



Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: <https://www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine>



Automated detection of mandibular landmarks in CT data using a dual-input approach in a two-stage design

Matthias Deitermann^{a,b}, Tobias Pankert^{a,c}, Srikrishna Jaganathan^a, Oliver Röhrle^d, Frank Hölzle^c, Ali Modabber^c, Stefan Raith^{a,c}*

^a Inzipio GmbH, Krantzstraße 7, Aachen, 52070, Germany

^b University of Stuttgart, Keplerstraße 7, Stuttgart, 70174, Germany

^c Department of Oral and Maxillofacial Surgery, University Hospital RWTH, Pauwelsstraße 30, Aachen, 52074, Germany

^d Institute for Modelling and Simulation of Biomechanical Systems, University of Stuttgart, Pfaffenwaldring 5a, Stuttgart, 70569, Germany

ARTICLE INFO

Keywords:

Landmark detection
Mandibular landmarks
Cranio-maxillofacial surgery
Surgical planning
CT
UNet
Deep learning

ABSTRACT

Background and Objective:

Identification of anatomical landmarks in 3D imaging data is an essential step in patient-specific cranio-maxillofacial surgery. Today, precise landmark localization remains largely manual, prone to inter-operator variability, and a bottleneck in streamlined workflows of digitalized preoperative planning, that have in recent years, become a key aspect of cranio-maxillofacial surgery. In clinical practice, bone segmentation and landmark detection in CT imaging is often avoided and automated solutions fall back to the analysis of 2D cephalograms.

Methods:

This work investigates different pipelines to automate the process of landmark localization in the mandible from volumetric CT imaging using convolutional neural networks. As a central element, a 3D U-Net architecture is employed to treat landmark localization and classification like a multi-label segmentation problem. We leverage a two-stage coarse-to-fine approach to tackle heterogeneous input data and preserve high resolution for the final prediction. Our primary innovation is a novel dual-input architecture for the second stage, which uses both the cropped CT data and a mandible segmentation to provide the model with explicit geometric priors for improved accuracy.

The method was developed and tested on a clinical dataset comprising 287 CT datasets to localize nine different landmarks on the human mandible, including the Condyles, Coronoids, Gonions, Pogonion, Gnathion and Menton.

Results:

On a test dataset of 29 CTs, landmarks were predicted with a mean absolute error of 1.40 ± 1.04 while successfully predicting 99.6% of all landmarks.

Conclusion:

The proposed method demonstrates high accuracy, robustness, and speed suggesting strong potential for integration into clinical workflows for automated, patient-specific surgical planning in cranio-maxillofacial surgery.

1. Introduction

The identification of anatomical landmarks in 3D imaging data is a key aspect in surgical planning in patient-specific craniomaxillofacial surgery. Besides its essential importance in preoperative planning and all recent advances in computer-aided planning, the precise localization

of these landmarks today still remains a largely manual task with inter-operator variance and thus a significant bottleneck in a streamlined process-chain of virtual surgical planning.

The cranio-maxillofacial (CMF) region comprising the mouth, jaw, midface, and skull is a challenging region for surgery with its complex geometrical shapes and thus the restoration of physiological masticatory functionality of the jaw is a task requiring expert knowledge

* Corresponding author at: Department of Oral and Maxillofacial Surgery, University Hospital RWTH, Pauwelsstraße 30, Aachen, 52074, Germany.
E-mail address: sraith@ukaachen.de (S. Raith).

<https://doi.org/10.1016/j.cmpb.2025.109113>

Received 11 June 2025; Received in revised form 11 September 2025; Accepted 6 October 2025

Available online 11 October 2025

0169-2607/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

and distinct operative skills. In the past, individual analysis of the mandibular bone (lower jaw) has become key in CMF surgery for appropriate virtual surgical planning. Computer-assisted planning leverages volumetric modeling of patient-specific morphology for virtual surgical planning and benefits from evolving (semi-)automatic image processing technologies [1,2].

CT scans provide high-resolution 3D data, allowing for the localization of anatomical landmarks and segmentation of malformed bones. However, fully automated technologies used in practice are often limited to 2D lateral Cephalograms to analyze individual geometries and measure landmark relations [3,4]. Multiple studies [1, 5–7] have shown how a 3D model of the mandibular bone increases the accuracy of reconstructive surgery and reduces the overall exposure time of the invasive procedure. Additionally, the spatial relations of anatomical landmarks allow for an accurate geometrical analysis as patient morphology is often asymmetrical, which is not encompassed in 2D cephalograms. Hence, clinical practice longs for rapid and fully automatic solutions to achieve precise analysis of the mandibular geometries from volumetric medical imaging.

To reduce time expenditure and individual bias in landmark annotation, knowledge-based semi-automatic, and fully automatic methods have been developed for volumetric imaging in the past [8–11]. However, the CMF bone geometry has been proven to be heavily patient-dependent [12–14] which often limits the capabilities of knowledge-based algorithms [11] as hand-crafted feature extraction often finds its limit in extreme cases and lacks performance on unseen datasets. In the past decade, deep learning (DL) applications have had notable success in localizing landmarks in both soft- and bone tissue due to their ability to encode powerful features from training data [15–18]. The U-Net, a convolutional neural network (CNN) with encoder–decoder architecture has proven to outperform manual feature engineering and other CNN architectures in 2D [19] and 3D [20,21] medical image segmentation tasks. Lately, state-of-the-art publications in mandibular landmark localization mostly use or adapt the U-Net structure [4,17, 22]. Despite the advanced capabilities of CNNs to learn feature representations from training datasets, locating anatomical landmarks in the complex anatomy of the skull still remains a difficult task. Many studies incorporate geometrical constraints into their algorithmic architecture due to the ideally symmetric relations of mandibular landmarks and placements on the bone surface. Evaluating a CT in 3D encapsulates the geometrical relation between all landmarks simultaneously, but increases the computational complexity. Though running their pipeline on 2D slices, Torosdagli et al. [17] achieved excellent accuracy by basing landmark localization on a previously segmented mandibular surface. In 3D [22] found landmarks by analyzing low-resolution global image information to identify sparsely related landmarks first and feeding the information into high-resolution localization of closely related landmarks.

Zhang et al. [23] performed joint segmentation and localization in 3D with high accuracy employing an encoder–decoder CNN architecture twice, first to learn displacement maps for all landmarks and second to perform mandible segmentation and landmark localization simultaneously. However, due to computational feasibility, they had to use a sliding window technique with a maximum resolution of $112 \times 112 \times 112$ voxels.

Landmark prediction accuracy often depends on previous acquisition parameters like the chosen field of view. When reducing the input dimensions to be computationally feasible while handling a large field of view, small targets like landmark features become increasingly difficult to detect. Additionally, the mandibular and maxillary bone structures include many prominent parts leading to a high likelihood of detecting false positives, thus, requiring high amounts of training data to build reliable applications.

In the present work, we apply a 3D U-Net architecture to build a two-stage pipeline able to localize and classify nine different landmarks on the human mandible in 3D CT data with varying field of view. The

anatomical landmarks include in bilateral the *Gonions*, *Coronoids* and *Condyles* and three unilateral landmarks at the chin, with the *Pogonion*, *Gnathion* and *Menton* (Fig. 1).

We incorporated work from [24] and aligned the inference structure along findings from different publications [17,18,22,23,25] addressing the issue of maximizing resolution while keeping global spatial information available. Further, we used a dataset comprising 287 CT scans in conjunction with semi-automatically derived landmark coordinates to train and test our pipeline to quantify the improvements of a two-stage pipeline and when providing mandible segmentation as prior information.

In order to achieve the given task of mandibular landmark detection, we implemented primarily a two-stage pipeline with a coarse-to-fine architecture that has proven its merits in related work [24] to handle large variations in the field of view. The versatility of the coarse-to-fine principle is also demonstrated by its successful application in domains such as 2D cephalometric analysis [26,27] and in architectures prioritizing computational efficiency [28]. Building upon this well-established foundation, our work investigates a novel dual input architecture that considers additional geometry information on the segmented bony geometry and thus could potentially benefit from this guidance to allocate the respective landmarks with superior accuracy. In this architecture, in the second stage, the model uses two distinct inputs: the high-resolution cropped CT image and a pre-existing segmentation of the mandible bone. This dual-input architecture provides the network with explicit anatomical context, bearing the potential of a more robust and accurate performance compared to a standard two-stage pipeline or other contemporary methods that rely on 2D slices or sliding-window techniques. Thus, the distinct innovation of our work is the implementation of this novel dual input pipeline that incorporates additional geometric information to guide the model and thus improve accuracy. The hypothesis of the present work was that the proposed two-stage pipeline is capable of detecting the relevant anatomical landmarks with sufficient precision for potential future integration into largely automated process chains of clinically valid surgical planning. Additionally, it is hypothesized that the additional information provided to the novel dual input approach is improving accuracy of the landmark detection.

2. Material and methods

In this study, we investigated the optimal use of the U-Net architecture to overcome previously named challenges. We aimed for an easily applicable pipeline to localize and identify landmarks efficiently and mainly targeted robust performance regarding three major aspects of our dataset: individual morphology, differing fields of view, and spatial orientation of the head. We trained and verified our approaches by using a large clinical CT dataset of the head and neck region to handle a large variety of real-world data with the same accuracy. We replicated successful implementations of a 3D U-Net for segmentation tasks in medical imaging [20,24] as a baseline model and built a coarse-to-fine pipeline connecting two 3D U-Net models sequentially for increased resolution. Anatomical landmark placements are directly connected to the geometrical shape of the bone, as they lie on their surfaces. Therefore, we provided additional segmentation data of the human mandible for training and inference of a second 3D U-Net in a dual-input architecture. Finally, we compared the achieved performances against each other and against a singular U-Net model.

2.1. Dataset

For training and testing, a base dataset comprising 500 multislice CT scans of the head and neck region was investigated. These scans featured a wide variety in field of view and spatial orientation of the head. The same dataset was previously examined in [14,29]. Data acquisition was completed at the RWTH Aachen University Hospital

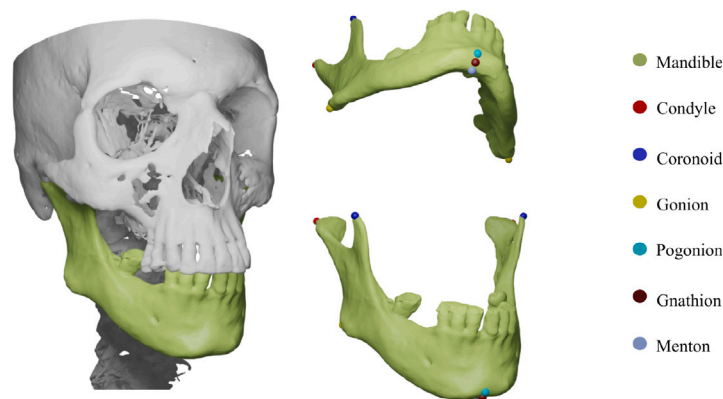


Fig. 1. Mandibular anatomy and landmark locations.

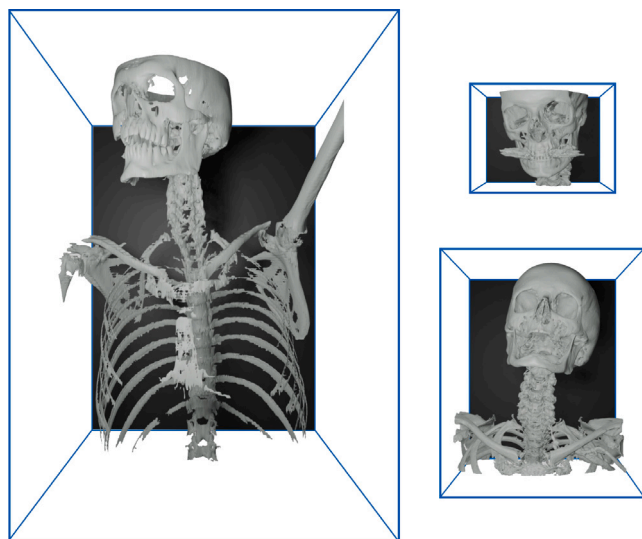


Fig. 2. Threshold segmentations of the largest, median, and smallest field of view from the dataset with original spatial orientation.

using the SOMATOM Definition Flash and SOMATOM Definition AS from Siemens (Erlangen, Germany).

All patients were scanned in a supine position, and the mandible was manually segmented by experts using a semi-automated thresholding approach with Mimics 14 (Materialise Inc., Leuven, Belgium). This party manual annotation represents the necessary, one-time data curation investment required to develop a fully automated pipeline that, once trained, offers rapid, reproducible landmarking without operator dependent variability.

Patients under the age of 20 and those with skeletal mandibular diseases were excluded. After applying further exclusion criteria, that were defined as geometrical flaws of the surface data, incomplete or inaccurate segmentation, and acquisitions with less than 5 mm margin from at least one landmark to the border of the scan, in total 287 CT scans were eligible for training and testing. The scanned volumes showed a wide variance in field of view from $195 \times 195 \times 146.3$ mm to $439 \times 439 \times 600$ mm and the respective voxel sizes ranging from $0.38 \times 0.38 \times 0.7$ mm to $0.86 \times 0.86 \times 4.0$ mm (Fig. 2). For the ground truth annotation of mandibular landmarks, we improved the surface-based algorithm from [14]. This algorithm uses the triangulated surface geometry of the mandible as an input and uses a cascade of geometric operations to align the mandible in space. First it uses a bounding sphere to access the mandibular center and salient points at the condyles and the chin. Using these points for a rough alignment, a subsequent identification of landmarks

based on unique geometric attributes becomes feasible. We refined the positioning of the *Pogonion* by identifying the *mandibular Symphysis* and used the new information to add the clinically relevant landmarks *Gnathion* and *Menton*. This specific set of skeletal landmarks was chosen to support our primary clinical application: the automation of virtual surgical planning for mandibular reconstruction [1], a task with distinct data requirements compared to other 3D landmarking studies focused on dental anatomy [30].

2.2. Implementation of the 3D U-Net

The U-Net CNN architecture was introduced by Ronneberger et al. [19], leveraging an encoder–decoder structure while copying feature maps on each convolutional level to perform accurate semantic segmentation in 2D medical imaging. We employed a 3D version of the U-Net based on fundamental works by Çiçek et al. [20] and Isensee et al. [21] to identify mandibular landmarks in 3D CT datasets, utilizing the abilities of the U-Net and preserving spatial 3D information for inference. We built the 3D U-Net model with an input resolution of $[144 \times 144 \times 144]$ while training on a batch size of 2 and an initial learning rate of $5 \exp -5$. The Adam optimizer [31] was applied to optimize the network on the Tversky loss [32]. Furthermore, we used Batch Normalization [33] for each layer.

2.3. Preprocessing and annotation

For the dataset, we used a randomized 8:1:1 split for training (230 samples), validation (28 samples), and testing (29 samples). An identical split was used for all training runs. Each CT dataset was preprocessed by clipping the intensity values between $[-1024, 3071]$ HU to reduce the impact of artificial material and subsequently normalized between $[0, 1]$. In training, we applied the augmentation methods blur, noise, brightness, and scaling from the batchgenerators library [34].

A binary sphere annotation for each landmark with a radius of 5 mm for the first-stage model and 3 mm for the second-stage model was chosen, based on the results of preliminary investigations. This annotation style was used to mask all training and testing data.

Thus, the U-Net models are trained to predict sphere-like segmentations around each landmark of interest. We used these voxel-wise predictions to calculate the locations by identifying their centroid.

2.4. Full pipeline setup

Recent research suggests superior results when cropping the inputs around the ROI [18,22,23]. Hence, we implemented a coarse-to-fine approach similar to the approach in [24] for the segmentation of the whole mandible, by training two models and using them in a sequence. While end-to-end, cascaded models have shown great success for high-resolution 2D cephalometric analysis [26], our sequential design is an

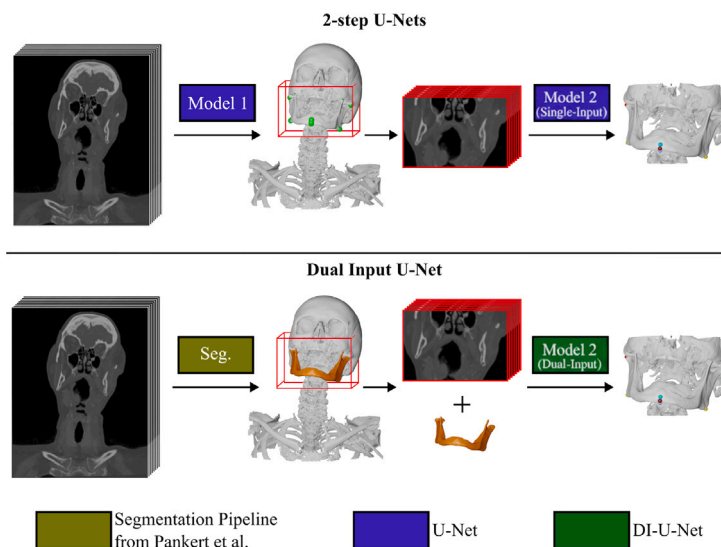


Fig. 3. Two-stage and dual-input pipelines. From left to right: CT input, identifying ROI by bounding box or segmentation input, cropped input for second-stage model, final landmark digitization.

adaptation of the established coarse-to-fine principle for the distinct challenges of 3D volumetric data.

The initial prediction from the first-stage model is used to create a bounding box around the ROI by cropping the original input to the boundaries of the first prediction. With a padding of 5 mm around the bounding box the cropped input is passed to the second-stage model. The second-stage model operates at a higher relative resolution compared to the first-stage model while both models have the same input and output dimensions, the first-stage model processes a significantly larger field of view.

We compared the first-stage model accuracy versus the two-stage pipeline accuracy. Additionally, we investigated a dual-input localization pipeline using mandible segmentation as additional information. For the training of the dual-input model, we used ground-truth volumetric annotations of the mandible bone for the second input scaled to the same dimensions as the CT data input and trained it only on data cropped around the ROI. For the dual-input pipeline's inference, we employed the robust mandible segmentation model described by Pankert et al. [24]. This model was previously validated and reported to achieve high segmentation accuracy, with a mean Dice score of 94.82%. The resulting segmentation mask was used to define the cropped ROI and to serve as the second input channel for the dual-input landmark localization model. Fig. 3 provides a schematic overview of the inference process for both pipelines. Fig. 3 gives an overview of the inference for both pipelines.

2.5. Implementation

Training and testing were executed on the same machine running Ubuntu 22.04 (RAM: 128 GB, GPU: 2x NVIDIA Quadro RTX 5000, CPU: Intel Xeon(R) Gold 5122) using Python 3.9 with Blender 2.93 for the annotation algorithm and Tensorflow 2.10 [35] for training and inference.

The end-to-end runtime for the Two-Step segmentation was on average 14.2 s per CT scan with a peak RAM consumption of 1.6 GB and a peak VRAM consumption of 13.4 GB. For the Dual-Input-Pipeline, the average run time was 15.2 s, also with a peak RAM consumption of 1.6 GB and a peak VRAM consumption of 15.0 GB. These times include initialization steps for TensorFlow and all networks, as well as preprocessing of the CT scan in DICOM format and post-processing. Despite using on more model pass step, the Dual-Input pipeline is only slightly slower than the Two-Step pipeline due to the comparably large overhead of initialization and I/O.

2.6. Evaluation metrics

The localization error was measured with the Euclidean distance d between the ground-truth position and the predicted position of each landmark defined by its centroid in x , y and z .

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

The intersection over union (IoU), also known as the Jaccard coefficient, compares the similarity of two objects and is used in this study to describe the similarity of the bounding box around all predicted landmarks against the bounding box around the ground-truth. In 3D the IoU is calculated as the volumetric intersection of two objects divided by the combined volume.

3. Results

All experiments were repeated three times with different random seeds. Each run was tested on 29 CTs with 9 landmarks each for a total of 261 landmarks per run. All three presented methods consistently identified over 99% of landmarks. Table 1 compares the measured mean absolute error (MAE). The first-stage model predicted locations for all nine landmarks with a MAE of 2.12 mm. This first-stage prediction was used to crop the original image, achieving a IoU of the cropped image patch of $93.38 \pm 2.94\%$ to the ground-truth patch. Employing the full two-stage pipeline reduces the MAE to 1.70 mm. The improvements are significant across all landmark types. Merely the Condyles showed a slightly higher standard deviation (SD). The dual-input two-stage pipeline outperformed the other pipelines in the MAE for each landmark type and achieved an overall MAE of 1.40 mm (Fig. 4).

Table 2 shows the percentages of predictions within the error ranges of 1, 2, 3, 4 and 5 mm. With the first-stage model 78.2% of predictions were performed within a 3 mm error. In comparison, the two-stage pipeline recorded 88.1% in the same error range. The dual-input pipeline outperformed the two-stage pipeline in all error limits and detected 93.7% of all predictions within the 3 mm error range. The two-stage pipeline recorded superior results for each error limit up to 4 mm against the first-stage model but performed slightly worse in the 5 mm confidence interval. Especially in the 1 mm interval the dual-input pipeline showed superior performance with 44.0% detected landmarks and proved overall to be the superior setup.

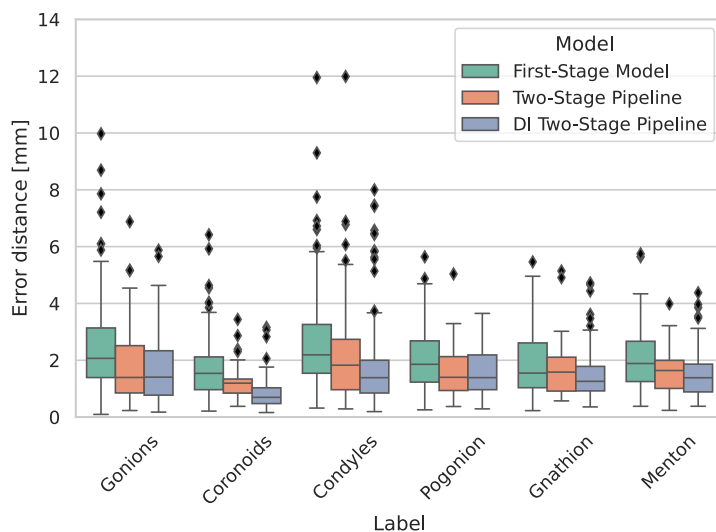


Fig. 4. Error distances per bone for the first-stage model, two-stage pipeline and dual-input two-stage pipeline.

Table 1

Overall and label-wise mean absolute error metric for all three setups on the test dataset. False negatives describe the sensitivity of each method.

	MAE \pm SD [mm]		
	First-stage	Two-stage	DI Two-stage
Overall	2.12 \pm 1.33	1.70 \pm 1.23	1.40 \pm 1.04
Gonions	2.41 \pm 1.58	1.89 \pm 1.47	1.68 \pm 1.14
Coronoids	1.70 \pm 1.01	1.20 \pm 0.52	0.80 \pm 0.49
Condyles	2.64 \pm 1.69	2.28 \pm 2.01	1.72 \pm 1.42
Pogonion	2.05 \pm 1.14	1.71 \pm 1.02	1.54 \pm 0.76
Gnathion	1.90 \pm 1.20	1.70 \pm 1.10	1.51 \pm 0.95
Menton	2.02 \pm 1.11	1.57 \pm 0.83	1.53 \pm 0.86
FNs	2 (0.76%)	1 (0.38%)	1 (0.38%)

Table 2

The confidence intervals describe the percentage of errors within the threshold distance. The thresholds are 1, 2, 3, 4 and 5 mm.

	Confidence intervals [%]				
	1 mm	2 mm	3 mm	4 mm	5 mm
First-stage	20.3	56.7	78.2	90.4	95.8
Two-stage	29.9	72.0	88.1	92.3	95.0
DI Two-stage	44.0	77.8	93.7	96.8	98.0

Fig. 5 shows the error distances from the dual-input pipeline. Breaking down the error distances in singular axes showed consistencies for all landmark areas. Four unique patterns stood out: The *Gonions* showed the largest error in posterior–anterior, a slightly smaller error in superior–inferior, and close to zero error in left–right. The *Condyles* showed a similar pattern with more equal error sizes in posterior–anterior and superior–inferior while the *Coronoids* showed the most equally distributed errors in all axes. However, a median error in lateral direction greater than zero on the left side and a median error smaller than zero on the right side showed the predicted *Coronoids* tended to lay more lateral than the ground truth. In contrast, all three landmarks on the *mandibular Symphysis* (*Pogonion*, *Menton*, and *Gnathion*) showed the largest error in lateral direction with small errors in the two other axes. With the median error below zero, all three showed a bias towards the left side of the patient.

4. Discussion

In this work, we propose a coarse-to-fine progression pipeline utilizing two 3D U-Net models to localize nine landmarks on the human

mandible and investigate implementations in three different variants. The two-stage and dual-input pipelines both showed to be sufficiently accurate and sensitive. Leveraging the sequential use of two separate models they can identify landmarks in CT with a small and large field of view. Transferring the identified spatial information gained from the first-stage model to the second-stage model ensured high-resolution predictions while preserving the entire mandible structure. Due to the geometrical distribution of all targeted landmarks, it is likely that even when the first-stage model missed a landmark, it is still encompassed in the subsequent cropping making the approach very robust.

4.1. Performance

The dual-input pipeline using additional segmentation data showed overall superior accuracy over the two-stage pipeline and achieved consistent MAEs for all landmark types between 1.51 and 1.72 mm. Only the *Coronoids* were detected far more accurately with an MAE of 0.80 mm which may be attributed to the salient geometry of most *Coronoids* facilitating their localization and can also be seen in comparable data by Torosdagli et al. [17] or Palazzo et al. [36].

Overall, the dual-input pipeline detected 93.7% of all landmarks with an error under 3 mm and 77.8% under 2 mm, which is superior or at least competitive to most literature. The multi-atlas method from [8] registered 63.57% of landmarks in the 3 mm confidence interval. O’Neil et al. [37] documented 90% of landmarks under a 4 mm error with their CNN-Atlas combination and Lu et al. [4] described 96.83% of all landmarks inside a 3 mm error range using their U-Net-like pipeline. *Condyles* and *Gonions* were the most challenging landmarks for our method with the largest singular error of 8 mm and 5.9 mm respectively. Torosdagli et al. [17] achieved significantly superior accuracy with their pseudo-3D approach recording an MAE under 1 mm for most landmarks but showed only a slight advantage for the *Pogonion* (1.36 mm MAE). However, in a more recent publication [38] their fully 3D approach improved the localization of the *Gnathion* and *Pogonion* while recording larger MAEs for *Coronoids* and *Condyles* than our pipelines. Both their publications do not localize the *Gonions*.

Breaking down the error distances in singular axes gave insights into the placements of each landmark type. The three landmarks on the chin (*Pogonion*, *Gnathion*, and *Menton*) were in most cases aligned on a line close to the actual *mandibular symphysis*. If displaced, all three landmarks showed consistent displacements majorly in the lateral direction. Similarly, the *Gonions* showed displacements mostly along the curvature of the bone in posterior–anterior and superior–inferior with negligible displacements in lateral. We argue that this may be attributed to the rather undefined positioning of the *Gonions* along the mandibular bone in cases with large mandibular angles.

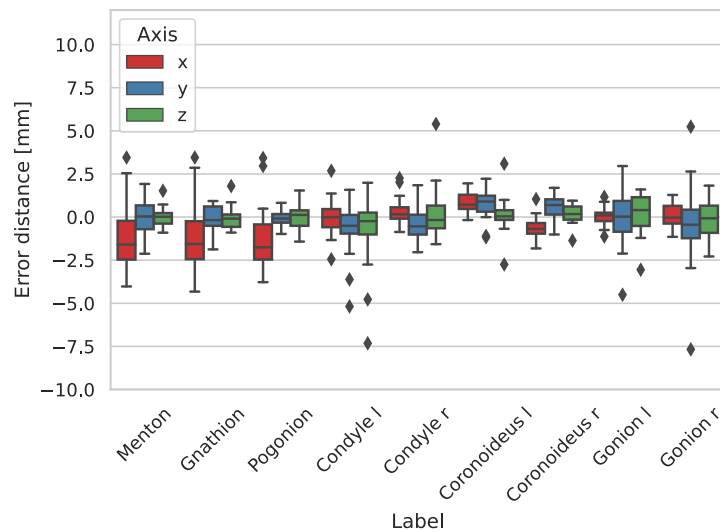


Fig. 5. Error distances for dual-input two-stage pipeline in lateral direction (x), anterior–posterior direction (y) and cranio-caudal direction (z).

4.2. Pipeline architecture

Overall, both presented coarse-to-fine approaches validate that two sequential U-Nets may provide a decent resolution while encompassing the whole mandibular geometry to perform highly accurate landmark localization no matter their proximity to each other. In contrast, other major publications employing modified encoder–decoder architectures as a base required additional steps to ensure high-resolution predictions. Zhang et al. [23] used a sliding window approach and Lian et al. [22] performed one dense second-step prediction per landmark. In another example, Torosdagli et al. [17] employed a long short-term memory (LSTM) network running on a single 2D slice in the sagittal plane for the identification of closely related landmarks. This procedure requires a highly accurate segmentation of the mandible for the consecutive prediction of landmarks along the *mandibular symphysis* where an error of the surface shape propagates directly into the localization of landmarks. Our dual-input approach, in contrast, provides a more robust structure to handle less accurate segmentations since predictions are calculated on both the input CT and the segmented mandible.

By exposing the network to the given real-world variability during training on our dataset, the model learned to be invariant to orientation implicitly, without requiring an explicit registration or data re-orientation step. This demonstrates the model’s capability and robustness to deal with the heterogeneous conditions of a real clinical workflow that are not curated towards specific fields of view or global spatial orientations.

4.3. Limitations and improvements

We successfully trained a robust and accurate deep-learning pipeline based on landmark annotations from a surface-based algorithm. In doing so we could leverage a validated algorithm in combination with a large clinical dataset to efficiently and consistently create training data. However, the algorithm may still insert bias into the learned locations. Exemplary, all *Condyle* landmarks were placed lateral-superior on the *Condyle* process following the convention of Raith et al. [14], this contrasts most of the literature placing the *Condyle* landmark in superior on top of the *Condyle* process [4,10,17].

One limitation is that the current model was developed using a non-pathological cohort. As explicitly defined by our exclusion criteria, the training and validation datasets intentionally omitted cases with skeletal mandibular diseases, such as significant deformities or fractures resulting from trauma or surgery. This deliberate focus allowed us to first establish a highly accurate and robust baseline for the core

methodology in a controlled setting. Consequently, the model’s performance on complex pathological anatomy is untested. Extending and validating the pipeline on these challenging cases is a critical next step to enhance its clinical utility and will be in focus of our future research.

While our results are contextualized with related literature, a direct quantitative benchmark against many state-of-the-art methods is challenging due to the specific clinical application of our work. The unique set of nine skeletal landmarks targeted for mandibular reconstruction differs from those in other studies, which may focus on different clinical tasks or utilize 2D imaging modalities [26,27]. Consequently, a direct comparison would not be scientifically appropriate. The future establishment of public benchmark datasets for this specific surgical task will be crucial and should be investigated as the field matures.

A potential concern with multi-stage pipelines is the risk of error propagation, where inaccuracies from an early stage could negatively impact subsequent predictions. However, our results demonstrate that the trade-off is highly favorable. The first-stage model proved reliable in localizing the mandible. This high fidelity ensures the second-stage model consistently receives a well-defined region of interest, mitigating the risk of error accumulation. The overall performance further validates this design choice, as both the two-stage and dual-input two-stage pipelines significantly outperformed the single-stage model. This marked improvement confirms that the precision gained from the high-resolution analysis in the second stage substantially outweighs any issue for propagated error in the investigated application.

Furthermore, compact training and postprocessing steps provide high flexibility for further adaptation to additional clinical requirements e.g. the inclusion of other landmarks. The employed surface-based algorithm for the ground-truth annotation offers a great opportunity for time-efficient adaptations in the same sense. In the future, the main improvements will focus on pushing the accuracy and sensitivity by optimizing aspects of the training e.g. using a Euclidean distance loss function and quantifying the robustness of the dual-input pipeline when dealing with flawed segmentations of the mandibular bone surface.

5. Conclusion

The proposed dual-input pipeline can efficiently replace manual digitization of landmarks by leveraging a coarse-to-fine setup starting with a mandible segmentation model to localize the *Gonions*, *Coronoids*, *Condyles*, *Pogonion*, *Gnathion*, and *Menton* from CT data. The fully automatic method demonstrates robustness with regards to variations in morphology, field of view and spatial orientations of the head. Due to its high accuracy and high sensitivity as well as its fast prediction times (less than a minute), the proposed method could be a valuable asset for patient-specific planning in the clinical context.

CRedit authorship contribution statement

Matthias Deitermann: Writing – original draft, Visualization, Validation, Software, Investigation, Data curation. **Tobias Pankert:** Writing – review & editing, Validation, Software, Formal analysis, Conceptualization. **Srikrishna Jaganathan:** Writing – review & editing, Visualization, Software, Investigation. **Oliver Röhrle:** Writing – review & editing, Supervision, Project administration, Methodology. **Frank Hölzle:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Ali Modabber:** Writing – review & editing, Supervision, Resources, Investigation. **Stefan Raith:** Writing – original draft, Visualization, Methodology, Data curation, Conceptualization.

Ethics statement

This study was conducted using retrospective, anonymized CT imaging data from clinical archives on CT data that was previously used in preceding studies with different scientific focus. No direct interaction with human subjects or interventions were performed. All data were handled in compliance with applicable data protection and privacy laws.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Stefan Raith reports a relationship with Inzipio GmbH that includes: employment and equity or stocks. Tobias Pankert reports a relationship with Inzipio GmbH that includes: employment and equity or stocks. Ali Modabber reports a relationship with Inzipio GmbH that includes: equity or stocks. Srikrishna Jaganathan reports a relationship with Inzipio GmbH that includes: employment. 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.

References

- [1] A. Modabber, A. Rauen, N. Ayoub, S. Möhlhenrich, F. Peters, K. Kniha, F. Hölzle, S. Raith, Evaluation of a novel algorithm for automating virtual surgical planning in mandibular reconstruction using fibula flaps, *J. Cranio-Maxillofacial Surg.* 47 (9) (2019) 1378–1386, <http://dx.doi.org/10.1016/j.jcms.2019.06.013>.
- [2] F. Probst, P.G. Liokatis, G. Mast, M. Ehrenfeld, Virtual planning for mandible resection and reconstruction, *Innov. Surg. Sci.* 8 (3) (2023) 137–148, <http://dx.doi.org/10.1515/iss-2021-0045>.
- [3] Z. Zhong, J. Li, Z. Zhang, Z. Jiao, X. Gao, An Attention-Guided Deep Regression Model for Landmark Detection in Cephalograms, *Medical Image Computing and Computer Assisted*, 2019, pp. 540–548, http://dx.doi.org/10.1007/978-3-030-32226-7_60.
- [4] G. Lu, H. Shu, H. Bao, Y. Kong, C. Zhang, B. Yan, Y. Zhang, J.-L. Coatrieux, CMF-Net: craniomaxillofacial landmark localization on CBCT images using geometric constraint and transformer, *Phys. Med. Biol.* 68 (2023) 95020, <http://dx.doi.org/10.1088/1361-6560/acb483>.
- [5] A. Modabber, M. Gerressen, M. Stiller, N. Noroozi, A. Füglein, F. Hölzle, D. Riediger, A. Ghassemi, Computer-assisted mandibular reconstruction with vascularized iliac crest bone graft, *Aesthetic Plast. Surg.* 36 (2012) 653–659, <http://dx.doi.org/10.1007/s00266-012-9877-2>.
- [6] A. Modabber, C. Legros, M. Rana, M. Gerressen, D. Riediger, A. Ghassemi, Evaluation of computer-assisted jaw reconstruction with free vascularized fibular flap compared to conventional surgery: a clinical pilot study, *Int. J. Med. Robot.* 8 (2012) 215–220, <http://dx.doi.org/10.1002/rcs.456>.
- [7] D.S. Shenaq, E. Matros, Virtual planning and navigational technology in reconstructive surgery, *J. Surg. Oncol.* 118 (2018) 845–852, <http://dx.doi.org/10.1002/jso.25255>.
- [8] S. Shahidi, E. Bahrapour, E. Soltanimehr, A. Zamani, M. Oshagh, M. Moattari, A. Mehdizadeh, The accuracy of a designed software for automated localization of craniofacial landmarks on CBCT images, *BMC Med. Imaging* 14 (2014) 32, <http://dx.doi.org/10.1186/1471-2342-14-32>.
- [9] M. Mestiri, H. Kamel, Reeb graph for automatic 3D cephalometry, *Int. J. Image Process.* 8 (2014) 17–29.

- [10] A. Gupta, O. Kharbanda, V. Sardana, R. Balachandran, H. Sardana, A knowledge-based algorithm for automatic detection of cephalometric landmarks on CBCT images, *Int. J. Comput. Assist. Radiol. Surg.* 10 (2015) <http://dx.doi.org/10.1007/s11548-015-1173-6>.
- [11] M. Ed-Dahraouy, H. Riri, M. Ezzahmouly, A. Elmoutaouakkil, F. Bourzgui, H. Aghoutan, Proposition of local automatic algorithm for landmark detection in 3D cephalometry, *Bull. Electr. Eng. Inform.* 10 (2021) <http://dx.doi.org/10.11591/eei.v10i5.3142>, URL <https://api.semanticscholar.org/CorpusID:228873578>.
- [12] J.J. Ferreira, C.M. Zagalo, M.L. Oliveira, A.M. Correia, A.R. Reis, Mandible reconstruction: History, state of the art and persistent problems, *Prosthet. Orthot. Int.* 39 (2015) 182–189, <http://dx.doi.org/10.1177/0309364613520032>.
- [13] W.K. Darkwah, A. Kadri, B. Buanya, G. Aidoo, Cephalometric study of the relationship between facial morphology and ethnicity: Review article, *Transl. Res.* 12 (2018) <http://dx.doi.org/10.1016/j.tria.2018.07.001>.
- [14] S. Raith, V. Varga, T. Steiner, F. Hölzle, H. Fischer, Computational geometry assessment for morphometric analysis of the mandible, *Comput. Methods Biomech. Biomed. Eng.* 20 (2016) 1–8, <http://dx.doi.org/10.1080/10255842.2016.1196196>.
- [15] Y. Zheng, D. Liu, B. Georgescu, H. Nguyen, D. Comaniciu, 3D deep learning for efficient and robust landmark detection in volumetric data, in: *Medical Image Computing and Computer-Assisted Intervention*, in: *Lecture Notes in Computer Science*, vol. 9350, Springer, 2015, pp. 565–572, http://dx.doi.org/10.1007/978-3-319-24553-9_69.
- [16] C. Payer, D. Stern, H. Bischof, M. Urschler, Regressing heatmaps for multiple landmark localization using cnns., in: S. Ourselin, L. Joskowicz, M.R. Sabuncu, G.B. Unal, W. Wells (Eds.), *MICCAI* (2), in: *Lecture Notes in Computer Science*, vol. 9901, 2016, pp. 230–238, http://dx.doi.org/10.1007/978-3-319-46723-8_27.
- [17] N. Torosdagli, D.K. Liberton, P. Verma, M. Sincan, J.S. Lee, U. Bagcı, Deep geodesic learning for segmentation and anatomical landmarking., *IEEE Trans. Med. Imaging* 38 (4) (2019) 919–931, <http://dx.doi.org/10.1109/TMI.2018.2875814>.
- [18] Q. Liu, H. Deng, C. Lian, X. Chen, D. Xiao, L. Ma, X. Chen, T. Kuang, J. Gateno, P.-T. Yap, J.J. Xia, SkullEngine: A multi-stage CNN framework for collaborative CBCT image segmentation and landmark detection., 2021, *CoRR*, [arXiv:2110.03828](https://arxiv.org/abs/2110.03828), URL <https://arxiv.org/abs/2110.03828>.
- [19] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III* 18, in: *Lecture Notes in Computer Science*, 9351, Springer, 2015, pp. 234–241, http://dx.doi.org/10.1007/978-3-319-24574-4_28.
- [20] Ö. Çiçek, A. Abdulkadir, S.S. Lienkamp, T. Brox, O. Ronneberger, 3D U-net: learning dense volumetric segmentation from sparse annotation, in: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17–21, 2016, Proceedings, Part II* 19, in: *Lecture Notes in Computer Science*, 9901, Springer, 2016, pp. 424–432, http://dx.doi.org/10.1007/978-3-319-46723-8_49.
- [21] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, K.H. Maier-Hein, Brain tumor segmentation and radiomics survival prediction: Contribution to the BRATS 2017 challenge., in: A. Crimi, S. Bakas, H.J. Kuijf, B.H. Menze, M. Reyes (Eds.), *BrainLes@MICCAI*, in: *Lecture Notes in Computer Science*, Vol. 10670, Springer, 2017, pp. 287–297, http://dx.doi.org/10.1007/978-3-319-75238-9_25.
- [22] C. Lian, F. Wang, H.-H. Deng, L. Wang, D. Xiao, T. Kuang, H.-Y. Lin, J. Gateno, S.G.F. Shen, P.-T. Yap, J.J. Xia, D. Shen, Multi-task dynamic transformer network for concurrent bone segmentation and large-scale landmark localization with dental CBCT, *Med. Image Comput. Comput. Assist. Interv.* 12264 (2020) 807–816, http://dx.doi.org/10.1007/978-3-030-59719-1_78.
- [23] J. Zhang, M. Liu, L. Wang, S. Chen, P. Yuan, J. Li, S.G.F. Shen, Z. Tang, K.C. Chen, J.J. Xia, D. Shen, Context-guided fully convolutional networks for joint craniomaxillofacial bone segmentation and landmark digitization, *Med. Image Anal.* 60 (2020) 101621, <http://dx.doi.org/10.1016/J.MEDIA.2019.101621>.
- [24] T. Pankert, H. Lee, F. Peters, F. Hölzle, A. Modabber, S. Raith, Mandible segmentation from CT data for virtual surgical planning using an augmented two-stepped convolutional neural network, *Int. J. Comput. Assist. Radiol. Surg.* 18 (2023) 1479–1488, <http://dx.doi.org/10.1007/s11548-022-02830-w>.
- [25] J. Zhang, Y. Gao, Y. Gao, B.C. Munsell, D. Shen, Detecting anatomical landmarks for fast alzheimer's disease diagnosis, *IEEE Trans. Med. Imaging* 35 (2016) 2524–2533, <http://dx.doi.org/10.1109/TMI.2016.2582386>.
- [26] T. He, J. Guo, W. Tang, W. Zeng, P. He, F. Zeng, Z. Yi, Cascade-refine model for cephalometric landmark detection in high-resolution orthodontic images, *Knowl.-Based Syst.* 265 (2023) 110332, <http://dx.doi.org/10.1016/j.knsys.2023.110332>.
- [27] T. He, J. Yao, W. Tian, Z. Yi, W. Tang, J. Guo, Cephalometric landmark detection by considering translational invariance in the two-stage framework, *Neurocomputing* 464 (2021) 15–26, <http://dx.doi.org/10.1016/j.neucom.2021.08.042>.

- [28] L. Cui, B. Liu, G. Xu, J. Guo, W. Tang, T. He, A pseudo-3D coarse-to-fine architecture for 3D medical landmark detection, *Neurocomputing* 614 (2025) 128782, <http://dx.doi.org/10.1016/j.neucom.2024.128782>.
- [29] V. Varga, S. Raith, C. Loberg, A. Modabber, A. Bartella, F. Hölzle, H. Fischer, T. Steiner, Classification of the level of mandibular atrophy– a computer-assisted study based on 500 CT scans, *J. Cranio-Maxillofacial Surg.* 45 (2017) <http://dx.doi.org/10.1016/j.jcms.2017.09.014>.
- [30] T. He, G. Xu, L. Cui, W. Tang, J. Long, J. Guo, Anchor ball regression model for large-scale 3D skull landmark detection, *Neurocomputing* 567 (2024) 127051, <http://dx.doi.org/10.1016/j.neucom.2023.127051>.
- [31] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, <http://dx.doi.org/10.48550/ARXIV.1412.6980>, URL <https://arxiv.org/abs/1412.6980>.
- [32] S.S.M. Salehi, D. Erdogmus, A. Gholipour, Tversky loss function for image segmentation using 3D fully convolutional deep networks, in: *International Workshop on Machine Learning in Medical Imaging*, Springer, 2017, pp. 379–387.
- [33] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015, CoRR, [arXiv:1502.03167](https://arxiv.org/abs/1502.03167), [arXiv:1502.03167](https://arxiv.org/abs/1502.03167), URL <http://arxiv.org/abs/1502.03167>.
- [34] F. Isensee, P. Jäger, J. Wasserthal, D. Zimmerer, J. Petersen, S. Kohl, J. Scheck, A. Klein, T. Roß, S. Wirkert, P. Neher, S. Dinkelacker, G. Köhler, K. Maier-Hein, Batchgenerators - A python framework for data augmentation, 2020, <http://dx.doi.org/10.5281/zenodo.3632567>, URL <https://doi.org/10.5281/zenodo.3632567>.
- [35] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, TensorFlow: Large-scale machine learning on heterogeneous systems, 2015, URL <https://www.tensorflow.org/>, Software available from tensorflow.org.
- [36] S. Palazzo, G. Bellitto, L. Prezzavento, F. Rundo, U. Bagci, D. Giordano, R. Leonardi, C. Spampinato, Deep multi-stage model for automated landmarking of craniomaxillofacial CT scans, in: 2020 25th International Conference on Pattern Recognition, ICPR, 2021, pp. 9982–9987, <http://dx.doi.org/10.1109/ICPR48806.2021.9412910>.
- [37] A.Q. O’Neil, A. Kascenas, J. Henry, D. Wyeth, M. Shepherd, E. Beveridge, L. Clunie, C. Sansom, E. Šeduikyt, K. Muir, I. Poole, Attaining human-level performance with atlas location autocontext for anatomical landmark detection in 3D CT data, in: *Computer Vision – ECCV 2018 Workshops: Munich, Germany, September 8–14, 2018, Proceedings, Part III*, Springer-Verlag, Berlin, Heidelberg, 2019, pp. 470–484, http://dx.doi.org/10.1007/978-3-030-11015-4_34.
- [38] N. Torosdagli, S. Anwar, P. Verma, D. Liberton, J. Lee, W. Han, Relational reasoning network for anatomical landmarking, *J. Med. Imaging* 10 (2023) <http://dx.doi.org/10.1117/1.JMI.10.2.024002>.