This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

From CT to Clinical Cranioplasty Inpainting using a Conditional Wavelet Diffusion Model

Ricardo Coimbra Brioso[1], Florentin Bieder[2], and Philippe Cattin[2]

[1] Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy
[2] Department of Biomedical Engineering, University of Basel, Basel, Switzerland

Abstract. Cranial implant design for cranioplasty remains a time-intensive task requiring specialized expertise, often leading to delays in treating traumatic brain injuries (TBI). While automating the implant design shows promise, there is a scarcity of diverse, annotated surgical datasets and models that generalize across different clinical domains (different acquitision machines and image processing techniques). We propose a pipeline to bridge this gap by enabling the generation of a low computational cost and large-scale, surgical training data from standard, non-defective CT scans. We introduce CranioImpInp, an end-to-end framework for AI-assisted implant inpainting in surgical planning. The pipeline starts with raw CT scans, which are segmented using the open-weights TotalSegmentator model. To address data scarcity, we developed a defect generation algorithm that injects realistic, diverse defects into binarized skulls, serving as a reproducible research resource for dataset-centric community building. Augmented datasets were used to train a 3D Conditional Wavelet Diffusion Model (cWDM), which leverages wavelet transforms for computationally efficient 3D inpainting without the need for intensive GPU-based latent space training. Our model was trained using the augmented CQ500 dataset and evaluated against the SkullBreak benchmark. The model SBCQ10 achieved a Dice Similarity Coefficient (DSC) of 90.3% and a 95th percentile Hausdorff Distance (HD95) of 2.15mm on the SkullBreak test set. Crucially, the data augmentation pipeline enabled superior generalization to out-of-distribution data, achieving 92.7% DSC and 6.31mm HD95 on the CQ500 test set, where the non-augmented baseline performed worse. Qualitative evaluation on three real clinical TBI cases demonstrated high-quality surgical planning proposals. This work provides a scalable framework for the usability of non-defective head CT data. By providing a method to transform public CT datasets into robust training resources, we move toward generalizable AI software that deceases the heavy workload in cranioplasty surgery planning.

Keywords: cranioplasty · surgical planning · dataset augmentation · image inpainting · conditional diffusion models · wavelet transform.

1 Introduction

Cranioplasty is a surgical procedure that repairs a defective segment of the patient's skull [1]. It is used to protect the brain's structure and function. After a decompressive craniotomy (DC), neuro-motor or cognitive impacts are consequences of the trephined syndrome. The probability of occurrence increases with the waiting time for cranioplasty. Cranioplasty prevents this by normalizing the cerebral blood flow and protecting the region of the brain where the skull segment was removed.

Materials used for cranioplasty can be categorized into biological and synthetic types. The biological materials are usually obtained from cadavers of animals (xenografts), cadavers of humans (allografts) or from your own body (autografts). Methyl methacylate and polyether ether ketone (PEEK) are popular materials for implants nowadays. PEEK is lightweight and does not interfere with image acquisition on the patient, for example, such as CT scans or MRI scans. These materials can be 3D-printed, allowing for increased speed in surgical planning when compared to other types of manufactoring [6]. Despite the advantages, designing implants using these materials still demands several hours from a trained doctor or biomedical engineer; therefore, using of artificial intelligence to generate implants could reduce the time devoted to implant design. The task of automatically creating the implants became popular with the AutoImplant 2021 challenge [10]. The challenge introduced binary skull volumes with defects and the goal was to generate the implant needed for each defect. Among the best entries in the challenge, the techniques used were based on the U-Net architecture, deep learning networks, or generative adversarial networks (GANs) for inpainting. In the area of image generation, diffusion models became increasingly popular due to their advantages over GANs, including higher image quality and diversity, and greater training stability,[7]. Despite their advantages, diffusion models are notoriously more computationally expensive to train and slower and GANs, especially when using 3D data. To solve this issue there are several approaches, such as dividing the input into smaller 3D patches, Variational Auto-Encoders (VAE) that transform the input space into its latent feature space [11], or Wavelet Diffusion Models (WDM) that apply Wavelet Transforms to the input [12] before feeding it to a diffusion model.

Although several models achieve promising performances on the automatic inpainting task, they cannot be applied to other datasets due to a lack of generalization arising from the scarse quantity of data. Our proposal is designed around this problem since it can create new defective skull samples from any CT dataset, moving the field towards a dataset and generalizable models for skull inpainting.

In this work, we present a complete pipeline that starts with a CT scan dataset containing the patient's head region. The cranium is segmented using an open-weight and open-source model called TotalSegmentator [14]. Using an algorithm that injects defects into the binarized cranium, we build several datasets with varying numbers of defects per patient that can be used to augment the training of a conditional WDM (cWDM) [4].

The contributions of the work are the following: competitive performance in the leaderboard of the AutoImplant2021 challenge and with recent published works on the topic; an end-to-end pipeline of CT scan to implant generation is proposed; a defect generation algorithm that can be used in any binarized skull dataset and allows the creation of multiple defects in the same skull; and, a qualitative and quantitative clinical evaluation of the implants generated using the proposed methodologies.

Acknowledgement: This work was financially supported by the Werner Siemens Foundation through the MIRACLE II project.

2 Related Work

The first edition of the AutoImplant Challenge in 2020 drew attention to the clinical application of computer vision in generating implants for defective skulls. Later, the follow-up challenge, AutoImplant 2021, introduced two datasets, SkullBreak and SkullFix [8, 10], which encouraged the development of deep learning solutions for the skull inpainting problem. The submission that came in first place used a 3D U-Net [15], trained with the SkullBreak and SkullFix datasets and achieved a performance of 87% DSC, in subsequent works [16, 17] the authors use the baseline of the first work and perform dataset augmentation using cross-case image registration and using the new training cases from the image registration to train a Variational Auto-Encoder (VAE) that would generate new 100.000 training samples. In their latest work [17], a detailed exploration of augmentation techniques is performed, adding GAN, Vector Quantisized-VAE, and a Latent Diffusion Model to the image registration and traditional VAE augmentation methods. The combination of all augmentation techniques culminated in a performance of 95% DSC and 1.13 mm HD95 in the SkullBreak dataset. Although the combination of these techniques performs well, the time required for training and sampling the new training samples might not be practical if, in the end, the model does not generalize to a different domain or clinical setting. The work of [5] explored a point-cloud diffusion model, achieving competitive of 87% DSC. Friedrich's algorithm avoided GPU memory limitations of 3d diffusion models by shifting the skull inpainting task from the voxel to wavelet domain. In [18], two 3D encoder-decoder networks are used to inpaint defective skulls. The first network is used to generate the implant in a lower resolution, and the second network increases the voxel resolution of the implant, making it more finer-grained. A clinical assessment compared geometric properties between of AI-generated and the manually generated implants, concluding that AI-generated implants exhibited greater skull symmetry than the manual implants and saved the clinician time in implant design. The work of [9] uses a private dataset comprising 167 healthy skulls and 9 different injection defects applied to each skull, resulting in 1503 defective skulls. The paired sets of healthy and defective skulls are used to train an encoder-decoder deep learning model that achieved a Dice Score of 77% in the implant region. The clinical applicability of these implants is studied and compared with that of implants created

using commercial software. In quantitative terms, the implants created by the model had higher DSC, precision, and recall than the manually-made implants. It is important to note that these metrics only measure the overlap and similarity between the implant and the original part of the healthy skull, and do not consider the clinical applicability of thinner implants or the implant's ability to fit within the skull.

3 Background and Methods

3.1 Datasets and Defect Generation Algorithm

The SkullBreak private dataset [8] consists of data derived from the CQ500 dataset [2]. The dataset includes 114 skulls, and, for each skull, five different synthetic defects: a unilateral defect in the parieto-temporal area, a unilateral defect in the fronto-orbital area, a bilateral defect, and two random defects. This resulted in 570 training cases. Each sample has dimensions of 512×512×512 and a voxel size of 0.4 mm. Pre-processing involved acquisition-geometry regularization and rigid alignment using landmarks that define the Frankfort horizontal plane. Skull segmentation from CT was performed using a convolutional neural network followed by a graph-cut refinement. The defect generation process involves subtracting a binary defect shape from the complete skull, followed by border smoothing using morphological operations. However, the precise algorithm is not publicly documented, and the corresponding code is not available.

The CQ500 dataset is publicly available and contains 491 head CT scans. Most of the scans are from patients with head trauma or stroke symptoms, such as different types of intracranial hemorrhages. Scans are in DICOM format and were collected from multiple CT scanner models with varying protocols, representing real-world diversity. Due to severe artifacts, registration failures, resolution discrepancies, or incomplete skull coverage, only 200 of the 500 scans were used for our experiments. We will refer to this dataset as the CQ dataset from now on. The SkullBreak dataset was also generated from the CQ500 but the binarization of the skulls, the patient selection and the implant generation is different, therefore, we assume these two datasets as if two different domains.

The CQ dataset was binarized by using the TotalSegmentator (an open-weights model that can segment several structures in CT and MRI scans) [14] to segment the skulls of all of the patients, as depicted in Figure 1. After skull segmentation, the defect-generation algorithm was applied to produce various defective versions of each skull.

The defects were generated by randomly sampling locations within the segmented skull; the randomly sampled voxels served as the defect centers. Two types of defects were used: spheres and rectangular blocks. Spherical defects were defined by a center and radius, while center coordinates and random dimensions defined rectangular defects. Defect masks were smoothed using a binary morphological opening and elastically deformed to emulate realistic defect boundaries.

To increase geometric variability, up to three additional border defects were sequentially added at randomly sampled points along the border of the initial defect. Each defect followed the same generation and smoothing process and was merged with the existing defects. To summarize, each skull had different defective versions; in each defective version, the skull could have up to n defects; and, finally, each of these defects could have smaller k defects on its borders.

The final defect volume was subtracted from the original binary skull volume to produce a defective skull, and the difference between the original skull and the defective one constituted the implant used for evaluation. For each patient, multiple defective volumes were created and saved along with the corresponding implant volumes and a CSV file containing detailed defect metadata.

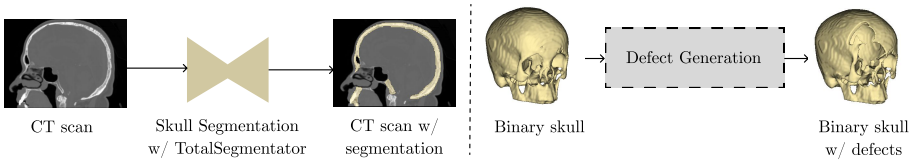


Fig. 1: CT scan to binarized image with defect generation.

The selection of patients from the CQ500 and generation algorithm can be found in our GitHub repository: `https://github.com/RicardoCQB/bc-wdm-3d`.

3.2 Conditional Wavelet Diffusion Model (cWDM)

Our framework utilizes a **Conditional Wavelet Diffusion Model (cWDM)** [4], a generative approach that produces data by reversing a progressive Gaussian noising process. Unlike Latent Diffusion Models [13] that rely on computationally expensive pre-trained autoencoders, the cWDM uses a 3D discrete Haar wavelet transform (DWT) to achieve dimensionality reduction without additional training. Let $y_0 \in \mathbb{R}^{D\times H\times W}$ denote the original binary skull and c_0 the defective condition. Both are decomposed via DWT into representations $x_0 = DWT(y_0)$ and $c_0' = DWT(c_0)$ in $\mathbb{R}^{8\times D/2\times H/2\times W/2}$, significantly lowering GPU memory requirements for 3D processing. The model learns a reverse Markovian process where a neural network ϵ_θ is trained to estimate the noise added at each step of the forward diffusion process [3]. During inference, a noise vector x_T is sampled from a standard normal distribution and iteratively denoised; at each step, the current sample is concatenated with the wavelet-transformed condition c_0' to guide the inpainting. The final predicted wavelet coefficients x_0^* are mapped back to the spatial domain using the inverse discrete wavelet transform (IDWT) as $y_0^* = IDWT(x_0^*)$ to produce the completed 3D skull volume.

4 Results and Discussion

In Figure 2 and Figure 3, we present the average Dice Similarity Coefficient (DSC) and 95th percentile Hausdorff Distance (HD95) values, respectively, for each trained model evaluated on the SkullBreak (SB) and CQ test sets. The models compared are: (i) SB, trained only on the SkullBreak dataset; (ii) SBCQ3, trained on SkullBreak and the CQ dataset where each patient had three different defective skulls generated; (iii) SBCQ5 (5 defects); (iv) SBCQ10 (10 defects); and (v) SBCQ20 (20 defects).

It is clear from the figures that the baseline SB model fails to generalize to the CQ test set, achieving a very low mean DSC of 0.305 and a very high mean HD95 of 105.66 ± 29.88 mm. In contrast, augmenting the training data with synthetically generated defects from the CQ dataset greatly improved performance on the CQ test set. As more defect examples are added (from SBCQ3 to SBCQ20), the mean DSC on the CQ test set increases significantly from 0.862 to 0.927, while the mean HD95 simultaneously decreased from 18.56 mm to 6.31 mm. The performance on the SB test set remained high across all models, with the mean DSC peaking at 0.903 for the SBCQ10 model. The performance of the final model, SBCQ20, appeared to stabilize on both test sets, achieving a DSC on the SB test set (0.889) similar to that of the SBCQ10 model (0.903), suggesting a point of diminishing returns with additional data augmentation.

The increments on the average DSC value in the CQ test set demonstrate the benefit of increasing the number of training cases using the defect generation algorithm. Relative to the SBCQ3 model, the mean DSC increased by approximately 3.4% for SBCQ5 to 0.891, 6.6% for SBCQ10 to 0.919, and 7.6% for SBCQ20 to 0.927. The benefits of increasing the number of training cases in our experiments appear to stabilize at around 10-20 defects per patient skull.

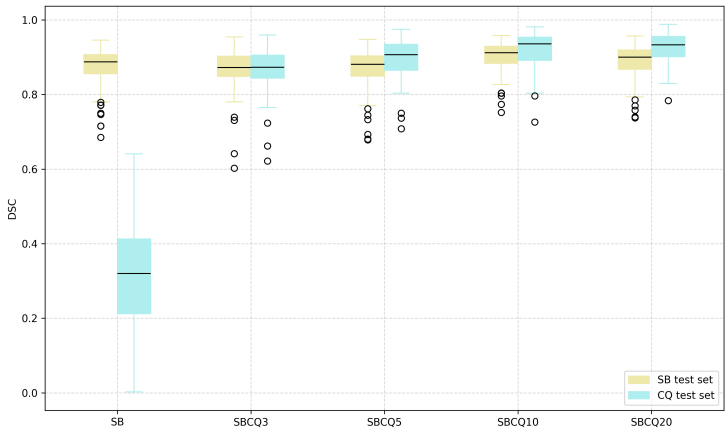


Fig. 2: Boxplot of the DSC performance distribution for each model on the SkullBreak (yellow) and CQ (blue) test sets.

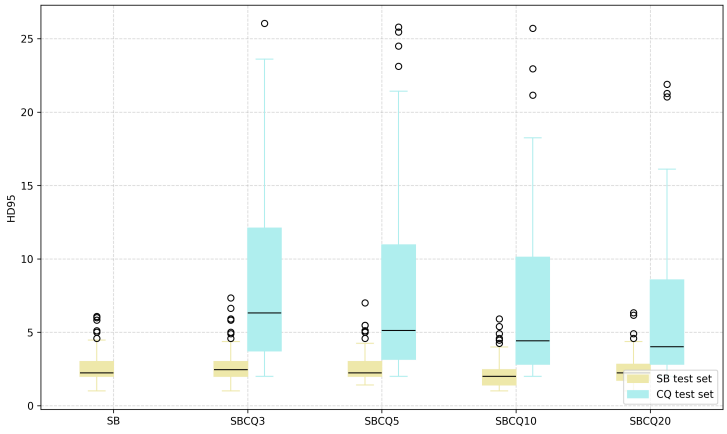


Fig. 3: Boxplot of the HD95 performance distribution for each model on the SkullBreak (yellow) and CQ (blue) test sets. The HD95 values for the baseline (SB) model were so high that we omitted them from the visualization.

In Figure 4, we observe that most models generate high-quality implants for the SkullBreak test set; the red and green circles indicate the greatest changes in implants for the CQ test set (blue implants). The SB model in the CQ test set erroneously filled skull cavities that should have remained open, such as the eye socket (orbita) and nose cavity. This proves that a model trained on the SB dataset alone does not generalize well enough to other datasets and that training a model with a small amount of newly generated data on a similar distribution can improve the implant generation greatly, as seen in the generated implants of the CQ test set obtained using the SBCQ3 model with only minor extra data.

Figure 5 demonstrates the use of our suggested pipeline in three patients who had part of their skull removed. The CT scans of the patients underwent TotalSegmentator segmentation for the cranium, and the binarized skull was then used as input to the SBCQ20 model to produce the resulting implants. The implants of Patient 1 and 3 show promising results for clinical usage as it fills the defect completely, thus providing the implant designer with a reasonable starting point towards the final implant. Alternatevily, the defect in the skull of Patient 2 is recognized but the inpainting model fails to completely fill the area of the defect.

5 Conclusions

This work presents CranioImpInp, an end-to-end framework that receives raw CT data, uses TotalSegmentator to do bone segmentation of the dataset's population and introduces new synthetic defects to the patient's skulls aiming to augment the training data and, finally, it trains a cWDM to generate skull implants that can help speed up the design process and thus, the surgical planning

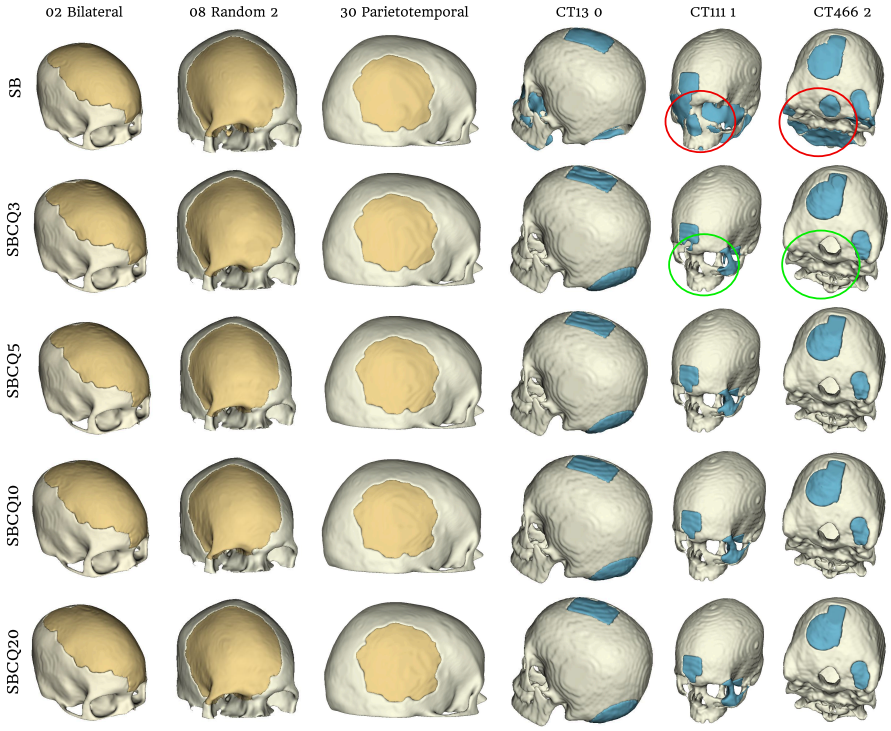


Fig. 4: Visualization of defective skulls with the generated implant from each trained model. The yellow implants correspond to the SB test set, and the blue implants to the CQ test set.

of cranioplasty. Results show higher generalization to out-of-distribution test sets. The SBCQ20 model achieved a DSC of 92.7% on the CQ test set, proving that a rule-based and inexpensive augmentation algorithm can improve performance in new clinical data, even if that data has no defected samples. While the current framework provides a logical solution for automatic implant generation, future work should focus on refining the low-cost synthetic data augmentation based on clinician recommendations and TBI defect statistics to further enhance surgical precision. By providing a scalable method to generate reproducible research resources, this work moves the field toward generalizable solutions that can significantly decrease the time and resources required for patient-specific implant design.

References

1. Alkhaibary, A., Alharbi, A., Alnefaie, N., Aloraidi, A., Khairy, S.: Cranioplasty: A comprehensive review of the history, materials, surgical as-

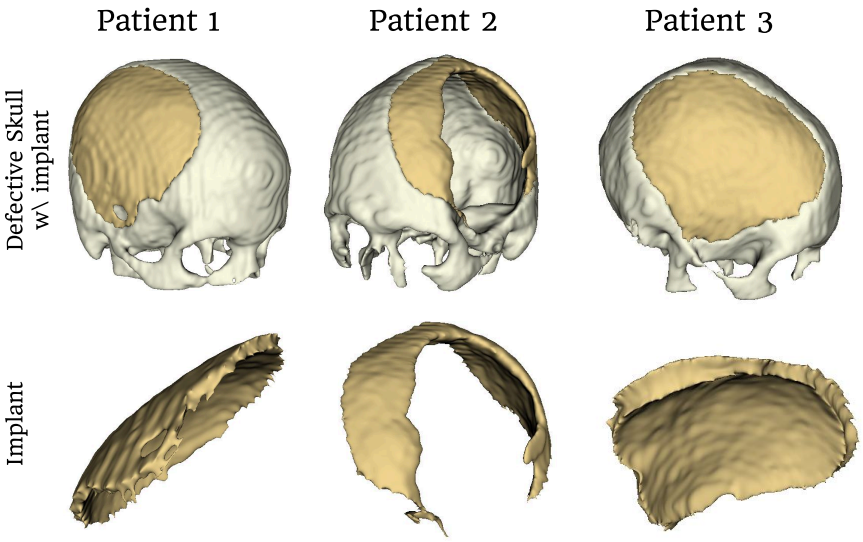


Fig. 5: Visualization of real defective skulls and the implant volumes.

pects, and complications. World Neurosurgery **139**, 445–452 (7 2020). https://doi.org/10.1016/J.WNEU.2020.04.211
2. Chilamkurthy, S., Ghosh, R., Tanamala, S., Biviji, M., Campeau, N.G., Venugopal, V.K., Mahajan, V., Rao, P., Warier, P., Ai, Q.: Development and validation of deep learning algorithms for detection of critical findings in head ct scans
3. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems **11**, 8780–8794 (5 2021), https://arxiv.org/pdf/2105.05233
4. Friedrich, P., Wolleb, J., Bieder, F., Durrer, A., Cattin, P.C.: Wdm: 3d wavelet diffusion models for high-resolution medical image synthesis (2 2024). https://doi.org/10.1007/978-3-031-72744-3_2, http://arxiv.org/abs/2402.19043http://dx.doi.org/10.1007/978-3-031-72744-3_2
5. Friedrich, P., Wolleb, J., Bieder, F., Thieringer, F.M., Cattin, P.C.: Point cloud diffusion models for automatic implant generation (3 2023). https://doi.org/10.1007/978-3-031-43996-4_11, http://arxiv.org/abs/2303.08061http://dx.doi.org/10.1007/978-3-031-43996-4_11
6. Haleem, A., Javaid, M.: Polyether ether ketone (peek) and its 3d printed implants applications in medical field: An overview. Clinical Epidemiology and Global Health **7**, 571–577 (12 2019). https://doi.org/10.1016/J.CEGH.2019.01.003, https://www.sciencedirect.com/science/article/pii/S2213398418303178
7. Kazerouni, A., Aghdam, E.K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., Merhof, D.: Diffusion models for medical image analysis: A comprehensive survey. arXiv pp. 1–25 (11 2022), https://arxiv.org/abs/2211.07804v3
8. Kodym, O., Li, J., Pepe, A., Gsaxner, C., Chilamkurthy, S., Egger, J., Španěl, M.: Skullbreak / skullfix – dataset for automatic cranial implant design and a benchmark for volumetric shape learning tasks. Data in Brief **35**, 106902 (4 2021).

https://doi.org/10.1016/J.DIB.2021.106902, https://www.sciencedirect.com/science/article/pii/S2352340921001864?ref=cra_js_challenge&fr=RR-1

9. Li, J., von Campe, G., Pepe, A., et al.: Automatic skull defect restoration and cranial implant generation for cranioplasty. Medical Image Analysis **73**, 102171 (10 2021). https://doi.org/10.1016/J.MEDIA.2021.102171, https://www.sciencedirect.com/science/article/abs/pii/S1361841521002176
10. Li, J., Ellis, D.G., Kodym, O., Rauschenbach, L., Rieß, C., et al.: Towards clinical applicability and computational efficiency in automatic cranial implant design: An overview of the autoimplant 2021 cranial implant design challenge. Medical Image Analysis **88**, 102865 (8 2023). https://doi.org/10.1016/J.MEDIA.2023.102865
11. Ntavelis, E., Siarohin, A., Olszewski, K., Wang, C., Van, L., Cvl, G., Zurich, E., Psi, C.H., Leuven, K.U., Sofia, U., Tulyakov, S.: Autodecoding latent 3d diffusion models. Advances in Neural Information Processing Systems **36** (7 2023), https://arxiv.org/abs/2307.05445v1
12. Phung, H., Dao, Q., Tran, A.: Wavelet diffusion models are fast and scalable image generators. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition **2023-June**, 10199–10208 (11 2022). https://doi.org/10.1109/CVPR52729.2023.00983, https://arxiv.org/pdf/2211.16152
13. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition **2022-June**, 10674–10685 (12 2021). https://doi.org/10.1109/CVPR52688.2022.01042, https://arxiv.org/pdf/2112.10752
14. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., Bach, M., Segeroth, M.: Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5** (9 2023), https://pubs.rsna.org/doi/10.1148/ryai.230024
15. Wodzinski, M., Daniol, M., Hemmerling, D.: Improving the Automatic Cranial Implant Design in Cranioplasty by Linking Different Datasets, pp. 29–44 (12 2021). https://doi.org/10.1007/978-3-030-92652-6_4
16. Wodzinski, M., Daniol, M., Socha, M., Hemmerling, D., Stanuch, M., Skalski, A.: Deep learning-based framework for automatic cranial defect reconstruction and implant modeling. Computer Methods and Programs in Biomedicine **226**, 107173 (11 2022). https://doi.org/10.1016/J.CMPB.2022.107173, https://www.sciencedirect.com/science/article/pii/S0169260722005545
17. Wodzinski, M., Kwarciak, K., Daniol, M., Hemmerling, D.: Improving deep learning-based automatic cranial defect reconstruction by heavy data augmentation: From image registration to latent diffusion models. Computers in Biology and Medicine **182**, 109129 (11 2024). https://doi.org/10.1016/J.COMPBIOMED.2024.109129, https://www.sciencedirect.com/science/article/pii/S0010482524012149
18. Wu, C.T., Yang, Y.H., Chang, Y.Z.: Creating high-resolution 3d cranial implant geometry using deep learning techniques. Frontiers in Bioengineering and Biotechnology **11**, 1297933 (2023). https://doi.org/10.3389/FBIOE.2023.1297933/FULL, https://pmc.ncbi.nlm.nih.gov/articles/PMC10750412/