EFFECTIVE 3D BOUNDARY LEARNING VIA A NONLOCAL DEFORMABLE NETWORK

Yueyun Liu[a], *Yu Wang*[b] *and Yuping Duan*[a*]

[a]Center for Applied Mathematics, Tianjin University, Tianjin, 300072, China
[b]DAMO Academy, Alibaba Group, China

ABSTRACT

Due to the unbalance between the boundary pixels and regional pixels, the accuracy of boundary prediction is a challenging issue for learning-based medical segmentation approaches. In this paper, we propose a two-stage segmentation method to identify and refine the object boundary accordingly. By modeling the boundary by the signed distance function, we develop a nonlocal deformable convolutional network to accurately predict the local geometry of boundaries. We also introduce an efficient loss function to enhance the learning ability in the boundary area. Experiments on two public spleen datasets can evidence the superior performance of the proposed model compared to the existing 2D, 3D, and boundary-based learning methods.

Index Terms— Signed distance function, boundary segmentation, deformable convolution, point cloud, refinement

1. INTRODUCTION

Medical image segmentation is a challenging task, but also a key component for smart medicine. With the development of the deep learning method, various Convolutional Neural Network (CNN) models have been devoted for medical image segmentation, such as U-Net [1], DeepLab [2], and so on.

Focusing on organ segmentation, a typical 3D segmentation problem, both 2D and 3D CNNs have been successfully used in literature. Milletari *et al.* [3] introduced a 3D CNN called V-Net to catch the spatial context. Yu *et al.* [4] proposed a two-stage segmentation model named recurrent saliency transformation network (RSTN) for small organ segmentation, which took multi-slices as input to emphasize the spatial information and updated the bounding box in an iterative way to produce better results. A common point of aforementioned models is that they are all region-based methods, leading to inaccuracies on boundaries. Fabian *et al.* [5] developed the nnU-Net, a self-configuring method that can automatically configures the preprocessing, network architecture, training and post-processing for any new task. However, a main limitation of nnU-Net is its high time consuming.

*Corresponding author: Yuping Duan (yuping.duan@tju.edu.cn). The work was supported by National Natural Science Foundation of China (NSFC 12071345, 11701418), Major Science and Technology Project of Tianjin 18ZXRHSY00160 and Recruitment Program of Global Young Expert.

(a) Raw image (b) Ground truth (c) SDF

Fig. 1. An illustration of the relationship between binary segmentation and Signed Distance Function (SDF).

Traditional boundary-based segmentation methods realize segmentation by tracking the explicit curve or surface for 2D and 3D problems, which are very efficient and require low memories. The representative geodesic active contour model [6] represented the contour implicitly by a level set function to allow topological changes such as splitting and merging, which have been widely used for various segmentation tasks and medical image segmentation. Inspired by that, Peng *et al.* [7] introduced a contour-based model with circular convolution to predict the offset of initial octagon points, and then deformed the points to achieve segmentation. Ni *et al.* [8] proposed an elastic boundary projection (EBP) model, which placed a number of pivot points in 3D space and evolved them to object boundaries along the predefined directions. Guo *et al.* [9] introduced a learned snakes model (LSM) consisting of surface initialization and evolution, which estimated the offsets by 2D U-Net. Although deforming the initial points to match the desirable boundary or surface shows a promising result with desirable boundaries, they require a series of post-processing to obtain pixel-wise segmentation results, which leads to an increase in computation time. As shown in Fig. 1, the segmentation can also be realized by learning the level set function rather than offsets. Ma, He and Yang [10] designed a geodesic active contour loss to learn SDF for naturally embedding the 3D contour as the zero level set. Xue *et al.* [11] used a 3D U-Net as the backbone network to learn SDF directly from medical scans for both large and small organs.

This work introduces a novel boundary learning method using a cascade architecture to initialize and refine the SDF accordingly. Unlike the existing EBP and LSM, which used 2D U-Net to estimate the offsets from image intensity, we use the deformable point convolution to learn the SDF and introduce a non-local module to leverage global spacial infor-

mation. By modeling the boundary by SDF, small changes of shapes can be easily identified and treated during the network training process. Meantime, the deformable kernel point convolution (KPConv) [12] is used to meet the geometric changes of boundaries. We also propose a novel loss function to guide the network to converge to the true boundary. Compared to the existing 2D, 3D and boundary-based learning methods, our model can better incorporate the local and global information and possess high efficiency.

2. OUR METHOD

Our method adopts the two-step framework, which can be regarded as a *coarse-to-fine* [4] or *crop-then-refine* strategy [13]. The first step can be realized by general CNN-based methods, where DeepLab-v3 is used to obtain coarse boundary location.

2.1. Our boundary learning framework

Boundary initialization: Suppose $\widetilde{\mathrm{P}}$ be the coarse segmentation result. We construct the initial boundary as a point set with bandwidth d defined by

$$\mathcal{B} = \{x \in \Omega \mid \min_{y \in \partial \widetilde{\mathrm{P}}} \|x - y\|^2 \leq d\}, \tag{1}$$

where $\|x - y\|^2$ presents the Euclidean distance. The narrow band strategy can reduce the dependency of the boundary refinement to the initialization and improve the robustness.

Boundary refinement: We use the KPConv operation to construct the boundary refinement network to learn the signed distance function for points on $\mathcal{B}$. In particular, the inputs are spheres centered on $x \in \mathcal{B}$, defined by $\mathcal{N}(x) = \{z_i \in \Omega \mid \|z_i - x\|^2 \leq r\}$ with r representing the radius of the spheres. The kernel point convolution is defined as

$$(\mathcal{F} * g)(x) = \sum_{z_i \in \mathcal{N}(x)} g(z_i - x) f_i, \tag{2}$$

and the kernel function g is given as

$$g(z_i - x) = \sum_{k<K} \max\left(0, 1 - \frac{\|(z_i - x) - y_k\|}{\sigma}\right) W_k, \tag{3}$$

where $\{y_k | \ \|y_k\|^2 \leq r, k < K\}$ denotes the kernel points, σ is the influence distance of kernel points and $\{W_k | \ k < K\}$ are the weight matrices to be trained from the data.

The architecture of our network is displayed in Fig. 2, where the KP residual blocks are designed similar to the bottleneck ResNet blocks with the image convolution replaced by the point convolution. There are four scales in the encoder-decoder process, where the channel numbers are of 64, 128, 256 and 512, respectively. For the decoder, we use the nearest upsampling operation to upsample the features, and concatenate the features of the encoder accordingly. The leaky ReLu is used as the activate function, and two shared multi-layer

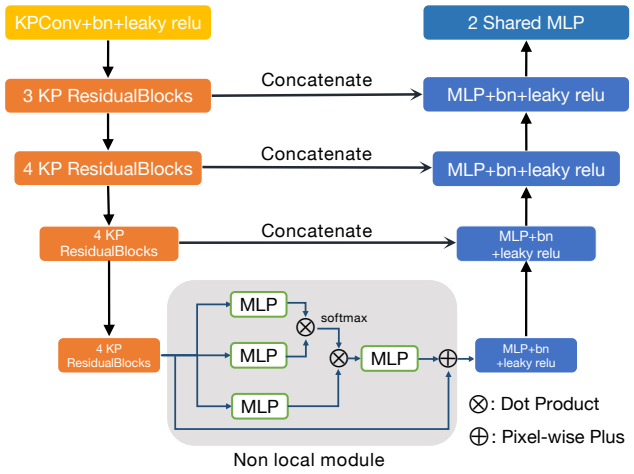


Fig. 2. An illustration of the network architecture, where the grey box is the non-local module.

perceptrons (MLP) are used to aggregate features to predict the signed distance function. Because the kernel point convolution only aggregates local features, we introduce the non-local module to capture long-range dependencies, where the multi-head self-attention block is the key part used to fuse the features similar to [14]. The non-local module can capture global information without building up a very deep network.

Inference: During inference, the evolved boundary points together with their neighboring points located in a narrow band with bandwidth d are assembled based on the coarse segmentation delivery. Then we use the Heaviside function to convert the signed distance function into a probability map, and obtain the binary segmentation by a thresholding. Finally, we apply the post-processing to remove the isolated points.

2.2. Our loss function

To guarantee the learning accuracy, we explicitly define the SDF as $\Phi : \Omega \subset \mathbb{R}^3 \to \mathbb{R}$ to describe the boundary position

$$\Phi(x) = \begin{cases} -\inf\limits_{y \in \partial \mathcal{D}} \|x - y\|_2, & \text{if } x \in \mathcal{D}; \\ 0, & \text{if } x \in \partial \mathcal{D}; \\ \inf\limits_{y \in \partial \mathcal{D}} \|x - y\|_2, & \text{if } x \in \Omega \backslash \mathcal{D}; \end{cases} \tag{4}$$

where $\mathcal{D}$ represents the region of interest and $\partial \mathcal{D}$ represents the boundary. The boundary can be well identified by the signed distance function, i.e., $\Phi(x) = 0$. We further rewrite the signed distance function by Heaviside function. In order to overcome the non-differentiability, we use the approximated Heaviside function [15] as follows

$$H_\epsilon(x) = \frac{1}{2}(1 + \frac{2}{\pi} \arctan(\frac{x}{\epsilon})), \tag{5}$$

where ϵ is fixed as $\epsilon = \frac{1}{32}$ in numerical experiments.

Region loss: We adopt the DSC loss together with the smooth ℓ_1 loss as the regional penalty, where DSC is used to

measure the overlap between the prediction and ground truth

$$L_{DSC}(\theta) = 1 - 2\frac{\sum_{\Omega} \mathbf{G} \odot H_{\epsilon}(\Phi_{\theta})}{\sum_{\Omega} \mathbf{G} + H_{\epsilon}(\Phi_{\theta})}, \tag{6}$$

with θ being the parameters of the network, and $\odot$ representing the pixel-wise multiplication. The smooth ℓ_1 loss $L_{\delta}(\theta) = \|\Phi_{\theta} - \Phi_G\|_1$ if $\|\Phi_{\theta} - \Phi_G\| \geq 1$, and $L_{\delta}(\theta) = \|\Phi_{\theta} - \Phi_G\|_2$ otherwise, used to drive the predicted signed distance function Φ to be as close as possible to the true SDF Φ_G. The DSC loss supervises on the probability map of the prediction centering on the overlap areas, while the smooth ℓ_1 loss takes the segmentation as a regression task, complementing the ignoring areas of the DSC loss.

Boundary loss: We also introduce the binary cross-entropy loss on the boundary, which is defined as

$$L_B(\theta) = \sum_{x \in \mathcal{B}} G(x) \log(1 - H_{\epsilon}(\Phi_{\theta}(x)) + (1 - G(x)) \log(H_{\epsilon}(\Phi_{\theta}(x)). \tag{7}$$

The boundary loss can drive the points with the signed distance function of value zero, i.e., the zero level set function, to match the organ boundary. To sum up, our loss function is defined as

$$L(\theta) = L_{DSC}(\theta) + \alpha L_{\delta}(\theta) + \beta L_B(\theta), \tag{8}$$

where the three terms are trained as a whole. In our experiments, we set the weights $\alpha = 0.1$ and $\beta = 1$ to balance the magnitude of each term.

3. IMPLEMENTATION AND EXPERIMENTS

For saving computational cost, we choose the simplest Deeplab-v3 with Resnet-18 as the backbone network for coarse segmentation. For the boundary refinement network, we set the narrow band width as $d = 1$, radius as $r = 20$ and kernel size as $K = 15$, and train about 90000 iterations with batch size being 6 and learning rate being 5e-5.

3.1. Dataset

The Medical Segmentation Decathlon (MSD) spleen dataset[1] contains 41 CT volumes. The number of 2D slices ranges between [31,168]. We clip the intensities of all images into [-125, 275] as EBP [8] suggests, and randomly divide the dataset into two groups, 20 and 21 volumes, respectively. We use cross-validation to evaluate the segmentation performance.

Another public dataset used for evaluation is TMI spleen dataset [16], consisting of two sub-datasets with 43 volumes selected from The Cancer Imaging Archive (TCIA) [17] and 47 volumes from the Beyond The Cranial Vault (BTCV) segmentation challenge [18]. Considering the thickness of the spleen in TCIA subset ranged in [70,160] is much larger than BTCV subset ranged in [16,38], we normalize the thickness of all volumes to 80 after the coarse segmentation. We clip the intensities of all images into [-50, 200], and randomly separate the volumes into two groups, one contains 21 volumes from TCIA and 24 volumes from BTCV, and the other one contains the rest volumes. The cross-validation is also used to evaluate the segmentation performance.

[1] http://medicaldecathlon.com/

Table 1. Comparison results on the MSD spleen dataset.

Methods	DSC	Max	Min	HD(mm)	B-box
nnU-Net	96.10±2.30	97.89	84.42	1.79	N
RSTN	91.80±9.91	97.26	46.85	7.71	N
V-Net	93.18±5.88	97.32	59.95	4.32	Y
EBP	90.54±6.34	96.75	68.91	6.69	Y
LSM	92.75±2.32	96.29	85.91	3.35	N
Initialization	89.99±2.68	95.86	82.83	3.55	N
Our Model	**96.61±1.30**	**98.20**	**92.13**	**1.13**	N

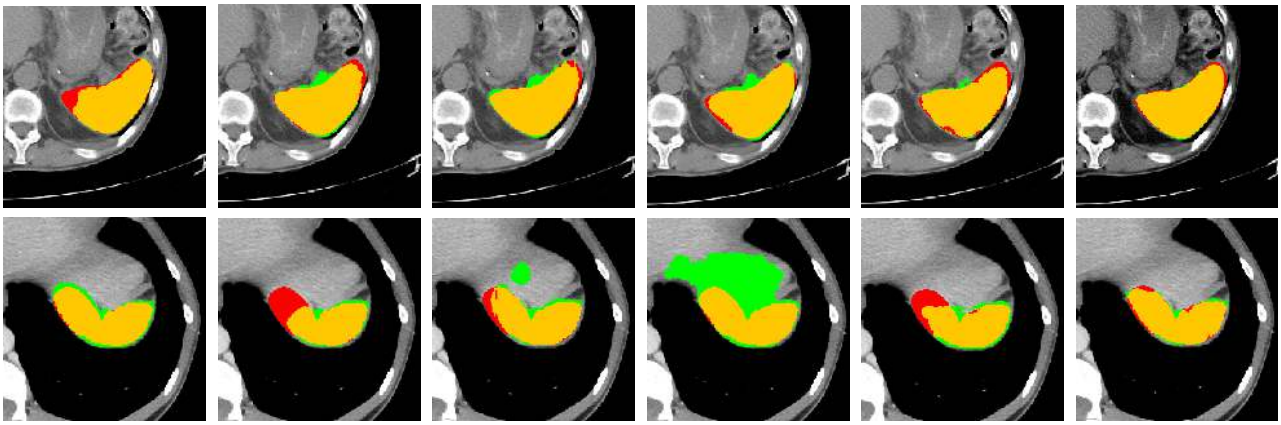
Table 2. Comparison of the baseline and our models with different loss terms combination.

Methods	DSC	Max	Min	HD(mm)
KP-FCNN	95.25±3.59	98.00	80.13	7.40
Ours+L_{DSC}	95.55±3.14	98.01	81.46	5.02
Ours+L_{DSC}+L_{δ}	96.02±2.33	97.95	86.87	2.79
Ours+L_{DSC}+L_{δ}+L_B	96.47±1.56	98.20	90.72	2.41

3.2. Results on MSD spleen dataset

Results comparison: We compare our model with several SOTA segmentation methods, including fully automatic nnU-Net [5], 2D RSTN [4], 3D V-Net [3], boundary-based EBP [8] and LSM [9]. The comparison methods are re-implemented by ourselves to obtain the cross-validation results. Specifically, we use the cascade 3D U-Net architecture for nnU-Net and train five folds for each cross-validation. The settings and parameters of nn-UNet are chosen automatically, and the fastest mode is used for testing. The V-Net is a patch-based model following the settings suggested in [8]. We first normalize all volumes along the long axis, and randomly crop $128 \times 128 \times 64$ patches for training. We use the same coarse segmentation for LSM and our model. As can be seen from Table 1, our model achieves the best segmentation accuracy in terms of DSC and HD. The DSC of our model surpasses LSM by more than 3% with the same coarse segmentation, which demonstrates the effectiveness of our nonlocal deformable network. One group of typical visual comparison results are displayed on the first row of Fig. 3.

Ablation Analysis: We further discuss the contribution of our network architecture and the three loss terms. We use the KP-FCNN network [12] as the baseline, and train our network with different combinations of loss functions on the MSD spleen dataset. To understand the model, we list the results without the post-processing in Table 2. By comparing the results in the first two rows, our model gives a better learning ability than KP-FCNN due to the nonlocal module, which helps the network to incorporate the global features.



(a) nnU-Net (b) RSTN (c) V-Net (d) EBP (e) LSM (f) Our model

Fig. 3. Visual comparison between our model and SOTA methods, where green, red and yellow color represent the prediction, ground truth and overlapped region, respectively. Here, the 1st and 2nd row are selected from MSD and TMI datasets.

Table 3. Comparison results on the TMI spleen dataset in terms of DSC, mean 95% HD(mm), parameters and run time(m).

Methods	TMI dataset				TCIA subset		BTCV subset	
	DSC	HD	parameters	Time(m)	DSC	HD	DSC	HD
nnU-Net	94.46±10.83	5.56	6.2e7	31.92	96.07±3.18	**2.93**	**92.99**±14.60	7.96
RSTN	93.50±10.93	5.75	8.1e7	1.1	95.79±7.31	4.56	91.41±13.15	6.84
V-Net	90.64±7.51	18.91	1.9e7	0.2	92.10±6.30	21.67	89.31±8.31	16.38
EBP	88.16±11.46	13.20	2.3e6	72.1	90.68±10.12	15.27	85.90±12.20	11.35
LSM	89.79±8.30	11.64	1.6e7	5.7	91.32±5.54	13.50	88.39±10.05	9.94
Initialization	89.19±7.87	13.47	1.5e7	0.1	91.00±4.21	15.02	87.54±12.06	9.89
Our Model	**94.54±5.63**	**3.78**	3.4e6	1.1	**96.41±2.53**	2.98	92.83±7.01	**4.52**

The last three rows demonstrate the effectiveness of different loss terms. By adding the smooth ℓ_1 term, the performance increases about 0.5% in terms of DSC, and the HD decreases from 5.02 to 2.79. After introducing the boundary term L_B, the DSC increases another 0.4% such that the boundary term works as supplementary to the region terms. The combined three term loss function leads to the best segmentation accuracy.

3.3. Results on TMI spleen dataset

We also perform the segmentation methods on the TMI spleen dataset, where DSC, HD, the number of parameters and inference time are shown in Table 3. As can be seen, our model produces promising segmentation results better than all comparative methods. Especially, compared to the other two boundary-based methods EBP and LSM, our model not only shows a significant superiority on segmentation accuracy, but also saves lot of computational time. The established nnU-Net incorporates the data augmentation techniques, including cropping data to the region of nonzero value, elastic deformations, mirroring and so on. What is more, the nnU-Net ensembles the predictions based on five well trained models to obtain high accuracy. Even so, our model gives the highest values of DSC and HD among all methods. More importantly, the proposed model consumes much less inference time than nnU-Net, saving about 30 minutes, which is also an important issue for real-world applications. Last but not the least, our model is shown to be more stable than others, which can be seen by the standard deviation on the TMI dataset. We also present a typical visual comparison results on the second row of Fig. 3.

4. CONCLUSION

In this paper, we promoted a cascade segmentation pipeline based on a nonlocal deformable convolutional neural network for realizing 3D organ segmentation. Our model learned the signed distance function to identify the boundary of the target, which can be performed as a refinement technique for the existing segmentation methods. Numerical experiments showed that our model gives promising segmentation results, much better than the existing 3D boundary learning approaches. Besides, the ablation study demonstrates the effectiveness of our nonlocal module and the multi-level structural loss function. This work only focused on the spleen segmentation. Our future work is to investigate its performance on more challenging segmentation tasks such as pancreas, gallbladder etc.

5. REFERENCES

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam, "Rethinking atrous convolution for semantic image segmentation," *https://arxiv.org/abs/1706.05587*, 2017.

[3] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*, 2016, pp. 565–571.

[4] Qihang Yu, Lingxi Xie, Yan Wang, Yuyin Zhou, Elliot K. Fishman, and Alan L. Yuille, "Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 8280–8289.

[5] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[6] Vicent Caselles, Ron Kimmel, and Guillermo Sapiro, "Geodesic active contours," *International Journal of Computer Vision*, vol. 22, no. 1, pp. 61–79, 1997.

[7] Sida Peng, Wen Jiang, Huaijin Pi, Xiuli Li, Hujun Bao, and Xiaowei Zhou, "Deep snake for real-time instance segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2020, pp. 8533–8542.

[8] Tianwei Ni, Lingxi Xie, Huangjie Zheng, Elliot K Fishman, and Alan L Yuille, "Elastic boundary projection for 3d medical image segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2109–2118.

[9] Lihong Guo, Yueyun Liu, Yu Wang, Yuping Duan, and Xue-Cheng Tai, "Learned snakes for 3d image segmentation," *Signal Processing*, vol. 183, pp. 108013, June 2021.

[10] Jun Ma, Jian He, and Xiaoping Yang, "Learning geodesic active contours for embedding object global information in segmentation cnns," *IEEE Transactions on Medical Imaging*, vol. 40, no. 1, pp. 93–104, 2020.

[11] Yuan Xue, Hui Tang, Zhi Qiao, Guanzhong Gong, Yong Yin, Zhen Qian, Chao Huang, Wei Fan, and Xiaolei Huang, "Shape-aware organ segmentation by predicting signed distance maps," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 12565–12572.

[12] Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, Francois Goulette, and Leonidas Guibas, "KPConv: Flexible and deformable convolution for point clouds," in *IEEE International Conference on Computer Vision*, Oct. 2019, pp. 6411–6420.

[13] Chufeng Tang, Hang Chen, Xiao Li, Jianmin Li, Zhaoxiang Zhang, and Xiaolin Hu, "Look closer to segment better: Boundary patch refinement for instance segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2021, pp. 13926–13935.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[15] Ping Hu, Bing Shuai, Jun Liu, and Gang Wang, "Deep level sets for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2300–2309.

[16] Eli Gibson, Francesco Giganti, Yipeng Hu, Ester Bonmati, Steve Bandula, Kurinchi Gurusamy, Brian Davidson, Stephen P. Pereira, Matthew J. Clarkson, and Dean C. Barratt, "Automatic multi-organ segmentation on abdominal CT with dense v-networks," *IEEE Transactions on Medical Imaging*, vol. 37, no. 8, pp. 1822–1834, 2018.

[17] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior, "The cancer imaging archive (TCIA): Maintaining and operating a public information repository," *Journal of Digital Imaging*, vol. 26, no. 6, pp. 1045–1057, July 2013.

[18] Zhoubing Xu, Christopher P. Lee, Mattias P. Heinrich, Marc Modat, Daniel Rueckert, Sebastien Ourselin, Richard G. Abramson, and Bennett A. Landman, "Evaluation of six registration methods for the human abdomen on clinically acquired CT," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 8, pp. 1563–1572, 2016.