$ synthetic lung nodule images accordingly.

Including the real data for training (1805 image pairs), the augmented training dataset has $1805 \cdot (K + 1)$ image pairs of lung nodule images and binary masks. We refer to the augmented training dataset as $\mathbf{TrainSet}_K$. For $K = 0$, the $\mathbf{TrainSet}_0$ only contains the real data without data augmentation. Figure 8 displays examples of segmentation data augmentation for $K = 3$.

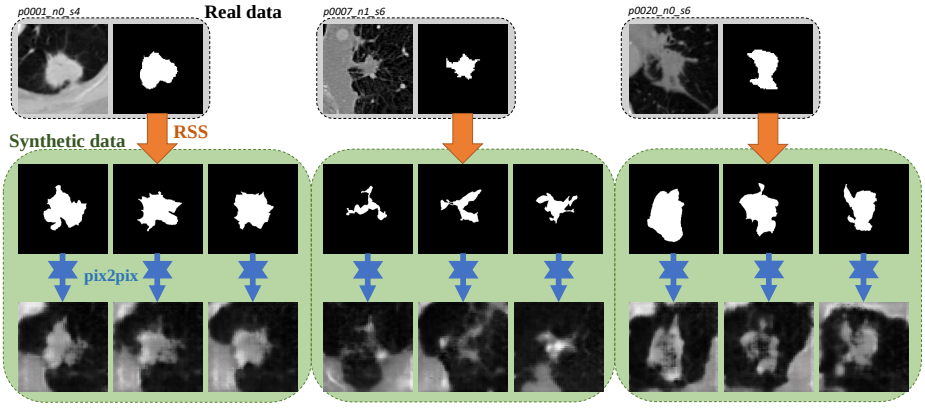

Fig. 8. Examples of segmentation data augmentation. The real data contain lung nodule images and truth masks from the LIDC-IDRI dataset. Three synthetic masks are generated from a truth mask by RSS. They are then transformed into synthetic lung nodule images via a trained pix2pix model.

To estimate the improvement from our data augmentation approach, the final step is to train the segmentation models (segmentors) on $\mathbf{TrainSet}_K$ and to obtain the test accuracy (Jaccard index or IoU [3]) of trained segmentors on the 774 real test image pairs. Specifically, the segmentor is the U-net model [7] (with ResNeXt encoder). The segmentor is **trained five times from scratch** on $\mathbf{TrainSet}_K$ ($K = 0, 1, 2, 3, 4$) for 60 epochs. The training accuracy and test accuracy (metrics: IoU score) are recorded after each training epoch. Table 3 shows the maximum and mean values of test IoU scores for segmentors trained on the five datasets. Adding synthetic data (data augmentation) improves the model's performance on the test data, and the more synthetic data added, the better the performance achieved. The results in Figure 9, not only improved the overall model performance, but also fewer training epochs to achieve a given performance level.

[3] https://en.wikipedia.org/wiki/Jaccard_index

Table 3. All five training datasets consist of 1805 real data, but different numbers of synthetic data. The U-net segmentation model is trained from scratch on each of the five datasets for 60 epochs, and 60 test IoU scores are recorded after each training epoch. Here are the maximum and mean values of the 60 test IoU scores.

Training Data	Real	Synthetic	max_Test-IoU	mean_Test-IoU
TrainSet$_0$	1805	0	0.6473	0.6250
TrainSet$_1$	1805	1805	0.6524	0.6318
TrainSet$_2$	1805	3610	0.6524	0.6352
TrainSet$_3$	1805	5415	0.6539	0.6371
TrainSet$_4$	1805	7220	0.6569	0.6444

Fig. 9. Plots of the training and test IoU scores after each training epoch. *Left*: the U-net segmentor [7] is trained from scratch on each dataset **TrainSet**$_0$, **TrainSet**$_2$, and **TrainSet**$_4$ for 60 epochs. *Right*: after each training epoch, the segmentor is tested on the test data to obtain test IoU scores.

4 Conclusion

In this study, we proposed the Restorable Segmentation Synthesis (RSS) method that can generate synthetic segmentation contours from given (truth) masks, with the property that averaging these synthetic segmentations can closely restore the original truth mask. Using the DSC, we verified the restorability and variability of generated synthetic segmentations. The RSS method allows for the creation of a benchmark dataset that can potentially be used to compare truthing methods for medical image segmentation. As a preliminary study, we used our RSS tool to analyze three truthing methods: MV, TESD, and STAPLE. By including the use of a generative model, the RSS can also be used to augment the training dataset for segmentation tasks. In another preliminary study, we transformed synthetic masks generated using the RSS tool into synthetic images using a pre-trained generative model (pix2pix). These results showed that our augmentation method can improve a segmentation model's performance.

Acknowledgments. The authors would like to thank Nicholas Petrick, Ph.D., for providing helpful review and comments on the manuscript. The mention of commercial products, their sources, or their use in connection with material reported herein is not to be construed as either an actual or implied endorsement of such products by the Department of Health and Human Services. This is a contribution of the U.S. Food and Drug Administration and is not subject to copyright.

References

1. Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Clarke, L.P., et al.: Data from lidc-idri. the cancer imaging archive. DOI http://doi. org/10.7937 K **9**(7) (2015)
2. Biancardi, A.M., Reeves, A.P.: Tesd: a novel ground truth estimation method. In: Medical Imaging 2009: Computer-Aided Diagnosis. vol. 7260, pp. 1116–1123. SPIE (2009)
3. Guan, S., Samala, R.K., Arab, A., Chen, W.: MISS-tool: medical image segmentation synthesis tool to emulate segmentation errors. In: Medical Imaging 2023: Computer-Aided Diagnosis. vol. 12465, pp. 273–281. SPIE (2023)
4. Guan, S., Samala, R.K., Chen, W.: Informing selection of performance metrics for medical image segmentation evaluation using configurable synthetic errors. In: 2022 IEEE Applied Imagery Pattern Recognition Workshop (AIPR). pp. 1–8. IEEE (2022)
5. Guan, S., Samala, R.K., Kahaki, S.M., Chen, W.: Restorable synthesis: average synthetic segmentation converges to a polygon approximation of an object contour in medical images. In: 2024 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI). pp. 77–80. IEEE (2024)
6. Hesamian, M.H., Jia, W., He, X., Kennedy, P.: Deep learning techniques for medical image segmentation: achievements and challenges. Journal of digital imaging **32**(4), 582–596 (2019)
7. Iakubovskii, P.: Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch (2019)
8. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
9. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.A., Arbel, T., Avants, B.B., Ayache, N., Buendia, P., Collins, D.L., Cordier, N., Corso, J.J., Criminisi, A., Das, T., Delingette, H., Demiralp, C., Durst, C.R., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci, A., Iftekharuddin, K.M., Jena, R., John, N.M., Konukoglu, E., Lashkari, D., Mariz, J.A., Meier, R., Pereira, S., Precup, D., Price, S.J., Raviv, T.R., Reza, S.M.S., Ryan, M., Sarikaya, D., Schwartz, L., Shin, H.C., Shotton, J., Silva, C.A., Sousa, N., Subbanna, N.K., Szekely, G., Taylor, T.J., Thomas, O.M., Tustison, N.J., Unal, G., Vasseur, F., Wintermark, M., Ye, D.H., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M., Van Leemput, K.: The multimodal brain tumor image segmentation benchmark (brats). IEEE Transactions on Medical Imaging **34**(10), 1993–2024 (2015). https://doi.org/10.1109/tmi.2014.2377694, https://dx.doi.org/10.1109/tmi.2014.2377694

10. Taha, A.A., Hanbury, A., Toro, O.A.J.d.: A formal method for selecting evaluation metrics for image segmentation. In: 2014 IEEE International Conference on Image Processing (ICIP). pp. 932–936 (2014). https://doi.org/10.1109/ICIP.2014.7025187
11. Taha, A.A., Hanbury, A.: Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. BMC medical imaging **15**(1), 1–28 (2015)
12. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE transactions on medical imaging **23**(7), 903–921 (2004)
13. Zheng, Y., Li, G., Li, Y., Shan, C., Cheng, R.: Truth inference in crowdsourcing: Is the problem solved? Proceedings of the VLDB Endowment **10**(5), 541–552 (2017)