Computerized Medical Imaging and Graphics 123 (2025) 102541

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

Automatic semantic segmentation of the osseous structures of the paranasal sinuses

Yichun Sun, Alejandro Guerrero-López, Julián D. Arias-Londoño, Juan I. Godino-Llorente *

Escuela Técnica Superior de Ingenieros de Telecomunicación, Universidad Politécnica de Madrid, Madrid, 28040, Spain

ARTICLE INFO

MSC:
0000
1111

Keywords:
Automatic semantic segmentation
CT
Osseous structures
Paranasal sinuses
U-Net
Neuronavigation
Robot-assisted surgery

ABSTRACT

Endoscopic sinus and skull base surgeries require the use of precise neuronavigation techniques, which may take advantage of accurate delimitation of surrounding structures. This delimitation is critical for robotic-assisted surgery procedures to limit volumes of no resection. In this respect, an accurate segmentation of the osseous structures of the paranasal sinuses is a relevant issue to protect critical anatomic structures during these surgeries. Currently, manual segmentation of these structures is a labour-intensive task and requires wide expertise, often leading to inconsistencies. This is due to the lack of publicly available automatic models specifically tailored for the automatic delineation of the complex osseous structures of the paranasal sinuses. To address this gap, we introduce an open source dataset and a UNet SwinTR model for the segmentation of these complex structures. The initial model was trained on nine complete ex vivo CT scans of the paranasal region and then improved with semi-supervised learning techniques. When tested on an external dataset recorded under different conditions, it achieved a DICE score of 98.25 ± 0.9. These results underscore the effectiveness of the model and its potential for broader research applications. By providing both the dataset and the model publicly available, this work aims to catalyse further research that could improve the precision of clinical interventions of endoscopic sinus and skull-based surgeries.

1. Introduction

Paranasal sinus surgery, particularly as a treatment for rhinosinusitis or removal of nasal polyps, has become frequent (Gupta et al., 2021). These surgeries use endoscopic techniques that are approached through the nasal cavity (Zhao et al., 2021). Alternatively, skull-based surgeries, such as pituitary tumours, meningiomas, or acoustic neuromas, are also procedures that are approached – in most cases – through the nasal cavity using endoscopic techniques (Martinez-Perez et al., 2021). Due to the proximity of the paranasal area to the ocular orbit and cranial nerves, precision in these surgeries is crucial to avoid major complications such as blindness, central nervous system injuries, trauma, and even death. This underscores the importance of high-precision navigation procedures to minimise risks and ensure optimal outcomes (Sieśkiewicz et al., 2009).

The support of computerised navigation systems to approach endoscopic sinus and skull base surgeries is a common approach to achieve the required level of precision (Fu et al., 2021). Modern neuronavigation systems allow real-time monitoring of the position of the endoscopic instrument relative to the three planes of the preoperative CT or Magnetic Resonance (MR) scans of the patient (Lauretti et al., 2018). The information provided by the neuronavigation tools can also be complemented with acoustic or haptic feedback when the boundaries of a certain region of interest are reached (Thatikunta et al., 2020).

CT imaging greatly improved the understanding of sinonasal anatomy and pathology since its introduction (Rao and El-Noueam, 1998). This high-resolution imaging technique is essential not only for identifying anatomical variants (Mokhasanavisu et al., 2019) but also for precise surgical planning in sinus and skull-based surgeries. Its application directly influences outcomes in treating conditions such as nasopharyngeal cancer (Tsang et al., 2022), obstructive sleep apnea (Lechien et al., 2021), and rhinosinusitis (He et al., 2020). Additionally, CT imaging is key for planning surgeries for skull base tumours, including fossa meningioma (Perin et al., 2021), enhancing surgical safety.

Detailed segmentation of the osseous structures surrounding the skull base and paranasal sinuses (illustrated in Fig. 1) can significantly improve the characteristics of current neuronavigators by establishing clear boundaries for volumes that cannot be resected, thus reducing

* Corresponding author.
E-mail address: ignacio.godino@upm.es (J.I. Godino-Llorente).

https://doi.org/10.1016/j.compmedimag.2025.102541
Received 30 September 2024; Received in revised form 6 March 2025; Accepted 25 March 2025
Available online 4 April 2025

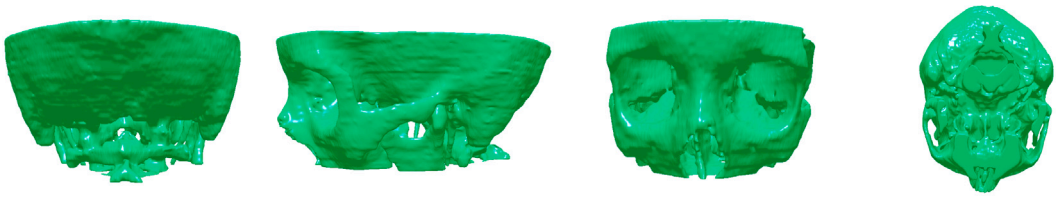

Fig. 1. 3D reconstruction of the OSPS. From left to right: back, lateral, front, and bottom views.

surgical complications. However, the complex anatomical structures of the paranasal sinuses and their significant variability between individuals (Mokhasanavisu et al., 2019) pose particular challenges not yet solved.

Furthermore, delimiting anatomical structures is especially critical for robot-assisted endoscopic sinus and skull base surgeries, where precise identification of 3D boundaries is essential (Yang et al., 2017). By accurately mapping osseous structures, surgeons can navigate complex anatomical volumes adjacent to the paranasal sinuses more effectively, thus minimising the risks associated with these invasive procedures (Singh et al., 2020). This task is typically carried out manually or semi-automatically (i.e., automatically but supervised by human experts), which is considered the gold standard for this purpose. This is a laborious procedure due to the complexity and size of the tiny structures to be delineated, and the large number of slices available for each CT scan (Heimann et al., 2009). Besides, it requires high levels of expertise (Cellina et al., 2021), is very time-consuming, and introduces a certain variability due to different delineation criteria introduced by different experts. Consequently, new automatic methods are required to segment the aforementioned structures.

In recent years, driven by advances in Deep Learning (DL) and image processing, substantial progress has been made in the automatic segmentation of many body structures and/or tissues using CT scans (Long et al., 2015). These advances are significant in delineating osseous structures (Ahmed and Mstafa, 2022), which in some cases are relatively straightforward to segment due to their size, density, and contrast with respect to their surrounding tissues. In this respect, large bones, such as the femur, tibia, or fibula, are typically simple to segment (Lu et al., 2024) achieving DICE scores up to 97.28 ± 1.73. Ribs present more complications (Yang et al., 2021), scoring with a DICE up to 94.9. Further complex targets include the segmentation of maxillofacial structures (Dot et al., 2022; Park et al., 2022), with DICE scores ranging from 82 to 94. In addition, studies in Ding et al. (2023) and Wang et al. (2021) present a much more challenging segmentation of the associated tiny structures in the inner ear, reporting scores that range from 56.0 for the steps to 95.2 for the labyrinth.

The works in Dot et al. (2022) and Park et al. (2022) present automatic models for the segmentation of the upper skull, including the entire volume from the crown to the maxilla. The authors used advanced neural networks such as nnU-Net and modified U-Nets, achieving DICE scores of 96.2 and 96.5, respectively. These scores are biased by the better precision obtained for large skull bones, such as frontal, parietal, and occipital. Consequently, they do not represent the specific precision achieved in the segmentation of the complex Osseous Structures surrounding the Paranasal Sinuses (OSPS).

In addition, Gillot et al. (2022) reports a DICE score of 78.8 ± 10.3 using a 3D U-Net transformer-based (UNETR), for complex structures of the cranial base that include some of the adjacent regions of the paranasal sinuses. Furthermore, a DICE score of 94.0 is reported for the skull base in Steybe et al. (2022) using a 3D U-Net.

In the context of sinus scan analysis, existing research focuses mainly on segmenting the upper airway rather than osseous structures. In this regard, (Choi et al., 2022) achieved a DICE score of 90.7 for the maxillary sinus. Other approaches reported DICE scores up to 93.0 for the entire upper airway (including all paranasal sinuses) using a 3D U-Net (Wu et al., 2021; Steybe et al., 2022). These results are of obvious interest, but the strong contrast and texture of the area corresponding to the upper airway significantly reduce the complexity of the challenge.

In summary, recent research on automatic segmentation of the osseous structures of the skull focuses on the large cranial bones. However, the specific challenges of delineating the OSPS are scarcely addressed. Thus, new automatic methods are required to segment these intricate structures. In addition, to our knowledge, there are no open datasets of CT scans specifically annotated with the OSPS. This is in part attributed to the specialised and hard work required to delineate each scan (Heimann et al., 2009). As an example and to illustrate the effort required to manually annotate sinus volume, the work in Pirner et al. (2009) reports an average of 13 h for a semi-automatic segmentation of the upper airway. Due to the complexity of the intricate osseous structures of the paranasal sinuses, the time required to segment them is expected to be much longer than that dedicated to the upper airway.

2. Material and methods

This section describes the material and methods, starting from the datasets, the annotation process to create the masks, the methods for pre-processing the images, the architectures used, and the experimental protocol.

2.1. Datasets

This section introduces two corpora used for each EP. These datasets were retrospectively collected in 2012, and have been used since then in different research projects and with different objectives. Data collection was carried out according to strict ethical protocols. In the context of this paper, the two corpora are next referred to as ID and ED.

All CT scans are axial views stored in Digital Imaging and Communications in Medicine (DICOM) format. The scanners – seven different scanners were used – are spiral and correspond to four different models from three manufacturers. The slices were reconstructed using six distinct kernels that were fixed during the acquisition process. All slices are 2D greyscale images, each with a resolution of 512×512 pixels. Table 1 summarises the most relevant characteristics of each dataset. Detailed information corresponding to each CT scan and patient can be accessed in the public GitHub repository.[1]

ID. Comprises full head CT ex vivo scans obtained post-mortem from twelve individuals who donated their organs for research purposes (see Fig. 2.Top). They were recorded using a Canon® Aquilion scanner. Tissues were treated with specific preservative treatments, which sometimes partially fill the upper airway but do not affect the cranial structures in any sense. Each scan contains slices with a thickness of 1.00 mm. These scans cover the whole head volume, from the crown to the neck.

ED. The second dataset includes in vivo CT sinus scans[2] (i.e., with a field of view restricted to the sinus area (Fig. 2.Down), and

[1] https://github.com/BYO-UPM/Craneal_CT.

[2] Sinus CT scans are recommended in case of sinus cancer and other malignant and metastatic tumours, inflammatory diseases of the sinuses, trauma, and preoperative assessments.

Table 1
Characteristics of the Internal and External datasets.

Dataset	Hospital	Model and manufacturer	Kernel	Subjects
ID	F	Canon Aquilion	FC30	12
ED	A	Canon Aquilion	FC10, FC12, FC30	17
	B	Canon Aquilion	FC10, FC12, FC30	4
		GE® LightSpeed VCT	BONE	2
	C	GE® LightSpeed VCT	BONEPLUS	1
	D	Siemens®	H70s	3
	E	GE HiSpeed Dual	BONE	1

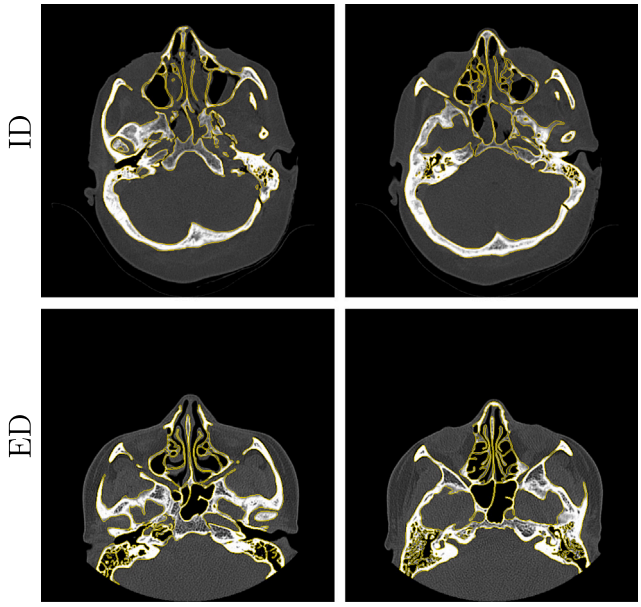

Fig. 2. Examples of manual segmentation of the OSPS. Figures in the top correspond with slices taken from the *ID* (full head scans). Figures in the bottom correspond with slices taken from the *ED* (sinus scans). Note the differences in the field of view.

with a range from the upper alveolar bone to the frontal sinuses), which were collected during clinical routines. They correspond to forty additional patients (with an average age of 29 ± 16 yrs. old) who agreed on the use of their data for research purposes. The scans were obtained using six different scanners from five different hospitals (as detailed in Table 1), and retrieved from their corresponding Picture Archiving and Communication System (PACS). The slice thickness ranged from 0.47 to 1.00 mm, with a mean value of 0.59 ± 0.17 mm.

2.2. Data annotation

An annotation procedure is required to create the masks needed for the training and validation of the models. The gold standard is a manual delimitation. This is a labour-intensive step that requires extensive knowledge of the anatomical structures of the skull, as stated above. The structures of interest were manually delineated with 1-pixel width contours (Fig. 3), and the bowels were subsequently filled to obtain the final masks. Fig. 3 exemplifies the complexity of the manual

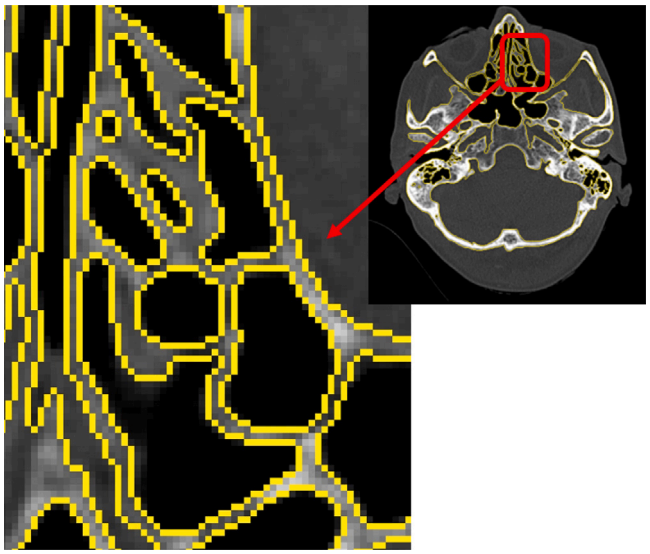

Fig. 3. Detail of the contours manually delineated in an area corresponding to the ethmoidal sinuses. The left image is a zoomed-in view of the red box on the right. The structures are carefully delineated using 1-pixel width contours. Note that some of the structures to be segmented are also 1 or 2 pixels wide.

annotation procedure. A different manual annotation was used for each dataset:

ID. Nine out of the twelve scans available in the ID were manually segmented from scratch by two experts (i.e., no semi-automatic methods were used). One of the experts carried out a reviewing process of all manual contours to ensure the consistency of the masks created. The expert coordinating the process went through each CT slice for each patient and modified (at pixel level) the contours provided by the remaining experts. Those structures missed were added, and those wrongly introduced were removed. Besides, the coordinator modified the contours in those structures that were significantly under or oversegmented.

Axial slices annotated are those in the volume of interest (i.e., from the upper alveolar bone to the frontal sinuses, including the entire paranasal sinuses region). This volume is covered by 64 adjacent slices. The slices of each of the nine subjects, along with their corresponding masks, are collectively designated as Subset 1 (S1). Due to the already commented complexity of the structures to be delineated, the annotation time ranged from 40 to 90 min. per slice (with an average time of 63 min.), i.e., 80 h per volume.

ED. For the ED, manual masks were delineated by seven different experts. The expert who carried out the reviewing process of the ID reviewed the consistency and accuracy of the masks created in this scenario following the aforementioned reviewing procedure. A total of 4 CT scans out of 28 were partially annotated. These scans were chosen to correspond to the 4 different CT machines available in the dataset and 5 different hospitals, specifically, A, B, C, D and E.

Aligned with certain trends in the state-of-the-art (Pirner et al., 2009; Xu et al., 2021), the annotation methods applied to the ED followed a semi-automatic procedure. The masks were extracted using the best-trained model with the ID (see Section 2.8), and manually corrected pixel-wise. The number of slices selected depends on the thickness: 64 for those scans with a thickness of 1.00 mm., 85 when it was 0.75 mm., and 128 when it was 0.50 mm. Subsequently, within this range, 30 discontinuous slices were manually annotated. These are spaced with 1–4 unannotated slices, but the volume covered is still from the upper alveolar bone to the frontal sinuses. These manually annotated slices are collectively designated as Subset 2 (S2). The manual annotation took between 15 and 50 min. per slice (25 min.

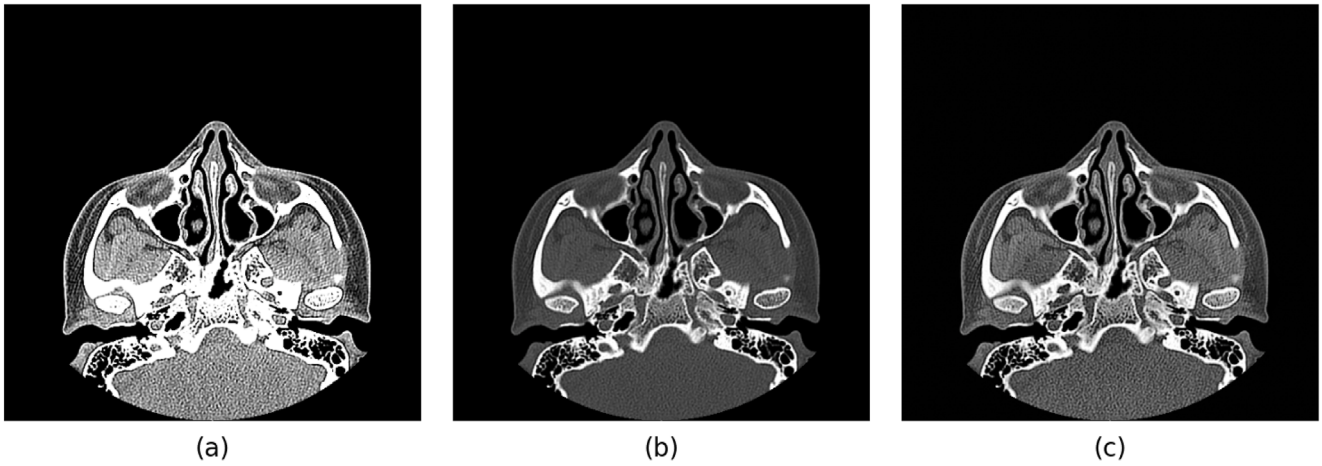

Fig. 4. Example of a preprocessed CT slice. (a) Original CT slice. (b) The windowed CT slice which was standardised with [WL, WW] = [500, 2000]. (c) The CT slice after applying preprocessing steps (normalisation and equalisation).

Table 2
Details of data annotation and distribution of subsets.

Dataset	Subjects	Subset	Slices/Patient	Manual masks/Patient
ID	9	S1	64	64 continuous
ED	4	S2	64	30 discontinuous
			64	30 discontinuous
			85	30 discontinuous
			128	30 discontinuous
ID & ED	27	S3	All	0

on average). The remaining slices in the ED —not included in S2—, have been automatically segmented by the model trained using S1 without manual correction. In contrast, the non-annotated slices of both datasets have been collectively designated as Subset 3 (S3). Table 2 provides detailed information about the annotated data and their distribution across subsets.

In summary, S1, derived from the ID, includes CT scans from 9 subjects obtained using the same equipment, containing a total of 576 slices and their corresponding manual masks. S2, from the ED, corresponds to scans obtained using four different CT scanners, and comprises 120 slices and their corresponding masks (excluding those not annotated). Together, this results in 696 samples available with ground truth.

2.3. Data preprocessing

All DICOM files were converted to 2D PNG images (one for each slice). In addition, all CT scans were standardised with the same window size (represented by its WL and WW). The window size was initially fixed as [WL, WW] = [500, 2000] for all CT scans in both datasets. Subsequently, the images were normalised and equalised. Greyscale images of one channel were scaled in the range [0,1]. Then, Contrast Limited Adaptive Histogram Equalisation (CLAHE) was used to enhance contrast (Pizer et al., 1987; Zuiderveld, 1994; Arias-Londoño and Godino-Llorente, 2024). The image was divided into 8×8 small blocks, each of which was equalised independently. In this respect, the contrast limit threshold was set to 2.0. Fig. 4 presents an example of a CT slice after preprocessing.

2.4. Data augmentation

Data Augmentation (DA) techniques have proven to be effective for improving DL models in many different applications (Chlap et al., 2021; Bansal et al., 2022). The DA methods used in this study follow the proposed in Isensee et al. (2020) and include: (i) spatial augmentations such as rotations (angles sampled from uniform distributions) and scaling (factors from 0.7 to 1.4), applied together with a probability of 0.2, along with centre-cropping of patches to the target size; and (ii) intensity-based augmentations, including Gaussian noise (variance up to 0.1), Gaussian blur (kernel width from 0.5 to 1.5 voxels), brightness adjustment (multiplication by factors between 0.7 and 1.3), contrast adjustment (factors between 0.65 and 1.5), low-resolution simulation (downsampling by factors up to 2), gamma transformation (exponent from 0.7 to 1.5), and mirroring (with a probability of 0.5 along all axes). These techniques were systematically applied to enhance the robustness and generalisability of the models across different configurations.

2.5. Evaluation metrics

Three evaluation metrcs were used to evaluate the results: the DICE, the Hausdorff distance (HD) and the Volume Similarity Index (VSI).

The DICE score was proposed to effectively address data imbalance since anatomical structures or lesions tend to be quite small compared to the background or the rest of the image. The DICE score measures the similarity between two structures, ranging from 0 to 1, with higher values indicating greater similarity between masks.

$$\mathrm{DICE} = \frac{2 \cdot |\mathrm{X} \cap \mathrm{Y}|}{|\mathrm{X}| + |\mathrm{Y}|}, \tag{1}$$

where X and Y represent the pixel values of the two images, respectively.

The HD quantifies the maximum discrepancy between the boundaries of two segmentations. Let $\mathrm{X_{HD}}$ and $\mathrm{Y_{HD}}$ be the sets of boundary points extracted from the ground truth and the predicted segmentation, $\mathrm{x_{hd}}$ and $\mathrm{y_{hd}}$ represent the points in the sets $\mathrm{X_{HD}}$ and $\mathrm{Y_{HD}}$, respectively. The directed Hausdorff distance from $\mathrm{X_{HD}}$ to $\mathrm{Y_{HD}}$ is defined as:

$$h(\mathrm{X_{HD}}, \mathrm{Y_{HD}}) = \max_{x_{hd} \in \mathrm{X_{HD}}} \min_{y_{hd} \in \mathrm{Y_{HD}}} |x_{hd} - y_{hd}|, \tag{2}$$

and similarly, the directed distance from $\mathrm{Y_{HD}}$ to $\mathrm{X_{HD}}$ is:

$$h(\mathrm{Y_{HD}}, \mathrm{X_{HD}}) = \max_{y_{hd} \in \mathrm{Y_{HD}}} \min_{x_{hd} \in \mathrm{X_{HD}}} |y_{hd} - x_{hd}|. \tag{3}$$

The symmetric HD is then given by:

$$H(\mathrm{X_{HD}}, \mathrm{Y_{HD}}) = \max h(\mathrm{X_{HD}}, \mathrm{Y_{HD}}), ; h(\mathrm{Y_{HD}}, \mathrm{X_{HD}}). \tag{4}$$

A lower HD indicates a closer match between the boundaries of the two segmentations.

The VSI measures the relative difference in the overall volume (i.e., total segmented pixels or voxels) between two images, irrespective of spatial overlap. It is computed as:

$$\mathrm{VSI} = 1 - \frac{|V_X - V_Y|}{V_X + V_Y}, \tag{5}$$

where V_X and V_Y represent the total number of positive (segmented) pixels in the ground truth and the predicted segmentation, respectively. A VSI value closer to 1 indicates that both volumes are very similar, whereas values deviating from 1 imply greater discrepancies.

For readability purposes, all DICE and VSI scores in this article were scaled to the range 0–100.

2.6. Network architectures

The network architectures in this work are based on U-Net (Ronneberger et al., 2015), which has consistently demonstrated strong performance in medical image segmentation (Dot et al., 2022; Park et al., 2022; Gillot et al., 2022; Steybe et al., 2022).

To determine whether 2D or 3D architectures are more suitable for this problem, a vanilla U-Net was first coded from scratch. The 2D U-Net was tested with greyscale input images, using a series of convolutional and activation layers, dropout, and pooling in the encoder and decoder. The same architecture was applied for the 3D U-Net using Conv3D layers instead, with adjustments to batch size and training epochs.

Once the optimal dimensionality (2D or 3D) was established, the chosen architecture was further compared with pre-trained backbones like ResNet50 (He et al., 2016a) and VGG16 (Simonyan and Zisserman, 2014), known for their effectiveness in medical image segmentation (Frid-Adar et al., 2018; Lefkovits and Laszlo, 2022).

These custom built models were then compared to the state-of-the-art nnUNet framework (Isensee et al., 2021), which includes various network architectures and backbones. Finally, these models were evaluated against large, pre-trained, plug-and-play models such as SAM (Kirillov et al., 2023) and MedSAM (Ma et al., 2024a), which are designed for general-purpose segmentation and adapted for medical images.

In all cases, the segmentation results were post-processed to remove small, isolated regions and fill in unwanted gaps, enhancing the quality and accuracy of the segmentation (Gillot et al., 2022).

2.7. Loss functions

Three different loss functions were used: nnUNet default loss, the Hybrid Loss (HL), and the Asymmetric Unified Focal Loss (AUFL).

2.7.1. BCE+DICE loss

The BCE+DICE loss is the default in nnUNet, and combines a Binary Cross-Entropy (BCE) loss with DICE loss to address class imbalances and improve segmentation accuracy.

BCE loss measures the difference between predicted probabilities and actual binary labels for each pixel:

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{N} \sum i = 1^{N} \left[y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \right] \tag{6}$$

where N is the number of pixels, y_i is the true label, and p_i is the predicted probability.

DICE loss, derived from the DICE score, optimises the overlap between the predicted segmentation and ground truth:

$$\mathcal{L}_{\mathrm{DICE}} = 1 - \mathrm{DICE} \tag{7}$$

Combining these, the BCE+DICE loss balances pixel-wise accuracy with overall segmentation quality:

$$\mathcal{L}_{\mathrm{BCE+DICE}} = \mathcal{L}_{\mathrm{BCE}} + \mathcal{L}_{\mathrm{DICE}} \tag{8}$$

This combined approach is particularly effective for complex medical image segmentation tasks.

2.7.2. Hybrid loss

HL combines the DICE loss with the Focal loss function. It is widely used in image segmentation (Yeung et al., 2023), such as in SAM (Kirillov et al., 2023).

To improve the stability of the gradient during training (Yeung et al., 2022), the Focal loss is incorporated. It introduces a modulating factor γ, which dynamically scales the cross-entropy loss ($\mathcal{L}_{\mathrm{CE}}$), reducing the weight of easily distinguishable samples and emphasising the hard-to-distinguish ones. The expression for the $\boldsymbol{\alpha}$-balanced version of the Focal loss is shown in Eq. (9), where $\boldsymbol{\alpha}$ represents the class weight vector, $\mathbf{p}_i$ is the predicted probabilities vector for each class $p_{i,c}$, where C is the number of classes and i the pixel pointer (Lin et al., 2017). Hence, $p_{i,c}$ is the probability of the ith pixel to belong to class c.

$$\mathcal{L}_{\mathrm{Focal}} = -\frac{1}{N} \sum_{i=1}^{N} \frac{1}{C} \boldsymbol{\alpha} \odot ((1 - \mathbf{p}_i)^{\gamma})^{T} \underbrace{\log(\mathbf{p}_i)}_{\mathcal{L}_{\mathrm{CE}}} \tag{9}$$

where N denotes the number of pixels in an image. Therefore, the HL function is defined as $\mathcal{L}_{\mathrm{Hybrid}} = \mathcal{L}_{\mathrm{DICE}} + \mathcal{L}_{\mathrm{Focal}}$.

2.7.3. Asymmetric unified focal loss

This loss function consists of a modified Focal loss and a modified Focal Tversky loss. According to Yeung et al. (2022), this loss function resolves the issue of excessive hyperparameters in the original Tversky and Focal loss functions, as well as the convergence problems of the Focal loss at the end of training. Furthermore, through selective enhancement or suppression by focal parameters, asymmetry allows for different losses to be assigned to each class, thereby overcoming the harmful suppression of rare classes and the enhancement of background ones. The expression for the modified Focal Loss ($\mathcal{L}_{mF}$) is defined in Eq. (10) (Yeung et al., 2022), where the parameter $\boldsymbol{\alpha}$ in Eq. (9) is replaced with a constant δ.

$$\mathcal{L}_{mF} = -\frac{1}{N} \sum_{i=1}^{N} \left(\mathbb{1}_{y_i} \delta \log\left(p_{i,r}\right) + (1 - \delta) \sum_{c \neq r} (1 - p_{i,c})^{\gamma} \log\left(p_{i,c}\right) \right) \tag{10}$$

where y_i is the target class for the ith sample (pixel) and $\mathbb{1}_{y_i}$ is an indicator function that takes value 1 if $y_i = r$ and 0 otherwise.

On the other hand, the modified focal Tversky loss ($\mathcal{L}_{\mathrm{mFT}}$) can be expressed as in Eq. (11) (Yeung et al., 2022), where $\gamma < 1$ increases the focus on harder examples, and mTI is the modified Tversky index (Yeung et al., 2022; Salehi et al., 2017).

$$\mathcal{L}_{\mathrm{mFT}} = \sum_{c \neq r} (1 - \mathrm{mTI}) + \sum_{c=r} (1 - \mathrm{mTI})^{1-\gamma} \tag{11}$$

where mTI is expressed as (Eq. (12)):

$$\mathrm{mTI} = \frac{\sum_{i=1}^{N} p_{i,1} y_i}{\sum_{i=1}^{N} p_{i,1} y_i + \delta \sum_{i=1}^{N} p_{i,1}(1 - y_i) + (1 - \delta) \sum_{i=1}^{N} p_{i,0} y_i} \tag{12}$$

where y_i is the pixel's target value, which takes 1 for the foreground and 0 for the background.

Using the previous definitions, the AUFL ($\mathcal{L}_{AUF}$) can be expressed as (Eq. (13)):

$$\mathcal{L}_{AUF} = \gamma \mathcal{L}_{\mathrm{mF}} + (1 - \gamma) \mathcal{L}_{\mathrm{mFT}} \tag{13}$$

In this work, the hyperparameters of the AUFL were set to $\lambda = 0.5, \delta = 0.6$ and $\gamma = 0.5$, as proposed in Yeung et al. (2022).

2.8. Experimental protocol

Different experiments were carried out to identify the most suitable scheme for automatic segmentation of the osseous structures of the paranasal sinuses. Three experimental phases were proposed, namely: EP A, EP B, and EP C. They are summarised in Table 3, and illustrated graphically in Fig. 5.

The experiments were carried out using a high-performance computer based on an ASUS® board, with an AMD® EPYC 7313 3.000 GHz CPU, 512 GB of RAM, and four NVIDIA® Quadro RTX A6000 GPUs with 48 GB of VRAM.

Table 3
Summary of the experimental protocol.

	Exp.	Subset	DL methods	Network architecture	Loss function	DA
EP A	E0	S1	Supervised	2D vs. 3D	HL	No
	E1	S1	Supervised	U-Net vs. nnUNet	All	No
EP B	E2	S1	Zero-shot	SAM & MedSAM	–	No
EP C	E3	S1 & S2	Supervised	Optimal	Optimal	Yes
	E4	S1 & S2	Semi-supervised	Optimal	Optimal	Yes
	E5	S1 & S2 & S3	Semi-supervised	Optimal	Optimal	Yes

2.8.1. Experimental Phase A

The goal of this phase was to identify the optimal architecture and loss function for the task. The experiments were conducted using S1 from the ID, consisting of 576 slices from 9 patients. A 9-fold Cross-Validation (CV) was performed with a 7:1:1 split for training, validation, and testing, ensuring that the same patients were never used across different sets.

E0. We first tested the differences of using 2D or 3D architectures training a U-Net from scratch.

E1. We tested seven different architectures without DA and evaluated them with three loss functions to determine the best combination. The architectures included a custom U-Net with a VGG16 backbone and several state-of-the-art models such as nnUNet (Isensee et al., 2021) with various backbones (ResNet50 (He et al., 2016b), Transformer (Hatamizadeh et al., 2022), Swin Transformer (Hatamizadeh et al., 2021), and U-MAMBA (Ma et al., 2024b)). Furthermore, we implemented and adapted the U-KAN (Li et al., 2024) network, based on the Kolmogorov–Arnold Network (Liu et al., 2024), specifically for our task, independent of the nnUNet framework. For the loss functions, we used the HL, AUFL, and the DICE+BCE loss from nnUNet, as detailed in Section 2.7.

2.8.2. Experimental Phase B

The goal of this phase was to compare the performance of the best model from EP A with respect to general-purpose pre-trained state-of-the-art segmentation solutions.

E2. In E2, the SAM and MedSAM models were tested in the same patients as in EP A, i.e., S1, in a zero-shot procedure.

2.8.3. Experimental Phase C

This EP aimed to test and improve the model's generalisation to an out-of-distribution dataset. Experiments 3, 4, and 5 (E3, E4 and E5) were conducted using all patients labelled from S1 and different selections from Subsets 2 (S2) and 3 (S3). To maximise the number of training samples, no validation set was included. All experiments in this phase used the best architecture and loss function identified from EPs A and B, i.e. nnUNet SwinTR with HL. The training time per epoch ranged between 5–16 h depending on the experiment.

E3. In E3, four models were trained using all samples from S1 plus three of the four available from S2. The remaining patient from S2 was used for testing, resulting in a 4-fold CV. In total, 696 slices from 13 patients were used for training and testing purposes. Two variants of E3 were tested to validate whether DA improved the generalisation of the model.

E4. In E4, and based on the results of E3, a semi-supervised learning strategy was developed. All unlabelled slices of S2 (221 slices) were pseudo-labelled using the model obtained in E3 for each fold. This semi-supervised approach expanded the available data to 917 slices from 13 patients. The strategy is similar to that followed in Lee (2013) and Arias-Londoño et al. (2023).

Table 4
DICE scores obtained for E0 in mean ± std.

Exp.	U-Net	Vanilla	ResNet50	VGG16
E0	2D	90.20 ± 2.3	91.55 ± 1.6	91.70 ± 1.6
	3D	90.71 ± 1.3	89.88 ± 1.3	90.77 ± 1.5

E5. In E5, 30 slices manually selected from each patient of S3 were pseudo-labelled with the E4 model obtaining 810 pseudo-masks for training. Then, following same CV as in E3 and E4, a final model was trained. To balance the number of labelled and pseudo-labelled samples while minimising error propagation, the training included 12 labelled patients and 12 pseudo-labelled. Each epoch randomly selected 12 pseudo-labelled patients from S3, thus increasing the variability. Furthermore, to decrease computational burden, a transfer learning strategy was used, starting from the models obtained in E4 for each fold.

3. Results

The results obtained for all experiments carried out in all phases (EP A, EP B and EP C) are presented below.

3.1. Experimental Phase A

Table 4 illustrates the results obtained for E0. According to the results, the use of 2D and 3D U-Net architectures has a negligible impact on the results, as the DICE coefficients are very similar for both architectures. However, in terms of training time, 2D architectures converged faster than 3D ones (0.15 min. per epoch for 2D models, and 0.3 min. for 3D ones). Thus, 2D architectures are chosen over 3D.

Table 5 presents the results obtained for E1. Most configurations of models and loss functions achieved DICE scores above 90, with the exception of U-KAN combined with AUFL, which scored the lowest at 89.81 ± 4.9.

In terms of loss functions, performance was generally comparable across the board, with each loss function outperforming the others in at least one model configuration. The DICE+BCE loss delivered the best results for three models (UNet VGG16, U-KAN, and U-MAMBA Enc), while AUFL performed the best for the three other models (nnUNet ResNet, nnUNetTR, and U-MAMBA Bottom). The HL outperformed the others in only one configuration, but notably with nnUNetSwinTR, which achieved the highest overall DICE score.

In terms of architectures, the U-MAMBA models were the only ones that consistently scored above 92 points, regardless of the loss function used.

However, the combination of nnUNetSwinTR with HL produced the overall best DICE score. Therefore, this combination is selected as the optimal architecture and loss function for subsequent experiments.

3.2. Experimental Phase B

The qualitative results obtained using SAM and MedSAM are shown in Fig. 6, where each colour represents a different class. The SAM model (Fig. 6.b) was able to segment the background and the external contour of the patient's head. However, it struggled to detect the osseous

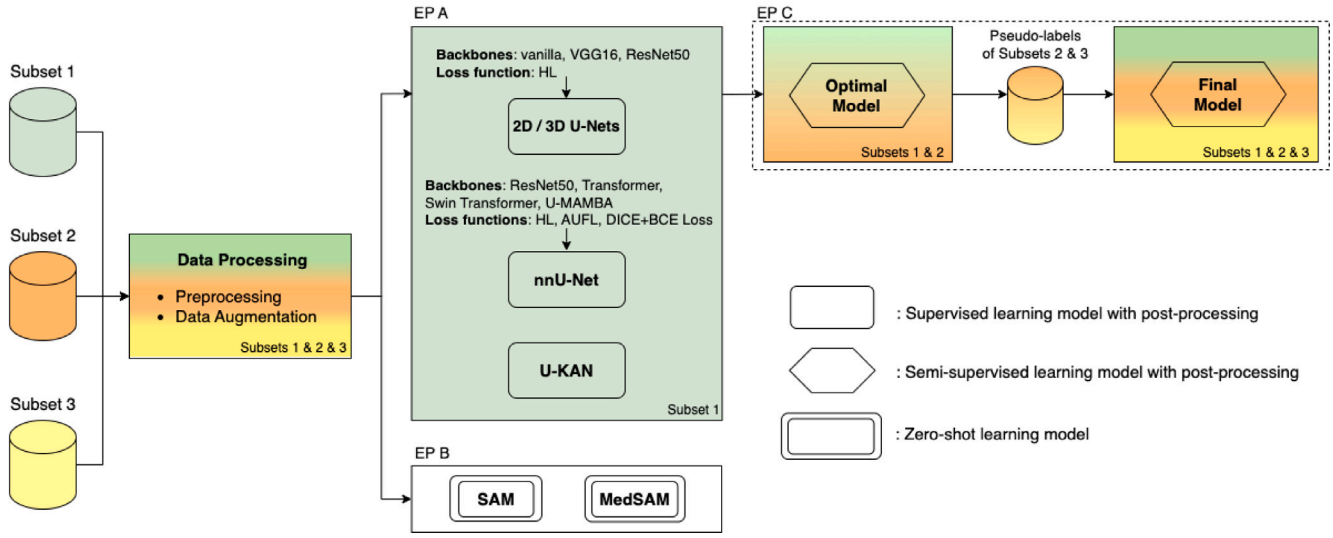

Fig. 5. Outline of the overall experimentation protocol. Colours are an indication of the subset used for each EP being S1 green, S2 orange and S3 yellow.

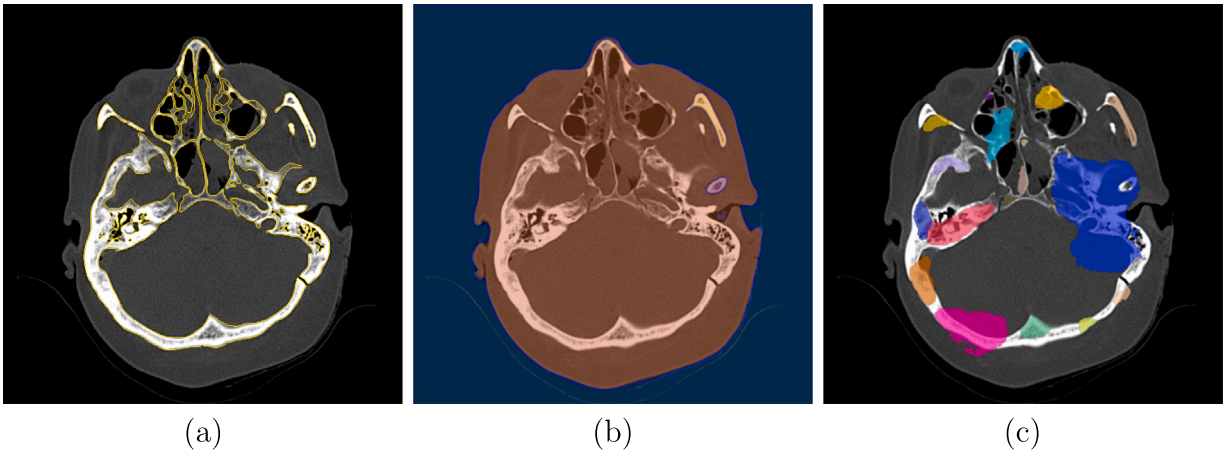

Fig. 6. Examples of masks generated by the SAM and MedSAM models. (a) Original CT slices with manual contour lines (ground truth). (b) Results provided by SAM. (c) Results obtained by MedSAM.

Table 5
DICE scores obtained for E1 from EP A in mean ± std for different architectures and loss functions. Best result is highlighted in bold.

Model	HL	AUFL	DICE+BCE loss
UNet VGG16	91.70 ± 1.6	91.73 ± 1.7	91.84 ± 1.6
nnUNet ResNet	92.42 ± 2.3	92.46 ± 2.4	91.74 ± 2.1
nnUNetTR	91.74 ± 2.4	91.96 ± 2.5	91.32 ± 1.3
nnUNet SwinTR	**92.73 ± 2.4**	92.27 ± 2.5	91.43 ± 1.5
U-KAN	90.26 ± 3.2	89.81 ± 4.9	91.41 ± 1.8
U-MAMBA Bot	92.21 ± 2.5	92.32 ± 2.4	92.12 ± 2.2
U-MAMBA Enc	92.10 ± 2.3	92.14 ± 2.5	92.19 ± 2.2

* Mean ± std.

structures. For example, only part of the frontal sinus was detected, while the other paranasal sinuses were not identified.

A similar pattern was observed with the MedSAM model (Fig. 6.c). Despite the visual prompts providing to MedSAM, the segmentations were not accurately aligned with the osseous structures.

3.3. Experimental Phase C

In this phase, the generalisability of the model was evaluated. DICE scores reported in this section were calculated using S2.

Table 6
DICE scores obtained for EP C, i.e., E3, E4 and E5.

Exp.	DA	Fold 1	Fold 2	Fold 3	Fold 4	Mean ± std
E3	No	92.51	88.65	94.94	96.42	93.89 ± 2.1
E3	Yes	94.56	93.75	94.92	96.65	95.16 ± 0.9
E4	Yes	94.80	94.88	95.12	96.73	95.45 ± 0.7
E5	Yes	98.80	98.53	99.16	96.82	**98.25 ± 0.9**

Table 6 presents those results obtained for E3, E4 and E5. These results show that the best model developed in E1 was able to generalise to the external S2: the DICE scores even improved with respect to those in E1. Fig. 7 graphically presents several examples of predictions from E3, E4 and E5.

Furthermore, with respect to E3, results in E4 and E5 showed an increase in the mean DICE and a reduction in its standard deviation. This suggests that the semi-supervised and transfer learning strategies followed in these experiments were able to improve the generalisability of the model.

A comparison between the results obtained for E4 and E5 was carried out using statistical analysis, including Welch's t-test and Wilcoxon–Mann–Whitney test. For Welch's t-test, the null hypothesis (H_0) stated that there was no significant difference in the means of E4 and E5. The test returned a p-value of 0.0055, leading to

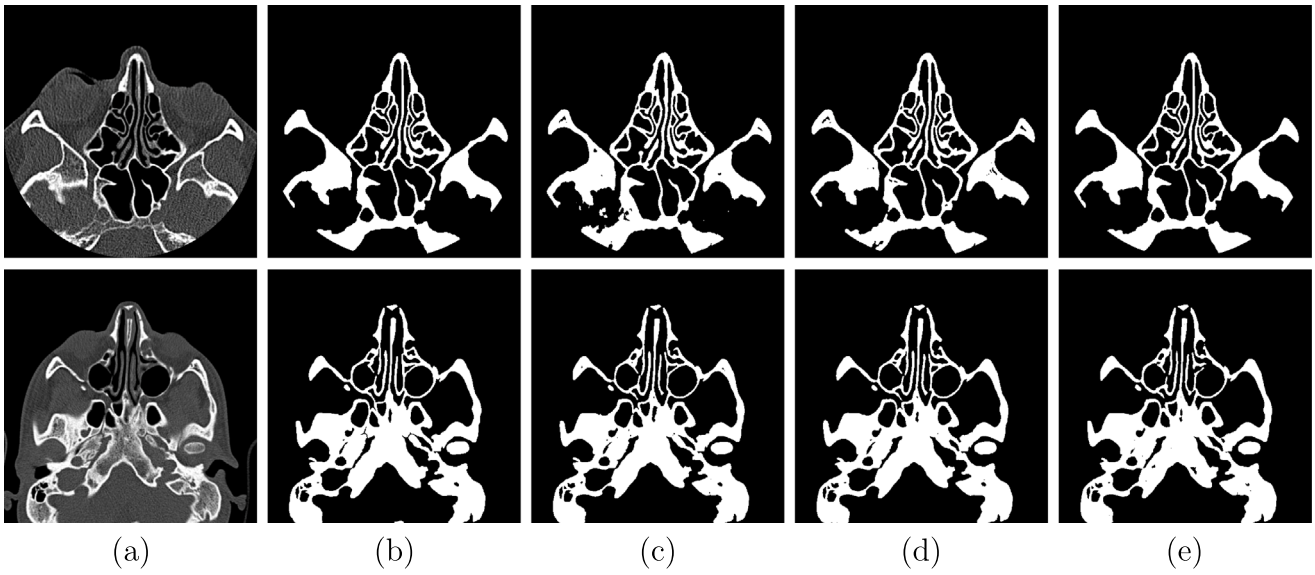

Fig. 7. Results from EP C. (a) Original CT slices. (b) Ground truth. (c) & (d) & (e) Masks obtained in E3, E4, and E5, respectively.

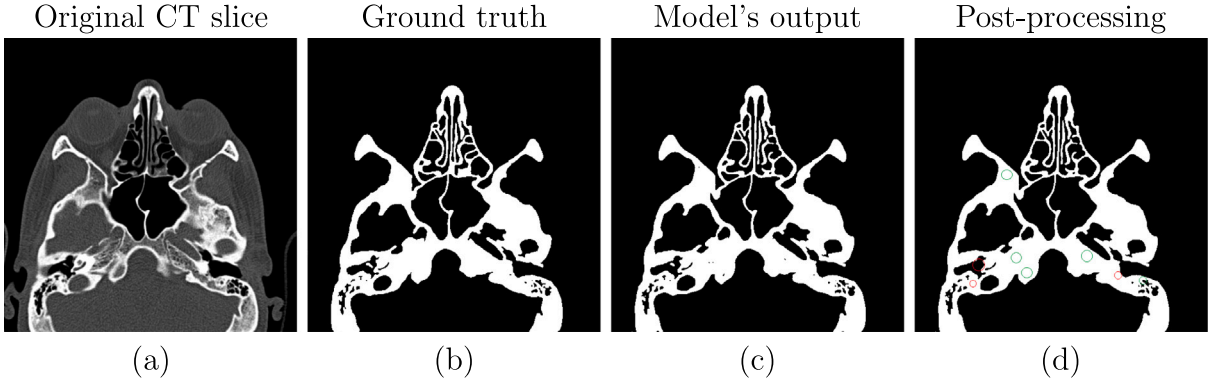

Fig. 8. Illustration of results from E5. (a) Original CT slice. (b) Ground truth. (c) Results without post-processing. (d) Results post-processed (green circles indicate areas where post-processing successfully addressed false positives or true negatives; red circles indicate errors).

Table 7
Different metrics obtained for E3 (with DA), E4 and E5.

	Fold 1			Fold 2			Fold 3			Fold 4			Mean ± std		
	DICE	HD	VSI	DICE	HD	VSI	DICE	HD	VSI	DICE	HD	VSI	DICE	HD	VSI
E3	94.69	23.43	98.64	94.37	32.32	96.61	94.90	18.50	99.78	96.68	18.57	99.14	95.16 ± 0.01	23.21 ± 6.50	98.54 ± 0.01
E4	94.85	10.71	99.63	95.16	22.28	98.22	95.09	17.80	99.96	96.71	19.40	98.79	95.45 ± 0.01	17.55 ± 4.92	99.15 ± 0.01
E5	98.80	4.62	99.82	98.49	6.76	99.98	99.10	8.89	99.90	96.60	17.40	99.45	98.25 ± 0.01	9.42 ± 5.60	99.79 ± 0.01

the rejection of H_0 and suggesting a significant difference. For the Wilcoxon–Mann–Whitney test, H_0 posited that the two samples, E4 and E5, come from the same distribution. The p-value in this case was 0.0286, prompting the rejection of H_0 and indicating a significant difference in the distributions of E4 and E5. Both tests showed significant differences, aligning with the main conclusions obtained through cross-validation, thus further validating the reliability of the findings.

Besides, Welch's and Wilcoxon–Mann–Whitney statistical tests were also carried out for E5 to evaluate potential differences due to the sex of the subject, and due to the manufacturer of the equipment. No significant differences were found, but the size of the dataset suggests prudence in this regard.

Alternatively, Table 7 presents a comparison of the results for the experiments: E3 with DA, E4 and E5. The metrics in the table further reinforce the superiority of E5. Specifically, the DICE and VSI values for E5 exhibit clear improvements over both E3 DA and E4, highlighting E5's enhanced performance. Moreover, E5's HD values are consistently lower, reflecting better precision and stability compared to the other experiments.

The impact of the post-processing stage in the E5 is illustrated in Fig. 8. Despite visible minor errors, post-processing effectively improved the reliability and thus is considered a crucial step.

Finally, Fig. 9 illustrates different errors committed by the system trained in E5. A comparison of the ground truth with the predicted masks reveals that although they were very similar for large structures, the predicted masks exhibit certain errors in fine detail. This discrepancy was particularly pronounced along the continuity of osseous structures around the ethmoid sinuses, and in other small anatomical structures.

Regarding training time, E5 required 150 min. per epoch. This is attributed to a larger training set – from 576 to 696 slices, resulting in 26,448 slices after DA –. In E4, training took 180 min. per epoch due to a larger training set – 917 slices, 34,846 after DA –. In E5, 360 slices from S3 were randomly added to the training set for each

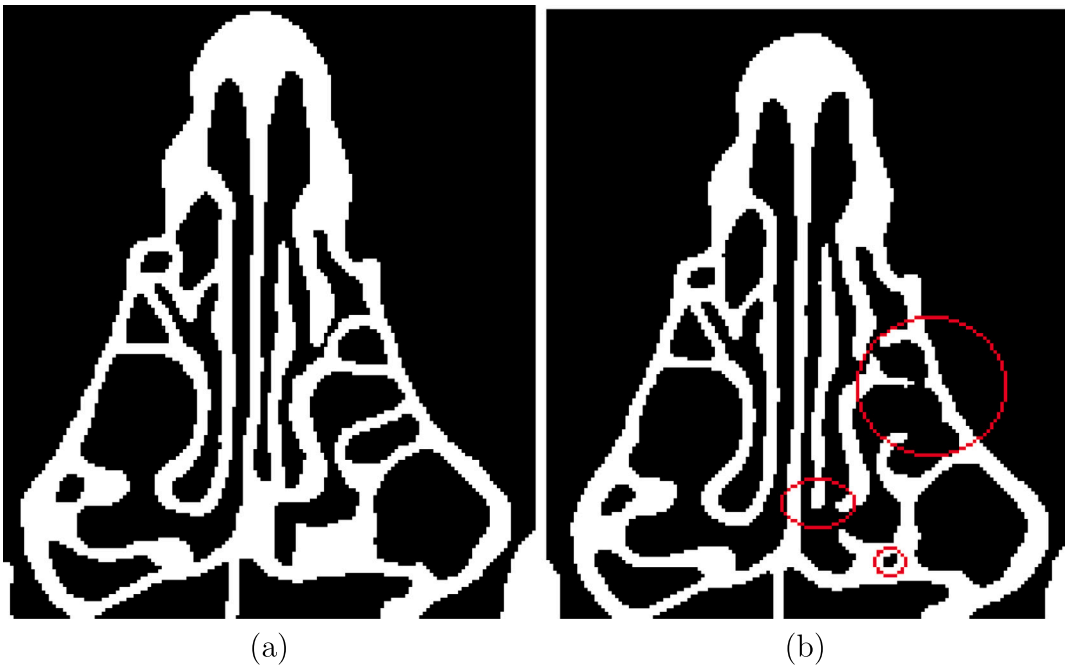

Fig. 9. Zoomed-in view of typical errors in the segmentation process. (a) Ground truth. (b) Results from E5.

epoch, increasing the number of slices – to 1277, 48,526 after DA –. Consequently, the training time per epoch increased to 270 min.

With respect to the inference time for E5, the 64 slices of a CT scan were segmented in 2 ± 1.1 s using only one A6000 GPU with 48 GB of VRAM of those available.

4. Discussion

Automated segmentation of the OSPS can enhance the computer-assisted workflow for preoperative virtual surgical planning, CAD/CAM-assisted or neuronavigated surgery, and postoperative verification of the outcomes. While accurate segmentation of osseous structures is often the starting point for virtual planning, it provides valuable information for computer-assisted planning and surgery. For example, in combination with neuronavigators, can offer crucial intraoperative information to the surgeon by providing visual and/or haptic feedback during surgery, or can provide extra information for the creation and precise placement of patient-specific implants (e.g., in reconstructing complex orbital defects). Additionally, in septoplasty, automated segmentation of the osseous structures can aid in virtual surgery planning and execution.

Besides, rapid advances in augmented and virtual reality applications are expected to further promote computer-assisted craniomaxillofacial surgery. To fully exploit these developments, detailed and precise segmentation procedures for the OSPS are crucial.

However, to our knowledge, no studies have specifically targeted personalised segmentation of the OSPS. Instead, existing research has concentrated mainly on segmenting the upper airway within the paranasal sinuses (Cellina et al., 2021), leaving a notable gap in the precise delineation of the osseous structures, which is considered a much more challenging task.

In addition, the absence of open datasets significantly limits not only the research about specific segmentation techniques for delineating these structures, but also the development of models specifically tailored to these anatomically complex regions. To our knowledge, the corpus used in this study is the first developed for this purpose that has been manually annotated and made openly available.

To this respect and to support diverse research needs, we have developed – and also made available – four different models: (i) trained on nine ex vivo patients; (ii) trained on all patients with manual segmentations (from both ID and ED); (iii) semi-supervised model incorporating manual and pseudo-labels from 13 patients; and (iv) a model trained on 13 manually segmented patients plus 27 pseudo-labelled.

In EP A with S1, the best DICE coefficient was 92.73 ± 2.4. These results were obtained using the nnUNet SwinTR architecture. However, the difference with respect to the other U-Net architectures tested, such as the state-of-the-art U-MAMBA, is scarce, as also reported in Park et al. (2022) for a different application domain.

In terms of comparison, these results are lower than those reported in Dot et al. (2022), where authors reached DICE values of 96.0 for the upper skull. However, the comparison is not straight because of differences in the target volume segmented. The volume reconstructed in Dot et al. (2022) covers the entire upper skull, including structures that are much easier to identify and segment, significantly biasing up the DICE. The work in Park et al. (2022) could also be used for comparison purposes. In this work, authors obtained DICE scores of 96.5 for the maxilla and 98.4 for the mandible. However, as in the previous case, the problem posed in the referenced work is less challenging because of the presence of large bones that are much easier to segment. Another potential comparison could be established with the results reported in Gillot et al. (2022), where the authors obtained a DICE score of 78.8 for the craneal base. The comparison, once again, is not straight due to significant differences in the target structures.

Regarding generalisation with an external dataset, the models developed have demonstrated robust capabilities with the ED in S2 and S3. These results contrast with those obtained in Park et al. (2022) (also using an external dataset), where the authors report a substantial decrease of 10 absolute points in the DICE score. This is attributed to the ambitious DA and semi-supervised learning strategies followed in this work, which did not significantly improve scores in EP A, but has enhanced the generalisation capabilities of the model. These techniques yielded an average increase of 4.36 for the DICE score.

In summary, the highest DICE coefficients achieved between the predicted mask and the ground truth were 92.73 ± 2.4 testing with S1, and 98.25 ± 0.9 with S2. In order to exemplify graphically the behaviour of the model, Fig. 10 presents two 3D renderings of the OSPS derived from E3.

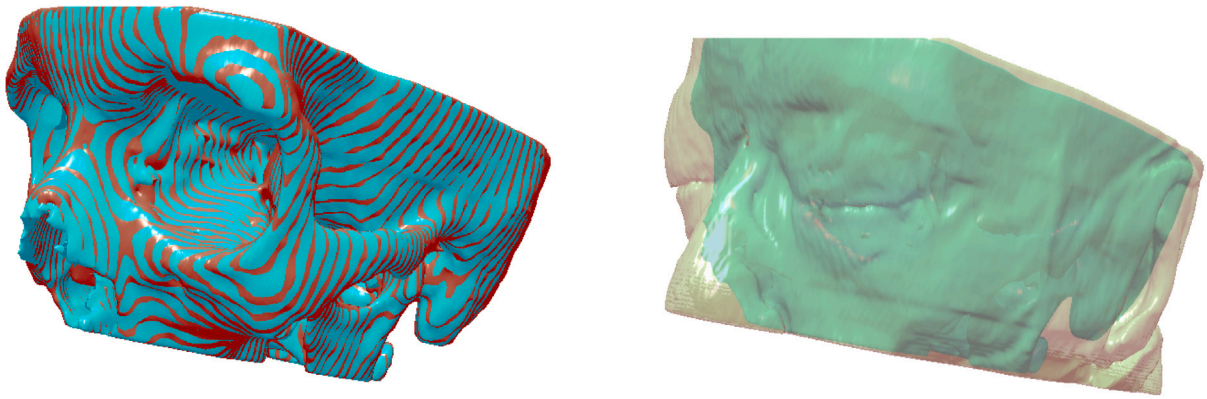

Fig. 10. Example of a 3D reconstruction using the models developed in E3. Left: Ground truth in red, and automatic delineation in blue. Right: Automatic delineation with the contour of the skin overlayed.

Segmentation modelling of the OSPS still has great potential for improvement. The corpus used is limited, so increasing the sample size with more manually annotated masks could potentially enhance the performance and generalisability to other corpora. In any case, new training data could also be generated using a pseudo-labelling strategy similar to that followed in this work.

Although demographic factors can influence model performance, osseous structures present a similar and high contrast for all subjects exhibiting a low inter-subject contrast variability. On the other hand, the segmentation architectures used are specifically developed to identify boundaries, not areas or regions. This suggests that the models developed might generalise effectively across diverse patient cohorts. In any case, despite initial results suggesting no bias due to the scanner model or manufacturer, and sex of the patient, in-depth specific studies should be carried out in this respect, also considering the race or the age of the patient. However, this study would require a much larger dataset than the one available.

It is also crucial to acknowledge that segmentation and validation were performed on healthy CT scans (even when the ID corresponds to ex vivo heads). Therefore altered anatomical structures, particularly those with cranial abnormalities due to pathologies (e.g. tumours) or trauma, may lead to biased outcomes, as these cases are not represented in the training data. In such instances, the model's accuracy could be compromised, highlighting the need for further investigation into these cases to improve the model's robustness.

It is also worth noting that, even when the initial labelling involved a group of up to seven experts, the manual segmentation of the dataset was reviewed by a single expert in the search for better consistency. Thus, expert-related biases, remain a possibility, and models are at risk to reproduce one single segmentation criterion. Including more labelling experts would lead to more variability, which might lead to more generalisable models. In any case, this aspect would need a more in-depth study but also, again, a much larger dataset.

On the other hand, although the current SAM and MedSAM did not perform well for the aimed segmentation task, they remain promising for future research, since they could be used for a transfer learning-based training strategy. But the absence of OSPS samples in the training set used to create these general-purpose segmentation models, might lead to limited results, not achieving those of this research.

In any case, despite the good results obtained, the deployment of the model in a clinical setting would require an interface for user interaction and manual adjustment of potential errors that the automatic algorithm could commit. The model could be integrated into a specific plugin for a general-purpose 3D medical image segmentation tool supporting different medical data formats and providing built-in tools for user interaction (e.g., drawing bounding boxes, adding/removing points, reviewing masks, etc.).

5. Conclusions

Nine different models based on the U-Net architecture were developed for the segmentation of the OSPS from axial CT scans. Although vanilla U-Net variants have the advantage of a simple structure and less training time, other variants, based on SwinTR and U-MAMBA, provided better results. In this regard, the best results were obtained using the U-NET SwinTR, closely followed by the recent U-MAMBA.

The impact of DA techniques and potential improvements due to loss functions were also evaluated. The results show that the wide range of applied DA techniques significantly improve the generalisation of the models. On the other hand, HL provided slightly better results for the problem posed.

The impact of a pseudo-labelling strategy on the generalisation ability of the models was also evaluated. The results obtained show that this technique significantly improves the accuracy of the models developed.

The models were trained and tested using a relatively small dataset of 13 patients. Despite this limitation and the complexity of the task, the results suggest that the models developed perform well even with such small sample size. Improving the model by scaling to a wider dataset is straight, but would require more expensive computational resources, since the best model of those trained required almost two weeks of uninterrupted computing.

Additionally, the SAM and MedSAM, recently released for general-purpose segmentation problems, exhibited poor performance. The limited ability of these models to address this task is attributed to SAM's limitations in recognising fuzzy boundaries and, to the lack of OSPS in MedSAM's training data (Ma et al., 2024a).

For the sake of reproducibility, we have made data and models trained in this work freely available, so they could be used as transfer learning backbones for other bone segmentation applications.

CRediT authorship contribution statement

Yichun Sun: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation. **Alejandro Guerrero-López:** Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Formal analysis. **Julián D. Arias-Londoño:** Writing – review & editing, Supervision, Methodology, Formal analysis, Conceptualization. **Juan I. Godino-Llorente:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Juan Ignacio Godino Llorente reports financial support and article publishing charges were provided by Spanish Agency for Research (Spanish Ministry of Economy and Competitiveness). If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors thank C. Hoyos-Barceló, G. Pérez-de-Arenaza-Pozo and J. C. Puerta-Acevedo for collaborating in the manual annotation of the images. The authors gratefully acknowledge the Universidad Politécnica de Madrid for providing computing resources on the Magerit Supercomputer.

This research was funded by an agreement between Comunidad de Madrid (Consejería de Educación, Universidades, Ciencia y Portavocía) and Universidad Politécnica de Madrid, to finance research actions on SARS-CoV-2 and COVID-19 disease with the REACT-UE resources of the European Regional Development Funds. This work was also supported by the Ministry of Economy and Competitiveness of Spain under Grants PID2021-128469OB-I00 and TED2021-131688B-I00, and by Comunidad de Madrid, Spain. Universidad Politécnica de Madrid supports J. D. Arias-Londoño through a María Zambrano UP2021-035 grant funded by European Union-NextGenerationEU. The authors also thank the Madrid ELLIS unit (European Laboratory for Learning & Intelligent Systems) for its indirect support.

Data availability

Open data.

References

Ahmed, S.M., Mstafa, R.J., 2022. A comprehensive survey on bone segmentation techniques in knee osteoarthritis research: From conventional methods to deep learning. Diagnostics 12 (3), 611.

Arias-Londoño, J.D., Godino-Llorente, J.I., 2024. Analysis of the clever hans effect in COVID-19 detection using chest X-Ray images and Bayesian deep learning. Biomed. Signal Process. Control. 90, 105831.

Arias-Londoño, J.D., Moure-Prado, Á., Godino-Llorente, J.I., 2023. Automatic identification of lung opacities due to COVID-19 from chest X-ray images—Focussing attention on the lungs. Diagnostics 13 (8), 1381.

Bansal, M.A., Sharma, D.R., Kathuria, D.M., 2022. A systematic review on data scarcity problem in deep learning: Solution and applications. ACM Comput. Surv. 54 (10s).

Cellina, M., Gibelli, D., Cappella, A., Toluian, T., Pittino, C.V., Carlo, M., Oliva, G., 2021. Segmentation procedures for the assessment of paranasal sinuses volumes. Neuroradiol. J. 34 (1), 13–20.

Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., Haworth, A., 2021. A review of medical image data augmentation techniques for deep learning applications. J. Med. Imaging Radiat. Oncol. 65 (5), 545–563.

Choi, H., Jeon, K.J., Kim, Y.H., Ha, E.-G., Lee, C., Han, S.-S., 2022. Deep learning-based fully automatic segmentation of the maxillary sinus on cone-beam computed tomographic images. Sci. Rep. 12 (1), 14009.

Ding, A.S., Lu, A., Li, Z., Sahu, M., Galaiya, D., Siewerdsen, J.H., Unberath, M., Taylor, R.H., Creighton, F.X., 2023. A self-configuring deep learning network for segmentation of temporal bone anatomy in cone-beam CT imaging. Otolaryngol.–Head Neck Surg. 169 (4), 988–998.

Dot, G., Schouman, T., Dubois, G., Rouch, P., Gajny, L., 2022. Fully automatic segmentation of craniomaxillofacial CT scans for computer-assisted orthognathic surgery planning using the nnu-net framework. Eur. Radiol. 1–10.

Frid-Adar, M., Ben-Cohen, A., Amer, R., Greenspan, H., 2018. Improving the segmentation of anatomical structures in chest radiographs using U-net with an ImageNet pre-trained encoder. In: Int. MICCAI. pp. 159–168.

Fu, Z., Jin, Z., Zhang, C., He, Z., Zha, Z., Hu, C., Gan, T., Yan, Q., Wang, P., Ye, X., 2021. The future of endoscopic navigation: a review of advanced endoscopic vision technology. IEEE Access 9, 41144–41167.

Gillot, M., Baquero, B., Le, C., Deleat-Besson, R., Bianchi, J., Ruellas, A., Gurgel, M., Yatabe, M., Al Turkestani, N., Najarian, K., et al., 2022. Automatic multi-anatomical skull structure segmentation of cone-beam computed tomography scans using 3D UNETR. PlosOne 17 (10), e0275033.

Gupta, K.K., Jolly, K., Bhamra, N., Osborne, M.S., Ahmed, S.K., 2021. The evolution of sinus surgery in England in the last decade–an observational study. World J. Otorhinolaryngol.- Head Neck Surg. 7 (3), 240–246.

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D., 2021. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI Brainlesion Workshop. Springer, pp. 272–284.

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D., 2022. Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 574–584.

He, Y., Zhang, P., Qi, X., Zhao, B., Li, S., Hu, Y., 2020. Endoscopic path planning in robot-assisted endoscopic nasal surgery. IEEE Access 8, 17039–17048.

He, K., Zhang, X., Ren, S., Sun, J., 2016a. Deep residual learning for image recognition. In: IEEE CVPR. pp. 770–778.

He, K., Zhang, X., Ren, S., Sun, J., 2016b. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Heimann, T., van Ginneken, B., Styner, M., Arzhaeva, Y., Aurich, V., Bauer, C., Beck, A., Becker, C., Beichel, R.R., Bekes, G., Bello, F., Binnig, G.K., Bischof, H., Bornik, A., Cashman, P., Chi, Y., de Cordova, A.S., Dawant, B.M., Fidrich, M., Furst, J.D., Furukawa, D., Grenacher, L., Hornegger, J., Kainmüller, D., Kitney, R.I., Kobatake, H., Lamecker, H., Lange, T., Lee, J., Lennon, B., Li, R., Li, S., Meinzer, H.-P., Németh, G., Raicu, D.S., Rau, A.-M., van Rikxoort, E.M., Rousson, M., Ruskó, L., Saddi, K.A., Schmidt, G., Seghers, D., Shimizu, A., Slagmolen, P., Sorantin, E., Soza, G., Susomboon, R., Waite, J.M., Wimmer, A., Wolf, I., 2009. Comparison and evaluation of methods for liver segmentation from CT datasets. IEEE Trans. Med. Imaging 28 (8), 1251–1265.

Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K., 2020. Nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 18, 203–211.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. Nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 18 (2), 203–211.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., Dollár, P., Girshick, R.B., 2023. Segment anything. In: IEEE/ CVF ICCV. pp. 3992–4003.

Lauretti, L., D'Alessandris, Q.G., Rigante, M., Ricciardi, L., Mattogno, P.P., Olivi, A., 2018. O-arm in endonasal endoscopic cranial base surgery: technical note on initial feasibility. World Neurosurg. 117, 103–108.

Lechien, J.R., Chiesa-Estomba, C.-M., Fakhry, N., Saussez, S., Badr, I., Ayad, T., Chekkoury-Idrissi, Y., Melkane, A.E., Bahgat, A., Crevier-Buchman, L., et al., 2021. Surgical, clinical, and functional outcomes of transoral robotic surgery used in sleep surgery for obstructive sleep apnea syndrome: A systematic review and meta-analysis. Head Neck 43 (7), 2216–2239.

Lee, D.-H., 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on Challenges in Representation Learning. Vol. 3, p. 896.

Lefkovits, S., Laszlo, L., 2022. U-net architecture variants for brain tumor segmentation of histogram corrected images. Acta Univ. Sapientiae Inform. 14, 49–74.

Li, C., Liu, X., Li, W., Wang, C., Liu, H., Yuan, Y., 2024. U-KAN makes strong backbone for medical image segmentation and generation. arXiv preprint arXiv:2406.02918.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: IEEE ICCV. pp. 2980–2988.

Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T.Y., Tegmark, M., 2024. Kan: Kolmogorov-arnold networks. arXiv preprint arXiv:2404.19756.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: IEEE CVPR. pp. 3431–3440.

Lu, X., Cui, Z., Sun, Y., Khor, H.G., Sun, A., Ma, L., Chen, F., Gao, S., Tian, Y., Zhou, F., et al., 2024. Better rough than scarce: Proximal femur fracture segmentation with rough annotations. IEEE Trans. Med. Imaging.

Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B., 2024a. Segment anything in medical images. Nat. Commun. 15 (1).

Ma, J., Li, F., Wang, B., 2024b. U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722.

Martinez-Perez, R., Requena, L.C., Carrau, R.L., Prevedello, D.M., 2021. Modern endoscopic skull base neurosurgery. J. Neurooncol. 151, 461–475.

Mokhasanavisu, V.J.P., Singh, R., Balakrishnan, R., Kadavigere, R., 2019. Ethnic variation of sinonasal anatomy on CT scan and volumetric analysis. Indian J. Otolaryngol. Head Neck Surg. 71, 2157–2164.

Park, S., Kim, H., Shim, E., Hwang, B.-Y., Kim, Y., Lee, J.-W., Seo, H., 2022. Deep learning-based automatic segmentation of mandible and maxilla in multi-center CT images. Appl. Sci. 12 (3), 1358.

Perin, A., Carone, G., Rui, C.B., Raspagliesi, L., Fanizzi, C., Galbiati, T.F., Gambatesa, E., Ayadi, R., Casali, C., Meling, T.R., et al., 2021. The "stars–CT-made" study: advanced rehearsal and intraoperative navigation for skull base tumors. World Neurosurg. 154, e19–e28.

Pirner, S., Tingelhoff, K., Wagner, I., Westphal, R., Rilk, M., Wahl, F.M., Bootz, F., Eichhorn, K.W.G., 2009. CT-based manual segmentation and evaluation of paranasal sinuses. Eur. Arch. Otrhinolaryngol. 266 (4), 507–518.

Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J.B., Zuiderveld, K., 1987. Adaptive histogram equalization and its variations. Comput. Vis. Graph. Image Process. 39 (3), 355–368.

Rao, V.M., El-Noueam, K.I., 1998. Sinonasal imaging: anatomy and pathology. Radiol. Clin. North Am. 36 (5), 921–939.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. arXiv preprint arXiv:1505.04597.

Salehi, S.S.M., Erdogmus, D., Gholipour, A., 2017. Tversky loss function for image segmentation using 3D fully convolutional deep networks. In: Int. Workshop on Machine Learning in Medical Imaging. Springer International Publishing, pp. 379–387.

Sieśkiewicz, A., Łysoń, T., Mariak, Z., Rogowski, M., 2009. Neuronavigation in transnasal endoscopic paranasal sinuses and cranial base surgery: comparison of the optical and electromagnetic systems. Pol. J. Otolaryngol. 63 (3), 256–260.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint abs/1409.1556.

Singh, A., Kumar, R., Thakar, A., Sharma, S., Bhalla, A., 2020. Role of image guided navigation in endoscopic surgery of paranasal sinuses: a comparative study. Indian J. Otolaryngol. Head Neck Surg. 72, 221–227.

Steybe, D., Poxleitner, P., Metzger, M.C., Brandenburg, L.S., Schmelzeisen, R., Bamberg, F., Tran, P.H., Kellner, E., Reisert, M., Russe, M.F., 2022. Automated segmentation of head CT scans for computer-assisted craniomaxillofacial surgery applying a hierarchical patch-based stack of convolutional neural networks. Int. J. Comput. Assist. Radiol. Surg. 17 (11), 2093–2101.

Thatikunta, M., Eaton, J., Nuru, M., Nauta, H.J., 2020. Intraoperative CT for neuronavigation guidance and confirmation of foramen ovale cannulation for glycerol trigeminal rhizotomy: a technical report and case series. Cureus 12 (5).

Tsang, R.K., Chan, W.C., Holsinger, F.C., Chung, J.C., Chow, V.L., Chan, J.Y., Ho, W.-K., Wei, W.I., 2022. Long-term results of robotic-assisted nasopharyngectomy for recurrent nasopharyngeal carcinoma. Head Neck 44 (8), 1940–1947.

Wang, J., Lv, Y., Wang, J., Ma, F., Du, Y., Fan, X., Wang, M., Ke, J., 2021. Fully automated segmentation in temporal bone CT with neural network: a preliminary assessment study. BMC Med. Imaging 21, 1–11.

Wu, W., Yu, Y., Wang, Q., Liu, D., Yuan, X., 2021. Upper airway segmentation based on the attention mechanism of weak feature regions. IEEE Access 9, 95372–95381.

Xu, J., Liu, J., Zhang, D., Zhou, Z., Jiang, X., Zhang, C., Chen, X., 2021. Automatic mandible segmentation from CT image using 3D fully convolutional neural network based on DenseASPP and attention gates. Int. J. Comput. Assist. Radiol. Surg. 16, 1785–1794.

Yang, J., Gu, S., Wei, D., Pfister, H., Ni, B., 2021. Ribseg dataset and strong point cloud baselines for rib segmentation from CT scans. In: Int. Conf. on MICCAI. Vol. Part I, Springer, pp. 611–621.

Yang, J., Yu, L., Wang, L., Tang, Z., Li, Y., 2017. Preoperative planning of a celiac minimally invasive surgery robot based on feature parameters and double collaboration space. Robot 39 (2), 230–238.

Yeung, M., Rundo, L., Nan, Y., Sala, E., Schönlieb, C.-B., Yang, G., 2023. Calibrating the dice loss to handle neural network overconfidence for biomedical image segmentation. J. Digit. Imaging 36 (2), 739–752.

Yeung, M., Sala, E., Schönlieb, C.-B., Rundo, L., 2022. Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. Comput. Med. Imaging Graph. 95, 102026.

Zhao, R., Chen, K., Tang, Y., 2021. Olfactory changes after endoscopic sinus surgery for chronic rhinosinusitis: a meta-analysis. Clin. Otolaryngol. 46 (1), 41–51.

Zuiderveld, K., 1994. Contrast limited adaptive histogram equalization. In: Graphics Gems IV. pp. 474–485.