Response to the reviewers

We appreciate the efforts by the editors as well as Reviewers 1 and 2 in assessing our manuscript.

Reviewer 2

The authors have made several important amendments to this manuscript. The results for the proposed technique have greatly improved from the 1st submission following the correction of a coding mistake. Although the revision and rebuttal have improved clarity, there remain major concerns about the fundamental basis for the proposed technique and its application in structural vs. functional images. Although these results are encouraging and the open-source package presented is promising, there is especially concern about widespread use of a data augmentation technique (1) using functional images as a basis for template construction and (2) without more rigorous evaluation against current techniques.

> *Again, we very much appreciate the time spent by the Reviewer in improving the manuscript. Please see below a point-by-point to the issues raised by the reviewers and note that textual changes in the manuscript are in blue font.*

Major comments

1. Structural vs. Functional Transformations

The authors concede that severe disease might necessitate using a structural basis for template-based data augmentation of functional images. The authors then argue that misalignments are not crucial to prediction and that defect shape is not an influential feature being learned. Admittedly, intensity variation is probably the most important feature of the data for assessing ventilation defect, and data augmentation need not preserve shape exactly to still provide benefit. Nevertheless the authors' premise is that shape is important for medical imaging in particular, and that data augmentation strategies for medical imaging should preserve "shape variation within the typical range exhibited by the population under study". This is in stark contrast to the rebuttal comment that defect shape is not important. If shape is not important, then the standard augmentation strategies would suffice for functional image analysis.

On the contrary, considering that shape may be important, what happens to the shape of these ventilation images after deforming through the template? The template shown at the center of Figure 4 is high-intensity inside the lung boundary and low-intensity outside, which seems to best represent the average healthy subject. If this is the case, then it seems that low-intensity ventilation defects must be mapped/warped/deformed outside of the "lung" boundary of the template. Accordingly, mapping from one subject to another may result in distortion of the lung boundary such that the boundary develops pockets/bumps/swirls. This may distort the shape and/or texture of the defects, such that the CNN may learn features that are not physiologically relevent or even typical. I maintain my initial concern that it is not meaningful to use functional images (with no discernible anatomical features) as a basis for template construction (especially

when the intended purpose of the template is to preserve anatomically/structurally relevant shape), and instead I will continue to suggest that the transformations obtained from structural template construction should be applied to both sets of images.

Importantly, there is a major difference between learned features at different size scales: voxelwise intensity, local texture, boundary smoothness, and large-scale anatomical structure. There may be little influence of anatomical factors on prediction of ventilation defect but there is very plausibly influence of defect morphology and texture. Unfortunately it is often difficult to quantitatively evaluate the relative importance of these various features at different size scales in a trained CNN, nevertheless it is ill-advised to disregard their importance. Otherwise, if intensity were the only important feature, a simple thresholding metric could be used for assessing ventilation defect--but this would produce noisy, rough defects also vulnerable to partial voluming. I find the authors' response on this concern inadequate without further evidence.

The utility of the novel template-based data augmentation strategy introduced in this manuscript stems from its ability to constrain shape variation within the population of interest. There are many medical imaging segmentation applications where this is potentially useful and, as the author suggests, this contribution could be applied to structural data, functional data or to other types of anatomical images. As we write in the introduction:

> While common approaches to data augmentation include the application of randomized simulated linear (e.g., translation, rotation and affine) or elastic transformations and intensity adjustments (e.g., brightness and contrast), we advocate a tailored paradigm to commonly encountered medical imaging scenarios in which data is limited but is assumed to be characterized by a population-wide spatial correspondence.

This is a generic claim which applies only to relevant image segmentation applications. Not every medical image segmentation problem is going to be facilitated by such template-based shape constraints. We would argue that functional lung image segmentation is one of those applications for which there is no evidence that such template-based shape information is useful for segmentation of ventilation defects. This should be fairly obvious from the fact that state-of-the-art methods produce results on par or exceeding those of human radiologists and from the reasonable performance of the relatively simple MRF-based technique (referred to as Atropos in the manuscript and figures). These state-of-the-art methods listed in the manuscript discard all shape information and only our previous method [8] considers basic spatial information (i.e., neighborhood label information). As we wrote in the subsection **Previous approaches from our group for lung and ventilation-based segmentation**,

> Unlike other methods that rely solely on intensity distributions, thereby discarding spatial information (e.g., K-means variants [9, 11] and histogram rescaling and thresholding [10]), our technique employs both spatial and intensity information for probabilistic classification.

Note that "our technique" is in reference to Atropos, our MRF-based approach [8] (described earlier in the paragraph) and not in reference to the methodology under consideration. As this might be a source of confusion, we changed "our technique" to "our previous MRF-based technique [8]" to make this point clear that we are not discussing the methodology proposed in the current manuscript.

However, to avoid the types of confusions described by the Reviewer, we re-analyzed the functional lung imaging by creating a template from the ventilation masks (instead of the ventilation images themselves) and re-evaluated the results. We also revised the figures accordingly. Please see revised Figures 1, 4, 6, and 8. Note that the differences are negligible relative to the previous iteration. We also revised the description of ventilation template construction in the text.

Additionally, the authors suggest that 2D processing may be accurate to a good degree, but the reason cited for avoiding 3D training of the functional CNN was computational efficiency. However, efficiency is not a valid excuse when CNN training is merely an upfront cost. The authors are already using 3D training on the structural CNN, but claim that 3D training for functional images is not justified because lung shape has no effect on the functional CNN performance. This justification is speculative, but more importantly disregards the breadth of spatial information considered by a CNN. If spatial information were not important, then intensity-based processing (thresholding) would suffice, but this is not the case. The authors' own data shows important spatial distinctions made by the CNN (e.g., the partial voluming artifact near the borders of the lung, as well as the spatial smoothness of estimated regions). The authors must then admit that improved spatial smoothness across multiple adjacent slices using 3D CNN vs. 2D CNN is certainly plausible. The authors make a point that 3D may perform better than 2D in the Discussion, but the aversion to 3D training of functional images does not appear justified especially given that the authors already demonstrate the ability to perform 3D training in structural images.

Although we still maintain that computational efficiency is a reasonable consideration, we neglected to mention the slice thickness associated with much of the functional data. In the **Image Acquisition** *subsection, we describe two sets of protocols. Protocol 1 uses isotropic spacing of 3.9 mm^3 whereas Protocol 2 uses an in-plane resolution of 2-4 mm with a slice thickness of 15 mm. The 2-D approach generates a model compatible for both protocols. We add relevant commentary to the manuscript that clarifies the additional rationale (low out-of-plane resolution) for using readily-available 2D convolutional models:*

> Even though the functional images are processed as 3-D volumes and a 3-D ventilation template is created for the template-based data augmentation, the generated U-net model is 2-D. This is due to lack of any discernible anatomical signatures available for learning especially for functional images obtained from Protocol 2 which have a slice thickness of 15 mm. This also makes model generation and prediction much faster. Previous work from members of our group [43] has shown that 2-D CNNs can achieve comparable performance as their 3-D analogs in certain problem domains.

Can the authors reference other work evaluating 2D vs 3D deep learning approaches in medical imaging? For example, this conference paper:

X Zhou et al, Performance evaluation of 2D and 3D deep learning approaches for automatic segmentation of multiple organs on CT images, SPIE Medical Imaging, 2018. https://doi.org/10.1117/12.2295178

In the context of the previous issue, we added the following paper:

> Cullen, N. C. and Avants, B. B. "**Convolutional Neural Networks for Rapid and Simultaneous Brain Extraction and Tissue Segmentation**" *Brain morphometry* 136, (2018): 13–36.

2. Evaluating the Template-Based Data Augmentation Technique

The authors have included a small study showing that the proposed data augmentation technique results in better performance compared to no augmentation. This is encouraging. However the authors do not compare to augmentation using the current standard techniques (e.g., affine/elastic transforms). It is already accepted that various augmentation strategies improve CNN performance, so it should be unsurprising that the proposed technique would provide benefit compared to no augmentation. But the novelty of the proposed template-based technique warrants comparison with existing methods. Most deep learning libraries, including Keras, offer built-in data augmentation techniques (e.g., in Keras, the imageDataGenerator class). It should be relatively straightforward to design a study evaluating the proposed data augmentation vs. the current standard, and would increase the impact of this study for the medical image processing community. It is not beyond the scope of this academic manuscript, especially given that the novel aspect of the proposed technique (and indeed, the title of the manuscript) is not merely the use of the U-Net architecture in medical imaging, but rather the medical-imaging-tailored template-based data augmentation technique.

We agree with this recommendation and have added to the previous footnote [1] outlining the results of the recommended experiment:

> Although the need for data augmentation techniques is well-established within the deep learning research community, we performed a smaller 2-D experiment to illustrate the potential gain using template-based data augmentation over augmentation using randomly generated deformable transforms. Training data consisted of 50 coronal proton lung MRI with lung segmentations. Data and scripts are available in the companion repository to the ANTsRNet package, called ANTsRNetExamples, which contains examples for available architectures. Accuracy, in terms of Dice overlap, achieved with template-based augmentation was left lung: 0.94 ± 0.02, right lung: 0.92 ± 0.04, and whole lung: 0.93 ± 0.03. Accuracy achieved without augmentation was left lung: 0.88 ± 0.13, right lung: 0.83 ± 0.21, and whole lung: 0.86 ± 0.16. Accuracy achieved with random deformation augmentation was left lung: 0.94 ± 0.03, right lung: 0.90 ± 0.06, and whole lung: 0.92 ± 0.04.

In addition, we added the following paragraph outlining limitations and future work relevant to this data augmentation approach:

> The template-based data augmentation strategy follows the generic observation in [48] where constrained augmentation to plausible data instances enhances performance over generic data augmentation. Although we find the presented framework to be generally useful for model training, further enhancements could increase utility. A template-based approach for continuous sampling of the population shape distribution could provide a potentially unlimited source of data for training. Also, further evaluation needs to be conducted to determine the performance bounds of these augmentation strategies (not just template-based) for a variety of medical imaging applications.

MINOR COMMENTS

1. Manuscript Structure - The authors disagreed with previous review on a few sections of the text, but did not provide rationale. I have provided more specific rationale for moving/removing some of these sections here:

1st paragraph under "Template-based data augmentation" in "Methods"
You describe a general need for augmentation within computer vision, you describe other methods for circumventing this need, and you then explain why such other methods may not be ideal for medical imaging. None of this paragraph describes what you actually did in this study. This material is used as introductory material before the next paragraph, in which you described what your group actually did. This material should be removed from Methods, and possibly included in Introduction or Discussion if relevant.

The following is the paragraph in question:

> The need for large training data sets is a well-known limitation of deep learning algorithms. Whereas the architectures developed for such tasks as the ImageNet competition have access to millions of annotated images for training, such data availability is atypical in medical imaging. In order to achieve data set sizes necessary for learning functional models, various data augmentation strategies have been employed [25]. These include application of intensity transformations, such as brightening and enhanced contrast. They might also include spatial transformations such as arbitrary rotations, translations, and even simulated elastic deformations. Such transformations might not be ideal if they do not represent shape variation within the typical range exhibited by the population under study.

We believe that this provides essential context for this specific subsection in motivating and understanding the proposed template-based data augmentation strategy. It would be much less effective in isolation in the Introduction or Discussion. Additionally, we do not find the specific rationale provided by the Reviewer to be convincing. Obviously the Reviewer disagrees but since no canonical source exists for writing scientific manuscripts (including, in this case, Academic Radiology

https://www.academicradiology.org/content/authorinfo?code=xacra-site#idp1359888

), the Reviewer and the authors will just have to agree to disagree. Although we have made several changes proposed by the Reviewer during the first round of reviews as they certainly improved the quality of the manuscript, per the Editorial office, we are not required to make all of the suggested changes ("Although we do not demand that you make all of the suggested changes, the reviewers' and editor's criticisms should be utilized to increase the quality of your manuscript."). We do not believe that this suggestion improves the quality of the manuscript so we choose not to make this change.

1st paragraph under "Proton MRI Lung Segmentation" in "Results"

You describe the dataset used, the two techniques applied to the data, and the process of centering/aligning/selecting the data. This is a description of what methods you performed, but not what results you found. This material should be moved to the Methods. Your Methods section describes your image acquisition, the two techniques used to process the images (with emphasis on the new proposed technique), but is missing a description of what evaluation methods were used to compare the techniques.

The following is the paragraph in question:

> After constructing the U-net structural model using template-based data augmentation, we applied it to the evaluation data consisting of the same 62 proton MRI used in [8]. We performed a direct comparison with the joint label fusion (JLF) method of [8] with an adopted modification that we currently use in our studies. Instead of using the entire atlas set (which would require a large number of pairwise image registrations), we align the center of the image to be segmented with each atlas image and compute a neighborhood cross-correlation similarity metric [27]. We then select the 10 atlas images that are most similar for use in the JLF scheme. The resulting performance numbers (in terms of Dice overlap) are similar to what we obtained previously and are given in Figure 5 along with the Dice overlap numbers from the CNN-based approach.

Isolating our brief mention of the technique for evaluative comparison in the **Results** *section helps to avoid confusion with our proposed methodology in the previous section. Again, the Reviewer would probably disagree but we do not believe that this suggestion improves the quality of the manuscript so we choose not to make this change.*