**Response to the reviewers**

*We appreciate the time spent by the editors and reviewers in assessing our manuscript. Below is a point-by-point response. Please note that major textual changes are highlighted in blue font in the current version of the manuscript.*

**Reviewer 1**

This paper mainly reviews deep learning based works for learning spatial correspondence between images (i.e., image registration). Since registration is a basic task in medical image analysis, this paper is of great importance. However, this paper needs more improvement.

*Again, we appreciate the time spent by the Reviewer in reading the manuscript as well as providing positive suggestions for improvement.*

I have the following comments:

1. In the related work part, it is good that the authors introduces many base networks, like Unet, STN and GAN so on. However, why do you choose only these networks? Are there any other base networks related to image registration?

*We clarified our selection of these basic architectural components with the following additional text placed after their listing at the beginning of Section 2.3:*

Although not exhaustive, many of these core networks feature prominently in the research reviewed below. Other networks and/or network components were chosen based on their potential future application in image registration. For additional information, we refer the interested reader to the deep learning reviews cited earlier in addition to pertinent textbooks (e.g., [59]).

*Based on our review of the literature, we believe that this list is relatively comprehensive for this deep learning-based image registration review. However, if the Reviewer feels that specific items that were inadvertently omitted which should have been included, we would be more than willing to include those if the Reviewer provides a relevant citation or other source of identifying information.*

2. In Table I, can you please describe the loss functions by equations?

*This comment and others below compelled us to completely revise much of the introductory material and we appreciate the Reviewer’s suggestions for motivating such changes as we feel that the current version of the manuscript is now more complete and better organized for the reader.*

*The set of loss functions is now included in the current version of the manuscript in Section 2.1. Inclusion directly within Table 1 was problematic due to space considerations. However, in mentioning the loss functions in Section 2.1, we referred to Table 1 appearing later in the manuscript.*

3. We need further comparison for these methods, I suggest the authors to consider the following factors:

a. advantage and disadvantages of different registration methods

b. what datasets do the methods use to validate in their papers? public or private datasets?

c. what about the performances for different methods?

d. what about the speed for training and inference?

e. Are there any deep learning based methods consistently working better than the traditional methods (for example, ants)?

f. other factors

And you can use table the compare about it.

*These are all great suggestions and we agree that such information can generally be useful in comparing image registration algorithms. The multiple review articles concerning traditional image registration which are cited in this manuscript bring up many, if not all, of these issues. We also touched on the importance of such issues in evaluating software in a joint editorial a few years back:*

*Nicholas J Tustison, Hans J Johnson, Torsten Rohlfing, Arno Klein, Satrajit S Ghosh, Luis Ibanez, and Brian B Avants. Instrumentation bias in the use and evaluation of scientific software: recommendations for reproducible practices in the computational sciences. Front Neurosci, 7:162, 2013.*

*In addition to issues of space (the manuscript is currently quite lengthy), there are other concerns which motivate us to \*not\* include such information on a paper-by-paper basis in this review, specifically within Table 1. We find that such concerns are 1) outside the scope of the deep learning focus of the current manuscript and 2) much of the requested information is simply not available within the reviewed literature making such comparisons terribly incomplete.*

*\*However\*, as we said, we completely agree with the Reviewer as to the utility of such information in general and have added the following text to the Discussion:*

Many of the additional challenges which concern traditional image registration methodologies and their introduction into the community persist with the deep learning expansion of the field. These challenges have been discussed at length in various articles, reviews, and editorials but it is worth reiterating due to their importance. Historically-rooted evaluation issues such as the use of public and/or private data sets and reproducibility concerns (e.g., published parameters, code availability) continue to be relevant for the deep learning shift. In addition, new concerns are salient such as the distinction in training and prediction speeds as well as possible hardware issues including GPU advancements and hardware availability.

*We also address these issues in one of our Discussion subsections:*

**4.2 Rethinking methodological reporting in the literature**

An additional, related challenge to assessing the literature of biomedical image registration is that the majority of technical papers do not report enough methodological detail to enable readers’ understanding of performance differences. In the context of public challenges hosted by Kaggle, participants work off of common baseline datasets, share all code as a prerequisite to involvement and are evaluated against hidden datasets provided by challenge organizers. Data, preprocessing, networks, postprocessing and results are transparent. In the context of the biomedical image registration literature, such transparency in terms of evaluation and development source code—and use of truly hidden data—is rarely present [121].

A recent review of “deep regression” [122] provides guidance on how such issues might be resolved in published work. The paper uses three public ground truth datasets that represent different forms of correspondence problems. The authors evaluate well-known VGG and ResNet regression architectures on these reference datasets. Notably, the authors of this paper are not promoting any particular architecture or method under evaluation. Common parameter variations of these networks are care- fully explored and results are reported in terms of the impact on confidence intervals. This study suggests that differences in performance due to preprocessing may exceed differences attributable to changes in network architecture. This finding, the objective approach and the reporting methods in this paper should be kept in mind when researchers and reviewers are considering new methodological efforts.

4. I suggest to add a subsection to describe the unsupervised deep learning based registration methods, for example, the series of work following this one "End-to-End Unsupervised Deformable Image Registration with a Convolutional Neural Network."

*We agree. This subsection has now been added as a* **Background** *section:*

**2.2 Supervised vs. unsupervised image registration**

The distinction between supervised and unsupervised methods is particularly salient within the deep learning community, the former characterized by training data that contains both inputs and desired outputs which is lacking in the latter. Traditional image registration can be generally viewed as an unsupervised perspective of transform optimization with deep learning expanding possibilities to include both supervised and unsupervised learning of spatial correspondence. Both are represented in the proposed techniques surveyed below.

Supervised image registration within the context of deep learning entails the employment of sufficiently large training data sets of input fixed and moving image pairs with their corresponding transformations. These data are used to train a designated network to learn those transformation parameters based on features discovered through the training process. The loss function quantifies the discrepancy between the predicted and input transformation parameters. Possibilities for obtaining the desired transformations used in the training data include output from traditional image registration algorithms as well as synthetically derived data sets.

Unsupervised deep learning-based approaches are more closely related to their traditional analogs in that they lack of the use of input transformation data. Optimization is driven via loss functions which incorporate intensity-based similarity quantification in learning the correspondence between the fixed and moving images. This is conceptually analogous to the classic neural network example of unsupervised learning—the autoencoder (cf [52])—where differences between the input and the network-generated predicted version of the input are used to learn latent features characterizing the data. In the case of unsupervised image registration, the optimal transformation is that which maximizes the similarity (as determined by the user-selected similarity loss function) between the input, specifically the fixed image, and the network-generated predicted version of the input, specifically the warped moving image as determined by the concomitantly derived transform.

5. Also, I suggest the authors to reformulate the discussion section into several subsections, so that we can understand easily which topics you're discussing.

*Great idea. We added the following subsection headings:*

**4.1 Improving evaluation strategies**

**4.2 Rethinking methodological reporting in the literature**

**4.3 Tailoring deep learning tools for medical imaging**