

Patch-Based Deep Learning Approaches for Automated 3D Skull Segmentation in CT Imaging

Fatih Ekrem ONAT^{2,3}
fonat@yildiz.edu.tr

Yiğit KAHRAMAN²
yigit.kahraman@std.yildiz.edu.tr

Berk ARIDAĞ²
berk.aridag@std.yildiz.edu.tr

Melih Fırat BUDAK²
firat.budak@std.yildiz.edu.tr

Caner BULUT²
caner.bulut@std.yildiz.edu.tr

Berke APAYDINLI²
berke.apaydinli@std.yildiz.edu.tr

Abdülkadir GÜNEY⁴
abdulkadir.gunay@std.yildiz.edu.tr

Yunus Emre ÇAKMAKLI²
emre.cakmakli@std.yildiz.edu.tr

Nuri Serdar BAŞ¹
nuriserdarbas@gmail.com

Hamza Osman İLHAN³
hoilhan@yildiz.edu.tr

Görkem SERBES²
gserbes@yildiz.edu.tr

Mihriğül EKŞİ ALTAN^{2,4}
meksi@yildiz.edu.tr

¹Sağlık Bilimleri Üniversitesi Kanuni Sultan Süleyman Eğitim ve Araştırma Hastanesi, İstanbul, Türkiye

²Biyomedikal Mühendisliği, Yıldız Teknik Üniversitesi, İstanbul, Türkiye

³Bilgisayar Mühendisliği, Yıldız Teknik Üniversitesi, İstanbul, Türkiye

⁴Makine Mühendisliği, Yıldız Teknik Üniversitesi, İstanbul, Türkiye

Abstract—Accurate and efficient skull segmentation from cranial CT scans is a critical prerequisite for computer-aided cranial implant design, surgical planning, and postoperative assessment. In this study, we comparatively evaluate three state-of-the-art deep learning architectures (U-Net, Attention U-Net, and UNETR) for automated 3D skull segmentation. Experiments were conducted on a clinically curated novel dataset of 120 CT volumes encompassing diverse age groups and anatomical variations. Both patched and non-patched inference strategies were investigated to assess their impact on segmentation accuracy and robustness. Quantitative evaluation using a subject-based five-fold cross-validation protocol revealed that patched inference consistently outperformed non-patched processing, with Attention U-Net achieving the highest mean Dice Similarity Coefficient. Qualitative analysis demonstrated that patched inference produced anatomically complete and visually coherent skull reconstructions, particularly in challenging pediatric cases. The proposed approach shows strong potential for integration into clinical workflows, providing high-fidelity 3D reconstructions suitable for patient-specific cranial implant manufacturing.

Index Terms—Skull segmentation, Cranial CT, Deep learning, Attention U-Net, UNETR, Patch-based inference

I. INTRODUCTION

Skull segmentation (SS) in computed tomography (CT) imaging refers to automated extraction of cranial bone structures—such as the calvaria, facial bones, and sutures—from volumetric scans. This task underpins clinical workflows such as cranial implant design and reconstruction, where accurate

skull delineation directly affects implant fit and surgical outcomes [1]. Besides, segmentation is an essential component in craniofacial and orthodontic planning, improving anatomical landmark localization and measurement consistency [2].

Image-processing techniques, including thresholding and active contour level-set, have been applied for bone extraction in CT. However, these approaches often underperform in the presence of noise, artifacts, or in pediatric scans with incomplete sutural fusion [3]. The advent of deep learning, especially convolutional neural networks (CNNs) like U-Net, has significantly advanced the field, thanks to their robust localization and efficient learning from limited data [4, 5].

Due to privacy regulations and the sensitivity of cranial imaging, particularly in children, publicly annotated SS datasets remain scarce. Most existing studies rely on small, private collections, making reproducibility and benchmarking challenging. For instance, Ghadimi et al. [8] proposed a method for segmenting and reconstructing neonatal skulls using coupled level sets on a dataset of 18 CT scans, highlighting the limited scale of available annotation resources. While later efforts such as Xi et al. [9] introduced a larger pediatric CT cohort, only few datasets contain manually labeled skull masks suitable for supervised learning. In SS process, transformer-based models such as UNETR have emerged to address the limitation of local receptive fields in CNNs by capturing long-range spatial dependencies in 3D volumes while retaining multi-scale context through a U-Net-like structure [6]. Attention-augmented CNNs, like Attention U-Net, further enhance SS performance by dynamically focusing on contextually relevant anatomical regions during feature fusion [7].

In cranial implant planning, accurate SS is a critical first

This study was supported by The Scientific and Technological Research Council of Turkey (TÜBİTAK) under the ARDEB 1001 - Scientific and Technological Research Projects Funding Program, with the project number 123E494.

979-8-3315-5565-8/25 \$31.00 © 2025 IEEE

step. Precise delineation ensures reliable reconstruction, supports anatomically consistent implant design, and reduces manual corrections. Thus, SS robustness directly impacts the success of the implant pipeline, especially where structural precision is essential. Our study conducts a comparative evaluation of three deep learning methods (U-Net, Attention U-Net, and UNETR) for SS. The models were analyzed in terms of their architectural design, spatial context modeling capabilities, and segmentation accuracy across a heterogeneous dataset encompassing a broad range of anatomical variations.

II. MATERIALS AND METHODS

A. Dataset Construction and Preparation

A total of 120 cranial CT scans were collected from Bağcılar Training and Research Hospital under an approved institutional ethics protocol (Approval No: 2024/01/14/014). The dataset comprises both pediatric and adult cases, capturing a wide range of anatomical variability. All scans were initially annotated by an experienced radiologist and subsequently reviewed by a second senior radiologist to ensure clinical validity of the ground truth masks for bone structures. This dataset was used for training, validation, and comparative evaluation of segmentation models throughout the study. The demographic distribution of the dataset, including age statistics and gender ratios, is summarized in Table I.

TABLE I
DEMOGRAPHIC PROPERTIES OF THE DATASET.

Variable	Property	Value
Count	# Samples	120
Age (year)	min	0.33
	max	85
	mean	31.26
	median	26.5
	std.	24.89
Sex	Male (M)	66
	Female (F)	54

B. Data Augmentation

To improve model generalization and reduce overfitting, we applied a set of medical image-specific data augmentation. Training augmentations included spatial normalization orientation standardization, isotropic resampling, intensity scaling, and foreground cropping. In augmentation, firstly, spatial transformations including random cropping of $32 \times 32 \times 32$ patches with balanced positive-negative sampling, flipping along each spatial axis (x, y, z), random 90-degree rotations, and intensity shifting were applied. Secondly, the images were resized into $192 \times 192 \times 128$ without being divided into patches. All transformations were applied on-the-fly during training. Validation data underwent the same normalization pipeline without augmentation to ensure consistency across evaluations.

C. Segmentation Methods

Three state-of-the-art deep learning architectures (U-Net, Attention U-Net, and UNETR) were employed for 3D cranial CT segmentation. To investigate the effect of inference strategy

on performance, each model was evaluated under two settings: patched inference, where the CT volume is processed in smaller subvolumes to preserve local anatomical detail, and non-patched inference, where the full volume is segmented in a single pass. This comparison enabled assessment of how localized context modeling impacts skull boundary delineation and overall SS quality.

1) *U-Net Architecture*: U-Net is a CNN that follows an encoder-decoder structure, originally designed for biomedical image segmentation [4]. Its simplicity, symmetry, and use of skip connections make it an effective architecture for segmenting anatomical structures such as the skull, where precise delineation of bony boundaries is required. The encoder path of U-Net consists of successive convolutional blocks, each containing two 3×3 convolution layers with ReLU activations, followed by 2×2 max pooling operations for spatial downsampling. The decoder mirrors this structure using 2×2 transposed convolutions to progressively upsample the feature maps. Skip connections directly link encoder and decoder feature maps at corresponding resolutions, facilitating the recovery of spatial details lost during downsampling. U-Net was selected in this study due to its strong localization capabilities, efficient learning on limited data, and proven effectiveness in biomedical segmentation tasks involving high-resolution structural boundaries. Skip connections reintroduce high-frequency boundary cues essential for thin cortical bone, making U-Net a strong baseline for cranial structures.

2) *Attention U-Net Architecture*: Attention U-Net extends the original U-Net architecture by introducing *Attention Gates* (AGs) into the skip connections [7]. These gates learn to suppress irrelevant feature responses and highlight salient structures relevant to the target task during feature fusion. This mechanism proves particularly beneficial in cranial segmentation tasks, where anatomical boundaries often exhibit low contrast and irregular morphology due to patient-specific anatomical variations or pathology.

3) *UNETR Architecture*: UNETR is a hybrid encoder-decoder architecture that integrates a Vision Transformer (ViT) as the encoder backbone within a U-Net-like framework [6]. Unlike conventional CNN encoders, UNETR leverages self-attention mechanisms to model long-range dependencies across volumetric inputs, which is particularly valuable for segmenting large or discontinuous cranial structures. The input volume is divided into non-overlapping patches of size $P \times P \times P$. Each patch is flattened, projected, and combined with positional embeddings to form a sequence of tokens. This sequence is processed by transformer blocks composed of multi-head self-attention and multilayer perceptron layers. The outputs from selected transformer layers are reshaped into spatial feature maps and integrated into the decoder through skip connections. The decoder follows the classical U-Net structure with transposed convolutions for upsampling.

D. Training Configuration and Evaluation Protocol

All deep learning models were trained using the AdamW optimizer with a learning rate of 10^{-4} and a weight decay

of 10^{-5} . The combined Dice and Cross-Entropy loss function was employed to supervise voxel-wise segmentation outputs, balancing overlap-based and probabilistic accuracy metrics. Each model was trained for 15,000 iterations with a batch size of 1 on 3D volumetric CT inputs. Training was performed on a workstation equipped with an NVIDIA RTX 4090 GPU (24 GB VRAM), 64 GB system RAM, and an Intel i9-13980HX CPU. Performance was evaluated using the Dice Similarity Coefficient (DSC) as the primary segmentation metric. A subject-based five-fold cross-validation protocol was implemented, ensuring that all scans from a patient were assigned exclusively to either the training, validation, or test set within each fold. This strategy prevented data leakage between folds and provided a realistic assessment of model generalizability across diverse patient anatomies. Final performance was reported as the average Dice score across all folds.

III. EXPERIMENTAL RESULTS

Table II summarizes the quantitative performance of the evaluated segmentation models (U-Net, Attention U-Net, and UNETR) across both patched and non-patched inference settings. The primary evaluation metric was the DSC, reported as mean, standard deviation, and range statistics over the test set. Across all models, patched inference consistently outperformed non-patched inference, yielding higher mean DSC values and reduced variance. This improvement is likely attributable to the reduced memory footprint and increased local context consistency in smaller subvolumes, which can enhance boundary delineation for fine anatomical structures.

TABLE II

DSC STATISTICS FOR DIFFERENT MODELS AND SEGMENTATION TYPES.

Model	Type	DICE				
		Mean	Std	Min	Median	Max
Attention U-Net	Patched	0.9335	0.0279	0.8135	0.9367	0.9777
U-Net	Non-Patched	0.7836	0.0604	0.6106	0.7982	0.8691
U-Net	Patched	0.9211	0.0357	0.7549	0.9293	0.9710
U-Net	Non-Patched	0.7865	0.0538	0.6378	0.8031	0.8645
UNETR	Patched	0.9239	0.0406	0.7560	0.9341	0.9775
UNETR	Non-Patched	0.7812	0.0630	0.6059	0.7981	0.8678

Attention U-Net achieved the highest mean DSC (0.9335 ± 0.0279), followed by UNETR (0.9239 ± 0.0406) and U-Net (0.9211 ± 0.0357). The improvement of Attention U-Net over the baseline U-Net suggests that the attention gating mechanism effectively suppresses irrelevant features and emphasizes boundary-relevant structures—particularly beneficial for cranial bone contours where local intensity transitions can be subtle. In the non-patched setting, all models experienced a performance drop, with mean DSCs ranging between 0.7812 and 0.7865. U-Net achieved the highest score (0.7865), followed closely by Attention U-Net (0.7836) and UNETR (0.7812). This aligns with findings in prior studies [6, 7], where non-patched inference was shown to dilute local contrast information due to the dominance of global context. Fig. 1 visually confirms these trends, illustrating both the superiority of patched inference and the incremental gains from attention

mechanisms. Notably, the performance bars demonstrate that patched models not only achieve higher mean DSC but also maintain more stable performance across the dataset.

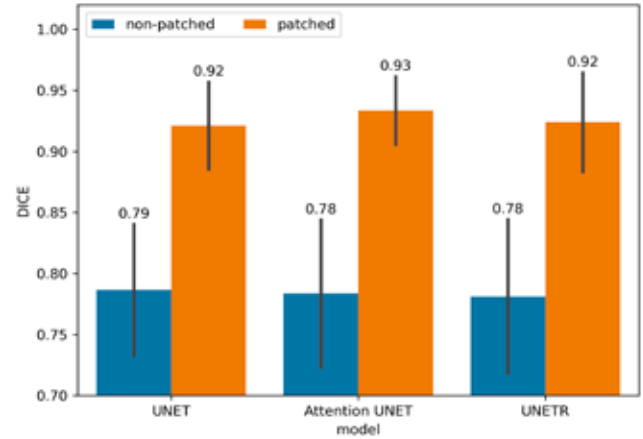


Fig. 1. The performance of all three models. Processing the images as patches has given significantly better performance than whole volumes.

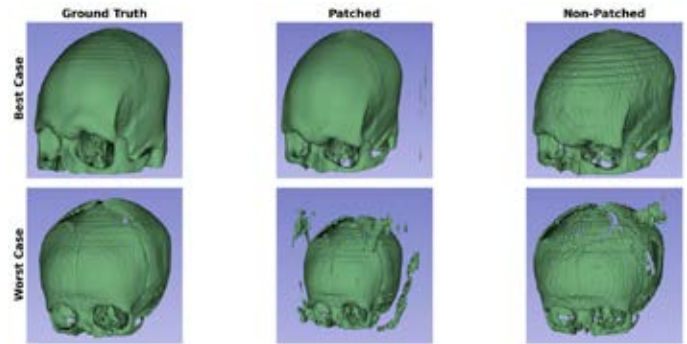


Fig. 2. The best case and the worst case for Attention UNET, compared to ground truth.

Given that Attention U-Net achieved the highest overall performance, qualitative results for this model are presented in Fig. 2. The best-case example corresponds to a 1-year-old infant scan, where the patched inference output exhibits near-perfect alignment with the ground truth. Cranial sutures and curved anatomical surfaces are reconstructed with high fidelity, and no visible over- or under-segmentation is observed. This demonstrates the model's ability to capture fine structural details when bone mineralization is relatively complete and sutures are well-defined. In contrast, the worst-case example involves a 4-month-old neonatal scan, where skull bones are thinner, sutures are wider, and fontanelles remain open. These characteristics result in reduced CT radiodensity contrast and more complex local intensity patterns, which present a significant challenge for automated SS. Additionally, in both patched and non-patched inferences, the presence of the infant's hand within the scan field further degraded segmentation quality by introducing non-cranial structures with similar intensity values. This led to false-positive regions and irregular boundaries in the reconstructed skull, particularly in the non-patched

setting where global context overshadowed fine local details. Patched inference partially mitigated these issues by providing the model with higher-resolution local context and reducing the influence of irrelevant anatomical regions.

IV. DISCUSSION AND CONCLUSION

The results demonstrate the effectiveness of deep learning methods for automated SS. Across three architectures (U-Net, Attention U-Net, and UNETR) and two inference strategies (patched and non-patched), all models produced anatomically coherent and clinically relevant skull masks. The patched Attention U-Net achieved the highest performance, delivering precise cranial boundary delineation while minimizing segmentation errors. Results indicate that, with suitable inference strategies, deep models can provide robust outputs for applications such as cranial implant design and surgical planning.

Among the evaluated architectures, the patched Attention U-Net achieved the highest mean DSC, confirming that attention mechanisms effectively suppress irrelevant background responses while enhancing boundary delineation. Although the transformer-based UNETR also demonstrated competitive results, the attention-augmented CNN maintained a slight advantage in this application, likely due to the localized and well-defined nature of cranial bone structures. These findings are consistent with prior work in cranial imaging: for instance, Xi et al. [9] reported a DSC of 0.96 in pediatric cranial CT segmentation using a ResU-Net, underscoring the feasibility of achieving near-expert accuracy. In our experiments, patched inference results ($DSC > 0.93$) confirm that CNN-based architectures with attention remain highly competitive, combining accuracy with computational efficiency. Qualitative analysis further highlighted the benefits of patch-based inference, which produced more complete and coherent segmentations in anatomically challenging cases such as thin cranial bones, wide sutures, or open fontanelles. Even in neonatal scans containing non-cranial structures (e.g., an infant's hand within the CT field), patched inference reduced false positives by leveraging finer local context. Patched inference consistently yielded more coherent reconstructions by allowing models to focus on fine boundaries (e.g., sutures, thin cortices) that were often blurred in non-patched predictions. The performance gain arises because patches preserve local contrast, reduce the downsampling needed for non-patched inputs, and—with overlap and blending—mitigate block artifacts at tile borders.

Beyond segmentation benchmarking, the findings of this work highlight the potential of automated skull reconstruction frameworks in clinical cranial implant manufacturing. In trauma, tumor resection, or congenital defects, accurate 3D skull models enable the design of patient-specific implants with precise anatomical fit. From a clinical perspective, reliable and consistent SS forms the foundation for applications such as implant design, surgical planning, and postoperative assessment. Our results indicate that patch-based strategies, combined with attention mechanisms, offer a practical balance between accuracy and computational efficiency, making them suitable for integration into such workflows.

A potential refinement to enhance clinical applicability is Connected Component Analysis (CCA), which retains only the largest connected structure in the predicted mask. By removing smaller noise-induced components, CCA produces cleaner reconstructions, reducing false positives and improving the reliability of outputs for applications such as implant generation. Regarding the limitations, the dataset, although diverse in age range and anatomical variability, remains limited in size. Additionally, only three architectures were evaluated; exploring more recent hybrid or transformer-based variants could provide further insights. Finally, while DSC is a robust metric, future studies could incorporate surface distance or shape similarity measures to better assess clinical applicability. Future work may also explore hybrid CNN–Transformer decoders and assess whether the proposed patch-based strategy transfers to other modalities (e.g., MRI). In conclusion, patch-based inference combined with attention-enhanced architectures, particularly Attention U-Net, provides accurate and robust SS. The approach achieves high Dice scores and generates anatomically precise reconstructions, making it well-suited as the initial step in cranial implant design.

REFERENCES

- [1] M. Wodzinski, M. Daniol, M. Socha, *et al.*, “Deep learning-based framework for automatic cranial defect reconstruction and implant modeling,” in *AutoImplant Challenge, MICCAI*, 2022.
- [2] M. Gillot, “Automatic multi-anatomical skull structure segmentation of CBCT volumes using UNETR and MONAI,” *PLOS ONE*, 2022.
- [3] L. Zhou, A. Toga, Z. Xue, “An overview of bone segmentation in clinical computed tomography,” *arXiv:1504.00221*, 2015.
- [4] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in *Proc. MICCAI*, 2015, pp. 234–241.
- [5] G. Litjens, T. Kooi, B. Bejnordi, *et al.*, “A survey on deep learning in medical image analysis,” *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.
- [6] A. Hatamizadeh, V. Nath, Y. Tang, *et al.*, “UNETR: Transformers for 3D Medical Image Segmentation,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2022, pp. 574–584.
- [7] O. Oktay, J. Schlemper, L. Le Folgoc, *et al.*, “Attention U-Net: Learning Where to Look for the Pancreas,” *arXiv preprint arXiv:1804.03999*, 2018.
- [8] S. Ghadimi, H. A. Moghaddam, R. Grebe, and F. Wallois, “Skull segmentation and reconstruction from newborn CT images using coupled level sets,” *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 2, pp. 563–573, 2016.
- [9] F. Xi, L. Tu, F. Zhou, Y. Zhou, J. Ma, and Y. Peng, “Automatic segmentation and quantitative analysis of brain CT volume in 2-year-olds using deep learning model,” *Frontiers in Neurology*, vol. 16, art. no. 1573060, 2025.