Computers in Biology and Medicine 137 (2021) 104766

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Deep learning for cranioplasty in clinical practice: Going from synthetic to real patient data

Oldřich Kodym [*], Michal Španěl, Adam Herout

Department of Computer Graphics and Multimedia, Brno University of Technology, Božetěchova 2, 612 66, Brno, Czech Republic

ARTICLE INFO

Keywords:
Cranioplasty
Skull reconstruction
Cranial implant design
3D convolutional neural networks

ABSTRACT

Correct virtual reconstruction of a defective skull is a prerequisite for successful cranioplasty and its automatization has the potential for accelerating and standardizing the clinical workflow. This work provides a deep learning-based method for the reconstruction of a skull shape and cranial implant design on clinical data of patients indicated for cranioplasty. The method is based on a cascade of multi-branch volumetric CNNs that enables simultaneous training on two different types of cranioplasty ground-truth data: the skull patch, which represents the exact shape of the missing part of the original skull, and which can be easily created artificially from healthy skulls, and expert-designed cranial implant shapes that are much harder to acquire. The proposed method reaches an average surface distance of the reconstructed skull patches of 0.67 mm on a clinical test set of 75 defective skulls. It also achieves a 12% reduction of a newly proposed defect border Gaussian curvature error metric, compared to a baseline model trained on synthetic data only. Additionally, it produces directly 3D printable cranial implant shapes with a Dice coefficient 0.88 and a surface error of 0.65 mm. The outputs of the proposed skull reconstruction method reach good quality and can be considered for use in semi- or fully automatic clinical cranial implant design workflows.

1. Introduction

Cranioplasty is a procedure that restores the aesthetic, mechanical, and protective function of a defective skull by implanting material into the defect area. Although autologous bone or pre-formed titanium meshes can be used as implants, 3D printable implants have been shown to be more versatile and to have several other advantages, such as a lower risk of complications or lower chance of requiring secondary surgery [1,2]. Manufacturing these implants requires modeling their shape in computer-assisted design (CAD) software as the first step. This virtual reconstruction, however, requires the human operator to have sufficient knowledge of skull anatomy as well as skill in 3D modeling. Even if these requirements are met, correctly modeling the implant is time-consuming even for a skilled operator, especially in cases of defects reaching into both lateral sides of the skull [3]. Automatically producing fast and precise estimations of the implant shapes could therefore lead to increased standardization and efficiency of cranioplasty clinical workflow. In recent years, skull shape reconstruction methods based on volumetric convolutional neural networks (CNNs) have shown great promise in this regard [4,5,6], yet they remain mostly untested on real patient data, which limits their potential of translation into clinical practice. This article deals with the issue of using these CNN-based models on real patient data and improving their performance with the use of multi-task learning, as illustrated in Fig. 1.

Most recent (semi-)automatic skull reconstruction methods aim to solve the task of finding the exact shape of the missing part of the skull. We refer to this type of reconstruction output as *a skull patch* in this article. The main criteria for a successful skull patch estimation is an anatomically plausible, symmetric shape with a smooth and seamless fit along the defect border. In clinical practice, this allows the operator to use the estimated skull patch as a template for the final cranial implant design in CAD software. Conventional skull reconstruction methods use mirroring of the healthy side of the skull onto the defective side [7], surface interpolations [8,9] or their combination [10] to estimate the skull patch. Statistical shape models [11] greatly expanded the range of skull defects that can be reconstructed automatically [12,13,14]. In recent years, the research focus shifted to volumetric convolutional neural networks (CNNs) which have shown great promise in fast and robust skull patch reconstruction [4,6,15] and became the method of choice in the 2020 AutoImplant challenge [5]. The CNN-based methods

[*] Corresponding author.
E-mail address: ikodym@fit.vutbr.cz (O. Kodym).

https://doi.org/10.1016/j.compbiomed.2021.104766
Received 15 April 2021; Received in revised form 10 August 2021; Accepted 10 August 2021
Available online 14 August 2021

are usually trained and evaluated using synthetic defects created by removing some part of a healthy skull, resulting in virtually an infinite amount of different samples.

The final shape of the cranial implant (referred to simply as *implant* for the remainder of this article) differs from the shape of the skull patch in several ways (see Fig. 2). The implants have a constant thickness different from the original bone and have some spatial tolerance along the defect border to account for scar tissue and continuing bone growth, ensuring implantability. The shape of the implant can also be estimated directly by a CNN model, provided that sufficient training data is available for training. Although it is more difficult to edit this kind of shape in CAD software due to fine details along the defect border, it has the potential to be used in a fully automatic setting when no human operator, or not enough time for a manual design, is available, for example in intra-operative rapid manufacturing of cranial implants [16].

Synthetic datasets for automatic estimation of skull patches recently became available because they are easy to create from public databases of healthy skulls, such as CQ500 [17]. However, they do not necessarily fully cover the defective skull shape distribution of target clinical data (i.e. different anatomical variability of the target population, defect shapes and sizes, complex morphology of defect border), which may affect the resulting reconstruction quality in practice [4]. Real clinical data with expert-designed implant models are, on the other hand, difficult to obtain. Furthermore, in our experience, the distribution of available clinical data is often biased towards simple unilateral defect cases and not easily extendable by synthetic defect and implant shapes. The more challenging bilateral and fronto-orbital defects are less common, yet it is in these challenging cases where correct automatic skull patch reconstruction or implant design can have the largest impact on clinical practice. It is therefore desirable to design a method that will be able to leverage both types of cranioplasty data.

The main contributions of this article are the following:

- A multi-branch CNN architecture is proposed as an extension to the cascaded CNN used for skull reconstruction in our previous work. The architecture allows for training on both synthetic and clinical data samples.
- The proposed CNN model is evaluated on a large dataset of real defective skulls with expert-designed implants for the first time. The positive effect of the proposed method on reconstruction performance is demonstrated.
- A novel metric based on Gaussian curvature is implemented to quantify surface imperfections along the defect border.

2. Materials and methods

2.1. Datasets

We use two different cranioplasty datasets in this work. The Skullbreak dataset [18] is a synthetic skull shape reconstruction dataset adapted from the CQ500 public database of head CT scans [17]. The CT scans were rigidly aligned and segmented to provide normalized shapes of healthy skulls. Then, synthetic defects were created by subtracting random shapes from several regions in each skull. Morphological operations were additionally used to mimic some degree of bone healing processes along the defect borders. The dataset contains 570 training and 100 testing pairs of defective skulls and corresponding skull patches.

The second in-house dataset was provided by TESCAN Medical company. It contains a total of 387 real patient cases indicated for cranioplasty. Each patient case consists of CT data with manual skull segmentation and a mesh model corresponding to an expert-designed cranial implant. 75 of these cases additionally contain expert-designed mesh models of patches covering the full area of the defects that were used as an initial template for the final implant design by an expert. Although these expert-designed patches have a different thickness from the original bone, their outer surface can be used as a reference for the outer surface of automatically reconstructed patches. This naturally led us to split the in-house dataset correspondingly into 312 training cases and 75 test cases, ensuring that a real clinical test set of reasonable size is available for the evaluation of both the skull patch shape estimation and the final implant shape estimation tasks. All implant and patch mesh

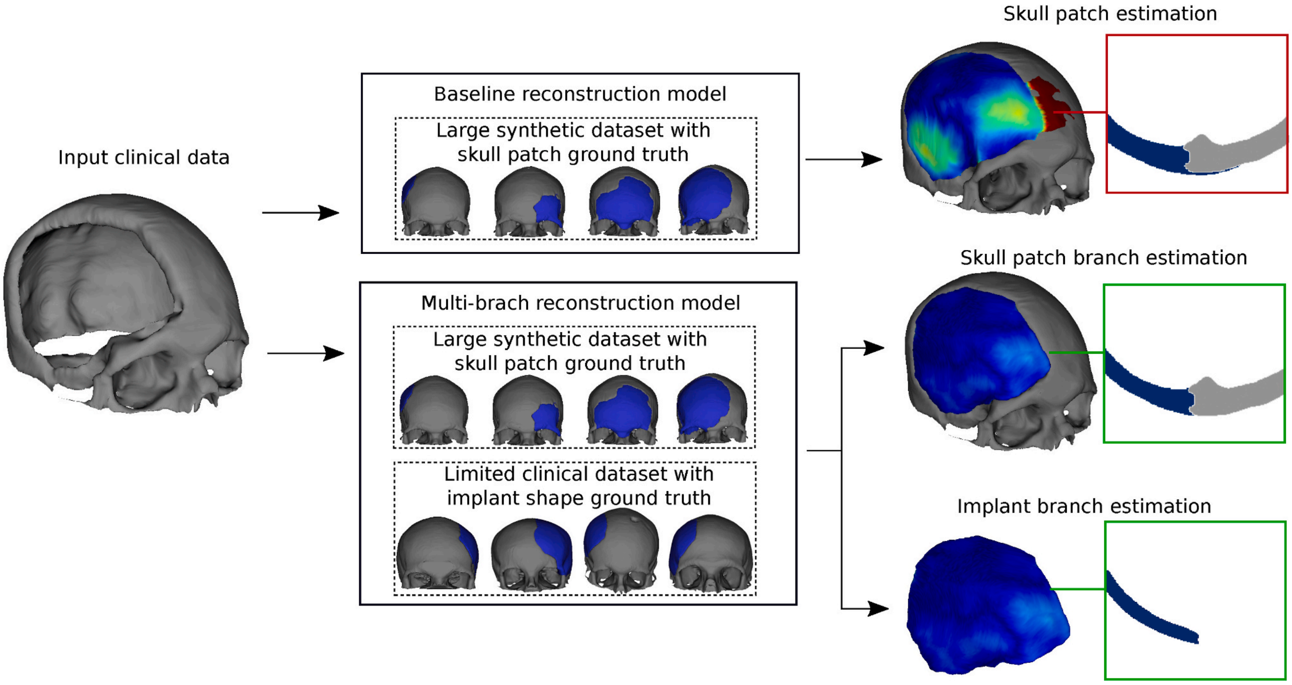
Input clinical data
Baseline reconstruction model
Large synthetic dataset with
skull patch ground truth
Multi-brach reconstruction model
Large synthetic dataset with
skull patch ground truth
Limited clinical dataset with
implant shape ground truth
Skull patch estimation
Skull patch branch estimation
Implant branch estimation

Fig. 1. The proposed multi-branch architecture makes use of multi-task learning on different skull reconstruction datasets. In addition to the higher overall accuracy and ability to directly output cranial implant shapes, the skull patch output of the multi-branch model also better fits the shape to complex defect borders in real clinical data.

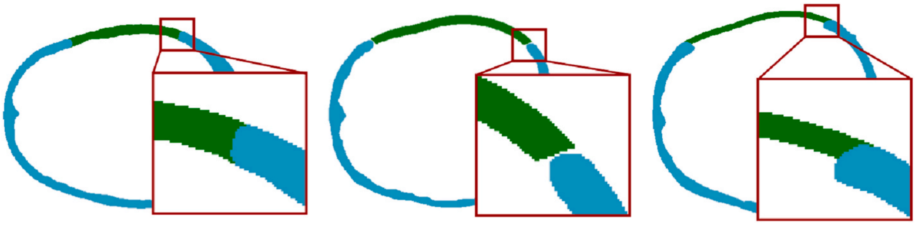

Fig. 2. Axial slices through samples from the datasets used in this work. From left to right: skull patch sample from a synthetic dataset, manually designed implant shape sample from an in-house clinical dataset, manually designed skull patch surface sample from an in-house clinical dataset.

models in the clinical in-house dataset were rasterized into voxel grids and the data were rigidly aligned to conform with the Skullbreak data. Several examples from all datasets can be seen in Fig. 2.

The two datasets also differ in several more aspects. Because they come from geographically distant sources, the average size and the anatomical variability of the skulls differ [19,20]. The scale and positional variability of the defects are also different. While the Skullbreak dataset was created specifically to contain a balanced amount of unilateral, bilateral, and fronto-orbital defects, the clinical in-house dataset contains a higher amount of unilateral defects with a larger size and reaching farther into lower parts of temporal and sphenoid bones. Although some of these differences could be addressed by tailoring the synthetic defects in the Skullbreak dataset to fit the distribution of clinical data more closely, some aspects such as skull shape variation and defect border complexity cannot be precisely emulated.

2.2. Baseline CNN models for shape estimation

We use the same baseline reconstruction method for both the skull patch estimation and the implant estimation tasks, with the only difference being the data used for training. The method is based on a cascade of two U-net-like volumetric CNNs proposed in our previous work [4]. The first, coarse CNN $g(\cdot)$ with weights θ_g takes a binary shape of defective skull in coarse resolution, denoted x_{coarse}, and produces an initial output shape estimate with the same resolution y_{coarse}:

$$y_{coarse} = g(x_{coarse}; \theta_g) \tag{1}$$

The second, high-resolution CNN $f(\cdot)$ with weights θ_f then takes a single crop of the upscaled coarse shape estimate y_{coarse} and a corresponding crop of the high-resolution defective skull $x_{high-res}$ and produces a high-resolution shape estimate $y_{high-res}$ of that crop, effectively performing super-resolution of the coarse shape estimate locally conditioned on the high-resolution defective skull:

$$y_{high-res} = f(y_{coarse}, x_{high-res}; \theta_f) \tag{2}$$

The coarse CNN model additionally uses a mirrored copy of the input volume, which has been shown to improve lateral symmetry of output shapes [4].

We use 12 initial feature channels and an input volume size of 128 × 128 × 128 for both the coarse and high-resolution CNNs. The final output is created by first inferring the coarse shape estimate and then inferring the high-resolution CNN in a sliding window manner. Both the original input and the final high-resolution output volumes have size 512 × 512 × 512 voxels with a resolution of 0.4 mm per voxel. We train the CNN cascade for 300 000 steps on mini-batches of size 4 using the soft Dice loss [21]. Each training step consists of updating the weights θ_g using the loss computed on coarse resolution and then updating both weights θ_g and θ_f using the loss computed on random high-resolution crops. More details about the CNN architecture and training procedure can be found in the original work [4].

2.3. Multi-branch CNN model for joint shapes estimation

To facilitate training of the CNN cascade using both the synthetic skull patch dataset and the clinical implant dataset simultaneously, we split the outputs of the model into a separate skull patch estimation branch and implant estimation branch at both coarse and high resolution. The shape estimation branches are formed by a single conv-ReLU-conv-softmax block with the convolutional layers having the same number of features as the last layer of the U-net backbone. We denote the weights of the U-net backbone θ^B, the weights of the skull patch estimation branch θ^{SP}, and the weights of the implant estimation branch θ^I. The outputs of the coarse CNN in the multi-branch model are given as

$$y^{SP}_{coarse} = g(x_{coarse}; \theta^B_g, \theta^{SP}_g) \tag{3}$$

$$y^{I}_{coarse} = g(x_{coarse}; \theta^B_g, \theta^{I}_g), \tag{4}$$

where y^{SP}_{coarse} denotes the coarse shape estimate of the skull patch and y^{I}_{coarse} denotes the coarse shape estimate of the implant. These two coarse shape estimates are then both used as an input into the high-resolution CNN, along with the high-resolution shape of the input skull. For the high-resolution CNN, the outputs are given as

$$y^{SP}_{high-res} = f(x_{high-res}, y^{SP}_{coarse}, y^{I}_{coarse}; \theta^B_f, \theta^{SP}_f) \tag{5}$$

$$y^{I}_{high-res} = f(x_{high-res}, y^{SP}_{coarse}, y^{I}_{coarse}; \theta^B_f, \theta^{I}_f) \tag{6}$$

Such architecture ensures that although two slightly different types of shape outputs can be produced by the model at both resolutions, the shared U-net backbone is forced to learn to extract meaningful local features that are suitable for correct shape estimation on both datasets.

During the training of the multi-branch CNN cascade, we use mixed mini-batches containing two samples from the Skullbreak dataset and two samples from the in-house dataset. Accordingly, two loss components are computed at each step: one for the skull patch estimation branch output y^{SP} using the Skullbreak samples, and one for the implant estimation branch output y^I using the in-house dataset samples. These loss components are then added together before updating the respective CNN weights. The iterative training of coarse and high-resolution model weights is otherwise the same as in the baseline method described in Section 2 and the multi-branch model overview is shown in Fig. 3.

2.4. Metrics

For the sake of the quantitative evaluation, we assume that the expert-designed shapes in the test set represent the only correct solution to the shape estimation tasks. This means that the quality of the output can be quantified using segmentation metrics such as volumetric overlaps (i.e. Dice coefficient) and surface distance [22]. However, it should be noted that the shape reconstruction task is specific in allowing some variability in the reconstructed shape in some cases, as long as there are no imperfections along the fit of the reconstructed shape to the input shape. See Appendix A for an illustration of how different segmentation metrics correlate with a subjective quality score of an expert implant designer. For these reasons, we evaluate the automatic reconstruction outputs using multiple different metrics in this work.

In the case of implant shape evaluation, we use the Dice coefficient and average surface distance for quantification of the estimated implant

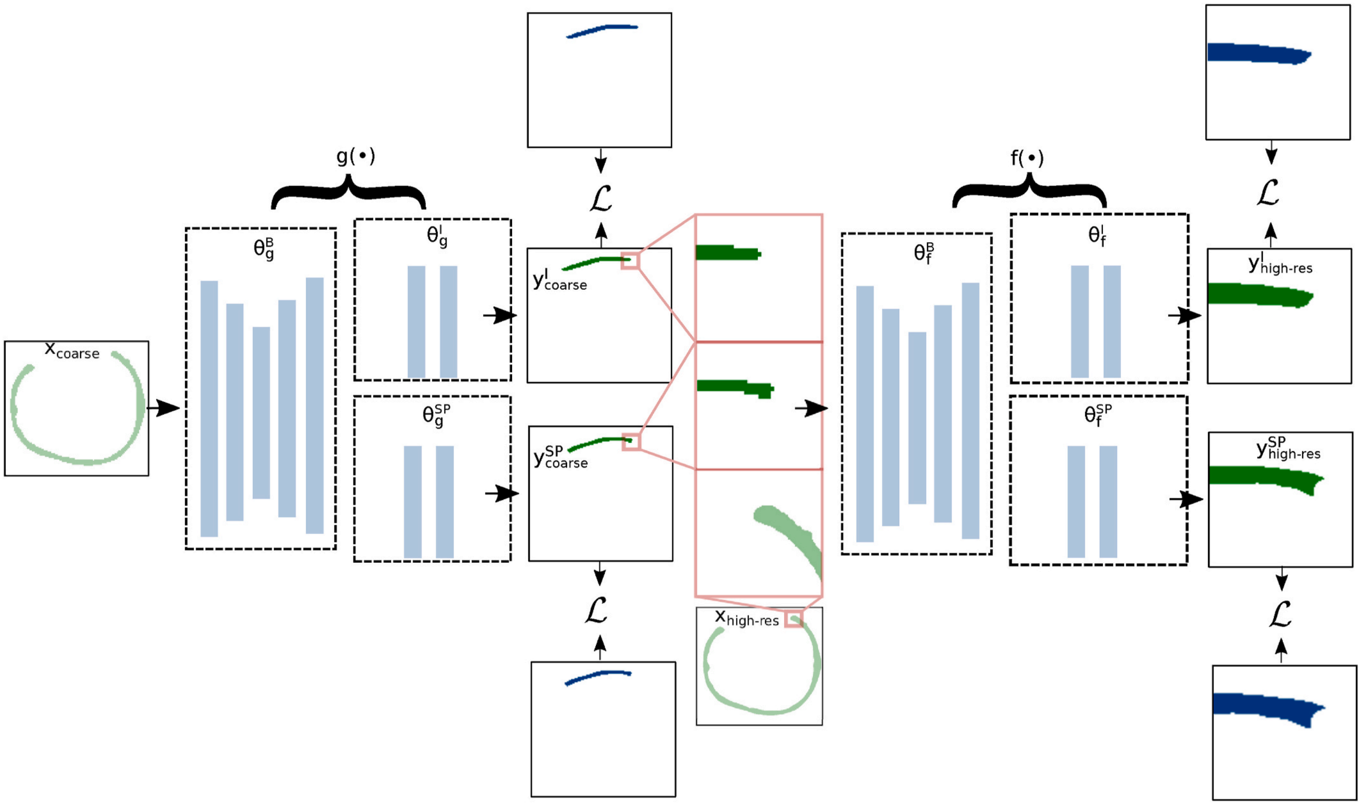

Fig. 3. Illustration of the multi-branch CNN cascade training process. Inputs and outputs of the network in light and dark green colors, respectively, and ground-truth shapes in blue. In each training step, the coarse network weights are first updated using the sum of the coarse losses, and then both coarse and high-resolution network weights are updated using the sum of the high-resolution losses.

shape quality, similarly to recent relevant works [12,23,6]. In the case of skull patch shape evaluation, however, the expert-designed ground truth patches and model outputs have different characteristics and this prevents us from using these metrics directly (see Fig. 2). Because the thickness of the ground-truth patch is different from the thickness of the original bone in the Skullbreak dataset, we measure average surface error only at the outer surface voxels of the skull.

We pay special attention to the quality of fit along the defect border of the skull patches. Similar to other authors [12], we report the outer surface distance computed along the defect border. However, this metric may not precisely convey some types of common errors of skull reconstruction which have an impact on the aesthetic outcome of cranioplasty, such as slight trenches or bumps on the surface along the defect border. To this end, we compare approximate Gaussian curvatures of reconstructed skulls and reference skulls along the defect border to supply this information.

Gaussian curvature is routinely used in 3D model surface analysis literature [24]. For simplification, we chose to approximate the Gaussian curvature error of the reconstructed skull shapes by first smoothing the binary images of skull shapes with a Gaussian blur with $\sigma = 5$, then normalizing back to a range between 0 and 1 and computing the Gaussian curvature K_i at each voxel i using the following equation:

$$K_i = -\frac{\begin{vmatrix} H(F_i) & \nabla F_i^T \\ \nabla F_i & 0 \end{vmatrix}}{|\nabla F_i|^4} \tag{7}$$

where F is the blurred volume containing the skull, ∇F_i is the vector of the first-order spatial differences in voxel i and $H(F_i)$ is the square matrix of the second-order spatial differences in voxel i [25]. The resulting Gaussian curvature volumes are then compared directly by computing voxel-wise squared error and we report the mean of this error computed along the defect border voxels as

$$MSE_K = \frac{1}{N_B}\sum_{i\in B}\left(K_i^{ref} - K_i^{pred}\right)^2 \tag{8}$$

where B is the set of outer border voxels of the predicted patch and N_B is their count. Although the exact result of this method is partially dependent on voxel resolution, value σ and on the absolute distance between the reconstructed and the reference skull surfaces, it eliminates the need for finding exact vertex correspondences and our experiments show that high resulting values correspond to dented or uneven parts of the surfaces.

3. Results

The baseline implant model was trained using the 312 training implant shapes from the in-house dataset and the baseline skull patch model was trained using the 570 Skullbreak training data samples. Because we noticed that the average size of the Skullbreak skulls differs from the average size of the in-house test skulls, we trained another baseline skull patch model on a modified version of the Skullbreak dataset that was rescaled to match the average height, length, and breadth of the in-house skulls. The multi-branch model was trained using a combination of the in-house and the rescaled Skullbreak dataset. Outputs of all models were morphologically denoised by removing smaller connected components and shape artifacts [26] before comparing them to the reference expert-designed shapes in the in-house test set. All models were implemented in Python programming language using the PyTorch[1] framework and the results were rendered using the

[1] https://pytorch.org.

Visualization Toolkit.[2]

3.1. Implant shape estimation performance

The implant shapes produced by the baseline implant model reached an average Dice coefficient of 0.85 ± 0.10 and average surface error of 0.77 ± 0.44 mm, confirming that it is possible to learn the direct mapping of defective skull shapes to the final cranial implant shapes using the CNN cascade. However, because central and fronto-orbital defects are not well represented in the in-house training dataset, the baseline implant model fails to correctly estimate implant shapes in these cases, as shown in Fig. 4. This issue may be amplified by the fact that the coarse CNN model learns to rely too much on the mirrored input to provide initial information about the missing shape, leading to overfitting and an inability to correctly deal with bilateral defects.

The implant estimates of the multi-branch model reached an average Dice coefficient of 0.88 ± 0.07 and an average surface error of 0.65 ± 0.33 mm, showing an increase in accuracy and decreased variability of output shape quality. Closer inspection of the outputs reveals an increased success rate of bilateral and fronto-orbital implant shape estimation. This can be attributed to better generalization of the U-net backbone which needs to account for more diverse defect positions in the Skullbreak dataset. Several example implant shape estimates from both the baseline implant model and the multi-head model implant estimation branch are shown in Fig. 4. The distribution of Dice coefficients and average surface distances achieved by both models can be found in Fig. 7 (top).

3.2. Skull patch estimation performance

The skull patches produced by the baseline skull patch model trained on the original Skullbreak data resulted in an average outer surface error of 0.98 ± 0.45 mm on the in-house test set. Rescaling the Skullbreak training skulls to match the average size of the in-house skulls decreased the error by 15% to 0.83 ± 0.38 mm, supporting the hypotheses that the model learns the average skull shape of the training data. However, the skull patch estimates still produced shapes with a high surface error and occasional artifacts such as holes and uneven surfaces, especially in cases of large defects. One of the causes may be the fact that the defects in the Skullbreak dataset do not fully cover the lower areas of the skull. This could be addressed by extending the dataset with additional synthetic defects, but Figs. 5 and 6 show that there are multiple different sources of error.

The skull patch estimates produced by the multi-branch CNN model further decreased the average surface error to 0.67 ± 0.37 mm. In addition to a lower amount of visible holes and artifacts in the estimated shapes, the multi-branch model also predicted the skull patches with an overall lower outer surface distance from the reference expert-designed patches, as shown in Fig. 5. The distributions of all error metrics for the three models are shown in Fig. 7 (bottom).

Interestingly, the multi-branch model output also reached a lower defect border surface error of 0.75 mm, compared to 0.96 mm and 0.94 mm for the baseline models trained on the original and the rescaled Skullbreak dataset, respectively. Similarly, the Gaussian curvature errors of the baseline skull patch model trained on the original Skullbreak and on the rescaled Skullbreak datasets also did not differ significantly, but the curvature error decreased by around 12% in the case of the multi-branch model skull patch branch outputs. This suggests that the multi-branch model learned to better fit the reconstructed skull patches to the more complex borders of the in-house defective skulls, despite only encountering the corresponding implant shapes with spatial tolerance along the border during training (see Fig. 2). Fig. 6 shows how the Gaussian curvature error reacts to different types of surface errors compared to the distance-based metrics, helping to visually identify problematic regions of the skull patch shape reconstruction outputs.

3.3. Statistical analysis

We performed a statistical analysis to report the significance of the performance gain achieved by the multi-branch CNN. The statistical significance levels are shown in Fig. 7.

A one-sided paired *t*-test was used to test the hypothesis that the error measurements of the multi-branch CNN outputs were significantly lower (or higher in the case of Dice coefficient) than in the case of the baseline CNN outputs. For the global metrics (i. e. surface distance and Dice coefficient), p-value was below the level of 0.05 for both the reconstructed skull patches and implants, which led us to accept the hypothesis that combining the data using the multi-branch CNN provides better global results when compared to the baseline models which use only one type of training data.

In the case of the border error metrics (i. e. border distance and Gaussian curvature error), the hypotheses could not be accepted using *t*-test as the p-values were over 0.05. This is likely because the shape artifacts along the border were often concentrated into a relatively small area (see Fig. 6), which resulted in a smaller quantitative difference. Therefore, we used a non-parametric Wilcox sign test to test whether the proposed approach lowers the border error when compared to the baseline methods. The hypothesis was accepted, showing that albeit small, the border error reduction is consistent across the test cases.

4. Discussion and conclusions

CNN-based skull reconstruction methods are becoming a hot topic in medical imaging. One of the major drawbacks in the current research is that the reconstruction outputs are most often evaluated on a held-out synthetic dataset in which similar anatomical variability and defect shape and type distribution can be ensured. One of the goals of this study was to illustrate the behavior of CNN-based skull reconstruction models trained on an easily accessible synthetic dataset when evaluated on real patient data. Our experiments showed that the transfer of the trained CNN model to a different population can negatively affect the reconstruction quality. Furthermore, by looking at differences in Gaussian curvature, we found that the shape complexity of the defect border in real clinical data can cause faults in the smoothness of the resulting surface.

We showed that when training the model on real clinical patient data, synthetic data can be effectively leveraged using the proposed multi-branch CNN model to significantly improve the model performance and compensate for common issues of clinical patient datasets (i. e. data scarcity and imbalance). Although a similar effect could possibly be achieved by collecting a vast amount of well-balanced clinical data, or by perfectly matching their distribution by meticulously tailoring synthetic data, we believe that the proposed approach of combining a large amount of imperfect synthetic data and a limited set of target clinical data is generally simpler and easily extendable to different types of cranioplasty data, for example, different population, additional defect areas such as the orbital floor or zygomatic bone or even different preferences for the final implant shape. The error of the outer surface of reconstructed skulls achieved by the proposed method is higher than some other recent works evaluated on synthetic defects [12,4]. However, we believe that factors such as a higher average area of the defects in our test set may be the cause and that the results are overall very promising.

The synthetic and clinical datasets used in this work contained different types of ground truths: the original missing skull patch shape and final cranial implant shapes. This allowed us to automatically produce 3D printable and directly implantable shapes, although this use case will require further evaluation of the clinical applicability in cooperation with experienced implant designers. More importantly, the

[2] https://vtk.org.

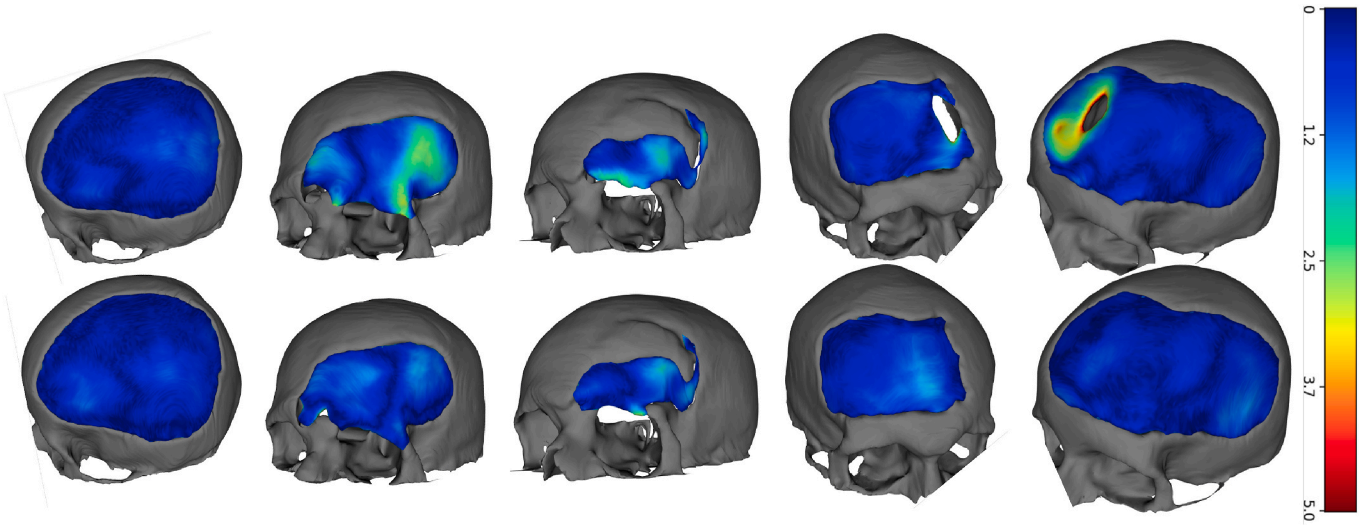

Fig. 4. Implant estimates of the baseline implant model (top) and the multi-branch model (bottom).

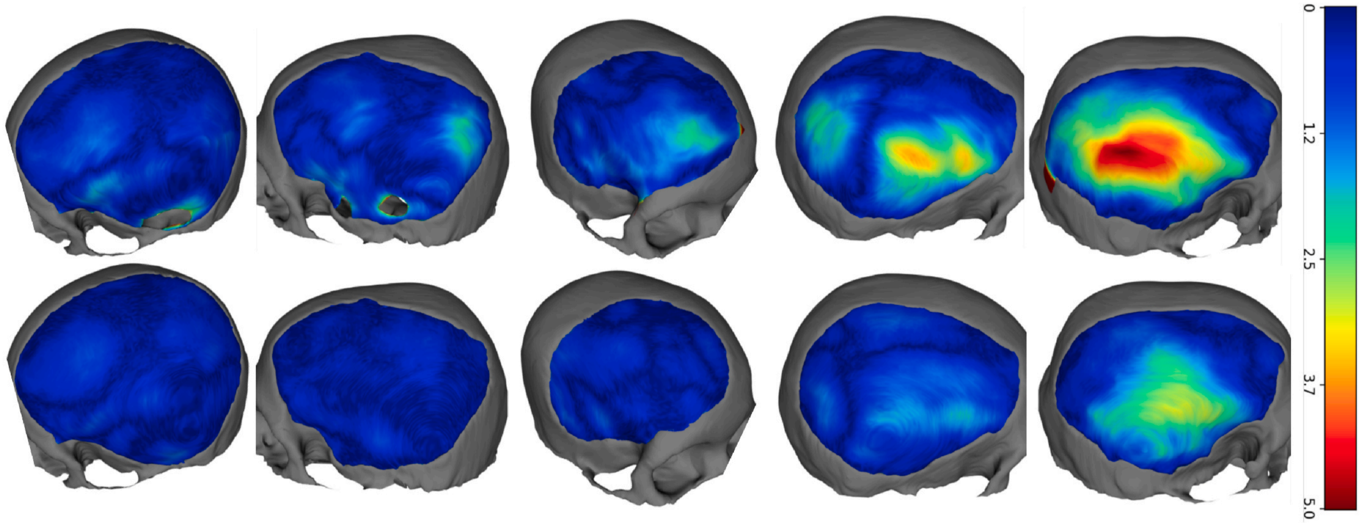

Fig. 5. Estimated skull patches of the baseline skull patch model (top) and the multi-branch model (bottom).

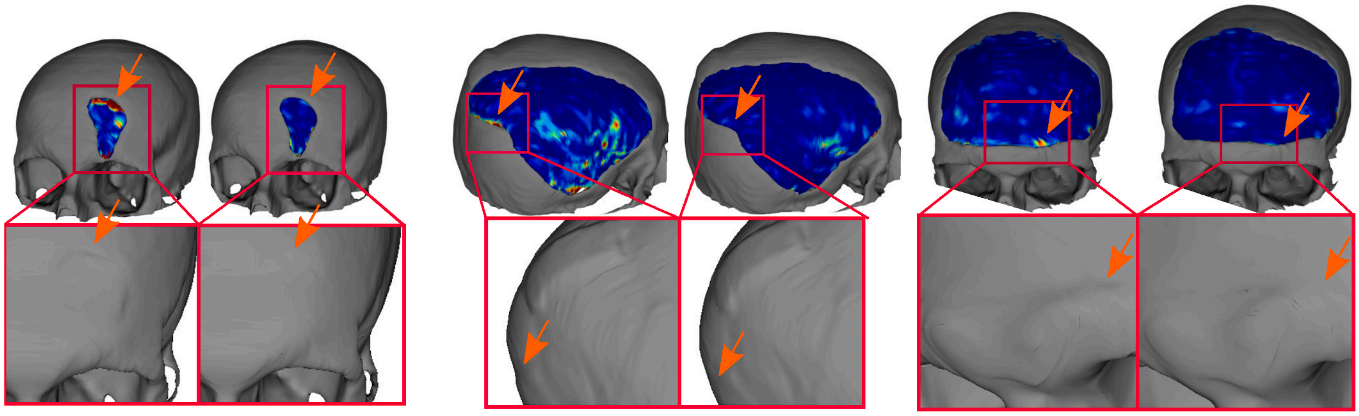

Fig. 6. Three example pairs of baseline skull patch model outputs and multi-branch model skull patch outputs, respectively, with color-coded Gaussian curvature error. The 3D models were rendered using the marching cubes algorithm and post-processed using quadratic decimation and normal smoothing. The multi-branch model can produce smoother results with lower curvature error. Note that we show the entire Gaussian curvature error maps for illustration while only defect border voxels are taken into account when computing the mean errors.

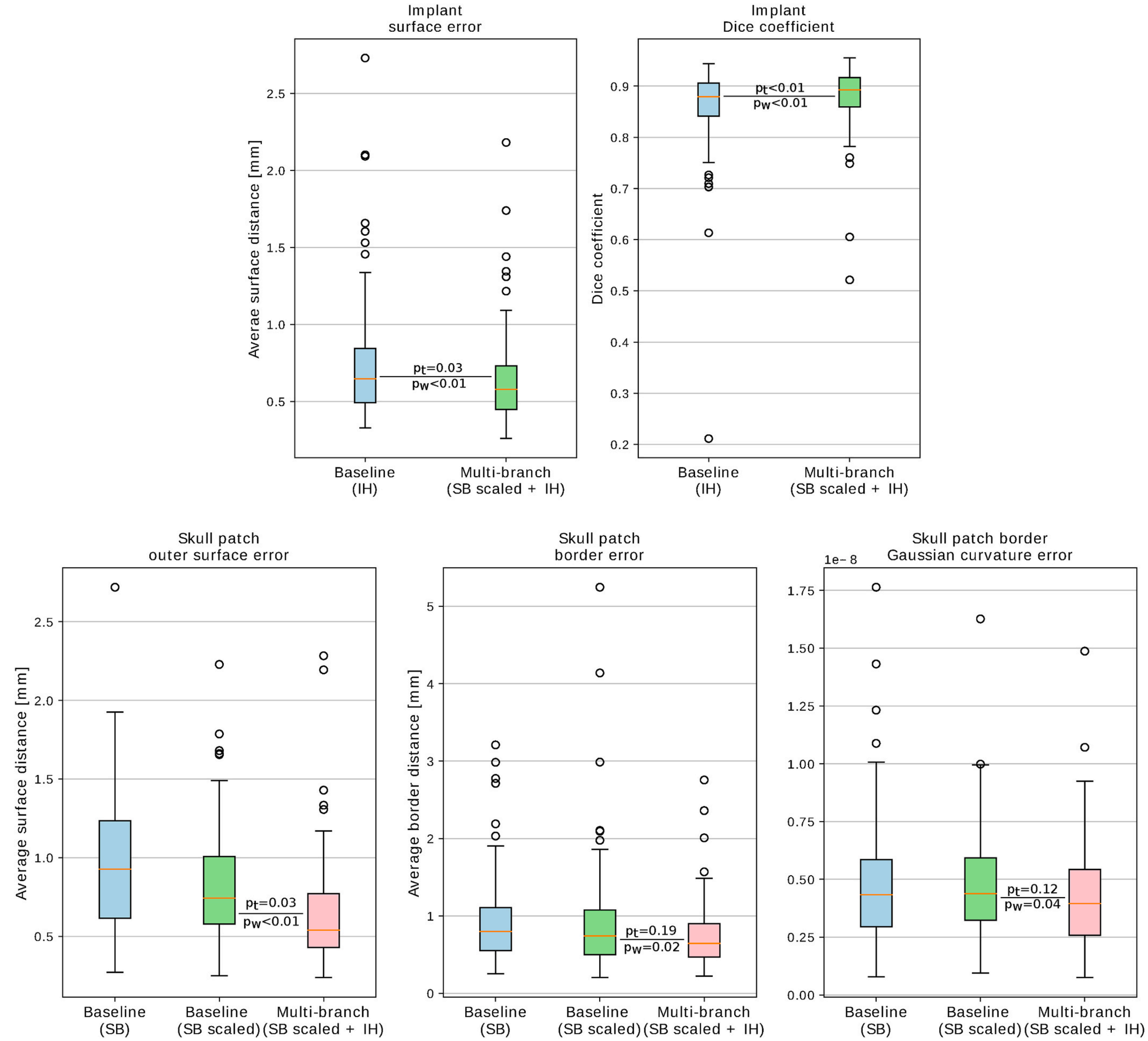

Fig. 7. Boxplots showing the error distributions of the evaluated models trained on corresponding datasets (Skullbreak - SB, in-house - IH). Average surface error and Dice coefficient of the implant estimates (top) and average outer surface error, border error, and Gaussian curvature error of the skull patch estimates (bottom). p_t and p_w values denote statistical significance of one-sided paired t-test and one-sided Wilcox test, respectively.

general ability of the model to combine cranioplasty data from different sources and of different types can accelerate the adoption of the automatic reconstruction methods by allowing training on specific target datasets while exploiting the advantages of available synthetic datasets.

To our best knowledge, this was the first study that evaluated CNN-based skull reconstruction on a real clinical dataset of this size. The proposed multi-branch CNN cascade increased the reconstructed shape quality by allowing training on more data when compared to the individual baseline models. Although the results of this study are promising from a quantitative perspective, they will need to be evaluated next by an experienced implant designer to ascertain their clinical value.

Funding

This work was partially funded by TESCAN 3DIM company.

Ethics approval

All studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent

All data were evaluated retrospectively and processed in compliance with General Data Protection Rule (GDPR). Formal consent is not required for this type of study.

Declaration of competing interest

Michal Španěl is the CEO of TESCAN 3DIM company at the time of writing this article.

Oldřich Kodym has no conflict of interest to declare.

Adam Herout has no conflict of interest to declare.

Acknowledgements

Part of the dataset was provided by TESCAN Medical company in the form of anonymized CT scans. We also gratefully acknowledge the support of the NVIDIA Corporation with the donation of one NVIDIA TITAN Xp GPU for this research.

Appendix A. Correlation Analysis of Quantitative Metrics and Subjective Expert Score of Automatic Skull Reconstructions

This section illustrates how well the quantitative segmentation metrics can predict the usability of automatic skull reconstruction results in clinical practice. We created a dataset of automatically reconstructed defective skulls and submitted it to an expert with experience in the field of skull reconstruction and implant design for subjective quality evaluation. Comparing these subjective expert scores with metrics of similarity between the reconstructed and the original shape can give an idea of what to look for when evaluating the reconstructions.

A.1 Skull Data and Reconstruction

The skull data come from the SkullBreak and SkullFix datasets [18], so the ground truth original shapes are available. A CNN-based reconstruction of the missing shape [4] was performed on each skull. Because we would ideally want to cover for this analysis the whole quality spectrum from bad reconstructions to very good reconstructions, we included the following types of reconstructed cases:

- SkullFix test case reconstructions
- SkullFix additional test case reconstructions
- SkullBreak test case reconstructions
- SkullBreak training case reconstructions (to include several close-to-perfect reconstructions)
- SkullBreak test case reconstructions using generative model [4] (to include multiple different reconstructions for a single case, including visibly bad ones)

This resulted in a total of 35 skulls. The expert assigned a score on a scale from zero to ten to each of the reconstructions, where zero corresponded to unacceptable reconstruction and ten to a nearly perfect result.

A.2 Global Metrics

We first computed correlation coefficients between the subjective expert score and routinely used segmentation metrics, including volumetric Dice coefficient and average symmetric surface distance. We also included the surface distance computed at the outer surface of the skull, since it is the most important aspect for subsequent implant modeling steps [27]. The outer surface was used in the evaluation of some previous works [12] and also in this work because of the shape characteristics of in-house ground truth data.

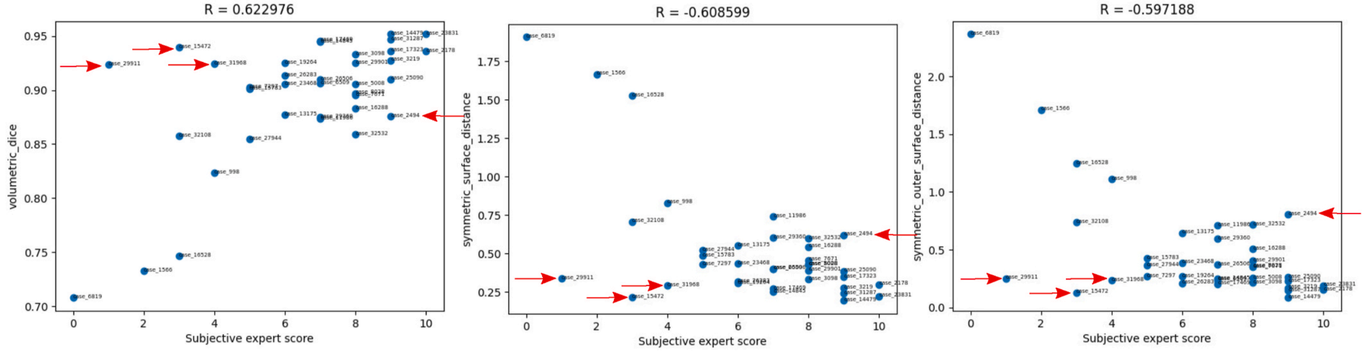

Fig. 8. Plots of the three global quantitative metrics plotted against the corresponding expert subjective score. Note that in some cases (highlighted by red arrows), the metrics failed to estimate the practical usability of the reconstruction result.

Figure 8 shows that these global metrics correlate with the expert subjective score with correlation coefficients around 0.6, confirming that they are appropriate for the comparison of different reconstruction methods. However, it can be noted that their correlation is weak when the subjective expert score is high, making it impossible to use them for discrimination between good and perfect results. Also, several cases satisfy the quantitative metrics while being seen as low-quality by experts and vice versa (see cases highlighted in red in Fig. 8).

A.3 Defect Border Metrics

The smoothness of the surface closest to the defect border has a significant impact on the aesthetic outcome of cranioplasty. We study two metrics that focus on this area: outer surface distance of the defect border and mean square error of Gaussian curvature. The defect border is defined as a set of outer surface voxels of the reconstructed skull patch shape in direct contact with the defective skull.

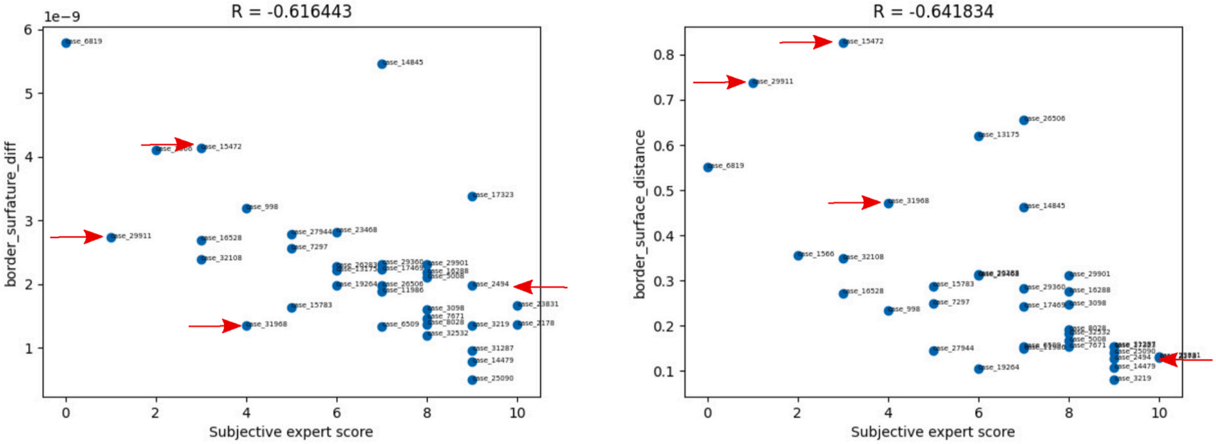

Fig. 9. Plots of border Gaussian curvature error (left) and border surface distance (right) plotted against the corresponding expert subjective score. The same cases are highlighted as in the case of global metrics, showing that the border metrics convey different yet relevant information about the reconstruction result.

Figure 9 shows that both of these metrics correlate with the subjective expert score similarly or slightly more than the global metrics. Most importantly, it can be seen that the border metrics indeed convey different information. Although the quantitative border metrics do not always agree with the subjective quality score, the correlation with the expert score was higher in the cases where the correlation of the global metrics was low.

This study was performed using only one type of automatic reconstruction method and the results were evaluated by a single implant design expert, which leaves much room for more extensive studies. However, it can be concluded that to best gauge the quality of results of automatic skull reconstruction, different types of quantitative metrics should be combined together, and both global and border metrics should be taken into account.

References

[1] M.S. Gilardino, M. Karunanayake, T. Al-Humsi, A. Izadpanah, H. Al-Ajmi, J. Marcoux, J. Atkinson, J.P. Farmer, A comparison and cost analysis of cranioplasty techniques, J. Craniofac. Surg. 26 (1) (2015) 113–117, https://doi.org/10.1097/scs.0000000000001305.

[2] E.B. da Silva Júnior, A.H. de Aragão, M. de Paula Loureiro, C.S. Lobo, A.F. Oliveti, R.M. de Oliveira, R. Ramina, Cranioplasty with three-dimensional customised mould for polymethylmethacrylate implant: a series of 16 consecutive patients with cost-effectiveness consideration, 3D Printing in Medicine 7 (1) (2021), https://doi.org/10.1186/s41205-021-00096-7.

[3] C.H. Cheng, H.Y. Chuang, H.L. Lin, C.L. Liu, C.H. Yao, Surgical results of cranioplasty using three-dimensional printing technology, Clin. Neurol. Neurosurg. 168 (2018) 118–123, https://doi.org/10.1016/j.clineuro.2018.03.004.

[4] O. Kodym, M. Španěl, A. Herout, Skull shape reconstruction using cascaded convolutional networks, Comput. Biol. Med. 123 (2020) 103886, https://doi.org/10.1016/j.compbiomed.2020.103886.

[5] J. Li, J. Egger, Towards the automatization of cranial implant design in cranioplasty: first challenge, AutoImplant 2020, held in conjunction with MICCAI 2020, Lima, Peru, october 8, 2020, proceedings, Lect. Notes Comput. Sci. (2021), https://doi.org/10.1007/978-3-030-64327-0. Springer International Publishing.

[6] F. Matzkin, V. Newcombe, S. Stevenson, A. Khetani, T. Newman, R. Digby, A. Stevens, B. Glocker, E. Ferrante, Self-supervised skull reconstruction in brain CT images with decompressive craniectomy, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, 2020, pp. 390–399, https://doi.org/10.1007/978-3-030-59713-9_38.

[7] X. Chen, L. Xu, X. Li, J. Egger, Computer-aided implant design for the restoration of cranial defects, Sci. Rep. 7 (1) (2017), https://doi.org/10.1038/s41598-017-04454-6.

[8] Y.W. Chen, C.T. Shih, C.Y. Cheng, Y.C. Lin, The development of skull prosthesis through active contour model, J. Med. Syst. 41 (10) (2017), https://doi.org/10.1007/s10916-017-0808-2. DOI 10.1007/s10916-017-0808-2, URL.

[9] Y. Volpe, R. Furferi, L. Governi, F. Uccheddu, M. Carfagni, F. Mussa, M. Scagnet, L. Genitori, Surgery of complex craniofacial defects: a single-step AM-based methodology, Comput. Methods Progr. Biomed. 165 (2018) 225–233, https://doi.org/10.1016/j.cmpb.2018.09.002.

[10] A. Marzola, L. Governi, L. Genitori, F. Mussa, Y. Volpe, R. Furferi, A semi-automatic hybrid approach for defective skulls reconstruction, Computer-Aided Design and Applications 17 (1) (2019) 190–204, https://doi.org/10.14733/cadaps.2020.190-204.

[11] A. Marzola, M. Servi, Y. Volpe, A reliable procedure for the construction of a statistical shape model of the cranial vault, in: Lecture Notes in Mechanical Engineering, Springer International Publishing, 2019, pp. 788–800, https://doi.org/10.1007/978-3-030-31154-4_67.

[12] M.A. Fuessinger, S. Schwarz, C.P. Cornelius, M.C. Metzger, E. Ellis, F. Probst, W. Semper-Hogg, M. Gass, S. Schlager, Planning of skull reconstruction based on a statistical shape model combined with geometric morphometrics, Int. J. Comput. Assist. Radiol.Surg. 13 (4) (2017) 519–529, https://doi.org/10.1007/s11548-017-1674-6.

[13] M.A. Fuessinger, S. Schwarz, J. Neubauer, C.P. Cornelius, M. Gass, P. Poxleitner, R. Zimmerer, M.C. Metzger, S. Schlager, Virtual reconstruction of bilateral midfacial defects by using statistical shape modeling, J. Cranio-Maxillofacial Surg. 47 (7) (2019) 1054–1059, https://doi.org/10.1016/j.jcms.2019.03.027.

[14] W. Semper-Hogg, M.A. Fuessinger, S. Schwarz, E. Ellis, C.P. Cornelius, F. Probst, M. C. Metzger, S. Schlager, Virtual reconstruction of midface defects using statistical shape models, J. Cranio-Maxillofacial Surg. 45 (4) (2017) 461–466, https://doi.org/10.1016/j.jcms.2016.12.020.

[15] A. Morais, J. Egger, V. Alves, Automated computer-aided design of cranial implants using a deep volumetric convolutional denoising autoencoder, in: Advances in Intelligent Systems and Computing, Springer International Publishing, 2019, pp. 151–160, https://doi.org/10.1007/978-3-030-16187-3_15. DOI 10.1007/978-3-030-16187-3_15, URL.

[16] G. von Campe, K. Pistracher, Patient specific implants (PSI), in: Towards the Automatization of Cranial Implant Design in Cranioplasty, 2020, pp. 1–9, https://doi.org/10.1007/978-3-030-64327-0_1.

[17] S. Chilamkurthy, R. Ghosh, S. Tanamala, M. Biviji, N.G. Campeau, V.K. Venugopal, V. Mahajan, P. Rao, P. Warier, Development and Validation of Deep Learning Algorithms for Detection of Critical Findings in Head Ct Scans, 2018 arXiv:1803.05854.

[18] O. Kodym, J. Li, A. Pepe, C. Gsaxner, S. Chilamkurthy, J. Egger, M. Španěl, SkullBreak/SkullFix – dataset for automatic cranial implant design and a benchmark for volumetric shape learning tasks, Data in Brief 35 (2021) 106902, https://doi.org/10.1016/j.dib.2021.106902.

[19] T.T. Le, L.G. Farkas, R.C. Ngim, L.S. Levin, C.R. Forrest, Proportionality in asian and north american caucasian faces using neoclassical facial canons as criteria, Aesthetic Plast. Surg. 26 (1) (2002) 64–69, https://doi.org/10.1007/s00266-001-0033-7. DOI 10.1007/s00266-001-0033-7, URL.

[20] P. Raghavan, D. Bulbeck, G. Pathmanathan, S.K. Rathee, Indian craniometric variability and affinities, Int. J. Evol. Biol. (2013) 1–25, https://doi.org/10.1155/2013/836738, 2013.

[21] F. Milletari, N. Navab, S.A. Ahmadi, V-net: fully convolutional neural networks for volumetric medical image segmentation, in: 2016 Fourth International Conference on 3D Vision, 3DV, 2016, https://doi.org/10.1109/3dv.2016.79.

[22] V. Yeghiazaryan, I. Voiculescu, Family of boundary overlap metrics for the evaluation of medical image segmentation, J. Med. Imag. 5 (1) (2018) 1, https://doi.org/10.1117/1.jmi.5.1.015006.

[23] J. Li, A. Pepe, C. Gsaxner, G. von Campe, J. Egger, A baseline approach for AutoImplant: the MICCAI 2020 cranial implant design challenge, in: Tanveer Syeda-Mahmood (Ed.), Multimodal Learning for Clinical Decision Support and Clinical Image-Based Procedures, Springer, Cham, 2020, pp. 75–84. https://doi.org/10.1007/978-3-030-60946-7_8.

[24] H. Yamauchi, S. Gumhold, R. Zayer, H.P. Seidel, Mesh segmentation driven by Gaussian curvature, Vis. Comput. 21 (8–10) (2005) 659–668, https://doi.org/10.1007/s00371-005-0319-x.

[25] R. Goldman, Curvature formulas for implicit curves and surfaces, Comput. Aided Geomet. Des. 22 (7) (2005) 632–658, https://doi.org/10.1016/j.cagd.2005.06.005. DOI 10.1016/j.cagd. 2005.06.005, URL.

[26] O. Kodym, M. Španěl, A. Herout, Cranial defect reconstruction using cascaded CNN with alignment, in: Towards the Automatization of Cranial Implant Design in Cranioplasty, 2020, pp. 56–64, https://doi.org/10.1007/978-3-030-64327-0_7.

[27] F.M.M. Marreiros, Y. Heuzé, M. Verius, C. Unterhofer, W. Freysinger, W. Recheis, Custom implant design for large cranial defects, Int. J. Comput.Assist. Radiol.Surg. 11 (12) (2016) 2217–2230, https://doi.org/10.1007/s11548-016-1454-8.