Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Deep learning-based framework for automatic cranial defect reconstruction and implant modeling

Marek Wodzinski[a,b,c,*], Mateusz Daniol[a,b], Miroslaw Socha[a], Daria Hemmerling[a], Maciej Stanuch[a,b], Andrzej Skalski[a,b]

[a] Department of Measurement and Electronics, AGH University of Science and Technology, Krakow, Poland
[b] MedApp S.A., Krakow, Poland
[c] Information Systems Institute, University of Applied Sciences Western Switzerland, Sierre, Switzerland

ARTICLE INFO

Article history:
Received 13 April 2022
Revised 19 August 2022
Accepted 2 October 2022

Keywords:
AutoImplant
Cranial implant design
Craniectomy
Deep learning
Image segmentation
Implant modeling
Mixed reality
Shape completion
Skull reconstruction
Personalized medicine

ABSTRACT

Background and Objective: This article presents a robust, fast, and fully automatic method for personalized cranial defect reconstruction and implant modeling.

Methods: We propose a two-step deep learning-based method using a modified U-Net architecture to perform the defect reconstruction, and a dedicated iterative procedure to improve the implant geometry, followed by an automatic generation of models ready for 3-D printing. We propose a cross-case augmentation based on imperfect image registration combining cases from different datasets. Additional ablation studies compare different augmentation strategies and other state-of-the-art methods.

Results: We evaluate the method on three datasets introduced during the AutoImplant 2021 challenge, organized jointly with the MICCAI conference. We perform the quantitative evaluation using the Dice and boundary Dice coefficients, and the Hausdorff distance. The Dice coefficient, boundary Dice coefficient, and the 95th percentile of Hausdorff distance averaged across all test sets, are 0.91, 0.94, and 1.53 mm respectively. We perform an additional qualitative evaluation by 3-D printing and visualization in mixed reality to confirm the implant's usefulness.

Conclusion: The article proposes a complete pipeline that enables one to create the cranial implant model ready for 3-D printing. The described method is a greatly extended version of the method that scored 1st place in all AutoImplant 2021 challenge tasks. We freely release the source code, which together with the open datasets, makes the results fully reproducible. The automatic reconstruction of cranial defects may enable manufacturing personalized implants in a significantly shorter time, possibly allowing one to perform the 3-D printing process directly during a given intervention. Moreover, we show the usability of the defect reconstruction in a mixed reality that may further reduce the surgery time.

1. Introduction

The process of modeling cranial implants is important for neurosurgery. The implants are used to repair defects induced by craniectomy or other sources of cranial damage. It is necessary to automate the process of implant design and production since nowadays the process takes days to weeks and induces the necessity of a follow-up intervention [1–6]. When designing the shape of the implant, it is important to adjust its surface shape and thickness to the location of the defect, as well as to adjust the edge trim, which will enable the formation of a bone-implant connection.

The process of cranial implant modeling can be improved and automated by deep learning (DL) [1,6]. The problem may be treated as a segmentation (shape completion) task where the segmented (completed) volume is the cranial defect [7]. In the DL-based approach, the computational load is transferred to the training phase and the inference process is usually fast enough to be performed in real-time. It enables one to reconstruct the defect during surgery and refine it to create an implantable model. The implant model

[*] Corresponding author at: Department of Measurement and Electronics, AGH University of Science and Technology, Krakow, Poland.
E-mail address: wodzinski@agh.edu.pl (M. Wodzinski).

https://doi.org/10.1016/j.cmpb.2022.107173

can be then 3-D printed directly during the intervention, thanks to the recent advances in medical 3-D printing [8–10].

This work focuses primarily on the automatic reconstruction of cranial defects and implant modeling, and presents applications of the calculated models in 3-D printing and mixed reality (MR) [11].

1.1. Related work

There are several important contributions to the automatic design of cranial implants in the literature. In this section, we focus on the most recent advances. We refer to the summary of the 1st edition of the AutoImplant challenge [1] for the detailed discussion about older contributions.

The 1st edition of the AutoImplant challenge inspired researchers to propose several unique solutions. The approaches were inspired mostly by DL [1]. The challenge organizers provided a baseline method [12]. The contributions were primarily based on an encoder-decoder or U-Net-like architecture [13–19], however, one method applied 2-D GANs [20] and presented that statistical shape models (SSM) may reduce the risk of overfitting and improve the generalizability. The best performing teams applied both the skull preprocessing and the defect augmentation [1,13–17]. Three contributions used shape priors with a convolutional neural network, U-Net architecture and variational auto-encoder applied with encoder-decoder network [13,14,18]. We hypothesize that the use of shape prior is beneficial for small datasets. However, with strongly augmented larger datasets the network should be able to learn it by themselves. The winner of the 1st edition presented that extending the training set by image registration may improve the results [15]. They concluded that augmenting the training set is crucial for the shape completion problem and we follow this idea in our research.

Several important contributions were introduced during the AutoImplant 2021 challenge [21]. The challenge consisted of three separate tasks: (i) Task 1 related to the reconstruction of prealigned, variously shaped defects, (ii) Task 2 considering the generalizability into real, clinical cases, and (iii) Task 3 related to the improvements in skull shape completion (using the same dataset as during the 1st edition of the challenge). One of the contributions decided to train two 3D U-Net-like networks separately on Task 1/3 [22]. They used the model trained using the Task 1 dataset in Task 2, to show its generalizability and used second-step filtering to increase the reconstruction of fine details. Another interesting contribution was based on implant prediction using the slice-by-slice approach [23]. The researchers used a recurrent neural network (RNN) to use the a priori knowledge about the continuity of the adjacent slices. The work presented in Yu et al. [24] addressed the challenge differently. The author argued that generalizability is crucial in cranial implant design, and proposed a method based on the principal component analysis (PCA), using only a subset of Task 3 data for training. Moreover, the authors used a registration-based approach to create a common image domain. Considering that a small training set was used, the method achieved superior results on Task 2, connected with the generalizability into real cranial defects. The authors of Pathak et al. [25] participated in Task 3 only and proposed a two-step procedure involving an initial bounding box search and refinement in higher resolution. Both the steps applied the 3-D V-Net. Our preliminary work [26] outperformed the other methods. We have shown that appropriate data augmentation and linking are crucial to obtaining reasonable results on different distributions. Two additional contributions came from the challenge organizers. The submissions attempted to solve the problem related to high GPU memory requirements of the DL-based contributions [27,28]. In [27], the authors observed that majority of the input voxels are uninformative. They argued that an autoencoder may be used to learn a voxel rearrangement limiting the required memory. On the other hand, in Kroviakov et al. [28], the authors decided to apply the spare convolutional neural networks [29]. Using this approach, they excluded empty voxels from computations and decreased memory consumption, allowing one to use a higher resolution input, without the necessity of downsampling.

There are also recent contributions not directly linked to AutoImplant challenge objectives [30]. In [31], the authors proposed a DL-based solution to perform the skull segmentation from computed tomography scans. The method is based on 3-D U-Net, with results further refined by graph-cut. The method was used to create the Task 1 dataset, in contrast to Task 2 and Task 3 datasets, which were created using global thresholding. Even though no significant benefits were observed related to cranial implant design, the proposed method may be beneficial for other, more irregular, and complex structures. Another work elaborates on the neurosurgeons' criteria for cranial implant feasibility [32]. The authors proposed a scoring system, used to qualitatively evaluate the Task 2 submissions. They studied the correlation between different qualitative measures and scores from experts. They concluded that the automatically modeled implants require further manual refinement and that additional post-processing may improve the defects to meet given clinical requirements. Another work [33] discussed the general insights into cranial implant design, motivation, manufacturing process, and currently used materials. The authors presented the idea that the implant may be modeled and fabricated before the intervention, based on patient-specific pathological conditions.

1.2. Contribution

The referenced works suffer from the generalizability-related issues [21]. This work presents our contribution to the automatic cranial defect reconstruction and implant modeling using a combination of traditional and deep-learning approaches. We introduce an automatic, complete pipeline allowing one to create 3-D printing files from binary volumes representing the skull with the cranial defect. We evaluate the method on public datasets introduced for the 1st and 2nd editions of the AutoImplant challenge [1,21]. The presented method scored the first prize in all the challenge tasks. The key finding is that combining different datasets by *imperfect* registration improves the training dataset diversity and outperforms other augmentation methods. Noteworthy, our method is trained purely on synthetic data, yet is still able to perform the skull reconstruction on real cranial defects. Moreover, we verify the 3-D printing capability, present a use case in mixed reality, and freely release the source code [34].

The article extends our initial contribution to the AutoImplant challenge by:

- Extending the image registration-based augmentation by the imperfect image registration.
- Proposing the VAE-based training set augmentation increasing the heterogeneity of the training set.
- Introducing the implant modeling step to make the reconstructed defects implantable.
- Completing the processing pipeline thus allowing one to generate mesh models (STL) ready for 3-D printing directly from the segmented skull with the defect.
- Performing several ablation studies comparing different training set augmentation methods.
- Presenting the 3-D printing and mixed reality use cases.
- Releasing the source code together will all experimental setups.

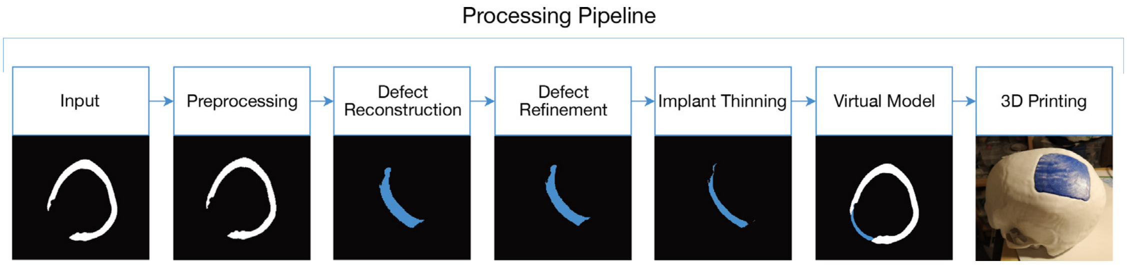


Fig. 1. Visualization of the processing pipeline. Please note that the defect reconstructions are zoomed in for presentation clarity.

2. Methods

2.1. Overview

The proposed method is a complete pipeline to perform cranial defect reconstruction, implant modeling, and preparation for 3-D printing. The pipeline consists of several steps: (i) data loading and preprocessing, (ii) deep learning-based defect reconstruction, (iii) deep learning-based defect refinement, (iv) optional implant thinning, and (v) the preparation for 3-D printing. Additionally, the steps performed to link and augment the datasets are discussed separately since they have a significant influence on the reconstruction quality. The overview of the processing pipeline is shown in Fig. 1.

2.2. Dataset

The method is evaluated on the public datasets introduced during the AutoImplant 2021 challenge [21]. Three datasets are used: (i) SkullBreak (Task 1), real cranial defects (Task 2) and SkullFix (Task 3) [35]. The Task 1 and Task 3 datasets were adapted from a public head CT collection CQ500 [36] by the challenge organizers by segmenting the skulls and creating synthetic defects [35]. The comparison of cases from all the datasets is shown in Fig. 2.

The Task 1 dataset is represented by 570 training cases created from 114 skulls and 100 testing cases created from 20 skulls. For each skull five defects are introduced: (i) bilateral, (ii) frontoorbital, (iii) parietotemporal, (iv) random, type 1, and (v) random, type 2. The goal of this task is to reconstruct defects with random shapes and positions, however, with already prealigned skulls. The ground truths for the dataset are the synthetic defects.

The Task 2 dataset consists of 11 skulls with real cranial defects. The dataset may be considered as a benchmark for the generalizability into different distributions and implant modeling. The dataset does not include any training cases. The ground truths for the dataset are the manually designed implant models.

The Task 3 dataset is created from 100 skulls for training and 110 skulls for testing. Each skull has deteriorated with one synthetic defect. The test set is further divided into 100 skulls with defects similar to the training set and 10 skulls with defects from different distributions. The ground truths for the dataset are the synthetic defects.

2.3. Preprocessing and dataset linking

The preprocessing starts with finding the boundaries of the defective skull. Then, the images are cropped with a predefined offset. The offset is required because the defect may extend beyond the bounding box of the defective skull. Then, the images are resampled to the same physical spacing (1 mm × 1 mm × 1 mm) and padded to the same, predefined shape (240 × 200 × 240). The steps are applied to cases from all datasets and allow one to use both the training sets together. The datasets are further offline augmented by cross-case image registration (IR) and variational autoencoder (VAE), discussed in Sections 2.5 and 2.6 respectively. The dataset linking and augmentation are crucial to improving the method performance, more important than the details of network architecture or the training hyperparameters.

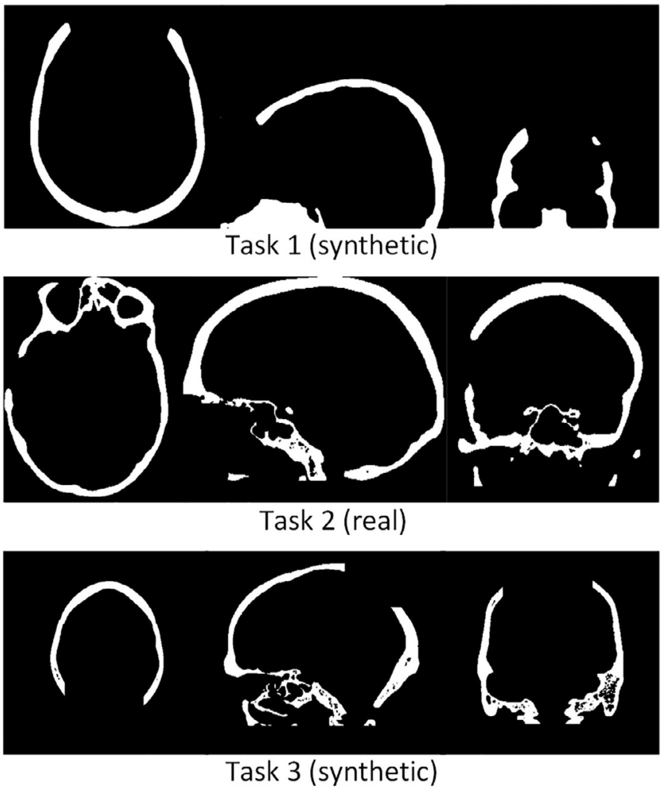


Fig. 2. Exemplary cases from the datasets introduced for Task 1, 2, and 3, respectively. Please note differences induced by distinct skull segmentation techniques and the defect synthesis (Task 1, 3).

2.4. Defect reconstruction and refinement

The defect reconstruction is performed by a U-Net-like network [37] with residual blocks. The network takes as the input the preprocessed defective skull and outputs the calculated defect. The architecture is shown in Fig. 3. We decided to calculate the defect instead of the complete skull due to slightly higher invariance to noise and smoother defect boundaries. The network is trained with the soft Dice loss as the cost function, defined as:

$$DC_{loss}(A,B) = 1 - 2\frac{A \cap B}{A + B + \alpha}, \tag{1}$$

where A, B, α are the calculated defect, the ground-truth defect, and the smoothing coefficient, respectively.

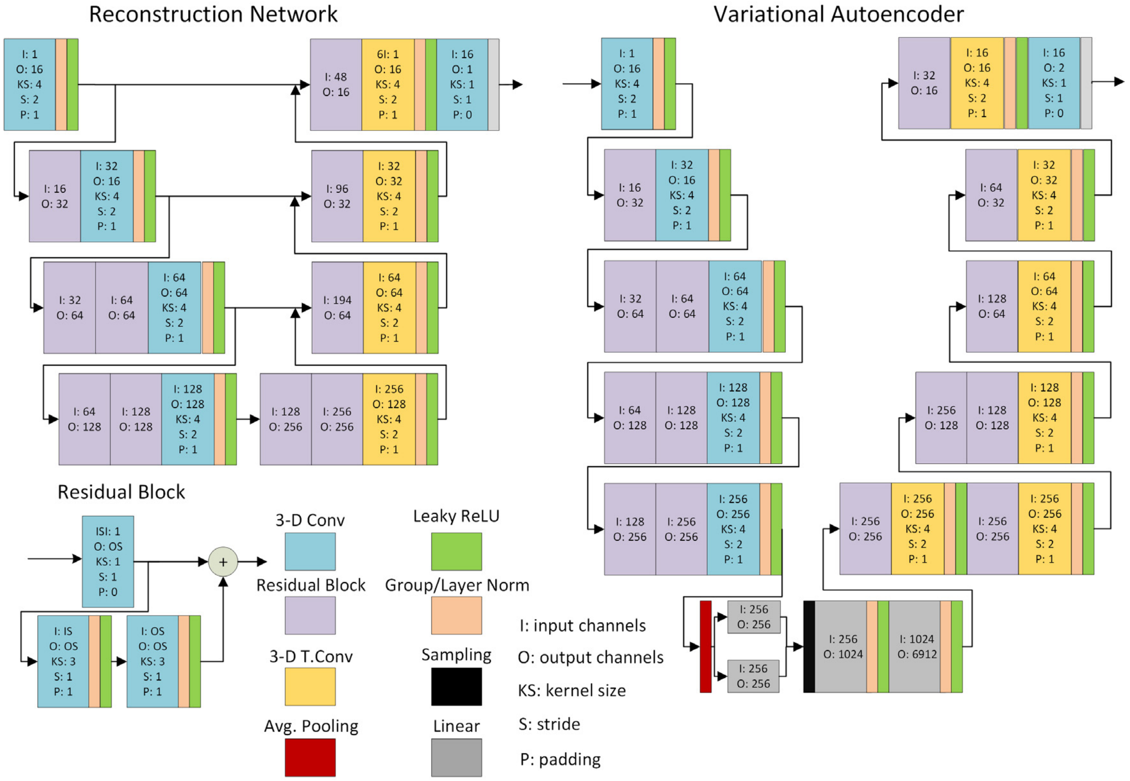

Reconstruction Network
I: 1 O: 16 KS: 4 S: 2 P: 1
I: 48 O: 16
6I: 1 O: 16 KS: 4 S: 2 P: 1
I: 16 O: 1 KS: 1 S: 1 P: 0
I: 16 O: 32
I: 32 O: 16 KS: 4 S: 2 P: 1
I: 96 O: 32
I: 32 O: 32 KS: 4 S: 2 P: 1
I: 32 O: 64
I: 64 O: 64
I: 64 O: 64 KS: 4 S: 2 P: 1
I: 194 O: 64
I: 64 O: 64 KS: 4 S: 2 P: 1
I: 64 O: 128
I: 128 O: 128
I: 128 O: 128 KS: 4 S: 2 P: 1
I: 128 O: 256
I: 256 O: 256
I: 256 O: 128 KS: 4 S: 2 P: 1
Variational Autoencoder
I: 1 O: 16 KS: 4 S: 2 P: 1
I: 32 O: 16
I: 16 O: 16 KS: 4 S: 2 P: 1
I: 16 O: 2 KS: 1 S: 1 P: 0
I: 16 O: 32
I: 32 O: 16 KS: 4 S: 2 P: 1
I: 64 O: 32
I: 32 O: 32 KS: 4 S: 2 P: 1
I: 32 O: 64
I: 64 O: 64
I: 64 O: 64 KS: 4 S: 2 P: 1
I: 128 O: 64
I: 64 O: 64 KS: 4 S: 2 P: 1
I: 64 O: 128
I: 128 O: 128
I: 128 O: 128 KS: 4 S: 2 P: 1
I: 256 O: 128
I: 128 O: 128
I: 128 O: 128 KS: 4 S: 2 P: 1
I: 128 O: 256
I: 256 O: 256
I: 256 O: 256 KS: 4 S: 2 P: 1
I: 256 O: 256
I: 256 O: 256 KS: 4 S: 2 P: 1
I: 256 O: 256
I: 256 O: 256 KS: 4 S: 2 P: 1
I: 256 O: 256
I: 256 O: 256
I: 256 O: 1024
I: 1024 O: 6912
Residual Block
ISI: 1 O: OS KS: 1 S: 1 P: 0
I: IS O: OS KS: 3 S: 1 P: 1
I: OS O: OS KS: 3 S: 1 P: 1
+
3-D Conv
Leaky ReLU
Residual Block
Group/Layer Norm
3-D T.Conv
Sampling
Avg. Pooling
Linear
I: input channels
O: output channels
KS: kernel size
S: stride
P: padding

Fig. 3. Visualizations of the architectures of the reconstruction/refinement network and the variational autoencoder.

The defect refinement is an optional step performed after the reconstruction. The goal of the refinement is to smooth the calculated defect, make the upsampling errors negligible, and learn how to recover fine details that are unavailable at lower resolution. The refinement model uses the same architecture as the reconstruction, however, takes as the input the already calculated defect with additional preprocessing. First, the calculated defect is resampled and unpadded back to the original volume shape. Then, the bounding box for the reconsidered defect is calculated, cropped with a predefined offset (equal to 10 voxels in each dimension), and resampled to the same shape (200^3). During training, the same bounding box, padding, and resampling are used for the ground-truth defect. We decided to use the defects only, without the surrounding part of the defective skull because it simplifies the problem, and all artifacts at the boundaries can be easily removed during the post-processing. The defect refinement is trained the same way as the defect reconstruction. The intuitive effect of the defect refinement is shown in Fig. 4.

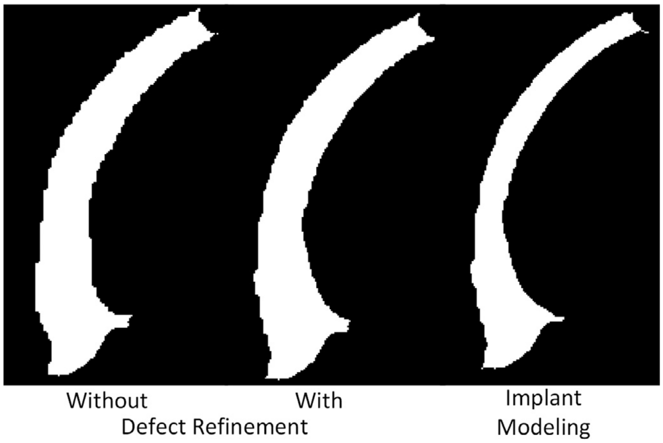


Fig. 4. Visualization of the defect refinement and implant modeling. Please note that the refinement makes the reconstruction boundaries not only smoother but also captures fine details. The modeling makes the reconstruction thinner and implantable into the cranial cavity.

2.5. Augmentation by image registration

The training sets are augmented by cross-case IR. All complete skulls from training sets are registered to each other. The registrations are performed using a multi-level, instance optimization-based approach [38]. First, an affine transformation is calculated and then followed by deformable IR. Since the registration is performed on binary images, the mean squared error (MSE) is used as the dissimilarity function. The diffusion regularization is used to regularize the nonrigid step. The cost function is defined as:

$$C_{REG}(S, T, u) = MSE(S \circ u, T) + \theta Reg(u), \tag{2}$$

where S, T are the moving and fixed complete skulls, u is the calculated displacement field, θ denotes the regularization coefficient, Reg is the diffusive regularization, and $\circ$ denotes the warping operation. The displacement field u is used to warp all images connected with the source image (the complete skull, the defective skull, and the implant) creating a new training case. The created cases are saved and added to the training set. The number of required registrations grows quadratically with the number of training cases.

Two training sets are created. One with registrations performed until convergence, with strong regularization and diffeomorphism enforcement by scaling and squaring [39] (further denoted as smooth and invertible IR), and the second performed with strongly relaxed regularization coefficient, without enforcing the deformation field invertibility, and for a predefined number of iterations

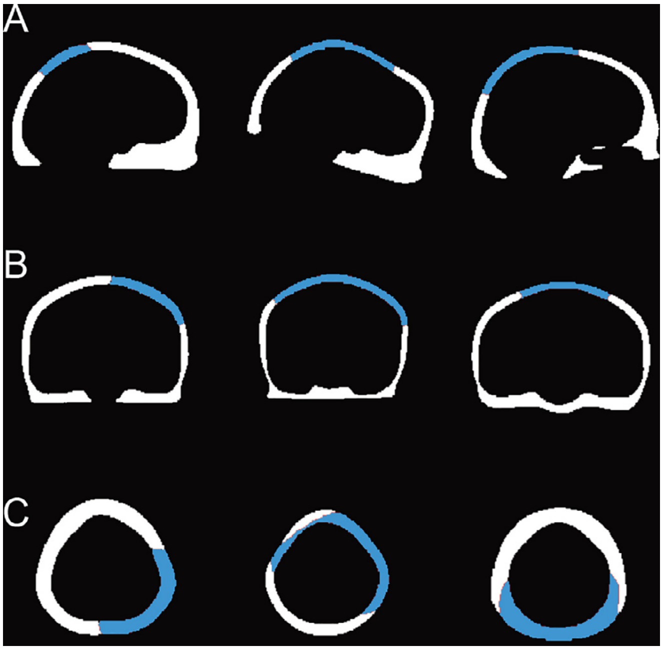


Fig. 5. Exemplary training cases generated by the VAE. (A) - side view, (B) - frontal view, (C) - top view.

(further denoted as imperfect IR). Importantly, the approach can be potentially used to augment training sets for other segmentation and classification problems, as long as the morphologies of the objects of interest are similar. This is, however, a subject of further research.

2.6. Augmentation by image generation

The training set created by smooth and invertible IR is used to train a VAE network to perform further augmentation by image generation. We use an encoder-decoder architecture with residual blocks, shown in Fig. 3. The network is trained until convergence with the objective function defined as:

$$C_{VAE}(\cdot) = DC_{loss}(I, G) + \beta KL(E, \mu, \sigma) - DC_{loss}(G_S, G_I), \quad (3)$$

where I, G are the input and the generated case respectively, KL denotes the Kullback–Leibler (KL) divergence between the latent space distribution and the normal distribution, β controls the influence of KL divergence, E is embedding sampled from the latent space distribution, μ, σ are the parameters of the latent space distribution, and G_S, G_I are the generated defective skull and the generated defect respectively.

We decided to use the VAE instead of other generative models because the desired similarity is well-defined and the generated images are binary. Therefore, the issues connected with blurring and incorrect fine details at image boundaries are not influential. The pretrained model is then used to create additional 100,000 training cases. The exemplary outputs of the VAE are shown in Fig. 5.

2.7. Postprocessing

The postprocessing is performed to reverse the geometry change during the initial preprocessing and to decrease the influence of noisy reconstructions or interpolation artifacts. First, the given case is unpadded using parameters calculated during the preprocessing. Then, it is upsampled back to the original voxel size and unpadded again to the original shape. Then, we perform morphological postprocessing consisting of binary closing and connected component analysis. The binary closing together with exclusive disjunction is used to improve the implant boundaries. The connected component analysis is used to find the largest defect. This step is optional and can be tuned to the desired number of defects. We introduced it because we found that the ground truth is always stated for only the largest defect, even if more are present.

2.8. Implant modeling

The reconstructed defect usually cannot meet the clinical requirements related to its thickness and borders regularity [32,40]. Therefore, the reconstructed defects should be additionally postprocessed to create implantable implants. We propose a simple, iterative, and tuneable procedure to refine the defect.

The procedure starts by calculating the binary contour of the defect using exclusive or between the defect and the defect after the binary erosion. Then, we calculate two centroids, one of the defective skull, and the second of the calculated contour, followed by the calculation of a normalized vector between them. The defect is iteratively transformed along the vector with a predefined step. During each iteration, after the transformation, the logical conjunction between the original defect and the transformed defect is calculated, followed by median filtering and exclusive disjunction. Finally, the connected component analysis is performed to delete potentially introduced artifacts and to calculate the relative volume ratio between the transformed and the original defect. The process terminates after the desired volume ratio is achieved. The effect of the implant modeling is shown in Fig. 4. In case of any errors, the implants may be further refined before manufacturing.

Intuitively, the method iteratively “pulls out” the calculated defect from the cranial cavity. Since the external part of the defect is deleted and the connections between boundaries are corrected by the morphological operations, the implant becomes thinner while maintaining the general shape, thus, making it correct and implantable. The desired volume ratio is set up by the end-user based on the desired implant thickness based on mechanical properties of the material that is going to be used during the 3-D printing, and other patient-specific conditions.

2.9. 3-D printing

We verify the designed defects and implants by 3-D printing. We perform the printing using the Fused Deposition Modeling (FDM) technique and the general purpose Cartesian 3-D printer Prusa i3 MK3S+. For test purposes, the objects are printed using PLA filament.

The models are post-processed to make the printing process faster and smoother. The binary outputs are smoothed using Gaussian kernel, followed by isosurface creation using Flying Edges algorithm [41]. Finally, the mesh quality is improved using sinc-based filtering and connected component analysis to delete small, unconnected artifacts.

The post-processed output of implant modeling is converted to STL geometrical representation. The STL files are processed by PrusaSlicer software to create the 3-D printer G-Code instructions. Due to the technical limitations of the 3-D printer, the skulls are divided into two parts. This enables not only an easier printing process but also the evaluation of the cranial implant from the inner of the skull. As a result, the analysis of implant boundaries and implantability is easier.

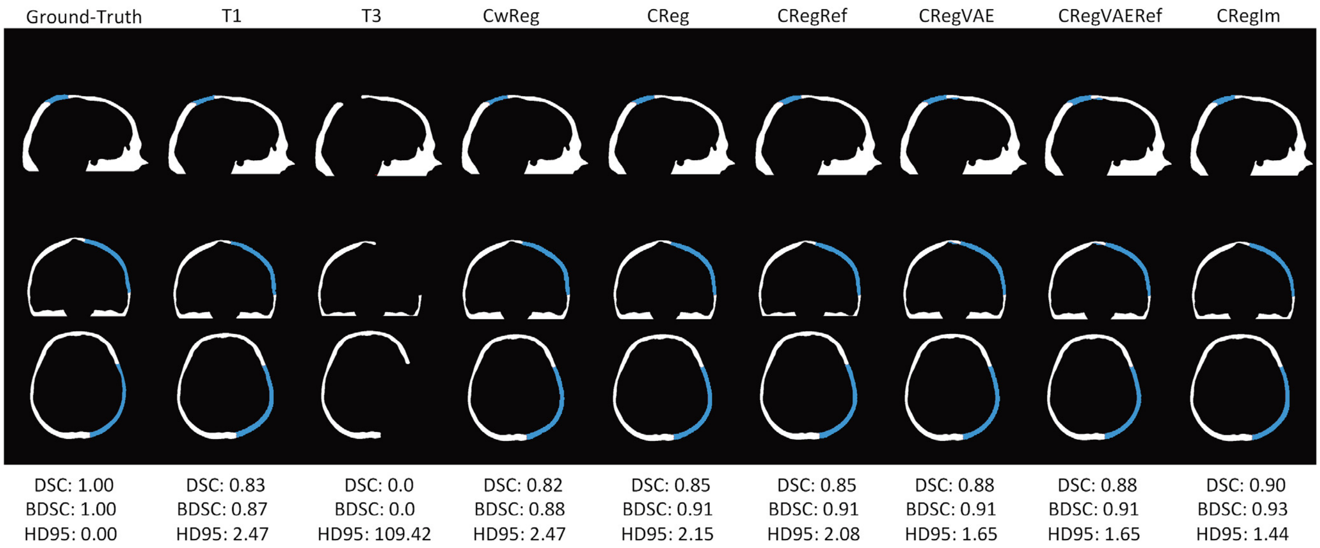


Fig. 6. Exemplary results for case from Task 1 dataset - ID 15, Defect: Random_2, (best viewed zoomed in electronic format).

2.10. Experimental setup

The pipeline is implemented in Python using PyTorch library [42], extended by PyTorch Lightning framework [43]. The postprocessing for 3-D printing is implemented using the Python VTK library. The training was performed using a cluster of 8x V100 GPUs (32 GB memory each) and a workstation with 2x RTX 3090 GPUs (24 GB memory each). The experiments were run in parallel utilizing a single GPU. The inference and evaluation were performed using a single RTX 3090 GPU. A given experiment required no longer than 120 h of single GPU training. The training cases are randomly split into training and validation sets with a 9:1 ratio. Only the training cases are used when combining the datasets or training the generative network. We perform several ablation studies concerning the training set structure:

- The Task 1 training set only (T1)
- The Task 3 training set only (T3)
- The Task 1 and Task 3 training sets combined (Cmb)
- The Task 1 and Task 3 training sets combined by smooth and invertible IR (CReg)
- The CReg with the defect refinement (CRegRef)
- The CReg further augmented by VAE (CRegVAE)
- The CRegVAE with the defect refinement (CRegVAERef)
- The Task 1 and Task 3 training sets combined by imperfect IR (CRegIm)
- An additional implant modeling (for Task 2 only) after the CRegIm (CImplant)

We also compare the presented pipeline to other state-of-the-art methods.

All the models are trained until convergence with batch size equal to 2, the number of cases per iteration is equal to 500, the initial learning rate is 0.003, and the decay rate varies from 0.95 to 0.99. Apart from the augmentation by IR or VAE, the training sets are online augmented by random affine transformations with scale, rotation, and translation ranging from 0.85 to 1.15, −15 to 15 degrees, and −10 to 10 voxels, respectively. Additional information about the source code, the experiments, and details on how to reproduce the results may be found in the repository [34].

3. Results

3.1. Defect reconstruction

The quantitative evaluation of the defect reconstruction is based on the Dice score (DSC), the boundary Dice score (BDSC), and the 95th percentile of Hausdorff distance (HD95). The BDSC is exclusively used for the AutoImplant challenge and calculates the DSC between the implant borders I_B, which are calculated as follows [21]:

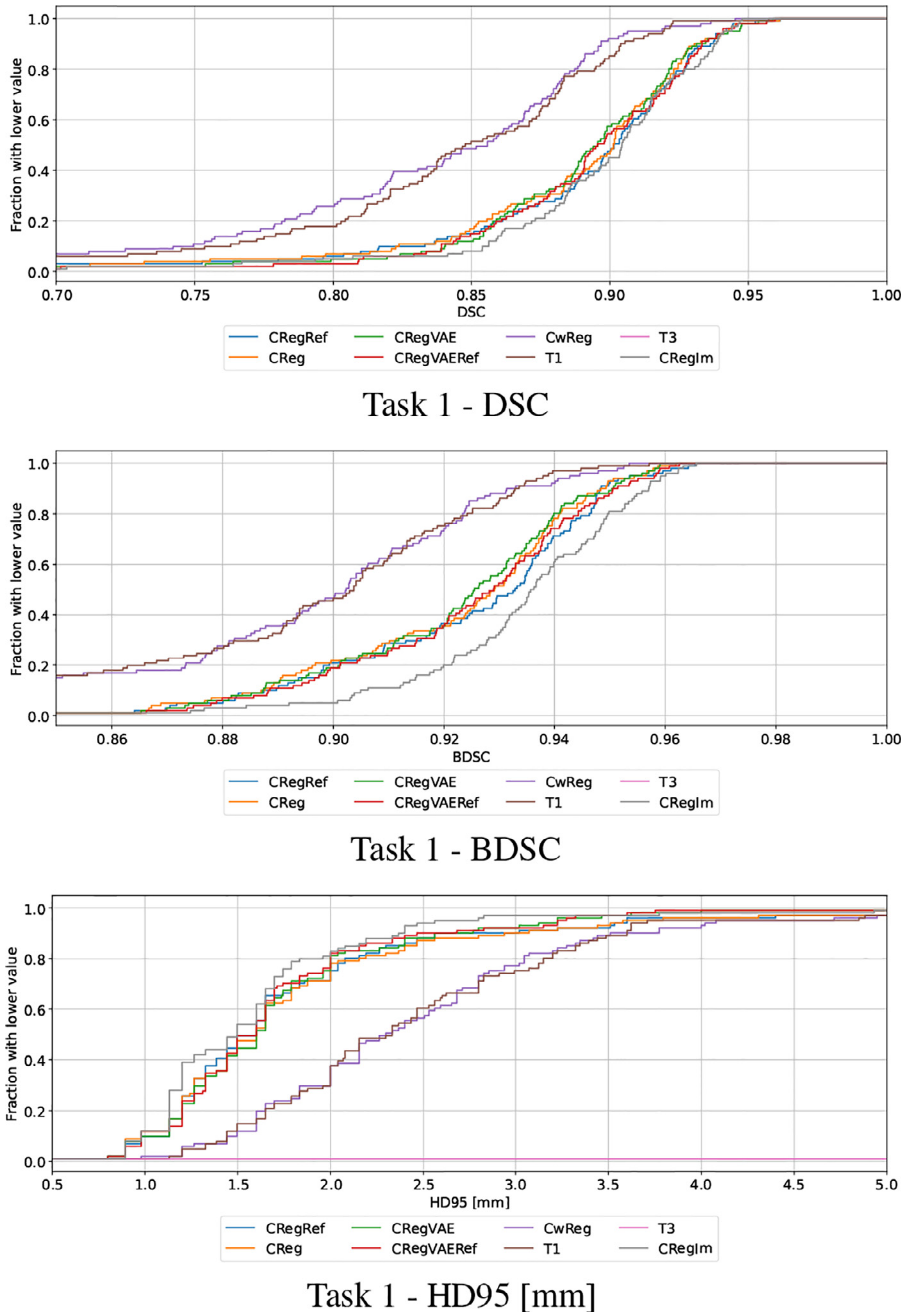
$$I_B = \begin{cases} I, & dt <= d \\ 0, & dt > d \end{cases} \tag{4}$$

where $dt = EDT(D)$ is the Euclidean distance transform (EDT) of the defective skull D, I denotes the corresponding implant, d is a pre-defined distance at which the voxels representing the implant I are considered as borders (equal to 10, defined by the AutoImplant challenge organizers). The BDSC is used to evaluate the quality of the borders reconstruction which is crucial for cranial implants.

The Figs. 7 and 8 present cumulative histograms of the DC, BDSC, and HD95 for the Task 1 and Task 3 test sets separately. Fig. 9 presents the quantitative results for Task 1, including the division into different defect types. Tables 1 and 2 show the quantitative statistics of different ablation studies, including other state-of-the-art methods. The exemplary visualization, presenting defective skulls together with the calculated defects, is shown in Fig. 6.

3.2. Implant modeling

The quantitative results of the implant modeling are presented in Table 3, for Task 2 only. There are no ground-truth implants for Tasks 1 and 3. It should be kept in mind that there may be multiple correct implants for a given defective skull (based e.g. on the desired implant thickness). Therefore the quantitative results are significantly lower than for the defect reconstruction. Additional expert assessment is required which was performed by the AutoImplant challenge organizers [21]. We show the effect of implant modeling in Fig. 10.



Task 1 - DSC

Task 1 - BDSC

Task 1 - HD95 [mm]

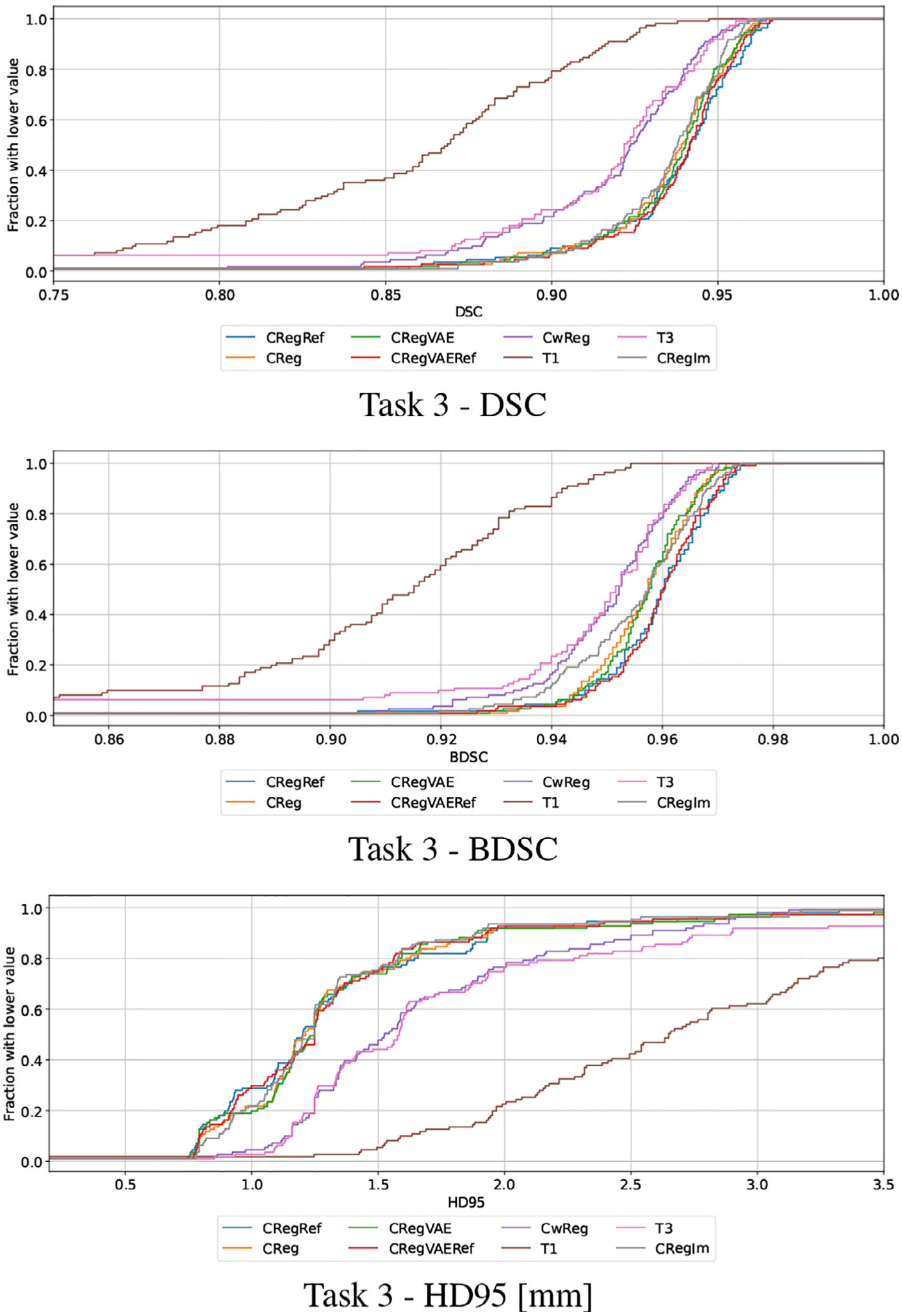
Fig. 7. Cumulative histograms for the Task 1 test set (presenting DSC, BDSC, and HD95 respectively). Note the direct impact of imperfect registration on BDSC and HD95.

Table 1
Quantitative results for Task 1 test set presenting ablation studies and comparison to other state-of-the-art methods.

Method	DSC	BDSC	HD95 [mm]	Method	DSC	BDSC	HD95 [mm]
T1	0.84	0.89	2.60	T3	0.06	0.09	98.15
CReg	0.88	0.92	1.87	CRegRef	0.89	0.92	1.86
CRegVAE	0.89	0.92	1.74	CRegVAERef	0.89	0.93	1.71
CwReg	0.84	0.89	2.51	CRegIm	**0.89**	**0.93**	**1.60**
Mahdi [22]	0.78	0.81	3.42	Yang [23]	0.85	0.89	3.52

3.3. 3-D printing

Fig. 11 presents the skulls and the corresponding defects/implants for three cases from the Task 2 set. We decided to print three cases based on the qualitative results: the best, the worst, and the moderate one. Please note that for the moderate and the best case, the implants fill the cranial defect very well and the boundaries are smooth. The reconstruction worst case was unsuccessful due to the skull being significantly different from the training distributions and large defect size. Nevertheless, it may be



Task 3 - DSC

Task 3 - BDSC

Task 3 - HD95 [mm]

Fig. 8. Cumulative histograms for the Task 3 test set (presenting DSC, BDSC, and HD95 respectively).

Table 2
Quantitative results for Task 3 test set presenting ablation studies and comparison to other state-of-the-art methods. Please note that the method by Kroviakov et al. [28] was evaluated only on a subset of the test set. Note that the results for Task 3 are indistinguishable for the majority of ablation studies since the synthetic defects are regular and similar.

Method	DSC	BDSC	HD95 [mm]	Method	DSC	BDSC	HD95 [mm]
T1	0.81	0.86	10.59	T3	0.87	0.91	4.40
CReg	0.93	0.95	1.96	CRegRef	0.93	0.95	1.93
CRegVAE	0.93	0.96	1.65	CRegVAERef	0.94	**0.96**	1.64
CwReg	0.92	0.95	2.00	CRegIm	**0.93**	0.95	**1.47**
Mahdi [22]	0.88	0.93	3.59	Yu [24]	0.77	0.77	3.68
Li [27]	0.81	–	–	Kroviakov [28]	0.85	0.95	2.64
Pathak [25]	0.90	0.95	2.02				

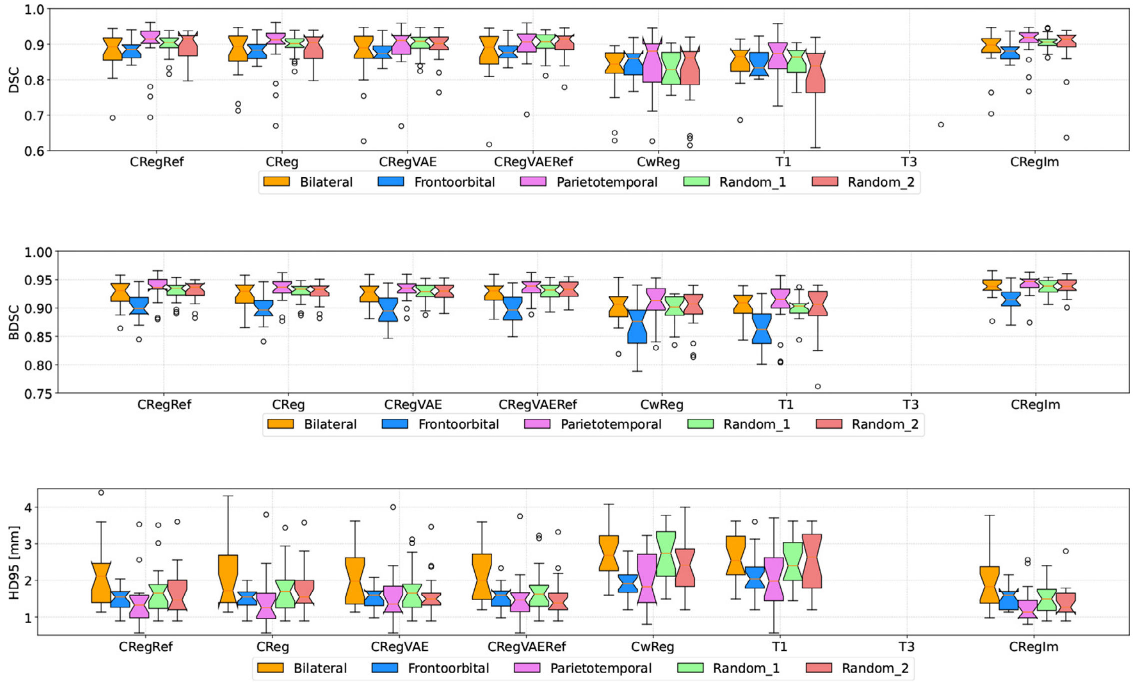


Fig. 9. Quantitative results for different defect types (Task 1). Note that the model trained using only the Task 3 dataset is completely unable to generalize into Task 1 dataset.

Table 3
Quantitative results for Task 2 (real cranial defects), including implant modeling. Please note that the evaluation metrics are calculated with respect to the expert-designed implant, not the cranial defect. The implant modeling step improves results for all cases except the one for which the defect reconstruction step failed (ID 6).

	Defect reconstruction (CRegIm)			Implant modeling		
Case	DSC	BDSC	HD95 [mm]	DSC	BDSC	HD95 [mm]
1	0.50	0.52	7.35	0.64	0.67	4.12
2	0.60	0.57	7.48	0.75	0.71	3.60
3	0.29	0.22	16.58	0.40	0.31	10.82
4	0.51	0.48	10.72	0.66	0.63	5.48
5	0.58	0.55	6.71	0.73	0.72	3.46
6	0.53	0.66	14.73	0.50	0.68	15.07
7	0.40	0.38	12.69	0.55	0.54	7.00
8	0.49	0.41	15.56	0.63	0.54	9.00
9	0.68	0.48	7.07	0.73	0.72	5.00
10	0.59	0.58	6.16	0.73	0.66	3.00
11	0.70	0.67	3.00	0.62	0.61	3.00
Mean	0.53	0.50	9.82	**0.63**	**0.61**	**6.32**
Std	0.12	0.13	4.48	0.11	0.12	3.86
Yu et al. [24]	0.52	0.45	8.34	–	–	–
Mahdi et al. [22]	0.38	0.33	51.24	–	–	–

observed that the implant is modeled correctly at the boundaries and requires just a minor manual modification to close the undesired hole.

3.4. Mixed reality

Fig. 12 shows a use case presenting the method outcome in mixed reality. The visualization is created using CarnaLife Holo software (MedApp S.A.). We show the implant hologram superimposed on the 3-D printed skull and both the defective skull and the calculated implant as holograms. This kind of visualization is created almost instantly in contrast to the 3-D printed model. It enables the operating team to analyze the proposed implant structure and to plan the procedure in 3-D. Thus, it is possible to propose adjustments based on the medical experience of the operating team before the implant is printed. The intuitive real-time planning should result in a reduction of possible errors. We also attach a supplementary movie presenting the procedure.[1]

3.5. Processing time

The average, minimum, and maximum times for the complete pipeline are 34.7, 28.9, and 44.6s respectively for the real cases from Task 2 dataset. The time includes the data input/output operations, initial preprocessing, inference using the defect reconstruction and defect refinement models, iterative implant thinning, and the preparation for 3-D printing. The most time-consuming part of the pipeline is the iterative implant thinning (∼60% of the total processing time), and the preparation for the 3-D printing (∼30% of the total processing time) while the input/output operations, pre-

[1] https://youtu.be/a1IMMtt3ovc

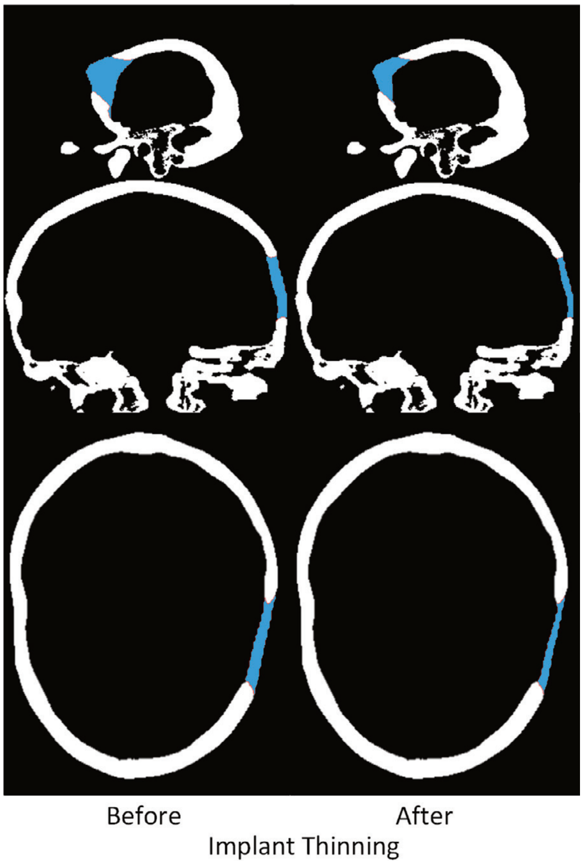


Fig. 10. Exemplary visualization of the implant modeling (Task 2 - Case 10). For the quantitative results we refer to Table 3. Note that before the implant thinning the reconstruction is not implantable.

processing and inference using deep models account for less than 10% of the total reconstruction time.

4. Discussion

4.1. General remarks

The key component to improving the reconstruction accuracy and generalizability is connected with combining the datasets from different distributions. The best performing setup achieves DC, BDC, and HD95 close to 0.91, 0.94, and 1.53 mm respectively, averaged across all test sets. These results are capable of automating the cranial implant modeling and are the best among currently proposed methods. The pipeline allows one to create a personalized implant directly from the segmentation of the defective skull.

The quantitative results vary between the three test sets. The reason for this is connected with the heterogeneity of the dataset. Task 3 is rather straightforward since the defects are regular, with similar sizes and shapes (except for the last 10 test cases). On the other hand, Task 1 introduces variability of the defect size and position. This, together with the fact that larger defects break the skull symmetry, leads to worse quantitative results. Interestingly, the variability between different defect types is minor and the method handles well all the defect variations. Only frontoorbital defects achieve considerably lower scores. It is related to breaking the skull symmetry. The quantitative scores for Task 2 are significantly different from Task 1 and Task 3. However, they are not related to the reconstructions, but the modeled implants. Since it is hard to define ground-truth implants, this evaluation requires the assessment of an expert neurosurgeon.

4.2. Ablation studies

The imperfect IR provides the most accurate results both from the quantitative and qualitative perspectives. The generalizability to real cranial defects is also the most promising. The augmentation by smooth and invertible IR or by the generative model improves the quantitative results on the Task 1/3 test sets, however, generalize worse into data from different distributions. It is because the smooth and invertible IR changes only the defect shapes while maintaining the original skull structure. On the other hand, imperfect registration creates skulls in between the original pairs, effectively creating a more heterogeneous training set. We hypothesized that similar results could be achieved by VAE by lowering the influence of the KL divergence, however, it turned out that it increased the variety of generated skulls but the defects were mostly incomplete and not anatomically plausible. The same issue arose when the VAE was trained on the training set created by imperfect IR. The naive combination of training sets by a simple resampling to the same voxel size and resolution does not improve the skull reconstruction significantly. It can be also observed that an attempt to use just a single training set results in a lack of generalizability. This is well-shown in experiments using only Task 3 training data, which provide good results for the Task 3 test set, however, are completely unable to reconstruct defects from the remaining datasets.

Several reasons are motivating the aforementioned ablation studies. First, we decided to not include ablation studies concerning the network architecture and training hyperparameters. We observed that straightforward modifications to the network architecture (e.g. adding additional skip connections, increasing the number of layers or filters) do not influence the results significantly. The different setups of training hyperparameters influence only the training time. Eventually, the models converge to similar states. On the other hand, the crucial aspect to improve the results is connected with preparing and augmenting the training set. It shows that the proposed method's performance is limited mostly by the training set size and its homogeneity.

4.3. Limitations

The proposed method has several limitations. It can be observed that the quantitative results for the implant modeling are not as good as for the defect reconstruction. This is caused by the difficulty with defining the ground-truth implant, in contrast with the ground-truth for the shape completion, which is straightforward. The optimal shape for a given implant depends on external geometric properties and the mechanical properties of the 3-D printing material. Moreover, there may be several implants that may be successfully implanted. Therefore, the implant evaluation should be performed by an expert neurosurgeon, as in the AutoImplant summary publication [21].

Another limitation is related to a significant difference between the DC and BDC for several cases. This happens when the reconstruction boundaries are well defined, however, the symmetry of the skull is lost. This is essential from an aesthetic point of view and could be addressed by further dataset extension or the use of statistical shape models [20].

In the current version, we assume that only a single implant is being modeled at a given time. However, this assumption is realized in postprocessing and may be easily relaxed to prepare several implants for a single skull with multiple defects.

4.4. Significance and practical applications

The automatic cranial implant design may improve several practical procedures. It may reduce the resources required for

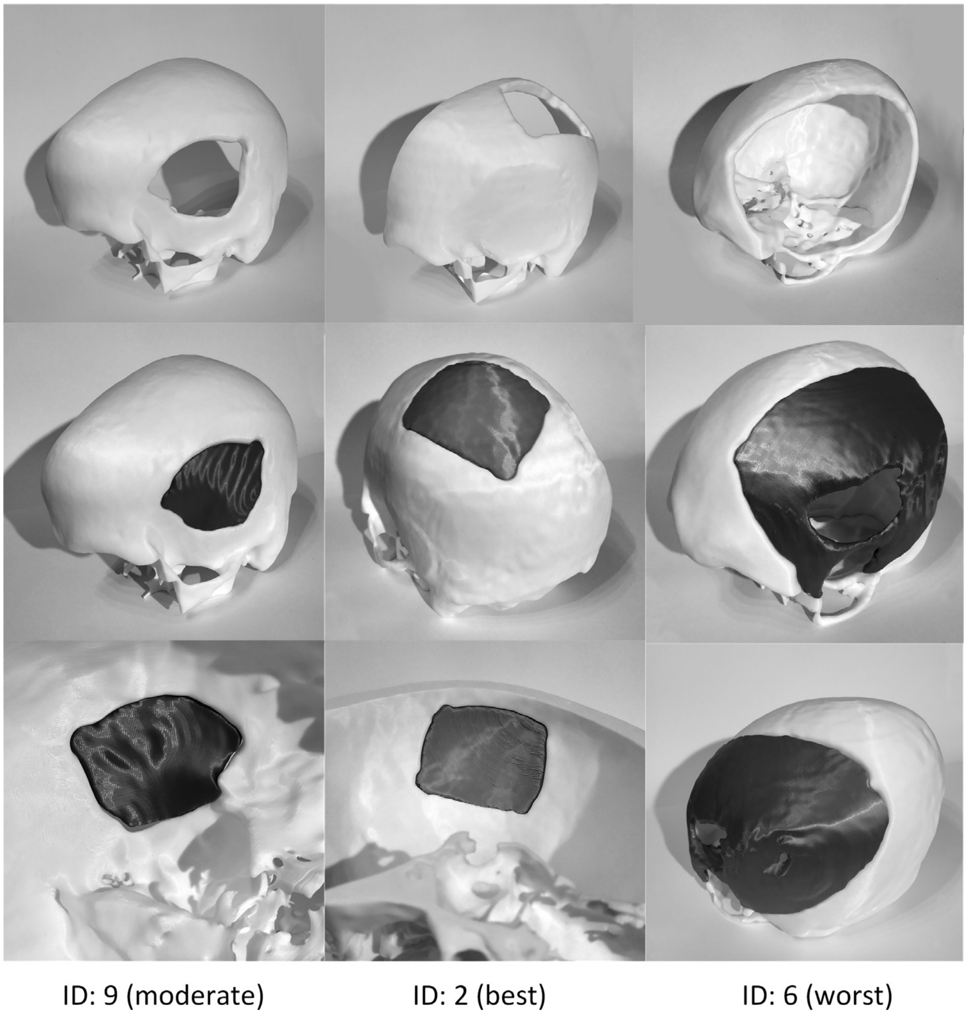


Fig. 11. Exemplary 3-D printed skulls together with the automatically designed implants for real, cranial defects (Task 2). Note that one case (ID: 6) is modeled incorrectly. However, the case strongly differs from all other real/synthetic defects in terms of size and shape.

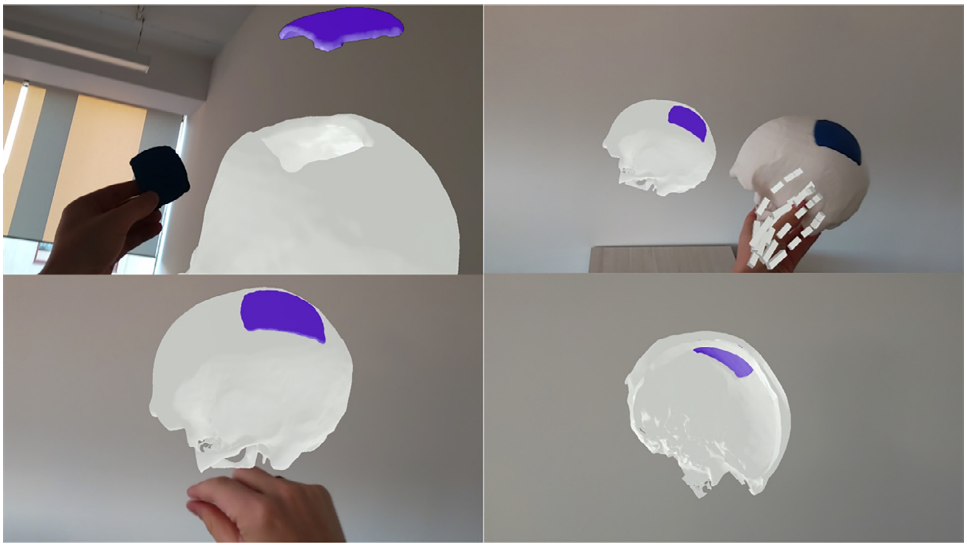


Fig. 12. Exemplary visualization of the reconstructed defect in mixed reality (real cranial defect, Task 2). We refer to a supplementary movie presenting the holographic interactions and system abilities.

manual modeling. Only small adaptations may be required, thus lowering the time required for implant manufacturing. As a result, the automatic design enables 3-D printing directly during the craniectomy, resulting in a potential unnecessity to perform the follow-up intervention to fill the defect. The surgery procedures may be further improved by the use of mixed reality systems. We attach a supplementary video presenting the hologram of the exemplary implant together with a corresponding 3-D printed defective skull. This technology may aid surgeons during the craniectomy and potentially reduce the surgery time [44]. Furthermore, we forecast that the presented pipeline and training strategy may be generalized into other areas, e.g. dental implants [45].

4.5. Future work

The most obvious way to proceed is to collect more clinical data samples and increase the training set variability. An example of such a dataset is MUG500+ [46]. We hope that in the next edition of the AutoImplant challenge, the organizers will preprocess the dataset by creating synthetic defects, and a unified representation of the ground-truth implants. Currently, the ground-truth implants are released in STL format. To make the research more reproducible, the ground-truth implants should be released in binary format, similarly to the defective skulls. However, as discussed in the Limitations section, it is hard to define which implant is the most suitable and it cannot be done at the level of the binary mask alone. Furthermore, the evaluation should be performed on cranial defects from different ethnic groups to ensure fairness and inclusivity.

The training set augmentation could be further explored. It would be interesting to compare several generative models to observe the influence on the results within the same and different distributions. In this research, we used VAEs due to the well-defined problem and their training stability. However, the generative adversarial networks (GANs) or GAN-VAEs could further increase the training set heterogeneity. Moreover, it would be beneficial to quantitatively describe the influence of the IR quality on defect reconstruction. However, this is computationally intensive since the IR-based augmentation requires thousands of 3-D registrations and increases quadratically with the number of skulls.

The use of deep reinforcement learning could be also interesting. This approach may limit the necessity of the shape completion step. The implant could be modeled directly using a set of pre-defined rules. These rules may be connected with both the geometric properties (e.g. minimum and maximum implant thickness, possibility to be implanted) and the mechanical properties of a given material. However, this approach requires significant computational resources during the training phase.

Another research area, strongly connected with the problem discussed, is finding the optimal scheme for the implants postprocessing to prepare them for 3-D printing. It would be beneficial to propose a pipeline to completely automate the printing procedure.

5. Conclusion

In this work, we proposed a complete pipeline to perform the automatic cranial defect reconstruction and implant modeling. We evaluated the proposed method on public datasets and obtained results significantly higher than the state-of-the-art methods. We released the source code to support the experiment's reproducibility. We performed several ablation studies, presented the method limitations, discussed practical use cases, and defined future research directions. The presented method is a significant contribution to the automatic design of cranial implants and may be potentially adapted to other implant types.

Declaration of Competing Interest

Authors declare that they have no conflict of interest.

CRediT authorship contribution statement

Marek Wodzinski: Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Project administration, Visualization, Writing - original draft, Writing - review & editing. **Mateusz Daniol:** Conceptualization, Formal analysis, Validation, Visualization, Writing - review & editing. **Miroslaw Socha:** Validation, Visualization. **Daria Hemmerling:** Supervision, Validation. **Maciej Stanuch:** Resources, Software. **Andrzej Skalski:** Supervision, Resources.

Acknowledgments

This research study was conducted retrospectively using human subject data made available in open access by AutoImplant 2021 challenge organizers [35]. The authors declare no conflict of interest.

We thank the AutoImplant organizers for evaluating our ablation studies on the closed test set, and the MedApp S.A for giving us access to the CarnaLife Holo software to create the visualizations in mixed reality. This research was supported in part by PLGrid Infrastructure.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.cmpb.2022.107173

References

[1] J. Li, et al., AutoImplant 2020-First MICCAI challenge on automatic cranial implant design, IEEE Trans. Med. Imaging 40 (9) (2021) 2329–2342.
[2] D. Bonda, S. Manjila, W. Selman, D. Dean, The recent revolution in the design and manufacture of cranial implants: modern advancements and future directions, Neurosurgery 77 (5) (2015) 814–824.
[3] F. Marreiros, et al., Custom implant design for large cranial defects, Int. J. Comput. Assist. Radiol. Surg. 11 (12) (2016) 2217–2230.
[4] W. Ameen, et al., Design, finite element analysis (FEA), and fabrication of custom titanium alloy cranial implant using electron beam melting additive manufacturing, Adv. Prod. Eng. Manag. 13 (3) (2018) 267–278.
[5] Y. Modi, S. Sanadhya, Design and additive manufacturing of patient-specific cranial and pelvic bone implants from computed tomography data, J. Braz. Soc. Mech. Sci. Eng. 40 (10) (2018).
[6] J. Li, et al., Automatic skull defect restoration and cranial implant generation for cranioplasty, Med. Image Anal. 73 (2021).
[7] A. Morais, J. Egger, V. Alves, Automated computer-aided design of cranial implants using a deep volumetric convolutional denoising autoencoder, Adv. Intell. Syst. Comput. 932 (2019) 151–160.
[8] P. Tack, J. Victor, P. Gemmel, L. Annemans, 3Dprinting Techniques in a medical setting: a systematic literature review, Biomed. Eng. Online 15 (115) (2016) 1–21.
[9] Q. Yan, et al., A review of 3D printing technology for medical applications, Engineering 4 (5) (2018) 729–742.
[10] N. DMA, Protolabs, 2022, (https://www.protolabs.co.uk/resources/case-studies/novax-dma/). [Online; accessed 05-Feb-2022].
[11] C. Gsaxner, U. Eck, D. Schmalstieg, N. Navab, J. Egger, Augmented reality in oral and maxillofacial surgery, Computer-Aided Oral Maxillofac. Surg. (2021) 107–139.
[12] J. Li, et al., A Baseline Approach for AutoImplant: The MICCAI 2020 Cranial Implant Design Challenge, in: Lecture Notes in Computer Science, vol. 12445, LNCS, 2020, pp. 75–84.
[13] H. Shi, X. Chen, Cranial Implant Design Through Multiaxial Slice Inpainting Using Deep Learning, in: Lecture Notes in Computer Science, vol. 12439, LNCS, 2020, pp. 28–36.
[14] F. Matzkin, V. Newcombe, B. Glocker, E. Ferrante, Cranial Implant Design via Virtual Craniectomy with Shape Priors, in: Lecture Notes in Computer Science, vol. 12439, LNCS, 2020, pp. 37–46.
[15] D.G. Ellis, M.R. Aizenberg, Deep learning using augmentation via registration: 1st place solution to the autoimplant 2020 challenge, in: Cranial Implant Design Challenge, 2020, pp. 47–55.
[16] O. Kodym, M. Španěl, A. Herout, Cranial Defect Reconstruction Using Cascaded CNN with Alignment, in: Lecture Notes in Computer Science, vol. 12439, LNCS, 2020, pp. 56–64.

[17] J. Mainprize, Z. Fishman, M. Hardisty, Shape Completion by U-Net: An Approach to the AutoImplant MICCAI Cranial Implant Design Challenge, in: Lecture Notes in Computer Science, vol. 12439, LNCS, 2020, pp. 65–76.
[18] B. Wang, et al., Cranial Implant Design Using a Deep Learning Method with Anatomical Regularization, in: Lecture Notes in Computer Science, vol. 12439, LNCS, 2020, pp. 85–93.
[19] Y. Jin, J. Li, J. Egger, High-Resolution Cranial Implant Prediction via Patch–Wise Training, in: Lecture Notes in Computer Science, vol. 12439, LNCS, 2020, pp. 94–103.
[20] P. Pimentel, et al., Automated Virtual Reconstruction of Large Skull Defects using Statistical Shape Models and Generative Adversarial Networks, in: Lecture Notes in Computer Science, vol. 12439, LNCS, 2020, pp. 16–27.
[21] J. Li, et al., Towards clinical applicability and computation efficiency in automatic cranial implant design: an overview of the AutoImplant 2021 cranial implant design challenge, (2022). In Submission.
[22] H. Mahdi, et al., A U-Net Based System for Cranial Implant Design with Pre-processing and Learned Implant Filtering, in: Lecture Notes in Computer Science, vol. 13123, LNCS, 2021, pp. 63–79.
[23] B. Yang, K. Fang, X. Li, Cranial Implant Prediction by Learning an Ensemble of Slice-Based Skull Completion Networks, in: Lecture Notes in Computer Science, 13123, LNCS, 2021, pp. 95–104.
[24] L. Yu, J. Li, J. Egger, PCA-Skull: 3D Skull Shape Modelling Using Principal Component Analysis, in: Lecture Notes in Computer Science, vol. 13123, LNCS, 2021, pp. 105–115.
[25] S. Pathak, et al., Cranial Implant Design Using V-Net Based Region of Interest Reconstruction, in: Lecture Notes in Computer Science, vol. 13123, LNCS, 2021, pp. 116–128.
[26] M. Wodzinski, M. Daniol, D. Hemmerling, Improving the Automatic Cranial Implant Design in Cranioplasty by Linking Different Datasets, in: Lecture Notes in Computer Science, vol. 13123, LNCS, 2021, pp. 29–44.
[27] J. Li, et al., Learning to Rearrange Voxels in Binary Segmentation Masks for Smooth Manifold Triangulation, in: Lecture Notes in Computer Science, vol. 13123, LNCS, 2021, pp. 45–62.
[28] A. Kroviakov, J. Li, J. Egger, Sparse Convolutional Neural Network for Skull Reconstruction, in: Lecture Notes in Computer Science, vol. 13123, LNCS, 2021, pp. 80–94.
[29] C. Choy, J. Gwak, S. Savarese, 4D spatio-temporal convnets: minkowski convolutional neural networks, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2019-June, 2019, pp. 3070–3079.
[30] A. Memon, et al., A review on patient-specific facial and cranial implant design using artificial intelligence (AI) techniques, Expert Rev. Med. Devices 18 (10) (2021) 985–994.
[31] O. Kodym, M. Španěl, A. Herout, Segmentation of Defective Skulls from CT Data for Tissue Modelling, in: Lecture Notes in Computer Science, vol. 13123, LNCS, 2021, pp. 19–28.
[32] D. Ellis, C. Alvarez, M. Aizenberg, Qualitative Criteria for Feasible Cranial Implant Designs, in: Lecture Notes in Computer Science, vol. 13123, LNCS, 2021, pp. 8–18.
[33] L. Rauschenbach, C. Rieß, U. Sure, K. Wrede, Personalized Calvarial Reconstruction in Neurosurgery, in: Lecture Notes in Computer Science, vol. 13123, LNCS, 2021, pp. 1–7.
[34] M. Wodzinski, Source Code, 2021https://github.com/MWod/AutoImplant_2021.
[35] O. Kodym, et al., SkullBreak / SkullFix - dataset for automatic cranial implant design and a benchmark for volumetric shape learning tasks, Data Brief 35 (2021).
[36] CQ500 Dataset, 2021, (http://headctstudy.qure.ai/dataset).
[37] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, in: MICCAI 2015, 2015, pp. 234–241.
[38] M. Wodzinski, et al., Semi-supervised deep learning-based image registration method with volume penalty for real-time breast tumor bed localization, Sensors 21 (12) (2021) 1–14.
[39] G. Balakrishnan, A. Zhao, M. Sabuncu, J. Guttag, A. Dalca, VoxelMorph: a Learning framework for deformable medical image registration, IEEE Trans. Med. Imaging 38 (8) (2019) 1788–1800.
[40] O. Kodym, M. Španěl, A. Herout, Deep learning for cranioplasty in clinical practice: going from synthetic to real patient data, Comput. Biol. Med. 137 (2021).
[41] W. Schroeder, R. Maynard, B. Geveci, Flying edges: a high-performance scalable isocontouring algorithm, in: 2015 IEEE 5th Symposium on Large Data Analysis and Visualization (LDAV), 2015, pp. 1–8.
[42] A. Paszke, et al., PyTorch: an imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8024–8035.
[43] W. Falcon, K. Cho, A Framework For Contrastive Self-Supervised Learning And Designing A New Approach, arXiv preprint arXiv:2009.00104(2020).
[44] R. Wierzbicki, et al., 3D mixed-reality visualization of medical imaging data as a supporting tool for innovative, minimally invasive surgery for gastrointestinal tumors and systemic treatment as a new path in personalized treatment of advanced cancer diseases, J. Cancer Res. Clin. Oncol. 148 (1) (2022) 237–243.
[45] S. Bayrakdar, et al., A deep learning approach for dental implant planning in cone-beam computed tomography images, BMC Med. Imaging 21 (1) (2021).
[46] J. Li, et al., MUG500+: database of 500 high-resolution healthy human skulls and 29 craniotomy skulls and implants, Data Brief 39 (2021).