Deep Learning in Automatic Cranial Implant Design: Insights from two years' MICCAI Grand Challenges

Dipi.Ing. Jianning Li, B.Eng.
Graz University of Technology &
IKIM, University Hospital Essen

University Medicine Essen
Institute for Artificial Intelligence in Medicine

Background

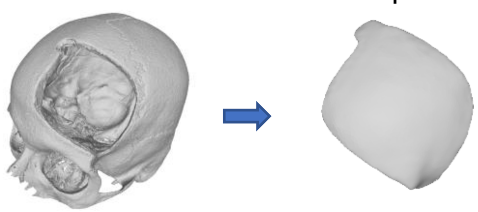
(Automatic) cranial implant design

using a patient-specific cranial implant (yellow) to repair a defective skull (gray)

problem formulation

shape completion subtraction

Defective skull Completed skull Cranial implant



Defective skull Cranial implant

Learning-based 3D shape completion/inpainting

MICCAI Challenges

- AutoImplant I (MICCAI, 2020, virtual)
- AutoImplant II (MICCAI, 2021, virtual)

Workload:

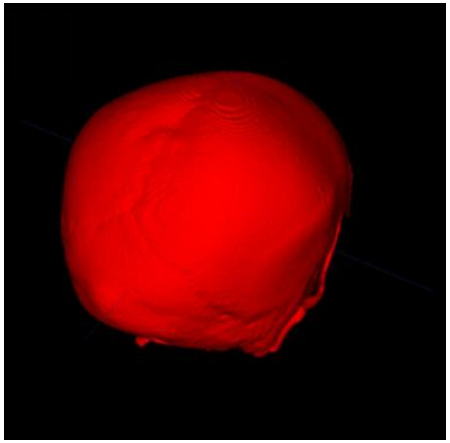
- **write a challenge proposal (peer-reviewed, rebuttal, revision, accept)**
- **set up challenge websites[1]**
- **prepare the datasets**
- **dissemination (call for participation, call for papers)**
- **process the submissions (calculate scores, ranking, paper review)**
- **set up challenge programs (organize presentation)**
- **post-challenge proceedings (Springer LNCS)**

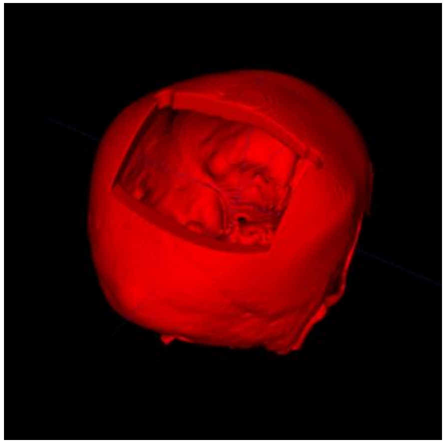
1 https://autoimplant2021.grand-challenge.org/ , https://autoimplant.grand-challenge.org/

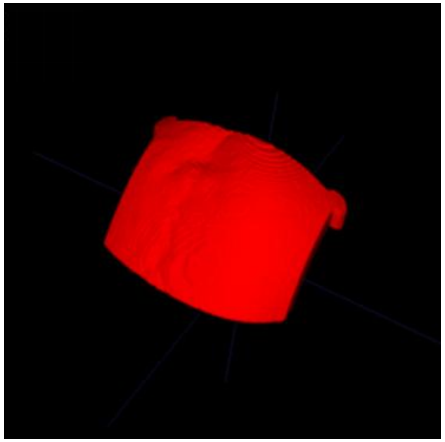
Background: *AutoImplant I,II* differences

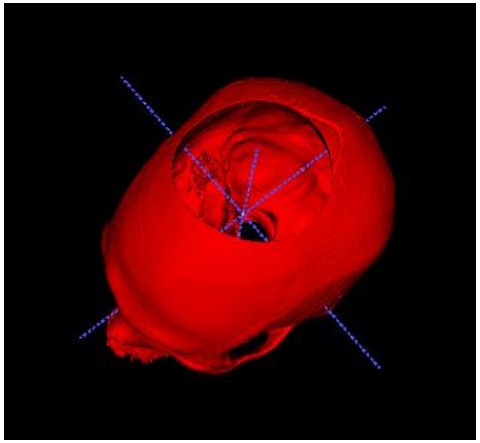
1. the datasets

AutoImplant I used synthetic defects (100 for training, 110 for evaluation)

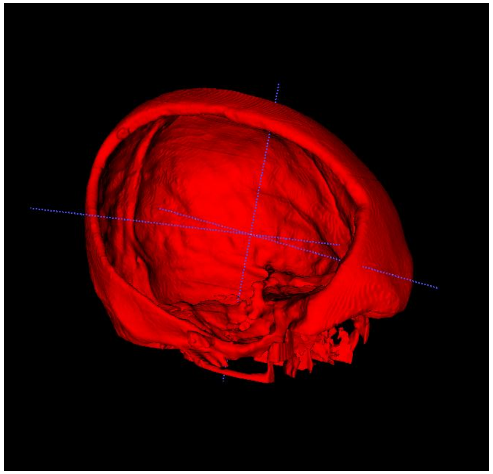


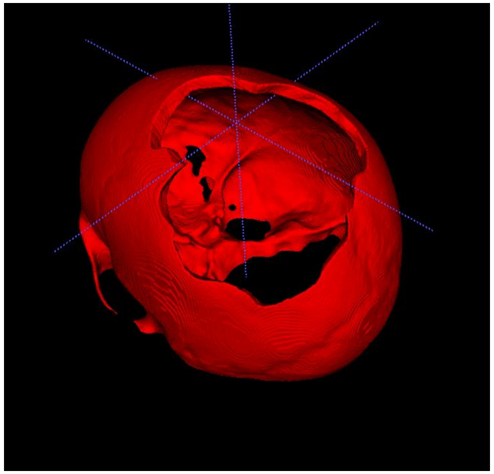






AutoImplant II provided both clinical defects and (more complex) synthetic defects





clinical (11, Task 2)

synthetic (114x5=570 for training, 20x5=100 for evaluation, Task 1)

2. evaluation and ranking

AutoImplant I used common quantitative metrics: DSC, HD

AutoImplant II used customized metrics: DSC, hd95, border DSC, quantified criteria from neurosurgeons → **ranking reflects the submissions' actual clinical usability**

Background: Network Architectures

''Details in method configuration have more impact on performance than do architectural variations''

nnU-Net. Isensee et al. (2021). nature methods

architectural variations (Encoder-Decoder 2D/3D):

AutoImplant I

- **ED+Squeeze-and-Excitation block (cvpr 2018)**
- **U-Net**
- **ED+ Residual blocks (ResNet, cvpr 2016)**
- **U-Net + Residual blocks (1st place submission)**
- **V-Net**
- **Residual Dense U-Net (DenseNet, cvpr 2017)**
- **Mesh-based statistical shape model (SSM, non learning-based method)**

AutoImplant II

- **ED**
- **U-Net**
- **U-Net+ Residual block (1st place submission)**
- **LSTM (2D)**
- **PCA (non learning-based method)**

Input: defective skull. Output: complete skull or implant

Pre-processing/ data augmentation/ method configurations contribute the most to the ranking variations of the submissions

observations from AutoImplant I and II.

preprocessing

- **background cropping**
- **skull registration & alignment**
- **skull cropping**

data augmentation

- **dataset linking**
- **augment the defects**
- **augment the skulls & defect via shape warping**

method configuration

- **coarse-to-fine framework**
- **shape prior**
- **regularization during training**

Background: Network Architectures

Pre-processing/ data augmentation/ method configurations contribute the most to the ranking variations of the submissions

observations from AutoImplant I and II.

preprocessing

- **background cropping**
- **skull registration & alignment**
- **skull cropping**

data augmentation

- **dataset linking**
- **augment the defects**
- **augment the skulls & defect via shape warping**

method configuration

- **coarse-to-fine framework**
- **shape prior**
- **regularization during training**

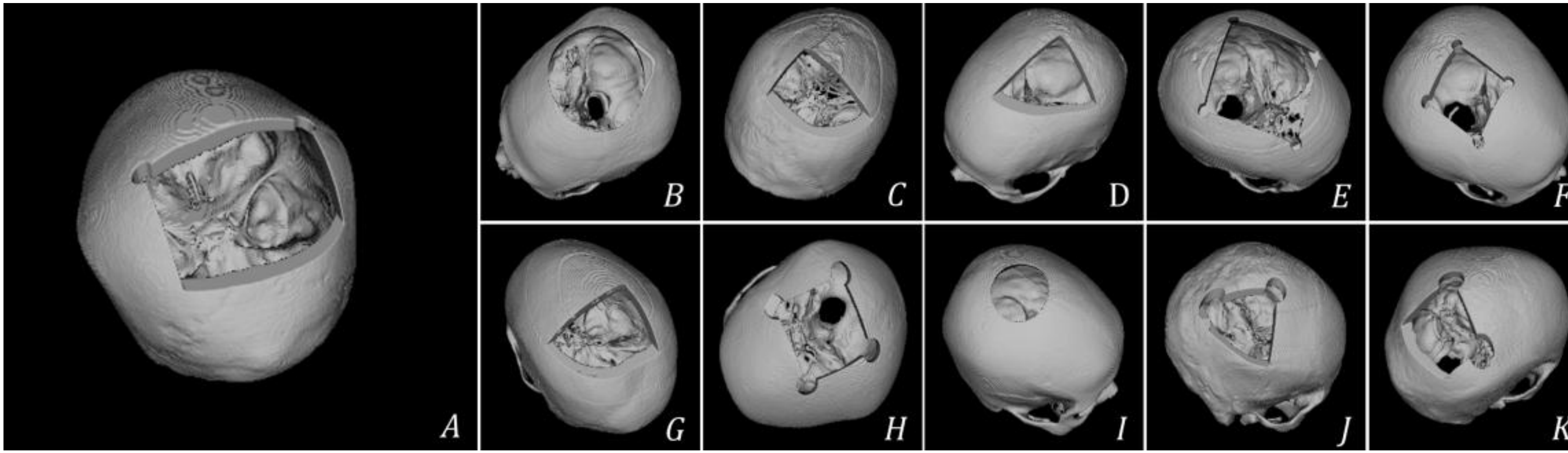
Non-trivial technical challenges:

- <u>Generalization/Domain shift</u>
  - generalize to various defect shapes
  - generalize to various skull shapes
  - generalize to clinical cases
- <u>High memory footprint</u>
  - skull images are large (512*512*Z)
  - desktop GPU memory is limited
  - training is slow
- <u>Clinical feasibility</u>
  - transfer models trained on synthetic data to clinical data
  - traditional quantitative metrics are not closely correlated to the submissions' actual usability
  - subjective quality measures are not standardized and quantified.

Generalization/Domain shift

problems

- There are 10 test cases with varied defect distributions (B-K) compared to the training defects (A)
- A network tends to overfit to the training defects and cannot generalize well to the 10 extra defects in the test set
- The clinical defects tend to be much more irregular and complex and therefore more difficult to complete



AutoImplant I: 5 out of 11 submissions failed on the 10 out-of-distribution cases.
AutoImplant II: only 3 teams attempted Task 2 (11 clinical cases).

solutions

- data augmentation & dataset linking
- preprocessing: skull registration & alignment
- using shape priors or regularization during training

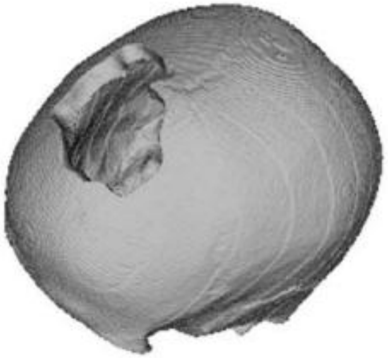
Those failed on the out-of-distribution cases did not use any of the methods above!

Generalization /Domain shift: data augmentation & dataset linking

AutoImplant I (2020)

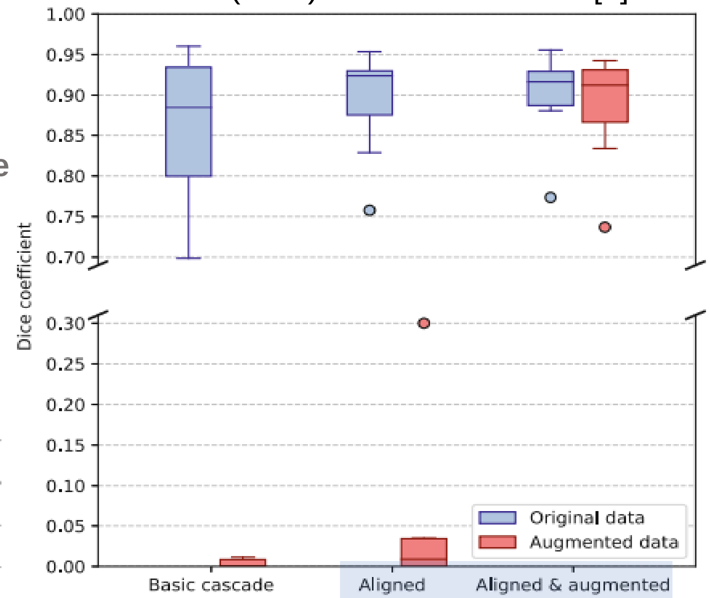
data augmentation:

- **to increase the varieties of the training samples to prevent overfitting (Kodym, O et al [1])**
- **to create training samples with similar distributions to the test samples (Matzkin, F et al [2])**



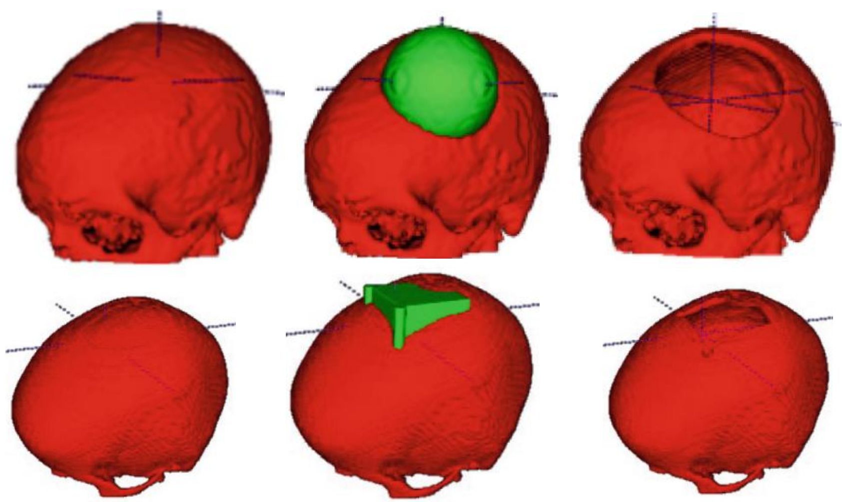
[1] Kodym, O et al: created 5 random defects for each complete skull in the training set

Results (DSC) on a validation set [1]



Results on the test set [1]: no major decline on the 10 out-of-distribution test samples

	Test case (100)	Test case (10)	Overall (110)
Mean DSC	0.920	0.910	0.919
Mean HD	4.137	4.707	4.189



Participants have access to all test samples.
We did not use a <u>hidden test set</u>.

[2] Matzkin, F et al: create defects that are similar to the out-of-distribution test defects, for training

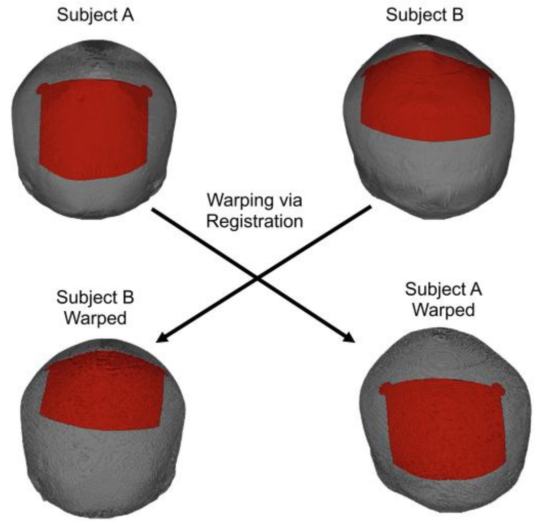
[1] Kodym, O., Španěl, M. and Herout, A., 2020, October. Cranial defect reconstruction using cascaded CNN with alignment. In Cranial Implant Design Challenge (pp. 56-64). Springer, Cham.

[2] Matzkin, F., Newcombe, V., et al., 2020, October. Cranial implant design via virtual craniectomy with shape priors. In Cranial Implant Design Challenge (pp. 37-46). Springer, Cham.

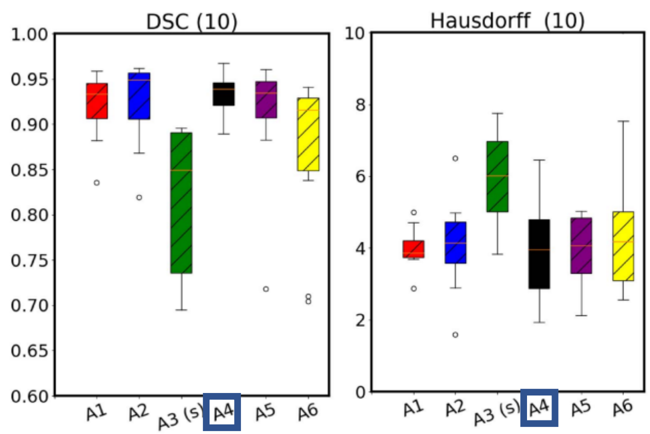
Generalization /Domain shift: data augmentation & dataset linking

- **intensive augmentation: to warp each training sample to the space of the rest samples (Ellis, D.G. et al[1])**
- **dataset linking: to combine datasets of different sources/distributions for training (Wodzinski, M et al [2])**

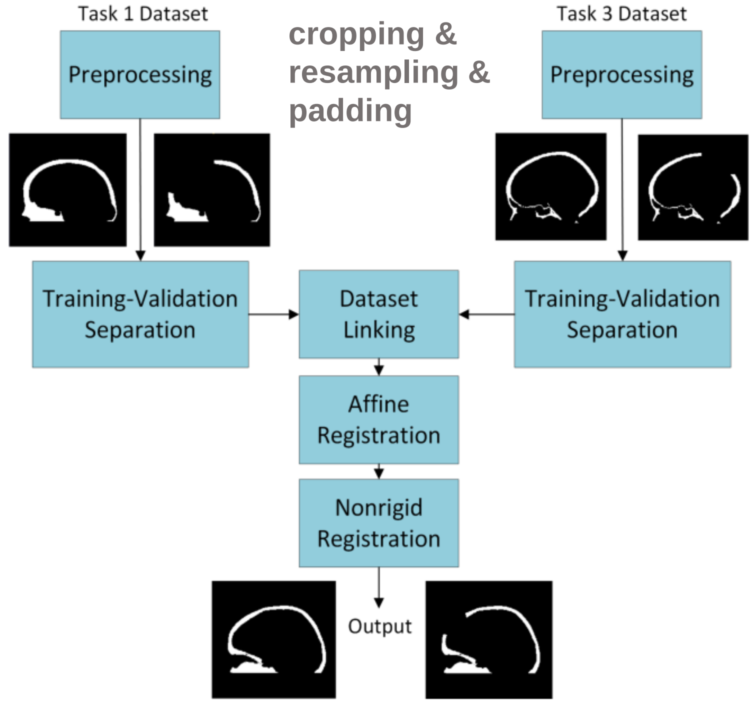
AutoImplant I (2020)



- **Ellis, D.G. et al [1]: 100 training samples augmented to 99*100+100=10000 samples (pair-wise registration & warping)**
- **Ranked 1st place in AutoImplant I**



AutoImplant II (2021)



- **Wodzinski, M et al [2]:AutoImplant II**
- **‘ all vs all ’ registration & warping as in [1] (different registration methods)**
- **Combine the dataset of Task 1 and 3**
- **Train a single model for all the 3 tasks**
- **Merge datasets by cropping, resampling, padding**
- **Train only on synthetic samples but work reasonably well on clinical test cases (Task2)**
- **Ranked 1st place in AutoImplant II (all 3 tasks)**

---

[1] Ellis, D.G. and Aizenberg, M.R., 2020, October. Deep learning using augmentation via registration: 1st place solution to the AutoImplant 2020 challenge. In Cranial Implant Design Challenge (pp. 47-55). Springer, Cham.

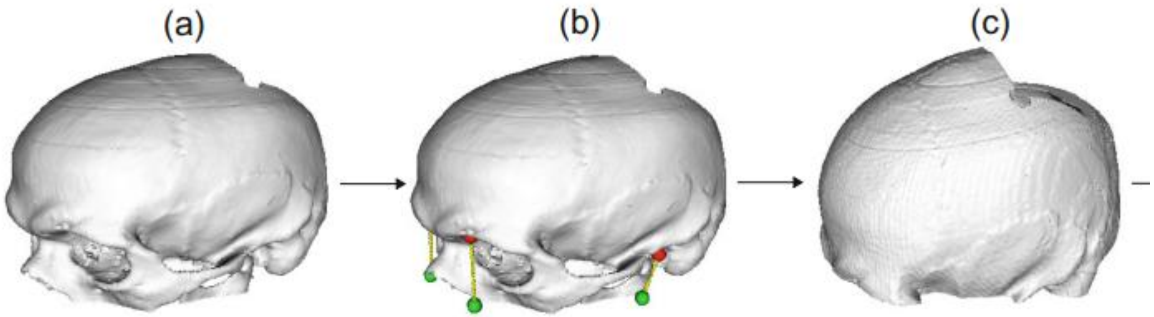
[2] Wodzinski, M., Daniol, M. and Hemmerling, D., 2021, October. Improving the Automatic Cranial Implant Design in Cranioplasty by Linking Different Datasets. In Cranial Implant Design Challenge (pp. 29-44). Springer, Cham.

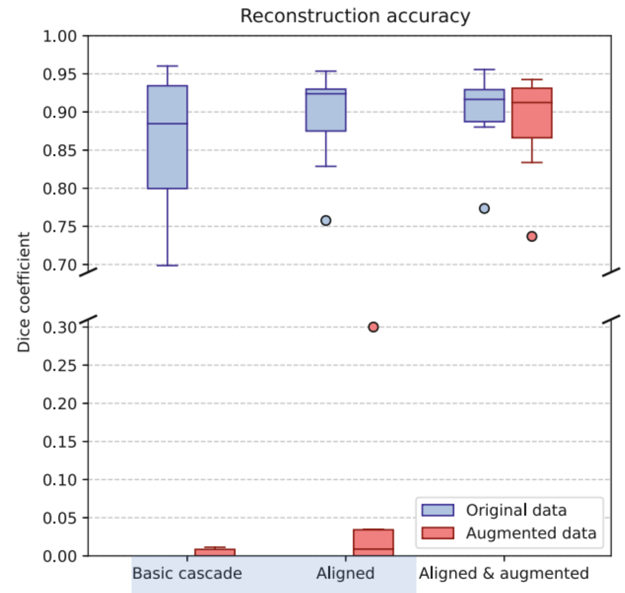
Generalization /Domain shift: skull registration & alignment

AutoImplant I (2020)

skull registration & alignment

- to make the training and test samples uniform (same orientation, position, etc)
- formally speaking, to reduce the difference between the distributions of the training and test sets

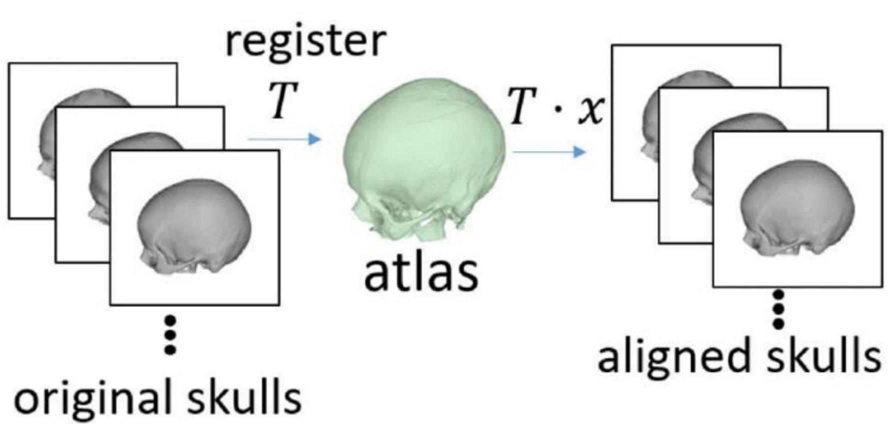




results on a validation set [1]

method [1]:

- manually place four landmarks on each skull in the training set
- align the skulls along the four landmarks using a similarity transformation (scale, rotation, translation), and discard the (facial) bones below the alignment plane
- align the test samples the same way as training samples



method [2]:

- register the training and test samples to a common (pre-selected) reference skull atlas
- the training and test samples have the same size, orientation, etc, and minimum differences
- the atlas is created by averaging several complete skull (a mean skull shape)

Method [1,2]: both used 3D registration & an inverse transform is needed to convert the results back to the original space

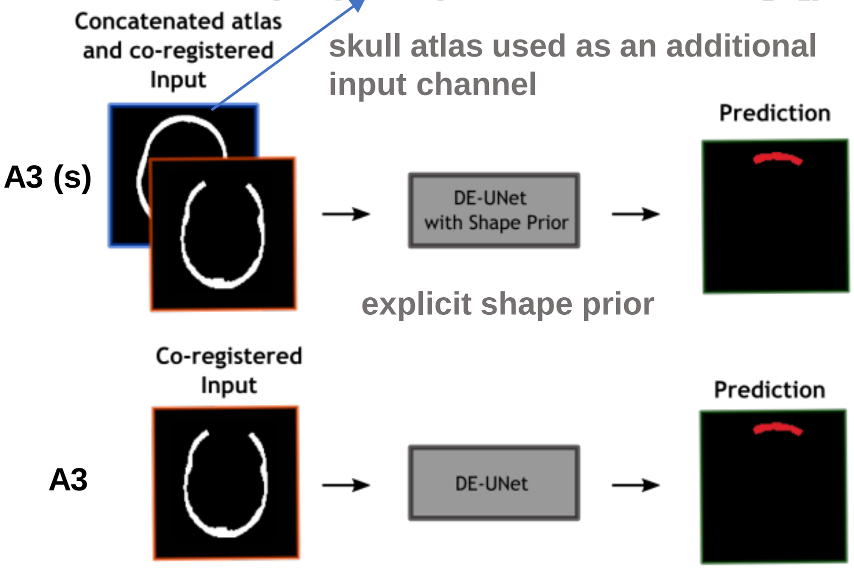
[1]. Kodym, O., Španěl, M. and Herout, A., 2020, October. Cranial defect reconstruction using cascaded CNN with alignment. In Cranial Implant Design Challenge (pp. 56-64). Springer, Cham.
[2]. Matzkin, F., Newcombe, V., et al., 2020, October. Cranial implant design via virtual craniectomy with shape priors. In Cranial Implant Design Challenge (pp. 37-46). Springer, Cham.

Generalization /Domain shift: shape prior & regularization

AutoImplant I (2020)

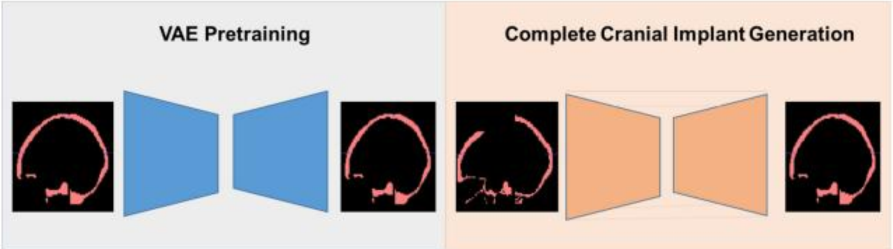
shape prior(Matzkin, F et al [1])

skull atlas used as an additional input channel



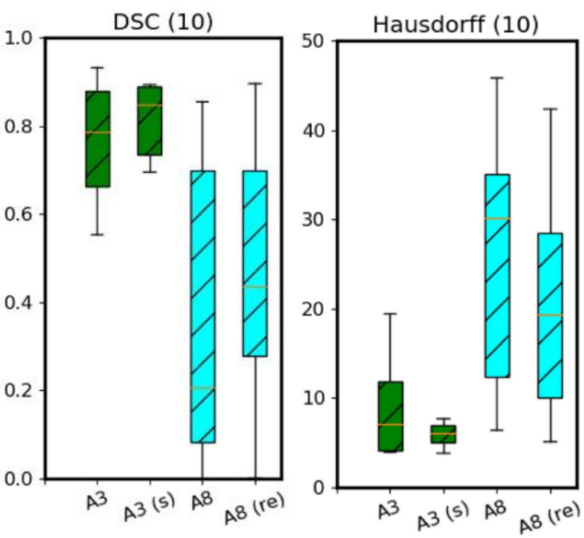
explicit shape prior

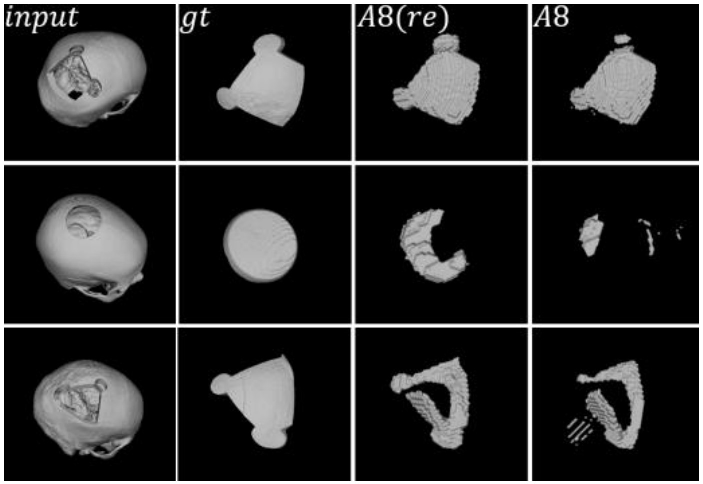
regularization(Wang, B. et al [2])



- train a VAE using only complete skulls
- minimize the difference between the latent variables of the ground truth and predictions during training of a shape completion network (a regularization term in the loss function $L = L_{dice} + \gamma||Z_{gt} - Z_{pred}||$

implicit shape prior





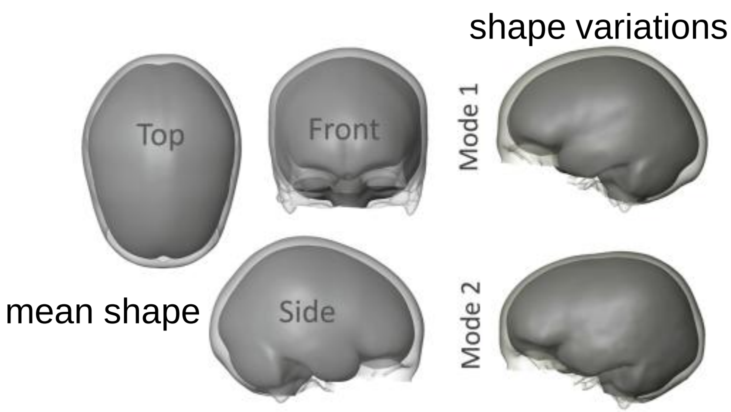
A3: defect augmentation, skull alignment
A3 (s): defect augmentation, skull alignment, shape prior

A8: unsuccessful (no augmentation, no preprocessing)
A8 (re): unsuccessful but better, quantitatively and qualitatively

[1]. Matzkin, F., Newcombe, V., et al., 2020, October. Cranial implant design via virtual craniectomy with shape priors. In Cranial Implant Design Challenge (pp. 37-46). Springer, Cham.
[2]. Wang, B., Liu, Zet al., 2020, October. Cranial implant design using a deep learning method with anatomical regularization. In Cranial Implant Design Challenge (pp. 85-93). Springer, Cham.

Generalization /Domain shift: statistical shape model

Unlike deep learning approaches that require *complete-defect* or *complete-implant* pairs for training, only complete skulls are needed to build a statistical shape model, so that it is not affected by the variations of the defects in the training and test sets



mesh-based SSM of the **complete** skulls [1]

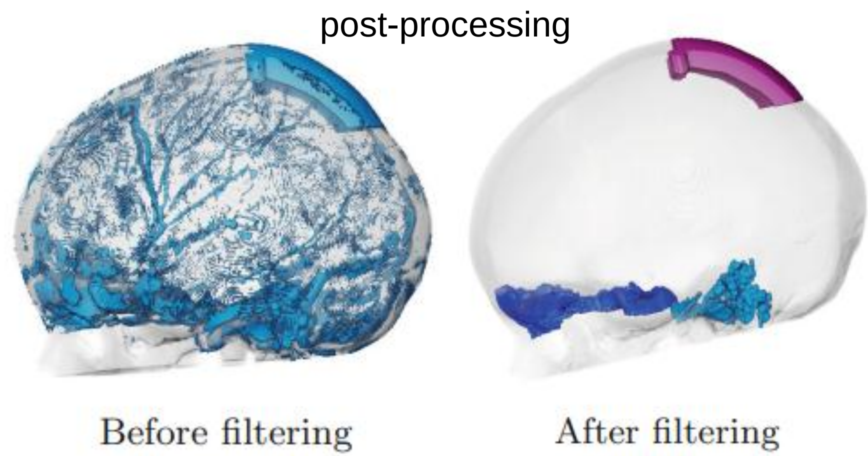


image-to-mesh -> mesh-to-image -> subtraction & post-processing (implant)

results: defect variations barely affect SSM's performance (Pimentel, P. [1])

	Test case (100)	Test case (10)	Overall (110)
mean DSC	0.917	0.919	0.917
mean HD	4.336	3.987	4.304

----

[1]. Pimentel, P., Szengel, A., Ehlke, M., Lamecker, H., Zachow, S., Estacio, L., Doenitz, C. and Ramm, H., 2020, October. **Automated virtual reconstruction of large skull defects using statistical shape models and generative adversarial networks**. In Cranial Implant Design Challenge (pp. 16-27). Springer, Cham.

Generalization /Domain shift: statistical shape model

1. **different methods for building and fitting an SSM**
2. **work directly on images instead of meshes: no image-mesh-image conversion needed**
3. **evaluated on both synthetic and clinical data (AutoImplant II & MUG500+**(Li, J. et al 2021 [2])**)**

workflow (Li, J. et al 2022 [1])

training samples x_i

T: scaling, rotation and translation

reference skull x_0 ← **a test sample y_j**

x'_i

y'_j

$$\bar{S} = \frac{1}{C}\sum_{i}^{C} x'_i$$

mean shape

PCA → $\mathbf{\Phi}$

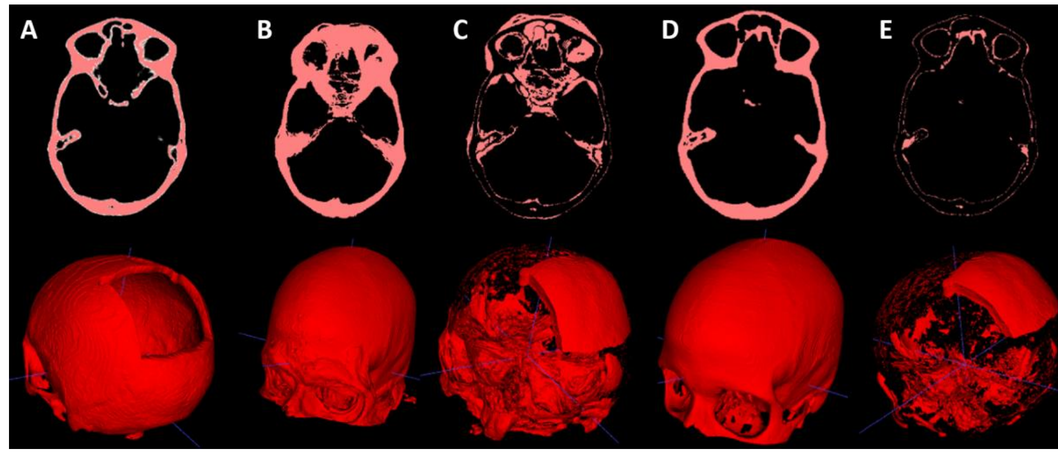
shape variations

$$\boxed{\bar{S} + y' \cdot \mathbf{\Phi}^T \cdot \mathbf{\Phi}}$$

new shape

T^{-1} **inverse transform**

$\bar{S}$ (30) x_i



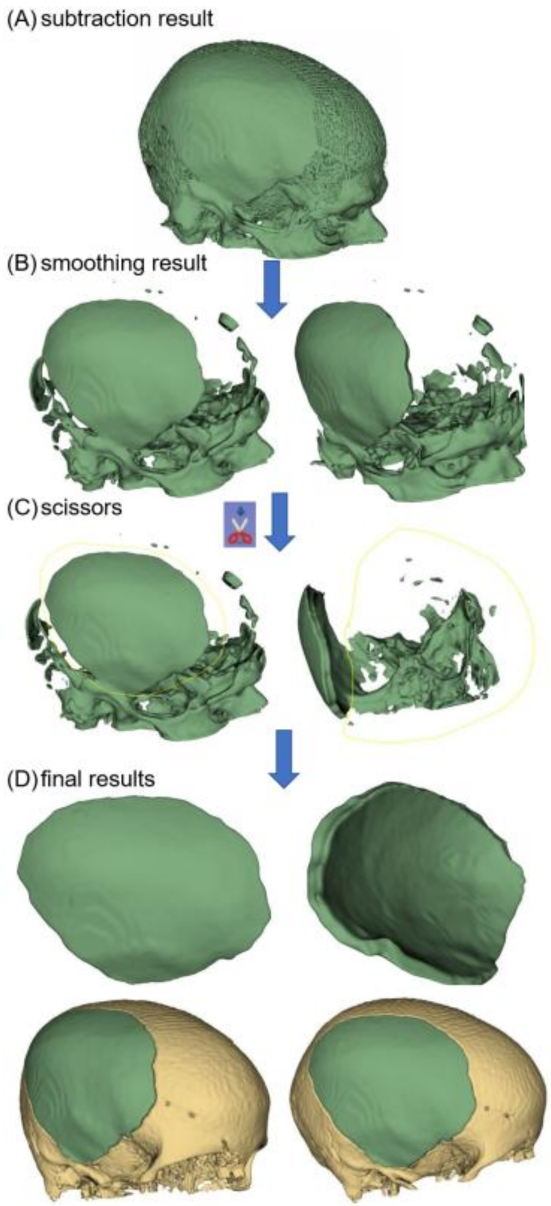
- the implant is easily separable from the subtraction result – cranium registration is accurate.
- noise occurs mainly in the facial area – subtle facial structures are not (or cannot be) registered properly
- using a mean shape or a single shape makes little difference on cranium reconstruction. A mean shape mainly adds to the complexity of the facial bones.

[1]. Li, J, Ellis, David G, et al. 2022, April, **Back to the Roots: Reconstructing Large and Complex Cranial Defects using an Image-based Statistical Shape Model.** arXiv:2204.05703
[2]. Li, J., Krall, M., et al., 2021. **MUG500+: Database of 500 high-resolution healthy human skulls and 29 craniotomy skulls and implants**. Data in Brief, 39, p.107524.

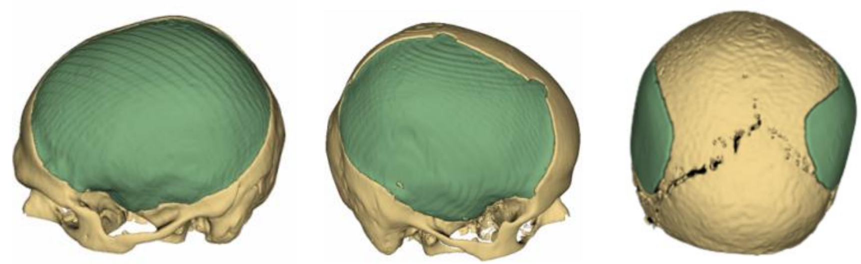
Generalization /Domain shift: statistical shape model

AutoImplant II (2021)

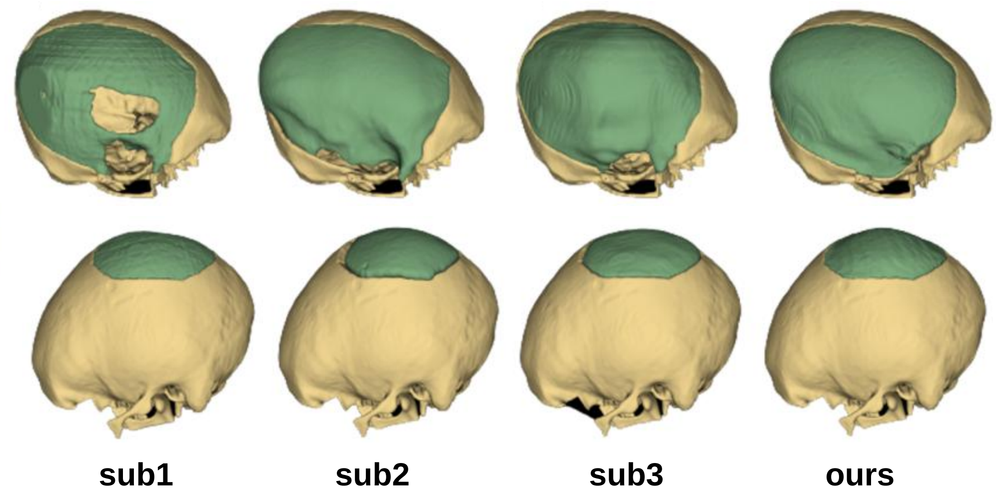
manual post-processing



results on mug500+ dataset (29 cases in total)



results on Task2@AutoImplant II dataset (11 cases in total)



neurosurgeons' evaluation Task2@AutoImplant II dataset

Methods \ Scores	Comp	FPA	Fit	Feasibility
$\bar{S}$ (50)	0.89	0.73	0.64	0.62
M. Wodzinski. et al.	0.93	0.57	0.55	0.42
L. Yu. et al.	0.80	0.59	0.36	0.42
H. Mahdi. et al.	0.76	0.43	0.45	0.33

Generalization /Domain shift: statistical shape model

Is deep learning too much for a ‘simple’ task as automatic cranial implant design?

Yes, since:

- a simple SSM produces better results on clinical cases than all previous deep learning approaches
- the reconstruction process of an SSM is transparent and interpretable (mean shape + shape variations)
- a SSM does not need clinical cases for training but still generalizes well to clinical cases in evaluation

And no, since:

- The reason why a CNN performs poorly on real cases is due to a lack of large quantities of annotated clinical cases
- the batch-wise training scheme enables deep learning to train on arbitrarily large datasets, while the number of images used to build an SSM is limited (the covariance matrix, matrix inverse etc is computationally intensive).
- state of the art deep learning approaches still far out-perform SSM on synthetic defects, by training on large quantities of synthetic data.
- the cranium is structurally simple so that registration accuracy is high. SSM might not perform as well on more complex structures such as the facial bones. (the registration step determines largely the quality of point correspondence, and hence the final results)

Table I: Quantitative results (mean DSC, bDSC and HD95) on the 110 test cases of Task 3.

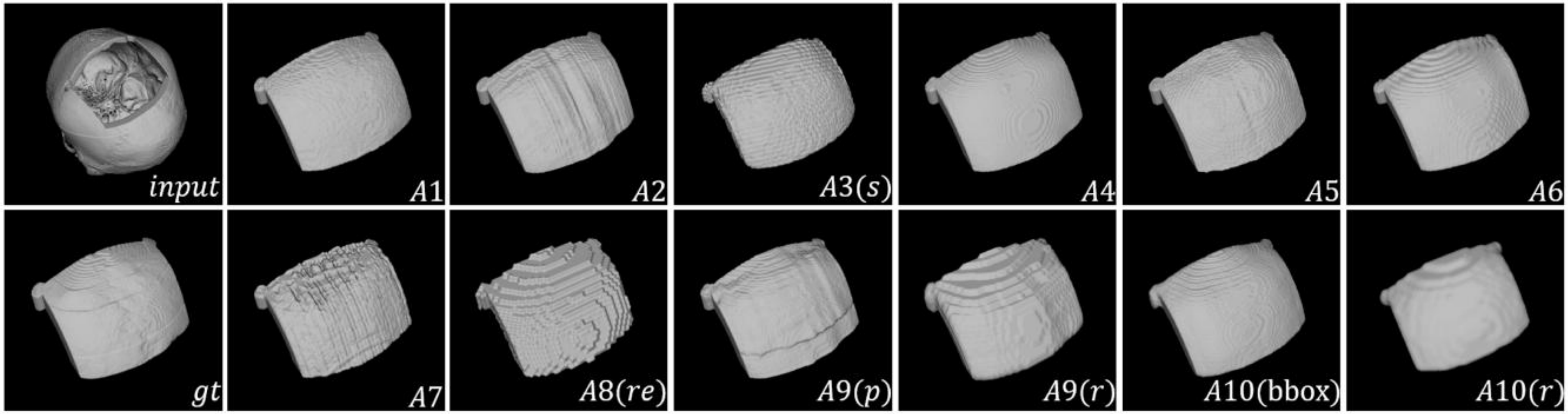
Methods \ Scores	DSC	bDSC	HD95 (mm)
$\bar{S}$ (30)	0.7840	0.8265	3.1989
$\bar{S}$ (50)	0.7853	0.8287	3.2447
x_j	0.7854	0.8285	3.1700
SSM (30)	0.7832	0.8255	3.2157
SSM (30) + DL	0.7830	0.8253	3.2170
DL [36, 27]	0.8058	0.7638	13.2891
$\sum_{i=1}^{d_0} \lambda_i \Phi_i$ ($\lambda_i = 1$)	0.7054	0.7403	3.6783
$\sum_{i=1}^{d_0} \lambda_i \Phi_i$	0.7064	0.7411	3.6601
L. Yu. et. al. [35]	0.7728	0.7716	3.6803
D. G. Ellis, et.al. [38]	0.9440	-	-
M. Wodzinski et. al. [32]	0.9329	0.9530	1.4781

Li, J, Ellis, David G, et al. 2022, April, **Back to the Roots: Reconstructing Large and Complex Cranial Defects using an Image-based Statistical Shape Model.** arXiv:2204.05703

High memory footprint/slow training

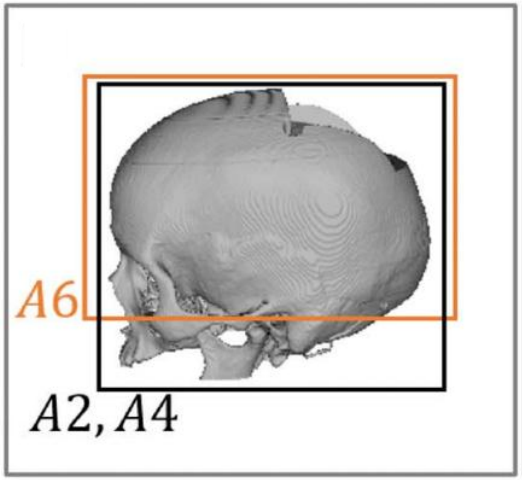
Input image is of high resolution (512*512*Z)

- **High memory footprint: GPU memory is limited**
- **Slow training (FLOPs): e.g., training takes seven days on two V100 GPUs for Ellis, D.G. et al (AutoImplant I, 1st place)**



Qualitative comparison of the implants produced by different methods (AutoImplant I)

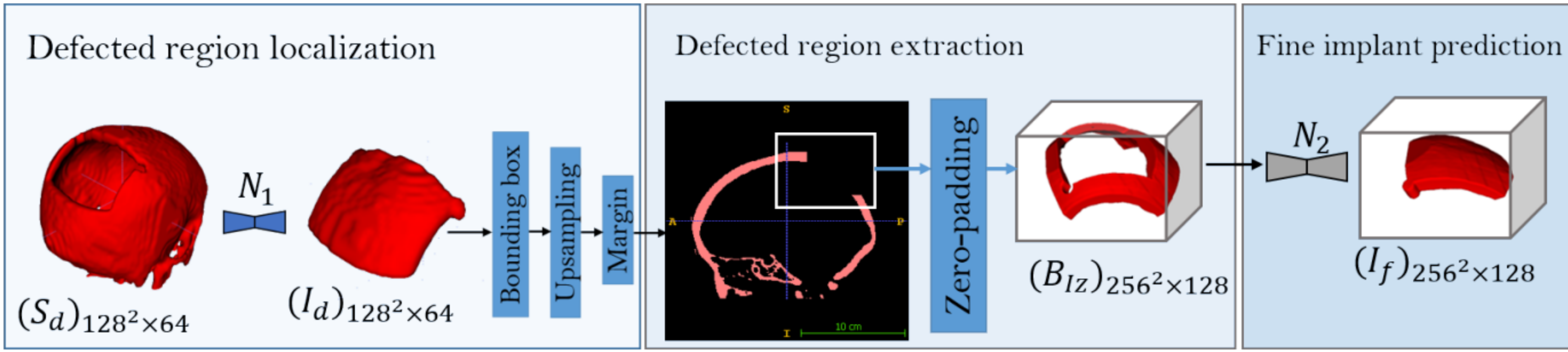
- **Obviously the quality of the implants varies**
- **Most methods downsample or resample the images to a smaller size, at the cost of loss of image quality (coarse input -> coarse output).**
- **The degree of down-sampling is negatively correlated with the implant quality: A8(re), A10 (r), 128*128*64. A9(r) 256*256*54**
- **To ease the negative effects of down/sampling, one can crop the image before down/sampling: crop the background and/or the facial area**
- **Other popular and effective approaches: coarse-to-fine, sparse CNN**

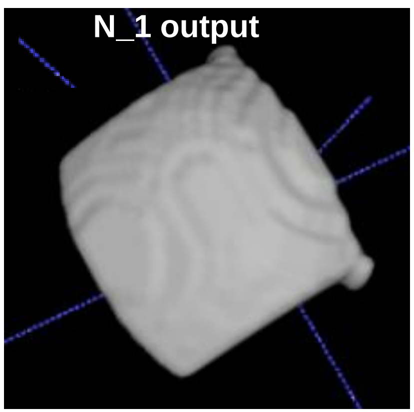


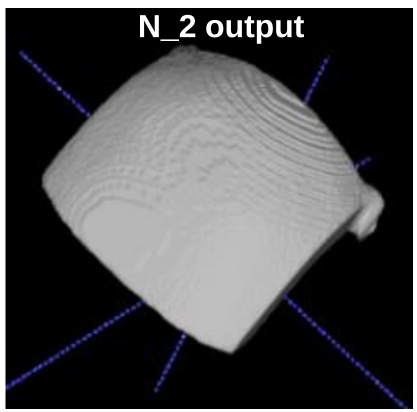
High memory footprint/slow training: coarse-to-fine prediction

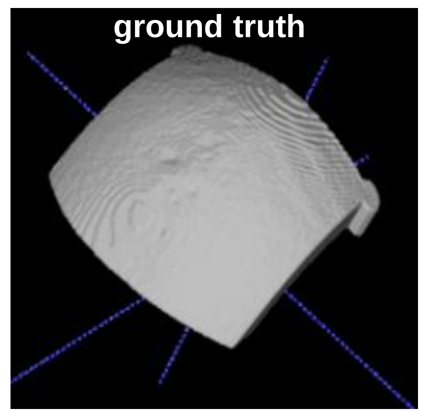
Two-step (Li, J .et al. [1])

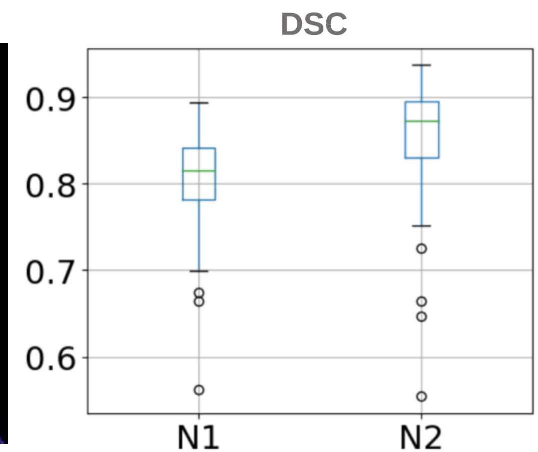
- **Step 1: use a CNN (N_1) to predict a coarse implant**
- **Step 2: use another CNN (N_2) to predict the fine implant based on a bounding box defined by the implant from Step 1**











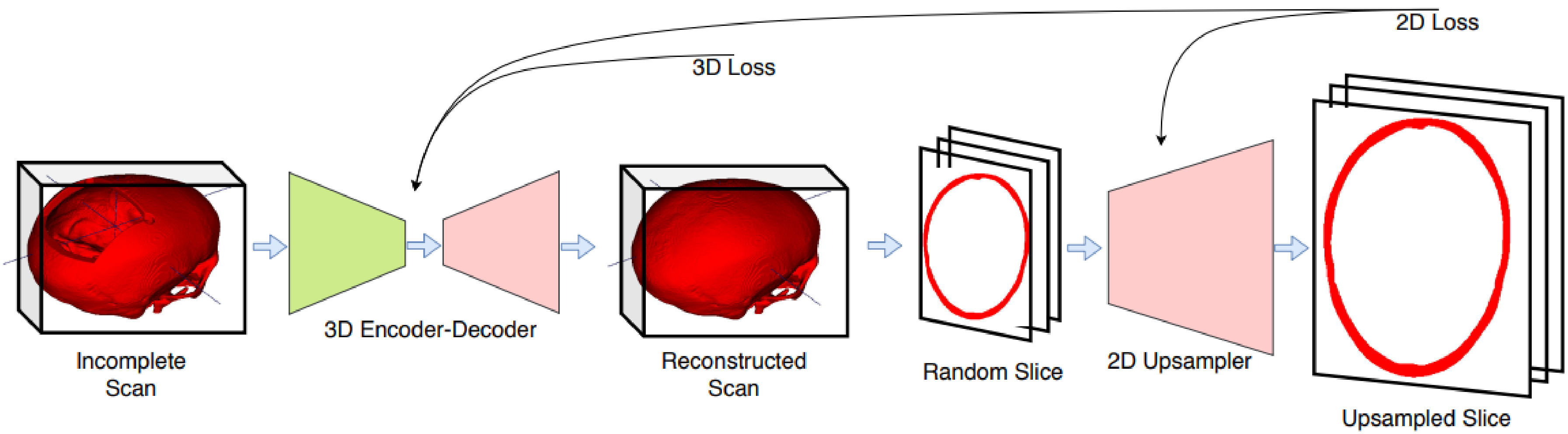
Additional remark: a CNN does not need to 'see' the entire skull to make reasonable predictions to the missing shape

[1]. Li, J., Pepe, A., Gsaxner, C., Campe, G.V. and Egger, J., 2020. A baseline approach for AutoImplant: the MICCAI 2020 cranial implant design challenge. In Multimodal Learning for Clinical Decision Support and Clinical Image-Based Procedures (pp. 75-84). Springer, Cham.

High memory footprint/slow training: coarse-to-fine prediction

Single-step, end-to-end (Bayat, A et al [1])

- **3D shape completion at low resolution + 2D super-resolution**
- **3D and 2D loss are combined to enable an end-to-end training**



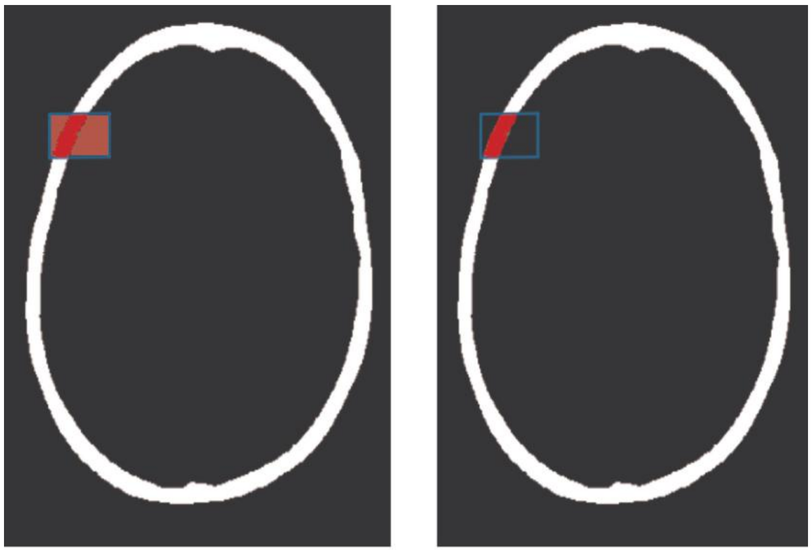
---

[1] Bayat, A., Shit, S., Kilian, A., Liechtenstein, J.T., Kirschke, J.S. and Menze, B.H., 2020, October. Cranial implant prediction using low-resolution 3D shape completion and high-resolution 2D refinement. In Cranial Implant Design Challenge (pp. 77-84). Springer, Cham.

High memory footprint/slow training: Sparse CNN

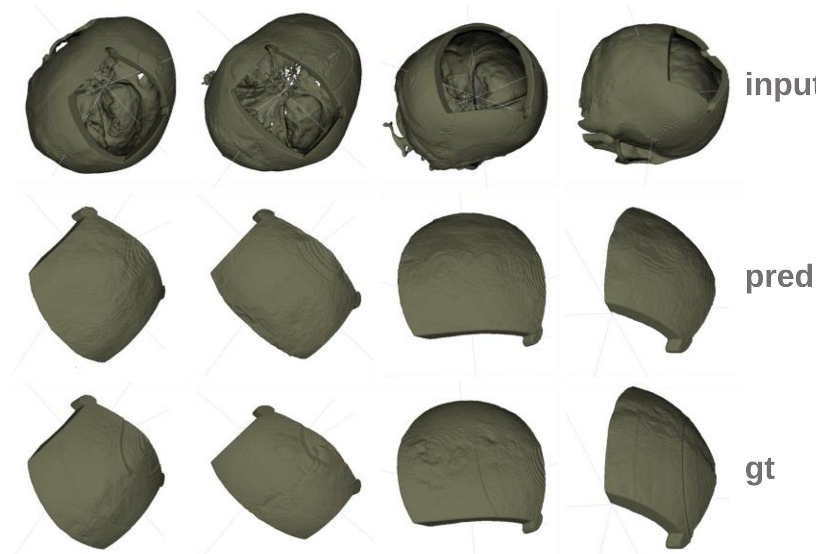
Sparse CNN for Medical Image Analysis (Li, J .et al. [1])

- **The skull voxels (‘1’) are sparsely distributed in the binary (‘0’ background, ‘1’ skull) image**
- **Skull images can be seen as ‘sparse tensors’, where most voxels are ‘0’**
- **Traditional convolutions are inefficient in processing sparse tensors (a)**
- **Minkowski Engine (Choy C. et al [2]) is designed specifically for sparse tensors (b)**



(a) Traditional convolution *(b) Sparse convolution*

- **Traditional convolution: consumes both ‘1s’ and ‘0s’**
- **Sparse CNN: consumes only the ‘1s’ -> low memory usage & low FLOPs**



Sparse CNN results:

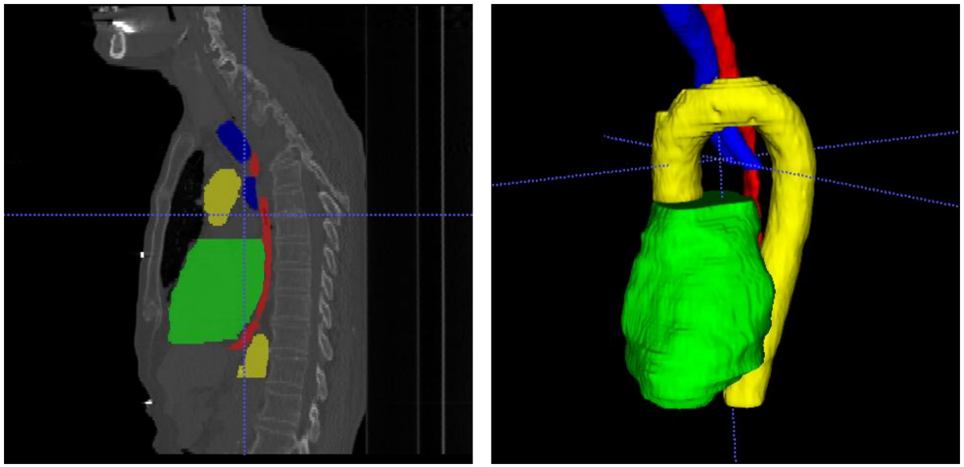
- **can take as input the original image (512*512*Z)**
- **can output the implants directly at 512*512*Z**
- **use about 11GB memory for training and 3GB for evaluation**
- **Fast: takes around 3 hours to train at full image resolution (512*512*Z)**

[1] Li, J., Gsaxner, C., Pepe, A., Schmalstieg, D., Kleesiek, J. and Egger, J., 2022. Sparse Convolutional Neural Networks for Medical Image Analysis.
[2] Choy, C., Gwak, J. and Savarese, S., 2019. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3075-3084).

High memory footprint/slow training: Sparse CNN

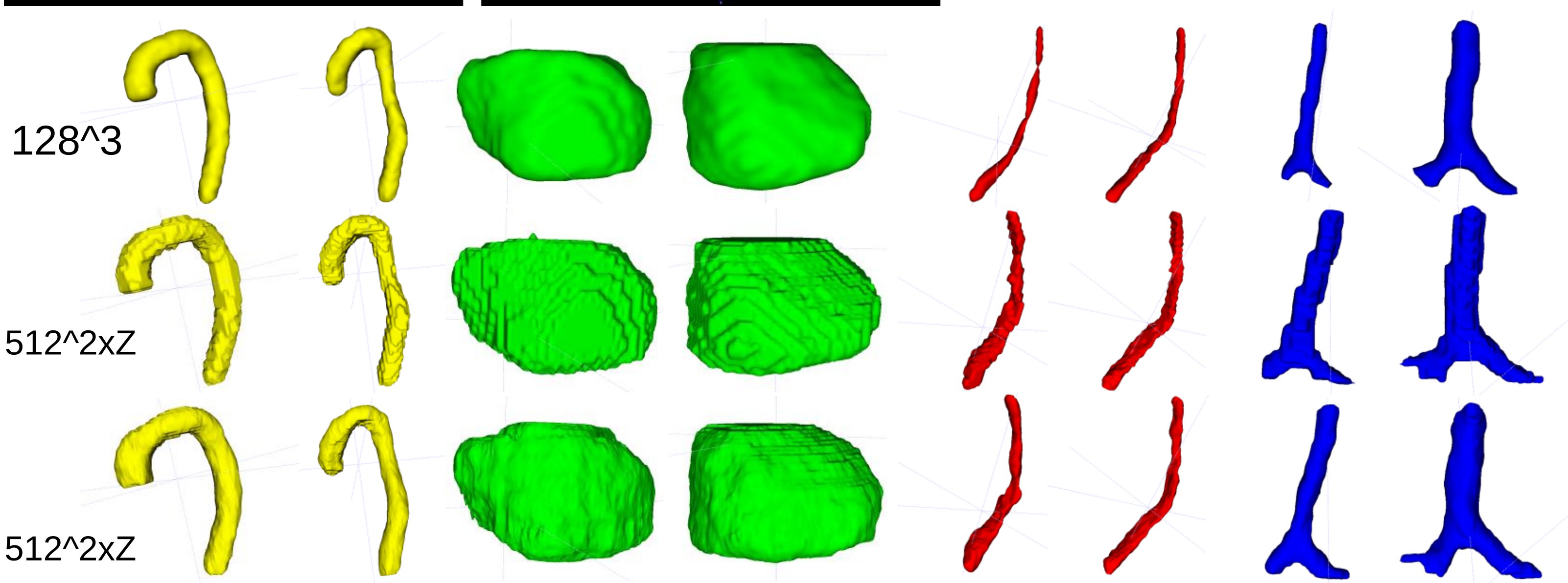
Sparse CNN can be used for the refinement of the segmentation masks (Li, J .et al. [1])

- **Image resolution: 512*512*Z**
- **Voxel occupancy rate of the organs is very low (see the table below)**
- **Workflow: dense CNN coarse segmentation (128^3) -> sparse CNN refinement (coarse-to-fine)**



organ	train	test	VOR (%)
aorta	2.05	1.75	0.20
heart	2.46	2.38	0.79
trachea	1.73	1.64	0.04
esophagus	1.77	1.64	0.05

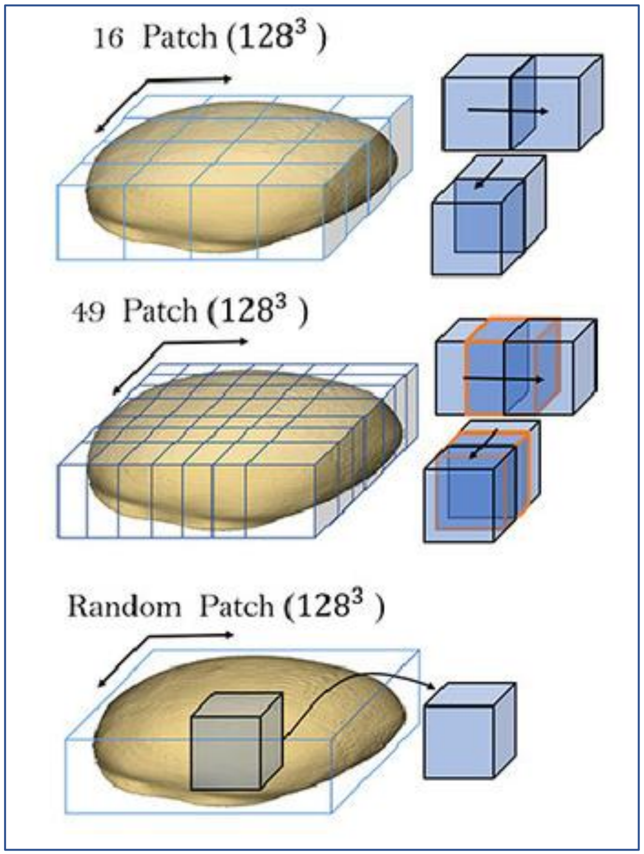
Table S1.Voxel occupancy rate (VOR) and the memory usage (in GB) during training and inference for different organs.



[1] Li, J., Gsaxner, C., Pepe, A., Schmalstieg, D., Kleesiek, J. and Egger, J., 2022. Sparse Convolutional Neural Networks for Medical Image Analysis.

High memory footprint/slow training: patch-wise training

Patch-wise training and inference (Li, J .et al. [1])
- **use image patches to avoid down-sampling**
- **tailored training strategies to maintain generalizability while training on patches**



[1] Li, J., von Campe, G., Pepe, A., Gsaxner, C., Wang, E., Chen, X., Zefferer, U., Tödtling, M., Krall, M., Deutschmann, H. and Schäfer, U., 2021. Automatic skull defect restoration and cranial implant generation for cranioplasty. Medical Image Analysis, 73, p.102171.

To summarize:

Major technical challenges/solutions in automatic cranial implant design:

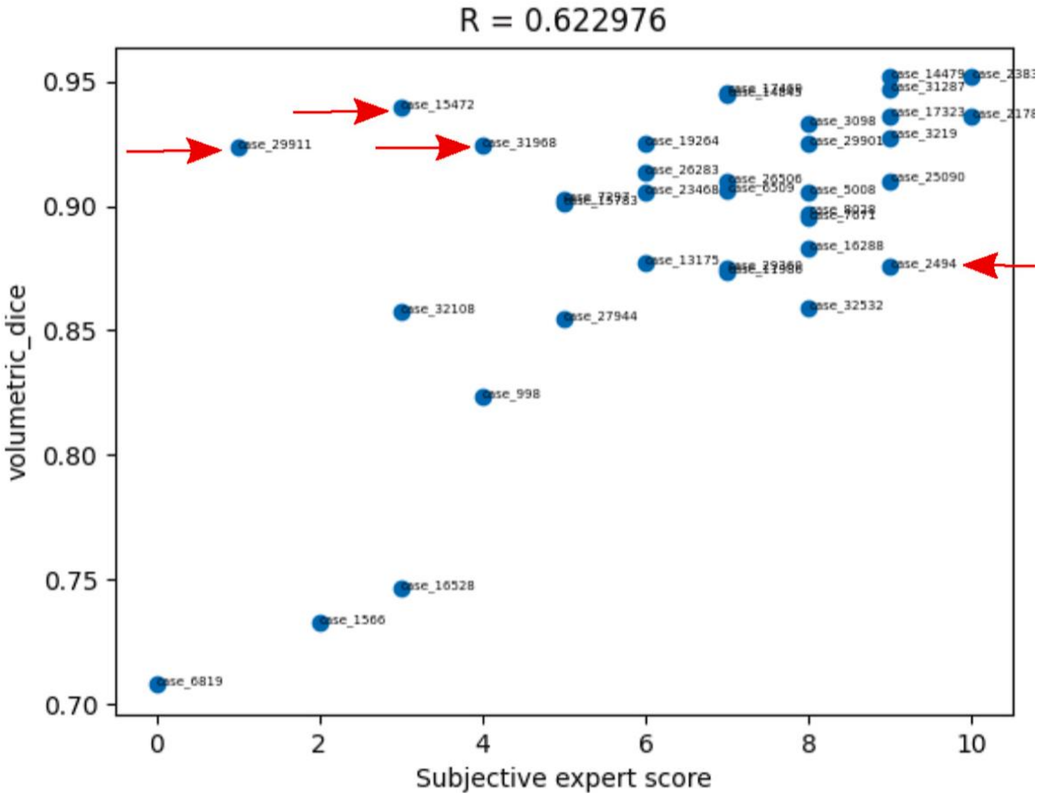
- Domain shift: data augmentation, skull registration & alignment, shape prior, regularization, statistical shape model
- High memory footprint & slow training: coarse-to-fine, patch-wise training and inference, sparse cnn

High-level Insights:

- Architectural variations do not make a big difference in performance and ranking.
- Top ranking submissions usually use a combination of the above mentioned methods.
- Data augmentation has the most influence on the generalizability and performance (Winners of both challenges used intensive data augmentation).

Issue 1: Quantitative evaluation of submissions' clinical feasibility

- traditional quantitative metrics (DSC, HD, HD95 etc) are not closely correlated with the usability of an implant



Kodym, O[1]:

- **correlation between quantitative scores and experts' evaluation is positive but weak.**
- **discrepancies: there are predictions with high dice scores but low usability and vice versa.**

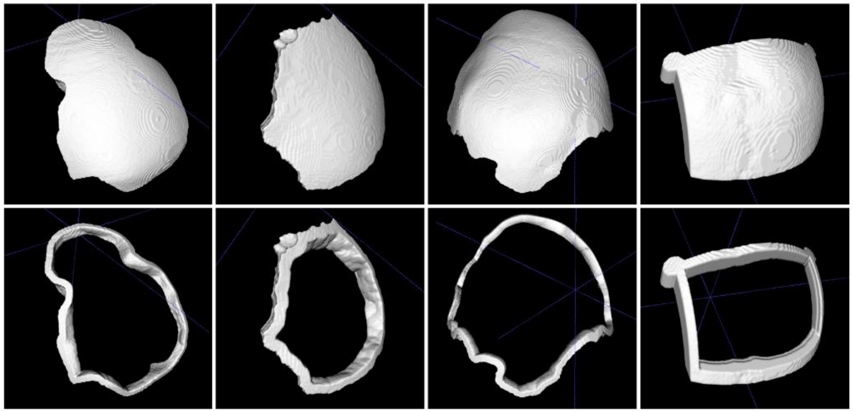
---

[1] Kodym, O., et al., 2021. Deep learning for cranioplasty in clinical practice: Going from synthetic to real patient data. Computers in Biology and Medicine, 137, p.104766.

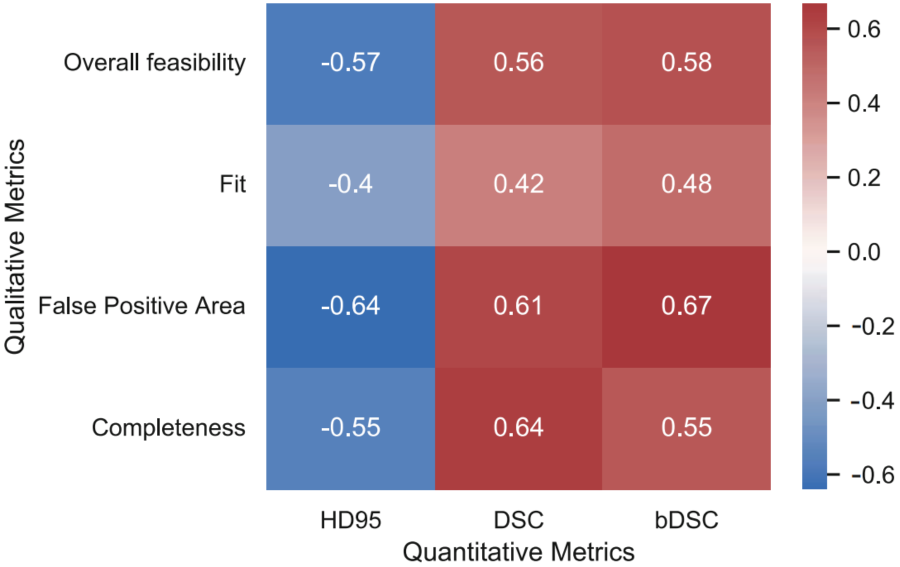
Issue 1: Quantitative evaluation of submissions' clinical feasibility

Solutions:

- customized quantitative metrics: border DSC (affected less by the overall thickness of the implants)
- quantification of experts' qualitative evaluation (Ellis, D G et al [1])



- border DSC measures the similarity only around the borders
- Border DSC has stronger correlation with experts' evaluation than traditional metrics (DSC, HD95)



the quantified scores reflect an expert's view on the submissions' feasibility

- compare different methods
- ranking

Table 2. Qualitative criteria for a feasible implant design.

Criteria	Description
Complete	The implant should cover the whole defect area
No false positive area	The implant should not extend beyond the defect area
Implantable	The implant should be able to be placed into the defect area
Restores skull shape	The implant should restore the expected skull shape
Smooth transition with skull	The area of transition between the skull and implant should be smooth
Minimal thickness	The implant must be thin enough as not to overly compress underlying tissue. Ideally, the implant should be at least 50% thinner than the skull

[1] Ellis, D.G et al., 2021, October. Qualitative Criteria for Feasible Cranial Implant Designs. In Cranial Implant Design Challenge (pp. 8-18). Springer, Cham.

Issue 1: Quantitative evaluation of submissions' clinical feasibility

Table III

Methods \ Scores	DSC	bDSC	HD95
$\bar{S}$ (50)	0.5007	0.4449	8.2539
SSM (30)	0.5055	0.4470	7.9042
M. Wodzinski. et al. [32]	0.5241	0.4823	54.5165
L. Yu. et al. [35]	0.5118	0.4547	8.3486
H. Mahdi. et al. [31]	0.3028	0.3092	71.4193

Table IV

Methods \ Scores	Comp	FPA	Fit	Feasibility
$\bar{S}$ (50)	0.89	0.73	0.64	0.62
M. Wodzinski. et al. [32]	0.93	0.57	0.55	0.42
L. Yu. et al. [35]	0.80	0.59	0.36	0.42
H. Mahdi. et al. [31]	0.76	0.43	0.45	0.33

	Completeness	False Positive Area	Fit	Overall implant feasibility*
1_li1	100	Minimal	Yes	Feasible with minor flaws
2_li1	100	Minimal	Yes	Feasible with minor flaws
3_li1	>75	None	Yes	Feasible with minimal modifications
4_li1	>75	Minimal	Yes	Feasible with minimal modifications
5_li1	>75	Moderate	N	Feasible with significant modifcations
6_li1	100	Moderate	N	Feasible with significant modifcations
7_li1	>75	None	Yes	Feasible with minimal modifications
8_li1	>75	Moderate	N	Feasible with significant modifcations
9_li1	>75	Minimal	N	Feasible with minimal modifications
10_li1	100	Minimal	Yes	Feasible with minor flaws
11_li1	100	Minimal	Yes	Feasible with minor flaws

evaluation of the results from Li, J et al [1] based on common quantitative metrics (DSC, HD95 Table III) and quantified qualitative metrics (Table IV)

- **ranked differently by different evaluation methods**
- **the quantitative scores (in Table III) alone cannot be used to judge the feasibility of an implant**

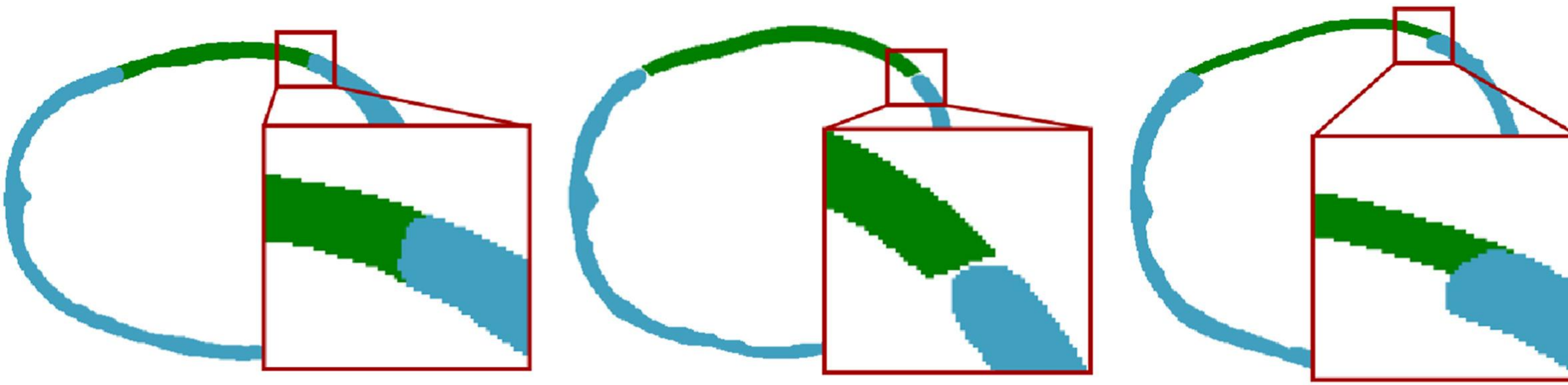
---

[1] Li, J, Ellis, David G, et al. 2022, April, **Back to the Roots: Reconstructing Large and Complex Cranial Defects using an Image-based Statistical Shape Model.** arXiv:2204.05703

Clinical Feasibility Evaluation

Issue 2: the synthetic defects used for training is defined differently from the clinical defects

- **AutoImplant I, II: all methods are trained on synthetic samples for a perfect fit, whether or not evaluated on clinical samples**



Kodym, O. et al [1]: (a) synthetic defetcs **(b) real defect + experts designed implant**

synthetic samples: the ground truth is simply the removed part. The implant fits the defect seamlessly in terms of borders and thickness

clinical samples: the ground truth is the implants manually designed by experts. the implant is thinner than the skull bones and implant does not necessarily fits the defect

The results from the automatic methods are not directly usable:

- **use the clinical cases for training directly: not enough training samples**
- **(manually) edit the automatic results to meet the clinical requirements**

----

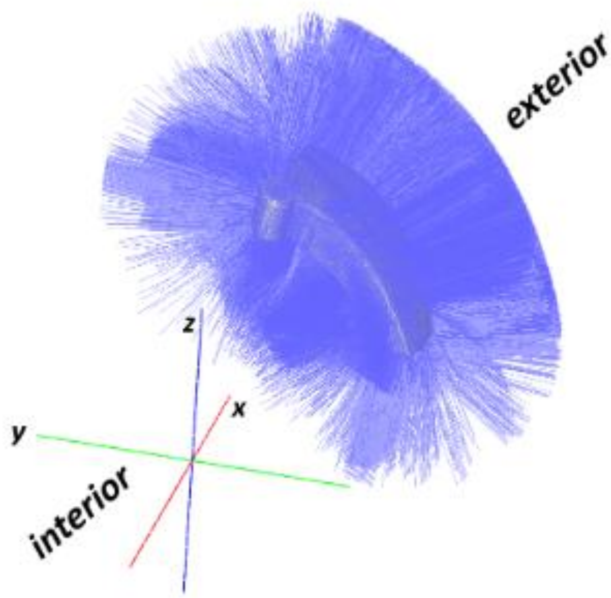
[1] Kodym, O., et al., 2021. Deep learning for cranioplasty in clinical practice: Going from synthetic to real patient data. Computers in Biology and Medicine, 137, p.104766.

Clinical Feasibility Evaluation

Issue 2: the synthetic defects used for training is defined differently from the clinical defects

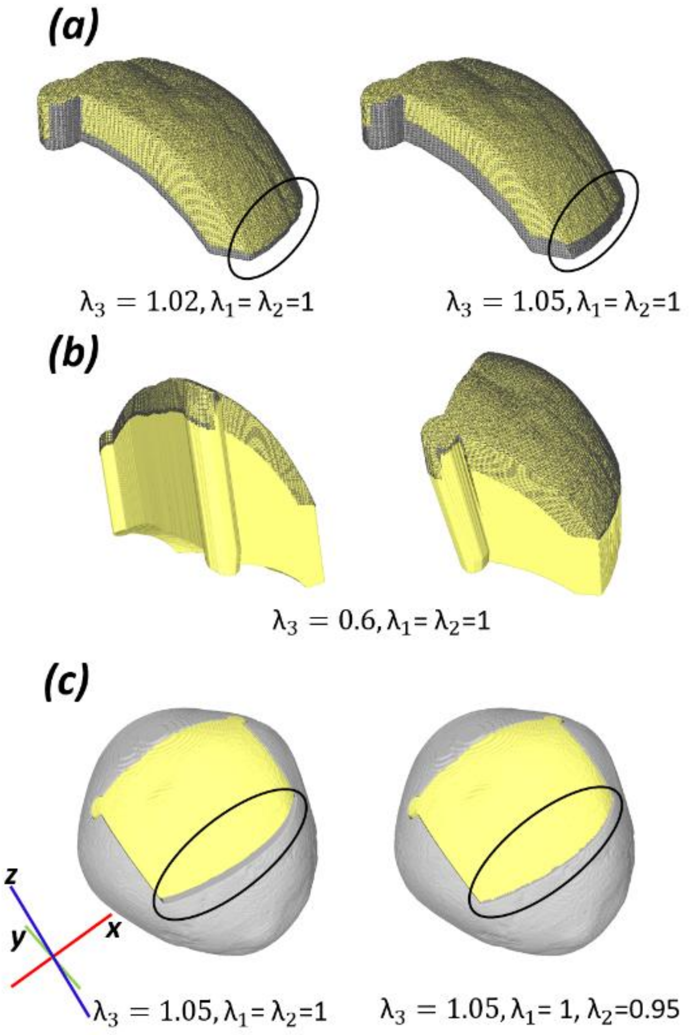
Solutions:

- **(manually) edit the automatic results to meet the clinical requirements**



$$\varphi(\lambda_1 x,\ \lambda_2 y, \lambda_3 z,)=0$$

rescale the coordinates of the inner implant surface to adjust the thickness and borders



Conclusions

Organizational efforts:

- Thanks to the challenge, there have been an increased interest in automatic cranial implant design in the community.
- Get to know / collaborate with other groups around the world working on the same problem.
- The submissions from different groups lay solid foundations for future studies on this or similar topics.

Technical contributions:

- The generalization/domain shift problem in 3D shape completion/inpainting.
- The sparse problem: how to efficiently process high-resolution but sparse data & Under a memory constrained environment, how to obtain high-resolution output.
- Limitations of common metrics in cranial implant design: how to quantify experts' qualitative assessment.
- The methods presented at the challenge is generalizable and applicable to other problems.

Thank You

Questions?