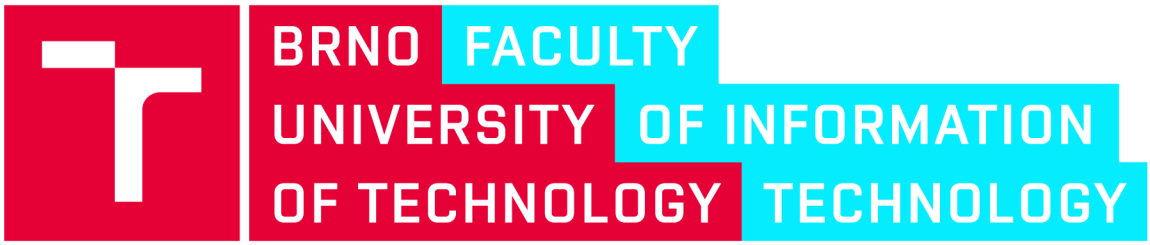

DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

DEEP LEARNING FOR VIRTUAL PATIENT-SPECIFIC SKULL MODELLING AND RECONSTRUCTION

HLUBOKÉ UČENÍ PRO VIRTUÁLNÍ PERSONALIZOVANÉ MODELOVÁNÍ LEBKY A JEJI REKONSTRUKCI

PH.D. THESIS

DISERTAČNÍ PRÁCE

AUTHOR

AUTOR PRÁCE

ING. OLDŘICH KODYM

SUPERVISOR

VEDOUCÍ PRÁCE

PROF. ADAM HEROUT, PH.D.

BRNO 2021

ABSTRACT

Skull segmentation from 3D patient data and virtual reconstruction of the defective skull shape are the most challenging steps required for creation of patient-specific models of skull. These models are used in cranioplasty practice for surgery planning, patient education and patient-specific implant design, but their utility is currently limited by the amount of manual processing time required to reach sufficient virtual model quality. This thesis aims to streamline this virtual workflow by utilizing deep learning methods.

The thesis proposes a novel solution that consists of an automatic skull segmentation method based on a combination of convolutional neural networks and graph cut algorithm and an automatic virtual skull reconstruction method based on convolutional network cascade. Both of these components are demonstrated to achieve state-of-the-art accuracy. This work also aims to improve reproducibility of the skull reconstruction research by providing a structured synthetic dataset for development and benchmarking of automatic methods.

The main focus of this work is on applicability in clinical practice. While the proposed skull segmentation method is already successfully deployed to clinical workflow, the integration of automatic virtual skull reconstruction presents some additional challenges, such as low tolerance towards shape imperfections around the defect border. This work therefore also proposes an extension of the skull reconstruction method that allows its adaptation to target population and the desired type of cranial implant shape, which can vary between different clinical sites. The results of expert's evaluation show that the shape outputs of this method reach enough quality to be deployed into clinical practice along with the segmentation method.

KEY WORDS

Deep learning; Medical imaging; Cranioplasty; Surgery planning; Skull segmentation, Virtual skull reconstruction.

ABSTRAKT

Segmentace lebky ze 3D pacientských dat a virtuální rekonstrukce tvaru lebek s defekty jsou nejnáročnějšími kroky potřebnými pro tvorbu lebečních modelů na míru pacienta. Tyto modely jsou v kranioplastice využívány pro plánování operací, poučení pacienta a design implantátů na míru, avšak jejich využitelnost je v současnosti limitována množstvím manuální práce potřebné pro dosažení dostatečné kvality virtuálních modelů. Tato teze má za cíl zefektivnění tohoto virtuálního pracovního postupu s využitím metod hlubokého učení.

Teze popisuje klinickou motivaci a současnou výzkumnou literaturu v oblasti automatizace virtuální kranioplastiky. Dále navrhuje nové řešení sestávající z metody automatické segmentace lebky založené na kombinaci konvoluční neuronové sítě a algoritmu graph-cut a metody automatické rekonstrukce lebky založené na kaskádě konvolučních sítí. Obě tyto komponenty demonstrují přesnost na úrovni vědeckého stavu poznání. Dále tato práce cílí na zvýšení reprodukovatelnosti výzkumu lebečních rekonstrukcí poskytnutím strukturovaného syntetického datasetu pro vývoj a srovnávání automatických metod.

Hlavním cílem této práce je využitelnost v klinické praxi. Zatímco navržená metoda segmentace lebek je již v klinické praxi využívána, integrace automatické virtuální rekonstrukce lebky představuje několik dalších překážek, jako nízká tolerance k nepřesnostem ve tvaru okolo hranice defektu. Tato práce proto také navrhuje rozšíření metody rekonstrukce lebky, které umožňuje její adaptaci na cílovou populaci a typ kraniálních implantátů, který se může mezi jednotlivými klinickými pracovištěmi lišit. Výsledky vyhodnocení experta ukazují, že výstupy této metody dosahují dostatečné kvality pro implementaci do klinické praxe společně s metodou segmentace.

KLÍČOVÁ SLOVA

Hluboké učení; Lékařské zobrazování; Kranioplastika; Plánování operací; Segmentace lebky, Virtuální rekonstrukce lebky.

BIBLIOGRAPHIC CITATION

Ing. Oldřich Kodym: *Deep Learning for Virtual Patient-Specific Skull Modelling and Reconstruction*, doctoral thesis Brno, Brno University of Technology, Faculty of Information Technology, 2021.

DECLARATION

I declare that this dissertation thesis is my original work and that I have written it under the guidance of prof. Ing. Adam Herout, Ph.D. and Ing. Michal Španěl, Ph.D. All sources and literature that I have used during my work on the thesis are correctly cited with complete reference to the respective sources.

Brno, 2021

Ing. Oldřich Kodym, June 14, 2021

ACKNOWLEDGMENTS

I would like to thank my supervisor Adam Herout and my consultant Michal Španěl for inspiration, guidance and great deal of support during my studies. I would also like to thank my colleague Michal Hradiš for the opportunity to discuss ideas and Michaela Novosadová and Libor Hrdý from TESCAN Medical company for valuable feedback regarding the clinical side of my research.

Finally, my thanks belong to my parents and, most importantly, to my dearest wife Kateřina for their relentless support, love, and enthusiasm at all times.

CONTENTS

1

INTRODUCTION

Virtual 3D patient-specific anatomical models are becoming an important factor in improving outcomes of surgical interventions as well as reducing the operating time. The goal of the research presented in this thesis is to accelerate the creation of these models with the use of deep learning algorithms. Although some of the approaches presented here are potentially applicable to various target domains, the main focus of this work is on craniomaxillofacial surgery, specifically treatment planning and reconstruction of skull defects. For this application, two types of anatomical models are typically used in the process of surgery planning. First, a precise model of a patient's defective skull serves for the planning of complex procedures, training, or patient education. Second, a virtual reconstruction of the defective tissue geometry provides a template for designing a patient-specific skull implant.

In correspondence with these practical applications, the methodological contributions of this work can be divided into two parts. The first area is precise segmentation of the skull in computed tomography data. Since a plethora of general-purpose segmentation algorithms can be applied to this problem, more attention is given to the cases where the available dataset used for the optimization of the segmentation algorithms cannot fully encompass the variability in data encountered in clinical practice and to the parts of the skull most difficult to segment due to their thinness and low visibility.

The second part of this work addresses the step of the virtual reconstruction of the full skull geometry. Given the shape representation of the patient's defective skull, the reconstructed part of the skull has to fully restore its protective and aesthetic function. In the current medical practice, semi-automatic methods with heavy use of manual refinement in computer-assisted design software are prevalent. As opposed to these often time-consuming approaches, fully automatic skull reconstruction methods can help streamline the clinical workflow. A novel deep learning-based, fully automatic skull reconstruction method is therefore proposed and evaluated. To improve the practical utility of the proposed method, this thesis further proposes an extension that allows adaptability to the target population and type of cranial implant.

Because automatic skull reconstruction is a rather novel area of research, some fundamental topics have not yet been completely addressed in the literature, such as: Is the task of skull shape reconstruction deterministic? Which metrics and what

datasets should be used to evaluate the clinical applicability of the reconstruction methods? To provide grounds for further research in these areas, this work additionally presents a public synthetic dataset of defective skulls. Furthermore, several different reconstruction metrics are studied in relation to how indicative they are of the actual performance of the reconstruction methods in clinical practice.

Several chapters of this thesis are adapted from articles published by the author. The articles are first-author publications and, unless stated otherwise, the research and writing were done primarily by the author with guidance and supervision of the thesis consultant and supervisor.

Part I

BACKGROUND

2

ANATOMICAL MODELS IN SURGERY PLANNING

This chapter serves as an overview of anatomical model applications in skull surgery. It also covers relevant topics of human skull anatomy and pathology as well as patient data acquisition. These topics are discussed to the extent necessary to illustrate both the motivation and current limitations of producing anatomical models for skull surgery planning.

2.1 GENERAL APPLICATIONS OF ANATOMICAL MODELS

3D patient-specific anatomical models, also called biomodels, improve spatial interpretation of patient data by surgeons [36, 102]. In contrast to conventional two-dimensional modalities, such as X-ray data or slices through computed tomography (CT) or magnetic resonance (MR) data, the use of 3D anatomical models brings more versatility to computer-assisted surgery. They may be used in pre-operative (surgical planning and implant design) or intra-operative (surgical navigation) applications, in combination with augmented or virtual reality, or surgical robot guidance systems [126, 105]. Furthermore, anatomical models are used in medical and dental education because they can depict anatomical structures in a more efficient way than cadavers [40].

The recent boom of 3D printing technology and its advancements in both precision and cost-effectiveness further added to the range of anatomical models application [121]. A straightforward application of 3D printed anatomical models is constructing a direct representation of patient's anatomical structures for pre-operative planning and training. Surgical interventions pose various risks, such as blood loss, infection, excessive trauma due to imprecise surgical technique, and time spent under anesthesia. Multiple studies show that using anatomical models during surgery planning reduces operative time and increase the procedure accuracy, effectively leading to better clinical outcome [58, 79, 138]. Such models are also suitable for pre-operative patient education which can increase patient's capacity to provide informed consent. This, in turn, leads to reduced legal risks for the surgeons [78].

However, exact replicas of the patient's anatomical structures are not the only convenient type of anatomical model. In dental, orthopedic, and craniomaxillofacial surgery, custom 3D printed surgical guides are becoming popular [121]. Of

most interest for this thesis, custom patient-specific implants (PSI) can be 3D printed using biocompatible materials such as titanium, polyether ether ketone, or other compounds [96, 50, 10]. As an alternative to directly using the target implant shape, an implant mold can be printed to be used for implant manufacturing later [132].

The aforementioned benefits of using anatomical models for surgical support usually come at the price of the increased cost of the procedure [121]. This is mostly caused by the fact that creating a precise patient-specific anatomical model is a tedious and time-demanding task with little room for inaccuracies and producing the models poses a challenge even for experienced clinical experts. Developing automated solutions to the most challenging steps of the anatomical model creation pipeline and integrating them into the clinical workflow will be crucial for enabling wider patient access to these procedures in the future.

2.2 REPAIRING THE SKULL: CRANIO-MAXILLOFACIAL SKELETON SURGERY

Cranio-maxillofacial (CMF) surgery deals with congenital and acquired deformities of the head, face, and jaw. Although CMF surgery generally deals with all types of tissue, this work focuses on reconstructive bone surgery. The main challenges of this domain lie in reestablishing anatomically correct shape, consistently restoring orbital volume, and accurately repositioning CMF skeleton components into optimal spatial relationships [13].

Suboptimal results of these interventions may lead not only to further clinical problems, but also societal issues related to loss of facial symmetry and aesthetics. This puts additional emphasis on the precision of pre-operative planning. It then comes as no surprise that CMF surgery served as a pioneer in clinical applications of anatomical models [36].

2.2.1 *Overview of Skull Anatomy*

The human CMF skeleton, or skull, is a bony complex that maintains the facial structure and mechanically protects the brain. The two main parts of the skull are the *neurocranium*, consisting of the cranial vault, and *viscerocranium* which encompasses the facial bones. There is a total of 22 bones that comprise the skull (see Figure 2.1) and with the exception of the mandible (the lower jaw), these individual bones are attached by connective regions known as sutures. While the sutures accommodate movement during development to enable the growth of the internal organs, increased level of interdigitation and bone bridging prevents the movement in adult skulls [30].

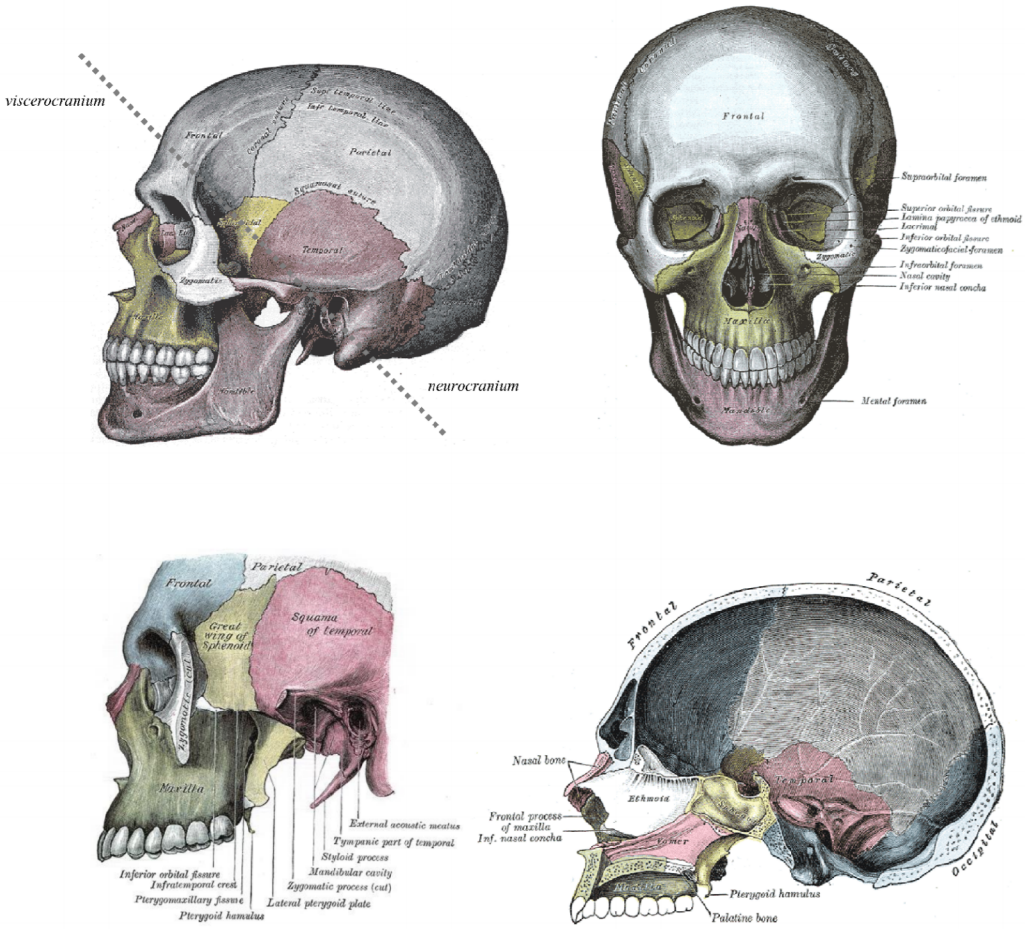

Figure 2.1: Frontal and lateral views of the skull bones. Taken from Gray's Anatomy of the Human Body [46]

Anatomical landmarks provide another way to navigate the complex skull anatomy. Many anatomical landmarks have been defined on the human skull for use in cephalometry which studies variability in human skulls through defined measurements of distances and angles. They can be broken down into two groups, as seen in Figure 2.2: the medial landmarks that lie on the mid-sagittal plane, and the bilateral landmarks that occur laterally on both sides of the skull.

Some of these landmarks also define anatomical planes that can be used for aligning the target skull into a standardized position and for evaluating skull and facial symmetry. The *mid-sagittal plane* (MSP) is anatomically defined as a vertical plane passing through the skull midline and it can be defined by any three medial landmarks, such as the basion, nasion, and sella turcica [8]. This plane marks the theoretical symmetry axis of the skull, but it is worth noting that no skull is perfectly symmetric and slight asymmetries are a natural part of the skull anatomy [67]. The *Frankfort-horizontal plane* (FHP) defined by the orbitale and the porion landmarks then defines the standard horizontal inclination of the skull [97, 84]. In clinical practice, this plane is also preferred for the identification

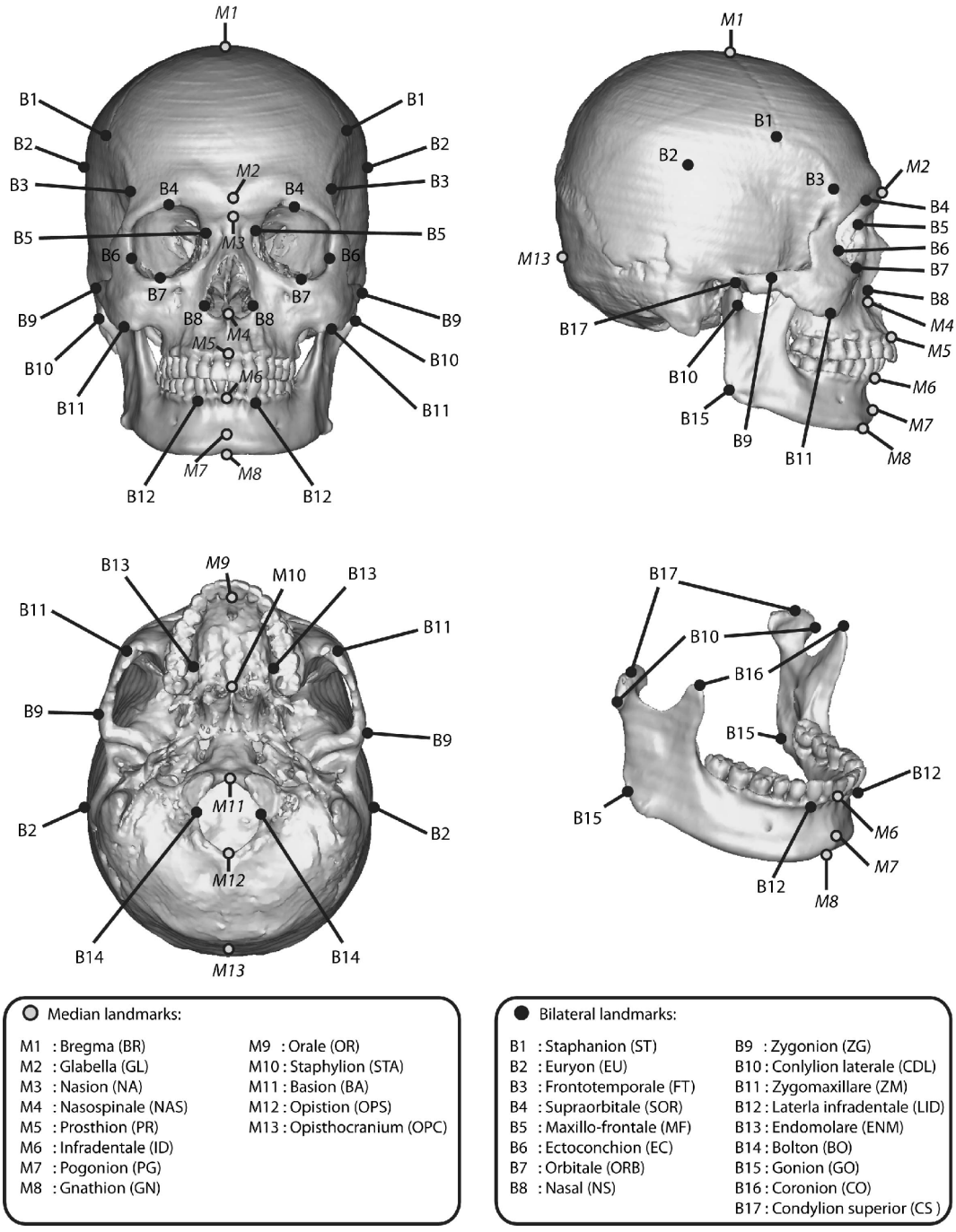

Figure 2.2: Cephalometric landmarks defined on the human skull [109].

of the optimal skull symmetry plane, which is defined as the plane perpendicular to the Frankfort-horizontal plane going through one of the medial landmarks. This definition of the symmetry plane tends to less underestimate the facial asymmetry when compared to the mid-sagittal plane [7]. See Figure 2.3 for an illustration of the anatomical planes of the skull.

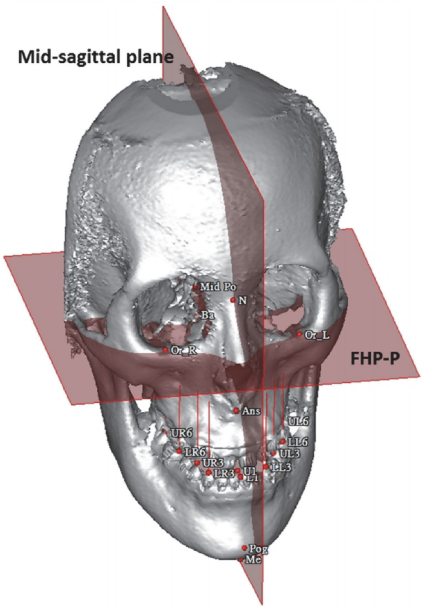

Figure 2.3: The mid-sagittal and the Frankfort-horizontal plane defined by the Orbital (Or) and the Porion (Po) landmarks. Note that three points are required to define the plane and therefore the mid-point between left and right porion is used in this example. Image adapted from Lin et al. [80]

2.2.2 *Cranial and Facial Bone Defects*

Skull defects are pathological changes in the physiological anatomy of the skull described in the previous section. *Congenital skull defects* are also known as birth defects. Congenital defects include both facial deformities, such as cleft lip, cleft palate, and cleft jaw, and cranial deformities, such as craniosynostosis [1, 3]. Although CMF surgery deals with many more congenital skull defects, these are not in the scope of this short overview and not directly addressed in this thesis [4].

As opposed to congenital skull defects, *acquired skull defects* are caused by factors not directly related to the skull developmental processes. These include mostly head trauma, infection, and tumors [114]. In the remainder of this work, the term *skull defect* will refer specifically to the acquired skull defects. Clinical treatment of acquired skull defects usually involves reconstructing the original skull shape in order to restore the aesthetic, masticatory, and protective functions of the skull. In the case of the neurocranium, restoring the skull shape (referred to as *cranioplasty*) can also lead to alleviating neurological impairment [2]. Two approaches can be taken to incorporating material into the skull defect to restore its shape: using the original part of the patient's bone (also referred to as autologous bone implant or bone graft) or constructing an implant from synthetic material. Appropriate use of anatomical models further improves the clinical outcome and reduces operat-

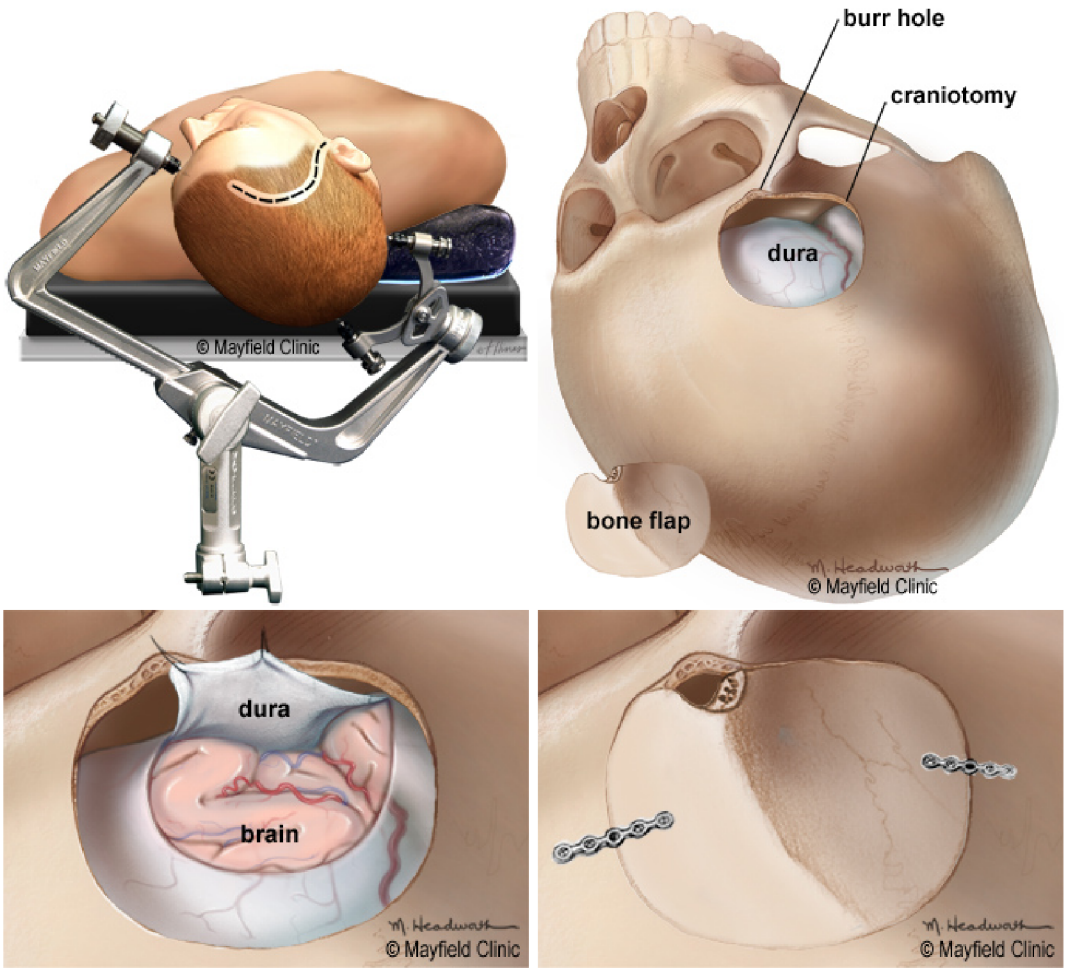

Figure 2.4: Illustration of individual steps of craniotomy procedure [117].

ing time of both of these methods, but the effect is most evident in the second case [121].

Using the autologous bone is straightforward in some cases of minor traumatic fractures, where the surgeon only needs to slightly reshape and reposition the bone before reattaching it to the rest of the skull using titanium mini plates or clamps [60]. Another procedure where the use of the autologous bone is common is *a craniotomy*, during which a part of the skull is temporarily removed to access the tissues underneath. The clinical implications for craniotomy include treating brain hemorrhage, infection of intracranial tissues, or resection of a tumor. The craniotomy procedure is illustrated in Figure 2.4. Similarly, during *decompressive craniectomy*, part of the skull is removed in order to decrease intracranial pressure caused by brain swelling due to head trauma or a stroke. The bone graft then needs to be stored for long enough for the swelling to subside, which may take weeks or months [71].

Although the autologous bone may seem like the perfect implant, it can often become infected or degraded either during the bone preservation stage or during

patient rehabilitation for various reasons [122, 26]. The incidence of these failure cases is as high as 10% of all cranioplasties [123]. Furthermore, the original bone can be unusable straight away, for example in cases where the bone is resected to remove a bone tumor, which also applies to maxillar and mandibular bones.

In cases where the use of autologous bone is not possible, the use of alloplastic materials is implied to replace the removed part of the bone [68]. Multiple different material options exist and the final choice is usually down to the personal preference of the surgeon. Some materials such as titanium mesh or polyether-ether-ketone (PEEK) can be manufactured pre-operatively based on a patient scan using rapid prototyping. On the other hand, bone cement based on polymerized methyl-methacrylate (PMMA) usually needs to be manually fitted to the patient's skull, which may increase the surgery time if done intra-operatively. They can, however, be loaded with a bioactive component such as antibiotics [61]. Most importantly, the risk of the implant having to be eventually removed due to complications is significantly lower when using these alloplastic materials and they are therefore often chosen in favor of the autologous bone not only in revision cranioplasties but in primary cranioplasties as well [61]. That is, however, conditioned on the fact that the implants are modeled to fit perfectly to the patient's skull.

2.3 PRODUCING PRECISE ANATOMICAL MODELS FOR SKULL SURGERY

Patient-specific anatomical models can be used in several different ways in the context of reconstructive skull surgery planning.

- Defective skull model for planning of the intervention
- Directly 3D printed implant (titanium, PEEK)
- Negative implant-shaped block used for molding the implant (PMMA)

The process of producing any of the mentioned types of physical anatomical models of the skull can be broken down into four steps: Data acquisition, image processing, virtual reconstruction, and manufacturing. The modern manufacturing methods achieve satisfactory accuracy for use in skull reconstruction and the majority of imprecision of the results comes from the three preceding steps [125].

2.3.1 *Data Acquisition*

Today's radiologists have a broad range of different medical imaging modalities at hand, from low-end ultrasound systems to magnetic resonance capable of displaying structural as well as functional tissue properties and positron-emission

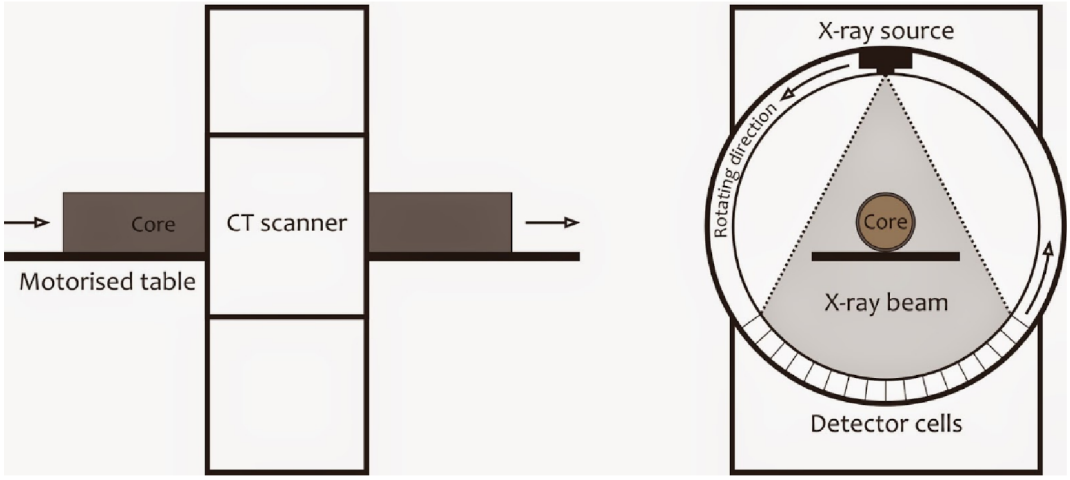

Figure 2.5: Scheme of a standard CT scanning system with a side view (left) and a cross-view (right) [16].

tomography used for the identification of areas with heightened metabolism. In the case of hard tissue surgical planning, CT is usually the modality of choice because of its unparalleled hard tissue contrast, short acquisition times, and no clinical contraindications except for radiation dose.

CT systems are, in the physical sense, similar to the X-ray radiography in that they measure the attenuation of X-ray photon beam on the way from the radiation source to the scintillation detectors through the scanning area in which the patient is positioned. The beam attenuation on the detector is then calibrated using the known attenuation coefficient of water which results in the Hounsfield Unit HU scale measurement. The construction of most current CT scanners is such that several transversal slices are simultaneously acquired by measuring several fan-like projections of the patient body under different angles as shown in Figure 2.5. Alternatively, cone beam computed tomography (CBCT) is often used in CMF surgery because it significantly decreases the patient radiation dose and acquisition time. Iterative mathematical reconstruction or filtered back-projection is then employed to transform the set of these projections into 3D data volume. The intensity of each voxel (the volumetric equivalent of a pixel) in this data volume can be interpreted as HU measurement in the area of the voxel. The acquired patient CT data are usually stored in the Digital Imaging and Communications in Medicine (DICOM) format which comprises the header containing the meta-information and the image intensity data stored in a voxel grid.

Both the scanner construction and the reconstruction method can lead to specific artifacts encountered exclusively in this modality. While some common artifacts such as motion blur, noise, and ring artifact can be avoided during the acquisition, others cannot, especially in the case of present metal artifacts. To truly reconstruct

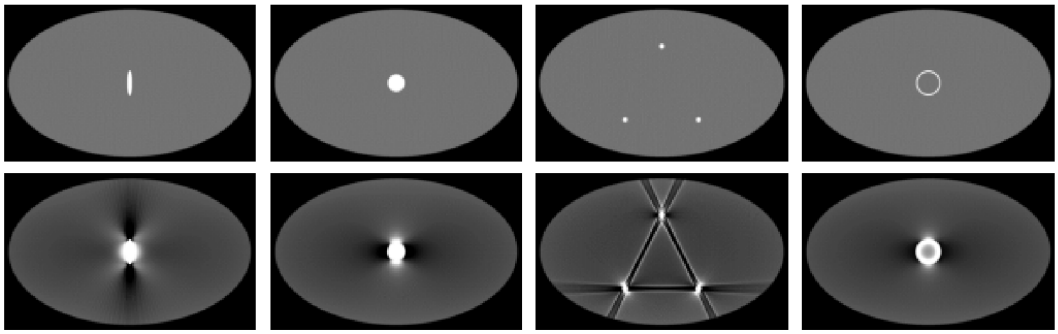

Figure 2.6: Examples of synthetic CT data without (upper row) and with beam hardening effect (bottom row). Scatter produces a similar type of artifact [16].

the volume, a theoretically infinite amount of the projections in different angles would be required. The above-mentioned methods used for approximate reconstruction from a limited number of projections are unable to correctly reconstruct the image in the case of small extremely high-intensity objects (see Figure 2.6), typically metal objects, due to beam scatter and hardening phenomena. This is particularly troublesome in surgical planning as the target tissue often contains metal implants such as fixation materials or dental implants [16]. Partial volume artifact is another source of imprecision in the CT data. In places where multiple tissues with different densities occupy the space of a single voxel, the resulting attenuation is proportional to their quantities. In the case of very thin bone structures, such as skull orbits, this means that the bone has very little definition in the data. Both of these issues cause complications in the subsequent step of the anatomical model production pipeline [53].

2.3.2 *Image Processing and Segmentation*

Before the CT data is converted to the standard tessellation language (STL) format used for manipulating the skull geometry and manufacturing, image processing methods are used for tasks such as re-sampling and cropping the data, mid-sagittal plane detection, and skull segmentation.

Correct skull segmentation, in other words correctly selecting the voxels that belong to the skull, is critical for final anatomical model precision [39]. Errors in segmentation lead to complications during the surgery, especially if the model is used for manufacturing of PSIs [51]. Because bone tissue has high image intensity in the CT image data due to its calcium content, HU value thresholding is often used for its segmentation from other tissues. However, this is not a sufficient criterion to obtain a segmentation of the skull acceptable for manufacturing anatomical models for several reasons, some of which are illustrated in Figure 2.7.

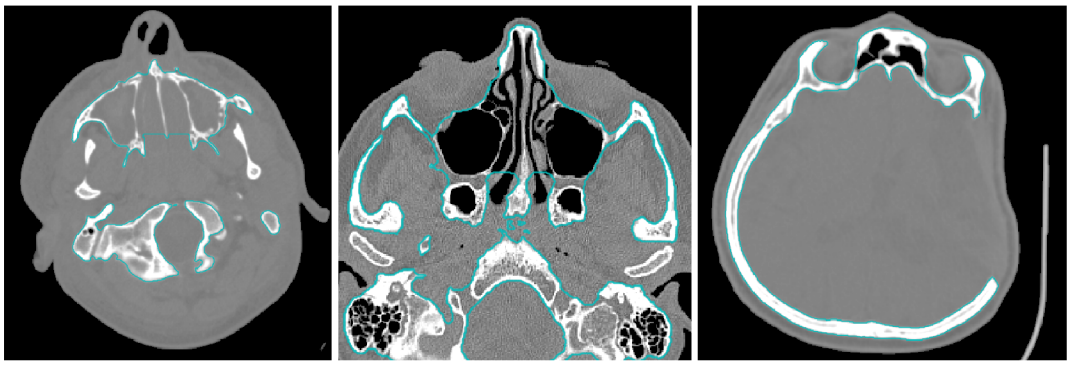

Figure 2.7: Examples of slices through a head CT, with expert-made segmentation in cyan. From left to right: the maxillary sinus, thin bone in orbits, skull defect.

First, the paranasal sinuses (i. e. the cavities in frontal, maxillary, and ethmoidal bones, see Figure 2.1) are often included in the skull segmentation to facilitate better stability of the manufactured model [120]. Second, the bone tissue forming sinusoidal and orbital walls is too thin to attain high intensity values due to partial volume artifact and CT slice thickness. Finally, the noise in form of adjacent tissues, such as vertebrae and bone fragments, instrumentation, such as fixation material and CT scanner table, or artifacts caused by metal implants need to be removed. Several software packages approved for medical use such as Mimics (Materialise®, Belgium), and Osirix® MD (Pixmeo, Switzerland) therefore come equipped with global and local thresholding tools in combination with manual post-processing tools. The manual post-processing step that addresses these issues is, however, very time-consuming [125]. A review of skull segmentation methods can be found in Section 3.1.

2.3.3 *Virtual Skull Reconstruction*

After the segmentation, the binary segmented mask is converted to the polygonal STL format using triangulation (typically marching cubes or Delaunay with mesh smoothing) [125]. Virtual reconstruction of the skull is the process of designing the shape that completes the defective skull using computer-assisted design (CAD) software. This is a different process from the final virtual implant design, which usually takes place after the reconstruction.

The reconstruction is usually done by using the mirrored healthy side of the skull or a similar case from a patient database as a template, and then manually fine-tuning the shape to fit seamlessly to the defective skull. However, a considerable amount of manual work is required to reach enough precision along the entire defect border while preserving the physiological shape and symmetry of the skull and that, as in the case of the image segmentation step, reduces the cost-

effectiveness of using the final anatomical model. A review of skull reconstruction methods can be found in Section 3.2.

The goal of this thesis is to propose a robust and fully automatic skull segmentation and reconstruction method that will lead to standardization and increased efficiency of clinical workflow of cranioplasty planning.

3

LITERATURE REVIEW OF SKULL SEGMENTATION AND RECONSTRUCTION METHODS

Eijnatten et al. summarize the cranial PSI design workflow by dividing parameters affecting the resulting quality into several categories as illustrated in Figure 3.1 [125]. This thesis directly addresses two main bottlenecks in the time effectiveness of this workflow: the skull segmentation and the computer-aided design of the implant shape, both of which require a substantial amount of the operating expert's time. This chapter provides an overview of existing semi-automatic and automatic segmentation and virtual reconstruction methods applied to the skull and serves as a background and reference for the rest of this thesis.

3.1 SKULL CT SEGMENTATION

Skull segmentation quality is most often measured either using average[1] surface distance or volumetric overlaps, such as the Dice coefficient. Authors of a meta-study that reviewed 36 CT bone segmentation articles in the context of anatomical models creation suggest reporting the average surface distance as a quantitative metric as it is easy to interpret. However, at the same time, the authors advise complementing these results with rendered 3D models of the segmented objects with color-coded surface errors because local characteristics play a crucial role in the manufacturing of precise anatomical models, and especially if they are being created for use in PSI design where the accuracy requirements are higher [125].

Thresholding the HU values is currently the most commonly used method of bone segmentation in CT data, including the segmentation of skull for anatomical model printing [125]. Although the precision of as much as 0.2 mm in terms of average surface distance can be achieved using thresholding as an initial segmentation estimate, the post-processing required to reach the final result can take several hours of manual work, which drastically increases the cost of the resulting anatomical model [107, 38]. Additionally, the optimal threshold HU value is subjective and depends on the CT scanner [99, 100, 124]. In some cases, global thresholding may provide sufficient approximation of the skull shape, even for subsequent virtual reconstruction steps. For example, the SkullFix dataset [72] that is used for the development of automatic skull reconstruction methods was created

[1] Average surface distance and mean surface distance are used interchangeably in this work.

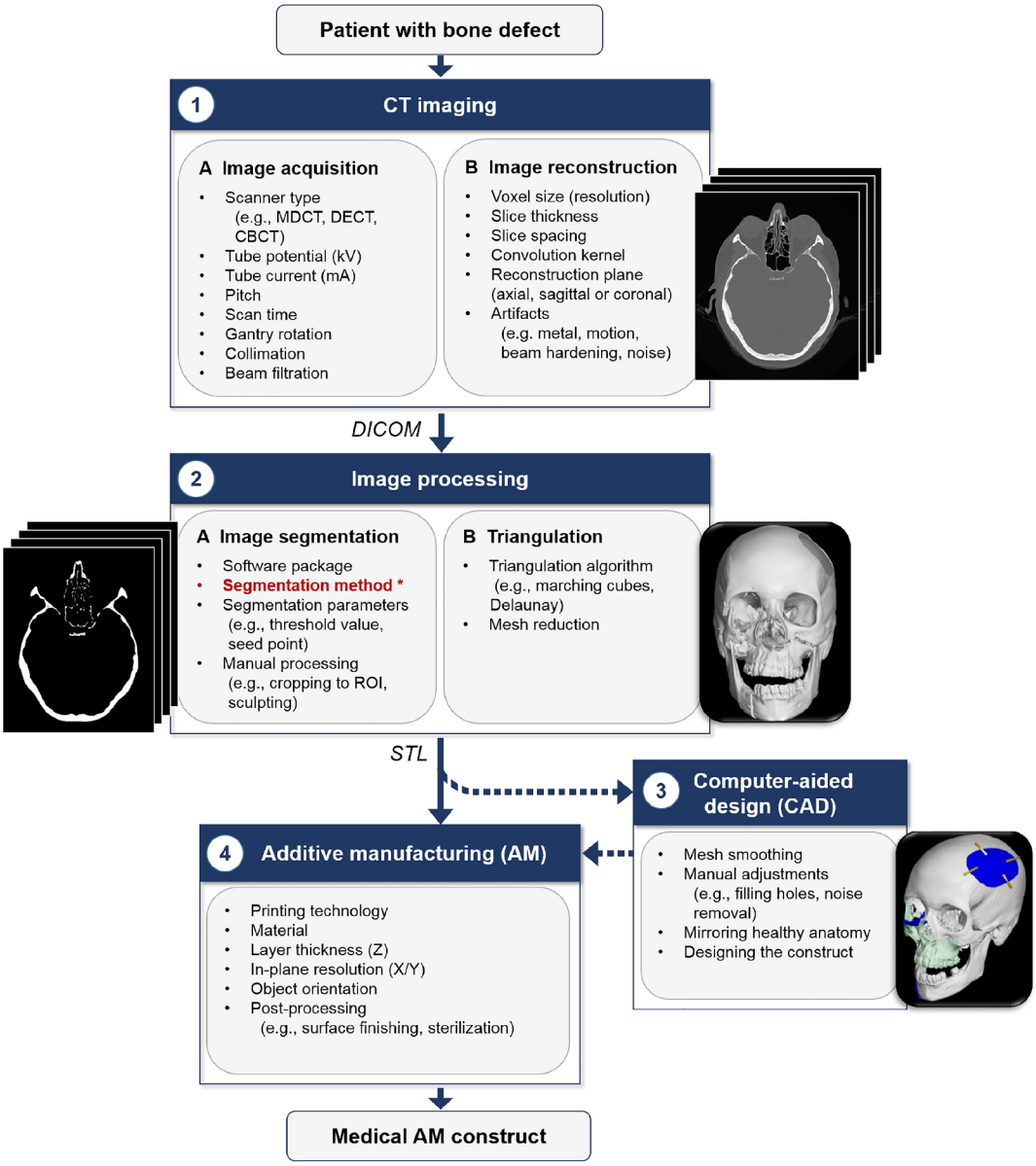

Figure 3.1: Overview of the parameters that can influence the accuracy of medical AM constructs. Adapted from [125]

using this method to provide reference shapes of the skulls The SkullFix dataset focuses on defects of parietal bone on top of the skull where thresholding usually provides sufficient segmentation accuracy [72]. However, it does not provide satisfactory results in other parts of the skull (see Figure 3.2 a).

The sinus and orbital walls are some of the most challenging parts of the skull to be segmented due to their sub-voxel thickness. Linares et al. [81] use a sheetness filter based on a Hessian matrix decomposition to locate the thin structures. Clustering of the volume into super-voxels is then performed and the final segmentation is obtained using interactive graph-cuts with user-defined seeds. The authors show that they can achieve high similarity to expert-made ground-truth

segmentation with a mean Dice coefficient of 0.915. An example of the result of this method can be seen in Figure 3.2 middle.

In the first attempt to use a deep learning approach for segmentation of skull from CT data, Minnema et al [95]. trained a 4 layers deep convolutional neural network (CNN) model that classifies the CT volume voxel by voxel in a sliding window manner. The authors used a dataset of 20 defective skulls with gold standard anatomical models that were created manually by an expert and leave-2-out cross-validation. The method reaches an average Dice coefficient of 0.92 and an average surface distance of 0.44 mm. The example renders of the skull segmentation results (see Figure 3.2 right) show that the method often misclassifies skull voxels as background in thinner parts of the skull, which may have been caused by the faulty conversion of the gold-truth mesh model to voxel grid before the training. These errors are typically not in the area of interest of the authors (i. e. close to the skull defect).

Metzger et al. [93] use an anatomical atlas template created from a manually segmented CT dataset of a healthy patient. The template is then registered to the target patient to provide segmentation using rigid and non-rigid registration based on mutual information criterion. This method achieves good stability even for the thin structures in orbital areas but it comes at the cost of lower accuracy in ethmoidal, sphenoidal, and temporal bones, where mean surface distances are over 0.5 mm. The authors also propose a semi-automatic post-processing thresholding step to improve accuracy in these larger bones but the outputs are not preferred by the clinicians in subjective evaluation because it, once again, introduces holes into thinner skull structures.

Statistical shape models (SSM) form an important group of CT bone segmentation algorithms that could be considered for anatomical bone model creation. SSMs model the general bone shape using mean and variation of corresponding vertex or landmark positions of ground truth bone models in the training set. Principal component analysis (PCA) or an alternative method reduces the dimensionality by modeling only the main eigenmodes of shape variability. Segmentation of new CT data is then obtained by finding an optimal linear combination of these eigenmodes that fits the CT volume intensity values by some criteria [49]. SSMs have been successfully used for segmentation of multiple bones in the hip area such as the pelvis, hip joint, and femur [111, 29, 136]. In the context of skull segmentation, Chang et al. [21] used SSM to segment the anterior surface of the maxilla bone and reached an average surface distance of 0.2 mm. However, due to the dimensionality reduction and limited training datasets, SSMs are not well suited for segmentation of large structures with high anatomical variability [125]. This makes them unsuit-

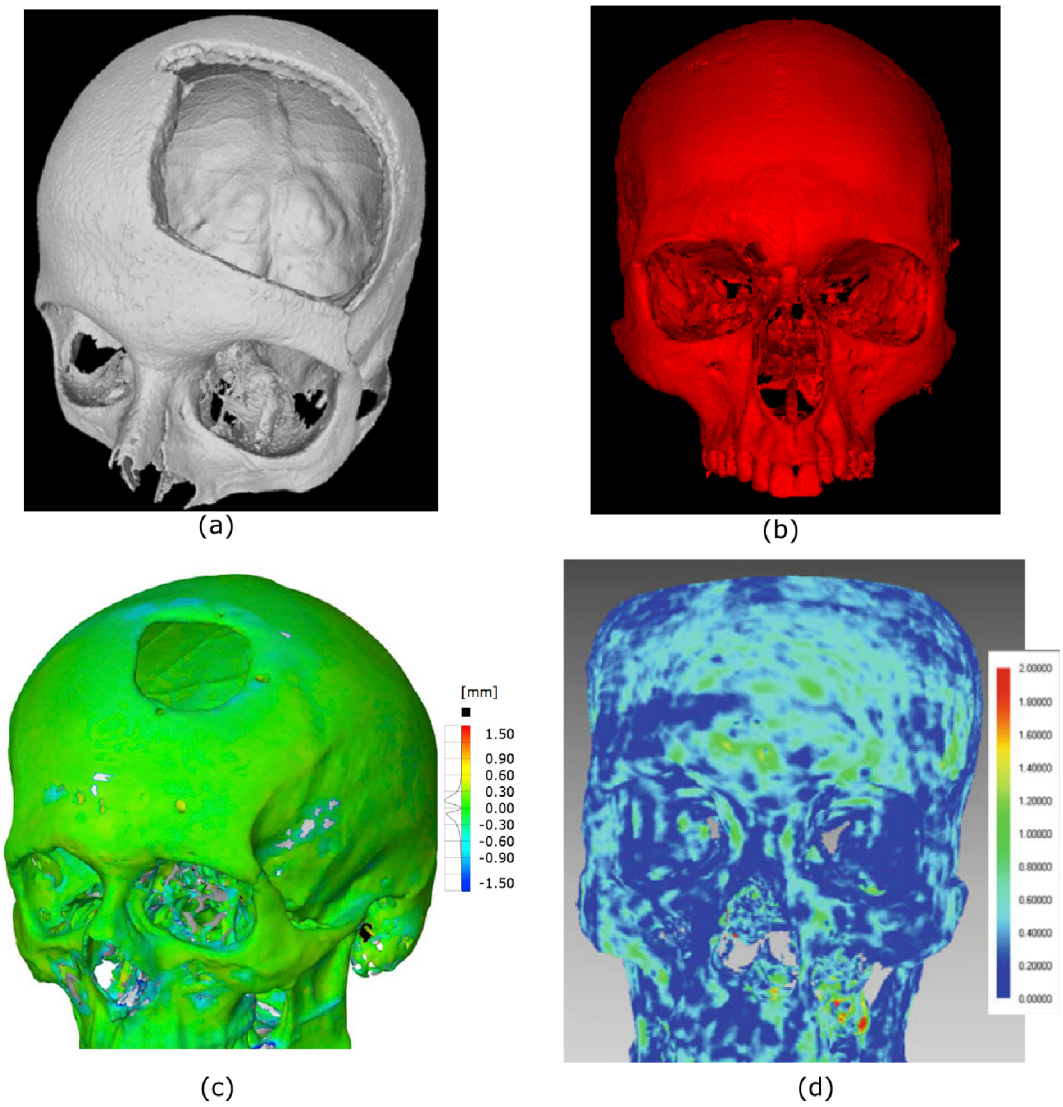

Figure 3.2: Skull models resulting from existing segmentation methods. Global thresholding [72] (a), Linares et al. [81] (b), Minnema et al. [95] (c), Metzger et al. [93] (d). Note that none of the methods achieves satisfactory accuracy of both the large bones and orbits.

able for automatic segmentation in skull PSI design workflow where high local precision is required.

A summary of the current skull segmentation methods can be found in Table 3.1. It is worth noting that the overall performance of segmentation methods is usually evaluated in the context of a specific target use. Although the methods mentioned in this section perform well in their respective areas of interest and even in the presence of defects, the segmentation performance is often limited in places where the skull bone is very thin, such as orbits and sinus walls. This is arguably due to the fact that the target use cases of these methods do not require optimal results in this area and the time cost of acquiring such ground truth data is simply not worth its value. However, this limits the range of applications of such segmentation

Table 3.1: Comparison of segmentation methods using average surface distance (ASD) and volumetric Dice coefficient (VDC). Thresholding is not included in the methods because the quantitative results depend on amount manual post-processing.

Authors	ASD [mm]	VDC [%]	Note
Metzger et al. [93]	≤ 0.6	-	Trade off between holes and generally lower accuracy.
Minnema et al. [95]	0.44	92.0	Occasional holes, often in orbital area.
Linares et al. [81]	-	91.5	Holes in orbital area, requires manual input.

outputs. For example, skull orbital floors and walls may be the target of virtual skull reconstruction [42], yet the segmentation outputs of the presented methods could not be used as a shape template for creating the PSI due to the segmentation containing holes. It is therefore desirable to develop methods able to consistently provide high-quality segmentation even in these challenging areas of the skull. Manual post-processing might still be required in many cases, but the time cost decreases with the higher quality of the initial automatic segmentation output.

3.2 SKULL SHAPE RECONSTRUCTION

Skull reconstruction can have multiple meanings in the literature, such as the process of repositioning bone fragments for fracture reduction [59] or creating a patient-specific anatomical model by means of segmentation and stacking of the CT slices [57]. In this thesis, skull reconstruction and skull segmentation are considered two different tasks. Reconstruction is understood as the task of mapping an input *shape* of the defective skull (represented as a binary voxel grid or a surface mesh model) to the corresponding reconstructed shape (a shape of complete skull, the missing patch, or the shape of final cranial implant). Although some methods in the literature are proposed to work directly with the HU values of the voxel grid, there is usually a thresholding step that discards them during preprocessing [20, 85, 41]. The soft tissues are therefore completely ignored during the skull reconstruction process. Most current methods focus on reconstructing the missing skull patch that perfectly completes the defective skull, which is subsequently used as a template for the final cranial implant design in a manual step. However some methods also automatically finalize the shape by drawing the de-

fect perimeter away from the skull defect border and accounting for the defect border shape [85].

Similar to the segmentation task, mean surface distance and Dice coefficient are often used to evaluate the quality of the reconstructed shape, with another popular choice being the percent of vertices or voxels with less than 1 mm from the target surface. However, using such metrics imposes some assumptions. These metrics compare the reconstructed shape to the original shape of the healthy skull that has to be available. Since pre-injury CT scan is rarely available, this limits the evaluation to artificially created skull defects. Second, it does not account for the shape variability of clinically viable reconstructions. For example, a bilateral defect implant shape can be slightly different from the original skull shape, as long as it is symmetric and fits well to the defective skull. The 3D model renders with color-coded surface distances are also a popular way to present the result. If reference shape is not available, as is usually the case for real defective patient skulls, the 3D rendered model can be used to visually evaluate anatomical feasibility, symmetry, and smoothness of fit. A considerable part of published works leaves the entire evaluation to this subjective visual assessment.

3.2.1 *Surface Interpolation Methods*

Methods based on surface interpolation aim to reconstruct holes by mathematically defining the surface over the defect area while maximizing the smoothness on the bone-implant interface. The first surface interpolation of a skull defect was proposed by Carr et al. [20], who proposed using radial basis function (RBF) for interpolation of 2D depth-map representing the skull surface around the defect. The algorithm starts by manually delineating the defect area. The depth map is then generated by casting parallel rays from a user-defined direction. The missing depth-map pixels are then interpolated using two forms of RBF; linear or thin-plate spline (TPS). TPS was found more appropriate because it enforces C1 continuity along the defect border, ensuring the smoothness of fit (see Figure 3.3).

The same authors showed that RBFs can be used to repair holes in 3D point clouds directly by interpolating a 3D signed distance function that represents the surface as a zero level set [19]. Zhou et al. [139] extend the RBF surface interpolation method by automatic detection of the defect border. Sing et al. [116] instead repair the defect using Beziér patch with the defect border additionally smoothed using a combination of Laplacian smoothing and mean curvature flow algorithm. Non-uniform rational basis spline (NURBS) representation is used by Chen et al. [22] to model the missing surface patch, which allows them to directly manufacture the corresponding titanium implant using a multi-point forming process.

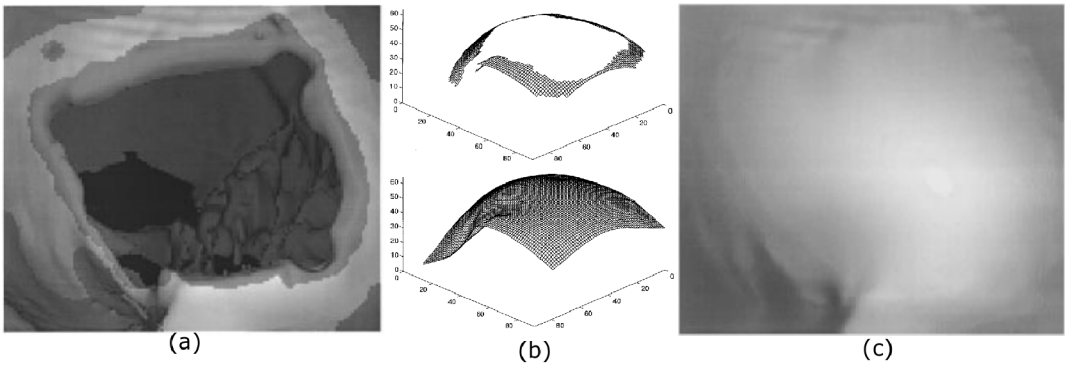

Figure 3.3: Illustration of skull surface interpolation using TPS. Defect area is selected on the input skull (a) and after interpolating the depth-map (b), the final surface can be merged with the skull (c). Adapted from Carr et al. [20]

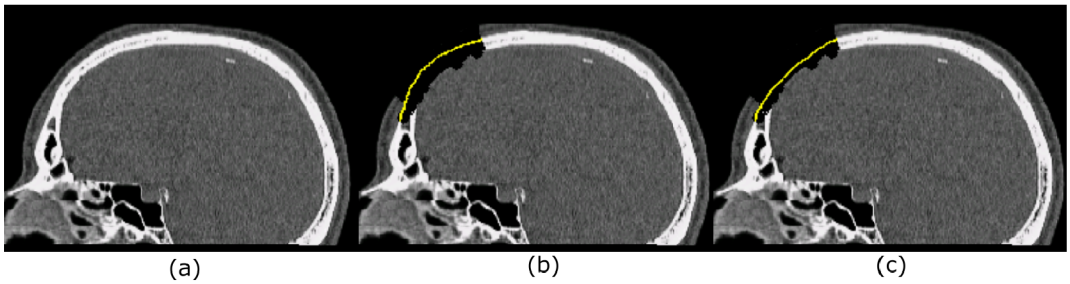

Figure 3.4: An example of reconstruction of an artificial defective skull (a) using anatomical constrains (b) and minimizing bending energy using TPS only (c). Anatomical constraints help to avoid undesired flattening of the surface close to the defect center. Adapted from Wu et al. [133]

The methods in this category are suitable for reconstructing defects in any convex part of the skull, including bilateral defects, and can be evaluated at arbitrary resolution. However, they are only constrained by the skull surface in close proximity to the defect border which leads to flattening in central regions of the reconstructed surfaces with large areas (see Figure 3.4) [20, 116]. Although this may be a desirable property for some use cases such as manufacturing titanium meshes with hydraulic presses, it ultimately damages the natural skull shape [20, 133]. They are also naturally unable to reconstruct complex shapes such as orbits. Despite these limitations, some elements of these methods, such as using TPS interpolation for the final fitting of the reconstructed surface patch, were successfully used in more recent hybrid skull reconstruction methods discussed in the following sections.

3.2.2 *Methods Based on Anatomical Templates*

Beyond surface smoothness, further anatomical constraints can be incorporated into skull reconstruction methods by using anatomical templates. Prior knowledge

about the missing skull shape can come in form of a similar patient skull found in a database of healthy skull scans or, in case of unilateral defects, from mirroring the opposing healthy part of the target skull. The methods in this category typically include two main steps; identifying the optimal template (by searching the database or by identifying the optimal symmetry plane for mirroring) and fitting the template to the target skull.

Repairing unilateral skull defects by mirroring is the most documented application of computer-assisted surgery in cranioplasty in the current literature [54, 47]. In the simplest case, the mirroring plane is identified manually by specifying three or more points on the mid-sagittal plane [47], although, as mentioned in Section 2.2.1, the optimal symmetry plane can actually differ from the mid-sagittal plane due to skull asymmetry. Recent methods that automate the symmetry plane detection step instead rely on different approaches such as template registration, feature-based point matching between the opposing sides of the skull, or CNN-based rotation estimation [27, 133, 77, 31]. After the healthy side of the skull is mirrored, the final fitting to the patient's defect is often performed manually [70, 55]. To reduce the time cost, the fitting can also be done automatically. Chen et al. [23] guide the final fitting using contours defined manually in two axial slices and report over 80% reduced time cost as opposed to fully manual fitting. Gall et al. [43] report similar results using Laplacian smoothing of the skull-implant interface after mirroring (see Figure 3.5). Direct comparison of manual fitting and automatic fitting using ANTs diffeomorphic registration [11] was performed by Wagner et al [129]. The authors found that the resulting reconstruction of an artificial defect of zygomatic bone overlapped with the original shape with a Dice coefficient of 0.87 for both manual and automatic reconstruction. However, visual inspection of the automatic fitting result reveals visible asymmetry.

In cases of bilateral defects where mirroring does not provide sufficient anatomical information, anatomical templates can be used instead. For example, Gall et al. [43] use a simple sphere as an anatomical shape prior. For robustness, templates acquired from healthy skulls are more appropriate, but they require a suitable template to be identified because the anatomical variability of human skull shape is substantial. The most suitable template can be chosen from a database manually by visual inspection or by using geometric or demographic descriptors [133]. Simply using the average skull shape computed from a database of multiple patients is another possibility [34]. However, more recent methods take advantage of an automatically identified optimal template chosen from a database of healthy skulls using Procrustes analysis [85] or SSMs [112, 41, 42].

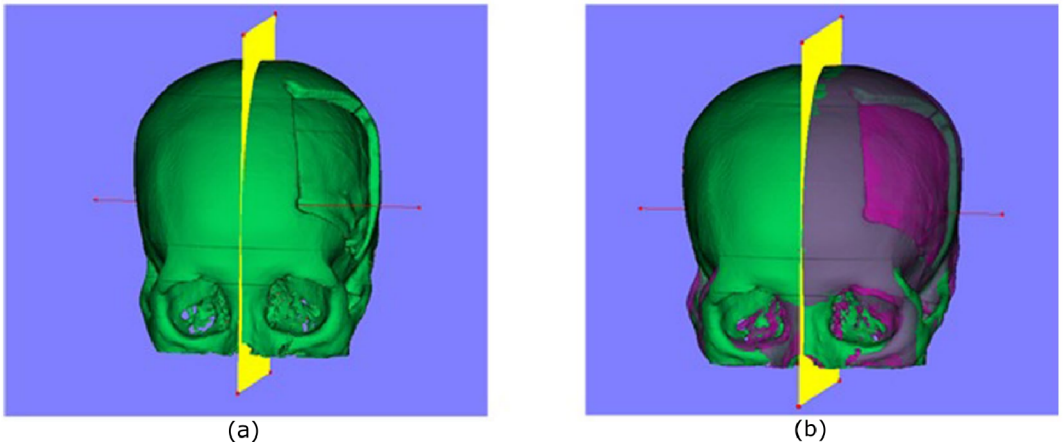

Figure 3.5: Illustration of mirroring-based reconstruction. The optimal symmetry plane has to be identified (a) and the mirrored object is fitted to the defect (b). Image adapted from Gall et al. [43]

3.2.3 *Methods Based on Statistical Shape Modelling*

Beyond segmentation, SSMs have been successfully used to study skull variability and symmetry, quality assurance and evaluation of skull fracture reduction outcome, or evaluation of craniosynostosis in pediatric patients. In terms of skull reconstruction, SSMs can serve as a straightforward way of finding an optimal skull shape template for reconstruction, provided that a large enough database of healthy skulls is available.

Semper-Hogg et al. [112] use a database of 175 healthy patients to construct SSM for skull mid-face area reconstruction. They begin by rigidly registering each skull to a single skull template using 4 manually defined anatomical landmarks. Vertex correspondences are then found by diffeomorphic registration of the template to each skull in the database. The authors use artificial defects with a diameter of 3 cm placed in zygomatic bone for evaluation. The defective skull is first rigidly aligned to the average shape and the SSM is then modified to only include patients with the similar placement of the 4 initial landmarks, resulting in a posterior SSM. Finally, the defective skull is registered to the average posterior shape using the non-elastic iterative closest point (ICP) algorithm and projected into the model PCA space where the best fitting SSM instance is found. The authors report the average surface distance of the result from the original shape of 0.85 mm, outperforming an alternative manual mirroring method.

The same authors later evaluated the same method on artificial bilateral defects of orbital floors and naso-orbital-ethmoid area, reaching average surface error of 0.75 and 0.81 mm, respectively [42]. However, visual inspection shows that despite the high similarity of the result to the original shape, there are noticeable errors

of 1-2 mm along some parts of the defect border, which could be attributed to the fact that the quality of SSM fit is limited by the variability of input healthy skull database and its loss during the PCA projection.

3.2.4 *Hybrid Methods*

Methods in this category use a combination of two or more previously described approaches, which allows them to address some of their mentioned limitations. They usually comprise of identifying and registering a correct anatomical template and fitting it to the defective skull using surface interpolation methods.

A straightforward combination of automatic mirroring and surface interpolation was proposed by Marzola et al. [88]. The method uses automatic symmetry plane detection [9] to mirror the healthy part of the skull and provide anatomical constraints in the defect area. Part of the mirrored points closest to the defect border is then discarded and TPS depth-map interpolation method [20] is used to connect the mirrored points and the defective skull. Although authors focus on uni-lateral defects, the method also demonstrated the ability to reconstruct defects reaching slightly into both lateral sides of the skull.

Dean at al. [34] start the semi-automatic reconstruction process by manually setting several points along the defect border. These points are then connected by using the Dijkstra algorithm that finds the optimal path connecting these points while avoiding parts of the surface with a high principal curvature value. This path is considered the defect border where the implant should connect to the defective skull. A template in form of the mirrored skull or average skull shape is then fitted to the defect border in two stages. In the first stage, several manually input correspondences are used as an input for TPS warp algorithm to initially register the template to the target skull and to find correspondences of the defect border points [17]. Then, TPS warp is used again, this time using the correspondences on the defect border, ensuring a better fit.

Similarly, Wu et al [133]. also begin the reconstruction by using paired point matching to find correspondences between the template and target skull. Next, the authors transform both shapes into spherical coordinates with origin in the mid-point between the two porion landmarks (i. e. the theoretical center of the Frankfort-horizontal plane). Then, scaling factors of the radial distance coordinates between the template and the target at user-defined defect border points are interpolated and used for the deformation of the template area inside the defect. The authors report 81% of the surface having an error of 0 or 1 voxel, but the resolution of the CT data is not disclosed.

Marreiros et al. [85] use Procrustes analysis to identify the optimal skull template from a database using 17 manually defined anatomical landmarks and 700 additional pseudo-landmarks. To improve the reproducibility of manual annotations, all landmarks are first slid along the skull surfaces to minimize bending energy. Target skull pseudo-landmarks corresponding to the defect area are then selected manually by the user and the corresponding pseudo-landmarks of the template are relaxed to minimize the bending energy again. Finally, the pseudo-landmarks are interpolated using depth-map TPS interpolation similarly to Carr et al. [20]. The method also estimates the inner surface and shape of the implant along the defect border automatically, although this aspect of the method is not evaluated quantitatively. The method reaches high accuracy of 0.4 mm on large artificial defects, including bilateral defects.

A similar method proposed by Fuessinger et al. [41] instead uses posterior SSM to find the optimal skull template using 6 anatomical landmarks and 2900 pseudo-landmarks. The pseudo-landmarks corresponding to the defect area are detected automatically based on their distance from corresponding template landmarks, surface normal difference, and surface curvature. A precise fit is then achieved by TPS warping the template using the point correspondences along the defect border. The authors report good accuracy of 0.47 and 0.5 mm for unilateral and bilateral artificial spherical defects, respectively.

The hybrid approaches generally reach the lowest surface distances from the original skull shapes, are suitable for large defects, and have been shown to provide a good fit along the defect border. Their downside is that they are often dependent on an exact definition of defect border and that they make several assumptions about the shape of the target anatomy (i. e. mostly spherical or elliptical shape of neurocranium), which makes them generally unsuitable for the reconstruction of orbits and zygomatic bones (see Figure 3.6).

3.2.5 *Deep Learning-Based Methods*

Although CNNs have overtaken most conventional medical image analysis domains such as classification, detection, and segmentation, it only made its way into anatomical shape completion research in recent years. The most common formulation of the task is mapping the input binary image with defective shape to the binary image of either the reconstructed or the missing shape. This allows for use of network architectures and loss functions well known from segmentation literature.

Using deep learning for general 3D shape reconstruction (also referred to as shape inpainting or shape completion) is a well-studied research topic in the lit-

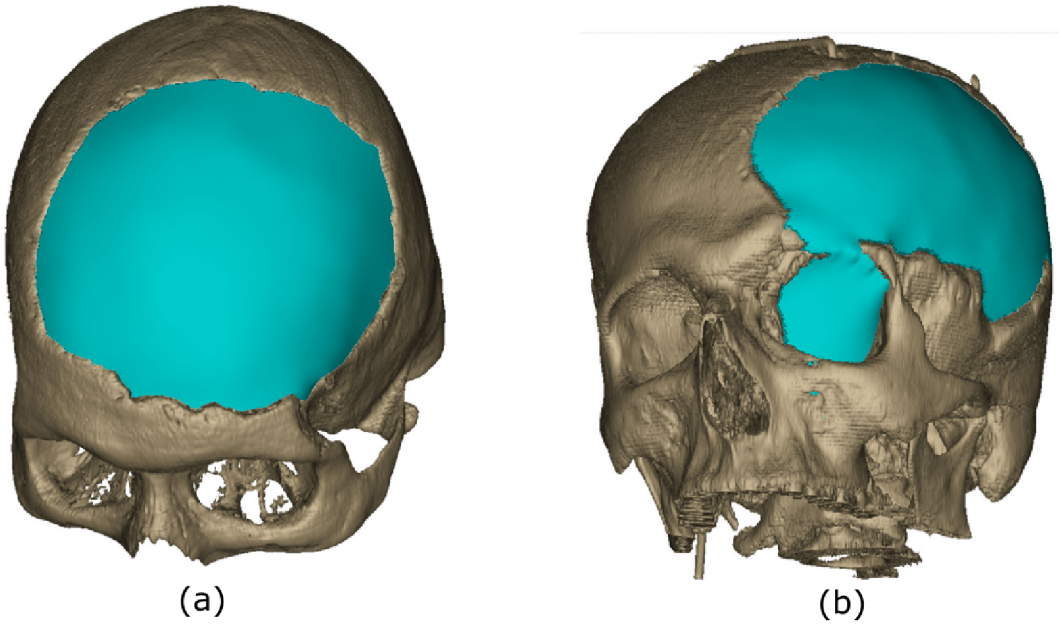

Figure 3.6: Examples of skulls reconstructed using the approach proposed by Marreiros et al. [85]. High-quality reconstructions can be achieved even for bilateral defects (a) but more complex anatomy of orbits causes failures and "leaking" effect (b).

erature outside of the medical domain, although the target application is usually filling incomplete sensor data such as LiDAR. The basic approach is to represent the incomplete input shape as a binary voxel grid and train a 3D CNN with a deterministic denoising auto-encoder architecture to output the completed binary shape [113].

Since the computational cost of such methods scales cubically with the target working resolution, they quickly become impractical. For example, for application in skull reconstruction, volumes of up to 512^3 voxels need to be processed in order to reach the desired precision. To correctly account for large defects, the output neurons should also have their receptive field spanning the whole input volume, resulting in a CNN so deep that it would not be possible to implement due to GPU memory limitations. To tackle the issue of memory requirements, several ways to make the training and inference more efficient have been studied. The first group of methods exploits different data representations such as graphs or point clouds [118, 131]. The second group uses the 3D CNN only for coarse shape estimation, refining the result in the post-processing step. Some authors use database lookup to identify a similar high-resolution object and then find the correspondences with the incomplete input shape or use it as the output directly [33, 82]. A second CNN, based on recurrent 2D convolutions has also been used to increase the output resolution [130].

An orthogonal research direction in this area led to substituting the purely discriminative CNN models with generative models such as generative adversarial

networks (GANs) or variational auto-encoders [130, 82, 118], suggesting that the shape completion task actually has multiple correct solutions conditioned on a single input. This issue of one-to-many mapping has also been raised by authors in the context of anatomical shape reconstruction [5, 6] where it is used to take the acceptable inter-expert variability of the resulting shape into account. However, the argument that the variability of the output should be enforced at the cost of precision measured against the original shape is in direct conflict with the current literature on skull reconstruction where the original shape is usually considered to be ground truth.

The first documented use of CNNs for the reconstruction of incomplete skull shape images was reported by Morais et al. in 2019 [98], who used a database of 1113 segmented MR scans of the skull with artificial defects placed in the neurocranium to train a denoising auto-encoder. The model only uses two convolutional layers and one fully connected layer in the encoder and two deconvolutional layers in the decoder parts of the architecture. The experiments were limited to infer shapes of maximum 60^3 voxels due to memory restrictions, but the approach was proven to be effective in reconstructing the defective skulls. Voxel reconstruction error of 3.2% is reported by the authors, demonstrating the feasibility of using CNN for 3D skull shape reconstruction.

Matzkin et al. [91] use a similar approach to train several different models on segmented and CT datasets with the resolution of 2 mm per voxel. Each skull is first registered to an anatomical atlas. They use a reconstruction method based on direct PCA projection of the voxel grid as a baseline method and compare it to simple auto-encoder and U-net [108] models. They also compare results of direct missing shape estimation and of full skull shape reconstruction. The direct missing shape estimation using U-net architecture is shown to yield the best results with Dice coefficient 0.8, showing the importance of skip-connections in the U-net architecture. The authors also report similar results for real craniectomy cases, but only several cases of post-operative CT scans without any further description are used. The same authors later also attempted to use anatomical atlas as a shape prior by using it as an additional input CNN channel [90]. Although it was shown to improve performance on out-of-distribution cases, the shape prior decreased the quality of the reconstruction results on average.

Li et al. [73] created a public dataset of defective skull shapes by thresholding the CQ500 head CT database and injecting artificial defects (referred to later as the SkullFix dataset [64]). The dataset was then used to organize the first AutoImplant challenge hosted by the MICCAI 2020 conference. The goal of the challenge was to reconstruct defects with shapes typical for craniectomy, positioned in the parietal

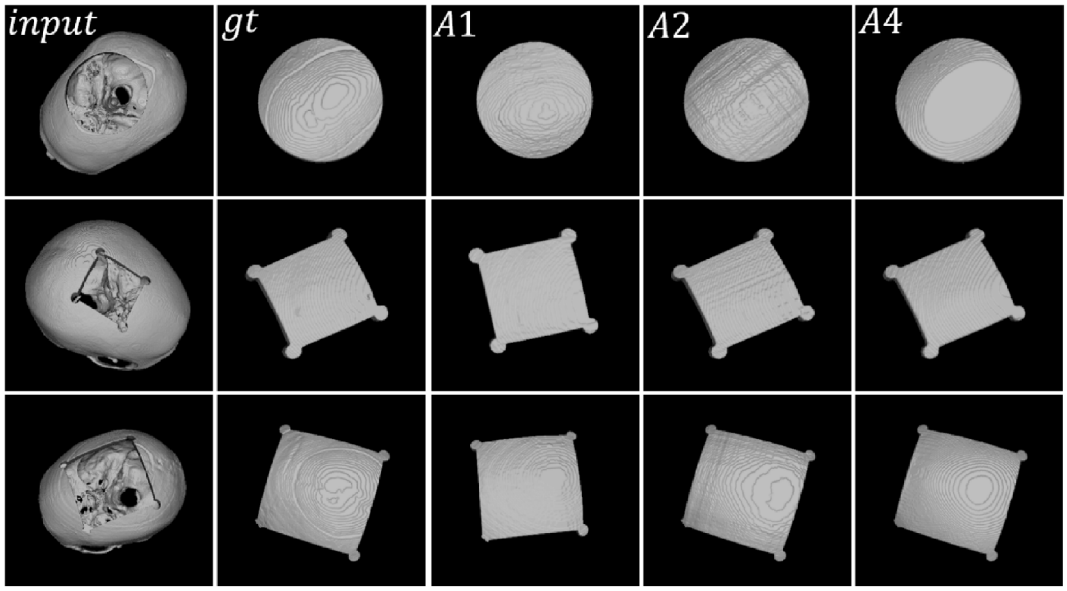

Figure 3.7: Examples of skull defect reconstructions from the AutoImplant challenge participants. Results by Pimentel et al. [104] (A1), Shi et al. [115] (A2) and Ellis et al. [37] (A4). Image adapted from Li et al. [76]. Criss-cross pattern caused by combining results from different slices in A2 and over-smoothing due to resampling to lower resolution in A4 can be noticed.

bone. Results were evaluated using the average of Dice coefficient and Hausdorff distance over 100 regular testing cases and 10 additional out-of-distribution cases.

The baseline approach proposed by the authors of the challenge [74] is based on two auto-encoders. The first auto-encoder reconstructs the missing skull patch on a low resolution. This first result is then used to compute the 3D bounding box, in which the second, high-resolution model estimates the final output. The authors reach a Dice coefficient of 0.856 on the regular test dataset. A total of 11 teams participated in the challenge and all of the methods were based on CNNs. Among the more interesting methods, Pimentel et al. [104] used a SSM of healthy skulls for prediction of the missing skull patch, which was then refined using 2D GAN and report results of 0.917 Dice coefficient. Shi et al. [115] showed that very competitive results in terms of Dice coefficient can be achieved by simply reconstructing 2D sagittal and frontal slices of the skull and then averaging the results, reaching a Dice coefficient of 0.931. The winning contribution by the team of Ellis et al. [37] used a very large 3D U-net model in combination with a novel data augmentation method based on mutual elastic registration of each pair of training data, resulting in the highest achieved Dice coefficient of 0.942, despite operating on less than half the resolution of the original ground-truth data.

Methods based on deep learning promise the ability to reconstruct any part of the skull present in the training dataset, including the challenging facial area, and the ability to seamlessly connect the reconstructed shape to any part of the

defect border. However, it is necessary to consider these requirements when designing the training process and data. The current methods were shown to achieve great performance when reconstructing artificial defects created as spheres or axis-aligned regular shapes in neurocranium (see Figure 3.7), but the ability to connect high-resolution reconstructed shapes to more complex border defect shapes in the facial area is yet to be demonstrated. This requires the development of appropriate datasets containing a balanced amount of defects with various shapes and characteristics. That, in turn, requires the datasets to be created from skulls without undesired holes and artifacts in areas of thin bone. Additionally, the existing approaches do not include any mechanism for taking skull symmetry into account in form of prior information, relying on the model having enough capacity to learn it implicitly instead.

A summary of the current literature is shown in Table 3.2. It shows that despite the recent advances in the research of automatic virtual skull reconstruction, no automatic method so far demonstrated the ability to provide high-quality reconstruction estimates of neurocranium as well as facial and orbital defects of the skull while ensuring the precise fit of the implant along the defect border. Also, most of the approaches are validated either using artificial defects only or using very few real cranioplasty cases and little attention is given to the fully automatic use case of direct intra-operative implant design and manufacturing. Finally, no work addressed the challenge of generating multiple missing shapes hypothesis for single defective skull as of time of writing this thesis.

Table 3.2: Skull reconstruction accuracy in terms of ASD [mm] and VDC [%]. Methods that do not report either of these metrics and methods that require large amount of manual fine-tuning are omitted. Asterisk denotes methods published concurrently with this thesis.

Authors	ASD	VDC	Note
Methods based on surface interpolation			
Carr et al. [20]	≤ 0.56	-	Limited to smaller defects of neurocranium. Evaluated on artificial defects.
Methods based on anatomical templates			
Wagner et al. [129]	-	87.0	Limited to unilateral defects. Evaluated on artificial unilateral facial defects.
Methods based on SSMs			
Semper-Hogg et al. [112]	0.85	-	Implant fit not guaranteed. Evaluated on artificial unilateral facial defects.
Fuessinger et al. [42]	0.81	-	Implant fit not guaranteed. Evaluated on artificial bilateral facial defects.
Hybrid methods			
Marreiros et al. [85]	0.40	-	Limited to neurocranium. Evaluated on large artificial bilateral and unilateral defects.
Fuessinger et al. [41]	0.50	-	Limited to neurocranium. Evaluated on artificial bilateral and unilateral defects.
Methods based on deep learning			
* Matzkin et al. [91]	-	80.0	Evaluated on artificial and real defects of neurocranium.
* Li et al. [72]	-	85.6	Evaluated on artifical bilateral defects of neurocranium.
* Pimentel et al. [104]	-	91.7	2D post-processing. Evaluated on artifical bilateral defects of neurocranium.
* Shi et al. [115]	-	93.1	2D slices reconstruction. Evaluated on artifical bilateral defects of neurocranium.
* Ellis et al. [37]	-	94.2	Limited resolution. Evaluated on artifical bilateral defects of neurocranium.

Part II

CT DATA PROCESSING FOR CRANIOPLASTY

Work presented in this part of the thesis addresses processing the patient CT data into shape representation of the patient's skull. This has two main applications: first, automatization of data processing steps of cranioplasty anatomical model design workflow, and second, efficient creation of a public dataset of defective skulls and corresponding missing skull patches that can be used for development and evaluation of virtual skull reconstruction methods.

CT scans of patients indicated for cranioplasty were provided for this work by TESCAN Medical company and included gold standard segmentation as well as reconstruction done by experts with extensive experience in PSI design. One of the drawbacks of the current skull reconstruction literature is that authors usually report results of several in-house skull samples which often cannot be published due to the nature of patient data and this limits the reproducibility of the proposed approaches. Because the same applies to the in-house patient data available for this work, we used the provided data to develop an automatic processing method that we then applied to the public CQ500 skull database [28] from which a public synthetic cranioplasty dataset was finally created.

Based on observations of the provided real cranioplasty patient cases, we concluded that the skull shape representation should have the following properties. First, the skulls should be *aligned to the natural head position* defined by the Frankfort-horizontal plane [92]. This removes several degrees of freedom in skull shape variability and allows the reconstruction methods to focus on modeling the anatomical variability instead. Second, the shapes should come from *precise segmentation of CT* data, including the orbital area which is often the target of the reconstruction. Additionally, as for the synthetic cranioplasty dataset, the *defects should match the shape distribution of real defects*, including both global characteristics and fine details along the defect border. Although the defect shape is often regular with craniotome drill holes along the edge, the defects can generally have arbitrary shapes. The ongoing bone remodeling and healing processes additionally increase the complexity of the defect border shape, which is usually not perfectly sharp. This chapter addresses each of these points in turn.

4

CNN-BASED SKULL ALIGNMENT

This chapter presents an automatic method for alignment of arbitrary skull CT data to the natural head position (i.e. FHP parallel with the axial plane and MSP parallel with the lateral plane [92]). The skulls in CT scans of patients, especially those indicated for cranioplasty, are often heavily rotated due to the patient's condition and medical instrumentation. The scans also span different parts of the patient's head depending on scanning protocol and type of examination. By moving the skull into the natural position, we correct for the resulting translation and rotation variability. We also define the FHP as the bottom margin of the area of interest for skull reconstruction. This technical choice lets us focus on defects of the neurocranium and the upper part of the viscerocranium including the orbital area. Although we do not directly address the excluded regions below the margin such as zygomatic arches, maxillar bone, or skull base, all methods proposed in this thesis are general and could be extended to include them.

Several works that present automatic methods for anatomical skull plane detection exist in the current literature. Cheng et al. [27] automatically detect the anatomical landmarks that define these planes. The authors first register a template to find initial landmark positions and then refine them using landmark-specific heuristics such as near extreme points or local maxima of Gaussian curvature. The method showed good performance but the evaluation was limited to three skull models and requires a healthy skull template to be correctly registered. Silva et al. [31] use a rotation-invariant 2D CNNs to estimate the optimal alignment directly instead. Both methods assume that skull segmentation is available from a pre-processing step. The method used in this work is also based on anatomical landmarks. To avoid the need for segmentation or template registration, we detect the anatomical landmarks of interest directly using a heatmap regression CT similar to the method proposed by Payer et al. [103]. The skull is then transformed into the natural position using a reference set of the landmarks defining FHP.

4.1 AUTOMATIC LANDMARK DETECTION

For the CT training and inference, the landmark detection task is defined as regression of heatmaps in form of 3D Gaussians centered around the landmark positions. We use a simple 3D U-net architecture [108, 101] with 16 initial features and the

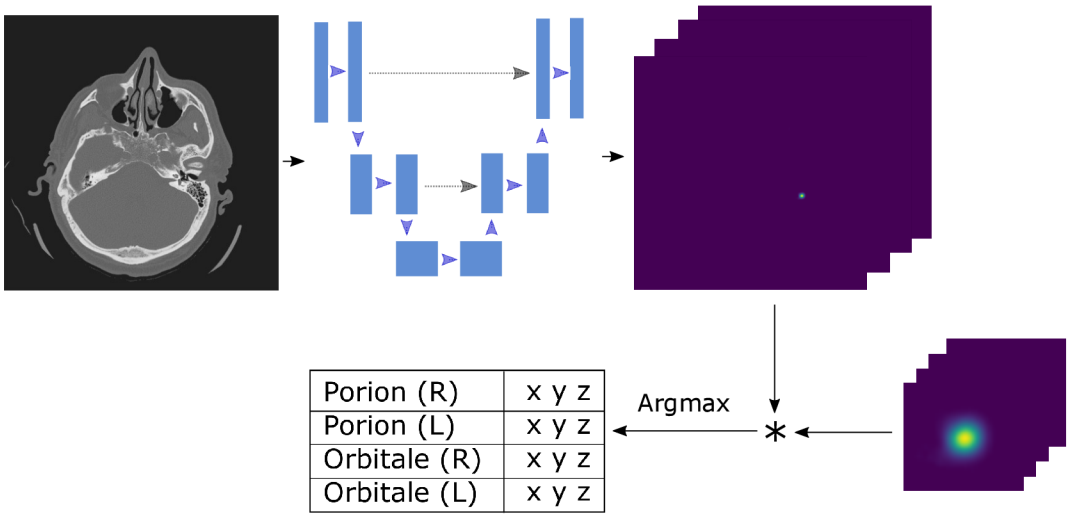

Figure 4.1: Overview of the automatic landmark detection method using heatmap regression U-net. Asterisk denotes convolution operation. Both the convolution and argmax operations are performed channel-wise. Note that 2D representation is used for illustration while the method operates fully in 3D.

number of the output channels equal to the number of target landmarks. The CT is trained for 100,000 iterations with Adam optimizer using mean squared error loss function. Rotation and translation invariance is important for the model to be able to correctly process data of patients in various positions. To account for this, we train the model using data crops with random offsets and strong data augmentation with random rotations. During inference, the positions of the landmarks are found in each channel as the coordinate of the maximum value after convolution with the same Gaussian kernel that was used for the training heatmaps generation. The method is illustrated in Figure 4.1.

The FHP is defined by the left and right Porion and Orbitale landmarks (see Figure 2.2 for reference). Because ground truth landmark annotations were not available from the experts, the annotations were done manually by the author of this thesis. This resulted in a dataset of 150 patients for the detection model training.

Figure 4.2 shows the accuracy of the trained landmark detection model on a test set of 37 patients. The median error of all four landmarks was under 2 mm, although slightly higher errors were measured in orbitale landmarks which could be partly attributed to imperfect manual annotations. The achieved accuracy is in agreement with other works using 3D CNNs for cranial landmark detection [137]. Most importantly, there were no significant outliers and the landmark detection provided stable detection on all test cases, including challenging cases with strong rotation and a high amount of noise in form of medical instrumentation and defects (see Figure 4.3).

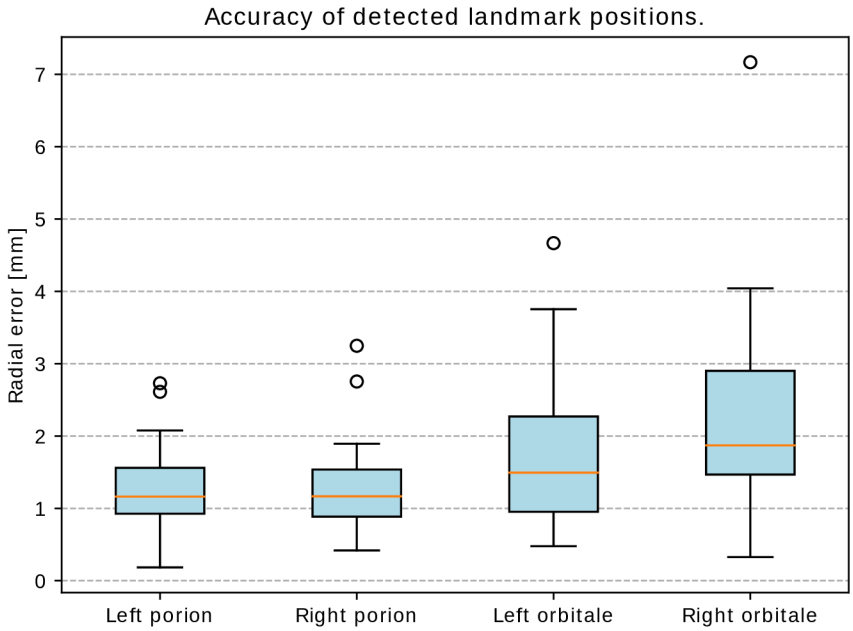

Figure 4.2: Test set accuracy of the detected landmark positions in milimeters.

4.2 CT VOLUME TRANSFORMATION INTO NATURAL HEAD POSITION

We define the target skull position by manually setting a reference set of Porion and Orbitale landmarks on the $z = 0$ plane (i. e. bottom of the target volume). Singular value decomposition, well known from the iterative closest point registration algorithm [15], is then used to compute the optimal rigid transformation matrix between the detected and reference set of landmarks. We finally align the CT volume by applying the transform to it and using linear interpolation. Several results of automatic skull CT volume alignment are shown in Figure 4.3.

Although we do not directly evaluate the accuracy of this alignment, it is shown in later sections of this thesis that normalizing the skull position using this automatic approach improves the performance of subsequent skull reconstruction methods by reducing the degrees of freedom of the skull shapes.

4.3 CONCLUSION

The proposed method is able to bring arbitrary skull CT data into the natural head position. We showed that it works well for skulls with large defects as well as patients in a strongly rotated position. The method assumes only that the skull and the CT volume contains both Porion and Orbitale landmarks. If this is not the case, a different set of landmarks defining a different anatomical plane can be used for normalization of the skull position instead. For example, in Section 9.6 of this thesis, we use Supra-orbital notches instead of Orbitale landmarks for align-

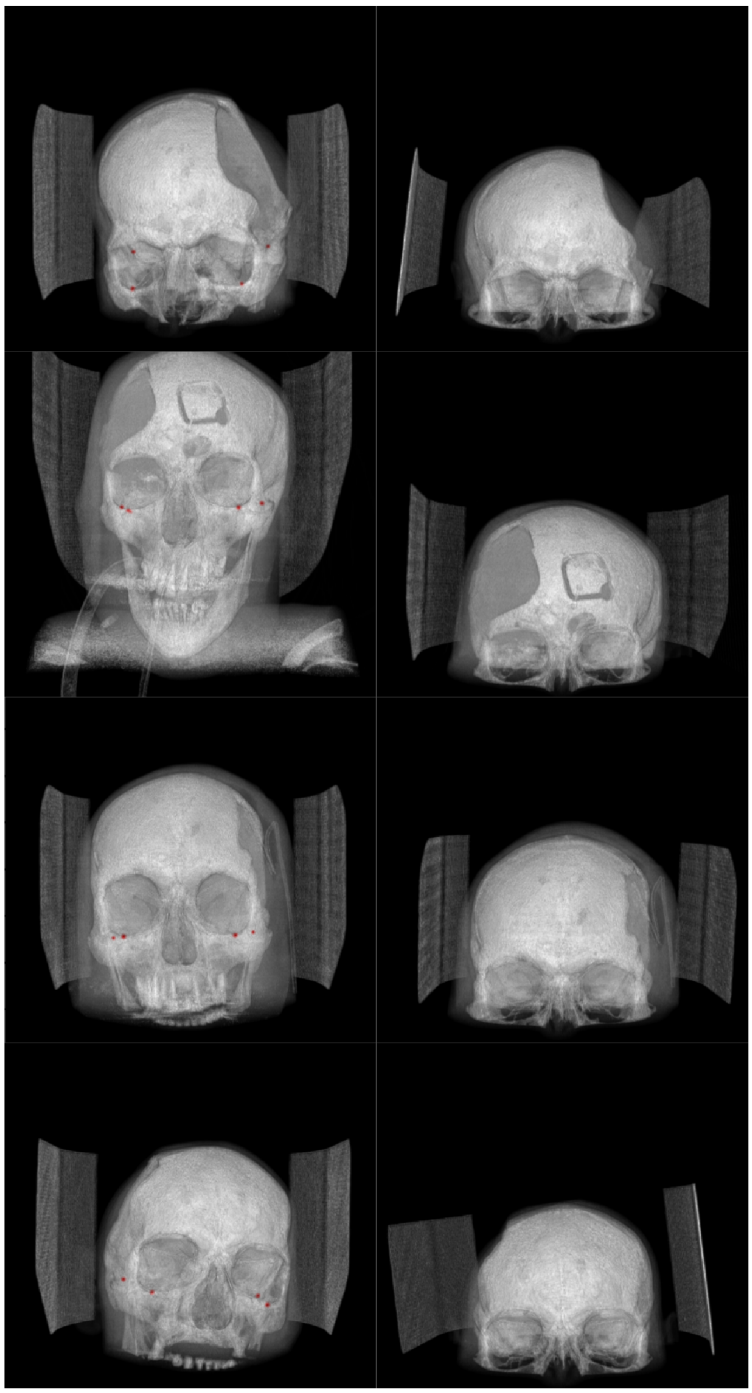

Figure 4.3: Skull CT volume alignment results. The original CT volume with the automatically detected landmarks highlighted in red (left) and the aligned skull CT volume (right).

ment because some data samples do not include them in an attempt to reduce unnecessary irradiation of optical nerves during CT scan.

Using the landmark-based alignment makes it easy to include user interaction through manual landmark position correction. Drevicky et al. [35] showed that the value of the landmark detection model output at the detected landmark position correlates with the certainty of the model estimate. The outputs of the model can therefore also be used to indicate an uncertain landmark detection and prompt the user for manual correction if required. However, we encountered no cases where significant correction would be required in our experiments.

5

SEGMENTATION OF DEFECTIVE SKULLS USING CNNS AND GRAPH-CUTS

This chapter presents the paper "Segmentation of defective skulls using CNNs and graph-cuts" presented at the MICCAI 2019 Muskuloskeletal Imaging workshop [65]. The text has been edited by removing the section with related work, which can be found in Chapter 3.1.

The main purpose of this work was to provide a fast and precise data segmentation method for anatomical model designers in TESCAN Medical company. Because of the way the provided gold-truth models were created for anatomical modeling, the ethmoidal and zygomatic bones forming the orbital area were also properly modeled which resulted in good performance of the segmentation model in this area. The maxillary sinuses were filled to increase mechanical model stability (see Figure 5.6 b). This may have contributed to the relatively high reported Dice coefficient values because it is easier to achieve higher overlap with such large and compact structures as opposed to precisely segmenting thin maxillar bones as done in other methods in the literature. Nevertheless, the method also achieved good performance in thin bones of orbital areas and since its deployment, it has been successfully used to produce anatomical models of skulls, achieving a substantial reduction of the time required for manual post-processing.

The method was additionally used for segmentation of the CQ500 public skull CT database [28]. High accuracy was achieved on this data as well, although occasional holes in the orbital area could be spotted. This made it possible to efficiently create a public, large-scale skull shape dataset which was later published and used for the development and benchmarking of skull reconstruction methods, as described in the following sections.

5.1 INTRODUCTION

Computer-assisted pre-surgical planning using generated 3D tissue models is seeing increasing use in personalized medicine. In the context of craniofacial surgery, the applications range from patient education, diagnosis, and operative planning [36] to patient-specific implant design [55], mostly in the cranial area. The latter had been accelerated by the advent of additive manufacturing (AM), also known as 3D printing in recent years [96]. A typical workflow of producing a pre-

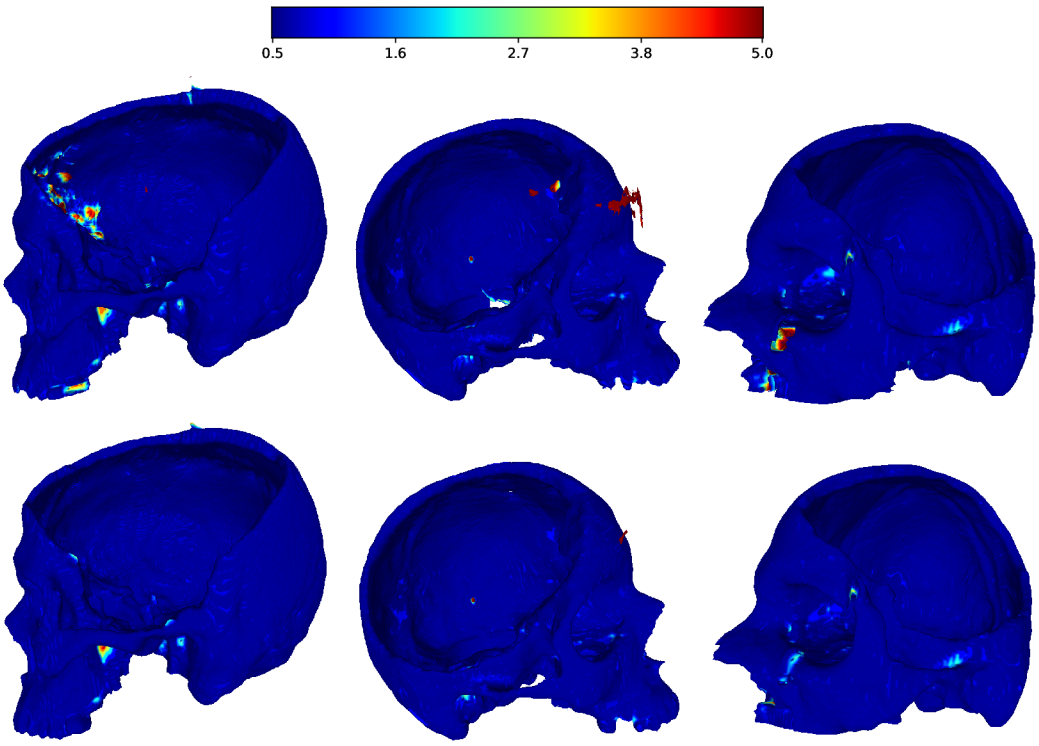

Figure 5.1: Example renders of segmented skulls with the distance to the ground-truth surface in mm coded in color. Multi-view CT segmentation outputs (top) and multi-view CutCNN segmentation outputs (bottom) are shown. To better display the differences, voxels with the surface error of less than 0.5 mm are left dark blue.

surgical 3D tissue model consists of data acquisition, converting the data into a patient model, and optionally printing the model. Computed tomography (CT) is usually the modality of choice because of its unparalleled hard tissue contrast required for precise model shape extraction. As the manufacturing process is usually able to produce the model with satisfactory precision, converting the raw CT data into an accurate patient model remains the most crucial step [87].

In this work, we propose an improved segmentation method that extracts region and boundary potentials using CNN and then uses graph-cut for globally optimal segmentation. The method outperforms methods based on conventional deep learning and other state-of-the-art methods of skull segmentation, and it produces results acceptable for the targeted use of 3D tissue modeling in clinical practice. Furthermore, we directly compare 2D and 3D CNNs for segmentation and demonstrate that the benefit of using the 3D approach is not unequivocal.

5.2 PROPOSED METHOD

We use the well-known U-net model [108] as a baseline method for our segmentation experiments. We experimented with both multi-view (MV) ensemble of 3

orthogonal 2D U-nets as used in [24] and fully 3D U-net [101] since to the authors' best knowledge, the current literature lacks direct comparison between the two approaches. The applied U-net slightly differs from the original architecture by using batch normalization and padding during convolutions, replacing the up-conv layers with bilinear up-sampling, and reducing the initial number of convolutions to 16. The architecture of the 3D model is identical except that each convolution, max-pooling, and up-sampling operation is replaced by its 3D equivalent. The networks are trained until convergence using mini-batches of shape $24 \times 128 \times 128$ in case of 2D and $4 \times 128 \times 128 \times 64$ in case of the 3D model using the Dice loss function [94].

To improve segmentation performance on slightly out-of-distribution data (such as previously unseen medical material or defect shapes), we opted to apply 3D graph-cut segmentation on the CNN output. While this approach has been taken by other authors before, we train the [83], we also modify our CT model to output an edge probability for each voxel in addition to the object probability. Thus, the final layer of the CNN has 3 channels instead of the standard 2. Figure 5.2 illustrates how this step can provide additional boundary information to the graph-cut in comparison to simply using the conventional intensity or probability gradient. Another advantage of this approach is that since both region and boundary terms have similar dynamic ranges, finding the optimal λ parameter of the graph-cut algorithm is simplified. We leave $\lambda = 1$ throughout our experiments.

We train the network using the following form of the Dice loss function:

$$\mathcal{L} = 1 - 2 \cdot \frac{\sum_{m=1}^{M} \left(p_0^m g_0^m + p_1^m g_1^m + p_e^m g_e^m \right)}{\sum_{m=1}^{M} \left(p_0^m + g_0^m + p_1^m + g_1^m + p_e^m + g_e^m \right)}, \tag{5.1}$$

where p_0^m and p_1^m are the probabilities of voxel m belonging to class background and object respectively, and g_0^m and g_1^m are the corresponding ground-truth labels. Analogously, p_e^m and g_e^m are the probability and the ground-truth label of voxel belonging to the object edge. Edge map ground truth is obtained by subtracting the binary object from its morphologically dilated version, leaving a surface with single voxel thickness. Note that edge voxels overlap with the background voxels and the edge probability channel is therefore not included in the final softmax activation layer of the CNN.

Next, the output maps are converted into a 6-connected graph structure with the region terms $R(a)$ for voxel a given by

$$R^{obj}(a) = -\ln(p_1^a), \quad R^{bkg}(a) = -\ln(p_0^a) \tag{5.2}$$

and the boundary term $B(a, b)$ between neighbouring voxels a and b given by

$$B(a, b) = -\ln[\max(p_e^a, p_e^b)]. \tag{5.3}$$

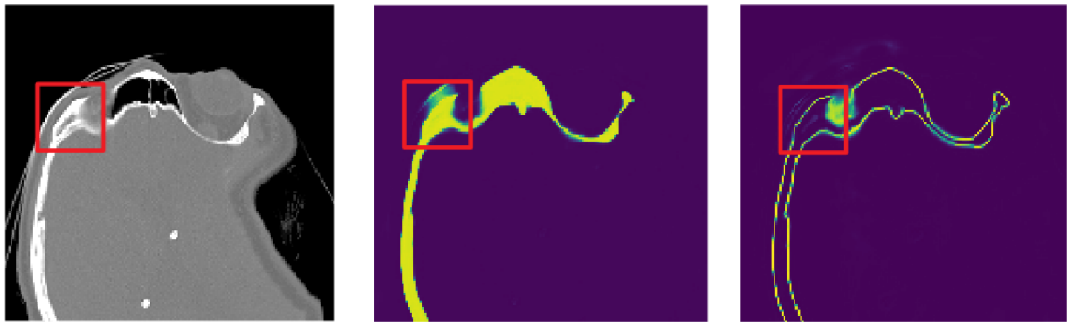

Figure 5.2: Example CT output slice, from left to right: Data, object probability map, edge probability map. Notice the segmentation error caused by an external object with a density similar to that of the skull in the upper left. The error is correctly separated by its detected edge.

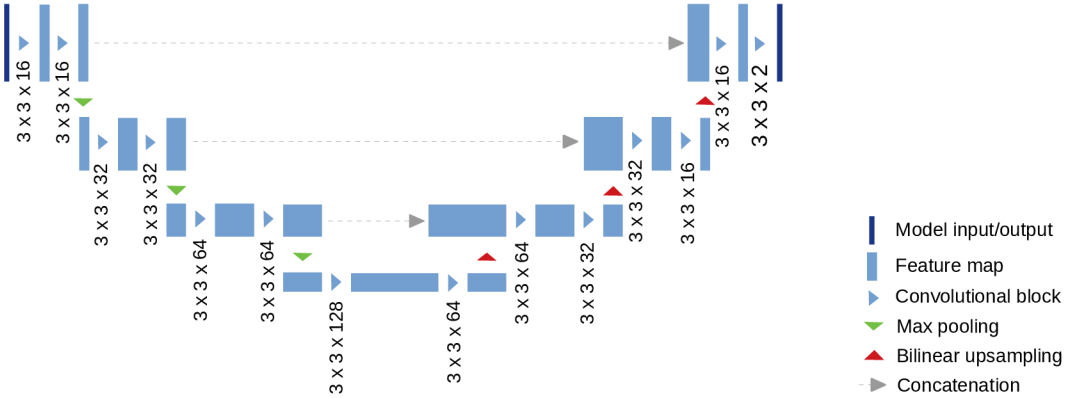

Figure 5.3: 2D Segmentation model architecture. Series of two convolutions, batch normalizations and ReLU activations are used in each block. In 3D model each 2D operator (convolutions, max-pooling and bi-linear upsampling) is replaced with its 3D equivalent.

Finally, globally optimal 3D segmentation can be obtained by finding minimum cut through this graph [18]. This method will be referred to as CutCNN in the remaining parts of the paper. Note that while the CT can be either multi-view (MV) or 3D, the graph-cut segmentation is always 3D. The 2D CNN architecture can be seen in Figure 5.3 and the complete method is summarized in Figure 5.4.

5.3 EXPERIMENTS

In this section, we present the skull tissue dataset on which the segmentation methods were evaluated. Then, we present the results of different segmentation methods on the dataset.

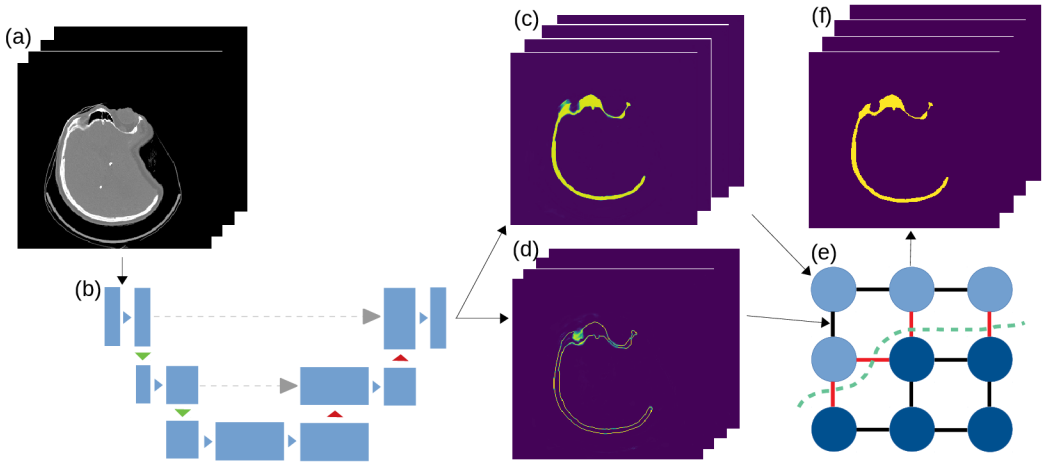

Figure 5.4: Scheme of the proposed segmentation framework. Input data (a) are processed by a CT model (b) to produce a probability map (c) and an edge strength map (d). These provide the boundary and region term for the graph-cut optimization step (e) which produces the binary output segmentation (f).

5.4 DATASET

Head CT scans of 199 different patients were available for this study. The scans were acquired for the purpose of patient skull modeling and its additive manufacturing or further patient-specific implant design. Therefore, pixel-wise ground-truth segmentation done by an experienced radiologist was also available for model training. The scans were acquired on multiple CT scanners using a variety of different acquisition protocols. The voxel size varied from $0.38 \times 0.38 \times 0.38$ mm to $0.5 \times 0.5 \times 1.5$ mm. All volumes were re-sampled to the isometric resolution of 1 mm per voxel for the experiments.

As the majority of these scans were acquired prior to surgery, the skulls often contained various defects, fixation materials, and other external objects. This makes fully automatic segmentation of these scans a challenging task, because many of these structures were only present in a single patient scan, making generalization difficult.

5.5 METRICS

Multiple metrics were used to quantitatively compare outputs of different segmentation methods used in the study. Inspired by the MICCAI 2018 Medical Segmentation Decathlon challenge[1], volumetric Dice coefficient and surface Dice coefficient were chosen. Furthermore, mean surface distance has been also included in the metrics as this is the recommended measure in the area of medical tissue model

[1] http://medicaldecathlon.com/

preparation [125]. Implementations of the metrics used in this work are publicly available[2].

Volumetric Dice coefficient (VDC) is a well-known metric in the medical segmentation domain. Given a number of true positive (TP) samples, false positive (FP) samples, and false negative (FN) samples, the coefficient is given by

$$VDC = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}. \tag{5.4}$$

In the case of volumetric Dice coefficient, the number of voxels assigned an object label in output segmentation as well as in the ground-truth segmentation is used to compute TP while FP + FN corresponds to the number of voxels assigned a different label.

To compute a surface Dice coefficient (SDC), the output and the ground-truth binary segmentation volumes are converted to polygon meshes. Each surface element in the output segmentation mesh is then considered a TP sample if the distance to the closest point on the ground-truth surface is lower than threshold t and vice-versa. The surface elements in output and ground-truth meshes that do not fall under this threshold are counted as FN and FP, respectively. We chose the threshold to correspond to the voxel size in our experiment.

5.6 EXPERIMENTAL DESIGN AND RESULTS

The performance of four different models has been evaluated in this study. Both 3D and MV CNN models and their CutCNN counterparts were implemented in the TensorFlow framework. PyMaxflow library was used for the implementation of the graph-cut optimization. All experiments were run on a desktop system equipped with Nvidia Titan Xp GPU, an i5 intel core processor, and 16GB RAM.

22 scans were randomly selected as test subjects for the experiment, leaving 177 skulls for model training. Using convolutional kernels of size 3 in all the CNN models results in the 3D model having the same number of trainable parameters as the sum of the three orthogonal 2D models. The comparison between the MV ensemble and the 3D approach can therefore be considered an ablation study to an extent. CutCNN models also have a similar number of parameters, the only difference being the final edge probability output layer. A quantitative comparison of the results of each method is presented in Figure 5.5 and Table 5.1. Further qualitative results are shown in Figure 5.6 and 5.1.

[2] https://github.com/deepmind/surface-distance

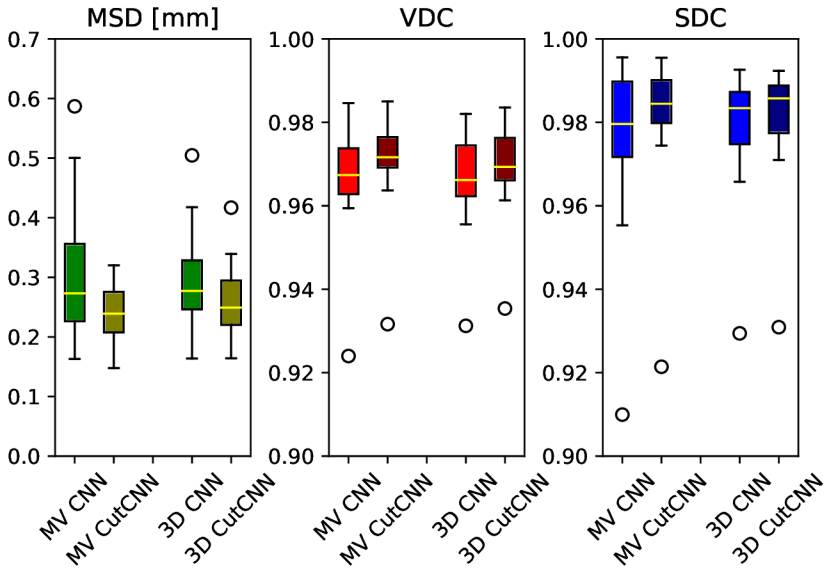

Figure 5.5: Accuracy of standard multi-view (MV) and 3D CNN and their CutCNN counterparts. Results are shown in terms of mean surface distance (MSD), volumetric Dice coefficient (VDC), and surface Dice coefficient (SDC).

5.7 DISCUSSION

CutCNN segmentation framework resulted in a performance gain in all cases in terms of every metric used in the experiment over standard CNN approaches. The output of CNN object probability map often contains errors near external objects or smaller tissue defects as these are scarce in the training data distribution. However, the graph-cut optimization guides the resulting binary segmentation towards a spatially consistent and compact shape, often eliminating these artifacts if a detected edge corresponds mostly to the correct object boundary. This effect is further illustrated in Fig. 5.1.

Our second observation is that using 3D convolutional kernels has a rather small effect on the final segmentation precision quantitatively compared to the MV approach. However, although the quantitative difference is small, for applications in medical additive manufacturing, it is important to avoid ragged segmentation output which may result from MV CNN in areas of lower model certainty. These include for example teeth, which are challenging to detect, especially when the lower and upper teeth are in contact (see Figure 5.6 a), or maxillary sinus, which is often enclosed in order to improve the mechanical stability of the manufactured model (see Figure 5.6 b). Therefore, 3D U-nets are often considered necessary to avoid these discontinuities caused by slice-by-slice processing.

However, this artifact can also be addressed by employing the CutCNN framework since the ragged segmentation boundary introduces a high boundary-term penalization during optimization and it is therefore avoided in the final binary

Table 5.1: Comparison of segmentation methods using mean surface distance (MSD) [mm], volumetric Dice coefficient (VDC) and surface Dice coefficient (SDC).

Method	**MSD**	**VDC**	**SDC**
MV CT	0.37	96.7	97.1
3D CT	0.35	96.7	97.0
MV CutCNN	**0.31**	97.7	**98.3**
3D CutCNN	0.32	**98.0**	98.1
* Minnema et al. [95]	0.44	92.0	-
* Linares et al. [81]	-	91.5	-

* Results obtained on different datasets

segmentation. Thus, employing CutCNN allows the decision between the 3D and multi-view approach to be merely a technical choice. Using 2D models can offer some advantages, such as faster training of deeper models with less overfitting [24].

We also evaluate the performance of the proposed method in the context of existing related work in skull segmentation. In terms of volumetric Dice coefficient, the proposed method achieved performance of 0.977 ± 0.019 in the multi-view scenario and 0.980 ± 0.013 in the 3D scenario. This result is considerably higher than that of 0.92 ± 0.04 reported by Minnema et al. [95]. This is probably caused by several limiting factors in the other works, including the small training set that only allowed for a smaller CNN architecture and employing a patch-based approach. To our best knowledge, the presented work is the first to apply a fully automatic segmentation approach to a pathological skull dataset of this size. Furthermore, we also achieve a low mean surface distance with the proposed method, namely 0.31 ± 0.33 mm.

We also trained the multi-view CutCNN model with an isometric resolution of 0.5 mm per voxel to facilitate enough precision for clinical practice with almost no loss in accuracy. Preliminary testing of the proposed method by experts in medical tissue modeling practice showed that the results of this model are accurate enough to substantially reduce the amount of time spent by creating the model in practice when compared to currently used semi-automatic segmentation methods.

5.8 CONCLUSIONS

In this work, we presented CutCNN, an improved hard tissue segmentation method that integrates the CNN output with graph-cut segmentation. The results of such a method surpassed the commonly used CNN architectures such as 3D

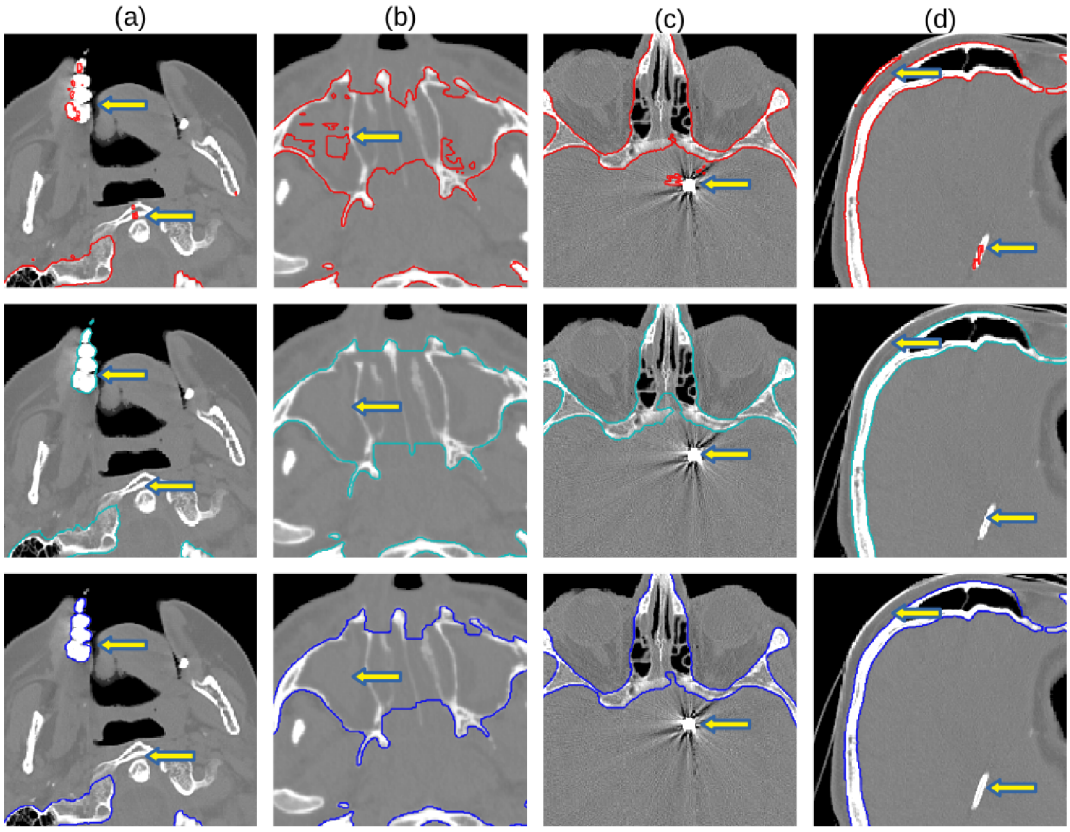

Figure 5.6: Qualitative results shown for several chosen axial slices. From top to bottom: Multi-view CT output (red), ground-truth (magenta), multi-view CutCNN output (blue).

and multi-view U-nets as well as other competitive methods in the skull segmentation domain. The object and edge probability maps in combination with graph-cut provide a compact and smooth final tissue segmentation while adding very little computational cost. This method could therefore be used to improve the performance of any semantic segmentation task given that the edges are well defined in the data. In the future, to deal with any remaining segmentation errors, user interaction can be introduced to the method on both CNN and graph-cut level as the output of both steps can be improved through user scribbles in an iterative fashion. This will further reduce the time spent producing accurate tissue models.

6

SKULLBREAK DATASET DEVELOPMENT

This chapter presents the original SkullBreak data description paper. It was published along with the work presented in Section 8 of this thesis.

Although a dataset of defective skulls along with their expert-designed ground truth implants and skull patches was provided by TESCAN Medical for this work, it had several limitations. For one, because the skulls came from patients indicated for cranioplasty, they already contained a cranial defect and therefore could not be easily extended by artificial defects for data augmentation purposes. Because they mostly contained uni-lateral defects of the neurocranium and only a small portion of bi-lateral defects or defects reaching into the orbital area, it was not suitable for training deep CNN models for generic skull reconstruction. Second, the dataset could not be published for the reasons mentioned in Section 5.

The SkullBreak dataset was created and published to address these limitations. It could be used to expand the training dataset by a virtually unlimited number of artificial defects. It was also intended to help standardize the data used for benchmarking different skull reconstruction methods.

6.1 INTRODUCTION

The final objective of research in the area of automatic skull reconstruction is providing a method that is able to complete the correct shape of missing anatomy for any possible defect instance. Such reconstructed skull then serves as a template for designing the patient-specific implant.

Despite the increasing availability of medical datasets, providing a dataset of real defective skulls with enough variability to truly gauge the effectiveness of automatic methods is very challenging. Gathering enough data to consider supervised training of deep models is even less feasible due to the fact that the most challenging cases such as big frontal defects or defects reaching into the orbital area are far less frequent than simpler cases of smaller unilateral defects of the cranium. Yet the effectiveness in dealing with these complex defects is the most desired property in automatic methods as the reduction of surgery planning time is accentuated.

To account for the lack of real defective patient scans, the practice in most of the current literature [112, 41, 42, 98] is to evaluate reconstruction methods on

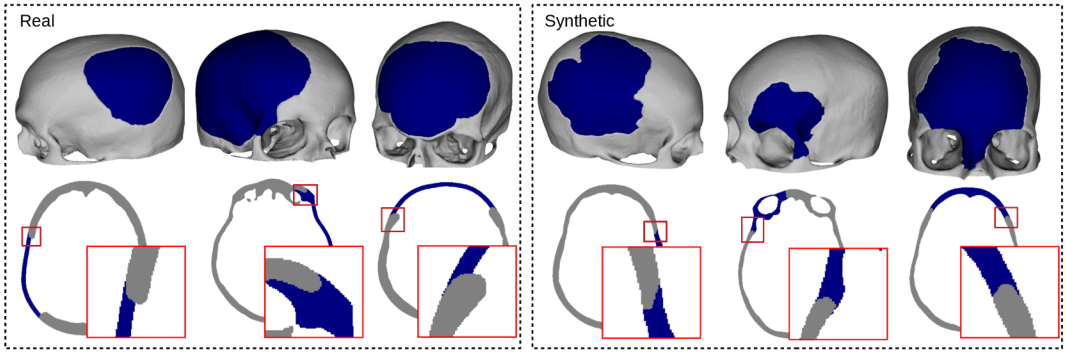

Figure 6.1: Examples of 3D models and slices through defective skulls from real patients (left) and synthetically generated defective skulls (right). The real patient samples include defects reconstructed by an experienced clinician.

skulls with simple synthetic defects created in healthy patient scans. However, these synthetic defects are usually very simplistic and are not publicly available to serve as a benchmark. In order to provide a suitable public dataset for training and validation of reconstruction methods, we created a realistic defective skull dataset from a population of healthy skulls available in the CQ500 dataset [28].

6.2 PRE-PROCESSING

Of the 491 patient scans, we filtered out those that contained skull fractures in expert reads or that were scanned with too thick slice spacing. Of the remaining 322 scans we further hand-picked only those that contain the whole cranial and orbital area and were overall suitable for further processing. The resulting 189 scans were used as the basis of our dataset. As the first step, we re-sampled each data volume to an isotropic voxel size of 0.4 mm and padded them to a size of 512^3 which ensures that the skulls can fit into the volume. Next, to remove unnecessary degrees of freedom in the skull position and rotation, we rigidly transformed each skull so that the Frankfort-horizontal plane (FHP), defined by orbital floors and auditory meatuses [84], was placed in the center of the bottom slice of the volume as illustrated in Figure 6.2.

For this purpose, we manually annotated the four landmarks defining this plane on 80 patient scans[1]. Then, we trained a CNN model to detect these landmarks in the CT data by the method of heatmap regression, which has been used extensively in anatomical landmark detection [103]. The achieved average error of the trained model on a held-out test set was under 4 mm for each of the four landmarks. We also manually placed a single set of reference landmarks on the ground

[1] This dataset was the first iteration of the landmark detection dataset described in Section 4. Although the model trained on this smaller dataset resulted in generally worse performance, the resulting alignment was successful in all cases.

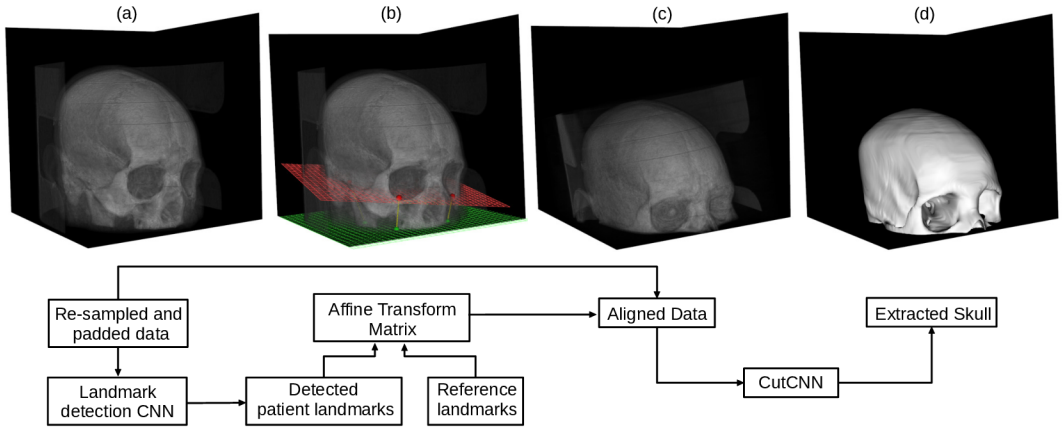

Figure 6.2: Overview of the skull extraction method. In the re-sampled and padded CT data (a), the four landmarks are first detected by CNN (b). The detected landmarks (red) and their position along with reference landmarks (green) are then used to compute an affine transform of the data into the reference coordinate system where FHP is placed in the center of the bottom of volume (c). Finally, the skull tissue is segmented using CutCNN [65] method (d).

plane of the target coordinate system. We use these reference landmark positions along with the detected landmark positions to compute the desired transformation matrix using singular value decomposition well known from the iterative closest point algorithm [15]. This automatic rigid alignment method was successful in compensating for rotation and translation on all of the patient scans.

SKULL SEGMENTATION After pre-processing all of the scans, skulls were segmented in the CT data to provide the binary shapes. For this, we used an internal dataset of defective patient scans with the ground-truth segmentation provided by an expert clinician. A combination of CNN models and graph-cut was used as the segmentation method. The method is described in detail in Section 5. The method generalized well to the population of healthy skulls in the CQ500 dataset and resulted in satisfactory segmentation of all of the scans.

6.3 DEFECT GENERATION

The skull segmentation masks in the reference coordinate system were designated a dataset of healthy skulls X_{healthy}. In current literature, synthetic defective skulls are created by subtracting a simple binary shape representing the defect area D such as cube [98, 6] or sphere [41, 42] at a random location, creating defective shape $X_{\text{defective}} = X_{\text{healthy}} - D$. While this allows for a simple and fast generation of artificial defects, such simple shapes cannot account for the variability of the shape of defect border caused by factors such as different causes of trauma or

healing stages in real patients as shown in Figure 6.1 (a). To address this issue, we generated a set of defect volumes D by first generating a union of spheres with a random radius located around a random location and then applying a random elastic transformation to the volume. The artificially defective skulls were then computed as $X_{\text{defective}} = r(X_{\text{healthy}} - D^i)$ where D^i is a randomly sampled defect volume instance and $r(\cdot)$ is a function that applies morphological opening operation with structuring element of a sphere with a random radius to the part of the healthy skull in close proximity to defect border. This operation mimics different stages of bone resorption. Examples of randomly generated defective skulls are shown in Figure 6.1 (b).

From each of the healthy skulls, we generated five instances of defective skulls. To allow for structured validation of the skull reconstruction methods, we constrain each defect into one of three areas as shown in Figure 6.3. The first area, designated *unilateral-frontal*, includes the frontal bone and parts of sphenoid and zygomatic bones and it is constrained to either the left or right side of the skull. The second area designated *unilateral-parietal* mostly includes parietal and nearest parts of frontal, temporal, and occipital bones and it is also constrained to one of the sides of the skull. The third area is designated *bilateral* and it is the combination of the first two areas, with the centers of defect concentrated near the mid-sagittal plane to ensure that defects in this group reach into both sides of the skull. Finally, two more random defects were created in each skull at completely random locations. The data were randomly split into 179 training and 10 testing samples of healthy skulls and corresponding 895 training and 50 testing samples of defective skulls. The testing skulls that had significant overlap with any of the training skulls (Dice coefficient > 0.75) were visually inspected to minimize the risk of a patient being present in both training and testing set due to follow-up scan. This test set allows for comprehensive evaluation of reconstruction methods thanks to its variability while still making it possible to measure performance even for semi-automatic methods. To our best knowledge, the biggest evaluation dataset in current literature so far contained 31 skulls [41]. The dataset is publicly available as the SkullBreak dataset[2].

[2] https://www.fit.vut.cz/person/ikodym/skullbreak/

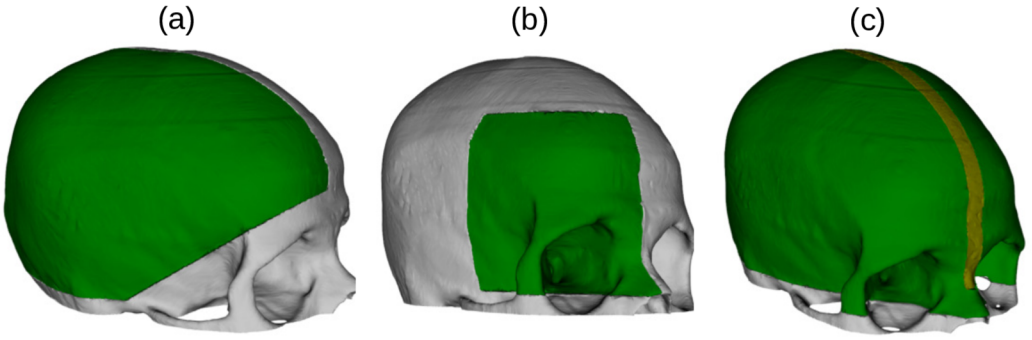

Figure 6.3: Defect area constrains. The unilateral-parietal (a) and unilateral-frontal (b) defects are constrained to one side of the skull. The random and bilateral defects are only constrained vertically (c). Bilateral defect center positions are further constrained into the yellow area.

7

SKULLBREAK DATASET IN THE CONTEXT OF THE AUTOIMPLANT CHALLENGE

The SkullBreak dataset was created concurrently and independently of the AutoImplant challenge dataset, which was used for similar goals. Due to the MICCAI 2020 event, the AutoImplant dataset (later referred to as SkullFix dataset) soon became better known in the research community. The SkullBreak dataset was therefore modified slightly to be compatible with the SkullFix data in terms of training/testing split and data format. Both datasets were then published jointly in the Data-in-brief journal [64] to serve as a combined benchmark for skull reconstruction methods. Both datasets will be used in the 2021 edition of the AutoImplant challenge that is being organized as of the time of writing this thesis.

This chapter presents the overview and mutual comparison of the two datasets as it was published in the Data-in-brief journal. The author of the thesis contributed by curating the SkullBreak part of the dataset, producing the describing data statistics, and contributing to the final article text.

7.1 DATA DESCRIPTION

The dataset described in this article was adapted from a public head CT collection from CQ500 (`http://headctstudy.qure.ai/dataset`), which was originally intended for the detection of critical findings in head CTs. The head CT collection is originally provided by the Centre for Advanced Research in Imaging, Neurosciences and Genomics(CARING), New Delhi, IN. In total, the adapted dataset consists of 880 triplets of a defective skull, a corresponding complete skull, and an implant, which is given by the difference between the defective and the complete skull. All data samples are represented as binary volumes and saved in NRRD format. The dataset is split into two tracks: the *SkullFix* track and the *SkullBreak* track.

The *SkullFix* track consists of the data used in the MICCAI 2020 AutoImplant challenge. This track is split into a training set with 100 triplets, a test set with 100 triplets, and a test set with 10 triplets. The defects in this track have a rectangular shape with craniotome drill holes in the corners, as often encountered in the craniotomy procedures. The defects are mostly located in the back of the skulls, depending on patient position during the CT data acquisition, with the exception

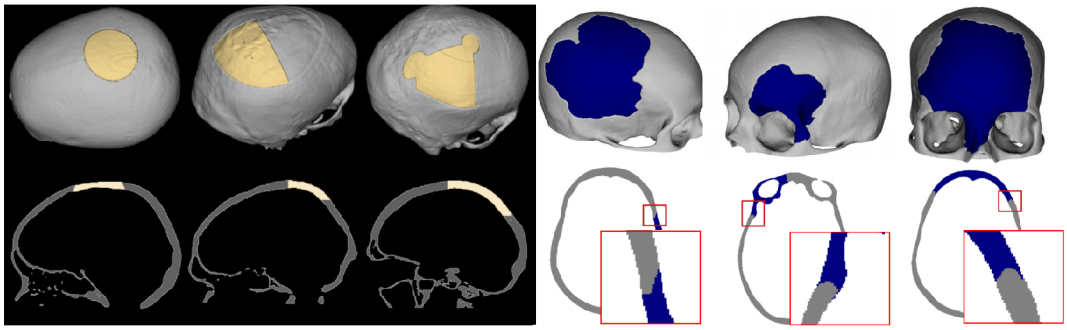

Figure 7.1: Examples of the 3D model renders (top) and slices (bottom) through the skull defect data. 3 defects from the 10 test data of the SkullFix track and 3 defects from the training data of the SkullBreak track (taken from [63]), respectively.

of the test set with 10 triplets, where the shapes, sizes, and positions of the defect are different as seen Figure 7.1. These 10 defects were used in the AutoImplant challenge to test the robustness of the proposed methods. The dimension of the data in *SkullFix* is $512 \times 512 \times Z$ and Z is the number of axial slices.

The *SkullBreak* track consists of the data originally used in Kodym et al. [63]. The dataset was adapted so that there is no overlap between training and test sets of both tracks presented in this article. 114 training and 20 test skulls were used to create this track. On each skull in this track, five different synthetic defects were created:

- unilateral defect in the parieto-temporal area,
- unilateral defect in the fronto-orbital area,
- bilateral defect,
- two random defects.

This resulted in a training set with 570 triplets and a test set with 100 triplets. The defects in this track were created with random shapes. Several examples from this track can be seen in Figure 7.1. The dimension of the data in *SkullBreak* is $512 \times 512 \times 512$.

In addition to the position and shape of the skull defects, the defect size is another factor that has an effect on the difficulty of implant design. Therefore, we report voxel occupancy rate (VOR) and approximate defect surface area (DSA) to illustrate the properties of the described dataset. Figure 7.2 shows the statistics of the dataset, including the VOR of the complete skulls and the implants and the DSA. The VOR is defined as the percentage of occupied voxels in the whole image volume: $\mathbf{VOR} = \frac{\sum \mathbf{V}}{N}$, where $\sum \mathbf{V}$ is the number of occupied voxels in the volume $\mathbf{V}$ and $N = 512 \times 512 \times Z$ is the total voxel number (occupied and unoccupied). The DSA was approximated by isolating only the implant voxels located on the

outer surface of the original complete skull and using the voxel size information to approximate the total area of the isolated surface. The differences in DSA between individual defect groups in the dataset are most apparent in the *SkullBreak* track, owing to the fact that the groups are constrained to different areas. The unilateral defects (parieto-temporal and fronto-orbital) are smaller because they are restricted only to one side of the skull, while the others are unconstrained. The fronto-orbital defects are further constrained to the space around orbits while the parieto-temporal can span the whole braincase area. The *SkullFix* defects have less variability, mostly being average in their size with the exception of the 10 test cases which are smaller.

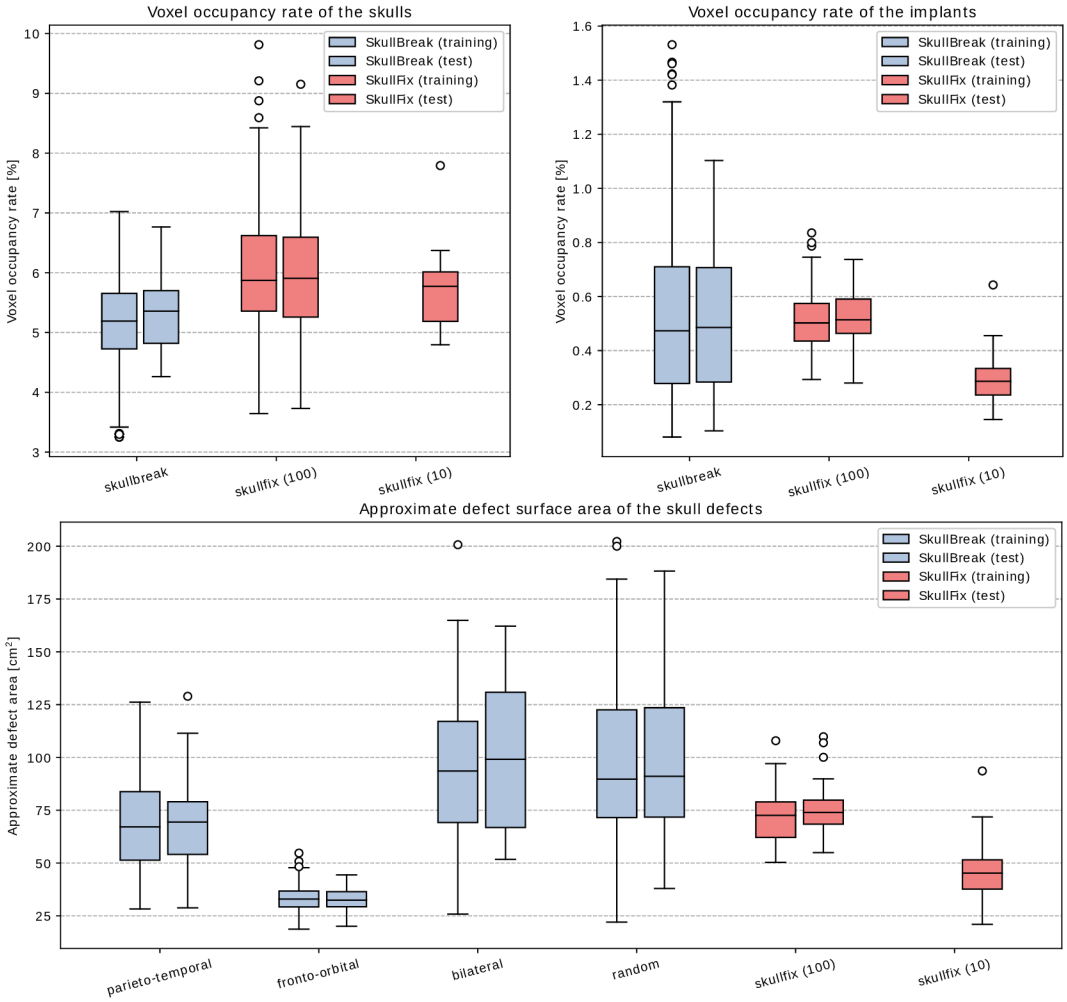

Figure 7.2: Boxplot of the data information for the training set and test set of both dataset tracks, including VOR of the complete skulls (top left), VOR of the implants (top right), and approximate defect area (bottom).

Table 7.1: Differences between the SkullFix and the SkullBreak dataset tracks.

	SkullFix	**SkullBreak**
Training/test split	$100/100 + 10$	$570/100$
Volume size	$512 \times 512 \times Z$	$512 \times 512 \times 512$
Voxel size	various	0.4 mm
Preprocessing	acquisition geometry regularization transform	acquisition geometry regularization transform and rigid alignment using the landmarks defining Frankfort-horizontal plane
Skull segmentation	thresholding at 150 HU, noise removal using connected components analysis	convolutional neural network and graph-cut [65]
Defect injection	binary defect shape subtraction from complete skull	binary defect shape subtraction from complete skull and defect border smoothing using morphological operations

7.2 EXPERIMENTAL DESIGN, MATERIALS, AND METHODS

Cranial implant design is the main bottleneck for an optimized workflow for cranioplasty [75]. The development of automated cranial implant design solutions, especially deep learning based solutions, has been hindered by the lack of public datasets. The real defective skulls from brain tumor surgery or trauma are difficult to obtain in large quantities and there often lack the ground truth, i.e., the complete skulls or the implants. We hence devised a pipeline to convert publicly available head CT collections into datasets suitable for training deep learning models. Artificial skull defects are used to simulate the surgical and traumatic process. The original CQ500 data contain head CT scans and expert annotations from the Centre for Advanced Research in Imaging, Neurosciences and Genomics, New Delhi, IN. The scans that contained skull fractures according to the expert annotations were removed as well as those that were not acquired with thin plate scanning protocol. The skulls in the presented dataset were sampled from the remaining scans. In cases of multiple series, the series with the highest resolution was used for further processing of each scan. The chosen scans contained between 211 and 394 axial slices. Acquisition geometry regularization transform was applied to each scan to correct deformations caused by gantry tilt and the data were converted to the NRRD format.

Both dataset tracks were created from the original CQ500 data using three main steps; CT data preprocessing, skull segmentation, and artificial defect injection. However, because they were originally created independently of each other, the individual processing steps differ. The different properties and the specific steps taken to create the two tracks are summarized in Table 7.1.

For preprocessing, the acquisition geometry regularization transform of SkullFix is performed using 3D Slicer to correct obvious deformation or tilting of the skull data. It involves several steps:

- Go to Edit -> Application Settings -> DICOM and change “acquisition geometry regularization” to “apply regularization transform”.
- While importing the DICOM, an item called “acquisition transform” is shown within the series.
- In the transform hierarchy, right-click on the deformed DICOM and press “harden transform”.
- Save the transformed data as NRRD.

For SkullBreak, the “dcm2niix” tool (`https://github.com/rordenlab/dcm2niix`) is used to convert DICOM series to 3D data. It corrects the deformation caused by

gantry tilt. On each skull, four landmarks defining the Frankfort-horizontal plane, i.e., the left and right auditory meatus and the left and right orbital floor, were aligned onto the $z = 0$ plane using rigid transformation. Finally, to be in conformity with the SkullFix dataset, the SkullBreak dataset is converted into NRRD format.

7.3 ARTIFICIAL DEFECT SHAPES

The skull defects in the presented dataset are created by subtracting a part of the healthy skull with the subtracted part serving as ground truth for the implant shape. In real patients, the defects can have different causes, such as craniotomy due to brain or bone tumors, brain swelling, or traumatic fractures. Consequently, the shape, position, and size of the defects can also vary.

During a craniotomy, the craniotome drill is often used to drill holes into the skull by the surgeon, leading to the defect having roundish holes in the corners. The defects in the *SkullFix* track emulate this property by injecting defects with such roundish corners into the back of the complete skulls. The defects in the *SkullBreak* track, on the other hand, have completely random shapes given by random combinations of elastically deformed spheres. The positions of the SkullBreak defects are also random, although they are structured into several categories, as discussed in the *Data Description* section. Furthermore, ongoing bone remodeling processes can deform the defect borders. To simulate this, the defect borders in the *SkullBreak* track are smoothed by morphological opening operation with structuring element in form of sphere with a radius between 2 and 7 voxels. Examples of the defects in both dataset tracks can be seen in Figure sh7.1.

Part III

VIRTUAL SKULL RECONSTRUCTION

This part of the thesis presents the development of an automatic virtual skull reconstruction method based on cascaded 3D CNN architecture. It follows up directly to the work described in the previous chapter by operating on the aligned and segmented volumetric skull shape representations of the SkullBreak and in-house cranioplasty data.

Experiments were performed on the synthetic SkullBreak and SkullFix data with ground-truth shapes to confirm that the proposed architecture is capable of performing the skull reconstruction task with sufficient accuracy and to study the effects of individual model components on the reconstruction performance. Then, an approach for combining the synthetic data with real data containing the expert-provided gold truth reconstructed shapes was proposed, providing a way to jointly design both the reconstructed skull shapes and the final cranial implant.

Attention is also given to the question of quantitative evaluation of skull reconstruction result and the choice of quantitative metrics. The topic is addressed by proposing a novel skull reconstruction metric and discussing the commonly used quantitative metrics in the context of subjective expert scoring of the automatic skull reconstruction outputs.

8

SKULL RECONSTRUCTION USING CASCADED CNNS

This chapter presents the paper "Skull Shape Reconstruction Using Cascaded Convolutional Networks" as published in the Computers in Biology journal [63]. The text has been edited by removing the chapter with related work, which can be found in Chapter 3.2 and the chapter describing the synthetic dataset, which is described in the Section 6.

The purpose of this work was to experimentally confirm the hypothesis that a CNN model is capable of learning to reconstruct generic skull defects with high enough resolution to be considered for clinical practice, as this was only previously done on very coarse data volumes [98]. A generative approach to model the skull reconstruction process was also explored. Although the achieved variability was limited, it was the first attempt to generate multiple reconstruction hypotheses for a single defective skull.

8.1 INTRODUCTION

Patient-specific implants (PSIs) are often used for the treatment of craniofacial defects. Especially in cases of larger defects caused by trauma, tumor resection, or decompressive craniectomy, it is usually required to reconstruct the original skull shape for aesthetic purposes and protection of intracranial structures against mechanical impact [68, 69].

Current state-of-the-art methods usually comprise of using a patient CT scan to design the implant pre-operatively and then 3D printing of the result using biocompatible materials such as titanium, porous polyethylene, or polyether ether ketone [55, 52]. Alternatively, implants can be cast in a 3D-printed mold from bone cement which can be loaded with antibiotics to decrease the risk of infection [122]. Such approaches lead to a reduction of operative time and improved patient results [110].

Provided that precise enough tissue segmentation is obtained from the CT data, the process of computer-aided design (CAD) of PSIs remains the most important step that affects the final quality and reproducibility of PSIs [125]. This presents a challenging and tedious task for the clinician or engineer designing the PSI. To ensure correct healing and prevent complications, the PSI must fit precisely to the defect border without any steps and with gaps of less than 0.8 mm between the

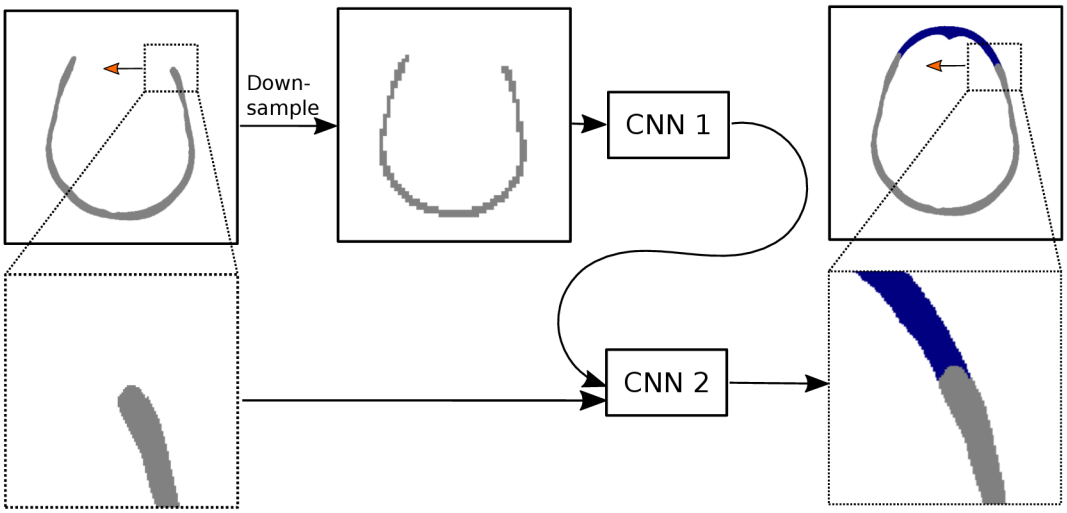

Figure 8.1: The proposed skull reconstruction method fully automatically produces a binary shape of the missing part of the skull using sliding-window approach with coarse-resolution middle step.

implant and remaining tissue [86]. Furthermore, smoothness and symmetry of the anatomy should be preserved to ensure a correct aesthetic result [12]. The first step of the implant design is a correct reconstruction of the missing shape of the skull anatomy from which the PSI is then derived.

CONTRIBUTIONS In this work, we design a cascaded CNN architecture for the estimation of a high-resolution 3D anatomy shape conditioned on the input defective skull. Although symmetry is used in the proposed method as an additional guiding signal, the method can successfully reconstruct defects reaching into both sides of the skull as well as into more complex anatomical regions such as orbitals. To the authors' best knowledge, this is the first deep learning-based method of 3D shape reconstruction that reaches a high enough resolution to be clinically viable for the skull reconstruction task. To address the issue of multiple possible correct solutions, we also experiment with a probabilistic generative version of the proposed model. Finally, in order to improve the reproducibility of research in the area of automatic skull reconstruction, we introduce an open dataset of skull shapes with synthetic defects. The dataset mimics the variability in shape, position, and bone resorption present among real patients and we show that a CNN model trained on this synthetic dataset also performs well on challenging real patient cases without any further pre-processing.

8.2 MATERIALS

For training and validation of the proposed method, we used the Skullbreak dataset described in Section 6. The synthetic dataset was split into 179 training and 10 testing skulls, resulting in 895 training and 50 testing defect shapes in total. To evaluate the ability of our approach to generalize, we also utilized an internal dataset of 9 real defective patients. For these patients, ground-truth skull reconstructions made by a clinician experienced in cranial implant design were available. Several samples from both datasets are shown in Figure 6.1.

8.3 METHODS

We formulate the skull reconstruction task as finding the missing part of the anatomy represented by binary volume $Y = X_{healthy} - X_{defective}$. Thus, we look for the function $f(\cdot)$ with parameters θ that maps the defective skull to an estimated shape $\hat{Y} = f_\theta(X_{defective})$ from distribution $P(\hat{Y}|X_{defective})$ of shapes that correctly complete it.

8.3.1 *Reconstruction Model Architecture*

We use a combination of two CNN models with a 3D U-net [101] backbone to approximate the function f, with parameters θ being the trainable weights of the CNN. The individual models differ from the original 3D U-net in several ways. Instead of up-convolutions, we use nearest-neighbor up-sampling followed by regular convolution in the decoder part of the model, as this has been shown to improve the model training process and performance in some cases [66]. The number of down-sampling and up-sampling layers is such that the bottleneck tensor has spatial dimensions of 4^3 as shown in Figure 8.2. This ensures that the output neurons of the CNN have a sufficient receptive field to correctly model the shape of missing anatomy in the case of defects with a large surface area.

Each of the models operates on a different resolution. The first model denoted f^{Lo} takes an input volume down-sampled to 64^3 voxels and is trained to output an estimate of missing anatomy $\hat{Y}^{Lo} = f^{Lo}(X^{Lo}_{defective})$ on an equivalent resolution of 3.2 mm per voxel. While this resolution is too low to model anatomy with enough precision, it can provide an initial estimate of the missing shape. The second model denoted f^{Hi} then takes a $128 \times 128 \times 128$ crop of the input data at the original high resolution concatenated to an up-sampled output of the first model. This model is trained to output the corresponding patch of the final missing anatomy estimate $\hat{Y}^{Hi} = f^{Hi}(\hat{Y}^{Lo}, X^{Hi}_{defective})$, which can be viewed as a super-

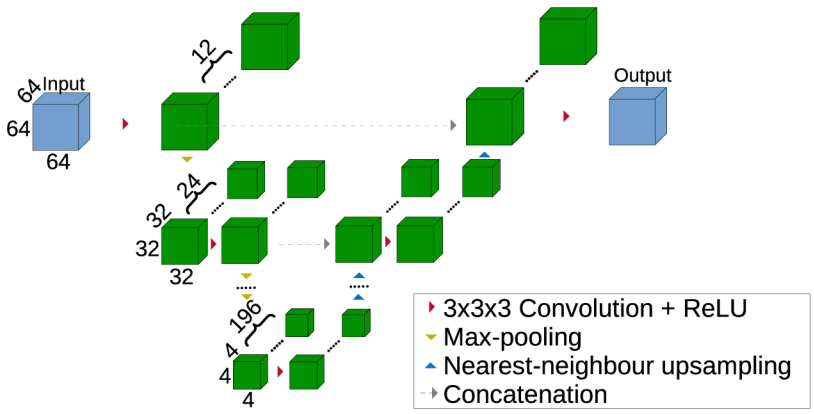

Figure 8.2: Overview of the 3D CNN backbone architecture example for an input of size 64^3. Note that for an input of size 128^3, the CNN is deeper and the bottleneck tensor has 384 channels.

resolution of the initial missing anatomy estimate conditioned on the remaining part of the skull at full resolution.

During inference, the first model provides enough contextual information about the overall shape of the defective skull while the second model can ensure precise contact at the defect border. The final estimate can be inferred by first computing the coarse estimate $\hat{Y}^{\text{Lo}}$ and then computing the final estimate $\hat{Y}^{\text{Hi}}$ using the sliding window approach, substantially reducing the memory footprint.

SYMMETRISED INPUT The chosen U-net architecture in the low-resolution CNN is in fact not well suited for transferring information from one side of the volume to the opposing side as this transfer can only happen in the deeper layers of the model where the shape information is already compressed. However, the ability to preserve anatomical symmetry is a critical part of the method. To this end, we concatenated a sagitally flipped copy of the volume to the low-resolution CNN input. This makes it easier to propagate the symmetry information using convolutional kernels and skip connections of the U-net architecture. The effect of symmetrizing input is demonstrated in Section 8.4.

8.3.2 *Optimization*

We optimize the CNNs using training batches of size 2, which fully utilize the available GPU. An Adam optimizer is used as it is currently one of the most widely used optimization algorithms suitable for most deep learning applications [119]. Although we train both CNNs with their respective loss functions $\mathcal{L}^{\text{Lo}}$ and $\mathcal{L}^{\text{Hi}}$, we train the cascade in an end-to-end manner. The training samples for the first

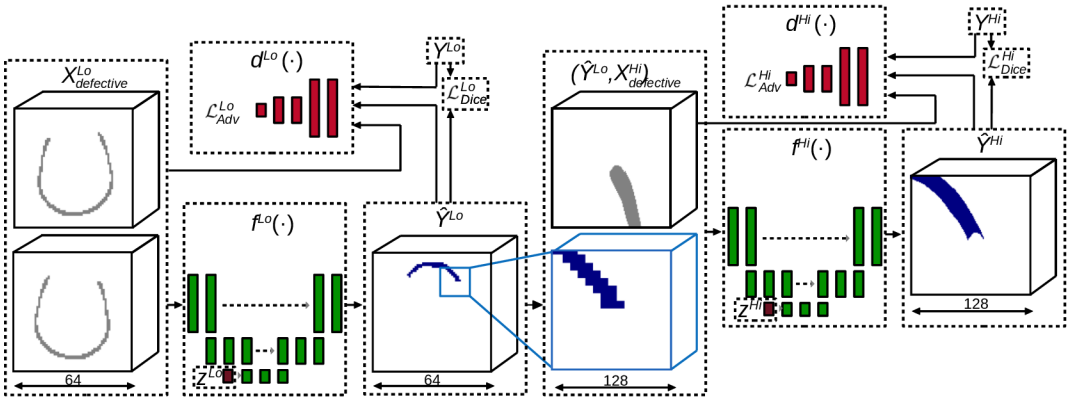

Figure 8.3: Overview of the proposed 3D CNN cascade. The symmetrized low-resolution input X^{Lo} is fed into the first model f^{Lo} to produce the missing shape estimation $\hat{Y}^{Lo}$. Then, it is concatenated to the high-resolution input X^{Hi} and fed into the second model f^{Hi} to produce the final high-resolution missing shape $\hat{Y}^{Hi}$. Discriminator CNNs $d(\cdot)$ and latent vectors z are only used in the generative model.

model $(Y^{Lo}, X^{Lo}_{defective})$ and the random training crops for the second model $(Y^{Hi}, X^{Hi}_{defective})$ are always sampled from the same skull volume.

DISCRIMINATIVE MODEL We first assume that the skull reconstruction task has a single correct ground-truth solution given by the original missing anatomy shape Y. This allows us to use a reconstruction loss similar to a segmentation task. We chose the soft Dice loss [94] due to its good performance in dealing with class imbalance. The two losses are defined as

$$\mathcal{L}^{Lo}_{Dice} = Dice(Y^{Lo}, f^{Lo}(X^{Lo}_{defective})), \tag{8.1}$$

$$\mathcal{L}^{Hi}_{Dice} = Dice\left(Y^{Hi}, f^{Hi}\left(f^{Lo}(X^{Lo}_{defective}), X^{Hi}_{defective}\right)\right), \tag{8.2}$$

and we optimize them iteratively for $300,000$ training steps. While it is possible to optimize the whole cascade using only the $\mathcal{L}^{Hi}$ loss, we found that using the auxiliary loss $\mathcal{L}^{Lo}$ is necessary for correct model behaviour.

GENERATIVE MODEL To make the described reconstruction model generative, we make two modifications well known from GAN literature [45] to both CNNs. Namely, we add the adversarial loss function $\mathcal{L}_{Adv}$ in the form of a discriminator CNN $d(\cdot)$, which allows the model to learn the distribution $P(\hat{Y}|X_{defective})$, and inject a random latent vector into the reconstruction CNNs, which allows them to randomly sample from this distribution. We concatenate the random latent vector with the bottleneck tensor of both CNNs as shown in Figure 8.3. The discriminator CNNs have the same architecture as the encoder part of the reconstruction CNNs with additional dense layers that output the discriminator scores. We use

the improved Wasserstein GAN formulation with gradient penalty [48] during the training. Given a combination of the defective skull shape and the missing anatomy shape, the discriminator is trained to assign a low score $d(Y, X_{defective})$ to the ground-truth missing shape and a high score to the reconstructed missing shape $d(\hat{Y}, X_{defective})$ at both a low and high resolution, using the low-resolution discriminator d^{Lo} and high-resolution discriminator d^{Hi}. To optimize the reconstruction CNNs in this case, we use a combination of the reconstruction and adversarial loss, similarly to Wang et al.[130]. The losses of the reconstruction CNNs, in this case, are defined as

$$\mathcal{L}^{Lo} = \mathcal{L}^{Lo}_{Dice} + \alpha \mathcal{L}^{Lo}_{Adv}, \tag{8.3}$$

$$\mathcal{L}^{Hi} = \mathcal{L}^{Hi}_{Dice} + \alpha \mathcal{L}^{Hi}_{Adv}, \tag{8.4}$$

where α is set to 10^{-2}. We again optimize d^{Lo}, f^{Lo}, d^{Hi} and f^{Hi} iteratively for $300,000$ training steps. For an overview of our method and both discriminative and generative models, see Figure 8.3.

8.4 EXPERIMENTAL RESULTS

The experiments discussed in this section were run on a system with 11 GB Titan Xp GPU and a quad-core i5 processor with 24 GB RAM. The complete training of the models took approximately 8 days. After the model is trained, the method is able to fully reconstruct each skull in under 5 seconds, which is important for its efficient use in clinical practice. This is achieved by first inferring the low-resolution model on the full down-sampled volume and then sequentially inferring the high-resolution model on positions where the low-resolution model predicted a defect until the whole estimated defect area is processed. For visualization, the voxel grid was converted into a polygonal mesh which was then smoothed using a two-step smoothing algorithm [14].

We measured the precision of each method as the average symmetric unsigned distance between the surface voxels of the output reconstruction and the original anatomy shape which we considered to be ground-truth. We only measured the error on the outer surface of the skull because the inner surface is not relevant for cranial implant design in clinical practice. To get more insight into the performance of our method, we divided the evaluation into four groups. The *unilateral-parietal*, *unilateral-frontal* and *bilateral* defect groups are described in Section 8.2 and the *combined* group includes all the defects, including random ones. Table 8.1 contains the average surface error for all models tested on each defect group.

We first evaluated the performance of the discriminative model and the effect of the symmetrized input on the error distribution in the testing set. The model with a simple input was able to reconstruct each testing skull successfully. However, we

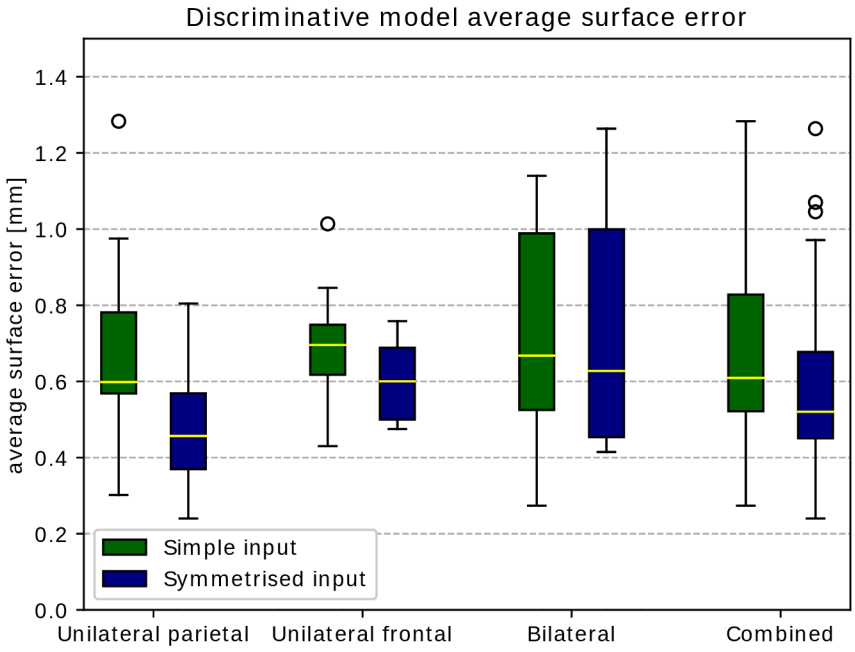

Figure 8.4: The overall performance of the discriminative model on different groups of testing defective skulls. Average surface error [mm] for a simple input (green) and symmetrised input (blue).

noticed that the errors in unilateral groups reached similar values as the bilateral group. This is in conflict with the expectation that while bilateral defects could allow for some variability in correctly completed shapes, the unilateral defects should be more directly constrained by the condition of symmetry and thus yield lower surface reconstruction errors. The effect of symmetrizing input as described in Section 8.3 was that the average measured error of the reconstructed unilateral defects dropped from 0.69 mm to 0.48 mm for parietal and from 0.69 mm to 0.60 mm for frontal defects. As expected, the bilateral defects group was less affected by the symmetrized input, although the error still slightly decreased since some bilateral defects are in fact partly constrained by the symmetry. The overall performance of the discriminative model for both simple and symmetrized input is shown in Figure 8.4. The overall average surface error of the discriminative model with a symmetrized input for the whole testing set was 0.59 ± 0.21 mm. Several examples of the discriminative model reconstructions are shown in Figure 8.5.

In order to explore the relationship between the discriminative model performance and the area of the reconstructed defects, we created an extra set of nine cranial defects in each of the ten designated test skulls. The defects were created by subtracting the same shape with different scales from each skull (see Figure 8.7 for their illustration). The surface area of the resulting skull defects ranged from 10 to 140 cm^2. The resulting surface errors of the discriminative model outputs are shown in Figure 8.6 in the form of a scatter plot. While there is an apparent correlation between the measured surface error and the reconstructed surface area, the

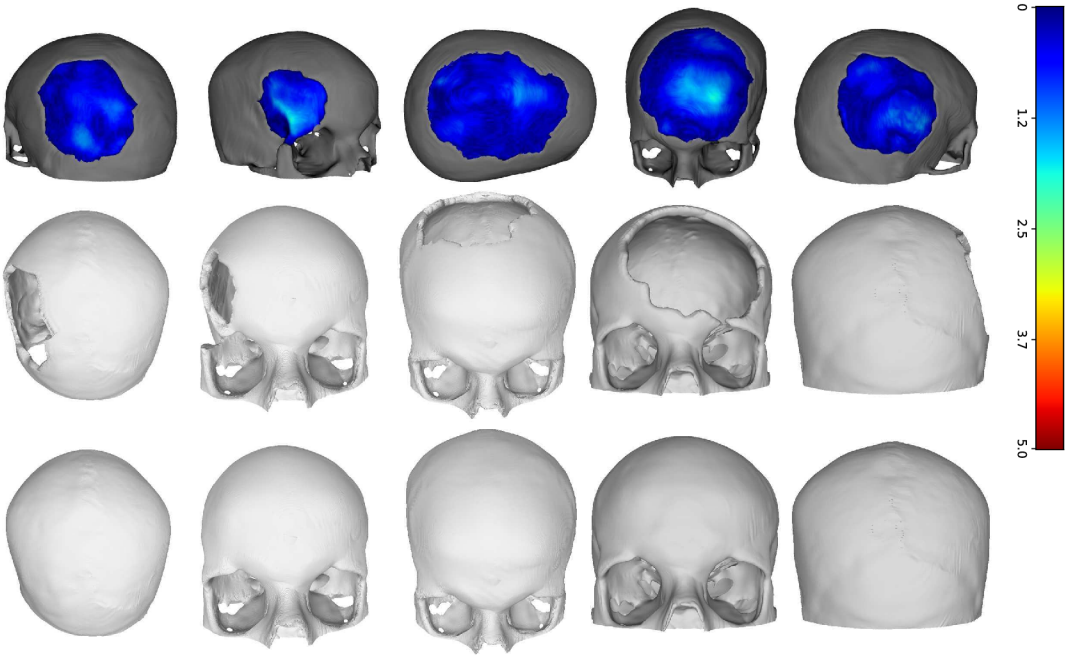

Figure 8.5: Examples of results of the discriminative model reconstructions for parietal, frontal, bilateral and random defects, respectively. From top to bottom: Surface error maps, input synthetically broken skulls and reconstructed skull shapes. Note that the majority of the reconstructed surface reaches errors of less than one millimeter in all cases.

average surface error was under 0.7 mm for all defects up to an area of 100 cm^2. For even larger defects, the average surface error exceeded 1 mm in several cases. However, for the majority of the cases, the surface errors of the results were still well under this value.

Next, we evaluated the performance of the generative model with symmetrized input volumes and random input latent vectors z. The overall average surface error was 0.68 ± 0.28 mm. For each defect group, the error of the generative model was higher than that of the discriminative model. However, it should be noted that since now we consider multiple correct reconstructions for a single skull defect, the error measured against the ground-truth shape might not be a good indicator of the method's performance. The generative model allows us to sample multiple different outputs for a single input defective skull by changing the input latent vectors. Therefore, we also experimented with generating multiple reconstructions and measuring the best achieved result. The overall average surface error when measuring the best-of-five sampled reconstructions for each testing skull was 0.56 ± 0.21 mm. The results for individual defect groups, as seen in Figure 8.8, were similar to the discriminative model in this case. However, a reduction of the error can be noticed in the *bilateral* group, with the error reduced from 0.73 mm to 0.65 mm when compared to the discriminative model. This might once again be explained by the fact that due to weaker symmetry constraints in this group,

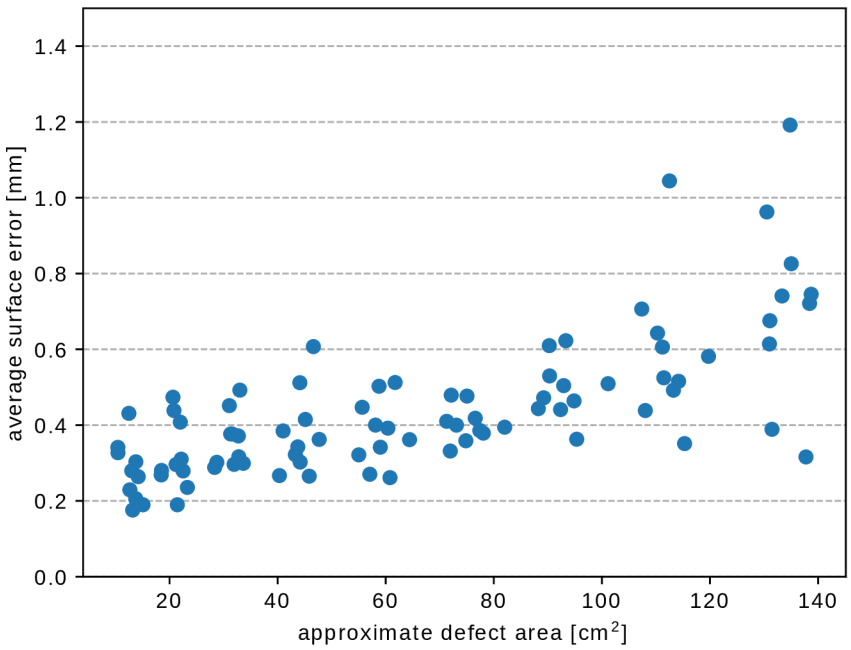

Figure 8.6: The performance of the discriminative model in the context of a reconstructed surface area. While the variability of the model output errors increases with larger defects, the results are within an acceptable range even for a majority of the larger defects.

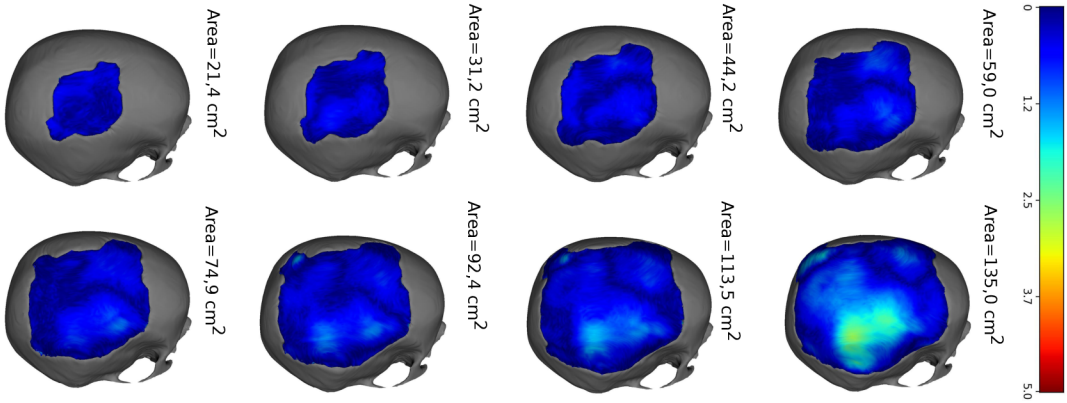

Figure 8.7: Examples of results of the discriminative model reconstructions for a set of defects with different scales on a single test skull. An area where the model output deviates from the original shape by more than 2 mm can be observed in the last case.

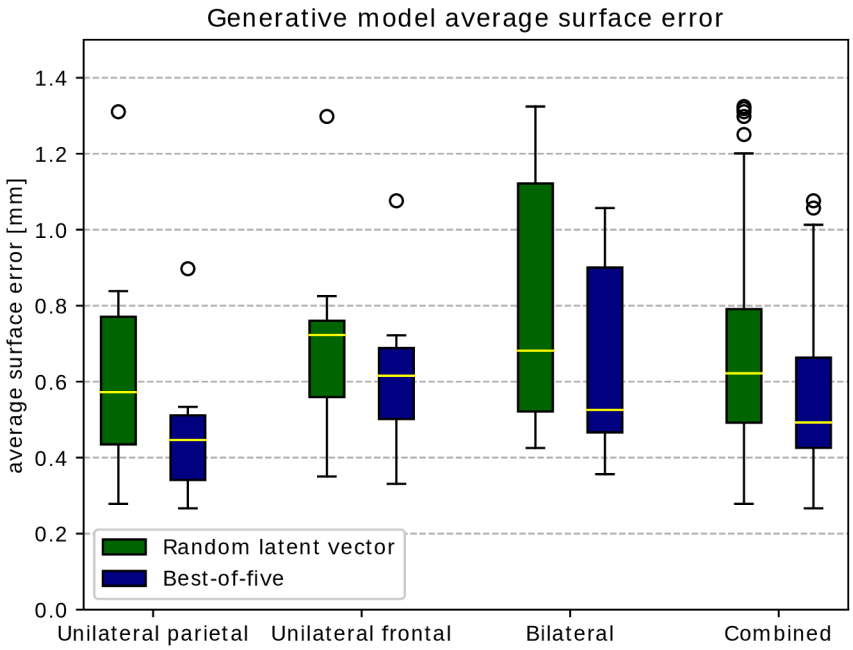

Figure 8.8: The overall performance of the generative model on different groups of testing defective skulls. Average surface error [mm] for reconstruction with random latent vector (green) and for best-of-five reconstructions (blue).

the variability of acceptable reconstructions is greater. Therefore, generating subsequent different samples constrained on the same input increases the probability of generating at least one sample close to the original ground-truth shape.

To further illustrate the behavior of the generative model, we also conducted an experiment with latent space interpolation for one bilateral defect. We set both latent vectors for low-resolution z^{Lo} and high-resolution z^{Hi} to only contain constant values $c/10$ and we generated samples for $c = 1, 2 \ldots 9$. Our experiments showed that the generative model responds to these changes in total latent vector energy the most and we leave investigation into the limits of achievable anatomical variability in the output for future work. The resulting reconstructions of the gen-

Table 8.1: Average surface error [mm] for individual defect groups.

Method	**Synthetic defects**				**Real defects**
	UP	UF	Bi	Total	
Statistical shape models [41]	0.47	-	-	-	-
Discriminative	0.69	0.69	0.78	0.68	-
Discriminative + symmetrised	0.48	**0.60**	0.73	0.59	0.80
Generative + symmetrised (random)	0.63	0.71	0.81	0.68	-
Generative + symmetrised (best of 5)	**0.46**	0.62	**0.65**	**0.56**	**0.69**

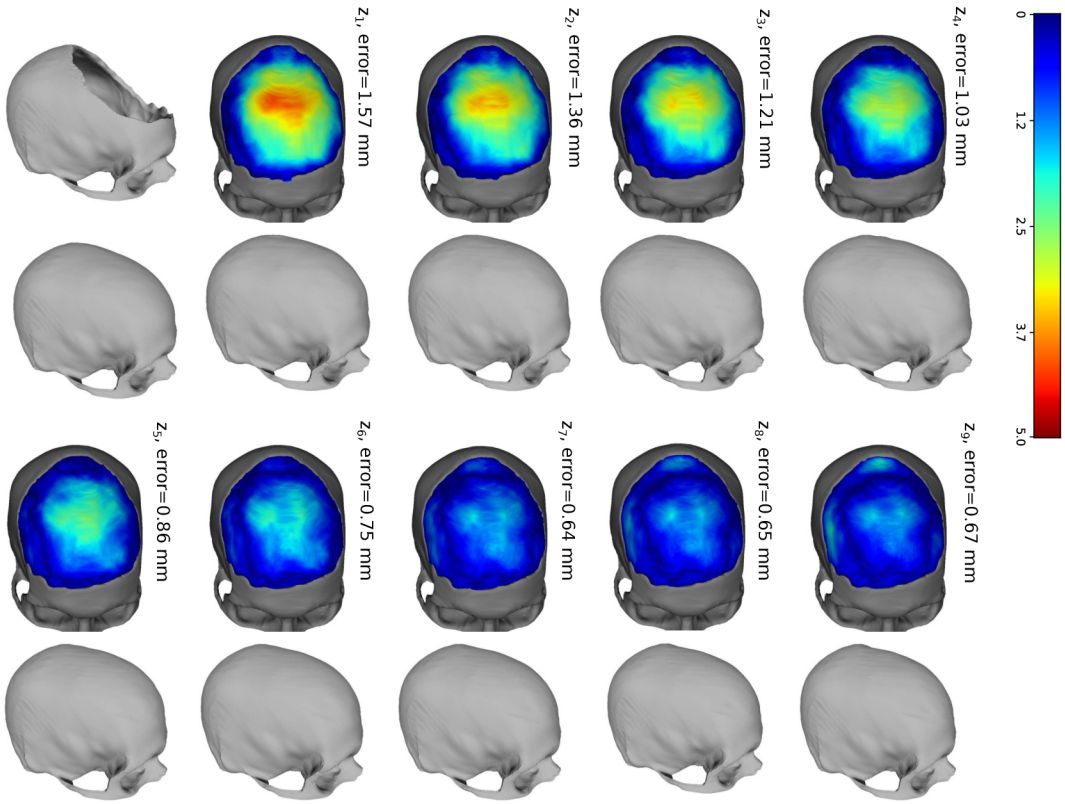

Figure 8.9: The input defective skull (first row, left) and original skull shape (second row, left) and examples of output reconstructions resulting from linear interpolation in the latent space superimposed onto the input defective skull. Surface distance from the original shape can be seen decreasing in the middle part of the defect, however, it increases in areas near the defect border where errors are unacceptable for aesthetic reasons.

erative model along with the ground-truth original shape are shown in Figure 8.9. We also reported the measured surface errors against the ground-truth shape for each sample. It can be seen that the model is able to sample from the learned manifold of solutions, allowing for manipulating the reconstructed shape while still keeping a seamless connection to the original bone.

Finally, in order to evaluate the ability of our approach to generalize, we also tested the performance of the models trained exclusively on our synthetic dataset on an internal dataset of real defective patients without any fine-tuning of the model. Both models reconstructed the real defects mostly successfully. However, there was an expected increase in the surface error in both the discriminative model output and the best-of-five generative model output. In some cases, there were also visible faults such as slight depressions or even holes as seen in Figure 8.10.

This could be partly attributed to the fact that the real testing patients come from a different geographic location, in which the anatomical variability of the skull is different [56]. Specifically, the differences in average shapes of the two datasets aligned using the same alignment method are illustrated in Figure 8.12.

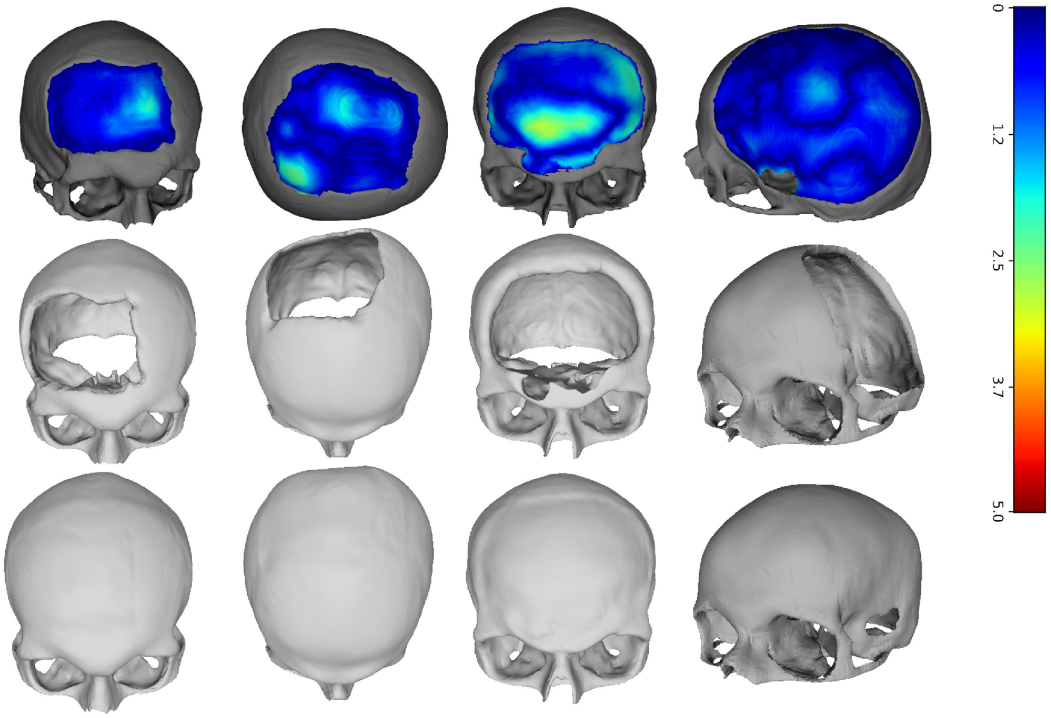

Figure 8.10: Example outputs of the discriminative model for real patient data. Although some reconstruction faults can be seen in the last two cases, suggesting that real training data of target population should be added to the model in the future, the reconstruction is usually correct. The surface distance to the original shape is well below one millimeter on average.

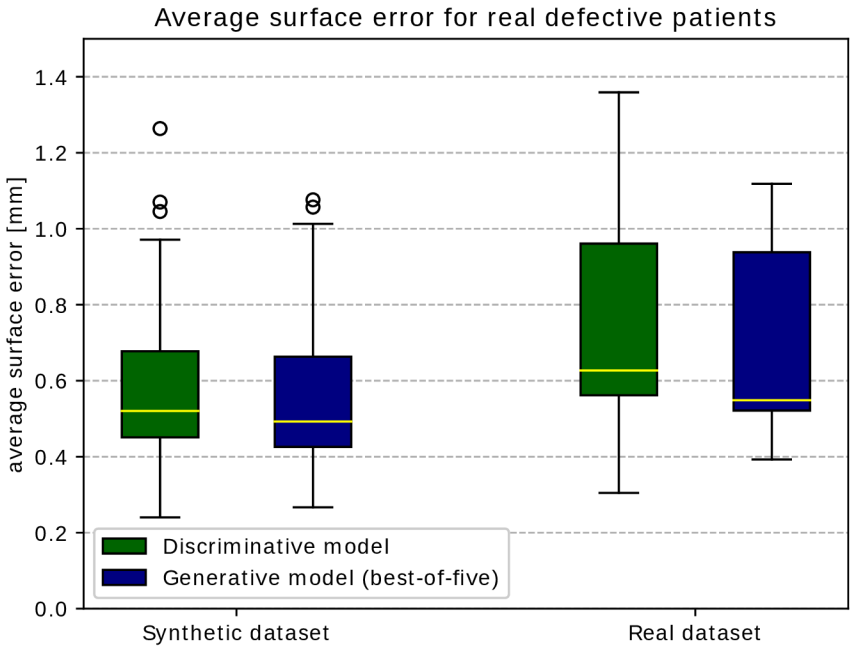

Figure 8.11: Comparison of performance of the models' average surface error [mm] for reconstruction of synthetic defects and real defects. Results shown for discriminative model (green) and for best-of-five outputs of generative model (blue).

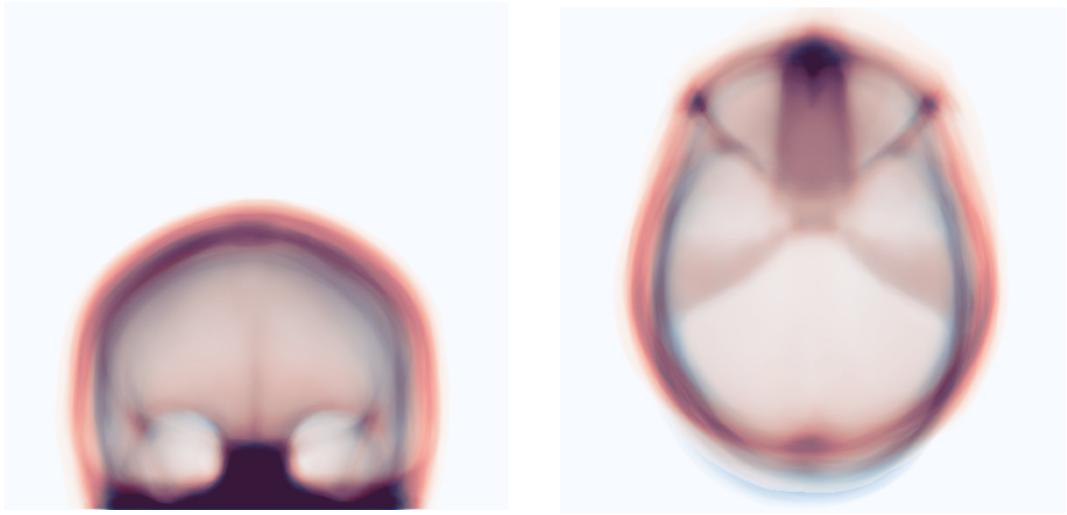

Figure 8.12: Superimposed frontal (left) and axial (right) projections of the segmented skulls. The 10 testing cases of the synthetic dataset rendered in blue and the 7 testing cases of the internal dataset in red. A difference in several shape characteristics of the skulls can be observed.

The fact that these basic shape characteristics are learned by the low-resolution reconstruction model may lead to the wrong estimation of the cranial volume in the frontal part of the skull and even holes in parts that extend significantly beyond the anatomical variability observed in the training dataset. Overall, the outcome of this experiment is encouraging, although real defective patient scans from the target population should be added to the training process before evaluating the method performance in a real clinical setting.

8.5 DISCUSSION

For the deployment of reconstruction methods into the clinical workflow, several conditions must be met. First, symmetry of the skull should be preserved as well as possible, including in cases where the patient's skull itself is partly asymmetric and where the defect reaches partly into both sides of the skull. Second, the automatic reconstruction should fit very precisely to the defect borders. Although the models presented in this work will occasionally produce a slightly asymmetric result or fail to avoid some depressions around the defect border, our results show that the proposed method can achieve an overall satisfactory performance in this regard, as illustrated by example reconstructions in Figures 8.5 and 8.10. The measured average surface errors shown in Figures 8.4 and 8.6 also show how the performance is affected by different shapes and sizes of the defects, including bilateral defects, orbital area reconstructions, and defects with a surface area of over 100 cm^2. The

implications of these results for the future implementation of the method into clinical practice should now be assessed by clinicians with experience in this area.

In the context of the current state of the art in the area of skull reconstruction, our approach differs from conventional mirroring-based and interpolation-based methods by its ability to reconstruct an arbitrary part of the skull present in the described dataset without requiring any parameter adjusting. Its ability to generalize to unseen skulls is, however, fully dependent on the variability of the training dataset used for model optimization. Figure 8.11 demonstrates how using the model on a population where shapes of the skulls come from different distribution causes occasional faults and a slight increase in the average surface error of the reconstructions. Nevertheless, this issue will be mitigated by introducing cases from the target population into the dataset and retraining the reconstruction model in the near future.

Methods based on statistical shape models also possess this dependency on training dataset variability and the potential ability to model any part of the skull. This makes them very similar to the approach proposed in this work in terms of possible target use cases. Fuessinger et al. [41] achieved an average surface error of 0.47 mm when reconstructing unilateral spherical defects of the cranial area with a radius of 5 cm. This could be compared to the performance of our discriminative and generative models reaching 0.48 and 0.46 mm average surface error, respectively, on the unilateral parietal defect group. In contrast, our method does not require any manual cleaning of the defect border as the seamless fit of the reconstructed part to the rest of the skull is handled by the CNN model. It would be interesting to see the performance of the statistical shape model on more challenging parts of the introduced dataset such as defects of the orbital area and larger bilateral defects.

A more general comparison is currently limited by the lack of standardized datasets and methodology to evaluate the anatomical reconstruction methods. Especially in the case of bilateral defects in which symmetry cannot be used to uniquely define the correct output, we argue that although the absolute distance from the ground-truth shape might give an adequate estimate of how well a method performs, it should not be used as the single criterion of correct reconstruction. In addition to variability in cranium shape, modeling structures such as skull protuberances, sutures, or uneven surface is unnecessary for means of PSI design. Therefore, the most relevant metric to measure the reconstruction method performance would be the amount of time required by the operating expert to design clinically acceptable PSI from the initial reconstruction. However, this is infeasible without the method being deployed into clinical practice.

Since the discriminative model outputs reach lower average surface error than the randomly sampled outputs of the generative model, it can be concluded that it is more suitable for a completely automatic setting. However, the generative model could alternatively be used in a semi-automatic setting. In case the initial reconstruction is not satisfactory for further processing, several subsequent samples from the generative model could be offered to the expert to increase the chance of avoiding falling back to a less efficient conventional workflow.

Finally, the reconstruction method is not limited to the skull reconstruction task or anatomical reconstruction in general. The method can potentially be applied to any shape completion task where both global contextual information, as well as fine structural details, need to be taken into account during the data volume reconstruction.

8.6 CONCLUSIONS

This work presented a multi-scale cascaded CNN architecture for general shape completion applied to the reconstruction of missing skull anatomy in a fully automatic manner. We also showed that symmetrized input can increase the performance in this task and that both discriminative and generative models can be used successfully. The proposed method reaches enough precision and robustness to be considered in clinical practice. Validation was done on a synthetic dataset that closely mimics real patient cases and this dataset was made public.

The model trained exclusively on synthetic data also performs well on real defective patient cases, but adding samples from the real target population to the training should be considered in order to improve the results. Further testing with more patient data in a clinical setting is now required to fully confirm its efficacy and identify any limitations.

Currently, the method is constrained to the cranial and orbital areas. We plan to extend the method and the dataset to include maxilla and zygomatic bones as well in the future. Adapting the method for alternative data representations, such as point clouds or graphs, could also be explored as a way to improve processing speed and precision.

9

AUTOIMPLANT CHALLENGE METHOD SUBMISSION

This chapter presents the paper "Cranial Defect Reconstruction using Cascaded CNN with Alignment" that describes a challenge submission to the MICCAI 2020 AutoImplant challenge [73]. The method reached 3rd place out of 11 participating teams in both challenge tracks. The text was edited by removing the section with related work which can be found in Section 3.2

9.1 INTRODUCTION

Craniectomy is a procedure during which a specific part of the skull is resected and eventually replaced with a cranial implant. When designing the implant, the correct skull shape reconstruction is critical for a satisfactory patient outcome. The shape of the implant should make it possible to restore the protective and aesthetic function of the skull and also fit very precisely along the border [68, 69]. A successfully reconstructed skull should be mostly indistinguishable from a healthy skull. The original skull shape before the resection is therefore often used as the golden standard of the target reconstructed shape [72].

This paper presents a BUT submission to the MICCAI 2020 AutoImplant Challenge [73]. The proposed method is an adaptation of the cascaded reconstruction CNN architecture that has been recently applied to the SkullBreak dataset [63]. Furthermore, the method is extended by an automatic landmark-based registration and a detail-preserving morphological post-processing step. In our experiments, we show how different components of the method affect the reconstruction accuracy on a validation dataset of defective skulls. Finally, we report the results on the full testing dataset of the AutoImplant Challenge.

9.2 PROPOSED METHOD

The proposed method consists of several steps as illustrated in Figure 9.1. The landmark detection step and the skull reconstruction step are handled by a 3D CNN model.

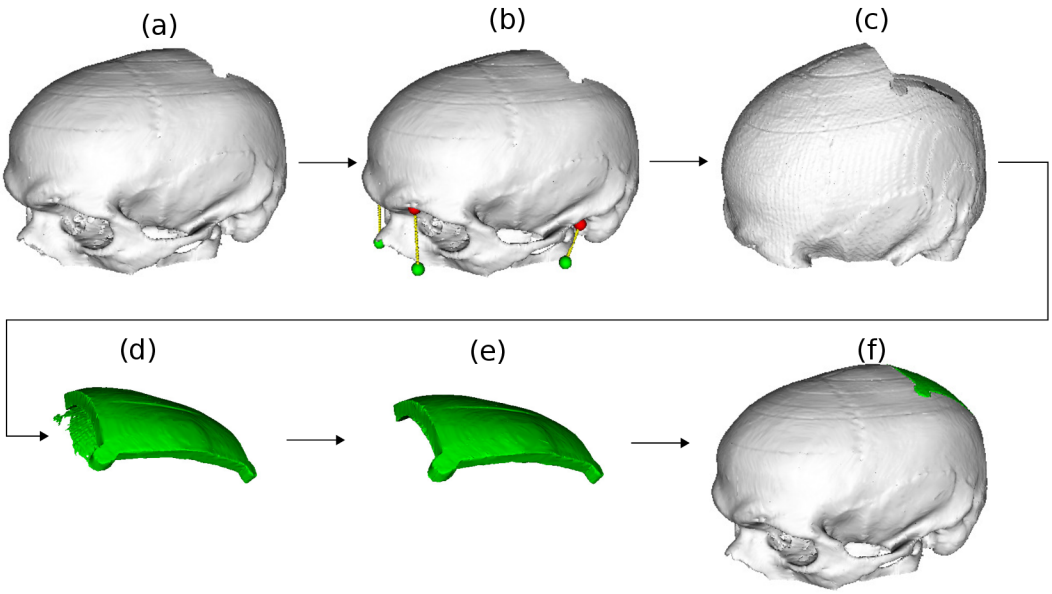

Figure 9.1: Overview of the proposed method. In the input skull volume (a) 4 landmarks are detected (b). The skull is transformed (c) so that the detected landmarks (red) are registered to the reference landmarks (green). Then, the skull is reconstructed by estimating the missing shape (d). Finally, the result is post-processed (e) and transformed back into the original skull coordinates (f).

9.3 SKULL ALIGNMENT

The defects in the AutoImplant dataset are generated on a static position inside the data volumes and the variability in their shapes and positions comes from the variability of positions of the skulls. When reconstructing the shapes with a volumetric CNN model, this introduces some difficulties. The reconstruction model needs to implicitly learn rotational and translational invariance and it also makes it cumbersome to exploit the symmetric properties of the skulls. To address this, we use scale and rigid transformations to normalize the scale and the position of the skulls.

Unlike the parameters of the scale transform that are known from the CT acquisition process, the parameters of the rigid transformation need to be inferred from the data. We use the positions of four anatomical landmarks, namely the left and right auditory meatus and left and right supraorbital notch (see Figure 9.1 b), to compute the transform. This allows us to avoid possible complications of using conventional registration methods, such as issues with substantial differences in initial positions of the data volumes and different anatomical regions present in the data.

We trained a simple 3D CNN model for landmark detection with a U-net architecture using the heatmap regression approach [103]. The detection model is illustrated in Figure 9.2 (left) and its training is further described in Section 9.6. After

detecting the landmarks, we find the rigid transform that moves these landmarks onto reference landmarks placed on the xy plane using singular value decomposition [15]. Even if one landmark is not detected either because of the detection model failure or because of a skull defect, such missing detection can usually be identified [35] and the missing landmark position can be computed from the other three landmarks.

9.4 SKULL RECONSTRUCTION

The skull reconstruction model takes the aligned binary defective skull data as an input and produces the missing part of the skull as an output. The model consists of two 3D CNNs with modified U-net architecture that are trained using the soft Dice loss. Both networks have additional max-pooling and up-sampling steps as compared to U-net to increase the field of view of the output neurons and only one convolutional layer at each resolution as shown in Figure 9.2 (right).

The first network takes a full data volume at a reduced resolution as input and produces an estimate of the missing shape with the corresponding resolution. A laterally flipped copy of the volume is also concatenated to the input of this network to facilitate easier propagation of information from one side of the skull to the other [63]. The second network takes a single patch of the original resolution input concatenated to the up-sampled patch of the low-resolution estimate at the corresponding position and produces the final missing shape estimate in this patch. Both networks are trained using their respective resolution ground-truth. Each training step comprises two updates. First, the low-resolution network weights are updated using the low-resolution ground truth. Next, both low- and high-resolution network weights are updated using the high-resolution ground truth. The patches are chosen randomly during the training. Evaluating the second network using a window sliding over all the positions in the low-resolution estimate produces the full missing shape at the original resolution.

The architecture of both of the reconstruction networks is shown in Figure 9.2 (right) and the training details can be found in Section 9.6. The reconstruction model is described in further detail by Kodym et al. [63].

9.5 SHAPE POST-PROCESSING

The reconstruction model will occasionally produce outputs that contain noise, such as disjoint objects or protuberances covering the healthy part of the skull as shown in Figure 9.3 (left). We make an assumption that the missing shapes should only consist of a single compact object. First, to isolate only the main missing

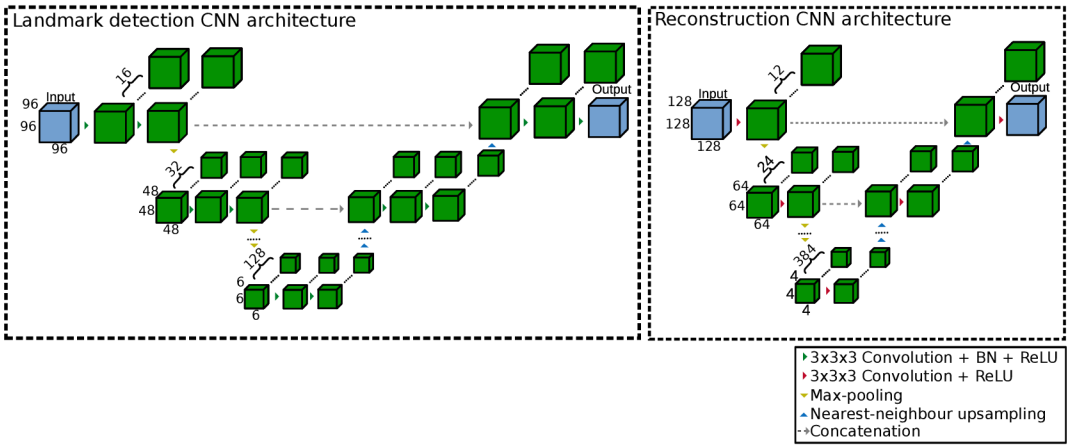

Figure 9.2: Architectures of the 3D CNN models used for landmark detection (left) and shape reconstruction (right).

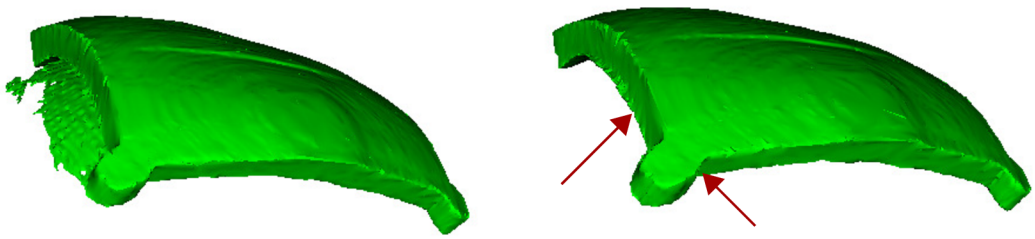

Figure 9.3: Example of the detail-preserving morphological post-processing of the estimated missing shape. Note that the undesired protuberance is removed while the fine details are preserved along the object border.

shape, we use connected component analysis and discard all objects except the largest one. Second, we use a morphological opening operation to remove any shape protuberances with less than desired minimum shape thickness.

However, the opening operation also tends to produce overly smooth shapes along the defect edges where it is desirable to keep the fine details produced by the reconstruction model. To address this, we keep both the original and morphologically open shapes. We then apply an additional morphological dilation to the open shape, producing a mask that is slightly bigger than the original shape but does not include the protuberances. Masking the original shape with such a mask results in a shape with the original fine details but without the larger protuberances as shown in Figure 9.3 (right).

9.6 EXPERIMENTS

In this section, we describe the experiments and show the effect of individual method components on the reconstruction outputs. All the experiments were run on a system with Titan Xp GPU with 12GB GRAM.

9.6.1 *Landmark Detection*

We manually annotated the four landmarks in all 100 training skull volumes. We trained the landmark detection CNN model on 90 samples, leaving 10 skulls for validation. The model was trained for $100,000$ iterations using Adam optimizer with training step 10^{-4} and the dataset was strongly augmented using random rotations to ascertain that the model is able to detect the landmarks in cases of arbitrary patient positions inside the scanner.

The results of the landmark detection on the 10 validation cases can be seen in Figure 9.4 (left). The auditory meatus landmarks were detected with an error of 1.22 ± 0.70 mm while the supraorbital notch landmarks achieved a slightly higher error of 1.84 ± 1.03 mm. An important observation is that the trained model also succeeded in the detection of all four landmarks in all the 110 testing cases as well, and every skull could be aligned fully automatically without any manual intervention at test time.

9.6.2 *Missing Shape Inference*

Similar to the landmark detection model, the reconstruction networks were also trained on 90 training samples. For the ablation experiments in this work, both low- and high-resolution networks were trained on batches of 4 samples using Adam optimizer with training step 10^{-4} for $50,000$ iterations using a resolution of 3.2 mm per voxel and 0.4 mm per voxel, respectively. All data volumes were padded to dimensions $512 \times 512 \times 512$ which means that the corresponding low-resolution samples had dimensions $64 \times 64 \times 64$. Random lateral flips were used to augment the dataset.

We trained three different reconstruction models. The *basic cascade model* is trained on the originally provided challenge data. The mirrored input channel is not used in the low-resolution network of this model as the sagittal plane is not known. The *aligned model* is trained on the data that have been previously aligned using the detected landmark positions. This also allows us to use the mirrored channel in this model. The *aligned and augmented* model is also trained on additional defective skulls that have been created from the training complete skulls.

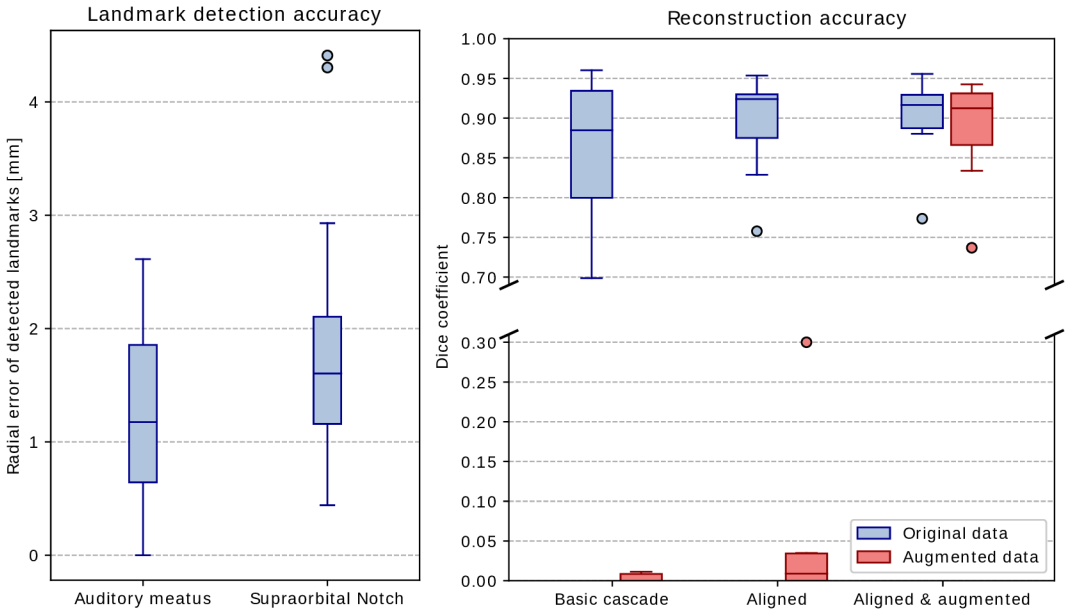

Figure 9.4: Accuracy of the landmark detection (left) and the reconstruction models (right) on the validation cases.

Five defects were created on each skull using random shapes, similar to the SkullBreak dataset, resulting in additional 450 training cases. We also created 10 additional validation cases using the same process.

The results of the reconstruction model on the validation cases are shown in Figure 9.4 (right). The basic cascade model had the worst performance on the validation cases, achieving an average Dice score of 0.835. Simply aligning the data and adding the mirrored input to the low-resolution network in the aligned model had a substantial effect on the model performance, reaching 0.895 Dice score and showing the benefit of reducing the degrees of freedom of the defects during the reconstruction. However, both models overfit strongly to the training dataset with specific shape and position of the defects and were unable to generalize to the additional augmented validation cases where the distribution of defect shapes and positions is different. The aligned and augmented model trained on the additional defective cases, on the other hand, was able to both reconstruct the additional validation cases and increase the original data accuracy to a Dice score of 0.903.

9.7 RESULTS

We aligned both subsets of the final 110 test cases of the AutoImplant challenge using the landmark detection model. For reconstruction, we used the aligned and augmented model that had been trained for 120,000 iterations. We also increased the first reconstruction network resolution to 1.6 mm per voxel, resulting in low-resolution volumes of dimensions $128 \times 128 \times 128$ voxels in the final evaluated

Table 9.1: The results of the proposed method on the AutoImplant Challenge test dataset in terms of Dice score and Hausdorff distance.

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

model. To discard the occasional artifacts, we used the post-processing method described in Section 9.5. Both standard and additional subsets of the test dataset were reconstructed completely automatically without any manual interactions. The landmark detection model, the aligned training dataset, and the augmented training dataset are publicly available[1].

The results of the proposed method on the challenge test dataset in terms of Dice coefficient and Hausdorff distance are shown in Table 9.1. Several qualitative examples of the reconstruction output on the standard subset, the additional subset, and also the augmented validation dataset are shown in Figure 9.5 where one case of reconstruction failure on the additional test set can also be observed.

9.8 CONCLUSION

Our experiments showed that the skull alignment and data augmentation techniques we used increased the accuracy of the skull reconstruction. These are general concepts that could be applied to any other reconstruction model. Although we only encountered one failure case in our experiments, it hints at the fact that more defect shape augmentations should be used to increase the robustness of the reconstruction model. It is currently unknown whether the achieved accuracy in terms of the Dice coefficient and Hausdorff distance could warrant clinical applicability of the method. However, visual inspection of the reconstructed defects shows no visible artifacts in most cases.

While the reconstruction method reaches good accuracy, the final shape will usually have to be further edited by an experienced clinician in medical practice. Therefore, it would be beneficial to explore ways to include interactivity in the implant design method, possibly drawing inspiration from interactive convolutional networks that have been successfully applied to segmentation tasks. Another interesting research direction is leveraging different data representations such as point clouds or level sets.

[1] https://github.com/OldaKodym/BUT_autoimplant_public

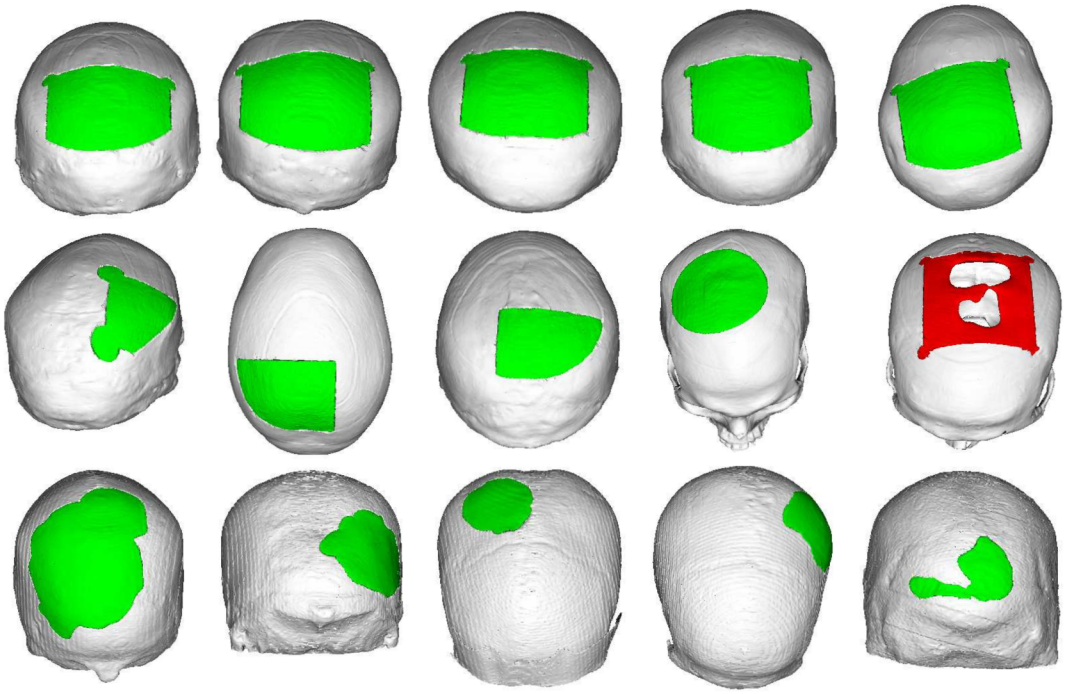

Figure 9.5: Examples of the reconstruction results. From top row to bottom: The standard test set, the additional test set and the augmented validation set. Reconstruction failure could be observed in the last case of the additional test set in red color.

10

GOING FROM SYNTHETIC TO REAL PATIENT DATA

This chapter presents the article "Deep Learning for Cranioplasty in Clinical Practice: Going from Synthetic to Real Patient Data" currently in the review process of the Computers in Biology journal. The work deals with the reconstruction of real patient data and with the design of directly printable cranial implants.

10.1 INTRODUCTION

Cranioplasty is a procedure that restores the aesthetic, mechanical, and protective function of a defective skull by implanting material into the defect area. Although autologous bone or titanium meshes can be used as implants, 3D printable implants have been shown to be more versatile and to have several other advantages, such as lower risk of complications or lower chance of requiring a secondary surgery [44, 32]. Manufacturing these implants requires modeling their shape in computer-assisted design (CAD) software as the first step. This virtual reconstruction, however, requires the human operator to have sufficient knowledge of skull anatomy as well as skill in 3D modeling. Even if these requirements are met, correctly modeling the implant is time-consuming even for a skilled operator, especially in cases of defects reaching into both lateral sides of the skull [25]. Automatically producing fast and precise estimations of the implant shapes could therefore mean increased standardization and efficiency of cranioplasty clinical workflow.

Most recent (semi-)automatic skull reconstruction methods aim to solve the task of finding the exact shape of the missing part of the skull. We refer to this type of reconstruction output as *a skull patch* in this article. The main criteria for a successful skull patch estimation is an anatomically plausible, symmetric shape with a smooth and seamless fit along the defect border. In clinical practice, this allows the operator to use the estimated skull patch as a template for the final cranial implant design in CAD software. Conventional skull reconstruction methods use mirroring the healthy side of the skull onto the defective side [23], surface interpolations [24, 127] or their combination [88] to estimate the skull patch. Statistical shape models [89] greatly expanded the range of skull defects that can be reconstructed automatically [112, 41, 42]. In recent years, the research focus shifted to volumetric convolutional neural networks (CNNs) which have shown great promise in fast and robust skull patch reconstruction[98, 63, 91] and became the method of

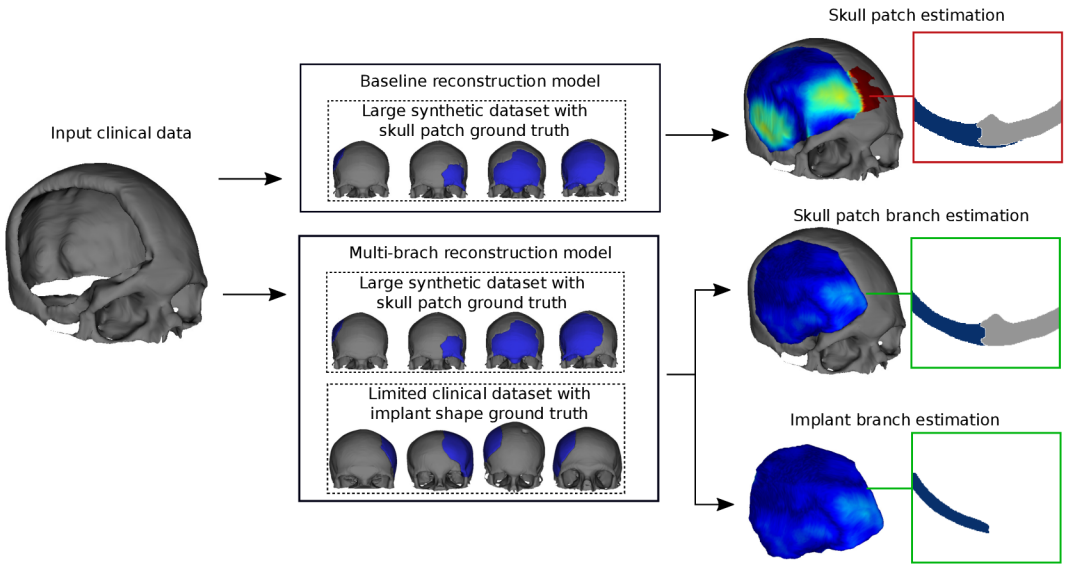

Figure 10.1: The proposed multi-branch architecture makes use of multi-task learning on different skull reconstruction datasets. In addition to higher overall accuracy and ability to directly output cranial implant shapes, the skull patch output of the multi-branch model also better fits the shape to complex defect borders in real clinical data.

choice in the 2020 AutoImplant challenge [73]. The CNN-based methods are usually trained and evaluated using synthetic defects created by removing some part of healthy, resulting in a virtually infinite amount of different samples.

The final shape of the cranial implant (referred to simply as *implant* in the remainder of this article) differs from the shape of the skull patch in several ways (see Figure 10.2). The implants have a constant thickness different from the original bone and have some spatial tolerance along the defect border to account for scar tissue and continuing bone growth, ensuring implantability. The shape of the implant can also be estimated directly by a CNN model, provided that sufficient training data is available for training. Although it is more difficult to edit this kind of shape in CAD software due to fine details along the defect border, it has the potential to be used in a fully automatic setting when no human operator or not enough time for manual design is available, for example in intra-operative rapid manufacturing of cranial implants [128].

Synthetic datasets for automatic estimation of skull patches recently became available because they are easy to create from public databases of healthy skulls, such as CQ500 [28]. However, they do not necessarily fully cover the defective skull shape distribution of target clinical data (i. e. different anatomical variability of the target population, defect shapes and sizes, complex morphology of defect border), which may affect the resulting reconstruction quality in practice [63]. Real clinical data with expert-designed implant models are, on the other hand, difficult to obtain. Furthermore, in our experience, the distribution of available clinical data

is often biased towards simple uni-lateral defect cases and not easily extendable by synthetic defect and implant shapes. The more challenging bilateral and fronto-orbital defects are less common, yet it is in these challenging cases where correct automatic skull patch reconstruction or implant design can have the largest impact on clinical practice.

The main contribution of this article is proposing a multi-branch CNN architecture capable of learning from the two types of cranioplasty data simultaneously. Given a defective skull, the model estimates the shapes of both the skull patch and the implant during inference, as illustrated in Figure 10.1. We compare the proposed multi-branch model to two baseline models trained on synthetic-only and clinical-only datasets and show that both outputs of the proposed multi-branch model reach comparatively higher quality and contain a lower amount of artifacts, which is crucial for deployment in clinical practice. Although a similar effect could possibly be achieved by collecting a vast amount of well-distributed clinical data or by perfectly matching their distribution by meticulously tailoring synthetic data, we believe that the proposed approach of combining a large amount of imperfect synthetic data and a limited set of target clinical data is generally simpler and easily extendable to different types of cranioplasty data (i. e. different population, additional defect areas such as the orbital floor or zygomatic bone, different preferences for the final implant shape). In addition to quantifying reconstruction quality using common binary shape reconstruction metrics such as Dice coefficient and surface distance, we also propose a new metric based on Gaussian curvature to quantify surface imperfections that have a direct impact on aesthetic patient outcome.

10.2 MATERIALS AND METHODS

10.2.1 *Datasets*

We use two different cranioplasty datasets in this work. The SkullBreak dataset [64] is a synthetic skull shape reconstruction dataset adapted from the CQ500 public database of head CT scans [28]. The CT scans were rigidly aligned and segmented to provide normalized shapes of healthy skulls. Then, synthetic defects were created by subtracting random shapes from several regions in each skull. Morphological operations were additionally used to mimic some degree of bone healing processes along the defect borders. The dataset contains 570 training and 100 testing pairs of defective skulls and corresponding skull patches.

The second, in-house dataset was provided by the TESCAN Medical company. It contains a total of 387 real patient cases indicated for cranioplasty. Each pa-

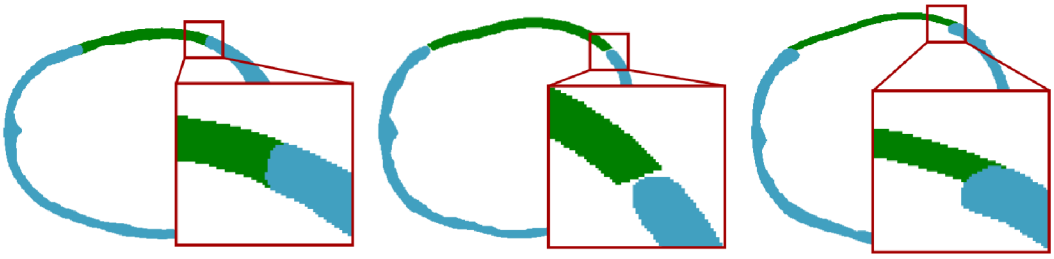

Figure 10.2: Axial slices through samples from the datasets used in this work. From left to right: skull patch sample from a synthetic dataset, manually designed implant shape sample from an in-house clinical dataset, manually designed skull patch sample from an in-house clinical dataset.

tient case consists of CT data with manual skull segmentation and a mesh model corresponding to an expert-designed cranial implant. 75 of these cases additionally contain expert-designed mesh models of patches covering the full area of the defects that were used as the initial template for the final implant design by the expert. Although these expert-designed patches have a different thickness from the original bone, their outer surface can be used as a reference for the outer surface of automatically reconstructed patches. This naturally led us to split the in-house dataset correspondingly into 312 training cases and 75 test cases, ensuring that a real clinical test set of reasonable size is available for evaluation of both the skull patch shape estimation and the final implant shape estimation tasks. All implant and patch mesh models in the clinical in-house dataset were rasterized into voxel grids and the data were rigidly aligned to conform with the SkullBreak data. Several examples from all datasets can be seen in Figure 10.2.

The two datasets also differ in several more aspects. Because they come from geographically distant sources, the average size and the anatomical variability of the skulls differ [106]. The scale and positional variability of the defects is also different. While the Skullbreak dataset was created specifically to contain a balanced amount of unilateral, bilateral, and fronto-orbital defects, the clinical in-house dataset contains a higher amount of uni-lateral defects with larger sizes and reaching farther into lower parts of temporal and sphenoid bones. Although some of these differences could be addressed by tailoring the synthetic defects in the Skullbreak dataset to fit the distribution of clinical data more closely, some aspects such as skull shape variation and defect border complexity cannot be precisely emulated.

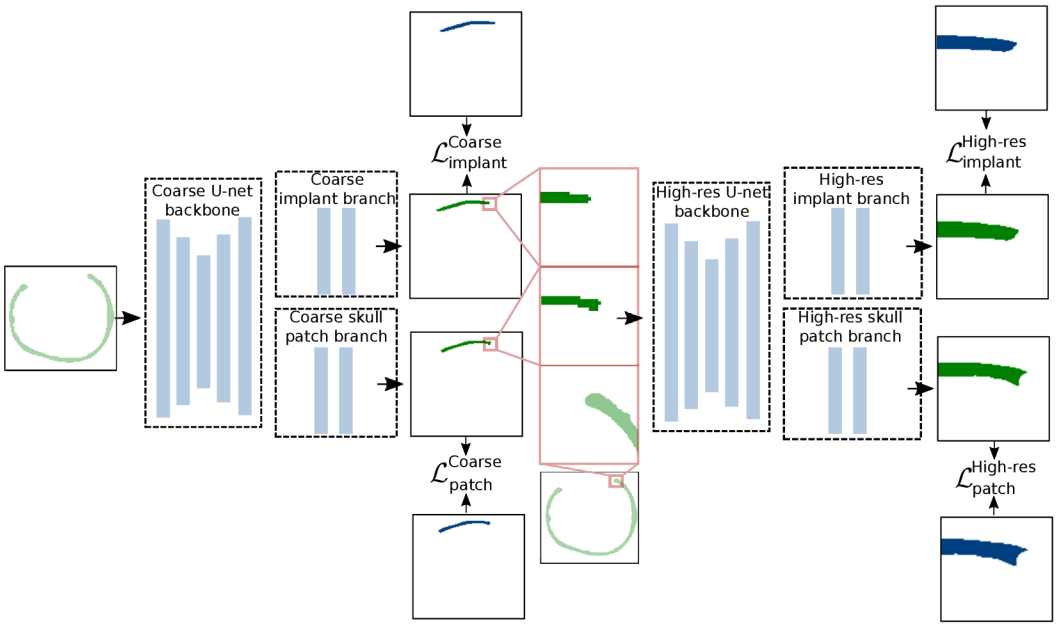

Figure 10.3: Illustration of the multi-branch CNN cascade training process. Inputs and outputs of the network in light and dark green colors, respectively, and ground-truth shapes in blue. In each training step, the coarse network weights are first updated using the sum of the coarse losses, and then both coarse and high-resolution network weights are updated using the sum of the high-resolution losses.

10.2.2 *Baseline CNN Models for Shape Estimation*

We use the same baseline reconstruction method for both the skull patch estimation and the implant estimation tasks, with the only difference being the data used for training. The method is based on a cascade of two U-net-like volumetric CNNs proposed in our previous work [63]. The first, coarse CNN takes a binary shape of the defective skull in coarse resolution and produces an initial output shape estimate with the same resolution. The second, high-resolution CNN then takes a single crop of upscaled coarse shape estimate and corresponding crop of the high-resolution defective skull and produces a high-resolution shape estimate of that crop, effectively performing super-resolution of the coarse shape estimate locally conditioned on the high-resolution defective skull. The coarse CNN model additionally uses a mirrored copy of the input volume, which was shown to improve lateral symmetry of output shapes [63].

We use 12 initial feature channels and an input volume size of $128 \times 128 \times 128$ for both the coarse and high-resolution CNNs. The final output is created by first inferring the coarse shape estimate and then inferring the high-resolution CNN in a sliding window manner. Both the original input and the final high-resolution output volumes have a size of $512 \times 512 \times 512$ voxels with a resolution of 0.4 mm per voxel. We train the CNN cascade for $300,000$ steps on mini-batches of size 4

using the soft Dice loss [94]. Each training step consists of propagating gradients of loss computed on coarse resolution through the coarse resolution CNN and propagating gradients of loss computed on one high-resolution crop through both the coarse and high-resolution CNNs. More details about the CNN architecture and training procedure can be found in the original work [63].

10.2.3 *Multi-branch CNN Model for Joint Shapes Estimation*

To facilitate training of both tasks on the two datasets simultaneously using a single CNN cascade model, we make the following changes to the baseline architecture. For the coarse CNN, instead of producing the output directly from the last layer of the U-net-like backbone, we branch the network into two separate shape estimation branches. The skull patch estimation branch is trained to output coarse shape estimates of the skull patch using the training samples from the SkullBreak dataset. Conversely, the implant estimation branch is trained to output coarse estimation of the shape of implants using the training samples from the in-house dataset. The two outputs are concatenated and used as an input for the high-resolution CNN. The high-resolution CNN output is again split into skull patch estimation and implant estimation branches trained on the corresponding dataset samples.

Each shape estimation branch is formed by a single conv-ReLU-conv-softmax block, with both convolutional layers having the same number of features as the last U-net layer. This means that while the shape estimation branches allow two slightly different outputs to be estimated by the CNN, the shared U-net-like backbone is forced to learn to extract meaningful local features that are suitable for correct shape estimation on both datasets.

During the training of the multi-branch CNN cascade, we use mixed mini-batches containing two samples from the Skullbreak dataset and two samples from the in-house dataset. Accordingly, two loss components are computed for both the coarse and high-resolution CNNs: one for the skull patch estimation branch output using the SkullBreak samples and one for the implant estimation branch output using the in-house dataset samples. These loss components are then added together before updating the according CNN weights. The multi-branch model overview is shown in Figure 10.3.

10.2.4 *Metrics*

For the sake of the quantitative evaluation, we assume that the expert-designed shapes in the test set represent the only correct solution to the shape estimation tasks. This means that the quality of the output can be quantified using segmen-

tation metrics such as volumetric overlaps (i.e. Dice coefficient) and surface distance [135]. However, note that the shape reconstruction task is specific in allowing some variability in the reconstructed shape in some cases, as long as there are no imperfections along the fit of the reconstructed shape to the input shape. See Section 11 for an illustration of how different segmentation metrics correlate with a subjective quality score of an expert implant designer. For these reasons, we evaluate the automatic reconstruction outputs using multiple different metrics in this work.

In the case of implant shape evaluation, we use the Dice coefficient and average surface distance for quantification of the estimated implant shape quality, similarly to recent relevant works [41, 91, 72]. In the case of skull patch shape evaluation, however, the expert-designed ground-truth patches and model outputs have different characteristics and this prevents us from using these metrics directly (see Figure 10.2). Because the thickness of the ground-truth patch is different from the thickness of the original bone in the SkullBreak dataset, we measure average surface error only at outer surface voxels of the skull.

We pay special attention to the quality of fit along the defect border of the skull patches. Similarly to other authors [41], we report the outer surface distance computed along the defect border. However, this metric may not precisely convey some types of common errors of skull reconstruction which have impact on the aesthetic outcome of cranioplasty, such as slight trenches or bumps on the surface along the defect border. To this end, we compare approximate Gaussian curvatures of reconstructed skulls and reference skulls along the defect border to supply this information.

Gaussian curvature is routinely used in 3D model surface analysis literature [134]. For simplification, we chose to approximate the Gaussian curvature error of the reconstructed skull shapes by first smoothing the binary images of skull shapes with a Gaussian blur with $\sigma = 5$, then normalizing back to range between 0 and 1 and computing the Gaussian curvature K_i at each voxel i using the following equation:

$$K_i = -\frac{\begin{vmatrix} H_i(F) & \nabla F_i^T \\ \nabla F_i & 0 \end{vmatrix}}{|\nabla F_i|^4} \tag{10.1}$$

where F is the blurred volume, ∇F is its gradient and H is Hessian matrix. The resulting Gaussian curvature volumes are then compared directly by computing voxel-wise squared error and we report the mean of this error computed along the defect border voxels as

$$MSE_K = \frac{1}{N_B} \sum_{i \in B} (K_i^{ref} - K_i^{pred})^2 \tag{10.2}$$

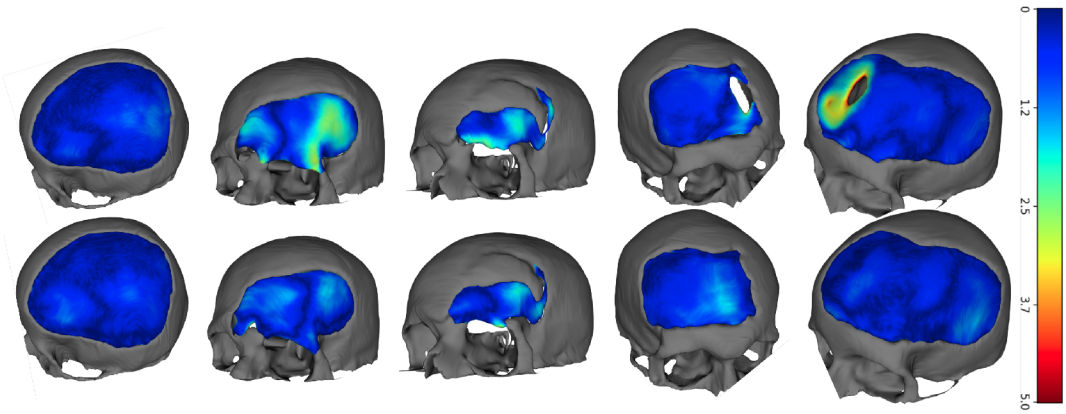

Figure 10.4: Implant estimates of the baseline implant model (top) and the multi-branch model (bottom).

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

10.3 RESULTS

The baseline implant model was trained using the 312 training implant shapes from the in-house dataset and the baseline skull patch model was trained using the 570 Skullbreak training data samples. Because we noticed that the average size of the SkullBreak skulls differs from the average size of the in-house test skulls, we trained another baseline skull patch model on a modified version of the SkullBreak dataset that was rescaled to match the average height, length, and breadth of the in-house skulls. The multi-branch model was trained using a combination of the in-house and the rescaled Skullbreak dataset. Outputs of all models were morphologically denoised by removing smaller connected components and shape artifacts [62] before comparing them to the reference expert-designed shapes in the in-house test set.

10.3.1 *Implant Shape Estimation Performance*

The implant shapes produced by the baseline implant model reached an average Dice coefficient of 0.85 ± 0.10 and average surface error of 0.77 ± 0.44 mm, confirming that it is possible to learn the direct mapping of defective skull shapes to the final cranial implant shapes using the CNN cascade. However, because central and

fronto-orbital defects are not well represented in the in-house training dataset, the baseline implant model fails to correctly estimate implant shapes in these cases, as shown in Figure 10.4. This issue may be amplified by the fact that, due to the data imbalance, the coarse CNN model learns to rely too much on the mirrored input to provide initial information about the missing shape due to overfitting, leading to the inability to correctly deal with bilateral defects.

The implant estimates of the multi-branch model reached an average Dice coefficient of 0.88 ± 0.07 and an average surface error of 0.65 ± 0.33 mm, showing an increase in accuracy and decreased variability of output shape quality. Closer inspection of the outputs reveals increased success rate of bilateral and fronto-orbital implant shape estimation. This can be attributed to better generalization of the U-net backbone which needs to account for more diverse defect positions in the SkullBreak dataset. Several example implant shape estimates from both the baseline implant model and the multi-head model implant estimation branch are shown in Figure 10.4. The distribution of Dice coefficients and average surface distances achieved by both models can be found in Figure 10.7.

10.3.2 *Skull Patch Estimation Performance*

The skull patches produced by the baseline skull patch model trained on the original SkullBreak data resulted in an average outer surface error of 0.98 ± 0.45 mm on the in-house test set. Rescaling the Skullbreak training skulls to match the average size of the in-house skulls decreased the error by 15% to 0.83 ± 0.38 mm, supporting the hypotheses that the model learns the average skull shape of the training data. However, the skull patch estimates still produced shapes with high surface error and occasional artifacts such as holes and uneven surfaces, especially in cases of large defects. One of the causes may be the fact that the defects in the SkullBreak dataset do not fully cover the lower areas of the skull. This could be addressed by extending the dataset with additional synthetic defects, but Figures 10.5 and 10.6 show that there are multiple different sources of error.

The skull patch estimates produced by the multi-branch CNN model further decreased the average surface error to 0.67 ± 0.37 mm. In addition to a lower amount of visible holes and artifacts in the estimated shapes, the multi-branch model also predicted the skull patches with overall lower outer surface distance from the reference expert-designed patches, as shown in Figure 10.5. The distributions of all error metrics for the three models are shown in Figure 10.8.

Interestingly, the multi-branch model output also reached a lower defect border surface error of 0.75 mm, compared to 0.96 mm and 0.94 mm for the baseline models trained on the original and the rescaled SkullBreak dataset, respectively.

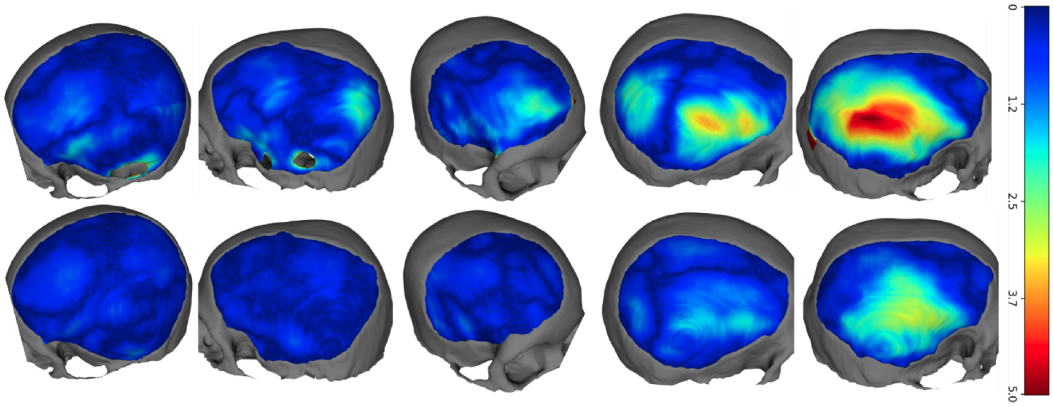

Figure 10.5: Estimated skull patches of the baseline skull patch model (top) and the multi-branch model (bottom).

Similarly, the Gaussian curvature errors of the baseline skull patch model trained on the original Skullbreak and on the rescaled SkullBreak datasets also did not differ significantly, but the curvature error decreased by around 12% in the case of the multi-branch model skull patch branch outputs. This suggests that the multi-branch model learned to better fit the reconstructed skull patches to the more complex borders of the in-house defective skulls, despite only seeing the corresponding implant shapes with spatial tolerance along the border during training (see Figure 10.2). Figure 10.6 shows how the Gaussian curvature error reacts to different types of surface errors compared to the distance-based metrics, helping to visually identify problematic regions of the skull patch shape reconstruction outputs.

10.4 DISCUSSION AND CONCLUSIONS

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

We showed that when training the model on real clinical patient data, synthetic data can be effectively leveraged using the proposed multi-branch CNN model to

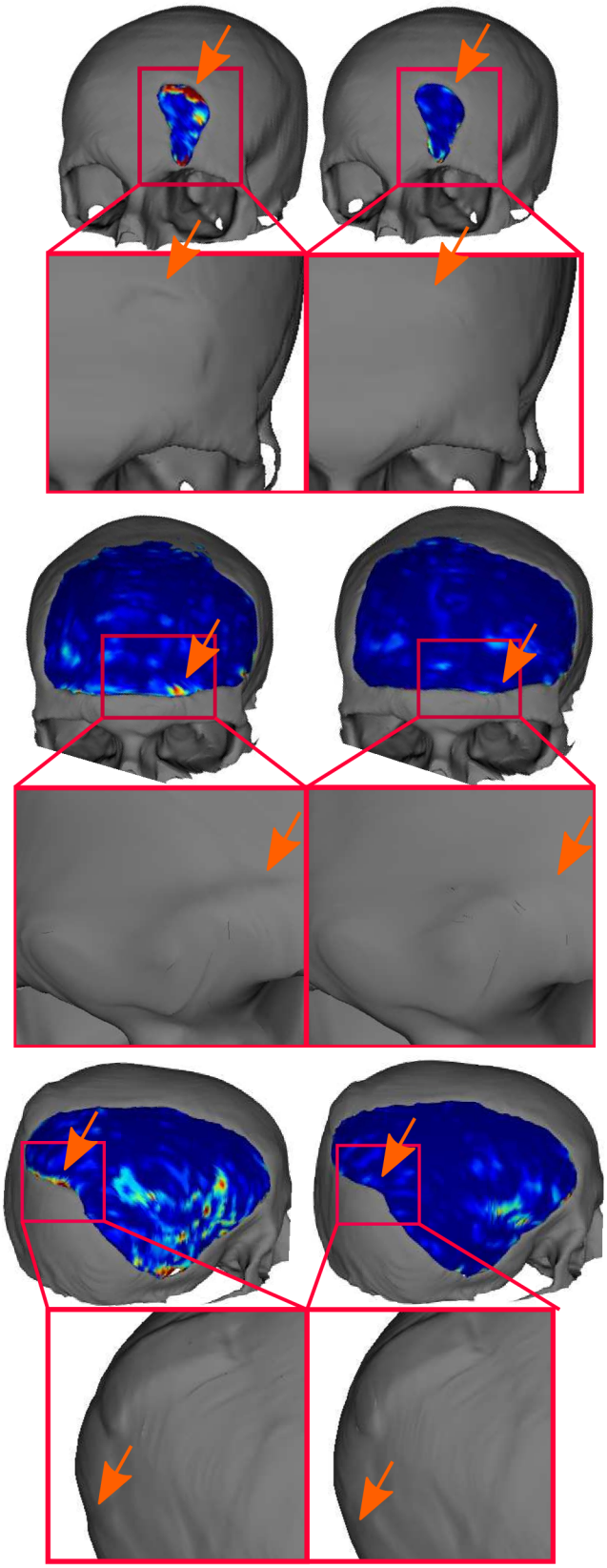

Figure 10.6: Three example pairs of baseline skull patch model outputs and a multi-branch model skull patch outputs, respectively, with color-coded Gaussian curvature error. The 3D models were rendered using marching cubes algorithm and post-processed using quadratic decimation and normal smoothing. The multi-branch model can produce smoother results with lower curvature error.

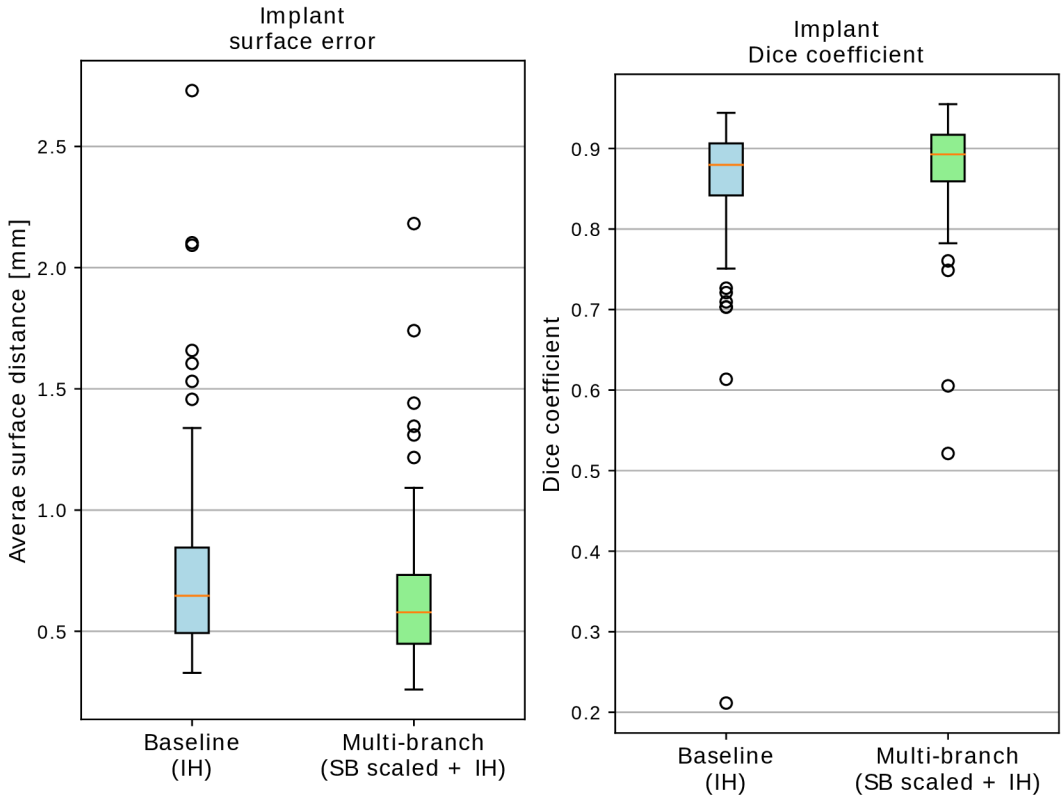

Figure 10.7: Boxplots showing the error distributions of the evaluated models trained on corresponding datasets (Skullbreak - SB, in-house - IH). Average surface error and Dice coefficient of the implant estimates.

significantly improve the model performance and compensate for common issues of clinical patient datasets (i. e. data scarcity and imbalance). The error of the outer surface of reconstructed skulls achieved by the proposed method is higher than some other recent works evaluated on synthetic defects [63, 41]. However, we believe that factors such as the higher average area of the defects in our test set may be the cause and that the results are overall very promising.

The synthetic and clinical datasets used in this work contained different types of ground truths: the original missing skull patch shape and final cranial implant shapes. This allowed us to automatically produce 3D printable and directly implantable shapes for the first time, although this use case will require further evaluation of the clinical applicability in cooperation with experienced implant designers. More importantly, the general ability of the model to combine cranioplasty data from different sources and of different types can accelerate the adoption of the automatic reconstruction methods by allowing training on specific target datasets while exploiting the advantages of available synthetic datasets.

To our best knowledge, this was the first study that evaluated CNN-based skull reconstruction on a real clinical dataset of this size. The proposed multi-branch CNN cascade increased the reconstructed shape quality by allowing training on more data when compared to the individual baseline models. Although the results

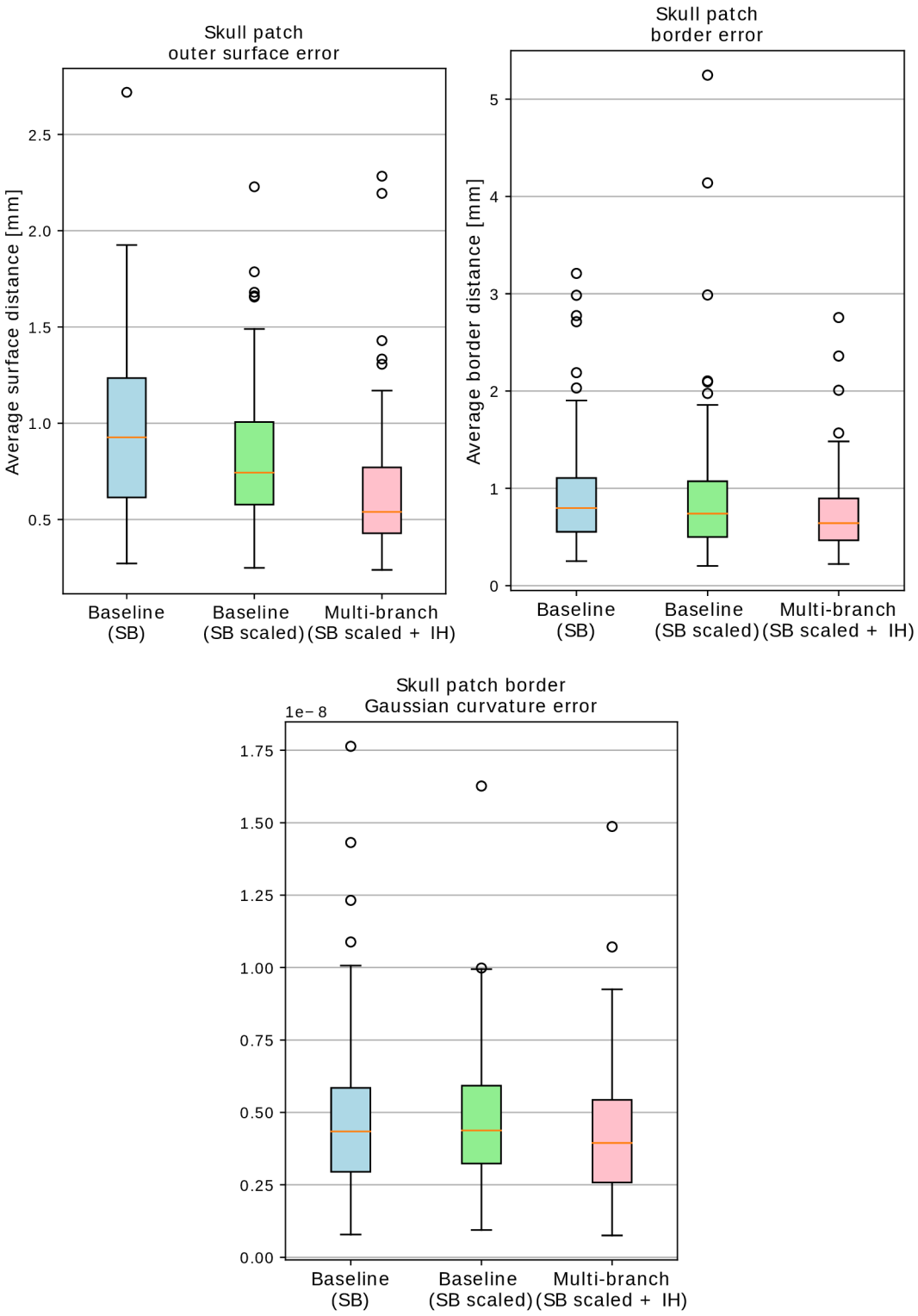

Figure 10.8: Boxplots showing the error distributions of the evaluated models trained on corresponding datasets (Skullbreak - SB, in-house - IH). Average outer surface error and border error (top) and Gaussian curvature error of the skull patch estimates (bottom).

of this study are promising from a quantitative perspective, they will next need to be evaluated by an experienced implant designer to ascertain their clinical value.

11

CORRELATION ANALYSIS OF QUANTITATIVE RECONSTRUCTION METRICS

This chapter illustrates how well can the quantitative segmentation metrics predict usability of automatic skull reconstruction results in clinical practice. It is currently under review in the Computers in Biology journal in form of an appendix to the article presented in the previous section.

We created a dataset of automatically reconstructed defective skulls and submitted it to an expert with experience in the field of skull reconstruction and implant design for subjective quality evaluation. Comparing these subjective expert scores with metrics of similarity between the reconstructed and the original shape can give an idea of what to look for when evaluating the reconstructions.

11.1 SKULL DATA AND RECONSTRUCTION

The skull data come from the SkullBreak and SkullFix datasets [64], so the ground truth original shapes are available. A CNN-based reconstruction of the missing shape [63] was performed on each skull. Because for this analysis we would ideally want to cover the whole quality spectrum from bad reconstructions to very good reconstructions, we included the following types of reconstructed cases:

- SkullFix test case reconstructions
- SkullFix additional test case reconstructions
- SkullBreak test case reconstructions
- SkullBreak training case reconstructions (to include several close-to-perfect reconstructions)
- SkullBreak test case reconstructions using generative model [63] (to include multiple different reconstructions for a single case, including visibly bad ones)

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

11.2 GLOBAL METRICS

We first computed correlation coefficients between the subjective expert score and routinely used segmentation metrics, including volumetric Dice coefficient and average symmetric surface distance. We also included the surface distance computed at the outer surface of the skull, since it is the most important aspect for subsequent implant modeling steps [85]. The outer surface was used in the evaluation of some previous works [41] and we used it in this work because of the shape characteristics of the ground truth data.

Figure 11.1 shows that these global metrics correlate with the expert subjective score with correlation coefficients around 0.6, confirming that they are appropriate for the comparison of different reconstruction methods. However, it can be noted that their correlation is weak in regions of higher subjective expert scores, making it impossible to use them for discrimination between good and perfect results. Also, several cases satisfy the quantitative metrics while being seen as low-quality by the expert and vice versa (see cases highlighted in red in Figure 11.1).

11.3 DEFECT BORDER METRICS

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

Figure 11.3 shows that both of these metrics correlate with the subjective expert score similarly or slightly more than the global metrics. Most importantly, it can be seen that the border metrics indeed convey different information. Although the quantitative border metrics also do not always agree with the subjective quality score, the correlation with the expert score was higher in the cases where the correlation of the global metrics was low.

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

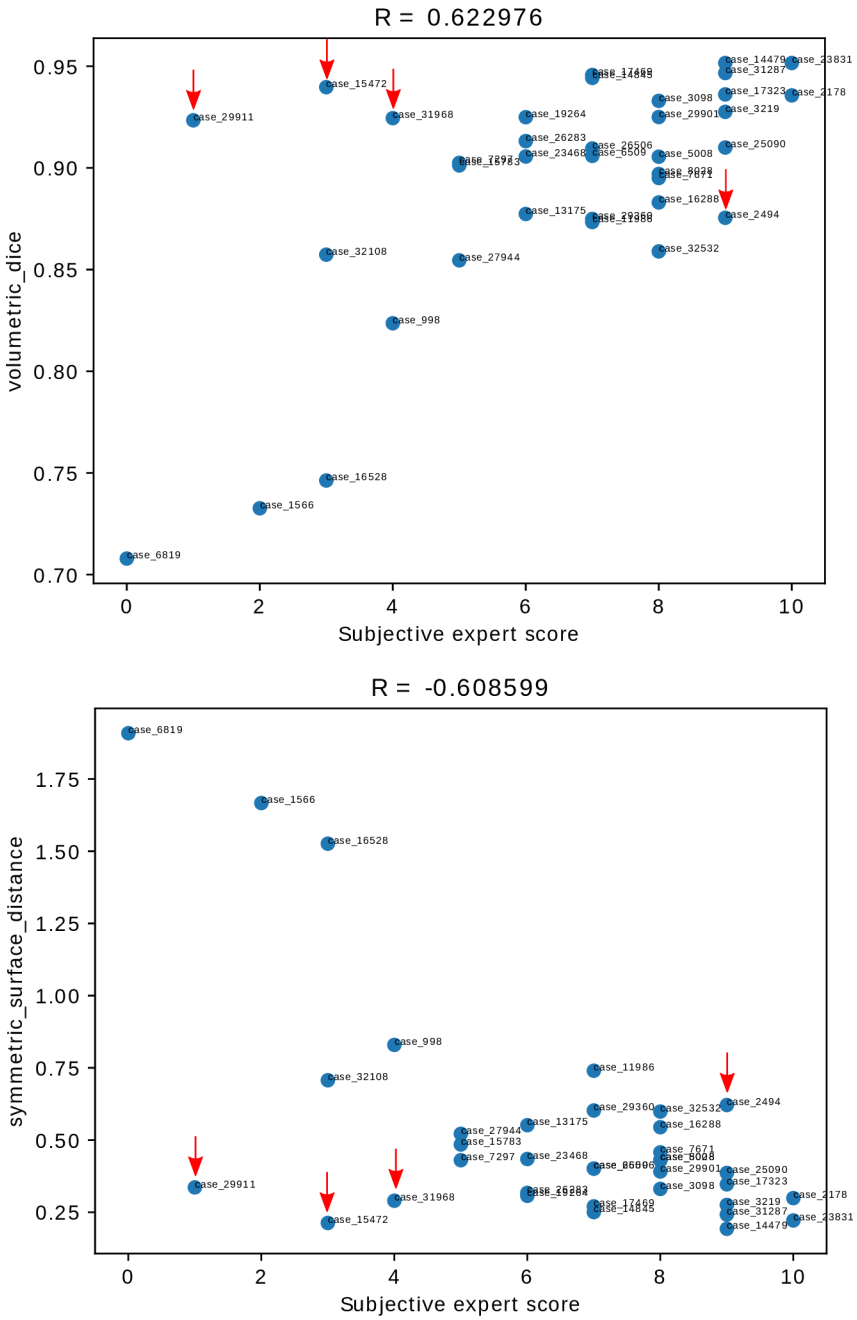

Figure 11.1: Plots of the three global quantitative metrics plotted against the corresponding expert subjective score. Note that in some cases (highlighted by red arrows), the metrics failed to estimate practical usability of the reconstruction result.

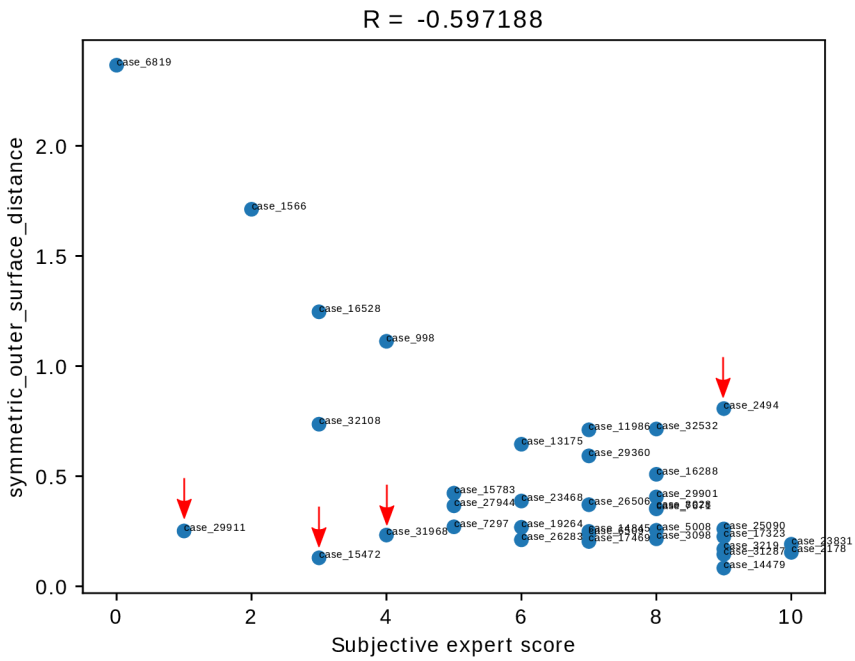

Figure 11.2: Global outer surface distance plotted against the corresponding expert subjective score. The same cases are highlighted as in the Figure 11.1.

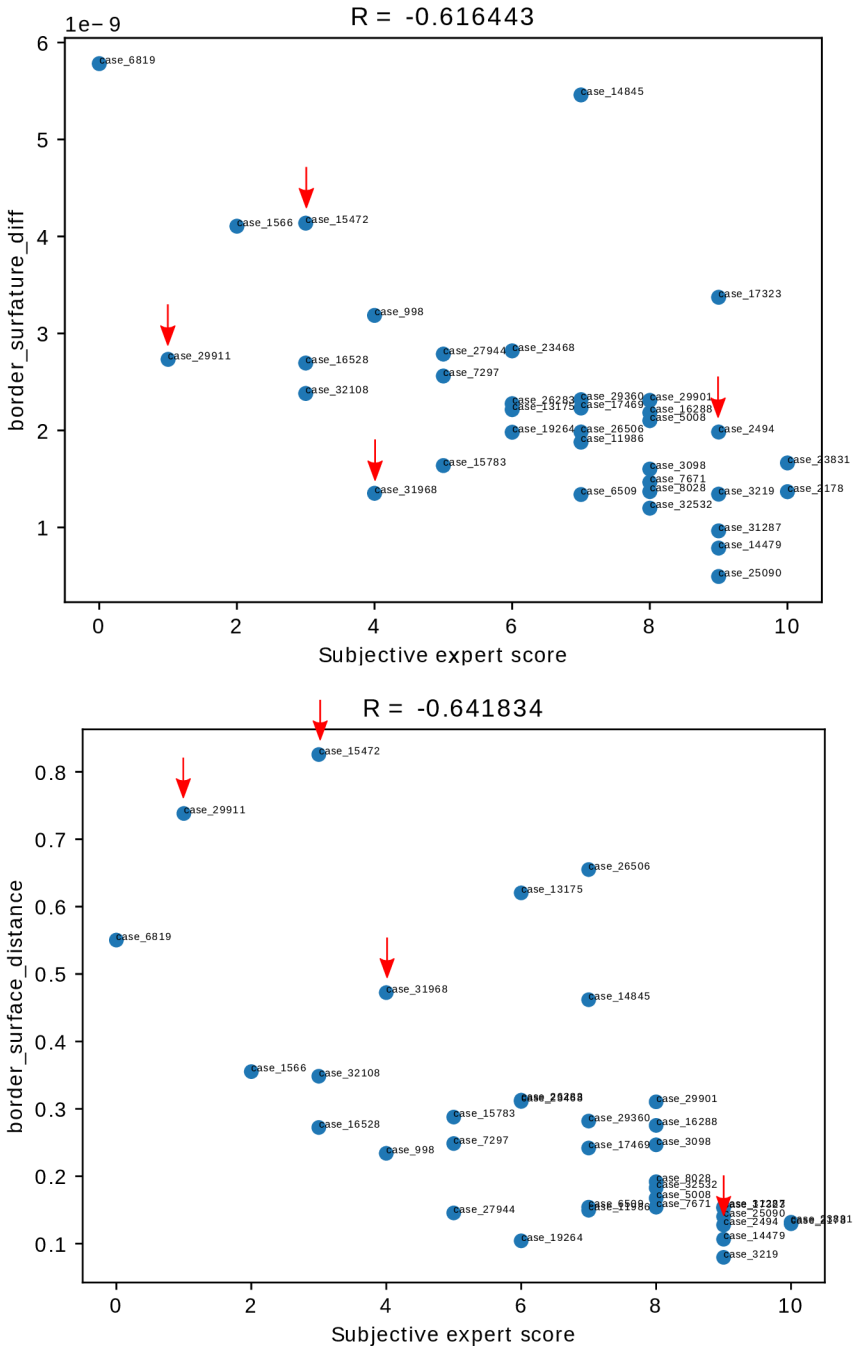

Figure 11.3: Plots of the two border quantitative metrics plotted against the corresponding expert subjective score. The same cases are highlighted as in the Figures 11.1 and 11.2, showing that the border metrics convey different yet relevant information about the reconstruction result.

Part IV

SUMMARY

12

CONCLUSION

The primary goal of this thesis was streamlining the clinical workflow of cranioplasty PSI design which will lead to higher quality and accessibility of state-of-the-art cranial implants. This task was approached by addressing two main bottlenecks of the current workflow: skull segmentation and virtual reconstruction. In both cases, novel approaches based on deep learning were proposed and implemented to reduce the amount of required human operator interaction during the design process. Combined with automatic skull alignment, the proposed methods form a complete cranioplasty PSI design system (see Figure 12.1). Although outputs of individual system components can be edited by a human operator in case of suboptimal results, the system is designed to be fully automatic.

Based on feedback from the experts of the TESCAN Medical company, the automatic CutCNN skull segmentation method already provides significant improvement and decreased time cost in practice when compared to the previously available set of conventional segmentation tools, although these can still be used to correct errors of the automatic method. This allows for more efficient creation of anatomy models for surgical planning or patient education, as well as preprocessing data for subsequent virtual skull reconstruction.

The automatic skull shape reconstruction is being experimentally implemented into the computer-assisted implant design practice as of the time of writing this thesis and therefore the feedback on the practical impact on clinical practice is limited. However, the approach was shown to achieve promising results. It ranked 3rd in the AutoImplant 2020 challenge, achieving highly competitive results in terms of Dice coefficient. It was also the only fully 3D, high-resolution approach, which allowed it to avoid some visible artifacts that can result from 2D slice-by-slice processing or resampling to lower resolution. A comprehensive expert evaluation is currently not available. However, 4 real cranioplasty test cases were evaluated by an expert in the same manner that was described in Section 11. All of the evaluated cases were labeled as bringing some practical utility to the design process (i. e. the expert would be willing to use them as a template for subsequent modeling). In 3 of the 4 reconstructed cases, the expert assigned a subjective score of 7 or more to the automatic skull reconstruction result, hinting at overall high quality.

The current research of automatic virtual skull reconstruction methods is limited by several factors. These include a lack of structured and standardized cranioplasty datasets and a lack of consensus regarding the use of quantitative metrics. Addi-

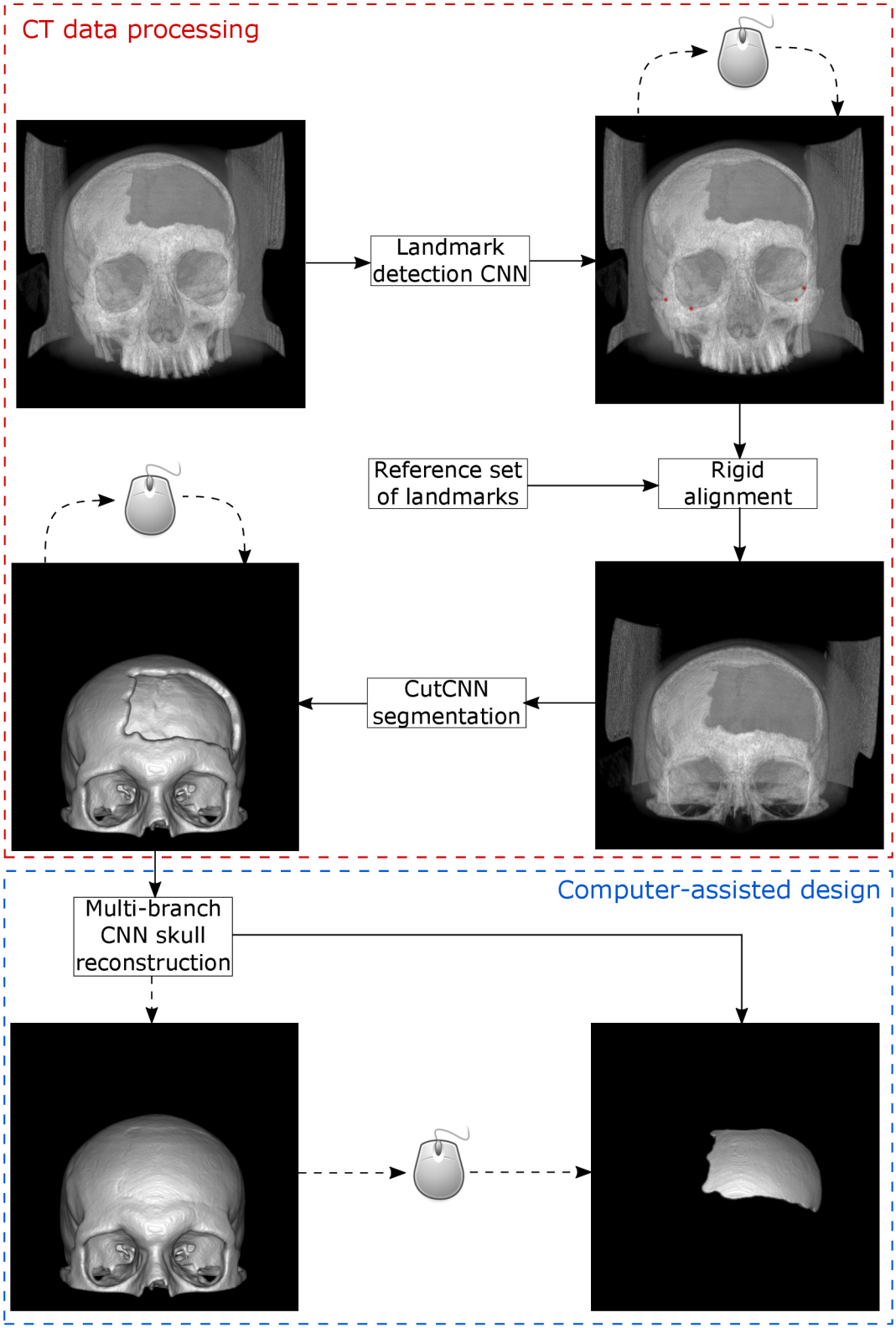

Figure 12.1: Overview of the implant design workflow proposed in this thesis.

tionally, little attention is given to the variability of feasible implant shapes. To address these issues, this thesis has several secondary contributions.

The SkullBreak dataset opened access to standardized virtual skull reconstruction data that could be used for development, reproducible evaluation, and comparison of automatic reconstruction methods. It was designed to focus on anatomical variability and variability of defect shapes and allows for the identification of strong and weak points of the reconstruction approaches by containing sets of strictly unilateral and bilateral defects, and also sets of cranial and orbital defects. The dataset is being used in the current edition of the AutoImplant challenge[1], serving as a benchmark for a variety of state-of-the-art skull reconstruction methods.

The discrepancy in the choice of quantitative metrics to be used for the evaluation of skull reconstruction methods does not have a straightforward solution. The single user study presented in Section 11 did not identify a single best quantitative metric. Instead, it was concluded that several different metrics should be combined to gauge the clinical usability of the skull reconstruction outputs. This work also expanded the pool of available metrics by the novel Gaussian curvature estimation error which can help visually identify bumps and trenches along the skull defect border.

Finally, this work briefly addressed the task of generating multiple reconstructed shape hypotheses for a single defective skull by using generative models. Although the achieved shape variability was limited and the option of generating multiple reconstructions is currently not required for medical practice, it is a promising area that could be explored further in the future.

Several additional topics should be addressed in future work. From an evaluation point of view, an extensive user study needs to be performed to ascertain the clinical impact of the proposed automatic cranioplasty pipeline. Two main sources of feedback will likely be available in near future: first, the experts from the TESCAN Medical will have access to the experimental implementation of the proposed framework, and second, the current edition of the AutoImplant challenge will include an evaluation dataset of real cranioplasty data which will be rated by experienced clinicians skilled in PSI design.

From a methodological point of view, future research will need to address shortcomings of the currently proposed automatic methods identified by the evaluating experts. Independently of what these shortcomings may be, one major topic could be incorporating user input into the inference mechanism of deep learning models. One possible approach proposed in Section 8 consists of generating a continuous spectrum of shape hypotheses and letting the user choose the most appropriate

[1] https://autoimplant2021.grand-challenge.org/Task1/

one. However, more sophisticated approaches could also be explored, such as expanding the proposed model to respond to user inputs in form of clicks or scribbles in 2D slices or operations defined in the 3D domain. Such a method would provide a powerful interactive tool capable of generating optimal PSIs with minimum time cost and almost no manual skill requirements, contributing greatly to the PSI quality as well as accessibility.

GLOSSARY

AM Additive Manufacturing. 38
ASD Average Symmetric Distance. 19, 30

CAD Computer-assisted Design. 13, 60, 83, 84
CBCT Cone Beam Computed Tomography. 11
CMF Craniomaxillofacial. 5, 8, 11
CNN Convolutional Neural Network. 17, 22, 25–28, 33, 34, 38–41, 43–49, 59–65, 73–79, 83–85, 87, 88, 90–92, 94, 97
CT Computed Tomography. 4, 11–13, 15, 17, 19, 20, 24, 27, 32–42, 45, 46, 48, 49, 52, 56, 60, 76, 85, 86

DICOM Digital Imaging and Communications in Medicine. 11, 56
DSA Defect Surface Area. 53, 54

FHP Frankfort-horizontal Plane. 6, 33, 34, 48, 49

GAN Generative Adversarial Network. 27, 64, 65

HU Hounsfield Unit. 11, 12, 15, 19, 55

ICP Iterative Closest Point algorithm. 23

MR Magnetic Resonance. 4, 27
MSD Mean Symmetric Distance. 44, 45
MSP Mid-sagittal Plane. 6, 33
MV Multi-view. 39, 41, 43–45

NURBS Non-uniform Rational Basis Spline. 20

Or Orbitale landmark. 8

PCA Principal Component Analysis. 17, 23, 24, 27
PEEK Polyether-Ether-Ketone. 10
PMMA Polymerized Methyl-Methacrylate. 10
Po Porion landmark. 8
PSI Patient-specific Implant. 5, 12, 15, 18, 19, 32, 60, 61, 73, 103, 105, 106

RBF Radial Basis Function. 20

SDC Surface Dice Coefficient. 43–45

SSM Statistical Shape Model. 17, 22–25, 28, 30

STL Standard Tessellation Language. 12, 13

TPS Thin Plate Spline. 20, 21, 24, 25

VDC Volumetric Dice Coefficient. 19, 30, 43–45

VOR Voxel Occupancy Rate. 53, 54

BIBLIOGRAPHY

[1] Cleft lip and cleft palate - Symptoms and causes. URL https://www.mayoclinic.org/diseases-conditions/cleft-palate/symptoms-causes/syc-20370985. Accessed: 2021-04-04.

[2] Cranioplasty Surgery, Bone Correction, . URL https://www.princetonneurologicalsurgery.com/our-services/brain-surgery/cranioplasty/. Accessed: 2021-04-04.

[3] Craniosynostosis and Craniofacial Disorders – Definitions, Diagnosis and Treatments, . URL https://www.aans.org/. Accessed: 2021-04-04.

[4] Plagiocephaly, Craniosynostosis, and Genetic Syndromes, . URL https://www.dukehealth.org/pediatric-treatments/pediatric-plastic-and-reconstructive-surgery/abnormal-head-shape-and-skull-deformity. Accessed: 2021-04-04.

[5] Amir H. Abdi, Heather Borgard, Purang Abolmaesumi, and Sidney Fels. Anatomygen: Deep anatomy generation from dense representation with applications in mandible synthesis. In *International Conference on Medical Imaging with Deep Learning*, London, United Kingdom, 08–10 Jul 2019.

[6] Amir H. Abdi, Mehran Pesteie, Eitan Prisman, Purang Abolmaesumi, and Sidney Fels. Variational shape completion for virtual planning of jaw reconstructive surgery. In *Lecture Notes in Computer Science*, pages 227–235. Springer International Publishing, 2019. doi: 10.1007/978-3-030-32254-0_26.

[7] SangIn An, Ji-Yeon Lee, Chooryung J. Chung, and Kyung-Ho Kim. Comparison of different midsagittal plane configurations for evaluating craniofacial asymmetry by expert preference. *American Journal of Orthodontics and Dentofacial Orthopedics*, 152(6):788–797, December 2017. doi: 10.1016/j.ajodo.2017.04.024.

[8] Kenneth Anderson. *Mosby's medical, nursing, & allied health dictionary*. Mosby, St. Louis, 1998. ISBN 978-0815148005.

[9] Luca Di Angelo, Paolo Di Stefano, Lapo Governi, Antonio Marzola, and Yary Volpe. A robust and automatic method for the best symmetry plane detection of craniofacial skeletons. *Symmetry*, 11(2):245, February 2019. doi: 10.3390/sym11020245.

[10] Sajad Arabnejad, Burnett Johnston, Michael Tanzer, and Damiano Pasini. Fully porous 3D printed titanium femoral stem to reduce stress-shielding following total hip arthroplasty. *Journal of Orthopaedic Research*, 35(8):1774–1783, October 2016. doi: 10.1002/jor.23445.

[11] Brian B. Avants, Nicholas J. Tustison, Gang Song, Philip A. Cook, Arno Klein, and James C. Gee. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage*, 54(3):2033–2044, February 2011. doi: 10.1016/j.neuroimage.2010.09.025.

[12] Mounir Bashour. History and current concepts in the analysis of facial attractiveness. *Plastic and Reconstructive Surgery*, 118(3):741–756, September 2006. doi: 10.1097/01.prs.0000233051.61512.65.

[13] R. Bryan Bell. Computer planning and intraoperative navigation in craniomaxillofacial surgery. *Oral and Maxillofacial Surgery Clinics of North America*, 22(1):135–156, February 2010. doi: 10.1016/j.coms.2009.10.010.

[14] Alexander Belyaev and Yutaka Ohtake. A comparison of mesh smoothing methods. *Israel-Korea Bi-National Conference on Geometric Modeling and Computer Graphics, Tel Aviv University, 83-87 (2003)*, 01 2003.

[15] P.J. Besl and Neil D. McKay. A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, February 1992. doi: 10.1109/34.121791.

[16] F Edward Boas and Dominik Fleischmann. CT artifacts: causes and reduction techniques. *Imaging in Medicine*, 4(2):229–240, April 2012. doi: 10.2217/iim.12.13.

[17] F.L. Bookstein. Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6):567–585, June 1989. doi: 10.1109/34.24792.

[18] Y.Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*. IEEE Comput. Soc. doi: 10.1109/iccv.2001.937505.

[19] J. C. Carr, R. K. Beatson, J. B. Cherrie, T. J. Mitchell, W. R. Fright, B. C. McCallum, and T. R. Evans. Reconstruction and representation of 3D objects with radial basis functions. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques - SIGGRAPH '01*. ACM Press, 2001. doi: 10.1145/383259.383266.

[20] J.C. Carr, W.R. Fright, and R.K. Beatson. Surface interpolation with radial basis functions for medical imaging. *IEEE Transactions on Medical Imaging*, 16 (1):96–107, 1997. doi: 10.1109/42.552059.

[21] Yu-Bing Chang, James J. Xia, Peng Yuan, Tai-Hong Kuo, Zixiang Xiong, Jaime Gateno, and Xiaobo Zhou. 3D segmentation of maxilla in cone-beam computed tomography imaging using base invariant wavelet active shape model on customized two-manifold topology. *Journal of X-Ray Science and Technology*, 21(2):251–282, 2013. ISSN 08953996. doi: 10.3233/XST-130369.

[22] Jian-Jun Chen, Wei Liu, Ming-Zhe Li, and Cheng-Tao Wang. Digital manufacture of titanium prosthesis for cranioplasty. *The International Journal of Advanced Manufacturing Technology*, 27(11-12):1148–1152, March 2005. doi: 10.1007/s00170-004-2309-y.

[23] Xiaojun Chen, Lu Xu, Xing Li, and Jan Egger. Computer-aided implant design for the restoration of cranial defects. *Scientific Reports*, 7(1), June 2017. doi: 10.1038/s41598-017-04454-6.

[24] Yani Chen, Bibo Shi, Zhewei Wang, Pin Zhang, Charles D. Smith, and Jundong Liu. Hippocampus segmentation through multi-view ensemble ConvNets. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE, April 2017. doi: 10.1109/isbi.2017.7950499.

[25] Cheng-Hsin Cheng, Hao-Yu Chuang, Hung-Lin Lin, Chun-Lin Liu, and Chun-Hsu Yao. Surgical results of cranioplasty using three-dimensional printing technology. *Clinical Neurology and Neurosurgery*, 168:118–123, May 2018. doi: 10.1016/j.clineuro.2018.03.004.

[26] Pengfei Cheng, Pei Han, Changli Zhao, Shaoxiang Zhang, Xiaonong Zhang, and Yimin Chai. Magnesium inference screw supports early graft incorporation with inhibition of graft degradation in anterior cruciate ligament reconstruction. *Scientific Reports*, 6(1), May 2016. doi: 10.1038/srep26434.

[27] Yuan Cheng, Wee Kheng Leow, and Thiam Chye Lim. Automatic identification of frankfurt plane and mid-sagittal plane of skull. In *2012 IEEE Workshop on the Applications of Computer Vision (WACV)*. IEEE, January 2012. doi: 10.1109/wacv.2012.6162994.

[28] Sasank Chilamkurthy, Rohit Ghosh, Swetha Tanamala, Mustafa Biviji, Norbert G. Campeau, Vasantha Kumar Venugopal, Vidur Mahajan, Pooja Rao, and Prashant Warier. Development and validation of deep learning algorithms for detection of critical findings in head CT scans, 2018.

[29] Chengwen Chu, Cheng Chen, Li Liu, and Guoyan Zheng. FACTS: Fully automatic CT segmentation of a hip joint. *Annals of Biomedical Engineering*, 43(5):1247–1259, November 2014. doi: 10.1007/s10439-014-1176-4.

[30] John D. Currey. The structure and mechanics of bone. *Journal of Materials Science*, 47(1):41–54, September 2011. doi: 10.1007/s10853-011-5914-9.

[31] Rodrigo Dalvit Carvalho da Silva, Thomas Richard Jenkyn, and Victor Alexander Carranza. Convolutional neural networks and geometric moments to identify the bilateral symmetric midplane in facial skeletons from CT scans. *Biology*, 10(3):182, March 2021. doi: 10.3390/biology10030182.

[32] Erasmo Barros da Silva Júnior, Afonso Henrique de Aragão, Marcelo de Paula Loureiro, Caetano Silva Lobo, Ana Flávia Oliveti, Rafael Martinelli de Oliveira, and Ricardo Ramina. Cranioplasty with three-dimensional customised mould for polymethylmethacrylate implant: a series of 16 consecutive patients with cost-effectiveness consideration. *3D Printing in Medicine*, 7 (1), February 2021. doi: 10.1186/s41205-021-00096-7.

[33] Angela Dai, Charles Ruizhongtai Qi, and Matthias NieBner. Shape completion using 3D-encoder-predictor CNNs and shape synthesis. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017. doi: 10.1109/cvpr.2017.693.

[34] David Dean, Kyoung-June Min, and Angus Bond. Computer aided design of large-format prefabricated cranial plates. *Journal of Craniofacial Surgery*, 14 (6):819–832, November 2003. doi: 10.1097/00001665-200311000-00002.

[35] Dušan Drevický and Oldřich Kodym. Evaluating deep learning uncertainty measures in cephalometric landmark localization. In *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies*. SCITEPRESS - Science and Technology Publications, 2020. doi: 10.5220/0009375302130220.

[36] Paul S. D'Urso, Timothy M. Barker, W. John Earwaker, Lain J. Bruce, R. Leigh Atkinson, Michael W. Lanigan, John F. Arvier, and David J. Effeney. Stereolithographic biomodelling in cranio-maxillofacial surgery: a prospective trial. *Journal of Cranio-Maxillofacial Surgery*, 27(1):30–37, February 1999. doi: 10.1016/s1010-5182(99)80007-9.

[37] David G. Ellis and Michele R. Aizenberg. Deep learning using augmentation via registration: 1st place solution to the AutoImplant 2020 challenge. In *Towards the Automatization of Cranial Implant Design in Cranioplasty*, pages 47–

55. Springer International Publishing, 2020. doi: 10.1007/978-3-030-64327-0_6.

[38] Jean H. D. Fasel, Jörg Beinemann, Karl Schaller, and Heinz-Otto Peitgen. Computer science tools for manual editing of computed tomographic images: Impact on the quality of 3D printed models. *Surgical Science*, 05(10): 439–443, 2014. doi: 10.4236/ss.2014.510068.

[39] Zacharias Fourie, Janalt Damstra, Rutger H. Schepers, Peter O. Gerrits, and Yijin Ren. Segmentation process significantly influences the accuracy of 3D surface models derived from cone beam computed tomography. *European Journal of Radiology*, 81(4):e524–e530, April 2012. doi: 10.1016/j.ejrad.2011.06.001.

[40] John R. Fredieu, Jennifer Kerbo, Mark Herron, Ryan Klatte, and Malcolm Cooke. Anatomical models: a digital revolution. *Medical Science Educator*, 25 (2):183–194, March 2015. doi: 10.1007/s40670-015-0115-9.

[41] 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(4):519–529, October 2017. doi: 10.1007/s11548-017-1674-6.

[42] Marc Anton Fuessinger, Steffen Schwarz, Joerg Neubauer, Carl-Peter Cornelius, Mathieu Gass, Philipp Poxleitner, Ruediger Zimmerer, Marc Christian Metzger, and Stefan Schlager. Virtual reconstruction of bilateral midfacial defects by using statistical shape modeling. *Journal of Cranio-Maxillofacial Surgery*, 47(7):1054–1059, July 2019. doi: 10.1016/j.jcms.2019.03.027.

[43] Markus Gall, Xing Li, Xiaojun Chen, Dieter Schmalstieg, and Jan Egger. Computer-aided planning and reconstruction of cranial 3D implants. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, August 2016. doi: 10.1109/embc.2016.7590915.

[44] Mirko S. Gilardino, Mihiran Karunanayake, Taghreed Al-Humsi, Ali Izadpanah, Hasan Al-Ajmi, Judith Marcoux, Jeffrey Atkinson, and Jean-Pierre Farmer. A comparison and cost analysis of cranioplasty techniques. *Journal of Craniofacial Surgery*, 26(1):113–117, January 2015. doi: 10.1097/scs.0000000000001305.

[45] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adver-

sarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[46] Henry Gray. Anatomy of the Human Body. URL https://www.bartleby.com/107/. Accessed: 2021-04-04.

[47] Haijun Gui, Huawei Yang, ShiLei Zhang, Steve G.F. Shen, Ming Ye, and Rainer Schmelzeisen. Mirroring tool. *Journal of Craniofacial Surgery*, 26(7): 2115–2119, October 2015. doi: 10.1097/scs.0000000000000913.

[48] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5767–5777. Curran Associates, Inc., 2017.

[49] Tobias Heimann and Hans-Peter Meinzer. Statistical shape models for 3D medical image segmentation: A review. *Medical Image Analysis*, 13(4):543–563, August 2009. doi: 10.1016/j.media.2009.05.004.

[50] Philipp Honigmann, Neha Sharma, Brando Okolo, Uwe Popp, Bilal Msallem, and Florian M. Thieringer. Patient-specific surgical implants made of 3D printed PEEK: Material, technology, and scope of surgical application. *BioMed Research International*, 2018:1–8, 2018. doi: 10.1155/2018/4520636.

[51] Eero Huotilainen, Risto Jaanimets, Jiří Valášek, Petr Marcián, Mika Salmi, Jukka Tuomi, Antti Mäkitie, and Jan Wolff. Inaccuracies in additive manufactured medical skull models caused by the DICOM to STL conversion process. *Journal of Cranio-Maxillofacial Surgery*, 42(5):e259–e265, July 2014. doi: 10.1016/j.jcms.2013.10.001.

[52] Ji hyeon Oh. Recent advances in the reconstruction of cranio-maxillofacial defects using computer-aided design/computer-aided manufacturing. *Maxillofacial Plastic and Reconstructive Surgery*, 40(1), February 2018. doi: 10.1186/s40902-018-0141-9.

[53] Jesper Jansen, Ruud Schreurs, Leander Dubois, Thomas J. J. Maal, Peter J. J. Gooris, and Alfred G. Becking. Orbital volume analysis: validation of a semi-automatic software segmentation method. *International Journal of Computer Assisted Radiology and Surgery*, 11(1):11–18, July 2015. doi: 10.1007/s11548-015-1254-6.

[54] Jesper Jansen, Leander Dubois, Ruud Schreurs, Peter J.J. Gooris, Thomas J.J. Maal, Ludo F. Beenen, and Alfred G. Becking. Should virtual mirroring be

used in the preoperative planning of an orbital reconstruction? *Journal of Oral and Maxillofacial Surgery*, 76(2):380–387, February 2018. doi: 10.1016/j.joms.2017.09.018.

[55] André Luiz Jardini, Maria Aparecida Larosa, Rubens Maciel Filho, Cecília Amélia de Carvalho Zavaglia, Luis Fernando Bernardes, Carlos Salles Lambert, Davi Reis Calderoni, and Paulo Kharmandayan. Cranial reconstruction: 3D biomodel and custom-built implant created using additive manufacturing. *Journal of Cranio-Maxillofacial Surgery*, 42(8):1877–1884, December 2014. doi: 10.1016/j.jcms.2014.07.006.

[56] Yasas S. N. Jayaratne and Roger A. Zwahlen. A systematic review of interethnic variability in facial dimensions. *Plastic and Reconstructive Surgery*, 129(1):164e–165e, January 2012. doi: 10.1097/prs.0b013e3182362e3f.

[57] Patrik Kamencay, Martina Radilova, Roman Radil, Miroslav Benco, Robert Hudec, and Roberta Vrskova. Innovative 3D reconstruction method based on patch based technique using neural network. In *2020 43rd International Conference on Telecommunications and Signal Processing (TSP)*. IEEE, July 2020. doi: 10.1109/tsp49548.2020.9163484.

[58] Eugene E. Keller, Evre Baltali, Xinhua Liang, Kristin Zhao, Marianne Huebner, and Kai-Nan An. Temporomandibular custom hemijoint replacement prosthesis: Prospective clinical and kinematic study. *Journal of Oral and Maxillofacial Surgery*, 70(2):276–288, February 2012. doi: 10.1016/j.joms.2011.06.202.

[59] Kevin J. Kelly, Paul N. Manson, Craig A. Vander Kolk, Bernard L. Markowitz, Michael C. Dunham, Thomas O. Rumley, and William A. Crawley. Sequencing LeFort fracture treatment (organization of treatment for a panfacial fracture). *Journal of Craniofacial Surgery*, 1(4):168–178, October 1990. doi: 10.1097/00001665-199001040-00003.

[60] Basel A. Khader and Mark R. Towler. Materials and techniques used in cranioplasty fixation: A review. *Materials Science and Engineering: C*, 66:315–322, September 2016. doi: 10.1016/j.msec.2016.04.101.

[61] Uwe Klammert, Uwe Gbureck, Elke Vorndran, Jan Rödiger, Philipp Meyer-Marcotty, and Alexander C. Kübler. 3D powder printed calcium phosphate implants for reconstruction of cranial and maxillofacial defects. *Journal of Cranio-Maxillofacial Surgery*, 38(8):565–570, December 2010. doi: 10.1016/j.jcms.2010.01.009.

[62] Oldřich Kodym, Michal Španěl, and Adam Herout. Cranial defect reconstruction using cascaded CNN with alignment. In *Towards the Automatization of Cranial Implant Design in Cranioplasty*, pages 56–64. Springer International Publishing, 2020. doi: 10.1007/978-3-030-64327-0_7.

[63] Oldřich Kodym, Michal Španěl, and Adam Herout. Skull shape reconstruction using cascaded convolutional networks. *Computers in Biology and Medicine*, 123:103886, August 2020. doi: 10.1016/j.compbiomed.2020.103886.

[64] Oldřich Kodym, Jianning Li, Antonio Pepe, Christina Gsaxner, Sasank Chilamkurthy, Jan Egger, and Michal Španěl. SkullBreak / SkullFix – dataset for automatic cranial implant design and a benchmark for volumetric shape learning tasks. *Data in Brief*, 35:106902, April 2021. doi: 10.1016/j.dib.2021.106902.

[65] Oldřich Kodym, Michal Španěl, and Adam Herout. Segmentation of defective skulls from ct data for tissue modelling, 2019. Presented at MICCAI-MSKI 2019.

[66] Martin Kolarik, Radim Burget, and Kamil Riha. Upsampling algorithms for autoencoder segmentation neural networks: A comparison study. In *2019 11th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*. IEEE, October 2019. doi: 10.1109/icumt48472.2019.8970918.

[67] Rotem Kowner. Facial asymmetry and attractiveness judgement in developmental perspective. *Journal of Experimental Psychology: Human Perception and Performance*, 22(3):662–675, 1996. doi: 10.1037/0096-1523.22.3.662.

[68] David B. Kurland, Ariana Khaladj-Ghom, Jesse A. Stokum, Brianna Carusillo, Jason K. Karimy, Volodymyr Gerzanich, Juan Sahuquillo, and J. Marc Simard. Complications associated with decompressive craniectomy: A systematic review. *Neurocritical Care*, 23(2):292–304, Oct 2015. ISSN 1556-0961. doi: 10.1007/s12028-015-0144-7.

[69] Ming-Yih Lee, Chong-Ching Chang, Chao-Chun Lin, Lun-Jou Lo, and Yu-Ray Chen. Custom implant design for patients with cranial defects. *IEEE Engineering in Medicine and Biology Magazine*, 21(2):38–44, March 2002. ISSN 1937-4186. doi: 10.1109/MEMB.2002.1000184.

[70] Ming-Yih Lee, Chong-Ching Chang, Chao-Chun Lin, Lun-Jou Lo, and Yu-Ray Chen. Custom implant design for patients with cranial defects. *IEEE Engineering in Medicine and Biology Magazine*, 21(2):38–44, March 2002. doi: 10.1109/memb.2002.1000184.

[71] Si Hoon Lee, Chan Jong Yoo, Uhn Lee, Cheol Wan Park, Sang Gu Lee, and Woo Kyung Kim. Resorption of autogenous bone graft in cranioplasty: Resorption and reintegration failure. *Korean Journal of Neurotrauma*, 10(1):10, 2014. doi: 10.13004/kjnt.2014.10.1.10.

[72] Jianning Li and Jan Egger. Dataset descriptor for the AutoImplant cranial implant design challenge. In *Towards the Automatization of Cranial Implant Design in Cranioplasty*, pages 10–15. Springer International Publishing, 2020. doi: 10.1007/978-3-030-64327-0_2.

[73] Jianning Li and Jan Egger, editors. *Towards the Automatization of Cranial Implant Design in Cranioplasty*. Springer International Publishing, 2020. doi: 10.1007/978-3-030-64327-0.

[74] Jianning Li, Antonio Pepe, Christina Gsaxner, Gord von Campe, and Jan Egger. A baseline approach for AutoImplant: The MICCAI 2020 cranial implant design challenge. In *Multimodal Learning for Clinical Decision Support and Clinical Image-Based Procedures*, pages 75–84. Springer International Publishing, 2020. doi: 10.1007/978-3-030-60946-7_8.

[75] Jianning Li, Antonio Pepe, Christina Gsaxner, and Jan Egger. An online platform for automatic skull defect restoration and cranial implant design. *Proc. SPIE 11598, Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling, 115981Q*, February 2021.

[76] Jianning Li, Pedro Pimentel, Angelika Szengel, Moritz Ehlke, Hans Lamecker, Stefan Zachow, Laura Estacio, Christian Doenitz, Heiko Ramm, Haochen Shi, Xiaojun Chen, Franco Matzkin, Virginia Newcombe, Enzo Ferrante, Yuan Jin, David G. Ellis, Michele R. Aizenberg, Oldrich Kodym, Michal Spanel, Adam Herout, James G. Mainprize, Zachary Fishman, Michael R. Hardisty, Amirhossein Bayat, Suprosanna Shit, Bomin Wang, Zhi Liu, Matthias Eder, Antonio Pepe, Christina Gsaxner, Victor Alves, Ulrike Zefferer, Gord Von Campe, Karin Pistracher, Ute Schafer, Dieter Schmalstieg, Bjoern H. Menze, Ben Glocker, and Jan Egger. AutoImplant 2020 -first MICCAI challenge on automatic cranial implant design. *IEEE Transactions on Medical Imaging*, pages 1–1, 2021. doi: 10.1109/tmi.2021.3077047.

[77] Xin Li, Zhao Yin, Li Wei, Shenghua Wan, Wei Yu, and Maoqing Li. Symmetry and template guided completion of damaged skulls. *Computers & Graphics*, 35(4):885–893, August 2011. doi: 10.1016/j.cag.2011.01.015.

[78] Yaoren Liew, Erin Beveridge, Andreas K. Demetriades, and Mark A. Hughes. 3D printing of patient-specific anatomy: A tool to improve patient consent

and enhance imaging interpretation by trainees. *British Journal of Neurosurgery*, 29(5):712–714, March 2015. doi: 10.3109/02688697.2015.1026799.

[79] Christopher G.T. Lim, Duncan I. Campbell, Nicholas Cook, and Jason Erasmus. A case series of rapid prototyping and intraoperative imaging in orbital reconstruction. *Craniomaxillofacial Trauma & Reconstruction*, 8(2):105–110, June 2015. doi: 10.1055/s-0034-1395384.

[80] Hsiu-Hsia Lin, Ya-Fang Chuang, Jing-Ling Weng, and Lun-Jou Lo. Comparative validity and reproducibility study of various landmark-oriented reference planes in 3-dimensional computed tomographic analysis for patients receiving orthognathic surgery. *PLOS ONE*, 10(2):e0117604, February 2015. doi: 10.1371/journal.pone.0117604.

[81] Oscar Cuadros Linares, Jonas Bianchi, Dirceu Raveli, João Batista Neto, and Bernd Hamann. Mandible and skull segmentation in cone beam computed tomography using super-voxels and graph clustering. *The Visual Computer*, 35(10):1461–1474, April 2018. doi: 10.1007/s00371-018-1511-0.

[82] Or Litany, Alex Bronstein, Michael Bronstein, and Ameesh Makadia. Deformable shape completion with graph convolutional autoencoders. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, June 2018. doi: 10.1109/cvpr.2018.00202.

[83] Fang Lu, Fa Wu, Peijun Hu, Zhiyi Peng, and Dexing Kong. Automatic 3D liver location and segmentation via convolutional neural network and graph cut. *International Journal of Computer Assisted Radiology and Surgery*, 12(2):171–182, September 2016. doi: 10.1007/s11548-016-1467-3.

[84] Anders Lundström and Fredrik Lundström. The frankfort horizontal as a basis for cephalometric analysis. *American Journal of Orthodontics and Dentofacial Orthopedics*, 107(5):537–540, May 1995. doi: 10.1016/s0889-5406(95)70121-4.

[85] Filipe M. M. Marreiros, Y. Heuzé, M. Verius, C. Unterhofer, W. Freysinger, and W. Recheis. Custom implant design for large cranial defects. *International Journal of Computer Assisted Radiology and Surgery*, 11(12):2217–2230, June 2016. doi: 10.1007/s11548-016-1454-8.

[86] Richard Marsell and Thomas A. Einhorn. The biology of fracture healing. *Injury*, 42(6):551 – 555, 2011. ISSN 0020-1383. doi: https://doi.org/10.1016/j.injury.2011.03.031. Bone Regeneration in the 21st Century.

[87] Nicolas Martelli, Carole Serrano, Hélène van den Brink, Judith Pineau, Patrice Prognon, Isabelle Borget, and Salma El Batti. Advantages and disad-

vantages of 3-dimensional printing in surgery: A systematic review. *Surgery*, 159(6):1485–1500, June 2016. doi: 10.1016/j.surg.2015.12.017.

[88] Antonio Marzola, Lapo Governi, Lorenzo Genitori, Federico Mussa, Yary Volpe, and Rocco Furferi. A semi-automatic hybrid approach for defective skulls reconstruction. *Computer-Aided Design and Applications*, 17(1):190–204, May 2019. doi: 10.14733/cadaps.2020.190-204.

[89] Antonio Marzola, Michaela Servi, and Yary Volpe. A reliable procedure for the construction of a statistical shape model of the cranial vault. In *Lecture Notes in Mechanical Engineering*, pages 788–800. Springer International Publishing, September 2019. doi: 10.1007/978-3-030-31154-4_67.

[90] Franco Matzkin, Virginia Newcombe, Ben Glocker, and Enzo Ferrante. Cranial implant design via virtual craniectomy with shape priors. In *Towards the Automatization of Cranial Implant Design in Cranioplasty*, pages 37–46. Springer International Publishing, 2020. doi: 10.1007/978-3-030-64327-0_5.

[91] Franco Matzkin, Virginia Newcombe, Susan Stevenson, Aneesh Khetani, Tom Newman, Richard Digby, Andrew Stevens, Ben Glocker, and Enzo Ferrante. Self-supervised skull reconstruction in brain CT images with decompressive craniectomy. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 390–399. Springer International Publishing, 2020. doi: 10.1007/978-3-030-59713-9_38.

[92] N Meiyappan, S Tamizharasi, KP Senthilkumar, and K Janardhanan. Natural head position: An overview. *Journal of Pharmacy and Bioallied Sciences*, 7(6):424, 2015. doi: 10.4103/0975-7406.163488.

[93] M. C. Metzger, G. Bittermann, L. Dannenberg, R. Schmelzeisen, N.-C. Gellrich, B. Hohlweg-Majert, and C. Scheifele. Design and development of a virtual anatomic atlas of the human skull for automatic segmentation in computer-assisted surgery, preoperative planning, and navigation. *International Journal of Computer Assisted Radiology and Surgery*, 8(5):691–702, February 2013. doi: 10.1007/s11548-013-0818-6.

[94] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE, October 2016. doi: 10.1109/3Dv.2016.79.

[95] Jordi Minnema, Maureen van Eijnatten, Wouter Kouw, Faruk Diblen, Adriënne Mendrik, and Jan Wolff. CT image segmentation of bone for

medical additive manufacturing using a convolutional neural network. *Computers in Biology and Medicine*, 103:130–139, December 2018. doi: 10.1016/j.compbiomed.2018.10.012.

[96] Dimitris Mitsouras, Peter Liacouras, Amir Imanzadeh, Andreas A. Giannopoulos, Tianrun Cai, Kanako K. Kumamaru, Elizabeth George, Nicole Wake, Edward J. Caterson, Bohdan Pomahac, Vincent B. Ho, Gerald T. Grant, and Frank J. Rybicki. Medical 3D printing for the radiologist. *RadioGraphics*, 35(7):1965–1988, November 2015. doi: 10.1148/rg.2015140320.

[97] Coenraad F. A. Moorrees and Martin R. Kean. Natural head position, a basic consideration in the interpretation of cephalometric radiographs. *American Journal of Physical Anthropology*, 16(2):213–234, June 1958. doi: 10.1002/ajpa.1330160206.

[98] Ana Morais, Jan Egger, and Victor Alves. Automated computer-aided design of cranial implants using a deep volumetric convolutional denoising autoencoder. In *Advances in Intelligent Systems and Computing*, pages 151–160. Springer International Publishing, 2019. doi: 10.1007/978-3-030-16187-3_15.

[99] Olivia Nackaerts, Frederik Maes, Hua Yan, Paulo Couto Souza, Ruben Pauwels, and Reinhilde Jacobs. Analysis of intensity variability in multislice and cone beam computed tomography. *Clinical Oral Implants Research*, 22(8):873–879, January 2011. doi: 10.1111/j.1600-0501.2010.02076.x.

[100] Olivia Nackaerts, Maarten Depypere, Guozhi Zhang, Bart Vandenberghe, Frederik Maes, and Reinhilde Jacobs and. Segmentation of trabecular jaw bone on cone beam CT datasets. *Clinical Implant Dentistry and Related Research*, 17(6):1082–1091, March 2014. doi: 10.1111/cid.12217.

[101] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 424–432. Springer International Publishing, 2016. doi: 10.1007/978-3-319-46723-8_49.

[102] William C. H. Parr, Joshua L. Burnard, Peter John Wilson, and Ralph J. Mobbs. 3D printed anatomical (bio)models in spine surgery: clinical benefits and value to health care providers. *Journal of Spine Surgery*, 5(4):549–560, December 2019. doi: 10.21037/jss.2019.12.07.

[103] Christian Payer, Darko Štern, Horst Bischof, and Martin Urschler. Regressing heatmaps for multiple landmark localization using CNNs. In *Medical Image*

Computing and Computer-Assisted Intervention – MICCAI 2016, pages 230–238. Springer International Publishing, 2016. doi: 10.1007/978-3-319-46723-8_27.

[104] Pedro Pimentel, Angelika Szengel, Moritz Ehlke, Hans Lamecker, Stefan Zachow, Laura Estacio, Christian Doenitz, and Heiko Ramm. Automated virtual reconstruction of large skull defects using statistical shape models and generative adversarial networks. In *Towards the Automatization of Cranial Implant Design in Cranioplasty*, pages 16–27. Springer International Publishing, 2020. doi: 10.1007/978-3-030-64327-0_3.

[105] Rosalie Plantefève, Igor Peterlik, Nazim Haouchine, and Stéphane Cotin. Patient-specific biomechanical modeling for guidance during minimally-invasive hepatic surgery. *Annals of Biomedical Engineering*, 44(1):139–153, August 2015. doi: 10.1007/s10439-015-1419-z.

[106] Pathmanathan Raghavan, David Bulbeck, Gayathiri Pathmanathan, and Suresh Kanta Rathee. Indian craniometric variability and affinities. *International Journal of Evolutionary Biology*, 2013:1–25, December 2013. doi: 10.1155/2013/836738.

[107] F. Rengier, A. Mehndiratta, H. von Tengg-Kobligk, C. M. Zechmann, R. Unterhinninghofen, H.-U. Kauczor, and F. L. Giesel. 3D printing based on imaging data: review of medical applications. *International Journal of Computer Assisted Radiology and Surgery*, 5(4):335–341, May 2010. doi: 10.1007/s11548-010-0476-x.

[108] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Lecture Notes in Computer Science*, pages 234–241. Springer International Publishing, 2015. doi: 10.1007/978-3-319-24574-4_28.

[109] Supakit Rooppakhun, Nattapon Chantarapanich, and Kriskrai Sitthiseripratip. Advanced medical imaging and reverse engineering technologies in craniometric study. In *Forensic Medicine - From Old Problems to New Challenges*. InTech, September 2011. doi: 10.5772/22792.

[110] Kelli Rudman, Craig Hoekzema, and John Rhee. Computer-assisted innovations in craniofacial surgery. *Facial Plastic Surgery*, 27(04):358–365, July 2011. doi: 10.1055/s-0031-1283054.

[111] Heiko Seim, Dagmar Kainmueller, Markus Heller, Hans Lamecker, Stefan Zachow, and Hans-Christian Hege. Automatic segmentation of the pelvic bones from CT data based on a statistical shape model, 2008.

[112] Wiebke Semper-Hogg, Marc Anton Fuessinger, Steffen Schwarz, Edward Ellis, Carl-Peter Cornelius, Florian Probst, Marc Christian Metzger, and Stefan Schlager. Virtual reconstruction of midface defects using statistical shape models. *Journal of Cranio-Maxillofacial Surgery*, 45(4):461–466, April 2017. doi: 10.1016/j.jcms.2016.12.020.

[113] Abhishek Sharma, Oliver Grau, and Mario Fritz. Vconv-dae: Deep volumetric shape learning without object labels. In Gang Hua and Hervé Jégou, editors, *Computer Vision – ECCV 2016 Workshops*, pages 236–250, Cham, 2016. Springer International Publishing. ISBN 978-3-319-49409-8.

[114] Ankit Sharma. Neuroprosthetic rehabilitation of acquired skull defects. *International Journal of Prosthodontics and Restorative Dentistry*, 1(1):65–70, 2011. doi: 10.5005/jp-journals-10019-1012.

[115] Haochen Shi and Xiaojun Chen. Cranial implant design through multiaxial slice inpainting using deep learning. In *Towards the Automatization of Cranial Implant Design in Cranioplasty*, pages 28–36. Springer International Publishing, 2020. doi: 10.1007/978-3-030-64327-0_4.

[116] Chong Chiet Sing, Lee Heow Pueh, and A Senthil Kumar. Automatic hole repairing for cranioplasty using bézier surface approximation. *Journal of Craniofacial Surgery*, 16(6):1076–1084, November 2005. doi: 10.1097/01.scs.0000180009.75857.81.

[117] Mayfield Brain & Spine. Craniotomy, Craniectomy | Mayfield Brain & Spine, Cincinati, OH. URL https://www.mayfieldclinic.com/pe-craniotomy.htm. Accessed: 2021-04-04.

[118] David Stutz and Andreas Geiger. Learning 3D shape completion from laser scan data with weak supervision. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, June 2018. doi: 10.1109/cvpr.2018.00209.

[119] Shiliang Sun, Zehui Cao, Han Zhu, and Jing Zhao. A survey of optimization methods from a machine learning perspective. *IEEE Transactions on Cybernetics*, pages 1–14, 2020. doi: 10.1109/tcyb.2019.2950779.

[120] Piotr Szymor, Marcin Kozakiewicz, and Raphael Olszewski. Accuracy of open-source software segmentation and paper-based printed three-dimensional models. *Journal of Cranio-Maxillofacial Surgery*, 44(2):202–209, February 2016. doi: 10.1016/j.jcms.2015.11.002.

[121] Philip Tack, Jan Victor, Paul Gemmel, and Lieven Annemans. 3D-printing techniques in a medical setting: a systematic literature review. *BioMedical Engineering OnLine*, 15(1), October 2016. doi: 10.1186/s12938-016-0236-4.

[122] Hilbrand van de Belt, Daniëlle Neut, Willem Schenk, Jim R van Horn, Henny C van der Mei, and Henk J Busscher. Infection of orthopedic implants and the use of antibiotic-loaded bone cements: A review. *Acta Orthopaedica Scandinavica*, 72(6):557–571, January 2001. doi: 10.1080/000164701317268978.

[123] Sophie E.C.M. van de Vijfeijken, Tijmen J.A.G. Münker, Rene Spijker, Luc H.E. Karssemakers, William P. Vandertop, Alfred G. Becking, Dirk T. Ubbink, A.G. Becking, L. Dubois, L.H.E. Karssemakers, D.M.J. Milstein, S.E.C.M. van de Vijfeijken, P.R.A.M. Depauw, F.W.A. Hoefnagels, W.P. Vandertop, C.J. Kleverlaan, T.J.A.G. Münker, T.J.J. Maal, E. Nout, M. Riool, and S.A.J. Zaat. Autologous bone is inferior to alloplastic cranioplasties: Safety of autograft and allograft materials for cranioplasties, a systematic review. *World Neurosurgery*, 117:443–452.e8, September 2018. doi: 10.1016/j.wneu.2018.05.193.

[124] Maureen van Eijnatten, Juha Koivisto, Kalle Karhu, Tymour Forouzanfar, and Jan Wolff. The impact of manual threshold selection in medical additive manufacturing. *International Journal of Computer Assisted Radiology and Surgery*, 12(4):607–615, October 2016. doi: 10.1007/s11548-016-1490-4.

[125] Maureen van Eijnatten, Roelof van Dijk, Johannes Dobbe, Geert Streekstra, Juha Koivisto, and Jan Wolff. CT image segmentation methods for bone used in medical additive manufacturing. *Medical Engineering & Physics*, 51: 6–16, January 2018. doi: 10.1016/j.medengphy.2017.10.008.

[126] Matthias N. van Oosterom, Henk G. van der Poel, Nassir Navab, Cornelis J.H. van de Velde, and Fijs W.B. van Leeuwen. Computer-assisted surgery. *Current Opinion in Urology*, 28(2):205–213, March 2018. doi: 10.1097/mou.0000000000000478.

[127] Yary Volpe, Rocco Furferi, Lapo Governi, Francesca Uccheddu, Monica Carfagni, Federico Mussa, Mirko Scagnet, and Lorenzo Genitori. Surgery of complex craniofacial defects: A single-step AM-based methodology. *Computer Methods and Programs in Biomedicine*, 165:225–233, October 2018. doi: 10.1016/j.cmpb.2018.09.002.

[128] Gord von Campe and Karin Pistracher. Patient specific implants (PSI). In *Towards the Automatization of Cranial Implant Design in Cranioplasty*, pages 1–9. Springer International Publishing, 2020. doi: 10.1007/978-3-030-64327-0_1.

[129] Maximilian Eberhard Hermann Wagner, Jürgen Thomas Lichtenstein, Marcel Winkelmann, Hoen oh Shin, Nils-Claudius Gellrich, and Harald Essig.

Development and first clinical application of automated virtual reconstruction of unilateral midface defects. *Journal of Cranio-Maxillofacial Surgery*, 43 (8):1340–1347, October 2015. doi: 10.1016/j.jcms.2015.06.033.

[130] Weiyue Wang, Qiangui Huang, Suya You, Chao Yang, and Ulrich Neumann. Shape inpainting using 3d generative adversarial network and recurrent convolutional networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, October 2017. doi: 10.1109/iccv.2017.252.

[131] Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. 3DN: 3D deformation network. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2019. doi: 10.1109/cvpr.2019.00113.

[132] William J. Weadock, Curtis J. Heisel, Alon Kahana, and John Kim. Use of 3D printed models to create molds for shaping implants for surgical repair of orbital fractures. *Academic Radiology*, 27(4):536–542, April 2020. doi: 10.1016/j.acra.2019.06.023.

[133] Ting Wu, Martin Engelhardt, Lorenz Fieten, Aleksandra Popovic, and Klaus Radermacher. Anatomically constrained deformation for design of cranial implant: Methodology and validation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006*, pages 9–16. Springer Berlin Heidelberg, 2006. doi: 10.1007/11866565_2.

[134] Hitoshi Yamauchi, Stefan Gumhold, Rhaleb Zayer, and Hans-Peter Seidel. Mesh segmentation driven by gaussian curvature. *The Visual Computer*, 21 (8-10):659–668, September 2005. doi: 10.1007/s00371-005-0319-x.

[135] Varduhi Yeghiazaryan and Irina Voiculescu. Family of boundary overlap metrics for the evaluation of medical image segmentation. *Journal of Medical Imaging*, 5(01):1, February 2018. doi: 10.1117/1.jmi.5.1.015006.

[136] Futoshi Yokota, Toshiyuki Okada, Masaki Takao, Nobuhiko Sugano, Yukio Tada, Noriyuki Tomiyama, and Yoshinobu Sato. Automated CT segmentation of diseased hip using hierarchical and conditional statistical shape models. In *Advanced Information Systems Engineering*, pages 190–197. Springer Berlin Heidelberg, 2013. doi: 10.1007/978-3-642-40763-5_24.

[137] Hye Sun Yun, Chang Min Hyun, Seong Hyeon Baek, Sang-Hwy Lee, and Jin Keun Seo. Automated 3D cephalometric landmark identification using computerized tomography, 2020.

[138] Yuan Z. Zhang, Bin Chen, Sheng Lu, Yong Yang, Jian M. Zhao, Rui Liu, Yan B. Li, and Guo X. Pei. Preliminary application of computer-assisted

patient-specific acetabular navigational template for total hip arthroplasty in adult single development dysplasia of the hip. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 7(4):469–474, October 2011. doi: 10.1002/rcs.423.

[139] Lei Zhou, Yang Song, Jianhui Wu, Huilan Li, Guobin Zhang, and Chunling Sun. Surface reconstruction of bilateral skull defect prosthesis based on radial basis function. In *Lecture Notes in Electrical Engineering*, pages 741–747. Springer London, December 2012. doi: 10.1007/978-1-4471-4802-9_98.

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". It is available for LaTeX via CTAN as `classicthesis`.

Final Version as of June 14, 2021 (`classicthesis` version 1.0).