UNCERTAINTY-BASED ADAPTIVE DATA AUGMENTATION FOR ULTRASOUND IMAGING ANATOMICAL VARIATIONS

Edward Chen, Howie Choset, and John Galeotti

The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, 15213, USA

ABSTRACT

Data augmentation remains to be a simple and inexpensive method for generalizing across unseen domains. Current data augmentation methods for ultrasound imaging involve simple image transformations - rotations, flips, skews, and blurs - but are not able to adapt to the current state of the deep learning model. We present the first online adaptive data augmentation method that is able to generate synthetic training data on-the-fly, enabling the model to adapt to countless spatial deformations. Our proposed method leverages prior work on uncertainty quantification to understand the model's weaknesses at any given stage. Our method is also able to then "spot-augment" subsets of regions within the ultrasound image, all in a real-time manner. We also show that our proposed method is able to perform significantly better in out-of-true-training distributions, when compared against models trained on the same dataset.

Index Terms— data augmentation, bayesian modelling, deep learning, ultrasound

1. INTRODUCTION

In the case of high-tempo, traumatic scenarios on the battlefield, real-time ultrasound (US) imaging serves as an enabler for countless possible robotic interventions. Having the ability to automatically segment anatomical landmarks in the body, such as arteries, veins, ligaments, and veins, for percutaneous procedures remains to be a difficult task when considering the countless domains across body types, potential traumatic injury scenarios, and imaging artifacts. Collecting data spanning all, or many, of the cases is tremendously time-consuming and expensive. *A key motivation of this work is to propose a method for enhancing deep learning models' generalization capabilities by generating synthetic data which is transformed in a manner designed to account for various body types, injury scenarios, and imaging features.*

A great amount of the focus in the medical imaging community has been towards engineering improved deep learning architectures, with the goal of learning improved features [1, 2, 3]. On the other hand, the data augmentation is a relatively simple and commonly used method for generating synthetic data to account for invariances in the data [4]. Unfortunately, many data augmentation procedures currently rely on domain expertise and manual tuning to be effective [5, 6, 7]. Under the manually designed image augmentations, the generated synthetic data may produce countless simple images which limit the model's ability to learn generalized features [5, 6]. In addition, many of the commonly used medical imaging data augmentation strategies still remains to be basic transformations such as flipping, rotating, shifting, and blurring. Although advanced data augmentation strategies do exist for natural images [5, 8, 6, 7], many of them are designed for cases which aren't as relevant for medical images. Our goal is to research a learning-based data augmentation method which can adaptively generate augmented images for learning invariances across various anatomical shapes and imaging artifacts.

In this paper, we propose a novel data augmentation technique for ultrasound images. Specifically, the method is designed to adaptively train a semantic segmentation network such that it's able to generalize to different anatomical variations, artifacts, and potentially backgrounds. We first employ a Bayesian temporal-based segmentation network which is able to output both the segmentation and epistemic uncertainty [9, 10, 11] maps. The epistemic uncertainty maps, which we use to obtain knowledge of the model's spatial weaknesses, are then passed into an augmentation module. Inspired by [12], the augmentation module then initializes a set of fiducial points across the training image and uses an agent feedforward neural network, with the unceraint maps as input, to create a final moving state for the fiducial points. Similar to [7], we use a similarity transformation based on the moving least squares transformation method [12]. The constructed images are designed to "spot-augment" the images in such a way that they challenge the models precisely where they already have learned good features for, and display less uncertainty. We train the agent neural network based on how much it's able to challenge the base segmentation network, which is measured by the loss on the augmented images. All of these steps are embedded within a unified real-time training framework, shown in Figure 1. Our primary contributions are: (1) incorporating the epistemic uncertainty map outputs from a Bayesian segmentation network for "spot-augmenting" areas where the model is currently strong, (2) an adaptive training pipeline which also incorporates the spatial relationship nuances between ultrasound imaging anatomical landmarks, and (3) experiments illustrating how the method is able to better enable medical segmentation networks to generalize across various vessel shapes and imaging features.

2. METHODS

2.1. Semantic Segmentation Neural Network

Our final task is multi-class, multi-instance segmentation of arteries, veins, ligaments, and nerves in ultrasound images. The deep learning network architecture we use is a dropout-based Bayesian formulation [9, 11] of the 3D U-Net encoder-decoder architecture [13]. The network consists of four convolutional blocks on the encoder side, with matching pairs on the decoder side. Each block consists of an input layer followed by 2 pairs of the following: convolutional layer, batch normalization, and ReLU. Within each block, we empirically determined to place a single dropout layer before the output layer, as opposed to other possible variations [14]. The model outputs two values (represented below), for both the predicted mean, $\hat{\mu}$, the segmentation map, and predicted variance, $\hat{\sigma}^2$, which we use for the epistemic uncertainty map:

$$[\hat{\mu}, \hat{\sigma}^2] = f^{\hat{W}}(x) \tag{1}$$

where f is the Bayesian 3D U-Net, in this case, parameterised by model weights $\hat{W}$. The epistemic uncertainty maps are obtained using test-time stochastic forward passes, also referred to as Monte Carlo dropout [11]:

$$\frac{1}{T}\sum_{t=1}^{T}(\hat{\mu}_t - \bar{\mu})^{\otimes 2} \tag{2}$$

where T is the total number of Monte Carlo samples and $\bar{\mu} = \sum_{t=1}^{T} \frac{\hat{\mu}_t}{T}$. Despite not actively using the logits variance output for computing the aleatoric uncertainty, which is the statistical uncertainty inherent in the data, empirical trials and [11] both illustrated that having the variance output was still necessary. Without the logits variance output, we empirically found that the epistemic uncertainty tended to overcompensate for that fact and obtained poor performance.

2.2. Augmentation Module

The augmentation module is responsible for generating the synthetic images for further training. It does so by using a variation of the moving least squares deformation method [12, 15] which first generates a set of control points p around the border of the overall image as well as of the individual anatomical classes. The immediate next step is then to also generate a set of deformed control points q. The points q govern the image deformation using the best affine transformation $l_v(x)$ which minimizes the following [12] :

$$\sum_i w_i |l_v(p_i) - q_i|^2 \tag{3}$$

where w_i represents the set of deformation weights, which are dependent on the point of evaluation v [12].

To generate the set of points q, we use a convolutional neural network, which we refer to as the augmentation agent neural network, which takes as input the epistemic uncertainty map and outputs a set of directions to shift the original points p. We output the directions rather than individual shift magnitudes because we found it empirically simpler to train and to avoid potential negative image augmentations. Currently, we only use straight top-down and left-right directional shifts but this can be expanded into additional degrees of freedom as well.

We then apply points p and q to the original training images in the batch to output a new batch with transformed versions of the images. The moving least squares image deformation method is notoriously known to take a long time to compute [15]. To enable it for real-time capabilities in our method, we rearrange the mathematical relationships in such a way that we can pre-compute most of the expensive matrix multiplications prior to the training process. We first represent $l_v(x)$ as an affine transformation with a linear transformation matrix, M, and a translation value, T. According to the proof in [12], we can solve for T by rearranging the relationship as such:

$$T = q_* - p_* M \tag{4}$$

where q_* and p_* are the weighted centroids used for the linear moving least squares deformation, represented as such:

$$p_* = \frac{\sum_i w_i p_i}{\sum_i w_i} \tag{5}$$

$$q_* = \frac{\sum_i w_i q_i}{\sum_i w_i} \tag{6}$$

We re-formulate the above relationships by splitting our initial and final control points, p and q, respectively into a set for the border of the ultrasound image, p_B and q_B, and another set for the anatomical classes within, p_i and q_i. The reason we have a set of dedicated control points for the border is to prevent the sides of the image from folding in, creating holes in the image. We show an example of a proper deformation in Figure 2. Furthermore, since the borders of the images stay constant throughout the training process, we can re-arrange the equation for q_* as such:

$$q_* = \frac{\sum_i w_i q_i + w_B q_B}{\sum_i w_i + w_B} \tag{7}$$

where p_* would follow a similar structure. We further pre-compute the values of w_i, represented as:

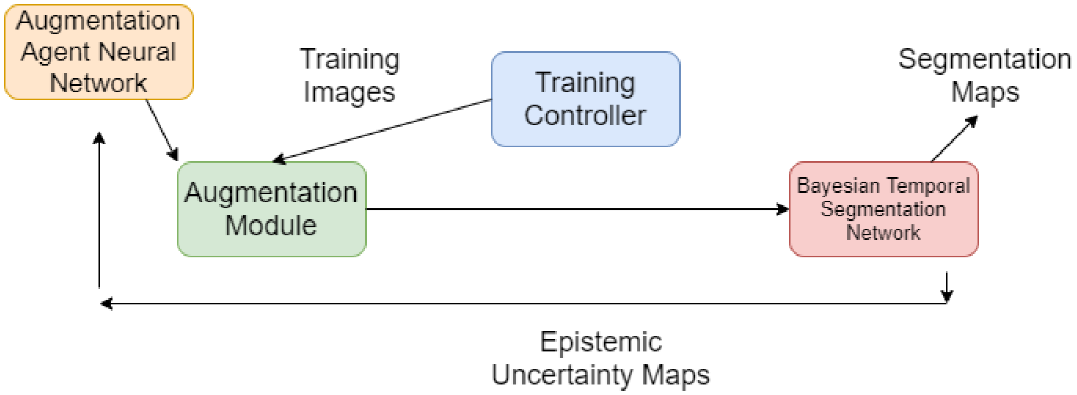

Augmentation Agent Neural Network
Training Images
Training Controller
Segmentation Maps
Augmentation Module
Bayesian Temporal Segmentation Network
Epistemic Uncertainty Maps

Fig. 1. Flow diagram of the overall uncertainty-based augmentation pipeline. The uncertainty maps from the segmentation network get passed into the agent neural network, which then generates the control points for deforming the image in the augmentation module. Those are then fed back into the training process.

$$w_i = \frac{1}{|p_i - v|^{2\alpha}} \tag{8}$$

Additional values for computing the affine transformation represented in [12], which do not depend on each individual image, are also pre-computed. Finally, to customize the speed of the image deformations, we set as hyperparameters the count of the control points p_i and p_B.

2.2.1. Augmentation Agent Neural Network

The augmentation network is a simple, custom 3D convolutional neural network which consists of 5 convolutional blocks. The first 3 convolutional blocks each consist of, sequentially, a 3D convolutional layer, a batch normalization layer, ReLu activation, and then a max pooling layer. The last 2 convolutional blocks do not have the max pooling layer. The final convolutional block outputs to a fully connected layer which then outputs a set of points classifying whether to go up or down for each control point.

To train the augmentation agent neural network, we generate a random set of points, signalling up or down for the deformations. We then compute the moving least squares image deformations [12] for both the agent-generated and randomly-generated points and compute the segmentation loss for both sets. If the agent-generated points resulted in a lower segmentation loss, we assume the randomly-generated points as more difficult and assign those as the label for training this network. If the randomly-generated points resulted in a lower loss, however, we assign the opposite direction of the agent-generated points as the label similar to [7]. The rest of the training process is completed as normal. A diagram detailing the entire pipeline is shown in Figure 1.

3. EXPERIMENTS AND RESULTS

3.1. Data

The data was acquired from a CAE Blue Phantom femoral vascular access lower torso ultrasound training model. To collect the data, we used the Fukuda Denshi portable (i.e. Point of Care Ultrasound, *POCUS*) scanner with a 5-12 MHz transducer imaging a maximum depth of 5cm and 10cm, respectively. The main dataset, which we refer to as *standard-POCUS*, used for training consisted of 12 sequences, each containing 50 frames totalling 600 frames. The other dataset, which we refer to as *higher-depth*, consisted of 3 sequences, with each sequence containing 50 frames totalling 150 frames. It was collected using a higher depth, which deformed the anatomy. All classes (arteries, veins, ligaments and nerves) were present in (at least some frames of) every sequence in every category, and each sequence was annotated by a single expert.

3.2. Training Details

Due to the memory restrictions presented by our precomputation details, we only train with a batch size of 1, with 8 images within each sequence. Each of the images are resized to 256x256 pixels. Also a consequence of the memory-hungry characteristic of our method, we do not perform initial significant offline data augmentations to the training set. For training, we used the aforementioned Bayesian temporal segmentation network with a stochastic version of the cross entropy loss, similar to [11]. To compute the uncertainty maps, use 10 Monte Carlo samples [11]. We set the learning rate to 10^{-3} and trained the network until the validation loss did not improve for 5 epochs. Lastly, the validation set consisted of about 10% of the original training dataset.

3.3. Experiments

We compare our proposed pipeline against the same Bayesian temporal segmentation network across 2 cases: 1) without any spatial data augmentations, so using the same initial training set as our pipeline (*non-aug*), and 2) using a full set of spatial augmentations (*full-spatial-aug*). The full set of spatial augmentations consisted of horizontal and vertical flips, gamma adjustment, Gaussian noise addition, Gaussian blurring, Bilateral blurring, cropping, affine transformations, and shear transformations. In terms of the control points, we use 128 and 2 points for the border and class, respectively. To limit the up/down directions, we use 20, 40, 80, 40 pixels as the maximum shifts for the arteries, veins, ligaments, and nerves, respectively. All of the experiments are evaluated against *standard-POCUS* and *higher-depth* across 2 trials. The results are shown in Table 1.

3.4. Metrics

To evaluate our experiments, we computed the following metric - region similarity. Region similarity measures how similar the ground-truth label and predicted images are in terms of their inner contour regions. We calculate region similarity as how it is used in [16], using the intersection-over-union between the ground-truth label and predicted image.

Approach	*standard-POCUS*	*higher-depth*
non-aug	.534 ± .010	.238 ± .004
full-spatial-aug	.564 ± .012	.453 ± .009
Ours	**.571 ± .019**	.419 ± .013

Table 1. Region similarity metric comparing the best-performing methods, across the 2 datasets. The bolded values represent our methods which surpassed the baseline results.

4. DISCUSSION

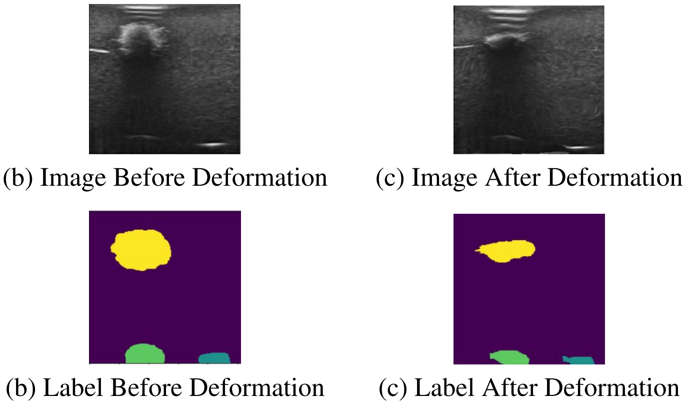
Based on the results presented in Table 1, our pipeline is able to enhance the generalization performance of segmentation models due to the diversity of images which it generates. We would like to emphasize that the performance of our method in Table 1 only uses the original training set, without any prior spatial augmentations - which is a 28x difference. Even with the lack of augmentation, the method is able to approach the out-of-domain generalization abilities of the fully augmented version, reflected by the score on *higher-depth*. Some samples of the image deformations are shown in Figure 2. In terms of computation time, the pipeline is able to completely generate the augmented batch of images in, on average, 10.576 seconds. To completely perform the moving least squares image deformation [12], it takes, on average, 5.201 seconds - which we have to compute twice for both the images and the labels. Those times are all based on the 128 border control points we use, which affects it the most. We noticed that switching it to 256 border points nearly doubles the computation time, hence why we reduced to 128. We reduced the number of class control points to 2, for each present class, for the same reason. The reason for the greater number of border points is because we empirically noticed that reducing the number of border points too much created visible holes near the borders and made the images look unrealistic.

The biggest limitation for this pipeline right now is the aforementioned computational time and memory requirements. In order to pre-compute all of the matrices mentioned, it takes roughly 260 MB of hard drive space for each image, hence our memory restrictions during training it (not being able to use the full spatial augmentations). As with the computation time, this memory requirement also decreases in proportion to the number of border points.

Another potential limitation is that this method may miss image deformations in areas of the image where the segmentation ground-truth label does not cover. As a result, there may be cases where traditional, uniform, spatial augmentations will maintain better generalization capabilities.

5. CONCLUSION AND FUTURE WORK

To the best of our knowledge, we have presented the first online, adaptive data augmentation pipeline for adapting to different anatomical variations with ultrasound imaging. According to our results, the pipeline is able to enhance the generalizability of the segmentation network, but with a cost due to the memory and computational time. In the future, we plan to improve upon this by expanding the degrees of freedom with which the pipeline is able to modify the features in the ultrasound images. We also plan to improve the computational memory and time requirements of this overall pipeline, in order to make it more feasible for consistent training.


(b) Image Before Deformation

(c) Image After Deformation

(b) Label Before Deformation

(c) Label After Deformation

Fig. 2. Example ultrasound image undergoing the image deformation. The top images visualize the original and new images, whereas the bottom images visualize the same images' labels.

6. COMPLIANCE WITH ETHICAL STANDARDS

This is a numerical simulation study on phantoms for which no ethical approval was required.

7. ACKNOWLEDGMENTS

These awards helped us in gathering data and designing initial algorithms: NIH 1R01EY021641, DOD awards W81XWH-14-1-0371 and W81XWH-14-1-0370, and NVIDIA Corporation GPU donations. This present work was sponsored in part by US Army Medical contracts W81XWH-19-C0083, W81XWH-19-C0101, and W81XWH-19-C-0020. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. Specifically, it used the Bridges system, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC). We would also like to thank our collaborators at the University of Pittsburgh, Triton Microsystems, Inc., Sonivate Medical, and URSUS Medical LLC. Potential conflicts of interest include members of the organizations above and of Carnegie Mellon University. We do not anticipate any conflicts of interest beyond that.

8. REFERENCES

[1] F. Milletari, N. Navab, and S. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *International Conference on 3D Vision*, 2016.

[2] Z. Zhou et al., "Unet++: A nested u-net architecture for medical image segmentation," *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 2018.

[3] L. Chen et al., "Attention unet++: A nested attention-aware u-net for liver ct image segmentation," in *IEEE International Conference on Image Processing*, 2020.

[4] C. Shorten and T. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, 2019.

[5] E. Cubuk et al., "Autoaugment: Learning augmentation policies from data," *arXiV*, 2018.

[6] E. Cubuk et al., "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[7] C. Luo et al., "Learn to augment: Joint data augmentation and network optimization for text recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[8] S. Lim et al., "Fast autoaugment," in *Advances in Neural Information Processing Systems*, 2019.

[9] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, 2016.

[10] Y. Gal, "Uncertainty in deep learning," *PhD Thesis*, 2016.

[11] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?," in *Advances in neural information processing systems*, 2017.

[12] S. Schaefer, T. McPhail, and J. Warren, "Image deformation using moving least squares," in *ACM SIGGRAPH 2006 Papers*, 2006.

[13] O. Çiçek et al., "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*, 2016.

[14] Y. Kwon et al., "Uncertainty quantification using bayesian neural networks in classification: Application to biomedical image segmentation," *Computational Statistics Data Analysis 142*, 2020.

[15] J. Lee and H. Byun, "Image deformation using moving least squares based on unequal grids," *International Journal of Software Engineering & Its Applications 6*, 2013.

[16] F. Perazzi et al., "A benchmark dataset and evaluation methodology for video object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.