# Towards the Automatization of Cranial Implant Design in Cranioplasty: Structured description of the challenge design

Remark: This challenge have been slightly modified. All changes are highlighted in red.

## CHALLENGE ORGANIZATION

### Title

Use the title to convey the essential information on the challenge mission.

Towards the Automatization of Cranial Implant Design in Cranioplasty

### Challenge acronym

Preferable, provide a short acronym of the challenge (if any).

AutoImplant

### Challenge abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Cranioplasty is the surgical process where a skull defect, caused in a brain tumor surgery or by trauma, is repaired using a cranial implant, which must fit precisely against the borders of the skull defect as an alternative to the removed cranial bone. The designing of the cranial implant is a challenging task and involves several steps: (1) obtaining the 3D imaging data of the skull with defect from CT or MRI, (2) converting the 3D imaging data into 3D mesh model and (3) creating the 3D model of the implant for 3D printing. The last step usually requires expensive commercial software, which clinical institutions often have limited access to. Researchers have been working on CAD software as alternative to the commercial software for the designing of cranial implant whereas these approaches still involve human interaction, which is time-consuming and requires expertise of the specific medical domain. Therefore, a fast and automatic design of cranial implants is highly desired, which also enables in Operation Room (in OR) manufacturing of the implants for the patient. Centered around the topic, our challenge provides 200 healthy skulls acquired from CT scans in clinical routine and seeks data-driven approaches for the problem. We inject artificial defects into each healthy skull to create training pairs. The datasets are split into a training set and a testing set, each containing 100 healthy skulls and their corresponding skulls with artificial defects. Participants are expected to design algorithms (such as deep learning) based on these training pairs for an automatic cranial defect restoration and implant generation. In this sense, the problem is being formulated as a 3D volumetric shape completion task where a defected skull volume is automatically completed by the algorithm from the participants. The restored defect, which is in fact the implant we want, can be obtained by the subtraction of the defected skull from the completed skull. The implants reconstructed from the skulls with the artificial defects will be quantitatively evaluated using the Dice Similarity Score (DSC) and the Hausdorff Distance (HD).

### Challenge keywords

List the primary keywords that characterize the challenge.

Reconstruction, Cranioplasty, Cranial implant, Deep learning

### Year

The challenge will take place in ...

2020

## FURTHER INFORMATION FOR MICCAI ORGANIZERS

### Workshop

If the challenge is part of a workshop, please indicate the workshop.

There are no associated workshops for our challenge.

### Duration

How long does the challenge take?

Half day.

### Expected number of participants

Please explain the basis of your estimate (e.g. numbers from previous challenges) and/or provide a list of potential participants and indicate if they have already confirmed their willingness to contribute.

The expected number of participants is 15. We formulate the clinically relevant task to a common technical problem (3D shape completion) so that people from technical side can participant without needing in-depth medical knowledge of this field (cranioplasty). We expect that people can be attracted by our potentially interesting challenge that aims at solving a critical medical problem.

### Publication and future plans

Please indicate if you plan to coordinate a publication of the challenge results.

We plan to coordinate a publication of the challenge results and want to refine the challenge based on the results and the feedback of the participants for a subsequent challenge.

### Space and hardware requirements

Organizers of on-site challenges must provide a fair computing environment for all participants. For instance, algorithms should run on the same computing platform provided to all.

Our space/ hardware requirements are a standard computer with Nvidia 1080 Ti or higher GPU, a table and a monitor.

# TASK: Perfoming a Cranial Implant Reconstruction for a Skull Defect.

## SUMMARY

### Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

The goal of this task is to reconstruct a skull defect. Hence, the skull to reconstruct will have some kind of a hole, which was caused by an accident or a bone tumor. The cranial implant is in general the difference between the defected skull (with the hole) and the reconstructed (healthy) skull. There exist several semi-automatic software tools from the industry. However, this challenge focus on an automatic skull reconstruction and implant design. In doing so, these automatic methods can be used in the future for an fast in operation room (in OR) application for the 3D printing of implants. Semi-automatic approaches, which take at least 15-30 minutes are too slow for in OR 3D implant printing, as the patient should not be under anesthesia for such a long time during the cranioplasty.

### Keywords

List the primary keywords that characterize the task.

Skull reconstruction, Cranioplasty, Cranial implant, Deep learning.

## ORGANIZATION

### Organizers

a) Provide information on the organizing team (names and affiliations).

The information of the organizing team (names and affiliations) are as follows:

Priv.-Doz. Dr. Dr. Jan Egger
Graz University of Technology
Institute of Computer Graphics and Vision

Xiaojun Chen, Professor
Shanghai Jiao Tong University
School of Mechanical Engineering

Mr. Jianning Li
Graz University of Technology
Institute of Computer Graphics and Vision

Univ.-Prof. Dr. Ute Schäfer
Medical University of Graz
Department of Neurosurgery

Univ.-Ass. Priv-Doz. Dr. med. Gord of Campe
Medical University of Graz

Department of Neurosurgery

Mr. Marcell Krall
Medical University of Graz
Department of Neurosurgery

Ms. Ulrike Zefferer
Medical University of Graz
Department of Neurosurgery

Ms. Christina Gsaxner
Graz University of Technology
Institute of Computer Graphics and Vision

Mr. Antonio Pepe
Stanford University School of Medicine
Department of Radiology

Professor Dr. Dieter Schmalstieg
Graz University of Technology
Institute of Computer Graphics and Vision

b) Provide information on the primary contact person.

Jianning Li (jianning.li@icg.tugraz.at)
Jan Egger (egger@tugraz.at)

## Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.

Examples:

- One-time event with fixed submission deadline
- Open call
- Repeated event with annual fixed submission deadline

Repeated event open call.

## Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

MICCAI.

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

grand-challenge.org

c) Provide the URL for the challenge website (if any).

https://autoimplant.grand-challenge.org

## Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

Fully automatic.

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

Publicly available data is allowed.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

May participate but not eligible for awards and not listed in leaderboard.

d) Define the award policy. In particular, provide details with respect to challenge prizes.

The top 3 performing methods will be awarded with a certificate.

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

The top 3 performing methods will be announced publicly.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

At most two of the participating teams' members qualifies as authors and the participating teams may publish their own results separately. There is no embargo time, results can be published also before the challenge paper of the organizers.

## Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

Participants can submit their results via the grand-challenge platform.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

A pre-evaluation is possible until September 2020, please see Schedule (following section 9). We will report the Dice Similarity Coefficient (DSC) and the Hausdorff Distance (HD) back to the participants.

## Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

The release date of the training cases is scheduled for April/May 2020.
The registration period is scheduled for January 2020 - August 2020.
The release date of the test cases and validation cases is scheduled for September 2020.
The Submission date is scheduled for September 2020.

## Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

An already public dataset is used.

## Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

CC BY.

## Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

The evaluation code is available on the following GitHub site:
https://github.com/li-jianning/evaluation_metrics

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

The participants can choose to open source their code on GitHub for example. However, a code release is not a requirement for participants.

### Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

There is no conflicts of interest. The sponsoring comes from the organizers of the challenge. All data will be published online and is freely accessible by everyone who is interested.

## MISSION OF THE CHALLENGE

### Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

Surgery, CAD.

### Task category(ies)

State the task category(ies).

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Restoration.

## Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

The target cohort are patients undergoing brain tumor surgery when craniotomy is performed. After the surgery, the patients require a cranial implant for the restoration of the cranial defect (cranioplasty).

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

The challenge datasets are acquired from the patients using CT scans of the heads from the clinical routine. The datasets can be used to create an 3D skull atlas or create training pairs for cranial defect restoration as in our challenge.

## Imaging modality(ies)

Specify the imaging technique(s) applied in the challenge.

The imaging technique applied in the challenge are Computed Tomography (CT) scans of the head from the clinical routine.

## Context information

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

From the images only the skull data segmented from head CTs will be provided. The raw head CTs are not provided.

b) ... to the patient in general (e.g. sex, medical history).

Males and females each account for approximately half of the datasets.

## Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

The data origin are CT scans of the head from the clinical routine.

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

The algorithm target are skulls segmented from CT scans of the head.

### Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Integration in workflow, Feasibility, Usability, Applicability, Robustness, Accuracy.

## DATA SETS

### Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

The device used to acquire the challenge data was CT machines.

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

We will provide 200 high resolution healthy skull datasets (binary 3D volume) segmented from CT scans of the head. Each skull (bone) is segmented using thresholding (between 100 and 200). For each healthy skull, we inject artificial defects on the skull surface. The artificial defects resembles the defects manually injected by neurosurgeons in craniotomy but are simplified. Both, the original 200 healthy skulls and the corresponding skulls with artificial defects will be provided, which can be used for developing algorithms (training) and evaluation. We created on our GitHub page an illustration of healthy skull data, the skull with an artificial defect and how the implant generated from participants' algorithm should fit the defect on the skull: https://github.com/li-jianning/skulldataprocessing/wiki/data_formulation

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

All datasets were acquired from the QC 500 public dataset.

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

The CT scans of the head are selected from the public QC 500 dataset. The skulls are segmented from the selected CT scans using thresholding.

### Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

One case consists of two skull datasets (the healthy skull and the corresponding skull with an artificial defect) in NRRD format with a resolution of 512×512×Z, with Z corresponding to the number of axial slices in every singe dataset.

b) State the total number of training, validation and test cases.

In total, 200 skull datasets will be provided for the challenge. From this 200 datasets, 100 datasets are for training and validation, and the remaining 100 datasets are for testing.

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

Each dataset comes from a unique skull of a patient that has been CT scanned during the clinical routine.

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

Further important characteristics of the cases are not applicable for the challenge.

### Annotation characteristics

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

There are no human annotators involved.

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

There are no human annotators involved.

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

There are no human annotators involved.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

There are no human annotators involved.

### Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

A skull segmentation from the raw CT scans is performed before the datasets are provided to the participants. The segmentation is done through thresholding.

### Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

There are no human annotators involved.

b) In an analogous manner, describe and quantify other relevant sources of error.

There are no human annotators involved.

## ASSESSMENT METHODS

### Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

The metrics to assess a property of an algorithm are the Dice Similarity Coefficient (DSC) and the Hausdorff Distance (HD). The evaluation and the calculation of the metrics is taking place in actual 3D volumes. In addition, our clinical partner will check and and judge the implants with the highest DSC and HD score for the winning teams, and we will provide a written statement about the clinical plausibility from our clinical partner.

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

The Dice Similarity Coefficient (DSC) measures the overlap between two shapes. The similarity between the ground truth implant and the automatic results from participants' algorithm can be effectively measured using this metric. Hausdorff Distance (HD) measures the difference between the ground truth implant and the automatic results.
We have a previously published paper related to cranial implant design that uses DSC and HD as evaluation metrics:
Egger, J., Gall, M., Tax, A., Ücal, M., Zefferer, U., Li, X., Campe, G.V., Schäfer, U., Schmalstieg, D., & Chen, X. (2017). Interactive reconstructions of cranial 3D implants under MeVisLab as an alternative to commercial planning software. PloS one.

### Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.

For each test case, we will calculate the DSC and the HD between the ground truth and the results from the participants. The code for the metric calculation for each case can be found on our GitHub page: https://github.com/li-jianning/evaluation_metrics
We take the mean of the DSCs and the mean of the HDs over the test cases. The mean DSC and mean HD will be ranked separately among the teams. DSC is ranked in descending order and HD is ranked in ascending order. A final overall rank is given by taking the average of the two ranks.

b) Describe the method(s) used to manage submissions with missing results on test cases.

We will exclude the participants who fail to report on the whole testing set.

c) Justify why the described ranking scheme(s) was/were used.

We use two metrics (DSC and HD) for the evaluation of participants' algorithms. DSC and HD capture from different aspects important characteristics of how the implant from the participants and the ground truth match. The ranking for each metric (DSC and HD) will be given separately. However, we will also give a final overall ranking by taking the average of the two ranks from DSC and HD like the MICCAI 2019 challenge: StructSeg 2019 https://structseg2019.grand-challenge.org/Evaluation/

### Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

We will exclude the participants who fail to report on the whole testing set. Besides the statistical values such as mean, standard deviation of the DSCs and HDs, we use the p-value in t-test to assess whether the top performing/ranking algorithms are significantly better than the rest of algorithms. t test code is available at https://github.com/li-jianning/ttest.
To measure the variability, we will also consider variance, squared deviation, average absolute deviation and the inter-quartile range.

b) Justify why the described statistical method(s) was/were used.

The mean value of DSC and HD produced by the algorithms are indicators of their overall performance. The standard deviation measures the performance stability of the algorithms.

### Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

These further analyses will be discussed in a further publication after the challenge.

## ADDITIONAL POINTS

## References

Please include any reference important for the challenge design, for example publications on the data, the annotation process or the chosen metrics as well as DOIs referring to data or code.

Related references and publications that are important for the challenge design are:

[1] Markus Gall, Xing Li, Xiaojun Chen, Dieter Schmalstieg, and Jan Egger. Computer-aided planning and reconstruction of cranial 3D implants. 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 1179–1183, 2016.

[2] Xiaojun Chen, Lu Xu, Xing Li, and Jan Egger. Computer-aided implant design for the restoration of cranial defects. Scientific Reports, 2017.

[3] Antonio Marzola, Lapo Governi, L. Genitori, Federico Mussa, Yary Volpe, and Rocco Furferi. A semi-automatic hybrid approach for defective skulls reconstruction. Computer-Aided Design and Applications, 17:190–204, 2019.

[4] Jan Egger, Markus Gall, Alois Tax, Muammer Ucal, Ulrike Zefferer, Xing Li, Gord von Campe, Ute Schaefer, Dieter Schmalstieg, and Xiaojun Chen. Interactive reconstructions of cranial 3D implants under MeVisLab as an alternative to commercial planning software. PLoS ONE, 12:20, 2017.

[5] Ana Morais, Jan Egger, and Victor Alves. Automated Computer-aided Design of Cranial Implants using a Deep Volumetric Convolutional Denoising Autoencoder. World CIST, pages 151–160. 2019.

[6] Yuan-Lin Liao, Chia-Feng Lu, Yung-Nien Sun, Chieh-Tsai Wu, JiannDer Lee, Shih-Tseng Lee, and Yu-Te Wu. Three-dimensional reconstruction of cranial defect using active contour model and image registration. Medical Biological Engineering Computing, 49:203–211, 2010.

[7] Marc Anton Fuessinger, Steffen Schwarz, Carl-Peter Cornelius, Marc Christian Metzger, Edward Ellis, Florian Probst, Wiebke Semper-Hogg, Mathieu Gass, and Stefan Schlager. Planning of skull reconstruction based on a statistical shape model combined with geometric morphometrics. International Journal of Computer Assisted Radiology and Surgery, 13:519–529, 2017.

[8] Francesco Buonamici, Rocco Furferi, L. Genitori, Lapo Governi, Antonio Marzola, Federico Mussa, and Yary Volpe. Reverse engineering techniques for virtual reconstruction of defective skulls: an overview of existing approaches. Computer-Aided Design and Applications, 16:103–112, 2018.

Code:
We provide python scripts to segment human skull bones from CT scan, clean the segmented skull, convert the skull volume to mesh and inject artificial defects to the healthy skull on Github:
https://github.com/jianning-li/Skull-Data-Processing

## Further comments

Further comments from the organizers.

We have currently a Scientific Data (https://www.nature.com/sdata/) article in submission, which has a more detailed description of the skull dataset we provide. The article is expected to be published well before the

challenge.