IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 May 2024, accepted 11 June 2024, date of publication 17 June 2024, date of current version 24 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3415173

RESEARCH ARTICLE

CraNeXt: Automatic Reconstruction of Skull Implants With Skull Categorization Technique

THATHAPATT KESORNSRI[1], NAPASARA ASAWALERTSAK[2], NATDANAI TANTISEREEPATANA[3], PORNNAPAS MANOWONGPICHATE[2], BOONRAT LOHWONGWATANA[4], CHEDTHA PUNCREOBUTR[4], TITIPAT ACHAKULVISUT[2], AND PEERAPON VATEEKUL[1], (Senior Member, IEEE)

[1]Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok 10330, Thailand
[2]Department of Biomedical Engineering, Faculty of Engineering, Mahidol University, Nakhon Pathom 73170, Thailand
[3]College of Biomedical Engineering, Rangsit University, Muang, Pathum Thani 12000, Thailand
[4]Department of Metallurgical Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok 10330, Thailand

Corresponding authors: Peerapon Vateekul (peerapon.v@chula.ac.th) and Titipat Achakulvisut (titipat.ach@mahidol.edu)

This work was supported in part by the joint research collaboration between Chulalongkorn University and Meticuly Company Ltd. under Grant RES-6629721021, in part by the Program Management Unit Competitiveness (PMUC) under Grant C10F650050, and in part by the National Science and Technology Development Agency (NSTDA) Supercomputer Center (ThaiSC) under Project lt200097-rskull.

ABSTRACT Automatic cranial implant design aims to design a patient-specific implant where various machine-learning-based skull reconstruction techniques have been introduced to predict the implant. Despite the significant progress made in the previous research, the existing techniques often struggle to generalize to diverse clinical cases and may not fully leverage the latest advancements in deep learning architectures. Moreover, the limited availability of large-scale clinical datasets hinders the development of the models. In this paper, we represent a novel skull reconstruction model, CraNeXt, which utilizes a ConvNeXt backbone to achieve a 5.8x reduction in size when compared to 3DUNetCNN without sacrificing reconstruction quality. In addition, we introduce a novel method, skull categorization, to classify unlabeled skulls and determine the location of defects and the distribution of skull areas. We expand the training dataset by incorporating a larger collection of 328 in-house clinical cases, enabling the model to better capture the diversity of real-world cranial defects. CraNeXt demonstrates superior results with the skull categorization technique, achieving a dice score of 0.7969±0.13 on both public and in-house data. We perform a qualitative assessment of the predicted implants and discuss potential improvements to the skull reconstruction toward clinical use cases.

INDEX TERMS Skull reconstruction, deep learning, skull categorization, autoimplant, volumetric shape completion.

I. INTRODUCTION

Skull reconstruction, or autoimplant, the process of generating 3D data of the cranial structure, has garnered significant attention due to its role in neurosurgery [1]. Traditional methods for skull reconstruction often rely on manual intervention, human expertise, and computationally expensive applications [2]. However, these methods are frequently time-consuming and sensitive to image quality [3], patient demographics [4], and pathological conditions [5], highlighting the necessity for modern and efficient methodologies.

Deep learning techniques for medical images, such as convolutional neural networks (CNNs) [6] and UNet architectures [7], have emerged as powerful tools for automating and enhancing skull reconstruction. By exploiting the hierarchical representations within neural networks, deep learning models learn complex patterns and features from large datasets.

The associate editor coordinating the review of this manuscript and approving it for publication was Jinhua Sheng.

This ability is especially advantageous for the complex and variable structures of the human skull. One common use case for deep learning models is to generate a complete skull from defected skulls, known as volumetric shape completion, and then subtract them to obtain a generated implant. The utilization of deep learning in skull reconstruction offers several potential benefits, including increased accuracy, reduced processing time, and enhanced adaptability to diverse clinical scenarios [1]. Most current state-of-the-art skull reconstruction models, as proposed in AutoImplant Challenges [5], [8], use conventional CNNs as a backbone. Recent developments in transformer architectures such as vision transformers (ViT) [9] and improved CNNs such as ConvNeXt [10] show the potential to improve performance and generalization in various computer vision tasks [11]. ConvNeXt incorporates key designs from ViT, such as larger kernel sizes, inverted bottlenecks, and improved normalization techniques, while maintaining the simplicity and efficiency of traditional convolutional networks. ConvNeXt has demonstrated outstanding performance on image classification [10], object detection, and semantic segmentation benchmarks, often outperforming CNNs and ViT. We hypothesize that improving the architecture of the model can lead to improvements in the skull reconstruction.

An additional difficulty in implant generation is the work toward clinical translation [1]. Addressing these challenges is essential to ensuring the clinical reliability of deep learning-based skull reconstruction methods. To improve on this, there is a need for large and well-curated clinical skull datasets [12], robustness to anatomical variations [5], and the interpretability of results [1]. The present issues include data heterogeneity [13], generalization [1], limitations of the model's interpretability [1], and the model's robustness [1]. While numerous methods have been proposed to generate synthetic defected skull datasets in the AutoImplant challenges [1], [5], [8], the lack of comprehensive and diverse datasets derived from actual patient cases hinders the refinement and validation of existing methodologies, impeding the clinical translation of skull reconstruction. This matter is complicated as certain organizations may have internal data; these datasets frequently lack standardization and fail to offer complete information on various defect areas. Here, we believe that incorporating real clinical datasets and publicly available datasets can improve the model and its usage in real clinical applications. In addition, adding information about defect areas may help understand the distribution of defects and improve clinical applications.

Within the scope of this paper, we incorporate recent developments in computer vision models to enhance the autoimplant models. Specifically, we use ConvNeXt backbones to improve skull reconstruction and make the skull reconstruction model more efficient. Below are the primary contributions of our research:

- We propose CraNeXt, a novel UNet-based architecture for skull reconstruction, inspired by the success of ConvNeXt [10], which was constructed entirely from standard ConvNet modules [6] without using any specialized attention-based blocks like in vision transformers.
- We apply a skull categorization to label each part of the skull by matching the unlabeled input skull to a template. We then utilize the categorization technique to calculate the distribution of defect areas and incorporate categorization features during the model training, improving implant generation.
- We introduce the Surface Hausdorff distance (SdH) metric, which measures the surface distance between the predicted and actual implant and can be used in actual clinical and 3D printing setups.
- We incorporate 570 skulls from the public synthetic dataset and 328 skulls from the in-house clinical dataset. Our proposed models demonstrate the ability to generalize and exhibit compatibility with synthetic and clinical data, indicating their potential applicability in clinical settings.
- We evaluate the performance of the proposed model on both public and in-house datasets. The proposed model with the skull categorization technique achieves superior dice scores of 0.7969±0.13.

Our paper is structured as follows: In Section II, we present the skull reconstruction task and review related works. In Section III, we provide a thorough explanation of the entire pipeline and the proposed methods. We provide the evaluation criteria and experimental settings in Sections IV and V. The results of the experiment are detailed in Section VI. The discussion and conclusion are covered in Sections VII and VIII.

II. RELATED WORK

A. OVERVIEW OF SKULL RECONSTRUCTION

Skull reconstruction, or autoimplant, encompasses a range of methods and techniques aimed at rebuilding or restoring the skull (Fig. 1). The process begins with data acquisition, where we acquire an image of a defective skull from the patient using CT imaging [14]. Skull segmentation extracts the skull from a CT image [15]. It is registered in the desired format, including orientation, position, space dimensions, and size [16]. Skull categorization can also be applied to the input skull to categorize the skull parts based on their location (Section III-C). Then, skull reconstruction aims to predict implants using a technique such as volumetric shape completion [17]. A volumetric shape completion predicts the complete skull to fill in missing portions of 3D skull data in voxel-space [1] or point cloud space [18], [19] using conventional machine learning [8] or deep-learning approaches [19], [20], [21], [22], [23], [24]. A reconstructed or predicted-complete skull is subtracted from the defective skull to get an implant. Afterward, some post-processing can be involved to clean an implant, such as manual removal [13] or automatic removal of remaining voxels [22]. Ultimately, the implant can be converted into a 3D shape format for

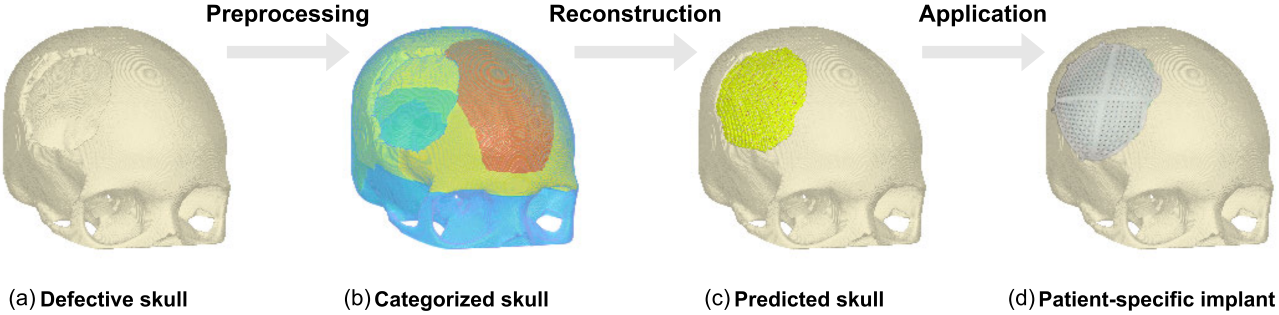

FIGURE 1. Overall procedures for the proposed skull reconstruction. In (b), different colors represent different skull regions. The colors in (c) and (d) represent the predicted implant and final titanium implant, respectively.

patient-specific implant design and then 3D printed for surgical procedures [2].

B. AUTOMATIC CRANIAL IMPLANT DESIGN: AUTOIMPLANT CHALLENGES

The automatic skull reconstruction has been brought to the attention of the biomedical deep learning community by the AutoImplant challenges [1], [5], [8]. Deep learning models have demonstrated their efficacy in reconstructing the initial shape of the skull. The AutoImplant I challenge (2020) formulated cranial implant design as a shape completion problem [1]. Both conventional and deep learning models are proposed for the challenge, which play roles as baseline approaches for skull reconstruction tasks. Ellis et al. [20], [21] proposed 3DUNetCNN, the U-Net-style convolutional neural network with residual connection [25], inspired by the AutoEncoder model for 3D tumor segmentation [26], [27], has an increased input size from 128 × 128 × 128 voxels to 176 × 224 × 144 voxels. The model consists of 2 blocks of ResNet [25] with a base width of 32 channels and a depth of 5 layers. The output of the final decoding layer went through a 1 × 1 × 1 convolution and sigmoid activation (Fig. 3(a)). This paper also uses the data augmentation technique to increase the dataset from 100 training sets of images to 9,900 additional training images using Advanced Normalization Tools (ANTs) [28] to calculate non-linear symmetric image normalization (SyN), warping transformations between original dataset pairs. Yu et al. [13] proposed PCA-skull, a data-driven approach using principal components analysis (PCA) to describe the shape of healthy human skulls. With the assumption that defective skulls and healthy skulls have similar shape distributions in a common principal component (PC) space, a defect would not alter the shape distribution of a human skull significantly in a compact PC space. Applying inverse PCA to the defective skull should result in a healthy version of the skull. To obtain the final implants, a subtraction operation between the reconstructed healthy cranium and the defect skulls is performed. The AutoImplant II challenge (2021) includes a more diverse synthetic dataset and focuses mainly on the clinical applicability of deep learning models. In the evaluation phase, submissions were quantitatively and qualitatively evaluated by experts using real clinically defective skulls. Wodzinski et al. [22] proposed two-step U-Net-like networks with 3D convolutional residual blocks. The first reconstruction network utilizes the preprocessed defective skull as input and generates the predicted defect as output. Then the second Variational AutoEncoder (VAE) network will perform the refinement process by smoothing the predicted defect and recovering the fine details. The Wodzinski U-Net model outperformed 3DUNetCNN with a smaller model architecture. In addition to the methods proposed in AutoImplant challenges, Friedrich et al. [19] proposed a point-cloud diffusion model for skull reconstruction in point-cloud space, which requires additional computational steps to convert from point-cloud to voxel space.

C. MODERN CNN ARCHITECTURES

Islam et al. [23] proposed a 3D attention-based U-Net architecture for brain tumor segmentation and survival prediction from MRI scans. They integrate channel and spatial attention mechanisms into the decoder blocks of the 3D U-Net to enhance segmentation performance. The attention module consists of parallel channel and spatial attention branches, along with a skip connection to reduce feature redundancy and sparsity. Their experiments demonstrate improved segmentation accuracy compared to the standard 3D U-Net. This work highlights the potential of integrating attention mechanisms into 3D U-Net. Liu et al. [10] proposed the ConvNeXt. It is a convolutional neural network (CNN) architecture designed to compete favorably with hierarchical vision transformers across multiple computer vision benchmarks while retaining the simplicity and efficiency of standard CNNs. Across multiple tasks like ImageNet classification, object detection, and semantic segmentation, ConvNeXt achieves competitive or even better performance than similarly-sized hierarchical vision transformers [9] like swin transformers [29]. The primary differences between the ResNet and ConvNeXt blocks are the elimination of batch normalization in favor of layer normalization and the widening of the convolutional stride. Lastly, Woo et al. [30] proposed the ConvNeXt predecessor, ConvNeXtV2, and introduced new normalization techniques by replacing layer scaling before the skip connection with Global Response Normalization (GRN) to enhance inter-channel feature competition. In essence, ConvNeXt demonstrates that a CNN architecture, when designed properly, can be as powerful and scalable as hierarchical vision transformers, challenging the notion that attention-based architectures are naturally superior for vision tasks.

By synthesizing these related works, our proposed approach attempts to address the complexities of skull reconstruction with a larger quantity of clinical data and explore modern CNN architectures to improve skull reconstruction. We consider 3DUNetCNN [21], PCA-skull [13], and Wodzinski UNet [22] as baselines and utilize ConvNeXt architecture [10] with GRN [30] for the proposed backbone.

III. METHODOLOGY

Here, we discuss our process for skull reconstruction. The process starts with acquiring data and standardizing data configurations from open and in-house clinical data sources. Data preprocessing includes skull registration (Fig. 2a), ensuring that the data for training are consistent regardless of the original configurations of the skulls. Additional data preprocessing (Fig. 2b) consists of proposed skull categorization, normalizing space dimensions between datasets, and resizing. The purpose of skull categorization is to register the skull template containing classified anatomical location with the input skull, providing explicit contextual information to the model. This distribution can be used to improve the explainability and analysis of defect distributions. Input resizing and foreground cropping ensure computational efficiency and the limitations of GPU memory. We train our proposed deep learning model, CraNeXt, using categorized skulls from data preprocessing to predict complete skulls (Fig. 2c). We then substitute a complete skull with a defective skull to get a predicted implant. Then, we apply post-processing techniques, including noise removal using erosion and dilation to remove small voxels, and select the largest connected components to remove irrelevant voxels after subtraction. The post-processing can improve the anatomical accuracy, smoothness, and compatibility of the predicted skull, improving the raw output of the model to a clinically usable implant. We evaluate output implants using dice coefficient similarity and Hausdorff distance. Lastly, implants are integrated into an existing clinical workflow for visualization and utilization of predicted skulls to manufacture patient-specific titanium implants (Fig. 2d).

A. DATASET

Our study uses two primary datasets for the development and evaluation of automated cranial implant design: (1) SkullBreak, which is a synthetic dataset, and (2) an in-house clinical dataset. Using a combination of these datasets enables a detailed evaluation of the proposed methods.

SkullBreak is a synthetically defective skull dataset generated from the CQ500 head CT collection originated by Kodym et al. [16]. The CQ500 collection [31] was created by the Centre for Advanced Research in Imaging, Neurosciences, and Genomics (CARING) in New Delhi, India, licensed under CC BY-NC-SA 4.0. SkullBreak is the primary dataset for skull reconstruction tasks in MedShapeNet [32], which proposes a large collection of 3D anatomical shapes such as bones, organs, and skulls, suggesting the importance of exploring tasks such as reconstruction in a larger and more diverse corpus.

SkullBreak consists of 570 training samples of 114 unique skulls, each with a defective skull, a corresponding complete skull, and the associated implant. The SkullBreak dataset has 512 × 512 × 512 voxels with a space dimension of 0.4 × 0.4 × 0.4 mm. Skulls were extracted and aligned to the Frankfort horizontal plane using a rigid transformation based on four anatomical landmarks. Finally, artificial defects were injected into the entire skull by subtracting randomly generated shapes, providing five types of defects: unilateral parieto-temporal, unilateral fronto-orbital, bilateral, and two random defects. The defect borders were smoothed with morphological operations to simulate current bone remodeling processes. The SkullBreak dataset provides a diverse range of synthetic skull defects across various defect shapes, sizes, and locations.

An "in-house" clinical dataset consists of 328 pairs of binary volumetric skulls and implants, acquired from patients who underwent implant modeling before cranioplasty procedures at Meticuly Co., Ltd. [33] under patient consent. Since this is a retrospective study, it does not require approval from an Institutional Review Board (IRB). All datasets are anonymized and renumbered to ensure the non-specification of patient information. The dataset was collected and anonymized to protect patient privacy, ensuring compliance with ethical guidelines for the use of medical data in research. The CT dataset has a fixed size of 500 × 530 × 465 voxels with a space dimension of 0.5 × 0.5 × 0.625 mm. The dataset includes a diverse range of cranial defects resulting from various etiologies, such as craniotomies due to brain tumors, traumatic brain injuries, and decompressive craniotomies. The binary skulls were preprocessed by manually segmenting the CT scans of patients, aligning to the skull template using rigid transformation, resampling the skull data into a 0.4 × 0.4 × 0.4 mm space dimension, and creating patient-specific implants under the supervision of experienced biomedical engineers.

To further standardize the alignment across both datasets, we perform additional affine registration and cropping on our in-house dataset to match the alignment of the SkullBreak dataset, ensuring alignment consistency between datasets. By integrating SkullBreak and in-house datasets, we have created a rich and diverse dataset that combines the strengths of both synthetic and clinical data and ensures that the two datasets are consistent and compatible.

Table 1 presents the distinctions and dissimilarities between the two datasets, including the file format, the process to label implants either by directly subtracting an implant from a healthy skull or by manually designing an implant from an actual defective skull, and dimensions obtained from CT imaging data. We measure the voxel

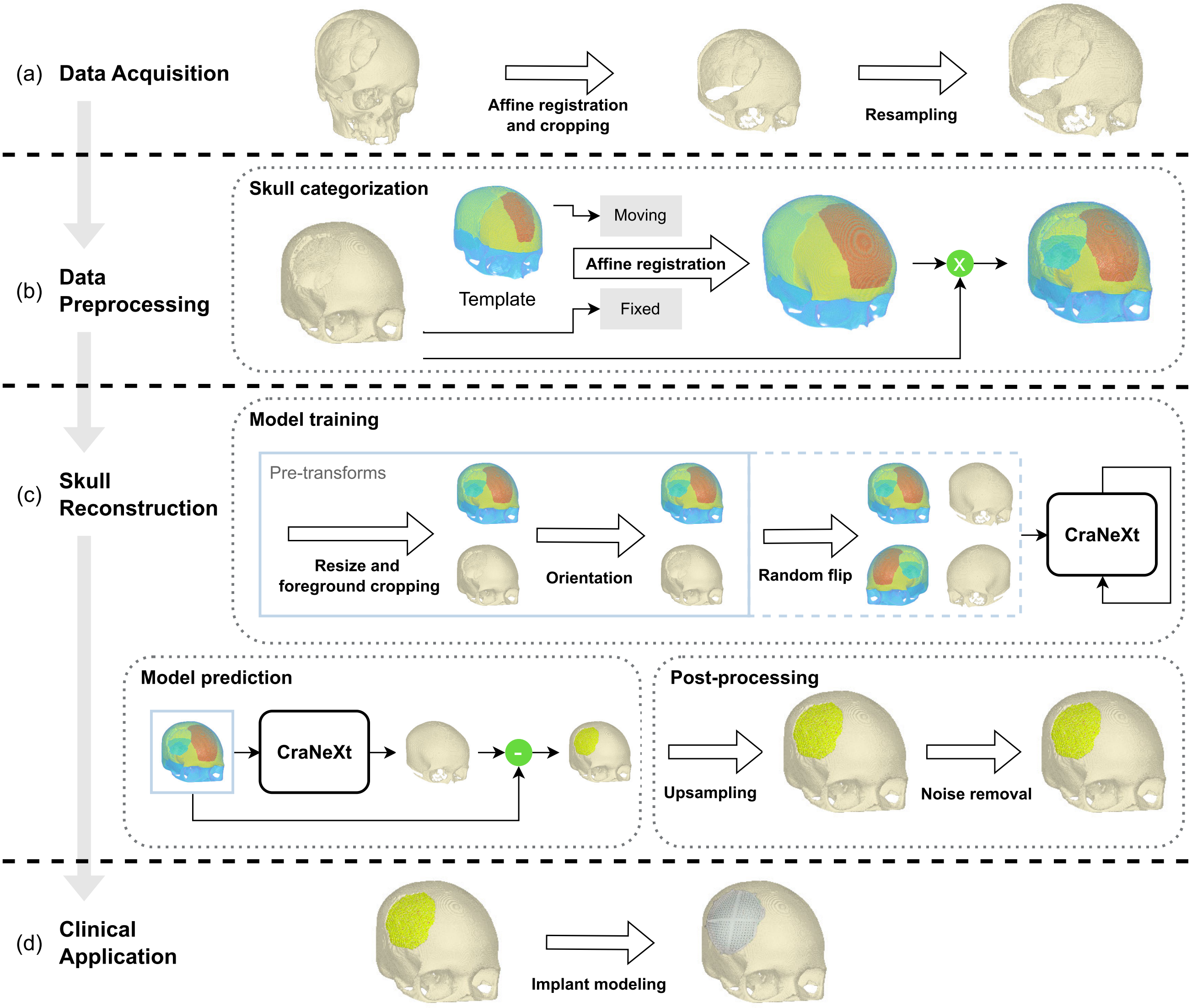

FIGURE 2. Overall diagram for proposed methods.

occupancy rate (VOR), the ratio of implants to complete skulls, which can be used as an approximation of the average volume to be filled by the model. The average VOR for complete skulls in the SkullBreak dataset is 5.21%, which is higher than the average VOR of 4.08% for complete skulls in the in-house dataset. However, the SkullBreak dataset has an average implant to complete skull VOR of 0.09%, while the in-house dataset has a higher average implant to complete skull VOR of 0.18%. This difference suggests that the implants in the in-house dataset tend to occupy a larger proportion of the skull volume compared to the implants in the SkullBreak dataset, indicating a higher implant volume to be predicted in clinical datasets.

B. CRANEXT: A PROPOSED MODEL

Deep learning models' efficiency and practicality are critical in the field of medical image processing, especially for tasks like skull reconstruction. Despite their promising results, advanced topologies like 3D U-Net have high computing costs and many parameters, which may limit their use in resource-constrained clinical situations. Furthermore, models with a large number of parameters are more likely to overfit, especially when working with small medical datasets. Therefore, there is a growing need for efficient and lightweight models that can achieve comparable or even better performance than their more complex counterparts.

In this section, we propose CraNeXt, a skull reconstruction model with fewer parameters and greater efficiency. As illustrated in Fig. 3, the CraNeXt model maintains the same architecture as 3DUNetCNN but with a modified stage compute ratio. The motivation behind changing the encoder stage ratio is introduced by the macro design of ConvNeXt, which aims to allocate more computational resources and representation power to the deeper stages of the encoder.

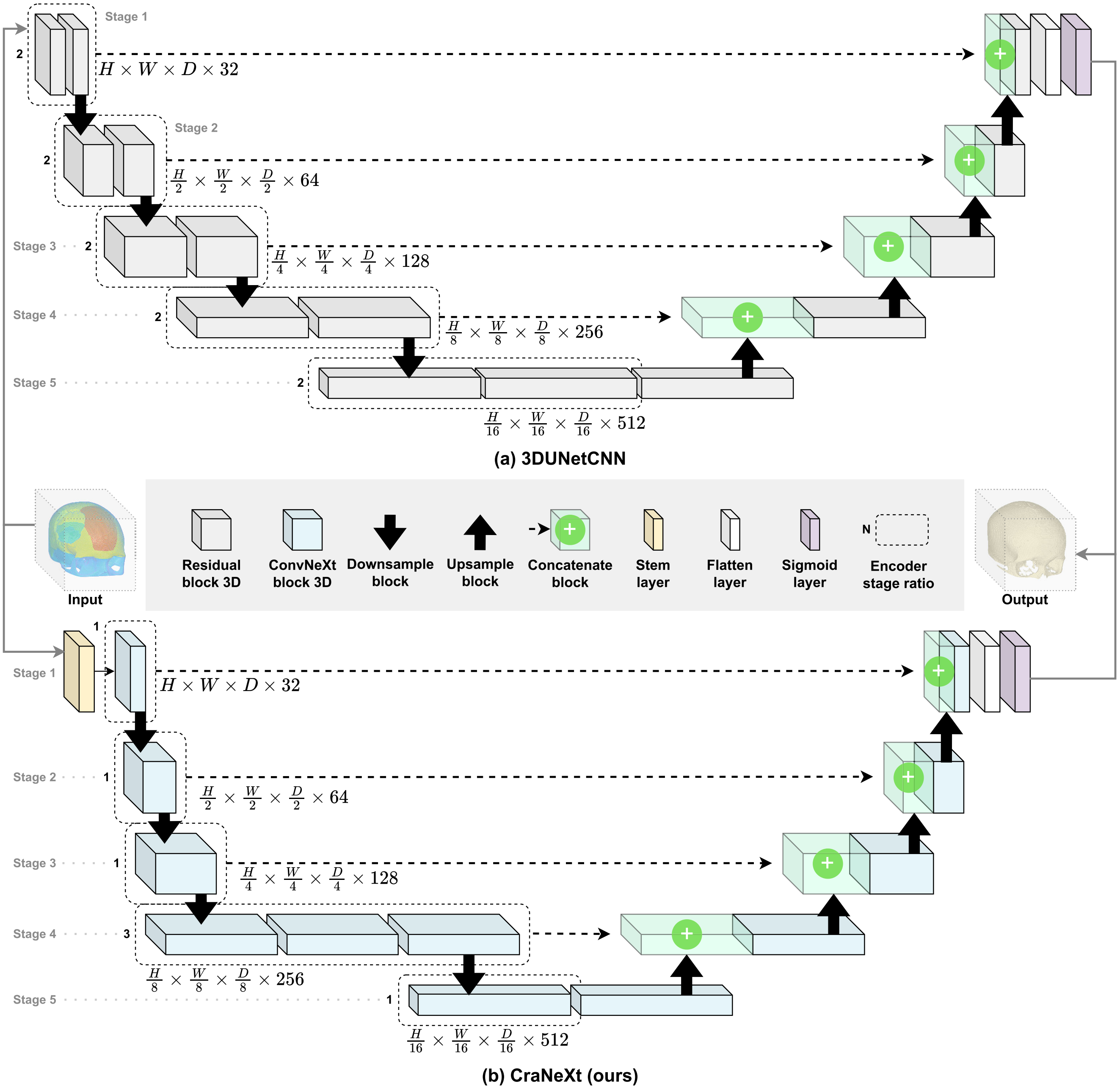
Stage 1
2
$H \times W \times D \times 32$
Stage 2
2
$\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times 64$
Stage 3
2
$\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 128$
Stage 4
2
$\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times 256$
Stage 5
2
$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 512$
(a) 3DUNetCNN
Input
Residual block 3D
ConvNeXt block 3D
Downsample block
Upsample block
Concatenate block
Stem layer
Flatten layer
Sigmoid layer
N
Encoder stage ratio
Output
1
Stage 1
$H \times W \times D \times 32$
Stage 2
1
$\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times 64$
Stage 3
1
$\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 128$
Stage 4
3
$\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times 256$
Stage 5
1
$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 512$
(b) CraNeXt (ours)

FIGURE 3. (a) The 3DUNetCNN architecture; and (b) Our proposed CraNeXt architecture, replacing residual block 3D with ConvNeXt block 3D and adjusting encoder stage ratios.

By reducing the number of ConvNeXt blocks in the earlier stages and increasing the number of blocks in the fourth stage (Fig. 3b), we allow the network to learn more complex and discriminative features at higher resolutions. We also introduce the stem layer to the model architecture. The stem layer consists of a convolutional layer with a kernel size and a stride equal to the patch size, designed to efficiently process and downsample the input data before passing it to the subsequent encoder stages. For the backbone part, we adapt the core backbone of 3DUNetCNN [21] from a modified 3D residual block of ResNet [25] to a 3D ConvNeXt block [10], [30] (Fig. 4). By leveraging the parameter-efficient design of ConvNeXt, we aim to create a more lightweight and computationally efficient model for skull reconstruction. The integration of ConvNeXt blocks into the 3D U-Net architecture allows us to benefit from the advantages of both the U-Net's hierarchical structure for capturing multi-scale features and ConvNeXt's efficient feature extraction capabilities. We replace the original convolution with a depthwise convolution with a bigger kernel size, resulting in a bigger receptive field and more contextual information. We apply the Global Response Normalization

TABLE 1. Dataset specification and format.

Dataset	SkullBreak	In-house
# Unique skull	114	328
Defective skull	570	328
Data type	3D Binary voxel	3D Binary voxel
Size	512 × 512 × 512	500 × 530 × 465
Space dimension	0.4 × 0.4 × 0.4 mm	0.5 × 0.5 × 0.625 mm
Implant label	Generated synthetically by subtracting an implant from a healthy skull	Manufactured an implant from a defective skull using 3D software
Format	NRRD	NIfTI
Defect type	Single-label (Manual)	Multi-label (Categorization technique)
Defect definition	Bilateral Fronto-orbital Parietotemporal Random type 1 Random type 2	Maxilla & Mandible Temporal Occipital Lower Parietal Middle Parietal Upper Parietal Lower Frontal Upper Frontal Undefined
Voxel occupancy rate (VOR) [%]		
Complete skull	5.2112±0.77	4.0825±0.88
Implant	0.4977±0.28	0.7536±0.41
Implant/complete	0.0955±0.37	0.1846±0.46

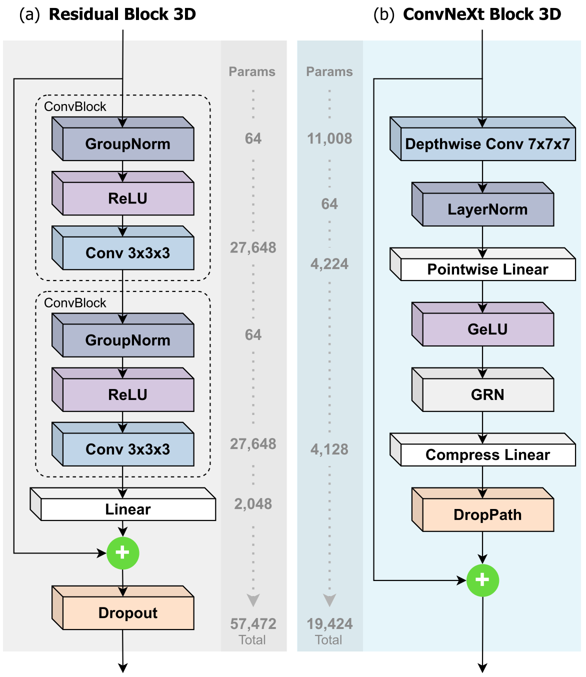

FIGURE 4. (a) Block Design for Residual Block 3D (b) Block design for ConvNeXt block 3D. And parameter comparison of 3D input of size 32 x 32 x 32. The ConvNeXt block 3D differs from the residual block 3D by employing depthwise convolutions with a larger receptive field and inverted bottleneck structure, a normalization layer, and a GeLU activation layer.

(GRN) layer [30] after the GELU activation layer to perform global aggregation of features. Additionally, we incorporate DropPath regularization within the ConvNeXt block to randomly drop out entire paths within the block during training, forcing the network to learn redundant and diverse features. The output of the ConvNeXt block is then added to the input through a residual connection, similar to the original residual block. Furthermore, we also compare CraNeXt to the traditional 3D attention U-Net architecture [23] in the experiment section to assess the performance between attention-based and convolutional architectures.

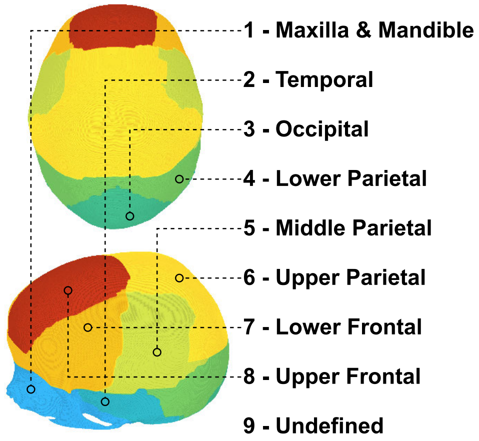

FIGURE 5. Categorized skull template.

C. SKULL CATEGORIZATION

The skull categorization approach aims to provide the skull data with additional semantic information, enabling it to learn more precise and context-aware reconstructions. We hypothesize that the model can learn to handle defects depending on their locations, e.g., by applying different strategies for reconstructing defects in the frontal bone versus the parietal bones. Here, we propose a skull categorization technique to label the binary skull with the different anatomical regions of the skull. We transform it into multi-label skull data before inputting it into the CraNeXt model.

We first create the categorized skull template using 3DSlicer software [34] by manually segmenting the healthy skull into 8 regions (Maxilla & Mandible, Temporal, Occipital, Lower Parietal, Middle Parietal, Upper Parietal, Lower Frontal, and Upper Frontal) based on cranial bone anatomy and directly assigning label values into the skull, converting skull data from binary to multi-label values ranging from 1 to 8, as shown in Fig. 5. Secondly, we apply the ANTs [28] 3D affine registration function to the binary input skull as a fixed voxel and the template skull as a moving voxel. The registration method will align the skull template to match the input skull's scale and offset. However, the affine registration is robust in terms of accuracy, but it is insufficient to ensure that the registered skull template completely covers the area of the fixed input skull, resulting in an undefined label (value 9) in the registered skull. We use scikit-image's dilation morphology [35] to enlarge registered skulls, minimizing the non-coverage area before final label mapping. This conventional dilation technique was chosen as the dilated template skull significantly fills the entire binary skull and makes efficient computation. Lastly, we categorize the binary skull input by performing a dot-product operation, mapping non-zero values of the input voxel to the value of the skull template.

In summary, categorized skull data can facilitate the design of modular or multi-part implants, where each part is optimized separately based on its anatomical characteristics. Incorporating categorized skull part input can provide additional anatomical context, improve defect localization, and facilitate more precise and anatomically accurate reconstructions.

IV. EVALUATIONS

A. EVALUATION METRICS

To assess the binary 3D implants generated by skull reconstruction techniques, we leverage quantitative evaluation metrics including the dice similarity coefficient (DSC), the 95$^{\text{th}}$ percentile Hausdorff distance (dH95), proposed surface Hausdorff (SdH), and the border DSC (bDSC) [1]. We also incorporate the false positive rate (FPR) and the false negative rate (FNR) into the metrics. More DSC and bDSC mean better-generated implants compared to the ground truth. In contrast, a lower dH95, FPR, and FNR indicate that the generated implant closely matches the ground truth.

The voxel-by-voxel binary analysis of the complete or labeled implant (y) and predicted implant ($\hat{y}$) is mainly used for the evaluation, including

$$TP = y \land \hat{y} \tag{1}$$

$$FP = (y \lor \hat{y}) \land \neg y \tag{2}$$

$$TN = \neg y \land \neg \hat{y} \tag{3}$$

$$FN = (y \lor \hat{y}) \land \neg \hat{y} \tag{4}$$

where conjunction ($\land$), disjunction ($\lor$), and negation ($\neg$) are boolean algebra operations.

The dice similarity coefficient (DSC) measures the similarity between two 3D volumes. It assesses the overlap between the predicted and ground-truth implant volumes and ranges from 0 to 1, where 1 indicates perfect overlap. DSC can be calculated as,

$$DSC = \frac{2TP}{2TP + FP + FN}. \tag{5}$$

The boundary dice coefficient (bDSC) (6) measures the DSC between the borders of the implants [1] where

$$I_B = \begin{cases} I, & dt \leqslant d \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

represents the implant border. $dt = EDT(D)$ is the Euclidean distance transform (EDT) of the defective skull D, I is the implant, and d is a distance parameter. We follow the AutoImplant challenge [1] by setting a distance parameter d equal to 10.

False positive rate (FPR) and false negative rate (FNR) are used to evaluate the performance of binary classification models. We utilize FPR to measure additional predicted voxels and FNR to measure missing regions. FPR and FNR can be calculated as

$$FPR = \frac{FP}{FP + TN} \quad \text{and} \quad FNR = \frac{FN}{FN + TP}. \tag{7}$$

The Hausdorff distance (dH) [36] metric calculates the maximum distance between the predicted and ground truth points, providing insight into the model's ability to capture surface differences between predicted and labeled implants on an actual millimeter scale. Given two finite point sets from a predicted implant A and a ground truth implant B,

$$A = \{a_1, \ldots, a_p\} \text{ and } B = \{b_1, \ldots, b_q\}, \tag{8}$$

the Hausdorff distance is defined as

$$dH(A, B) = \max\{h(A, B), h(B, A)\} \tag{9}$$

where

$$h(A, B) = \max_{a \in A} \min_{b \in B} \|a - b\|. \tag{10}$$

However, Hausdorff distance (dH) alone may provide the most clinical assessment as it applies to the entire surface of predicted implants. To address this limitation, we propose the surface Hausdorff distance (SdH), which calculates the Hausdorff distance exclusively on the surface of the ground truth and predicted implant. By focusing on the surface area, SdH provides a meaningful assessment aligned with the actual printing and evaluation process of the implant. To compute SdH, we employ a parallel projection algorithm that extracts the 3D implant surface from five directions: superior to inferior, left to right, right to left, anterior to posterior, and posterior to anterior [37] and calculate dH based on these extracted surfaces instead. The surface extraction method is defined in Algorithm 1 in the Appendix.

$$SdH(A, B) = \max\{h(S_A, S_B), h(S_B, S_A)\} \tag{11}$$

where h is (10) and S_A, S_B are the projected surfaces of the predicted implant A and ground truth implant B, respectively.

B. QUALITATIVE EVALUATION

Performing the qualitative analysis is crucial for clinical implant generation by offering insights beyond quantitative metrics, especially concerning implant morphologies. To conduct the evaluation, we utilize 3D Slicer software [34] to compare the 3D shape of predicted and generated implants. We use 5 qualitative criteria adapted from prior research [38] to evaluate the predicted implant, including completeness, no false positive area, restored skull shape, smooth transition with the skull, and minimal thickness (Table 2). The first row of Figure 7 presents the graphical examples for each criterion. Additionally, we explore errors based on the characteristics of skulls to further understand the potential clinical variability of the dataset.

TABLE 2. Implant prediction qualitative criteria.

Criteria	Description
Complete	Assesses whether the implant covers the entire defective area without any gaps or holes.
No false positive area	Evaluates whether the implant is limited to the defective area and does not extend beyond it.
Restored skull shape	Determines if the reconstructed implant has a similar shape to the ground truth.
Smooth transition	Examines the smoothness of the transition between the defective skull and the reconstructed implant.
Minimal thickness	Assesses whether the thickness of the reconstructed implant is acceptable and consistent with the defective skulls.

V. EXPERIMENTAL SETTINGS

We conducted 3 experiments to address our contributions, as follows:

- **Experiment 1** aims to assess the enhanced backbone architecture compared to previous models. We compare the performances of the proposed backbone architecture with other state-of-the-art methods, including PCA-skull [13], a machine learning-based 3DUNetCNN [21] (baseline), and Wodzinski et al. [22] UNet. Furthermore, we apply the attention-based mechanism [23] to the baseline for better comparison to ConvNeXt approaches. All methods are trained with the same experimental configurations on the SkullBreak dataset and evaluated on the combined SkullBreak and in-house datasets (Table 3).
- **Experiment 2** aims to assess the generalizability of the model by comparing synthetic data with actual clinical data. To evaluate our proposed model's generalizability and find optimal utilization for different datasets, we conduct four distinct assessments utilizing both SkullBreak and in-house datasets (Table 4).
- **Experiment 3** aims to improve CraNeXt's performance and investigate the impact of proposed skull categorization (Table 5). We categorized 730 defective skulls from both the Skullbreak public and in-house datasets, creating a new dataset called the categorized dataset. We then train CraNeXt on both the original dataset and the categorized dataset and compare their performance on the combined Skullbreak and in-house test datasets.

The top-performing model from experiment 1 compared to the baseline will serve as the base model for experiments 2-3, while the optimal dataset configuration from experiment 2 will be utilized in experiment 3. Furthermore, across 3 experiments, the results are analyzed with a one-sided paired t-test to measure the differences of DSC, bDSC, dH95, and SdH on the test set between major approaches (Table 6). Section VI(A-C) will cover the results of experiments 1-3, accordingly.

For the implementation of the proposed experiments, we use TorchIO [39] to load and format datasets. The datasets are registered to a common reference space via rigid transformation (translation and rotation) using MONAI [40] built-in transform functions. The model training is done using the MONAI framework [40]. Then, the SkullBreak and in-house clinical datasets are randomly split into training, validation, and test sets with ratios of 470:50:50 and 260:34:34, respectively. We use the input voxel size of 176 × 224 × 144, which is the same as in 3DUNetCNN [21]. After autoimplant prediction, we utilize scikit-image's erosion dilation [35] and largest component selection to eliminate small artifacts from the predicted output. We subtract the predicted output from the defective input skull to obtain the volumetric implant data.

We use NVIDIA A100 GPUs during training. The preprocessing and data loader transformations use Pytorch [41] and the MONAI framework [40]. The data preprocessing is performed once and cached using the MONAI persistent data loader to store it in the system storage. The training consists of 300 epochs with an initial learning rate of 10^{-4} utilizing Adam optimizer. We choose the best-validated epoch as the final model. We use this configuration in all experiments.

VI. RESULTS

A. EXPERIMENT 1: MODEL ARCHITECTURE PERFORMANCE

In Table 3, we observe similar dice scores even after increasing the model size after incorporating the attention mechanism into the 3DUNetCNN decoder. Meanwhile, after replacing the backbone of 3DUNetCNN and Wodzinski UNet with the ConvNeXt backbone (CraNeXt) while keeping the stage ratio, the results from 3DUNetCNN with ConvNeXt 3D show a similar performance with a slight decrease in DSC when compared to the baseline. CraNeXt with a proposed stage ratio outperforms baseline 3DUNetCNN by improving DSC from 0.7389±0.17 to 0.7753±0.14 while reducing the model size from 262M to 43M (5.8x smaller). After replacing the CraNeXt backbone, the model size reduces from 262M to 69M in 3DUNetCNN and 69M to 34M in Wodzinski UNet, respectively. Wodzinski UNet and 3DUNetCNN with ConvNeXt 3D outperform the original backbones. CraNeXt outperforms most of the metrics, including DSC, dH95, and SdH, while Wodzinski UNet with ConvNeXt slightly outperforms in bDSC and FNR.

In statistical analysis, the CraNeXt model significantly outperformed the 3DUNetCNN baseline in all metrics, including DSC, bDSC, dH95, and SdH (p-value < 0.05, 3rd row in Table 6). This implies that the choice of backbone architecture has a significant impact on the model's overall performance. Replacing the 3D ConvNeXt block with the 3D residual block in 3DUNetCNN improves skull reconstruction performance over the baseline. Meanwhile, we found no significant differences between Wodzinski UNet and ConvNeXt 3D addition to Wodzinski UNet ($p_{DSC} = 0.2341$, 4th row in Table 6).

B. EXPERIMENT 2: PERFORMANCE COMPARISON OF CRANEXT ON SKULLBREAK AND IN-HOUSE DATASETS

In the first assessment (Table 4), we train CraNeXt exclusively on the SkullBreak dataset. While this approach yields

TABLE 3. Performance comparison of the proposed method with state-of-the-art methods trained on the SkullBreak dataset. The best value for each metric is highlighted in boldface. A gray background highlights the method that performs best overall.

Model	Size	DSC ↑	bDSC ↑	dH95 ↓	SdH ↓	FNR ↓	FPR ↓
PCA-skull [13]	-	0.0857±0.07	0.0460±0.04	121.27±13.8	143.40±13.5	0.9550±0.05	0.0048±0.00
Stage ratio: 2:2:2:2:2:1:1:1:1:1							
3DUNetCNN (baseline) [21]	262M	0.7389±0.17	0.7837±0.15	4.5800±6.35	11.406±10.6	0.2106±0.17	0.0021±0.00
3DUNetCNN + Attention [23]	263M	0.7287±0.17	0.7799±0.15	4.5233±6.09	10.582±9.89	0.2069±0.16	0.0022±0.00
3DUNetCNN + ConvNeXt 3D	61M	0.6789±0.14	0.5964±0.15	62.189±34.3	92.752±36.3	0.4065±0.16	**0.0012±0.00**
Stage ratio: 1:1:1:3:1:1:1:1:1:1							
CraNeXt (ours)	45M	**0.7753±0.14**	0.7943±0.14	**3.8151±6.05**	**9.8865±9.18**	0.1907±0.16	0.0017±0.00
Stage ratio: 0:1:2:3:3:2:1:1							
Wodzinski UNet [22]	69M	0.7489±0.19	0.7837±0.17	6.8761±16.1	12.661±19.9	0.1951±0.17	0.0020±0.00
Wodzinski UNet + ConvNeXt 3D	**43M**	0.7551±0.17	**0.7960±0.14**	5.4651±8.21	11.614±11.8	**0.1868±0.17**	0.0019±0.00

TABLE 4. Comparison of CraNeXt's performance on SkullBreak and in-house datasets. The best value for each model and dataset is shown in boldface, and the second-best value is shown in italic. The best-performing model on the in-house dataset is highlighted with a gray background.

Dataset	Train/Validate/Test	SkullBreak			In-house		
		DSC ↑	dH95 ↓	SdH ↓	DSC ↑	dH95 ↓	SdH ↓
SkullBreak only	470/50/50,34	**0.8552±0.06**	**2.1196±0.95**	**6.8408±4.04**	0.6577±0.15	6.3085±8.94	14.366±12.4
In-house only	260/34/50,34	0.4808±0.29	31.737±37.4	40.824±40.6	0.6664±0.18	8.2509±14.2	16.487±17.2
Skullbreak fine tune with in-house	260/34/50,34	0.7629±0.09	4.3585±3.73	10.983±7.39	*0.7136±0.17*	*5.2716±9.22*	*12.091±11.6*
Combined SkullBreak & in-house	730/84/50,34	*0.8435±0.07*	*2.1239±0.92*	*7.0336±4.16*	**0.7217±0.16**	**5.1262±7.92**	**12.084±10.4**

TABLE 5. Comparison of CraNeXt's performance on the original binary dataset and the categorized dataset. The best value for each metric is shown in boldface.

Dataset	DSC ↑	bDSC ↑	dH95 ↓	SdH ↓	FNR ↓	FPR ↓
Original skull	0.7942±0.13	0.8059±0.13	3.3391±5.33	9.0779±7.80	0.2071±0.16	0.0012±0.00
Categorized skull	**0.7969±0.13**	**0.8099±0.13**	**3.3300±5.37**	**8.4257±7.84**	**0.2064±0.16**	**0.0012±0.00**

the best results on the SkullBreak test dataset, it performs poorly on the in-house test dataset. This indicates a limited generalization to the specific characteristics of our in-house data.

In the second assessment, we train CraNeXt solely on the in-house dataset. Surprisingly, this results in suboptimal performance on both the SkullBreak and in-house test datasets, suggesting that the in-house dataset alone might not provide sufficient diversity and coverage for effective model training.

For the third assessment, we employ a transfer learning approach by pre-training CraNeXt on the SkullBreak and fine-tuning it with the in-house dataset. Despite the potential benefits of transfer learning, the results for both test datasets remain unsatisfactory, indicating that the pre-trained features from the SkullBreak dataset might not align well with the specific requirements of our in-house data.

Finally, in the fourth assessment, we combine both the SkullBreak and in-house datasets for training. This approach yields the best performance on the in-house test dataset, with a dice score of 0.7217±0.16. Furthermore, CraNeXt's performance on the SkullBreak test dataset is the second-best, with a lower dice score from 0.8552±0.06 to 0.8435±0.07, suggesting a slight trade-off in generalizability when incorporating in-house data.

For the clinical in-house dataset, we found that using the combined dataset differs significantly from using only SkullBreak or the in-house dataset alone, with the exception of the p-value for dH95 ($p_{dH95} = 0.0539$, 12th row in Table 6). However, we discovered that fine-tuning and combined data approaches make no significant difference for the in-house dataset. Our results suggest that selecting training strategies, consolidating datasets, and fine-tuning with clinical dataset can affect CraNeXt's performance and generalization, especially when dealing with diverse characteristics of in-house data. We find that training the model by combining both synthetic and clinical datasets offers the best-performing metrics on an in-house dataset.

C. EXPERIMENT 3: PERFORMANCE COMPARISON OF CRANEXT ON ORIGINAL AND CATEGORIZED DATASETS

The results shown in Table 5 demonstrate that the model trained on the categorized dataset performs slightly better in all metrics than the model trained on the original binary dataset. The CraNeXt model, trained with a categorized dataset, achieved the highest dice score (0.7969±0.13), making it the best-performing model overall. Our results suggest that adding semantic information during training can slightly improve reconstruction performance. However, there are no significant differences in statistical analysis between original and categorized skulls in terms of DSC, bDSC, and dH95 ($p_{DSC} = 0.2109$, $p_{bDSC} = 0.0863$, and $p_{dH95} = 0.3987$, 14th row in Table 6). Despite this limitation, the model trained with the categorized dataset outperforms the

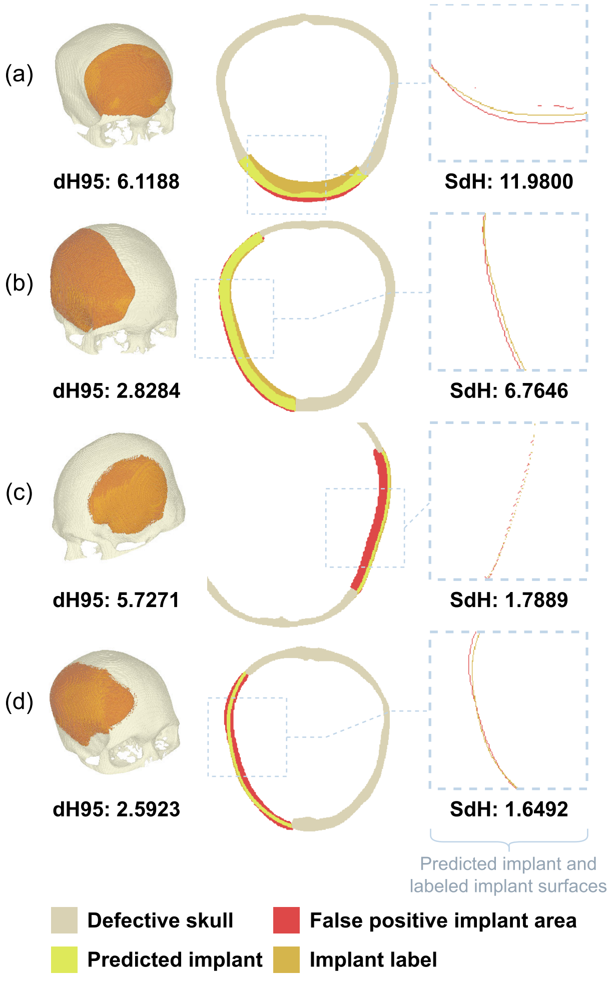

FIGURE 6. Comparison of predicted and labeled implant surfaces using the 95th percentile Hausdorff distance (dH95) and our proposed Surface Hausdorff distance (SdH). SdH outperforms in detecting surface errors, especially in case (b), where the predicted implant surface is smoother than (c). Despite the smoother surfaces, the dH95 values are higher than the SdH values, indicating that the surface Hausdorff distance is a better representative of contour errors.

uncategorized model, demonstrating the potential for skull categorization to improve model effectiveness in clinical applications.

D. EFFECTIVE MEASURING OF CONTOUR ERRORS USING SURFACE HAUSDORFF

We evaluate if the proposed SdH can be used as an additional metric for autoimplant. Fig. 6 presents a comparison between the predicted implant surfaces and the corresponding implant label surfaces for four different cases (a–d). The Surface Hausdorff distance (SdH) is introduced as a more effective measure of contour errors compared to the 95th percentile Hausdorff distance (dH95). This is highlighted in cases (Fig. 6b,c), where the predicted implant surfaces appear smoother. Despite the smoother surfaces, the dH95 values in these cases (2.8284 and 5.7271, respectively) are higher than what would be expected given the visual similarity between the predicted and actual surfaces. In contrast, the corresponding SdH values (6.7646 and 1.7889) provide a more accurate representation of the contour errors. This suggests that dH95 fails to capture the true nature of contour errors for smoother surfaces, while SdH offers a more reliable assessment. In conclusion, SdH can be used as an additional metric for autoimplant tasks in which the clinical application only uses the predicted implant surface for the patient-specific titanium implant designs.

E. QUALITATIVE EVALUATION

We investigate 50 and 34 predicted implants from SkullBreak and the in-house test dataset with CraNeXt by examining both 2D planes (coronal, sagittal, and axial planes) and 3D, then summarize them into 5 error categories based on the generated implant (Section IV-B). The most common error is the lack of completeness in the 42 predicting implants, meaning that half of them were unable to fully encompass the defective area, where they are mostly formed as tiny holes in the skull (Fig. 7a). The second most common error was related to the smoothness of the transition area between the skull and the implant (Fig. 7d), where error is found in 36 predicted implants. It may not give a smooth transition due to different defect edges. Another issue is found in 21 implants that create false positive areas, indicating that the model may over-predict some implants on the given skulls. However, we found that 63 predicted implants have the ability to cover the defective space with no additional prediction on skulls, while the rest are commonly found in the defects at the eye socket region (Fig. 7b). Additionally, we observed that 14 skulls contain improper curves, which generally occur when the segmented skull has curvature around the edges, leading to the non-restored skull shape causing the incorrect curvature from the model prediction (Fig. 7c). Lastly, 13 out of 84 dataset skulls are described as not passing the criteria of minimal thickness since they have either a too thick or too thin reconstructed implant (Fig. 7e).

In addition, we explore the error of an implant that may be influenced by the clinical variability of the skull itself into 4 main categories, including non-break-through skulls, bone artifacts, complexed defective areas, and defect edges (Table 7). For non-break-through skulls, they are characterized by a defect edge that is either completely or nearly fully connected to the skull, making it challenging for algorithms to estimate the boundary of the defective areas (Fig. 8a). In some cases, bone artifacts are present in defective skulls, causing the models to predict incorrectly. The predicted areas may still contain bone growth occurring due to calcification, alloplastic, and autograft, which is considered the most difficult artifact closing off a defective area with a patient's bone observing in those patients with prior implants during CT scan (Fig. 8b). The complexity of defective areas is also found in specific defective regions such as the frontal, temporal, and occipital areas, depending on the variety of patients' skull morphology (Fig. 8c). These complex defects may not be well represented in the training set and require high prediction accuracy to create the

TABLE 6. One-sided paired *t*-test between the approaches across 3 experiments, compared regarding DSC, bDSC, dH95, and SdH, with the assumption that approaches B on the right outperform those A on the left. The *p*-values smaller than 0.05 are highlighted in boldface.

Exp.	Approach A	Approach B	*p*-value			
			DSC	bDSC	dH95	SdH
1	3DUNetCNN (baseline)	3DUNetCNN + Attention	**0.0376**	0.7388	0.4043	0.0772
	3DUNetCNN (baseline)	3DUNetCNN + ConvNeXt 3D	**<0.001**	**<0.001**	**<0.001**	**<0.001**
	3DUNetCNN (baseline)	CraNeXt (ours)	**<0.001**	**0.0185**	**<0.001**	**0.0111**
	Wodzinski UNet	Wodzinski UNet + ConvNeXt 3D	0.2341	**0.0448**	0.1581	0.2633
	Wodzinski UNet	CraNeXt (ours)	**0.0117**	0.0972	**0.0299**	0.0700
2	**Data:** SkullBreak					
	SkullBreak only	Combined SkullBreak & in-house	**<0.001**	**<0.001**	0.4753	0.2952
	In-house only	Combined SkullBreak & in-house	**<0.001**	**<0.001**	**<0.001**	**<0.001**
	SkullBreak fine tune with in-house	Combined SkullBreak & in-house	**<0.001**	**<0.001**	**<0.001**	**<0.001**
	Data: In-house					
	SkullBreak only	Combined SkullBreak & in-house	**<0.001**	**<0.001**	**<0.001**	**0.0109**
	In-house only	Combined SkullBreak & in-house	**0.0075**	**0.0048**	0.0539	**0.0259**
	SkullBreak fine tune with in-house	Combined SkullBreak & in-house	0.2132	0.1609	0.2668	0.4898
3	Original skull	Categorized skull	0.2109	0.0863	0.3987	**<0.001**

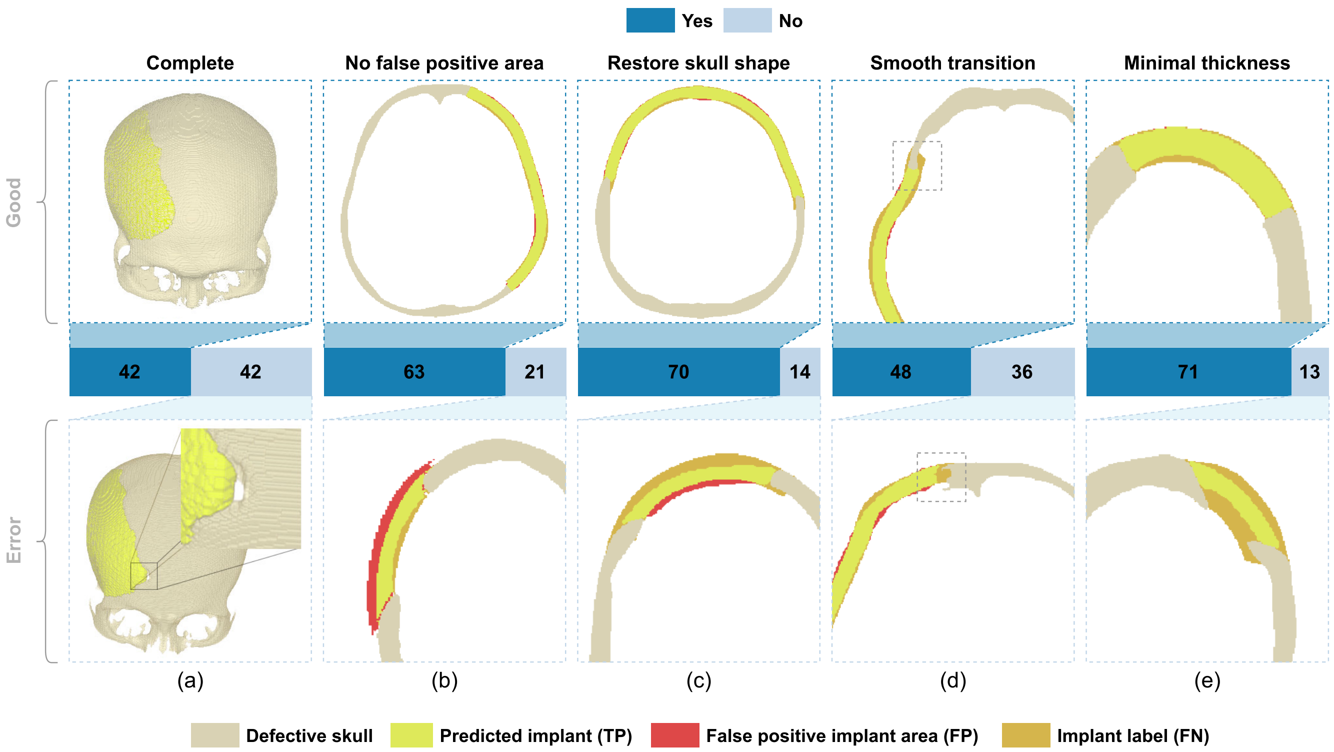

FIGURE 7. Qualitative evaluation of skull reconstruction results on 84 test cases, with "yes" indicating compliance with the criteria and "no" indicating non-compliance with the criteria. The top row shows representative cases that meet the criteria. The bottom row shows examples of predicted error cases. (a) The incomplete implant observed in 3D shows the hole displayed between the predicted implant (yellow area) and defective skull (bone area). (b) The false positive area (red area) is observed in the coronal plane where the predicted implant exceeds the defective area. (c) The predicted implant with a non-restored skull shape can be observed due to its small curve compared to the labeled implant (orange area). (d) The predicted skull with a non-smooth transition is shown between the boundaries of the implant and the defective skull. (e) The predicted skull is thin when compared to the defective thickness.

proper curves. Finally, defect edges, which are rounded cross sections transitioned between the skull and the defect area, create a large defect and may pose difficulties in generating the implant that entirely covers the surface area (Fig. 8d). This results in the implant protruding through the inner side of the outer surface that contracts to the brain for titanium mesh regeneration.

VII. DISCUSSION

In this paper, we have proposed CraNeXt, a novel convnext-based 3D skull reconstruction model. We use categorized skulls to both improve the reconstruction and help analyze the statistics of the missing skull. We incorporate a large number of 328 in-house clinical dataset with a public SkullBreak dataset for autoimplant. We found that our

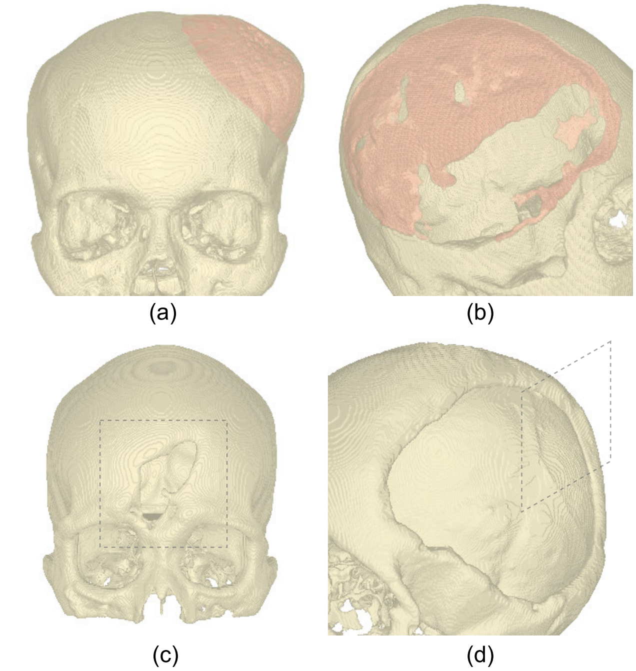

FIGURE 8. Examples of defective skull errors. (a) Not-break-through skulls arise from bone surface fractures (red area) that do not create open wounds. This is an example case from a patient with large brain tumors that visibly pushed the skull outward. (b) Bone artifacts (red area) are the result of calcification. (c) The complexity of the defective area is depicted in the frontal area. (d) A defect edge arises from the slope edges.

proposed model gained accuracy on the skull reconstruction while using 5.8x fewer parameters. We explore the errors made by the model on clinical and SkullBreak datasets and discuss challenges in clinical translation.

The size of the 3D backbones for skull reconstruction is significantly larger compared to their 2D counterparts, presenting challenges in terms of computational resources and efficiency. By optimizing the model's backbone architecture with ConvNeXt block designs [10], we have reduced the model size while simultaneously improving the generalizability of the skull reconstruction. The reduced model size also improves computational efficiency, leading to faster execution times for clinical applications. Future research may explore the development of architecture backbones and optimization methods to create more compact and expressive models capable of representing complex skull data [22].

To our knowledge, our collected data is one of the largest clinical defective skull datasets [1]. We found that combining synthetic datasets with clinical datasets can improve the generalizability of the model. Synthetic data often have more controlled and standardized defect patterns while lacking the variations that are found in clinical data. Meanwhile, clinical data have different characteristics, allowing models to learn complex defective areas. However, they may contain complex skull shapes and complexity, which makes the model learning step more challenging. One approach to improving the model is to increase the availability of real clinical data to cover the variability of defects. Another approach is to explore clinically relevant synthetic generation of defective skulls [16]. Nevertheless, there is another challenge relating to the labeling process when using clinical datasets. As the data were collected from subjects with pre-existing complications, they usually lack actual healthy skull information. During data preparation, design engineers should standardize the labeling process across all data and ensure that the designed implant label closely matches the missing part of the skull. We think that adding more clinical data and translating these clinical defects for synthetic data generation with a greater diversity of cranial anomalies could help improve the generalizability of the model.

TABLE 7. Type of error based on characteristics of skulls.

Type of labeled skull error	Description
Non-break-through	The defective areas can cover the skull with no holes that allow a vision passing through skulls.
Bone artifact	The defective areas contain obstruction of reconstruction skulls due to calcification, autograft, and alloplastic.
Complexed defective area	The defect areas are sophisticated for skull reconstruction.
Defect edge	Errors caused by the characteristics or shape of defective areas.

In our research, we propose SdH to measure the surface distance between the predicted and actual implant since most patient-specific titanium implants only consider implant contour for 3D printing. Although SdH has a close correlation with dH, we believe that using SdH to consider surface distance can be used for additional interpretation. While the metrics used in previous research [1] are useful for evaluating the overall performance of the skull reconstruction, they may not fully capture the clinical relevance of the reconstructed implant. In practice, the contour of the reconstructed skull is critical to achieving a satisfactory aesthetic outcome while also ensuring a proper fit with the surrounding anatomical structures. As a result, we believe that incorporating a more meaningful metric designed specifically to assess the accuracy of the skull contour will provide a more clinically relevant assessment of reconstruction quality.

Qualitative evaluation of the generated implants using CraNeXt reveals several areas for improvement, including practical-generated implants and variability in clinical patterns. We believe that pre- and post-processing of clinical skulls, including clinical-related loss functions, may improve generation tasks. For example, smoothness issues at the transition area between the skull and the implant can be mitigated by exploring post-processing techniques or incorporating smoothness constraints into the loss function during training. To improve the completeness of the implants and complex defective regions, future work could investigate the use of shape priors or topology-aware loss functions to ensure that the predicted implants fully encompass the defective areas [42], [43].

Our study demonstrates the potential for enhancing the clinical applicability and usability of modern deep learning architectures in skull reconstruction. It addresses

Algorithm 1 Implant Surface Extraction Method

```
Input  : 3D voxel of skull and implant $I$ with
         $W \times H \times D$ shape, where value of
         background = 0, original skull = 1, and
         implant = 2.
Output: Binary 3D voxel of implant surface $I_{Surface}$
         with the same input shape, where value of
         background = 0 and implant surface = 1.

$W, H, D \leftarrow$ shape ($I$)
$I_{Surface}(w, h, d) \leftarrow 0, \forall w \in W, \forall h \in H, \forall d \in D$
// projection paths
$P \leftarrow array[]$
// Left←→Right
for $h \in H, d \in D$ do
    Append $P$ with $path(w, h, d)forw = 0, \dots, W$
    Append $P$ with $path(w, h, d)forw = W, \dots, 0$
end
// Anterior←→Posterior
for $w \in W, h \in H$ do
    Append $P$ with $path(w, h, d)ford = 0, \dots, D$
    Append $P$ with $path(w, h, d)ford = D, \dots, 0$
end
// Superior→Inferior
for $w \in W, d \in D$ do
    Append $P$ with $path(w, h, d)forh = H, \dots, 0$
    continue
end
for $path \in P$ do
    for $w, h, d$ in $path$ do
        // hit skull
        if $I(w, h, d) = 1$ then break
        // hit implant
        if $I(w, h, d) = 2$ then
            $I_{Surface}(w, h, d) \leftarrow 1$
            break
        end
    end
end
return $I_{Surface}$
```

the challenges in craniofacial reconstruction and paves the way for the increased translation of skull reconstruction techniques into broader clinical practice. Future research can focus on further refining models and validating their performance in larger and more diverse patient populations.

VIII. CONCLUSION

We introduced CraNeXt, a novel model designed for skull reconstruction that combines the ConvNeXt backbone with a 3D U-Net AutoEncoder architecture. Our model outperforms the previous model architecture in autoimplant while using 5.8x fewer model parameters, with an improved dice score from 0.7389 to 0.7753. Our model's efficiency was improved by combining synthetic defects and real clinical datasets, elevating the dice score to 0.7942. We also developed a novel method of skull categorization that helps label different parts of the skull, enhances explainability, and improves model performance. We demonstrated that using categorized data leads to an improvement in model performance, with a final dice score of 0.7969. This paper translates autoimplant into practical clinical application.

APPENDIX

The surface Hausdorff distance (SdH) is a metric used to quantify the distance between two surfaces. The implant surface extraction method is outlined in Algorithm 1.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

DATA AVAILABILITY

A portion of the in-house dataset will be made available upon reasonable request. The source code will be publicly available on https://github.com/guitared/CraNeXt.

ACKNOWLEDGMENT

The authors acknowledge Chulalongkorn University Technological Center (UTC) for providing valuable computing resources for this work. They are also grateful to Kantapat Phakdeewisetkul and Tanya Charoensrisomboon from Meticuly Company Ltd. for their contribution to the in-house clinical dataset preparation. All data is collected under consent forms and in compliance with Thailand's Personal Data Protection Act (PDPA). Finally, they thank the public SkullBreak dataset with a CC BY-NC-SA 4.0 license.

REFERENCES

[1] J. Li et al., "Towards clinical applicability and computational efficiency in automatic cranial implant design: An overview of the AutoImplant 2021 cranial implant design challenge," *Med. Image Anal.*, vol. 88, Aug. 2023, Art. no. 102865, doi: 10.1016/j.media.2023.102865.

[2] S. Aydin, B. Kucukyuruk, B. Abuzayed, S. Aydin, and G. Z. Sanus, "Cranioplasty: Review of materials and techniques," *J. Neurosci. Rural Pract.*, vol. 2, no. 2, pp. 162–167, Jul. 2011, doi: 10.4103/0976-3147.83584.

[3] X. Chen, L. Xu, X. Li, and J. Egger, "Computer-aided implant design for the restoration of cranial defects," *Sci. Rep.*, vol. 7, no. 1, p. 4199, Jun. 2017, doi: 10.1038/s41598-017-04454-6.

[4] L. Bobinski, L.-O.-D. Koskinen, and P. Lindvall, "Complications following cranioplasty using autologous bone or polymethylmethacrylate—Retrospective experience from a single center," *Clin. Neurol. Neurosurgery*, vol. 115, no. 9, pp. 1788–1791, Sep. 2013, doi: 10.1016/j.clineuro.2013.04.013.

[5] J. Li, O. Kodym, D. G. Ellis, M. Spanl, M. R. Aizenberg, V. Alves, G. Von Campe, and J. Egger, Mar. 2, 2021, "Towards the automatization of cranial implant design in cranioplasty: 2nd MICCAI challenge on automatic cranial implant design," *Zenodo*, doi: 10.5281/zenodo.4577269.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*. Cham, Switzerland: Springer, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[8] J. Egger, J. Li, X. Chen, U. Schäfer, G. Campe, M. Krall, U. Zefferer, C. Gsaxner, A. Pepe, and D. Schmalstieg, Mar. 19, 2020, "Towards the automatization of cranial implant design in cranioplasty," *Zenodo*, doi: 10.5281/zenodo.3715952.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Represent. (ICLR)*, Austria, May 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[10] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11966–11976, doi: 10.1109/CVPR52688.2022.01167.

[11] S. Roy, G. Koehler, C. Ulrich, M. Baumgartner, J. Petersen, F. Isensee, P. F. Jäger, and K. H. Maier-Hein, *MedNeXt: Transformer-Driven Scaling of ConvNets for Medical Image Segmentation*. Cham, Switzerland: Springer, 2023, pp. 405–415, doi: 10.1007/978-3-031-43901-8_39.

[12] O. Kodym, M. Španěl, and A. Herout, "Deep learning for cranioplasty in clinical practice: Going from synthetic to real patient data," *Comput. Biol. Med.*, vol. 137, Oct. 2021, Art. no. 104766, doi: 10.1016/j.compbiomed.2021.104766.

[13] L. Yu, J. Li, and J. Egger, *PCA-Skull: 3D Skull Shape Modelling Using Principal Component Analysis* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, 2021, pp. 105–115, doi: 10.1007/978-3-030-92652-6_9.

[14] J. Li, M. Krall, F. Trummer, A. R. Memon, A. Pepe, C. Gsaxner, Y. Jin, X. Chen, H. Deutschmann, U. Zefferer, U. Schäfer, G. V. Campe, and J. Egger, "MUG500+: Database of 500 high-resolution healthy human skulls and 29 craniotomy skulls and implants," *Data Brief*, vol. 39, Dec. 2021, Art. no. 107524, doi: 10.1016/j.dib.2021.107524.

[15] O. Kodym, M. Spanel, and A. Herout, *Segmentation of Defective Skulls From CT Data for Tissue Modelling* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, 2021, pp. 19–28, doi: 10.1007/978-3-030-92652-6_3.

[16] O. Kodym, J. Li, A. Pepe, C. Gsaxner, S. Chilamkurthy, J. Egger, and M. Španěl, "SkullBreak/SkullFix—Dataset for automatic cranial implant design and a benchmark for volumetric shape learning tasks," *Data Brief*, vol. 35, Apr. 2021, Art. no. 106902, doi: 10.1016/j.dib.2021.106902.

[17] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3D ShapeNets: A deep representation for volumetric shapes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1912–1920, doi: 10.1109/CVPR.2015.7298801.

[18] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 77–85, doi: 10.1109/CVPR.2017.16.

[19] P. Friedrich, J. Wolleb, F. Bieder, F. M. Thieringer, and P. C. Cattin, "Point cloud diffusion models for automatic implant generation," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*, vol. 14228. Cham, Switzerland: Springer, p. 112, doi: 10.1007/978-3-031-43996-4_11.

[20] D. G. Ellis and M. R. Aizenberg, *Trialing U-Net Training Modifications for Segmenting Gliomas Using Open Source Deep Learning Framework* (Lecture Notes in Computer Science), D. G. Ellis and M. R. Aizenberg, Eds. Cham, Switzerland: Springer, 2021, pp. 40–49, doi: 10.1007/978-3-030-72087-2_4.

[21] D. G. Ellis and M. 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* (Lecture Notes in Computer Science), D. G. Ellis and M. R. Aizenberg, Eds. Cham, Switzerland: Springer, 2020, pp. 47–55, doi: 10.1007/978-3-030-64327-0_6.

[22] M. Wodzinski, M. Daniol, M. Socha, D. Hemmerling, M. Stanuch, and A. Skalski, "Deep learning-based framework for automatic cranial defect reconstruction and implant modeling," *Comput. Methods Programs Biomed.*, vol. 226, Nov. 2022, Art. no. 107173, doi: 10.1016/j.cmpb.2022.107173.

[23] A. Myronenko, *3D Brain Tumor Segmentation and Survival Prediction Using 3D Attention UNet* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, 2020, pp. 262–272, doi: 10.1007/978-3-030-46640-4_25.

[24] H. Mahdi, A. Clement, E. Kim, Z. Fishman, C. M. Whyne, J. G. Mainprize, and M. R. Hardisty, *A U-Net Based System for Cranial Implant Design With Pre-Processing and Learned Implant Filtering* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, 2021, pp. 63–79, doi: 10.1007/978-3-030-92652-6_6.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[26] A. Myronenko, *3D MRI Brain Tumor Segmentation Using Autoencoder Regularization* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, 2019, pp. 311–320, doi: 10.1007/978-3-030-11726-9_28.

[27] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, *3D U-Net: Learning Dense Volumetric Segmentation From Sparse Annotation* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, 2016, pp. 424–432, doi: 10.1007/978-3-319-46723-8_49.

[28] N. J. Tustison, P. A. Cook, A. J. Holbrook, H. J. Johnson, J. Muschelli, G. A. Devenyi, J. T. Duda, S. R. Das, N. C. Cullen, D. L. Gillen, M. A. Yassa, J. R. Stone, J. C. Gee, and B. B. Avants, "The ANTsX ecosystem for quantitative biological and medical imaging," *Sci. Rep.*, vol. 11, no. 1, p. 9068, Apr. 2021, doi: 10.1038/s41598-021-87564-6.

[29] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002, doi: 10.1109/ICCV48922.2021.00986.

[30] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "ConvNeXt v2: Co-designing and scaling ConvNets with masked autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 16133–16142, doi: 10.1109/CVPR52729.2023.01548.

[31] S. Chilamkurthy, R. Ghosh, S. Tanamala, M. Biviji, N. G. Campeau, V. K. Venugopal, V. Mahajan, P. Rao, and P. Warier, "Deep learning algorithms for detection of critical findings in head CT scans: A retrospective study," *Lancet*, vol. 392, no. 10162, pp. 2388–2396, Dec. 2018, doi: 10.1016/s0140-6736(18)31645-3.

[32] J. Li et al., "MedShapeNet—A large-scale dataset of 3D medical shapes for computer vision," 2023, *arXiv:2308.16139*.

[33] *Meticuly—Meticulously Crafted Bones for Better Lives*. Accessed: Jun. 16, 2024. [Online]. Available: https://www.meticuly.co.th

[34] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka, J. Buatti, S. Aylward, J. V. Miller, S. Pieper, and R. Kikinis, "3D slicer as an image computing platform for the quantitative imaging network," *Magn. Reson. Imag.*, vol. 30, no. 9, pp. 1323–1341, Nov. 2012, doi: 10.1016/j.mri.2012.05.001.

[35] S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu, "Scikit-image: Image processing in Python," *PeerJ*, vol. 2, p. e453, Jun. 2014, doi: 10.7717/peerj.453.

[36] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the Hausdorff distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 9, pp. 850–863, 1993, doi: 10.1109/34.232073.

[37] L. Axel, "Orientation of magnetic resonance images," *Radiology*, vol. 151, no. 2, p. 534, May 1984, doi: 10.1148/radiology.151.2.6709933.

[38] D. G. Ellis, C. M. Alvarez, and M. R. Aizenberg, *Qualitative Criteria for Feasible Cranial Implant Designs* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, 2021, pp. 8–18, doi: 10.1007/978-3-030-92652-6_2.

[39] F. Pérez-García, R. Sparks, and S. Ourselin, "TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning," *Comput. Methods Programs Biomed.*, vol. 208, Sep. 2021, Art. no. 106236, doi: 10.1016/j.cmpb.2021.106236.

[40] MONAI Consortium. (2023). *Monai: Medical Open Network for AI*. [Online]. Available: https://zenodo.org/record/4323058

[41] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Red Hook, NY, USA: Curran Associates, 2019.

[42] S. Paul, B. Jhamb, D. Mishra, and M. S. Kumar, "Edge loss functions for deep-learning depth-map," *Mach. Learn. Appl.*, vol. 7, Mar. 2022, Art. no. 100218, doi: 10.1016/j.mlwa.2021.100218.

[43] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE Trans. Med. Imag.*, vol. 32, no. 7, pp. 1153–1190, Jul. 2013, doi: 10.1109/TMI.2013.2265603.

THATHAPATT KESORNSRI received the bachelor's degree (Hons.) in computer engineering from Chiang Mai University, Thailand, in 2016. He is currently pursuing the master's degree in computer engineering with Chulalongkorn University. He joined ExxonMobil, Ltd., in 2016. He is also the DevOps and Data Integration Platform Engineer, with the responsibility of maintaining on-premises and cloud containerized infrastructures for enterprise data analytics and deep learning. His current research interests include deep learning, medical imaging, and data analytic fields.

NAPASARA ASAWALERTSAK is currently pursuing the B.Eng. degree in biomedical engineering with Mahidol University, Thailand. She has actively engaged in projects related to medical image analysis and machine learning algorithms, particularly focusing on their application in clinical settings. Her research interests include medical imaging and machine learning applications for clinical diagnostics and healthcare.

NATDANAI TANTISEREEPATANA is currently pursuing the bachelor's degree in biomedical engineering with Rangsit University. In 2023, he joined Meticuly Company Ltd., where he currently holds the position of Full Stack Machine Learning Engineer. His professional experience encompasses CT skull image registration, an area in which he has acquired expertise. His research interests include leveraging artificial intelligence techniques for medical imaging applications and exploring novel ways to enhance diagnostic capabilities, and patient outcomes through innovative technological solutions.

PORNNAPAS MANOWONGPICHATE is currently pursuing the bachelor's degree in biomedical engineering with Mahidol University. She is deeply interested in applying machine learning to real medical challenges, especially in medical imaging for diagnostics and text mining. Her goal is to bring about innovations in healthcare technology that would improve patient care and optimize workflows.

BOONRAT LOHWONGWATANA received the Ph.D. degree in materials science from California Institute of Technology. He is currently a Faculty Member with the Department of Metallurgical Engineering and the Director of the Biomedical Engineering Research Center, Chulalongkorn University, Thailand. He is also the Founder of the Biomechanics Research Center, Meticuly Company Ltd., Thailand. His expertise includes precision manufacturing, 3D printing of titanium, biomechanical designs, and standardized testing conforming to ISO 13485 and ASTM standards. He was a recipient of Thailand's Young Technologist Award from HRH Princess Sirindhorn, in 2013, and Thailand's Young Metallurgist Award, in 2015.

CHEDTHA PUNCREOBUTR received the Ph.D. degree in materials engineering from Imperial College London. He is currently a Faculty Member with the Department of Metallurgical Engineering and the Department of Materials Engineering, Chulalongkorn University. He also serves as the Co-Founder of the Biomechanics Research Center, Meticuly Company Ltd. He has pioneered the metal additive manufacturing and materials modeling for orthopedics and biomedical applications. He has received numerous awards, including Thailand Young Outstanding Metallurgist Award and Excellence Research Awards.

TITIPAT ACHAKULVISUT received the B.Eng. degree in electrical engineering from Chulalongkorn University, the M.S. degree in biomedical engineering from Northwestern University, and the Ph.D. degree in bioengineering from the University of Pennsylvania, specializing in the application of natural language processing (NLP) and machine learning (ML) to enhance scientific processes. He leads the Biomedical and Data Laboratory, Mahidol University, aiming to apply machine learning and natural language processing to understand biomedical data and images.

PEERAPON VATEEKUL (Senior Member, IEEE) received the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Miami (UM), Coral Gables, FL, USA, in 2012. He is currently an Associate Professor with the Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Thailand. His research interests include machine learning, data mining, deep learning, text mining, and big data analytics. To be more specific, his works include variants of classification (hierarchical multi-label classification), natural language processing, data quality management, and applied deep learning techniques in various domains, such as healthcare, bioinformatics, and hydrometeorology. Some examples of AI-assisted medical diagnoses are real-time polyp detection from colonoscopy, gastrointestinal metaplasia segmentation from gastroscopy, dyssynergic defecation classification, depressive scoring from interview videos, Parkinson's face classification, and movement disorder diagnosis.