2.4 Step B:

Question:

How to split the training data set in each fold ?

1. Initially, the number of folder is set to 5, k =5. Then, find the size per fold by $\frac{\text{number of images,n}}{\text{the number of folder,k}}$, where size_fold = 14.
2. Randomisation: The training data was divided to 5 folders, loop over the 5 folds and the image indices is suffered every time when we start a new fold. We import random function and set the random seed to 3 to increase the reproducibility, use random.shuffle to suffer the indices.
3. Splitting the training data set in each fold:
   - the test group indices sets are the 14 indices within the range of from idx_fold*14 to the (idx_fold+1)*14, where idx_fold is the index of fold within range(5). The remaining indices is our training group, its first half elements are the indices for the moving image set, and the rest are the indices for the fixed image set.
   - Use the predefined DataFeeder class to load the images data for moving image, fixed image and test group, based on their respective indices sets.
   - Set up the placeholders for moving image, fixed image set and test group
   - Feed the grouped data sets with their corresponding placeholders
4. Evaluation: root mean square error (RMSE) metric will be used to evaluate the performance of the existing network in the cross validation experiment, as It represents the sample standard deviation of the differences between predicted values and observed values (called residuals)..

   Initially, evaluate the fitting model to obtain the target value for the predicted(t_i) and true values(t_observed[indices_i]), where i = label of (moving image, fixed image or test), by using the moving image, fixed image and test group data respectively. Then, the residuals can be calculated as (t_i -t_observed[indices_i]), equivalent to (yj-y_hatj) in equation 1. Then, compute the individual RMSE value for each group, by using eqn 1, where n is the number of observed sample data. Append the computed components in each group to obtain RMSE_i.

   $$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$ (equation 1)

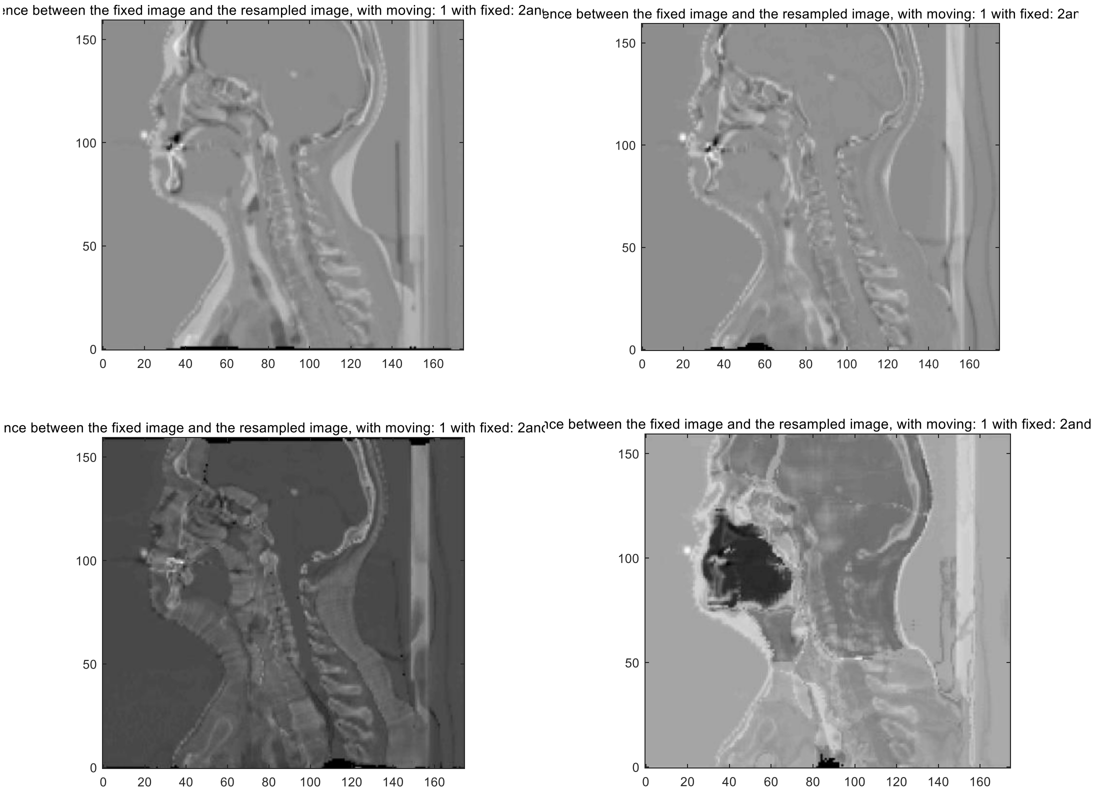
5. Potential problems with the choosing network:
   (1) With small amount of data, the existing network will easily run into the problems of overfitting. To achieve a better fitting, a better hardware(i.e. GPU) will be required.
   (2) The performance of the network is severely dependent on the initial parameter tuning, and this means expert level of knowledge in the domain will be required.
   (3) The physical implication is unclear, like a ‘blackbox’ model, and the network itself is hard to explain.

3.6 Reporting evaluation results

Note: For all evaluations, hyperparameters used was [10,1,0.1,0.01] and the moving-fixed image pairs are [0-1,0-2,1-0,1-2,2-0,2-1].

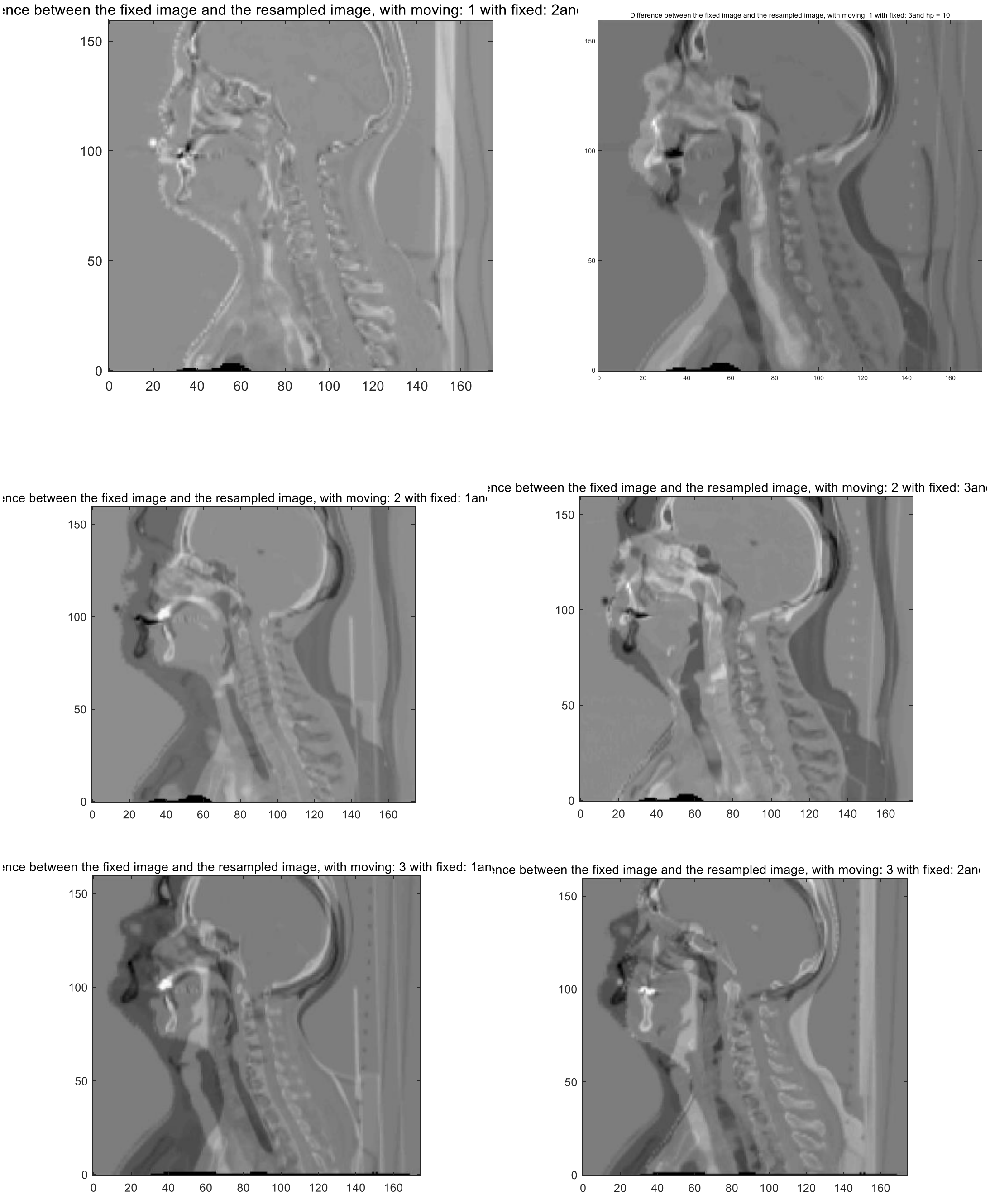
1. Visual assessment evaluation results

For registrations with descending hyperparameter (from 10 to 0.01) using same moving and fixed image pair, like the 4 figures shown below， we can find that the registration at hyperparameter of 1 gives the best alignment result. Then, hyperparameter [10,0.1, 0.01], tend to have better alignment from right to left. Figure set 2 indicates that, within same hyperparameter, the first image-pair[0-1] give the best alignment result. The inner soft（less dense）tissues like brain, brain stem and spinal cord have the best alignment across all registration, whereas the body surface(or contour), neck and the dense tissues like bone have the worst alignment.



ence between the fixed image and the resampled image, with moving: 1 with fixed: 2an
ence between the fixed image and the resampled image, with moving: 1 with fixed: 2ar
nce between the fixed image and the resampled image, with moving: 1 with fixed: 2anc
ce between the fixed image and the resampled image, with moving: 1 with fixed: 2and
0
50
100
150
0
20
40
60
80
100
120
140
160

Figure set 1



ence between the fixed image and the resampled image, with moving: 1 with fixed: 2an

Difference between the fixed image and the resampled image, with moving: 1 with fixed: 3and hp = 10

ence between the fixed image and the resampled image, with moving: 2 with fixed: 1an

ence between the fixed image and the resampled image, with moving: 2 with fixed: 3an

ence between the fixed image and the resampled image, with moving: 3 with fixed: 1an

ence between the fixed image and the resampled image, with moving: 3 with fixed: 2an

Figure set 2.

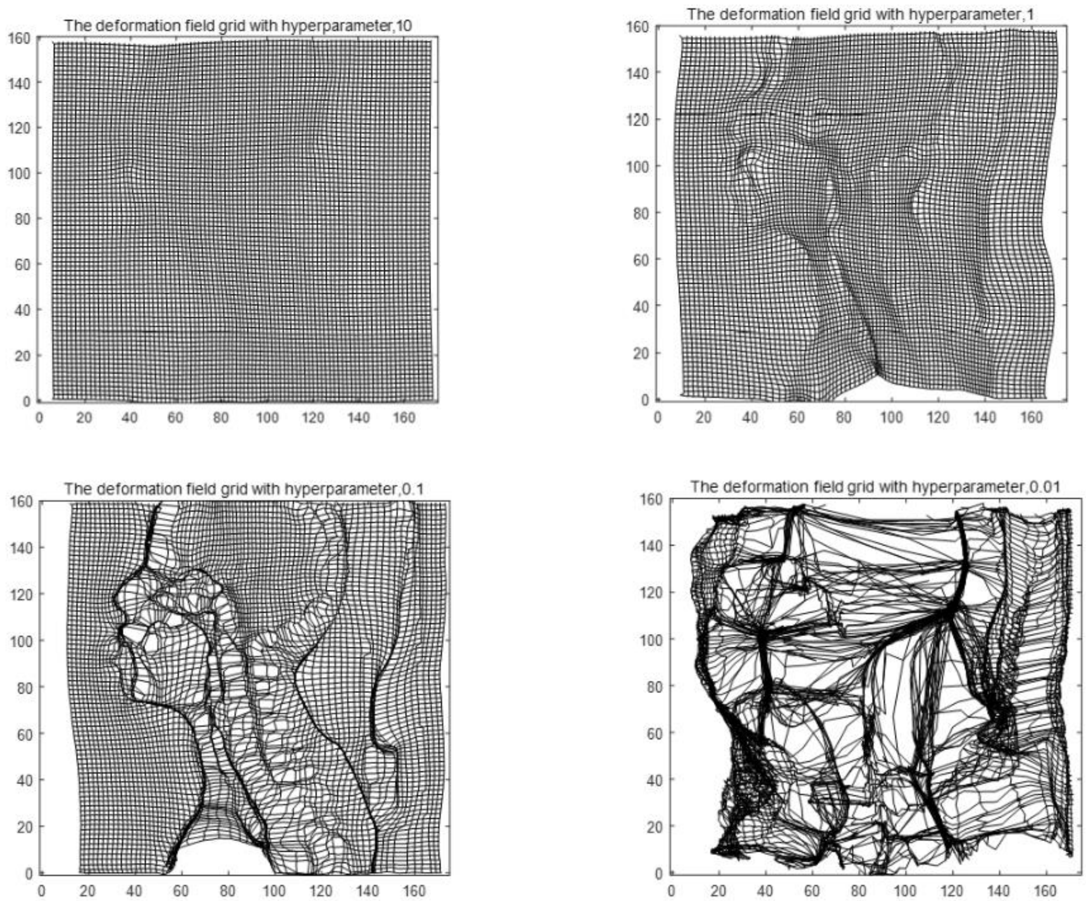


Figure set 3. Visualisation of deformation fields as grids for hyperparameter 10-0.01

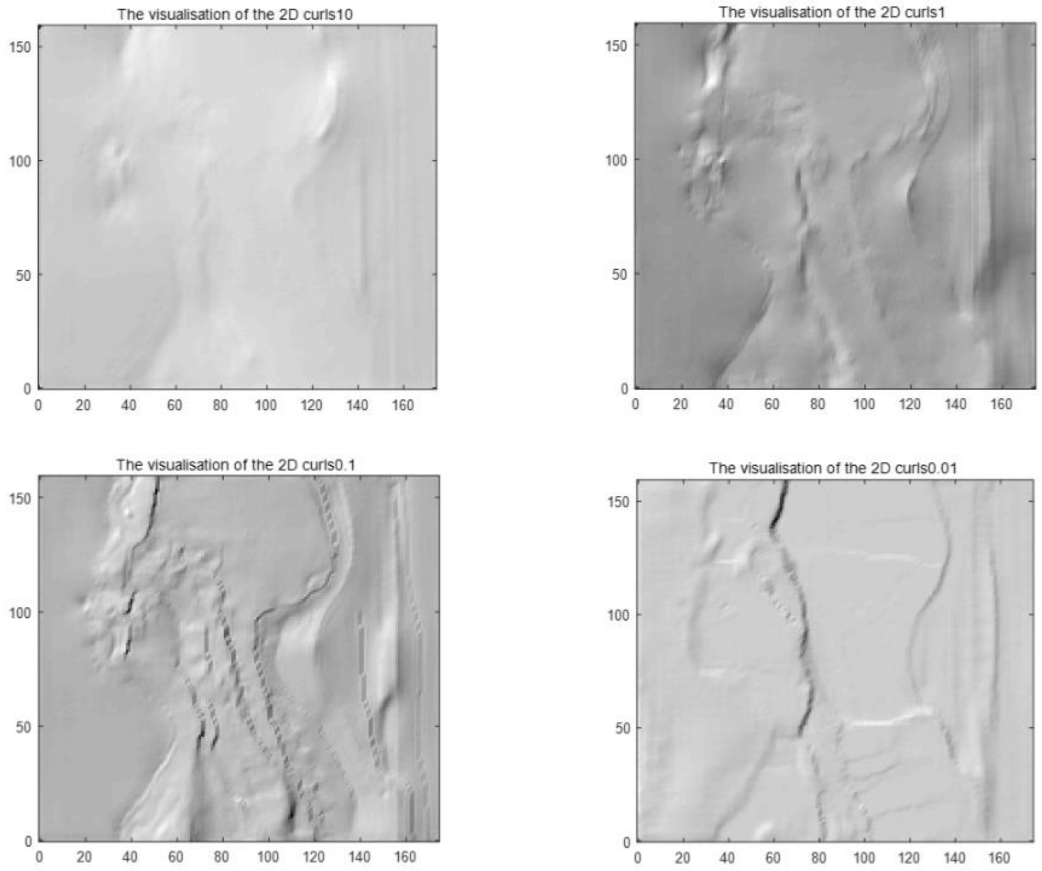


Figure set 4. Visualisation of 2D curls for hyperparameter 10-0.01

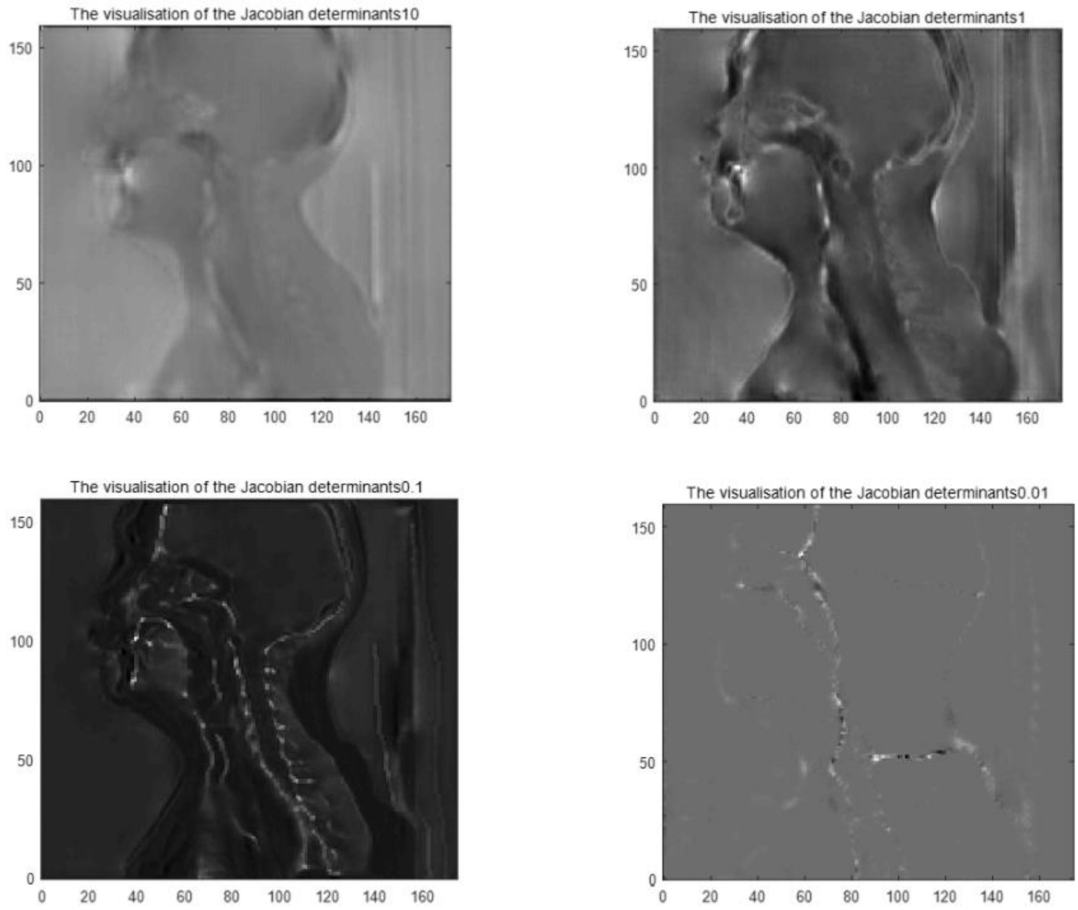
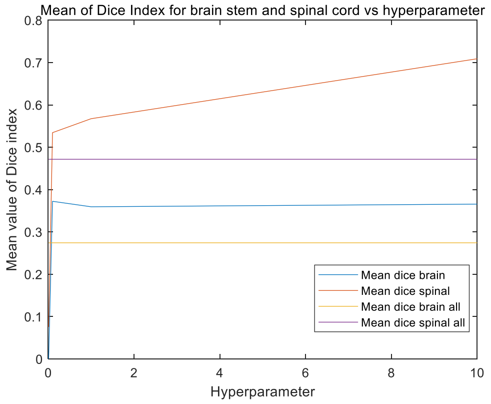
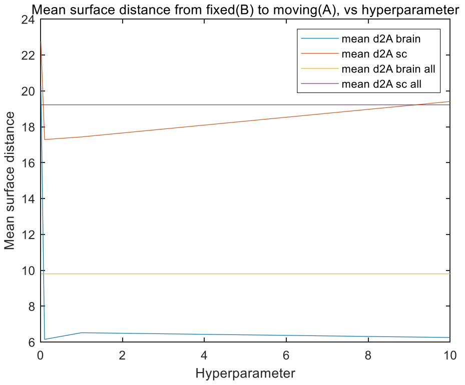

The visualisation of the Jacobian determinants10
The visualisation of the Jacobian determinants1
The visualisation of the Jacobian determinants0.1
The visualisation of the Jacobian determinants0.01
0
50
100
150
0
20
40
60
80
100
120
140
160

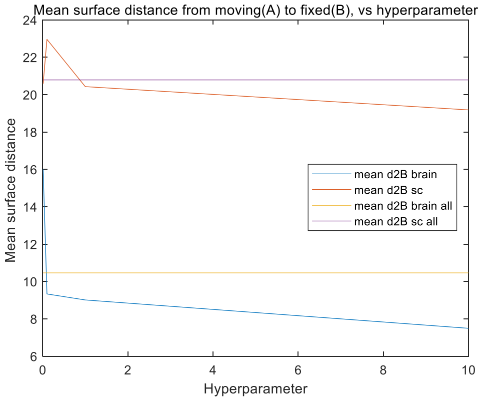
Figure set 5. Visualisation of Jacobian determinants for hyperparameter 10-0.01

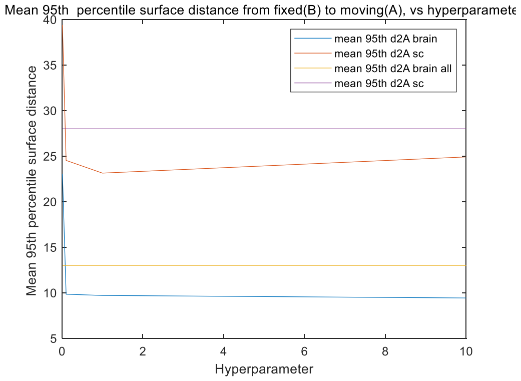
2. Obtained quantitative evaluation results based on different methods

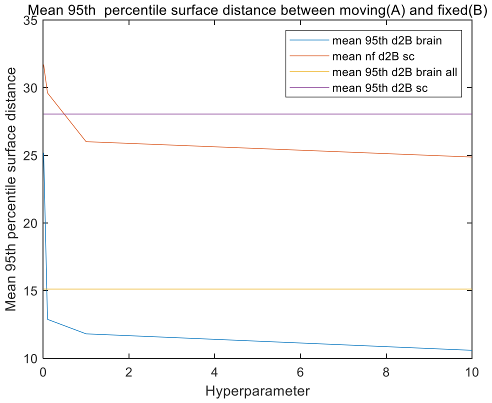
(1) Segmentation based evaluation results









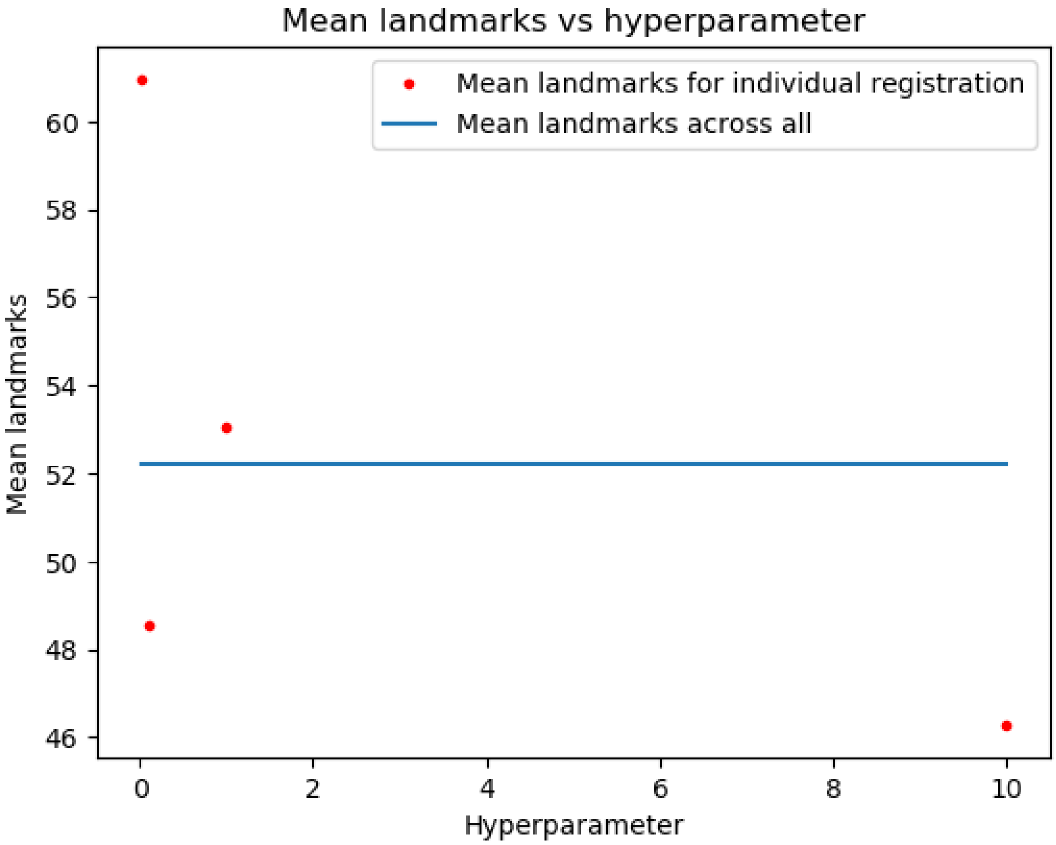


Dice index is a statistical term measures the similarity between two terms. In the first figure, the mean dice index measures the similarity between the warped image using the moving image and the fixed image across all the moving-fixed image pairs(see notes). It can see that mean dice index for the spinal cord is increasing with the weights(or hyperparameter) and higher than that for brain stem. This means that the spinal cord segmentations are better aligned compared to brain stem segmentations, and the overlap of spinal cord part between image pairs is getting bigger by increasing the hyperparameter. This trend matches with our observation in Section 3.6.1, where more pixels in the spinal cord parts overalap with increased hyperparameter. The brain stem part has similar dice index after hyperparmeter = 0.01, hence the overlapping pixels in the other 3 weights are similar. These three give a good alignment of brain stem, which matches our observation.

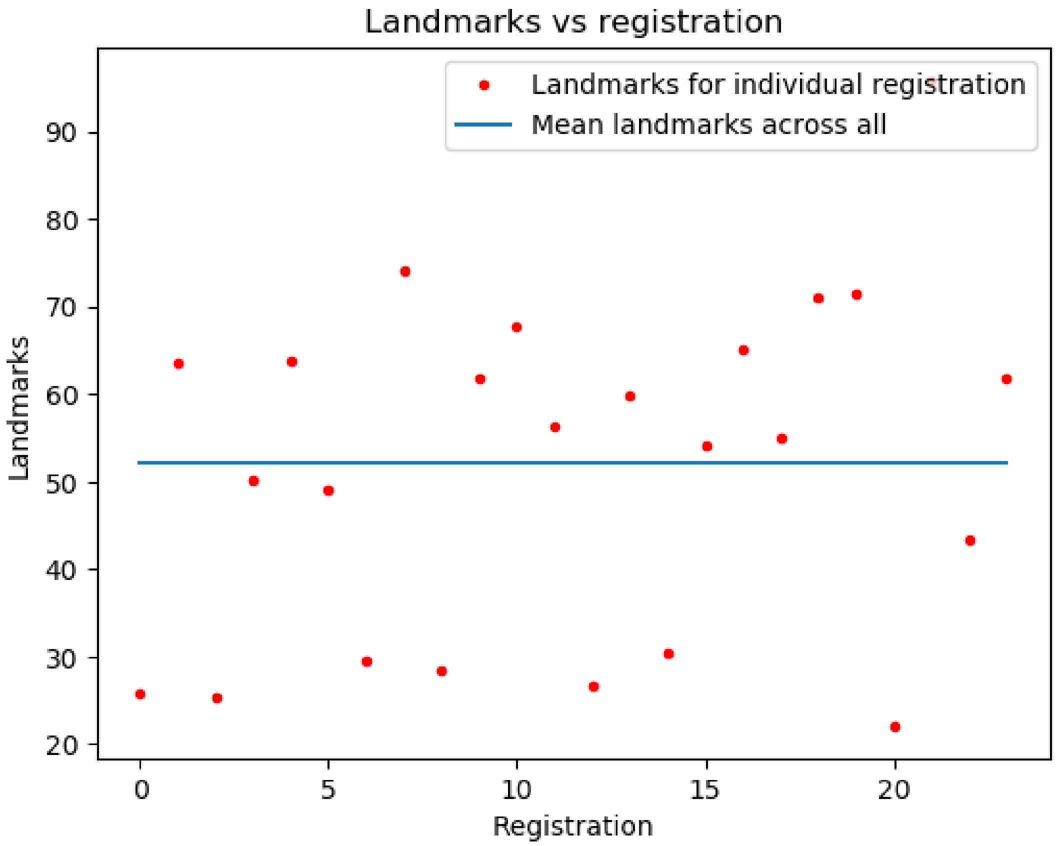
The forward (warped using moving to fixed) and backward(fixed to warped using moving) surface distances also measure the similarity between the image pair. The smaller the distance, the higher similarity between the image pair is. However, the interference tend to be bigger and higher error will be produced. From mean surface distance, the forward and backward surface distance are inverse to each other for spinal cord, but the trend for the brain stem stays the same except at weight=0.1 . Both mean and 95th percentile forward surface distance tend to decrease with hyperparameter, which means that the warped image will be more similar to the fixed image at higher hyperparameter. This is generally true according to our observation.

For the results observed as above, it shows that weight at 10 gives the best result.

(2) Landmarks based results evaluation

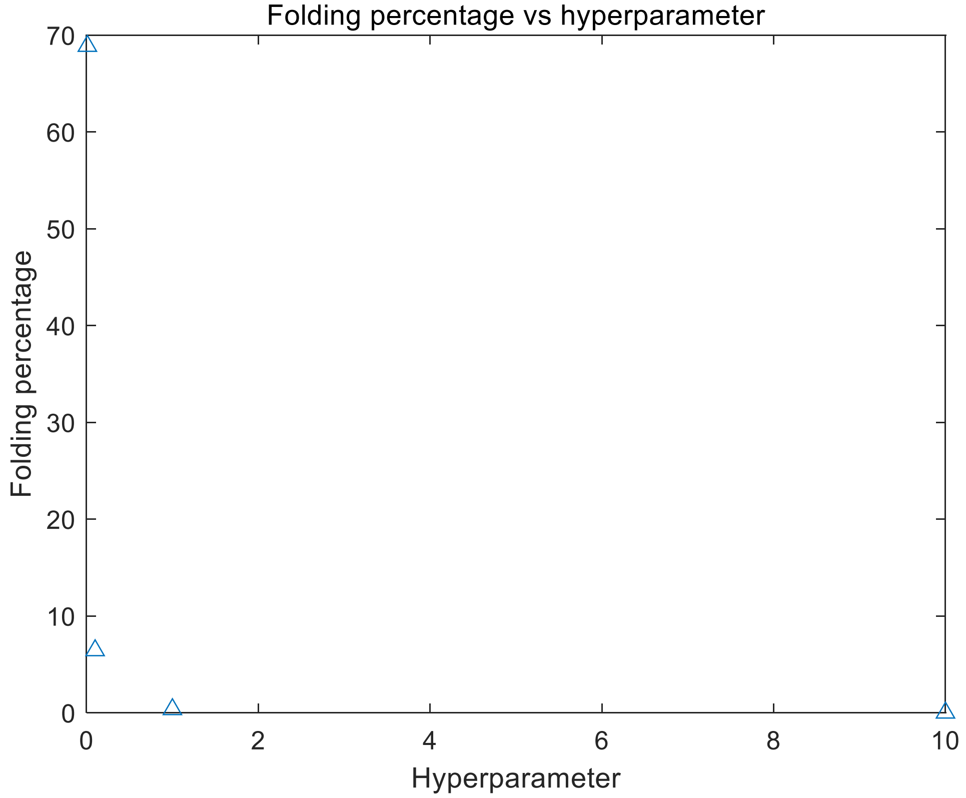


If we discard the hyperparameter at 0.01, we can observe a exponential decay shape across the weights. The hyperparameter at 1 is observed as the best estimator among all landmarks, as it is closest to the mean landmarks.

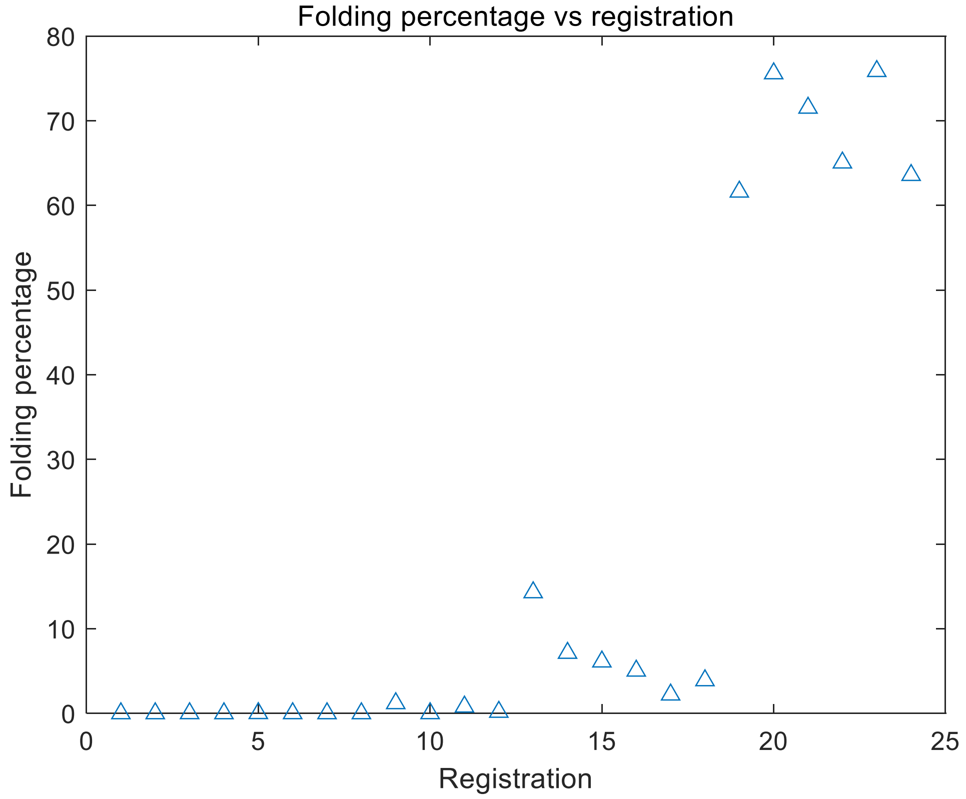


Across all the registrations, the 16th registration performs best as it is closest to the mean. 22nd performed the worst as it is most far away from the mean.

(3) Results of deformation field based evaluation

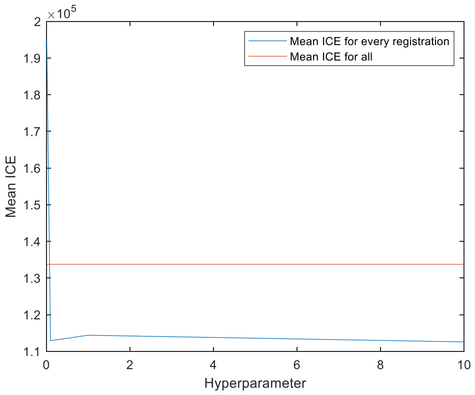


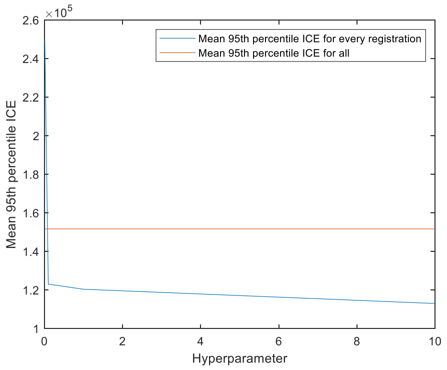
In the above figure, it shows that the trend of the folding percentage is exponentially decreasing, and there is no folding at hp = 1 and 10. This matches with the visualization of deformation fields (see Figure set 3), as the deformation fields at hp = 1 and 10 are much more smoother than that at others.

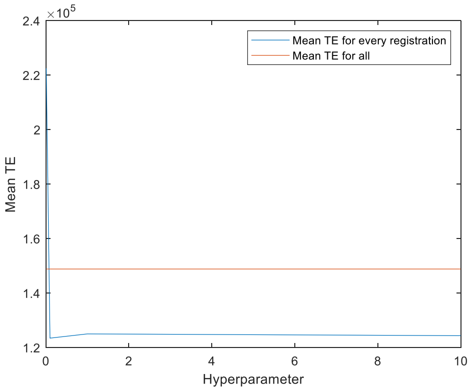


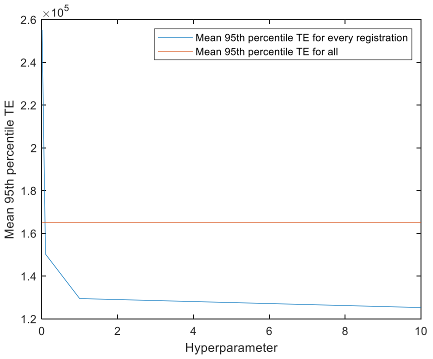
In the above table, 23rd registration has the most folding percentage, which represents the worst performance. The first 8 has the most smooth deformation fields. The overall trend is that the folding percentage increases with registration number(smaller weights).

(4) Results of transformation based evaluation

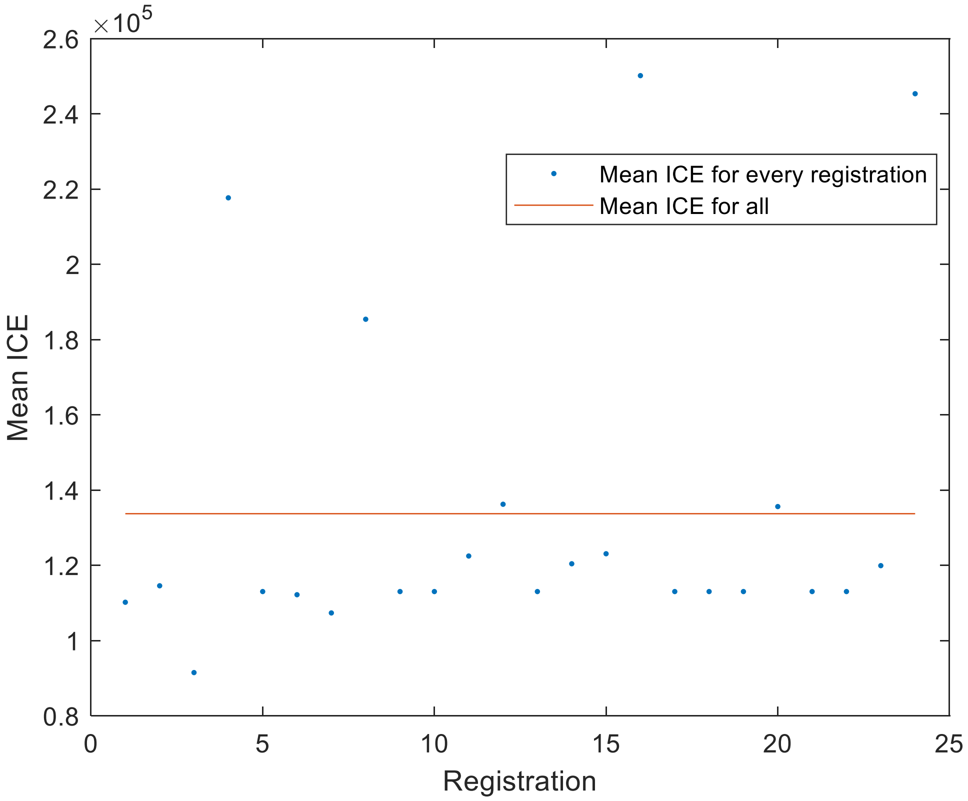


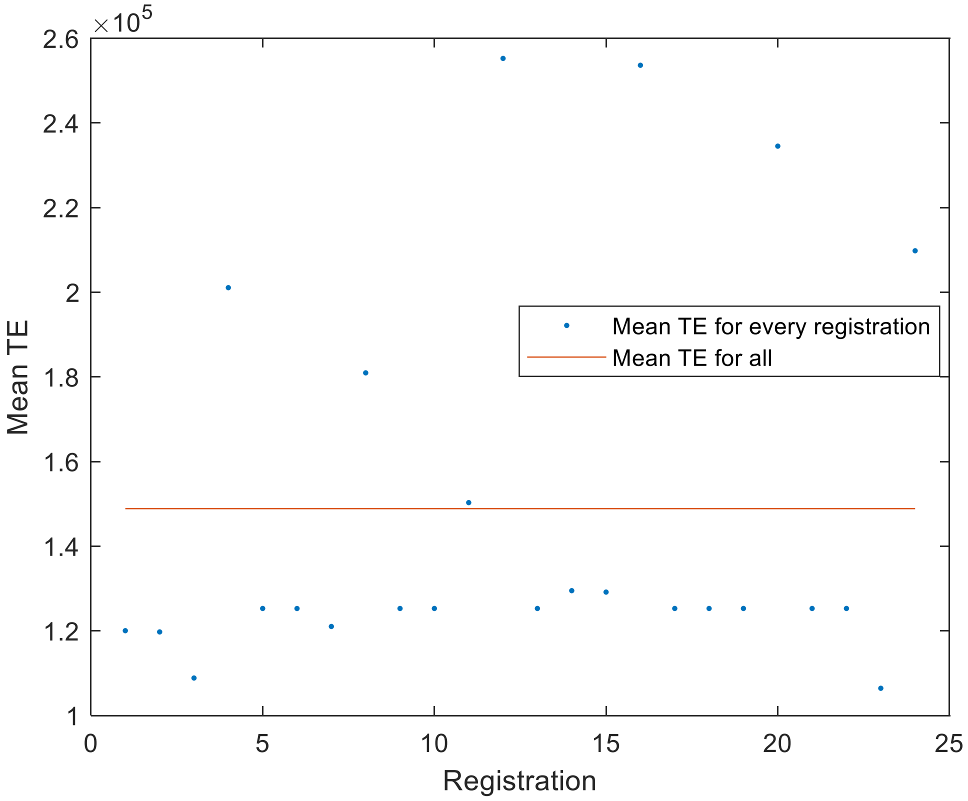






A larger inverse consistency error (ICE) means that the correspondence between the image pair is inversely consistent. The above figures show that both ICE and transitivity error (TE) are smallest at hyperparameter =10. This means that the similarity will be highest at weight of 10. It also show that the worst performance is given by the weight of 0.01. This agrees with the visualization that the best is given by hyperparameter at 10.





From above figure, we can see that the ICE tend to converge toward the direction of the mean ICE across all registrations. At registration 16, a highest ICE is observed, which means that 16 has the most inconsistent image pairs. The 3rd registration is observed to be the best as its magnitude is smallest. TE is highest at 12(worst) and lowest at 24(best). This is not agreed with the visual assessment.

3. Comparison

Most of the evaluation methods indicate that registrations performed with weight of 10 gives the best performance, except landmarks based evaluation where the weight of 1 is considering to give the best performance. Compare with the visual assessment, landmarks give a better evaluation result. Other methods are also accurate, as weight =10 also ranks the second best.

Visual assessment:

Advantage: More intuitive and allow the user to see the changes.

Disadvantage: Not very accurate, and it is pretty likely to ignore the local intensity changes. Also, computational expensive.

Combine with other methods: we can see whether other methods gave us the sensible results by comparing the particular image.

Segmentation based evaluation:

Advantage: This gives us a better observation at the specific changes at the autonomy parts that we are interested.

Disadvantage: Not very comprehensive, which may lead the user ignore the inference coming from the surrounding changes.

Landmark based evaluation:

Advantage: allow us to see the movement before and after transformation, for a particular point

Disadvantage: we cannot know much how the whole image changes

Contribution to multi-atlas: Less computational expensive, so that the bottleneck of the multi-atlas segmentation registration can be solved.

Combine with deformation field evaluation: This allows us to perform transformation of landmarks on the deformation fields, so that the change in the deformation fields in specific points can be seen.

Deformation field based evaluation

Advantage: we can see how the deformation fields change

Disadvantage: Expensive computational cost

Contribution to other methods: The deformation fields provide valuable information in knowing the training results, so that we can spot the errors in the deformation fields before performing other evaluatin.

Transformation based evaluation

Advantage: we can know how the distance between the start point and the end point changes.

Disadvantage: computationally expensive