Deep Learning Using Augmentation via Registration: 1st Place Solution to the AutoImplant 2020 Challenge

David G. Ellis(✉) and Michele R. Aizenberg

Department of Neurosurgery, University of Nebraska Medical Center,
Omaha, NE, USA
david.ellis@unmc.edu

Abstract. Automatic cranial implant design can save clinicians time and resources by computing the implant shape and size from a single image of a defective skull. We aimed to improve upon previously proposed deep learning methods by augmenting the training data set using transformations that warped the images into different shapes and orientations. The transformations were computed by non-linearly registering the complete skull images between the 100 subjects in the training data set. The transformations were then applied to warp each of the defective and complete skull images so that the shape and orientation resembled that of a different subject in the training set. One hundred ninety-seven of the registrations failed, resulting in an augmented training set of 9,803 defective and complete skull image pairs. The augmented training set was used to train an ensemble of four U-Net models to predict the complete skull shape from the defective skulls using cross-validation. The ensemble of models performed very well and predicted the implant shapes with a mean dice similarity coefficient of 0.942 and a mean Hausdorff distance of 3.598 mm for all 110 test cases. Our solution ranked first among all participants of the AutoImplant 2020 challenge. The code for this project is available at https://github.com/ellisdg/3DUnetCNN.

Keywords: Deep learning · Shape completion · Augmentation

1 Introduction

Automatic cranial implant design can save clinicians time and resources by computing the implant shape and size needed by a specific patient based on computed tomography imaging of their head [5,6,8,11]. The AutoImplant 2020 Cranial Implant Design Challenge seeks to test varying methods for designing an implant based on an image of a skull with a defect such that part of the skull is missing [7]. The challenge organizers submitted a baseline solution for this challenge in which they experimented with two deep learning solutions [10]. The first solution was to use a cascade style set of models where one model predicts the implant's shape at low-resolution, and another model subsequently refines

J. Li and J. Egger (Eds.): AutoImplant 2020, LNCS 12439, pp. 47–55, 2020.
https://doi.org/10.1007/978-3-030-64327-0_6

that shape at high-resolution. This cascade style solution has the advantage of limiting memory use by only computing the high-resolution implant shape on a cropped image rather than the whole image of the skull. However, the authors noted that overfitting to the training set caused the model to have two key limitations in its ability to generalize to cases outside of the training set [10]. The first limitation was that the model tended to predict the same implant shape for a given skull, even when the location of the defect had been changed. The second limitation was that the model was not able to accurately predict implants for defects that were in different locations and differently shaped than the defects in the training set images. The authors also experimented with using a deep learning network trained to predict the shape of the skull without defects and gave illustrations of how the model appeared to generalize well to cases outside the training set [10].

Inspired by these results, we aimed to implement a deep learning solution that would perform skull completion while learning from a heavily augmented training set. Augmentation is a common approach to expand the size of a training set of data so that a model training on the data will avoid overfitting and will generalize well to cases outside of the training set. Research by Zhao et al. has previously shown that registrations, along with other transformations, can be highly effective at augmenting small training sets of medical images [13]. Zhao et al. showed that the models trained on these augmented training sets performed much better than the models trained without such augmentations [13]. Therefore, we hypothesized that training models on a data set augmented using registrations would produce models that are highly accurate at predicting implant shapes from defective skull images.

2 Methods

2.1 Data

All data was provided by the organizers of the 2020 AutoImplant challenge. A training set was provided with images for 100 subjects along with a test set with images for 110 subjects. The training set consisted of binary images of the complete skull, the defective skull, and the implant for each subject. Renderings of these images are shown in Fig. 1 for an illustrative subject. The testing set consisted of only the defective skulls. One hundred of the testing set subjects had defects similar in location and shape to the training set, while 10 of the testing set subjects had defects with shapes and locations that varied from the defects of the training set.

2.2 Augmentation

Registration. To augment the training set of images, automatic registrations were computed between the skull images for each pair of subjects. This augmentation increases the size of the training set and allows for similarly shaped

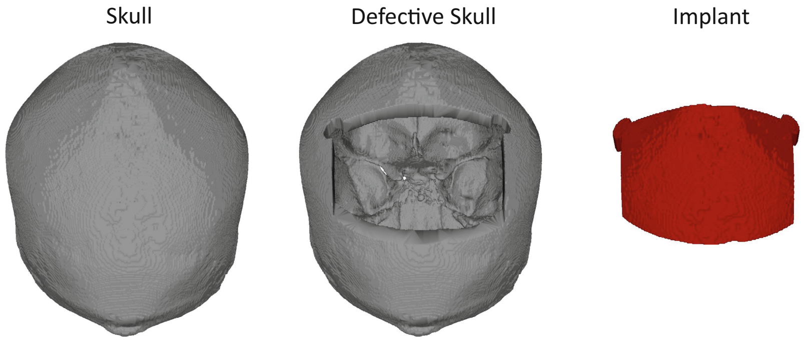

Fig. 1. Three-dimensional mesh renderings of the skull, defective skull, and implant images for a single subject from the training set.

skulls to have varying defect locations. With a training set of 100 images, each individual image can be registered with and warped into the space of 99 other images. Therefore, we attempted to warp the images via registrations and create 9,900 additional training images. Combined with the original training images, this would make an augmented training set of 10,000 images.

The "antsRegistrationSyNQuick.sh" script from the Advanced Normalization Tools (ANTs) package was utilized to compute the combined rigid, affine, and non-linear symmetric image normalization (SyN) warping transformations between skull images [2,3]. For each pair of subjects, the skull of the first subject would be used as the moving image, and the skull of the second subject would be used as the fixed image. The script then computed the transformations to warp the moving image into the fixed image space. The transformations were then applied to the complete and defective skull images to warp the moving images into the fixed image space, and the inverse transforms were used to warp the fixed images into the moving image space. This was repeated for every pair of subjects in the training data set.

Permutation. In order to enhance the model's ability to predict complete skulls for various defect locations, the images were mirrored along the anteroposterior and horizontal directions with a 50% probability for each training iteration.

Scaling. In order to make the model robust to variations in image scaling, the training images were randomly zoomed in and out with a 75% probability for each training iteration.

Translation. In order to make the model robust to variations in the position of the skull within the image, the training images were randomly translated with a 75% probability for each training iteration.

2.3 Preprocessing

All of the images were cropped to remove extra background padding in the images so that only one voxel of background padding around the non-background area remained [1]. In order that all of the data had the same orientation prior to being input into the model, the orientation of the images was set to Right, Anterior, Superior (RAS) [1,4]. After augmentation, the images were resampled down to a size of 176 × 224 × 144 voxels. Apart from the registrations, all preprocessing and augmentation steps were performed at run time.

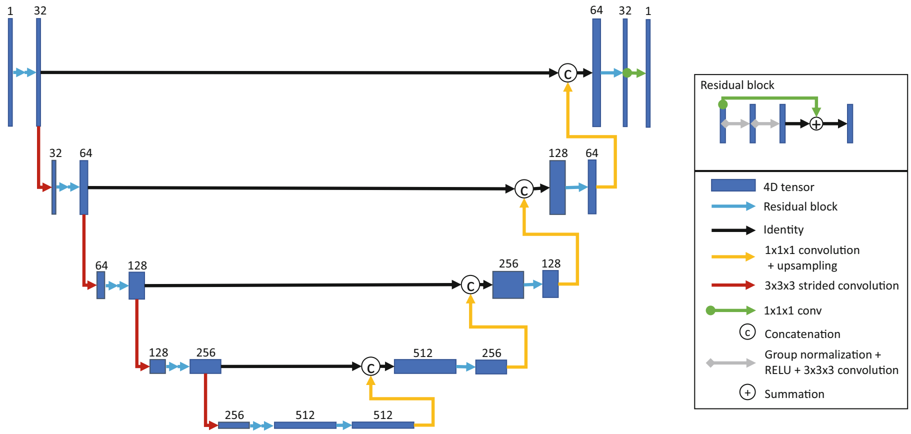

Fig. 2. U-Net model architecture. The number of channels at each step is shown in black.

2.4 Model

We used a U-Net-style convolutional neural network model with residual connections, as shown in Figure 2 [9,12]. Inspired by the research of Myronenko showing that a large receptive field and shallow decoder performed well in the automatic segmentation for brain tumors [12], our model used a large receptive field of 176 × 224 × 144 voxels. Comparatively, the baseline approach used a receptive field of 128 × 128 × 64 voxels [10]. The encoder to the model consisted of five layers, two ResNet style blocks per layer, a base width of 32 channels, dropout, and group normalization. The outputs of each encoding layer were downsampled using a strided convolution before being input into the next layer. The number of channels was doubled at each consecutive layer. Each decoding layer consisted of a single ResNet style block and took as input the output of the previous decoding layer concatenated with the output of the encoding layer at the same resolution. A 1 × 1 × 1 convolution and sigmoid activation were applied to the output of the final decoding layer.

2.5 Training

An ensemble of four models was trained to predict the complete skulls from the defective skulls of the augmented training set using four-fold cross-validation. Each model was trained using two NVIDIA V100 GPUs with 32 gigabytes of memory each. Due to limits on computing resources, training was stopped after seven days.

2.6 Testing

All four models were used to predict the complete skull for all 110 defective skulls from the test set, and the results were averaged across all four models. In order to derive the implant shape from the predicted skull shape, the defective skull was subtracted from the predicted skull. The difference image was then thresholded at 0.5. In order to remove spurious voxels from the predicted implant image, one iteration of morphological opening was performed, and all voxels not connected to the largest connected component were automatically removed.

3 Results

One hundred ninety-seven of the registrations failed, resulting in an augmented training set of 9803 sets of complete and defective skull images. Figure 3 shows an example of augmentations via registrations between two illustrative subjects.

The evaluation of the Dice similarity coefficients (DSC) and Hausdorff distances (HD) was computed by the organizers and reported in Table 1 and Fig. 4. The results are shown for the 100 test cases with defects resembling that of the training cases as well as the 10 test cases with defect shapes and locations that varied from the training set. Qualitatively, the predicted implants matched the shape of the holes in the defective skulls well regardless of defect shape and location, as shown in Fig. 5.

Table 1. Mean Dice similarity coefficient (DSC) and Hausdorff distances (HD) for the 100 test cases with defects resembling the training set as well as for the 10 test cases with defects that varied from the training set in shape and location.

		Test 100	Test 10	Overall
Baseline [10]	DSC	0.856	–	–
	HD (mm)	5.183	–	–
Ours	DSC	**0.944**	0.932	0.942
	HD (mm)	**3.564**	3.934	3.598

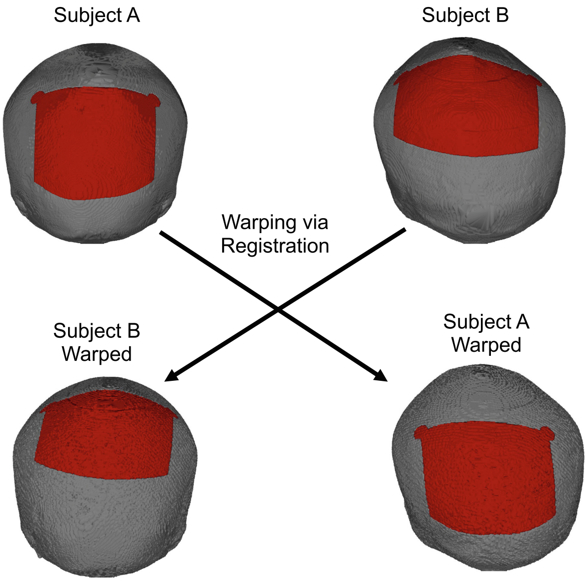

Fig. 3. Example of augmentations produced via registration. The non-linear registration is computed between the skull images of Subject A and Subject B. Then, the defective skull, shown in gray, as well as the implant, shown in red, are warped and translated using the computed non-linear registration. This produces two additional sets of defective skull and implant pairs from every pair of subjects in the training set. (Color figure online)

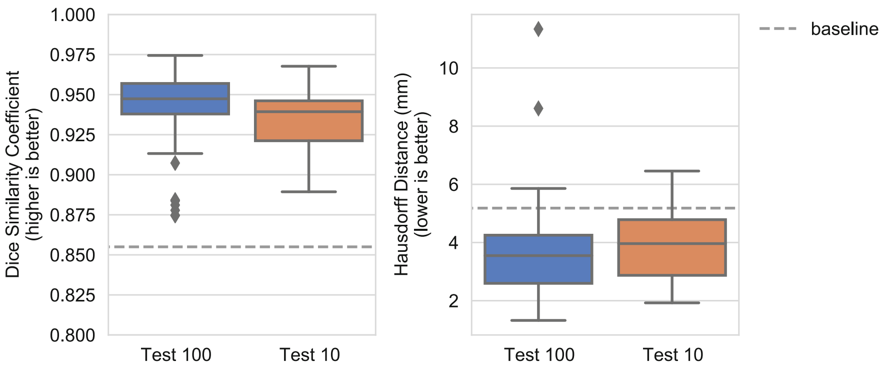

Fig. 4. Distribution of the Dice similarity coefficient and the Hausdorff distances for the 100 test cases and the 10 test cases with defects that varied from the training set in shape and location. For comparison, the average scores for the baseline method on the 100 test cases are shown as the dashed gray line [10].

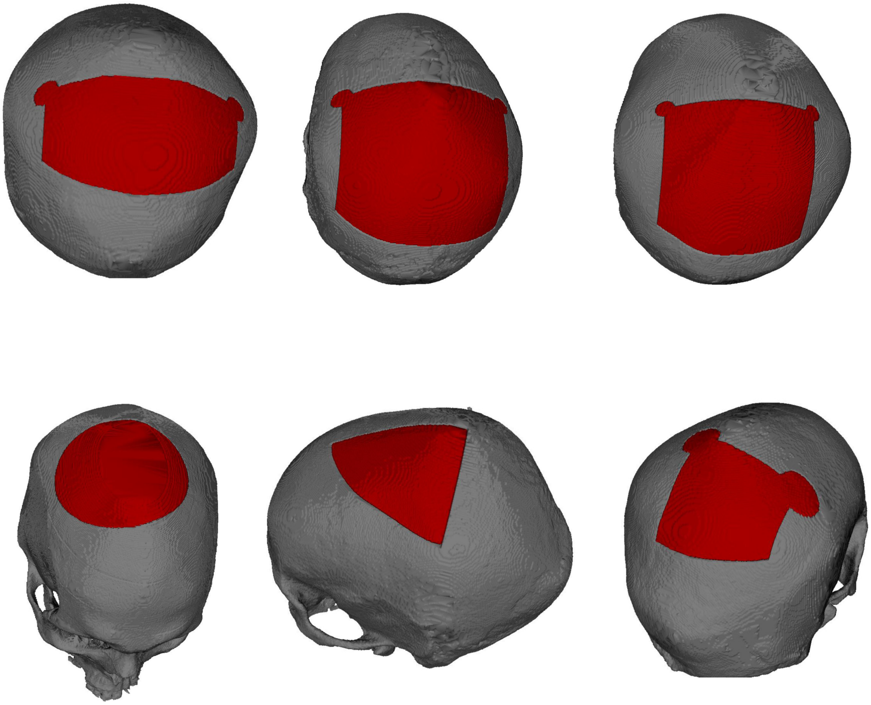

Fig. 5. Examples of defective skulls (gray) and the predicted implants (red) from cases in the 100-case test set (top row) and the 10-case test set (bottom row). The predicted implants fill the defects in the skulls well regardless of defect shape and location. (Color figure online)

4 Discussion

The evaluation metrics for our approach were much better than those reported using the baseline approach without augmentation [10]. For reference, the mean DSC and HD for the baseline approach on the 100 test cases was 0.855 and 5.44 mm [10], respectively, while our approach resulted in DSC and HD measures of 0.944 and 3.564 mm, respectively. This improvement in evaluation metrics likely resulted from a number of improvements we made over the baseline approach. The most prominent improvement was the automatic augmentation of the training data via registration. This augmentation prevents overfitting by allowing similarly shaped skulls to have varying defect locations. By using registrations, we were able to increase the size of the training set from 100 image pairs to 9803 image pairs. While manually creating a training set of this size would require an incredible amount of time, using registrations allows for the augmentations to be computed automatically without manual intervention. Furthermore, the permuting, scaling, and translating augmentations performed during training may have also improved performance by keeping the models from overfitting to the training set.

The DSC and HD scores were only slightly worse for the test set of 10 subjects with unique defects, as shown in Table 1 and Fig. 4, despite the models not being trained on cases with similar defects. The ability of the models to generalize to these cases is likely the result of using a shape completion strategy along with permutation augmentations. The shape completion approach forced the model to focus on generating a complete skull, regardless of the defect shape. The permutation augmentations allowed the model to train on defects that were flipped in orientation from the locations in the training set. Though these augmentation strategies performed very well, further improvement would likely be seen by adding skulls to the training set of images that have more variation in defect shape and location.

To remove spurious voxels from the implant predictions, we used a morphological opening procedure that removed voxels that were only partially or weakly connected to the predicted implant. This was effective but likely also resulted in a loss of edge voxels at the corners of the implant that should have been included in the prediction. Future work could focus on finding a more optimal way to remove these spurious predictions while retaining the correctly predicted corner voxels. One approach may be to experiment with threshold levels.

5 Conclusion

We demonstrated that registrations between anatomical CT images could effectively augment a training set of images for skull shape completion. The model trained on the augmented data set was able to accurately predict complete skull shapes much better than the baseline approach that was trained without using such augmentations.

Acknowledgments. This work was completed utilizing the Holland Computing Center of the University of Nebraska, which receives support from the Nebraska Research Initiative.

References

1. Abraham, A., et al.: Machine learning for neuroimaging with scikit-learn. Front. Neuroinform. **8**, 14 (2014). https://doi.org/10.3389/fninf.2014.00014. https://www.frontiersin.org/article/10.3389/fninf.2014.00014
2. Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C.: Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. **12**(1), 26–41 (2008)
3. Avants, B.B., Tustison, N., Song, G.: Advanced normalization tools (ANTS). Insight J. **2**(365), 1–35 (2009)
4. Brett, M., et al.: freec84: nipy/nibabel: 3.1.1 (2020). https://doi.org/10.5281/zenodo.3924343

5. Chen, X., Xu, L., Li, X., Egger, J.: Computer-aided implant design for the restoration of cranial defects. Sci. Rep. **7**(1), 1–10 (2017)
6. Egger, J., et al.: Interactive reconstructions of cranial 3D implants under MeVisLab as an alternative to commercial planning software. PLoS ONE **12**(3), e0172694 (2017)
7. Egger, J., et al.: Towards the automatization of cranial implant design in cranioplasty (2020). https://doi.org/10.5281/zenodo.3715953
8. Fuessinger, M.A., et al.: Planning of skull reconstruction based on a statistical shape model combined with geometric morphometrics. Int. J. Comput. Assist. Radiol. Surg. **13**(4), 519–529 (2017). https://doi.org/10.1007/s11548-017-1674-6
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
10. Li, J., Pepe, A., Gsaxner, C., von Campe, G., Egger, J.: A baseline approach for autoimplant: the miccai 2020 cranial implant design challenge. arXiv preprint arXiv:2006.12449 (2020)
11. Morais, A., Egger, J., Alves, V.: Automated computer-aided design of cranial implants using a deep volumetric convolutional denoising autoencoder. In: Rocha, Á., Adeli, H., Reis, L.P., Costanzo, S. (eds.) WorldCIST'19 2019. AISC, vol. 932, pp. 151–160. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-16187-3_15
12. Myronenko, A.: 3D MRI brain tumor segmentation using autoencoder regularization. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) BrainLes 2018. LNCS, vol. 11384, pp. 311–320. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11726-9_28
13. Zhao, A., Balakrishnan, G., Durand, F., Guttag, J.V., Dalca, A.V.: Data augmentation using learned transformations for one-shot medical image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8543–8553 (2019)