Context Aware Convolutional Neural Networks for Segmentation of Aortic Dissection

Authors removed for review

Editors: Under Review for MIDL 2020

Abstract

Three-dimensional (3D) reconstruction of patient-specific arteries is necessary for a variety of medical and engineering fields, such as surgical planning and physiological modeling. These geometries are created by segmenting and stacking hundreds (or thousands) of two-dimensional (2D) slices from a patient scan to form a composite 3D structure. However, this process is typically laborious and can take hours to fully segment each scan. Convolutional neural networks (CNNs) offer an attractive alternative to reduce the burden of manual segmentation, allowing researchers to reconstruct 3D geometries in a fraction of the time. We focused this work specifically on Stanford type B aortic dissection (TBAD), characterized by a tear in the descending aortic wall that creates two channels of blood flow: a normal channel called a true lumen and a pathologic new channel within the wall called a false lumen. While significant work has been dedicated to automated aortic segmentations, TBAD segmentations present unique challenges due to their irregular shapes, the need to distinguish between the two lumens, and patient to patient variability in the false lumen contrast. Here, we introduced a variation on the U-net architecture, where small stacks of slices are inputted into the network instead of individual 2D slices. This allowed the network to take advantage of contextual information present within neighboring slices. We compared and evaluated this variation with a variety of standard CNN segmentation architectures and found that our stacked input structure significantly improved segmentation accuracy for both the true and false lumen by more than 12%. The resulting segmentations allowed for more accurate 3D reconstructions which closely matched our manual results.

Keywords: Aortic dissection, segmentation, convolutional neural networks, deep learning, 3D reconstruction

1. Introduction

Reconstructing three-dimensional (3D) aortic geometries from computed tomography angiography (CTA) scans is important in several fields of study including vascular disease visualization, surgical planning, and physiological modeling (Meszaros et al., 2000). A single patient CTA scan of the aorta is comprised of a vertical stack of hundreds or even thousands of two-dimensional (2D) slices, depending on the resolution of the scan. To reconstruct 3D geometries, researchers segment the CTA scan which involves identifying and highlighting the anatomic region of interest within each slice (Pham et al., 2000).

In this work, we focus specifically on segmenting Stanford type B aortic dissections (TBADs), defined as a pathologic separation of the layers of the aortic wall. When a tear occurs in the aortic intima, pressurized blood is allowed to propagate within the aortic wall, creating a blood flow channel called the false lumen that differs from the normal channel of flow referred to as the true lumen (Meszaros et al., 2000). TBAD presents unique difficulties
during segmentation due to poor contrast in the false lumen seen in some patients resulting from thrombus formation, irregular shapes, and the need to distinguish between the two lumens (Kovács et al., 2006; Erbel et al., 1993) (figure 1). Despite the availability of commercial segmentation packages, segmenting TBAD slices is a tedious process, often requiring hours of work for a single geometry.

Figure 1: Examples of TBAD slices from three different geometries with a) a well-contrasted false lumen, b) a poorly contrasted false lumen, and c) an irregularly shaped false lumen.

1.1. Related work

With the rapid increase of deep learning research and development, convolutional neural networks (CNNs) have become a useful tool to automatically segment aortic scans (Noothout et al., 2018; López-Linares et al., 2017; Graffy et al., 2019). In 2015, Long et al. (2015) developed the fully convolutional network (FCN) which vastly improved the state-of-the-art for image segmentation by allowing CNNs to output dense image segmentations on arbitrarily sized inputs. Similarly, in 2015, Xie and Tu (2015) developed a deep learning framework called Holistically-nested Edge Detection that improved edge and boundary detection in images, an important component of aortic segmentation.

More recently, several works have adapted the FCN to automatically segment medical images (López-Linares et al., 2017, 2018). One of the most commonly used adaptations is 2D U-net, a deep CNN that allows pixel by pixel predictions for 2D image segmentations on smaller medical datasets (Ronneberger et al., 2015). This task is particularly important as large numbers of medical imaging datasets are difficult to acquire due to cost, patient safety, and strict patient privacy laws. Several modifications of 2D U-net adapted for 3D volumetric data are also widely used including 3D U-net and V-net (Çiçek et al., 2016; Milletari et al., 2016). Both of these architectures input whole volumes of medical image data to allow better contextual information for each segmentation. Another promising method for multiclass segmentation is called cascaded networks, whereby two or more CNNs are applied in series to segment multiclass images (Li et al., 2018b; Cao et al., 2019). Several works have recently used U-net for aortic aneurysm segmentation from CTA images with promising results (Zheng et al., 2018; Chen et al., 2019). While there has been significant work on deep learning for aortic segmentation (López-Linares et al., 2018; Zheng et al., 2018), the number of articles related to TBAD segmentation remains limited (Li et al., 2018a; Cao et al., 2019; Li et al., 2018b). In the work by Cao et al. (2019), the authors used
a 3D U-net to segment the true and false lumen. Similarly, Li et al. (2018a,b) developed a new architecture that relied on cascaded CNNs. Both of these works showed promising results for circular and well-contrasted aortas, but there remains a need to better distinguish between the true and false lumen for irregular shapes and low contrast CTA scans by incorporating knowledge of contextual information in the vertical direction.

1.2. Contributions
In this work, we proposed and tested a variation on U-net, called stacked U-net (SU-net), where we incorporate small vertical image stacks into the input data structure. We compared the performance of SU-net to three of the most common aortic segmentation architectures including 2D U-net, 3D U-net, and V-net using the Dice coefficient (DC) and Jaccard coefficient (JC) for image comparisons as well as the Hausdorff distance (HD) for 3D geometry comparisons. Our results showed significant improvements in segmentation accuracy depending on the vertical stack size. This work can improve TBAD segmentation accuracy and may be useful for other volumetric segmentations, alleviating the burden of manual segmentation by researchers and clinicians.

2. Methods
2.1. Patient dataset
21 adult patient CTA TBAD scans were retrospectively acquired from the LOCATION REMOVED FOR REVIEW under International Review Board (IRB) approved protocols. Patient CTA scans were in Digital Imaging and Communications in Medicine (DICOM) format and generally contained between 500 and 1,400 axial slices, where each slice captured the aortic region and the surrounding tissue in a 2D plane. All patient datasets were deidentified during retrieval. Two patient datasets contained a slice thickness of 1 mm, and the remaining 19 patient datasets contained a slice thickness of 0.5 mm. All DICOM slices were 512x512 pixels. Slices below the aortic bifurcation and above the aortic arch were removed resulting in a total of 12,355 slices between all 21 patients (figure 2 - step 1).

2.2. Ground truth segmentations
To obtain ground truth segmentations, each scan was segmented using a commercial image segmentation package called Materialise Mimics (Leuven, Belgium). This process involved identifying and highlighting the aortic domain in each slice and distinguishing between the true and false lumen. In well-contrasted regions, automated thresholding and dynamic region growing tools in Mimics were able to capture the aorta with minor revisions, but in regions of poor contrast, we found that these tools were unsuccessful and frequently captured significant amounts of background and surrounding tissue (supplemental figure 7). In these cases, fully manual segmentations were required. For high-resolution scans with 0.5 mm slice thickness, this process often took 10 to 12 hours per patient. Each segmentation was confirmed by a cardiac surgeon specializing in TBAD from the LOCATION REMOVED FOR REVIEW (figure 2 - step 2).
Figure 2: Automated segmentation pipeline from image retrieval to 3D reconstruction. The CNN architecture in step 3 is a generic U-net based architecture. Several different U-net based architectures were used throughout this work.

2.3. Data preprocessing

To preprocess the images, we converted each raw, unsegmented DICOM slice to an 8-bit gray scale image and brightened the image for better visualization of the aorta with Mimics. The labelled data from Mimics allowed us to create three-class one-hot-encoded images representing the true lumen, false lumen, and background. All slices were downsampled to 128x128 pixels with linear interpolation to reduce training time and memory requirements. Unsegmented images were used as inputs, and the one-hot-encoded images were used as labels in our networks. We tested the impacts of data augmentation by shifting and rotating the images as well as adding Gaussian noise. However, we found that these methods did not improve results and were not used.

2.4. Network architectures and training

We trained and tested our model with several different architectures, based on U-net, a popular CNN for medical image segmentation (figure 2 - step 3). U-net is a deep CNN that contains a contraction and expansion section. In the contraction section, convolutions and max pooling operations are used to reduce the image size and extract features. During expansion, images are up-sampled, and at each up-sampling layer, features from the contraction section are concatenated with the up-sampled layers (Ronneberger et al., 2015). This method is based on the skip connection architecture adopted from the FCN allowing features learned during the contraction section to be used during expansion, leading to better localized and general segmentations (Long et al., 2015).

The common architectures we used for comparisons were 2D U-net (Ronneberger et al., 2015), 3D U-net (Çiçek et al., 2016), and V-net (Milletari et al., 2016). We modified the 2D U-net by reducing the number of filters at each layer to better fit our dataset (figure 3). 3D U-net is a modification of 2D U-net where the entire geometry is passed into the CNN instead of individual slices, allowing better contextual information. V-net is a modification of 3D U-net where a residual function is learned at each convolution layer. When using 3D U-net and V-net, geometries were linearly downsampled to a depth of 128 pixels.
Figure 3: A description of 2D and 3D SU-nets with a sample stack size of three slices. The numbers above each block indicate the number of filters. Each slice was inputted with neighboring slices above and below. As shown in the slice stacks, slices were used for multiple inputs depending on the stack size. In 2D SU-net, when the stack size was one slice, the network is equivalent to the 2D U-net described by (Ronneberger et al., 2015) with fewer filters at each layer. When the stack size was greater than one slice, the stack size was set as channels. Thus, the input data array had a 3D shape (image height x image width x image stack size). For 3D SU-net, the implementation was similar to the 2D SU-net where 2D convolutions were replaced with 3D convolutions. The input array size was four-dimensions (image height x image width x image stack size x 1). Because the volume depth was shallow, our design did not alter the depth with max pooling or up-convolution operations.
2.4.1. Stacked 2D and 3D U-net

In the next set of architectures, we introduced a stacked slice input method that is fed into a modified 2D and 3D U-net. We refer to these architectures as 2D SU-net and 3D SU-net, respectively. In these architectures, each slice was passed into the network with its neighboring slices above and below. The network then estimates the segmentation for the entire stack. During training, the loss is computed across all outputs, but during testing, only the output from the middle slice was used. Therefore, each slice was used in multiple input data stacks depending on the stack size during training. Figure 3 shows an example of an input data structure with a stack size of three slices, and stacks of 5, 7, 9, 11, and 13 slices were also tested. With this method, we hypothesized that the network could better incorporate vertical contextual information. However, unlike 3D U-net and V-net, SU-net maintains a large training set and does not require downsampling in the vertical direction. This comes at the expense of longer training times, as each slice is used across multiple input stacks.

In 2D SU-net, input slices within each stack were concatenated to form channels, and the network used 2D convolutions. In 3D SU-net, the volume depth was set to the image stack size, and the network parameters were similar to the 2D U-net where 2D convolutions were replaced with 3D convolutions. Because these networks had a shallow depth equal to the stack size, the depth dimension was not altered at max pooling or up-convolution steps (figure 3). Finally, we tested the effects of cascading two SU-nets together. The output from the first SU-net contained a segmented whole aorta without distinguishing between the true and false lumen. After the first SU-net, the background of the output images was masked out so that the second SU-net could focus solely on the aorta. These images were passed into the second SU-net that we used to distinguish between the true and false lumen. Only the top performing SU-net on the whole aorta segmentation was used as the first network, and the top performing SU-net for true and false lumen differentiation was used for the second network. The cascaded SU-nets were trained individually. The pipeline for cascaded SU-nets is shown in supplemental figure 8.

During training, we used a DC loss function, and we tested each model using the DC (supplemental equation 1), JC (supplemental equation 2), and HD (supplemental equation 3) on each geometry in the test set. Further details on training and analysis metrics for all networks can be found in Appendix A.

3. Results

Three commonly used CNN segmentation architectures - 2D U-net, 3D U-net, and V-net - were compared to our introduced 2D SU-net and 3D SU-net. All output metrics were computed by averaging the results of three-fold cross-validation. Our results showed that a stack size of three slices performed optimally for the 2D SU-net. However, all stack sizes greater than one demonstrated improved performance (figure 4). Additionally, 2D SU-net improved performance compared with 3D SU-net (supplemental figure 9), 3D U-net, and V-net. The cascaded SU-nets used a stack size of three as that input structure proved to be optimal for the whole aorta as well as the true and false lumen. We saw approximately equal results for the cascaded SU-net as our top performing SU-net (table 1).
Table 1: Dice coefficients, Jaccard coefficients, and Hausdorff distances (mm) for the 3D U-net, V-net, 2D U-net, optimal 2D SU-net (three slices), and cascaded 2D SU-net.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Whole aorta</th>
<th>True lumen</th>
<th>False lumen</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DC</td>
<td>JC</td>
<td>HD</td>
</tr>
<tr>
<td>3D U-net</td>
<td>0.90</td>
<td>0.81</td>
<td>30.59</td>
</tr>
<tr>
<td>V-net</td>
<td>0.80</td>
<td>0.68</td>
<td>30.14</td>
</tr>
<tr>
<td>2D U-net</td>
<td>0.91</td>
<td>0.84</td>
<td>5.80</td>
</tr>
<tr>
<td>2D SU-net (stack size of 3)</td>
<td>0.92</td>
<td>0.86</td>
<td>1.42</td>
</tr>
<tr>
<td>Cascaded 2D SU-net (stack size of 3)</td>
<td>0.92</td>
<td>0.86</td>
<td>1.44</td>
</tr>
</tbody>
</table>

3D U-net and V-net exhibited the lowest segmentation accuracy on the whole aorta, true lumen, and false lumen. The 2D U-net and 2D SU-net showed similar performance for the whole aorta, but the ability to distinguish the true and false lumen improved significantly as the stack size increased to three. Figure 5 shows sample segmentation results using the 2D U-net, 3D U-net, V-net, and SU-net with three slices. From the images, it is clear that SU-net better distinguishes the true and false lumen compared with all other architectures tested. Further results demonstrating the networks ability to segment irregular shaped aortas and poor contrast in the false lumen are shown in supplemental figure 10. We note that the true lumen often over-predicts and the false often under-predicts the segmented area in V-net, 3D U-net, and 2D U-net. Finally, to compare whole geometries, we reconstructed 3D models of the aortas by combining segmentation results from each individual CTA slice. Visual inspection as well as computed HD indicated accurate matching compared with manual segmentations using 2D SU-net (figure 6, table 1).
Figure 5: Sample segmentation results. a) In some well-contrasted aortas, we found that V-net, 3D U-net, and 2D U-net were able to capture the whole aorta but could not distinguish between the true and false lumen. b) SU-net is the only architecture that was able to capture slices with somewhat irregular shapes and poorer contrast in the false lumen.

4. Discussion

Segmenting TBAD CTA scans presents unique difficulties due to the need to distinguish between the true and false lumen, patient variability in the degree of false lumen contrast, and irregular aortic shapes (Kovács et al., 2006; Erbel et al., 1993). During TBAD segmentations, researchers can often distinguish between the true and false lumen based on their unique shapes. However, in some instances, the shapes of the lumens are similar and difficult to distinguish by observing the single slice in isolation (figure 1). Therefore, researchers often track the geometries from slices above and below to better distinguish between the two lumens. These manual segmentation experiences led us to hypothesize that vertical contextual information was crucial for accurate TBAD segmentations. Thus, in this work, we proposed a stacked input data structure, called 2D and 3D SU-net, consisting of small sets of neighboring slices, and we inputted these stacks into a modified U-net architecture. We compared the results to several popular CNN architectures for aortic segmentation and found that 2D SU-net significantly improved segmentation accuracy in the true and false lumen compared with the 2D U-net, 3D U-net, and V-net (Ronneberger et al., 2015; Çiçek et al., 2016; Milletari et al., 2016). 3D U-net and V-net also contain vertical contextual information but performed significantly worse than our proposed structures. One possible explanation for this reduced performance is that the sample size is limited to the number of geometries instead of the number of images. Additionally, we needed to downsample in the vertical direction for 3D U-net and V-net, whereas downsampling was not needed.
in 2D SU-net or 3D SU-net. Downsampling in the vertical direction loses important and significant variations seen in aortic dissections that occur along their depth.

Figure 6: Three sample geometries generated with 2D SU-net (stack size of three slices) compared with geometries from manual segmentations.

4.1. Study limitations

One limitation is that manual segmentations often vary depending on the software and the person performing the segmentations. In fact, Cao et al. (2019) showed that segmentations on the same geometry by two different experts had DCs ranging between 0.92 and 0.94 indicating that our networks may be limited by their labels. Another limitation was the downsampling of each image in the horizontal direction to improve training speed and reduce memory restrictions. Although our 3D reconstructions were accurate with 2D SU-net, future work can train and test without downsampling in the horizontal and vertical direction on more powerful GPUs.

5. Conclusions

In this work, we developed a deep learning pipeline that takes advantage of spatial context to robustly automate TBAD segmentations. We found that inputting small stacks of slices, called 2D SU-net, improves segmentation accuracy compared with other commonly used CNN architectures.

Acknowledgments

REMOVED FOR REVIEW

References

Long Cao, Ruiqiong Shi, Yangyang Ge, Lei Xing, Panli Zuo, Yan Jia, Jie Liu, Yuan He, Xinhao Wang, Shaoliang Luan, et al. Fully automatic segmentation of type b aortic


