



# nnLandmark: A Self-Configuring Method for 3D Medical Landmark Detection


Alexandra Ertl<sup>1,2,3</sup> 

ALEXANDRA.ERTL@DKFZ-HEIDELBERG.DE

Stefan Denner<sup>1,4</sup> 

Robin Peretzke<sup>1,2,3</sup> 


Shuhan Xiao<sup>1,4</sup> 


David Zimmerer<sup>1</sup> 


Maximilian Fischer<sup>1,2</sup> 

Markus Bujotzek<sup>1,2</sup> 

Xin Yang<sup>6</sup> 

Peter Neher<sup>1,3,5</sup> 

Fabian Isensee<sup>\*1,7</sup> 

Klaus H. Maier-Hein<sup>\*1,2,3,4,5,7,8,9</sup> 

<sup>1</sup> German Cancer Research Center (DKFZ) Heidelberg, Division of Medical Image Computing, Heidelberg, Germany

<sup>2</sup> Medical Faculty Heidelberg, Heidelberg University, Heidelberg, Germany

<sup>3</sup> Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg University Hospital, Heidelberg, Germany

<sup>4</sup> Faculty of Mathematics and Computer Science, Heidelberg University, Heidelberg, Germany

<sup>5</sup> German Cancer Consortium (DKTK), DKFZ, core center Heidelberg, Germany

<sup>6</sup> School of Biomedical Engineering, Shenzhen University Medical School, Shenzhen University, Shenzhen, Guangdong, China

<sup>7</sup> Helmholtz Imaging, DKFZ, Heidelberg, Germany

<sup>8</sup> National Center for Tumor Diseases (NCT), NCT Heidelberg, A Partnership Between DKFZ and The University Medical Center Heidelberg, Heidelberg, Germany

<sup>9</sup> HIDSS4Health, Heidelberg, Germany

**Editors:** Accepted for publication at MIDL 2026

## Abstract

Landmark detection is central to many medical applications, such as identifying critical structures for treatment planning or defining control points for biometric measurements. However, manual annotation is labor-intensive and requires expert anatomical knowledge. While deep learning shows promise in automating this task, fair evaluation and interpretation of methods in a broader context, are hindered by limited public benchmarking, inconsistent baseline implementations, and non-standardized experimentation. To overcome these pitfalls, we present nnLandmark, a self-configuring framework for 3D landmark detection that combines tailored heatmap generation, loss design, inference logic, and a robust set of hyperparameters for heatmap regression, while reusing components from nnU-Net’s underlying self-configuration and training engine. nnLandmark achieves state-of-the-art performance across five public and one private dataset, benchmarked against three recently

---

\* Supervised equally

published methods. Its out-of-the-box usability enables training strong landmark detection models on new datasets without expert knowledge or dataset-specific hyperparameter tuning. Beyond accuracy, nnLandmark provides both a strong, common baseline and a flexible, standardized environment for developing and evaluating new methodological contributions. It further streamlines evaluation across multiple datasets by offering data conversion utilities for current public benchmarks. Together, these properties position nnLandmark as a central tool for advancing 3D medical landmark detection through systematic, transparent benchmarking, enabling to genuinely measure methodological progress. The code is available on GitHub: <https://github.com/MIC-DKFZ/nnLandmark>.

**Keywords:** 3D Medical Landmark Detection, Self-Configuration, Benchmarking.

## 1. Introduction

The task of medical landmark detection concerns the prediction of coordinates of predefined, anatomical keypoints. Accurate localization is critical for several medical imaging applications, including diagnosis, treatment planning, and navigation. In practice, these include the detection of anatomical reference points for image registration, control points for biometric measurements, small critical structures for surgical planning or fetal pose estimation in ultrasound (Taha et al., 2023; He et al., 2024; Chen et al., 2020; Gong et al., 2025c). Annotating such keypoints is highly dependent on detailed anatomical knowledge. For example, for fetal brain biometry, this requires reliably localizing the cerebellar landmarks, which is complicated by densely folded cortical structures and low tissue contrast (Gong et al., 2025c). Further, medical landmark annotation often involves between 10 and 50 landmarks, resulting in a time-consuming process, especially in 3D imaging data. Efforts in automating this task based on deep learning have already shown promising results (Schwendicke et al., 2021; Serafin et al., 2023; Singh et al., 2020). While earlier approaches aimed at directly predicting coordinate values, the current state-of-the-art formulation is heatmap regression. Thereby, each landmark is represented by a Gaussian-like blob in a dedicated output channel. During prediction, the coordinates are derived via channel-wise maximum (Payer et al., 2016; Pfister et al., 2015). The default base architecture for pixel-wise heatmap regression is the U-Net (Ronneberger et al., 2015; Çiçek et al., 2016). Many efforts have been made to optimize the architecture, aiming to effectively integrate global context or maintain high spatial resolution or super-resolution for sub-pixel accuracy of localization (Huang et al., 2025; Zhang et al., 2024). However, despite active methodological research in 3D medical landmark detection, progress is still affected by various pitfalls concerning benchmarking and usability of methods (Figure 1), identified from an extensive list of relevant, recent publications (Annex A).

**Pitfall 1: Insufficient public benchmarking.** The 3D landmark detection domain suffers from a lack of established and commonly used public benchmarks (Figure 1, left). Frequently, new developments only target single, often private datasets, leaving the generalizability of these methods to other tasks in question. This hinders a transparent interpretation of the results in a broader context and limits fair comparison to other methods (He et al., 2024; Schwendicke et al., 2021). While broad public benchmarking is already the standard in segmentation and has greatly propelled the field forward (Isensee et al., 2021, 2024), in the landmark detection domain, universal insights, generalizable solutions and the broader impact of new developments are often left unexplored by focusing on single and

private datasets.

**Pitfall 2: Inconsistent baseline implementations** Many publications compare their methods to a 3D U-Net baseline (Ronneberger et al., 2015; Çiçek et al., 2016). However, variations in hyperparameters, implementations, and training setups can substantially change performance, even when using the same underlying architecture. As depicted in Figure 1, center column, this is also evident for landmark detection based on reported U-Net results on the Mandibular Molar Landmark (MML) dataset, with mean radial errors (MRE) ranging from 1.9 mm to 2.7 mm (Huang et al., 2025; He et al., 2024; Zhang et al., 2024). In absence of a strong, commonly adopted baseline, the field lacks essential context for interpreting the results of new methods and assessing progress across datasets.

**Pitfall 3: Limited out-of-the-box usability.** The lack of comprehensive benchmarking and the use of non-standardized, custom code bases have led to dataset-specific implementations. Many methods are still published without code or clear instructions on how to adapt them to new datasets, for example when dealing with different modalities or image geometries (Figure 1, right). Applying such methods to new datasets can therefore require substantial expert knowledge in model development and resource-intensive hyperparameter tuning, which complicates broader application and increases the risk of reimplementations errors. The reliance on custom code further introduces potential confounding factors, obscuring the true performance of baseline architectures and leading to unclear conclusions about new developments. A standardized environment that works out-of-the-box across datasets is therefore crucial to enable transparent, systematic evaluation of new methods and quantify true methodological progress.

To counteract these pitfalls, we make the following contributions:

- We present a comprehensive benchmark study for 3D medical landmark localization, evaluating three recent state-of-the-art methods across five public and one private datasets that span different imaging modalities and anatomical regions.
- We introduce **nnLandmark**, a fully self-configuring framework for 3D heatmap-based landmark detection that builds on the nnU-Net infrastructure to automatically derive dataset-specific preprocessing and training hyperparameters, enabling robust out-of-the-box generalization to new datasets without manual intervention.
- We show that nnLandmark consistently achieves state-of-the-art performance across all six benchmark datasets, surpassing existing methods and establishing a strong, reproducible baseline for future developments in 3D landmark detection.
- We demonstrate that nnLandmark serves as a flexible, standardized environment for method development by integrating the H3DE architecture into the framework, yielding clear performance gains over the official implementation and highlighting the value of leveraging a proven experimental infrastructure for systematic ablations and fair evaluation of new methodological contributions.

## 2. Method

Tackling the current pitfalls in landmark detection we derive four practical requirements for a newly proposed framework: (1) Provide easy access to public benchmarking datasets;

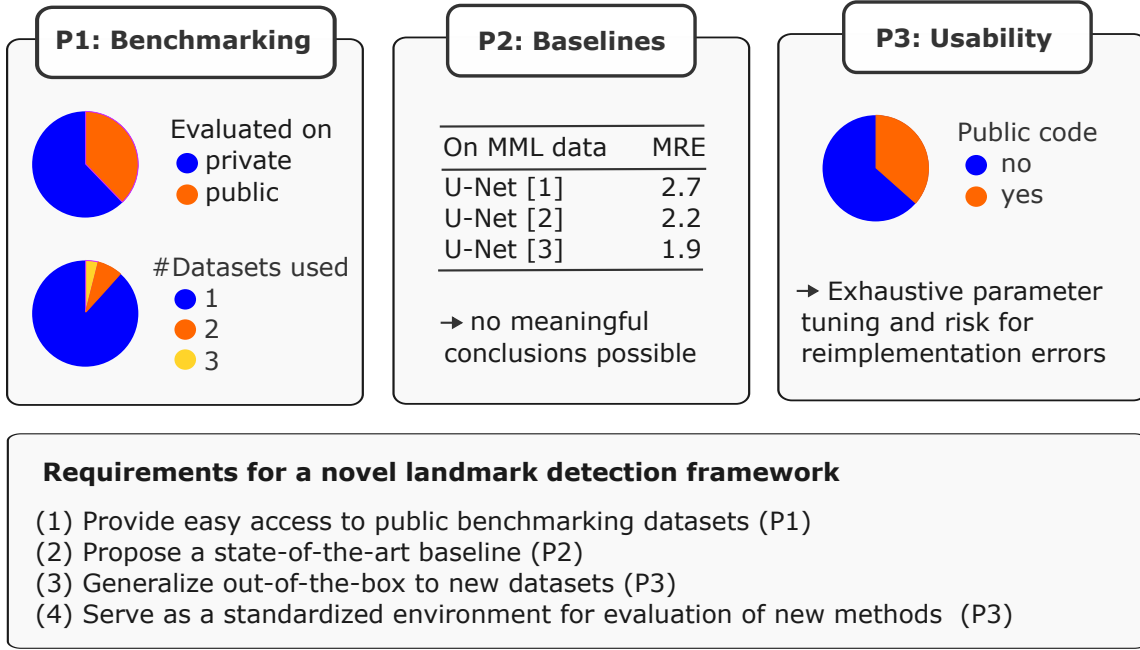


Figure 1: We identified three key pitfalls in the current landmark detection literature regarding benchmarking, baseline comparison and usability, and formulated four practical requirements for a new framework tackling these shortcomings. [1] (Zhang et al., 2024) [2] (He et al., 2024) [3] (Huang et al., 2025)

(2) propose a strong, common baseline, achieving state-of-the-art accuracy across datasets; (3) provide out-of-the-box generalizability for training on new datasets without the need for manual intervention; (4) serve as a flexible and standardized environment for evaluation of new methodological developments. In segmentation, these requirements have long been understood and are addressed by the well-established nnU-Net framework, which consistently delivers state-of-the-art performance across various datasets (Isensee et al., 2021, 2024). The key concept of nnU-Net is its self-configuration to the task at hand by automatically deriving dataset-specific properties and adjusting preprocessing and hyperparameters for a (residual) U-Net architecture. Further, nnU-Net has implemented many best practices of image processing for example regarding its data augmentation pipeline and sliding window prediction, which allow translation to heatmap regression. It is therefore well-motivated to reuse nnU-Net as a self-configuration and training engine. In the following we explore how nnLandmark can be built on this existing infrastructure to arrive at a state-of-the-art, generalizable framework for landmark detection.

To leverage nnU-Net’s data loading pipeline, which expects segmentation inputs, we initially store the landmarks in a multi-label segmentation map, with each landmark represented by a  $3 \times 3 \times 3$  voxel label. This way nnLandmark can exploit the fully automatic preprocessing and self-configuration machinery, which has been extensively tuned for 3D medical segmentation. The conversion from this multi-label representation to heatmap regression happens

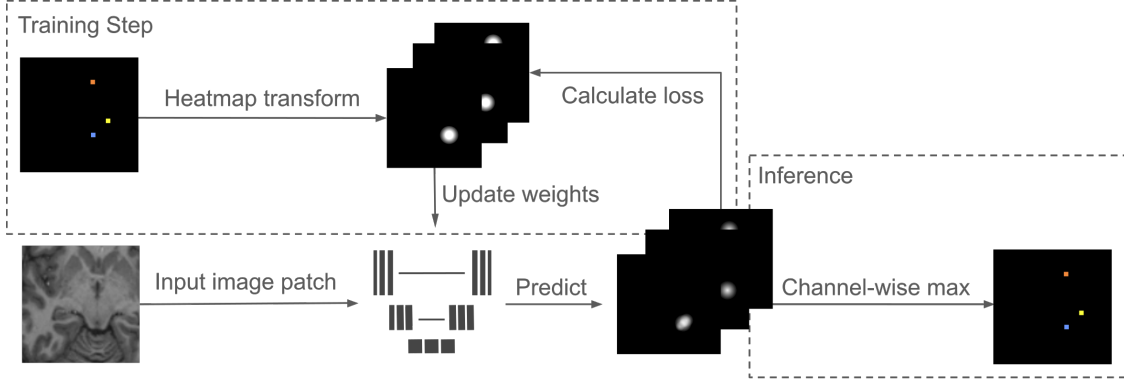


Figure 2: Leveraging nnU-Net’s data loading and augmentation pipeline, landmark segmentation maps are transformed to heatmaps only during loss computation, each landmark represented by a EDT in a dedicated channel. In inference, landmark coordinates are identified by the channel-wise maximum.

after data augmentation, directly inside the loss computation. For each foreground label, i.e. landmark, the target coordinate is obtained as the center of mass of the corresponding label region. Around this point, an Euclidean distance transform (EDT) with a radius of 15 voxels is injected into a dedicated output channel as the regression target, yielding a smooth distance-based heatmap (Figure 2). This on-the-fly transformation further avoids memory- and CPU-intensive storing and loading of large heatmaps. During prediction, a sigmoid activation in the final layer constrains voxel intensities to  $[0,1]$ , stabilizing training. Heatmap regression is trained with a Binary Cross-Entropy (BCE) TopK20 loss, which focuses the gradient signal on the most challenging voxels. Concretely, for every patch the voxel-wise BCE values are ranked and only the voxels with the highest 20% loss values contribute to the final loss, which can mitigate the foreground-background imbalance inherent to sparse landmark heatmaps. Ablation studies on the size of the EDT and choice of the loss function to evaluate the robustness of our selected parameters can be found in Annex B. For inference, we adapt nnU-Net’s sliding window prediction and derive landmark coordinates by taking the channel-wise maximum of the heatmap.

### 3. Experiments

#### 3.1. Metrics

The **Mean Radial Error (MRE)** measures the average Euclidean distance between the predicted and ground truth landmark coordinates. It is defined as:

$$\text{MRE} = \frac{1}{N} \sum_{i=1}^N \|\mathbf{x} - \hat{\mathbf{x}}\|_2 \quad (1)$$

where  $\mathbf{x}$  and  $\hat{\mathbf{x}}$  represent the ground truth and predicted coordinates for the  $i$ -th landmark, respectively, and  $N$  is the total number of landmarks. Lower MRE values indicate higher

localization accuracy.

The **Success Detection Rate (SDR)** within a tolerance range quantifies the proportion of detected landmarks that fall within a specified distance threshold  $t$  from their ground truth positions, defined as:

$$\text{SDR}@t = \frac{\# \text{ landmarks with MRE} \leq t}{\# \text{ landmarks}} \times 100. \quad (2)$$

### 3.2. Datasets

We evaluate five public and one private dataset spanning various 3D imaging modalities and anatomical regions. All test splits were used as hold-out test data.

The **Mandibular Molar Landmarking (MML)** dataset (He et al., 2024) provides 648 CT images along with annotations of 14 dental landmarks targeting the crowns and roots of the second and third mandibular molars. The dataset comes with the challenge of missing landmarks in cases where teeth are absent, structurally damaged, or have irregular root anatomy. However, we focus on predicting complete landmark annotations and use a subset that only included fully annotated cases, hereafter referred to as the complete MML subset. In accordance with the official split, the complete subset contains 283 training, 56 validation, and 60 test cases. We used the official train/test split; the validation split was not used in this study.

The **Anatomical Fiducials (AFIDs)** dataset (Taha et al., 2023; Abbass et al., 2022) consists of 132 T1 brain MRI images with 32 annotated brain landmarks as anatomical fiducials. AFIDs is a collection of 4 subsets: (1) the AFIDs-HCP30 dataset (n=30), 3T scans from the Human Connectome Project (HCP) (<https://ida.loni.usc.edu/login.jsp>) (Van Essen et al., 2012); (2) the AFIDs-OASIS30 dataset (n=30), 3T scans from the Open Access Series of Imaging Studies OASIS-1 (Marcus et al., 2007); (3) the London Health Sciences Center Parkinson’s disease (LHSCPD) dataset (n=40) containing gadolinium-enhanced images from a 1.5 T scanner (Abbas et al., 2023), and (4) the Stereotactic Neurosurgery (SNSX) (Lau et al., 2023) dataset (n=32) acquired with a 7T head-only scanner. Thus, this dataset is highly heterogeneous, with subdatasets differing in origin and imaging protocols. The human error on this dataset is reported as 0.99 mm with an inter-rater variability of 1.48 mm. As there is no official split, we performed a random split stratified across the four subsets into 110 training and 22 test cases, which will be published with the code.

The **Public Domain Database for Computational Anatomy (PDDCA)** dataset (version 1.4.1) (Raudaschl et al., 2017) contains CT scans of 33 patients with annotations for 5 bony landmarks in the head and neck area. We randomly split the data into 26 train and 7 test images, challenging the method’s robustness in low data scenarios.

The **Fetal pose estimation** dataset (Chen et al., 2020, 2024) is a private dataset provided by Shenzhen University. It encompasses 1000 fetal ultrasound (US) images with 22 landmarks throughout the head, trunk and limbs, allowing to monitor fetal position and development. We used a split provided by the dataset authors, resulting in 800 training and 200 test cases.

The **Fetal Tissue Annotation Challenge 2022 (FeTA22)** dataset contains super-resolution reconstructed T2 weighted MRIs of human fetal brains (Payette et al.; Sanchez et al., 2024). Part of the data is provided by the University Children’s Hospital Zurich

(Kispi) and is publicly available for research. This subset includes 68 cases with annotations for 10 landmarks, i.e. control points for 5 biometric measurements of the skull, brain and cerebellum. As the official test set is not public, we performed a custom random split into 65 training and 15 test cases.

The **Fetal Cerebellum Landmark Detection (LFC)** dataset (Gong et al., 2025c,b,a) contains fetal brain T2 MRIs. The annotations target 12 landmarks, which represent control points for 6 biometric measurements concerning the skull, brain and cerebellum diameters, similar to FeTA22. We use the official train/test split of 120/60. For LFC, we evaluated both landmark detection and the resulting measurements, as the dataset primarily targets the downstream task of fetal brain biometry rather than precise landmark placement and some measurements can be taken in slightly shifted locations while still producing accurate biometry values.

### 3.3. Benchmarking and Training Details

For nnLandmark, preprocessing and hyperparameters, such as patch size, batch size, network topology, are configured automatically by the framework based on dataset-specific properties. All nnLandmark experiments were performed using the *3d\_fullres* preset and 5-fold cross-validation. We trained with the plain U-Net architecture, as well as variations with a ResNet-based encoder in two sizes, M and L (ResEncM/L), following official nnU-Net recommendations (Isensee et al., 2024). We compared nnLandmark to three recently published and state-of-the-art methods and toolkits. All compared methods were trained utilizing the respective official implementations and recommendations. The goal is to compare entire frameworks and repositories against each other to evaluate their generalizability to new datasets, without requiring any custom changes or elaborate hyperparameter tuning. We additionally integrated the H3DE architecture (Huang et al., 2025) into nnLandmark to demonstrate its utility as a powerful, standardized method development framework.

The **Hybrid-3D Network (H3DE-Net)** (Huang et al., 2025) integrates transformer-based attention modules within a U-Net-like CNN structure to effectively handle local feature extraction as well as global context modeling. The design of downsampling layers and window configuration restricts the input shape to be divisible by  $64 \times 64 \times 32$ . MML training was done using a random cropping data augmentation to  $128 \times 128 \times 64$  voxels, to also fit the test image shape. For the remaining experiments, images were resized to  $128^3$  voxels. **Landmarker** (Jonkers et al., 2025b,a) is a toolkit which offers useful modules for handling landmark data and frequently used architectures in the domain. They provide a set of default configurations that show promising results on the MML data (Jonkers et al., 2025a), using a Flexible U-Net with EfficientNet backbone. Following the practices of the authors, for MML, the training data were cropped based on the annotations to  $128 \times 128 \times 64$  voxels, to fit the field of view in the test data. For the remaining datasets, images were resized to  $128^3$ . The models were trained as an ensemble using five different seeds. The **Super-Resolution U-Net (SR-UNet)** (Zhang et al., 2024) adopts pyramid pooling and super-resolution blocks to better preserve details and mitigate the error caused by downsampling and upsampling operations during training. All data was preprocessed using the published heatmap conversion script, which also includes resizing to  $128^3$ . For MML they report cropping the train data to  $128 \times 128 \times 64$  voxels, so we use the same label-based crops as for Landmarker.



Table 1: Results of landmark localisation accuracy of nnLandmark compared to current state-of-the-art methods, evaluated by MRE and micro standard deviation (std) on the respective on hold-out test data splits. All models were trained using the official code.

Method	MRE $\pm$ Std [mm]			
	MML	AFIDs	Fetal pose	PDDCA
H3DE	1.81 $\pm$ 1.15	4.28 $\pm$ 2.09	6.07 $\pm$ 6.44	8.21 $\pm$ 4.62
SR-UNet	10.01 $\pm$ 10.37	3.37 $\pm$ 1.97	5.93 $\pm$ 6.09	7.74 $\pm$ 4.45
landmarker	10.58 $\pm$ 13.92	2.86 $\pm$ 4.12	5.37 $\pm$ 7.99	4.98 $\pm$ 2.71
nnLandmark H3DE	1.63 $\pm$ 1.16	1.79 $\pm$ 1.05	4.25 $\pm$ 6.33	3.31 $\pm$ 2.31
nnLandmark	<u>1.39<math>\pm</math>0.85</u>	<u>1.55<math>\pm</math>1.01</u>	3.15 $\pm$ 5.01	<b>2.51<math>\pm</math>2.53</b>
nnLandmark ResEncM	<b>1.36<math>\pm</math>0.88</b>	<b>1.46<math>\pm</math>1.01</b>	<u>3.06<math>\pm</math>4.51</u>	2.82 $\pm$ 3.27
nnLandmark ResEncL	1.56 $\pm$ 1.22	1.61 $\pm$ 1.06	<b>3.05<math>\pm</math>4.52</b>	<u>2.72<math>\pm</math>2.76</u>

Table 2: Results of the landmark localisation accuracy and resulting biometry measurements on the FeTA and LFC datasets.

Method	MRE $\pm$ Std [mm]		Biometry error $\pm$ Std [mm]	
	FeTA	LFC	FeTA	LFC
H3DE	3.61 $\pm$ 2.74	4.22 $\pm$ 4.21	3.33 $\pm$ 2.84	2.85 $\pm$ 2.32
SR-UNet	3.41 $\pm$ 2.73	3.92 $\pm$ 3.62	3.61 $\pm$ 2.70	5.92 $\pm$ 5.55
landmarker	3.26 $\pm$ 3.76	4.02 $\pm$ 5.13	2.15 $\pm$ 3.28	1.78 $\pm$ 4.25
nnLandmark H3DE	<b>2.71<math>\pm</math>3.09</b>	<b>3.72<math>\pm</math>4.47</b>	2.12 $\pm$ 1.85	1.52 $\pm$ 1.24
nnLandmark	<u>2.87<math>\pm</math>3.19</u>	3.82 $\pm$ 4.67	<b>2.01<math>\pm</math>0.71</b>	1.79 $\pm$ 1.94
nnLandmark ResEncM	4.03 $\pm$ 10.14	<u>3.75<math>\pm</math>4.77</u>	<u>2.03<math>\pm</math>2.62</u>	<u>1.20<math>\pm</math>0.93</u>
<b>nnLandmark ResEncL</b>	3.67 $\pm$ 8.13	<u>3.75<math>\pm</math>4.75</u>	2.41 $\pm$ 4.32	<b>1.17<math>\pm</math>0.94</b>

## 4. Results

nnLandmark, particularly in the ResEncM configuration, demonstrated the overall highest performance. For the MML dataset, reproducibility results are reported in Appendix E. For H3DE we only saw slight deviation from the reported results, attributable to some randomness during training (Huang et al., 2025). For SR-UNet and Landmarker however results could not be reproduced, with both yielding an MRE above 10 mm, despite using the official repositories. This might be due to differences in handling the shift in field of view from whole-head CT during training to already cropped images in the test set. nnLandmark handles this inherently due to its patch-wise training scheme and sliding window prediction, and H3DE added random cropping during training. Landmarker required cropping the training



images based on the labels to resemble the test shape. SR-UNet similarly reports random cropping of the images. Further, the reported Landmarker results were obtained on a custom randomized data split after cropping (Jonkers et al., 2025a), hindering comparability. On AFIDs, all baselines showed moderate MRE of 3 mm to 4 mm. nnLandmark ResEncM achieved an MRE of 1.46 mm, falling within the reported inter-rater variability of 1.48 mm (Taha et al., 2023). For the fetal pose estimation task, all compared methods showed a moderate MRE of 5 mm to 6 mm. All models show a high standard deviation on this dataset, reflecting the challenging nature of the task, where the fetus can have varying positions in the uterus, arms and legs can be crossed and the respective landmarks consequently easily confused. The PDDCA results reveal that nnLandmark is the most robust in extreme low data scenarios, showing an MRE of  $> 3$  mm. For fetal brain annotation on the FeTA and LFC datasets, all models show a moderate MRE of about 3 mm to 4 mm and comparably high standard deviation, which could be attributed to the shifted annotation of some control points. A large error in landmark localization can still lead to accurate biometry measurements, which is also reflected in our results.

Integrating the H3DE architecture in our nnLandmark framework strongly improved accuracy compared to using H3DE with the published repository on all datasets. While it even achieved a slight advantage in FeTA and LFC landmark localization compared to nnLandmark U-Nets, this didn’t show in the biometry measurement results. For all datasets, MRE for each landmark class individually are presented in Appendix D. Randomly chosen example predictions for the nnLandmark ResEncM model for all datasets are shown in Figure 9. Visual examples for all methods are given in the Appendix F.

## 5. Discussion

Current research in 3D landmark detection lacks the foundation needed for systematic progress, including transparent benchmarking, consistent baselines, and methods that reliably generalize across datasets (Figure 1). Consequently, new methods are often not evaluated in a broader, standardized context and their translation to new datasets can require substantial manual effort, leading to a gap of more general solutions and insights. Tackling these pitfalls, we introduce nnLandmark, the first self-configuring framework for 3D medical landmark detection. Leveraging the established infrastructure of nnU-Net we inherit extensively optimized components for preprocessing, data augmentation and training while extending the framework with a dedicated heatmap representation, adapted loss computation, and coordinate prediction logic. This combination creates the first solution in the field to automatically adapt to new datasets without the need for expert intervention. Thereby, nnLandmark occupies the unique position of serving as an out-of-the-box usable baseline as well as a flexible framework for method development and standardized evaluation, enabling transparent and comparable experimentation in the field.

To ensure a comprehensive evaluation, we assessed nnLandmark five public and one private dataset spanning different modalities and anatomical regions and compare against three recently published methods and frameworks, H3DE, SR-UNet and Landmarker (Huang et al., 2025; Zhang et al., 2024; Jonkers et al., 2025b). Although all three report strong performance on MML, systematic evaluation beyond this dataset has been missing. Our benchmarking closes this gap and highlights the need for broader evaluation as a standard practice in medi-

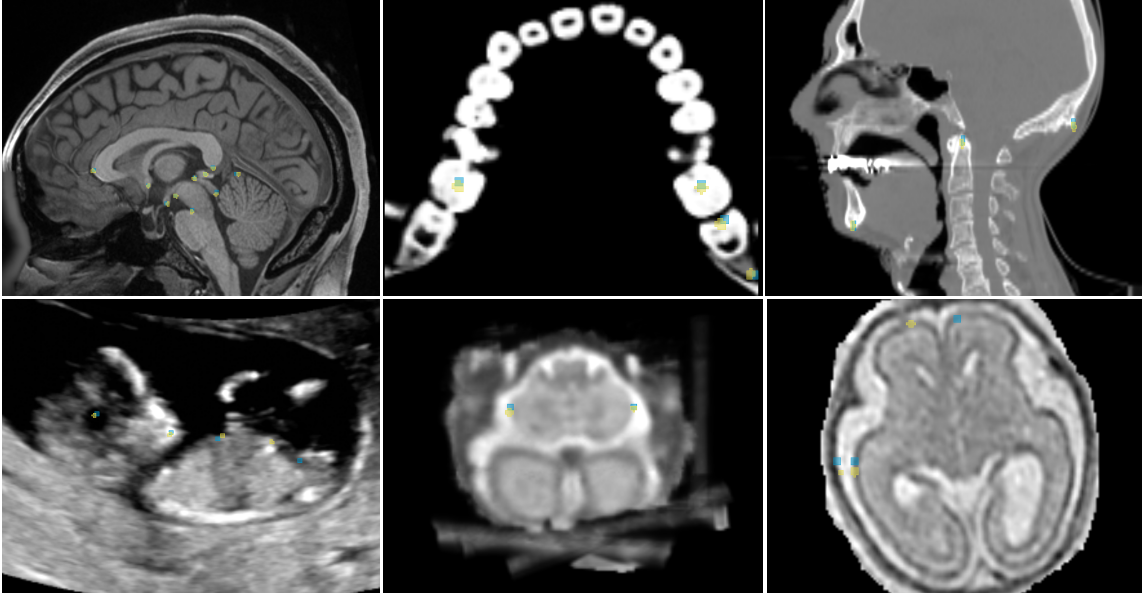


Figure 3: Visual examples for each dataset with ground truth (blue), nnLandmark ResEncM prediction (yellow). The ground truth is represented by a  $3 \times 3 \times 3$  voxel segmentation. For visualization purposes, the predicted segmentation is generated by taking the 27 voxels with the highest values from each channel in the heatmap.

cal landmark detection. The scarcity of established benchmarks and the common reliance on single, often private datasets have limited the comparability of methods. Our results show that performance shifts considerably across datasets, underlining the importance of designing and validating methods with generalization in mind. To facilitate broader adoption of multi-dataset evaluation, we release data conversion scripts for relevant public benchmarks, enabling straightforward use within the nnLandmark framework.

On the MML dataset, we also observed substantial variability among 3D U-Net baselines reported in the literature, with published MRE values ranging from 1.90 mm to 2.70 mm despite nominally identical architectures (He et al., 2024; Huang et al., 2025; Zhang et al., 2024). These discrepancies demonstrate that architectural design alone is insufficient; optimal preprocessing, hyperparameter configuration, and training practices are equally important to achieve reliable performance. nnLandmark addresses these issues as its self-configuring design eliminates the need for manual adjustments, providing a standardized, high-performing baseline for 3D landmark detection. nnLandmark’s automatic adaptation to the dataset at hand further makes it the first framework to allow training state-of-the-art landmark detection models on new datasets without the need for expert knowledge and manual tuning.

Finally, the framework provides a controlled environment for developing and ablating new methodological ideas, relieving researchers from the need to build custom experimentation code while substantially improving the comparability of results. We illustrate this by integrating the H3DE architecture into nnLandmark, which further achieved improved per-

formance compared to the official repository. These findings highlight the importance of a well-configured, standardized experimental environment for drawing meaningful, broadly applicable conclusions and reliably assessing methodological progress (Isensee et al., 2024). While nnLandmark addresses several long-standing challenges in landmark detection, some limitations remain. A limitation of storing labels as multi-label segmentation maps is the handling of closely spaced landmarks. Since each landmark is encoded as a  $3 \times 3 \times 3$  voxel cube, two landmarks must be separated by at least three voxels to avoid overlap that would distort their encoded locations. In addition, nnLandmark’s current inference design always predicts a complete set of landmarks by taking the argmax of each heatmap channel. This does not account for anatomically absent landmarks, which occur, for example, in the full MML dataset where teeth may be missing. Handling presence or absence could be incorporated by estimating a confidence threshold from cross-validation and suppressing predictions below that threshold. The same mechanism could extend the framework to small object detection, where thresholding would allow multiple instances per class. Implementing reliable presence detection and multi-instance prediction is left for future work.

## 6. Conclusion

nnLandmark is introduced as a self-configuring deep learning framework for 3D medical landmark localization based on heatmap regression. It addresses three key pitfalls of the current literature: the lack of public benchmarking, inconsistent baseline implementations, and limited out-of-the-box usability. By conducting a benchmarking study across five public and one private dataset, nnLandmark establishes a transparent reference for evaluating existing and future methods. Building on established components from nnU-Net, the framework translates self-configuration concepts to landmark detection, while providing a tailored heatmap generation, loss design, and inference logic. This enables robust generalization to new datasets without task-specific hyperparameter tuning or expert intervention. At the same time, nnLandmark offers a standardized, ready-to-use baseline and a flexible environment for method development, supporting reproducible experiments and systematic ablations. Together, these properties lay the groundwork for more rigorous and comparable research in 3D medical landmark detection, where novel ideas can be evaluated transparently and genuine methodological progress becomes measurable.

## Acknowledgments

Regarding the AFIDs-HCP dataset: Data collection and sharing for this project was provided by the Human Connectome Project (HCP; Principal Investigators: Bruce Rosen, M.D., Ph.D., Arthur W. Toga, Ph.D., Van J. Weeden, MD). HCP funding was provided by the National Institute of Dental and Craniofacial Research (NIDCR), the National Institute of Mental Health (NIMH), and the National Institute of Neurological Disorders and Stroke (NINDS). HCP data are disseminated by the Laboratory of Neuro Imaging at the University of Southern California. Regarding the AFIDs-OASIS dataset: Data were provided by OASIS-1: Cross-Sectional: Principal Investigators: D. Marcus, R. Buckner, J. Csernansky, J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382. This work was supported in part by HELMHOLTZ IMAGING, a platform of the Helmholtz Information and Data Science Incubator and by the Research Campus M2OLIE, which was funded by the German Federal Ministry of Research, Technology and Space (BMFTR) within the Framework 'Research Campus - Public-Private Partnership for Innovation' under the funding code 13GW0388A.

## References

- Mohamad Abbass, Greydon Gilmore, Alaa Taha, Ryan Chevalier, Magdalena Jach, Terry M Peters, Ali R Khan, and Jonathan C Lau. Application of the anatomical fiducials framework to a clinical dataset of patients with parkinson's disease. *Brain Structure and Function*, pages 1–13, 2022.
- Mohamad Abbass, Greydon Gilmore, Alaa Taha, Ryan Chevalier, Magdalena Jach, Terry M. Peters, Ali R. Khan, and Jonathan C. Lau. "london heath sciences center parkinson's disease dataset (lhscpd)", 2023.
- Sanjana Bakshi, Simon Freezer, Takeshi Matsumoto, and Craig Dreyer. Accuracy of an automated method of 3d soft tissue landmark detection. *European journal of orthodontics*, 43(6):622–630, 2021.
- Siavash Shirzadeh Barough, Catalina Ventura, Murat Bilgel, Marilyn S Albert, Michael I Miller, and Abhay Moghekar. Brainsignsnet: Deep learning-based 3d anatomical landmark detection in human brain imaging. *medRxiv*, pages 2025–07, 2025.
- Chaoyu Chen, Xin Yang, Ruobing Huang, Wenlong Shi, Shengfeng Liu, Mingrong Lin, Yuhao Huang, Yong Yang, Yuanji Zhang, Huanjia Luo, Yankai Huang, Yi Xiong, and Dong Ni. Region proposal network with graph prior and iou-balance loss for landmark detection in 3d ultrasound. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2020. doi: 10.1109/ISBI45749.2020.9098368.
- Chaoyu Chen, Xin Yang, Yuhao Huang, Wenlong Shi, Yan Cao, Mingyuan Luo, Xindi Hu, Lei Zhu, Lequan Yu, Kejuan Yue, Yuanji Zhang, Yi Xiong, Dong Ni, and Weijun Huang. Fetusmapv2: Enhanced fetal pose estimation in 3d ultrasound. *Medical Image Analysis*, 91:103013, 2024. ISSN 1361-8415. doi: <https://doi.org/10.1016/j.media.2023.103013>. URL <https://www.sciencedirect.com/science/article/pii/S1361841523002736>.

- Runnan Chen, Yuexin Ma, Nenglun Chen, Lingjie Liu, Zhiming Cui, Yanhong Lin, and Wenping Wang. Structure-aware long short-term memory network for 3d cephalometric landmark detection. *IEEE Transactions on Medical Imaging*, 41(7):1791–1801, 2022.
- Xiaoyang Chen, Chunfeng Lian, Hannah H Deng, Tianshu Kuang, Hung-Ying Lin, Deqiang Xiao, Jaime Gatenó, Dinggang Shen, James J Xia, and Pew-Thian Yap. Fast and accurate craniomaxillofacial landmark detection via 3d faster r-cnn. *IEEE transactions on medical imaging*, 40(12):3867–3878, 2021.
- Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf 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*, pages 424–432. Springer, 2016.
- Li Cui, Boyan Liu, Guikun Xu, Jixiang Guo, Wei Tang, and Tao He. A pseudo-3d coarse-to-fine architecture for 3d medical landmark detection. *Neurocomputing*, 614:128782, 2025.
- Juan Dai, Xinge Guo, Hongyuan Zhang, Haoyu Xie, Jiahui Huang, Qiangtai Huang, and Bingsheng Huang. Cone-beam ct landmark detection for measuring basal bone width: a retrospective validation study. *BMC Oral Health*, 24(1):1091, 2024.
- Gauthier Dot, Thomas Schouman, Shaole Chang, Frédéric Rafflenbeul, Adeline Kerbrat, Philippe Rouch, and Laurent Gajny. Automatic 3-dimensional cephalometric landmarking via deep learning. *Journal of dental research*, 101(11):1380–1387, 2022.
- Maxime Gillot, Felicia Miranda, Baptiste Baquero, Antonio Ruellas, Marcela Gurgel, Najla Al Turkestani, Luc Anchling, Nathan Hutin, Elizabeth Biggs, Marilia Yatabe, et al. Automatic landmark identification in cone-beam computed tomography. *Orthodontics & craniofacial research*, 26(4):560–567, 2023.
- Haifan Gong, Luoyao Kang, Yitao Wang, Yihan Wang, Xiang Wan, Xusheng Wu, and Haofeng Li. Nnmamba: 3d biomedical image segmentation, classification and landmark detection with state space model. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2025a. doi: 10.1109/ISBI60581.2025.10980694.
- Haifan Gong, Huixian Liu, Yitao Wang, Xiaoling Liu, Xiang Wan, Qiao Shi, and Haofeng Li. Fetal cerebellum landmark detection based on 3d mri: Method and benchmark. *IEEE Journal of Biomedical and Health Informatics*, 29(8):5712–5721, 2025b. doi: 10.1109/JBHI.2025.3559702.
- Haifan Gong, Yu Lu, Xiang Wan, and Haofeng Li. Domain generalized medical landmark detection via robust boundary-aware pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(3):3140–3148, Apr. 2025c. doi: 10.1609/aaai.v39i3.32323. URL <https://ojs.aaai.org/index.php/AAAI/article/view/32323>.
- Tao He, Guikun Xu, Li Cui, Wei Tang, Jie Long, and Jixiang Guo. Anchor ball regression model for large-scale 3d skull landmark detection. *Neurocomputing*, 567:127051, 2024.

- Zhen Huang, Ronghao Xu, Xiaoqian Zhou, Yangbo Wei, Suhua Wang, Xiaoxin Sun, Han Li, and Qingsong Yao. H3de-net: Efficient and accurate 3d landmark detection in medical imaging, 2025. URL <https://arxiv.org/abs/2502.14221>.
- Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.
- Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F Jaeger. nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 488–498. Springer, 2024.
- Yankai Jiang, Yiming Li, Xinyue Wang, Yubo Tao, Jun Lin, and Hai Lin. Cephalformer: Incorporating global structure constraint into visual features for general cephalometric landmark detection. In *International conference on medical image computing and computer-assisted intervention*, pages 227–237. Springer, 2022.
- Jef Jonkers, Frank Coopman, Luc Duchateau, Glenn Van Wallendael, and Sofie Hoecke. Reliable uncertainty quantification for 2d/3d anatomical landmark localization using multi-output conformal prediction, 03 2025a.
- Jef Jonkers, Luc Duchateau, Glenn Van Wallendael, and Sofie Van Hoecke. landmarker: A toolkit for anatomical landmark localization in 2d/3d images. *SoftwareX*, 30:102165, 2025b. ISSN 2352-7110. doi: 10.1016/j.softx.2025.102165.
- Sung Ho Kang, Kiwan Jeon, Sang-Hoon Kang, and Sang-Hwy Lee. 3d cephalometric landmark detection by multiple stage deep reinforcement learning. *Scientific reports*, 11(1): 17509, 2021.
- Yankun Lang, Xiaoyang Chen, Hannah H Deng, Tianshu Kuang, Joshua C Barber, Jaime Gateno, Pew-Thian Yap, and James J Xia. Dentalpointnet: landmark localization on high-resolution 3d digital dental models. In *International conference on medical image computing and computer-assisted intervention*, pages 444–452. Springer, 2022a.
- Yankun Lang, Chunfeng Lian, Deqiang Xiao, Hannah Deng, Kim-Han Thung, Peng Yuan, Jaime Gateno, Tianshu Kuang, David M. Alfi, Li Wang, Dinggang Shen, James J. Xia, and Pew-Thian Yap. Localization of craniomaxillofacial landmarks on cbct images using 3d mask r-cnn and local dependency learning. *IEEE Transactions on Medical Imaging*, 41(10):2856–2866, 2022b. doi: 10.1109/TMI.2022.3174513.
- Jonathan C. Lau, Yiming Xiao, Roy A. M. Haast, Greydon Gilmore, Kâmil Uludağ, Keith W. MacDougall, Ravi S. Menon, Andrew G. Parrent, Terry M. Peters, and Ali R. Khan. "stereotactic neurosurgery dataset (snsx)", 2023.
- Xiang Li, Songcen Lv, Minglei Li, Jiusi Zhang, Yuchen Jiang, Yong Qin, Hao Luo, and Shen Yin. Sdmt: Spatial dependence multi-task transformer network for 3d knee mri segmentation and landmark localization. *IEEE Transactions on Medical Imaging*, 42(8): 2274–2285, 2023. doi: 10.1109/TMI.2023.3247543.



- Jiawei Liu, Fuyong Xing, Abbas Shaikh, Brooke French, Marius George Linguraru, and Antonio R. Porras. Joint cranial bone labeling and landmark detection in pediatric ct images using context encoding. *IEEE transactions on medical imaging*, 42(10):3117–3126, 2023.
- Paula López Díez, Josefine Vilsbøll Sundgaard, François Patou, Jan Margeta, and Rasmus Reinhold Paulsen. Facial and cochlear nerves characterization using deep reinforcement learning for landmark detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 519–528. Springer, 2021.
- Gang Lu, Huazhong Shu, Han Bao, Youyong Kong, Chen Zhang, Bin Yan, Yuanxiu Zhang, and Jean-Louis Coatrieux. Cmf-net: craniomaxillofacial landmark localization on cbct images using geometric constraint and transformer. *Physics in Medicine & Biology*, 68(9):095020, 2023.
- Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9):1498–1507, 2007.
- Yijie Pang, Pujin Cheng, Junyan Lyu, Fan Lin, and Xiaoying Tang. Prior guided 3d medical image landmark localization. In Ipek Oguz, Jack Noble, Xiaoxiao Li, Martin Styner, Christian Baumgartner, Mirabela Rusu, Tobias Heinmann, Despina Kontos, Bennett Landman, and Benoit Dawant, editors, *Medical Imaging with Deep Learning*, volume 227 of *Proceedings of Machine Learning Research*, pages 1163–1175. PMLR, 10–12 Jul 2024. URL <https://proceedings.mlr.press/v227/pang24a.html>.
- Christian Payer, Darko Štern, Horst Bischof, and Martin Urschler. Regressing heatmaps for multiple landmark localization using cnns. In *International conference on medical image computing and computer-assisted intervention*, pages 230–238. Springer, 2016.
- Kelly Payette, Priscille de Dumast, Hamza Kebiri, Ivan Ezhov, Johannes C. Paetzold, Suprosanna Shit, Asim Iqbal, Romesa Khan, Raimund Kottke, Patrice Grehten, Hui Ji, Levente Lanczi, Marianna Nagy, Monika Beresova, Thi Dao Nguyen, Giancarlo Natalucci, Theofanis Karayannis, Bjoern Menze, Meritxell Bach Cuadra, and Andras Jakab. An automatic multi-tissue human fetal brain segmentation benchmark using the fetal tissue annotation dataset. 8(1):167. ISSN 2052-4463. doi: 10.1038/s41597-021-00946-3. URL <https://doi.org/10.1038/s41597-021-00946-3>.
- Tomas Pfister, James Charles, and Andrew Zisserman. Flowing convnets for human pose estimation in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 1913–1921, 2015.
- Patrik F. Raudaschl, Paolo Zaffino, Gregory C. Sharp, Maria Francesca Spadea, Antong Chen, Benoit M. Dawant, Thomas Albrecht, Tobias Gass, Christoph Langguth, Marcel Lüthi, Florian Jung, Oliver Knapp, Stefan Wesarg, Richard Mannion-Haworth, Mike Bowes, Annaliese Ashman, Gwenael Guillard, Alan Brett, Graham Vincent, Mauricio Orbes-Arteaga, David Cárdenas-Peña, German Castellanos-Dominguez, Nava Aghdasi,



- Yangming Li, Angelique Berens, Kris Moe, Blake Hannaford, Rainer Schubert, and Karl D. Fritscher. Evaluation of segmentation methods on head and neck ct: Auto-segmentation challenge 2015. *Medical Physics*, 44(5):2020–2036, 2017. doi: <https://doi.org/10.1002/mp.12197>.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. URL <https://arxiv.org/abs/1505.04597>.
- Soorena Salari, Amirhossein Rasoulzadeh, Michele Battie, Maryse Fortin, Hassan Rivaz, and Yiming Xiao. Uncertainty-aware transformer model for anatomical landmark detection in paraspinal muscle mris. In *Medical Imaging 2023: Image Processing*, volume 12464, pages 246–252. SPIE, 2023.
- Thomas Sanchez, Yvan Gomez, Roxane Licandro, Kelly Payette, Andras Jakab, Meriam Koob, and Meritxell Bach Cuadra. Fetal tissue annotation challenge (feta) biometry - miccai 2024, May 2024. URL <https://doi.org/10.5281/zenodo.11192452>.
- Falk Schwendicke, Akhilanand Chaurasia, Lubaina Arsiwala, Jae-Hong Lee, Karim Elhennawy, Paul-Georg Jost-Brinkmann, Flavio Demarco, and Joachim Krois. Deep learning for cephalometric landmark detection: systematic review and meta-analysis. *Clinical oral investigations*, 25(7):4299–4309, 2021.
- Marco Serafin, Benedetta Baldini, Federico Cabitza, Gianpaolo Carrafiello, Giuseppe Baselli, Massimo Del Fabbro, Chiarella Sforza, Alberto Caprioglio, and Gianluca M Tartaglia. Accuracy of automated 3d cephalometric landmarks by deep learning algorithms: systematic review and meta-analysis. *La radiologia medica*, 128(5):544–555, 2023.
- Kaibo Shi, Hairong Jin, and Youyi Zheng. End-to-end 3d tooth landmark detection with fuzzy tooth localization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 170–180. Springer, 2025.
- Satya P Singh, Lipo Wang, Sukrit Gupta, Haveesh Goli, Parasuraman Padmanabhan, and Balázs Gulyás. 3d deep learning on medical images: a review. *Sensors*, 20(18):5097, 2020.
- Jannik Stebani, Martin Blaimer, Simon Zabler, Tilmann Neun, Daniël M Pelt, and Kristen Rak. Towards fully automated inner ear analysis with deep-learning-based joint segmentation and landmark detection framework. *Scientific Reports*, 13(1):19057, 2023.
- Alaa Taha, Greydon Gilmore, Mohamad Abbass, Jason Kai, Tristan Kuehn, John Demarco, Geetika Gupta, Chris Zajner, Daniel Cao, Ryan Chevalier, et al. Magnetic resonance imaging datasets with anatomical fiducials for quality control and registration. *Scientific Data*, 10(1):449, 2023.
- Neslisah Torosdagli, Syed Anwar, Payal Verma, Denise K Liberton, Janice S Lee, Wade W Han, and Ulas Bagci. Relational reasoning network for anatomical landmarking. *Journal of Medical Imaging*, 10(2):024002–024002, 2023.
- David C Van Essen, Kamil Ugurbil, Edward Auerbach, Deanna Barch, Timothy EJ Behrens, Richard Bucholz, Acer Chang, Liyong Chen, Maurizio Corbetta, Sandra W Curtiss, et al.

- The human connectome project: a data acquisition perspective. *Neuroimage*, 62(4):2222–2231, 2012.
- Tai-Hsien Wu, Chunfeng Lian, Sanghee Lee, Matthew Pastewait, Christian Piers, Jie Liu, Fan Wang, Li Wang, Chiung-Ying Chiu, Wenchi Wang, Christina Jackson, Wei-Lun Chao, Dinggang Shen, and Ching-Chang Ko. Two-stage mesh deep learning for automated tooth segmentation and landmark localization on 3d intraoral scans. *IEEE Transactions on Medical Imaging*, 41(11):3158–3166, 2022. doi: 10.1109/TMI.2022.3180343.
- Runshi Zhang, Hao Mo, Weini Hu, Bimeng Jie, Lin Xu, Yang He, Jia Ke, and Junchen Wang. Super-resolution landmark detection networks for medical images. *Computers in Biology and Medicine*, 182:109095, 2024.
- Junjun Zhu, Qijie Zhao, Junhao Zhu, Anwen Zhou, and Hui Shao. A novel method for 3d knee anatomical landmark localization by combining global and local features. *Machine Vision and Applications*, 33(4):52, 2022.

## Appendix A. List of analyzed publications for Figure 1

We identified current pitfalls in the 3D medical landmark detection literature based on the following representative list of relevant methodological publications since 2021: (Huang et al., 2025; Zhang et al., 2024; Jonkers et al., 2025b; Gong et al., 2025a; He et al., 2024; Chen et al., 2021, 2024; Cui et al., 2025; Liu et al., 2023; Pang et al., 2024; Gong et al., 2025b; Barough et al., 2025; Chen et al., 2022; Kang et al., 2021; Jiang et al., 2022; Shi et al., 2025; Dai et al., 2024; Stebani et al., 2023; Baksi et al., 2021; Gillot et al., 2023; Salari et al., 2023; Wu et al., 2022; Li et al., 2023; Lu et al., 2023; Lang et al., 2022b; López Diez et al., 2021; Zhu et al., 2022; Lang et al., 2022a; Dot et al., 2022; Torosdagli et al., 2023)

## Appendix B. Ablation Study on Hyperparameters

Table 3: Ablation study assessing the robustness of our hyperparameter choices. EDT denotes the Euclidean distance transform radius used to generate the heatmap supervision for landmark localization during training. We additionally compare alternative loss functions, including MSE and BCE Topk with varying k. nnLandmark uses EDT 15 and BCE Top20 loss.

Parameter	MRE [mm]	
	Afids	Fetal Pose
EDT 7	1.65	17.59
EDT 11	1.59	3.21
EDT 15	1.66	2.87
EDT 19	1.64	2.85
EDT 23	1.68	2.82
MSE loss	12.06	52.48
BCE Top 10	1.63	2.85
BCE Top 20	1.66	2.87
BCE Top 30	1.62	3.00
BCE Top 40	1.62	3.00
BCE Top 50	1.67	3.02
BCE Top 60	1.68	3.21
BCE Top 70	1.69	3.15
BCE Top 80	1.72	3.15
BCE Top 90	1.76	3.23
BCE Top 100	1.76	3.19



## Appendix C. Extended Results Including Success Detection Rate (SDR)

Table 4: Extended results table with SDR. All results are done on the hold-out testsplits.

Method	MRE±Std	SDR [%]		
	[mm]	2 mm	3 mm	4 mm
AFIDs #samples 22, #landmarks 32				
H3DE (Huang et al., 2025)	4.28±2.09	13.07	28.98	51.28
SR-UNet (Zhang et al., 2024)	3.37±1.97	25.43	50.99	70.03
landmarker (Jonkers et al., 2025b)	2.86±4.12	46.16	62.22	74.57
nnLandmark H3DE	1.79±1.05	67.90	88.78	97.02
nnLandmark	1.55±1.01	76.85	93.61	97.87
<b>nnLandmark ResEncM</b>	<b>1.46±1.01</b>	<b>81.82</b>	<b>94.74</b>	<b>98.15</b>
nnLandmark ResEncL	1.61±1.06	75.71	92.90	97.44
MML complete subset #samples 60, #landmarks 14				
H3DE (Huang et al., 2025)	1.81±1.15	67.14	89.52	97.14
SR-UNet (Zhang et al., 2024)	10.01±10.37	5.24	13.93	24.17
landmarker (Jonkers et al., 2025b)	10.58±13.92	24.05	34.05	42.62
nnLandmark H3DE	1.63±1.16	72.50	91.19	97.26
nnLandmark	1.39±0.85	80.00	95.24	98.57
<b>nnLandmark ResEncM</b>	<b>1.36±0.88</b>	<b>82.02</b>	<b>95.48</b>	<b>98.69</b>
nnLandmark ResEncL	1.59±1.22	75.24	92.50	97.98
Fetal pose #samples 200, #landmarks 22				
H3DE (Huang et al., 2025)	6.07±6.44	11.21	27.81	45.02
SR-UNet (Zhang et al., 2024)	66.11±19.15	0.04	0.04	0.04
landmarker (Jonkers et al., 2025b)	5.37±7.99	36.45	56.16	67.36
nnLandmark H3DE	4.35±6.33	44.77	64.23	74.77
nnLandmark	3.15±5.01	49.77	70.86	82.16
nnLandmark ResEncM	3.06±4.51	50.80	70.64	81.68
<b>nnLandmark ResEncL</b>	<b>3.05±4.52</b>	<b>51.19</b>	<b>71.09</b>	<b>81.80</b>
PDDCA #samples 7, #landmarks 5				
H3DE (Huang et al., 2025)	8.21±4.62	2.86	5.71	11.43
SR-UNet (Zhang et al., 2024)	7.74±4.45	2.86	11.43	28.57
landmarker (Jonkers et al., 2025b)	4.98±2.71	2.86	11.43	37.14
nnLandmark H3DE	3.31±2.31	25.71	54.29	77.14
nnLandmark	2.51±2.53	45.71	74.29	91.43
<b>nnLandmark ResEncM</b>	<b>2.82±3.27</b>	<b>45.71</b>	<b>71.43</b>	<b>88.57</b>
nnLandmark ResEncL	2.72±2.76	40.00	68.57	88.57

Table 5: Extended results table with SDR. All results are done on the hold-out test splits.

Method	MRE±Std	SDR [%]		
	[mm]	2 mm	3 mm	4 mm
<b>FeTA</b> #samples 15, #landmarks 10				
H3DE (Huang et al., 2025)	3.61±2.74	23.33	53.33	74.67
SR-UNet (Zhang et al., 2024)	3.41±2.73	30.00	54.67	68.67
landmarker (Jonkers et al., 2025b)	3.26±3.76	54.00	73.33	78.00
nnLandmark H3DE	2.71±3.09	59.33	78.00	83.33
nnLandmark	2.87±3.19	58.67	73.33	80.00
<b>nnLandmark ResEncM</b>	<b>4.03±10.14</b>	<b>56.67</b>	<b>72.00</b>	<b>79.33</b>
nnLandmark ResEncL	3.67±8.13	58.00	73.33	80.00
<b>LFC</b> #samples 60, #landmarks 12				
H3DE (Huang et al., 2025)	4.22±4.21	34.03	55.42	67.78
SR-UNet (Zhang et al., 2024)	3.92±3.62	35.42	55.97	69.72
landmarker (Jonkers et al., 2025b)	4.02±5.13	50.56	68.19	76.67
nnLandmark H3DE	3.72±4.47	51.94	67.91	77.22
nnLandmark	3.72±4.47	51.94	67.91	77.22
<b>nnLandmark ResEncM</b>	<b>3.75±4.77</b>	<b>55.42</b>	<b>69.17</b>	<b>77.08</b>
nnLandmark ResEncL	3.75±4.75	54.44	68.61	76.53

## Appendix D. Individual Landmark Class Errors

Table 6: Landmark localization results of nnLandmark ResEncM for AFIDs dataset with #samples 22, #landmarks 32.

Landmark Class	MRE $\pm$ Std
AC [midline]	0.68 $\pm$ 0.34
PC [midline]	0.90 $\pm$ 0.39
Infracollicular sulcus [midline]	1.19 $\pm$ 0.39
Pontomesencephalic junction [midline]	1.38 $\pm$ 0.73
Superior interpeduncular fossa [midline]	0.87 $\pm$ 0.35
Right superior lateral mesencephalic sulcus	1.19 $\pm$ 0.50
Left superior lateral mesencephalic sulcus	1.06 $\pm$ 0.54
Right inferior lateral mesencephalic sulcus	1.43 $\pm$ 0.67
Left inferior lateral mesencephalic sulcus	1.35 $\pm$ 0.68
Culmen [midline]	1.79 $\pm$ 0.95
Intermamillary sulcus [midline]	1.10 $\pm$ 0.46
Right mamillary body	0.95 $\pm$ 0.42
Left mamillary body	1.12 $\pm$ 0.50
Pineal gland [midline]	1.61 $\pm$ 0.84
Right lateral aspect of frontal horn at AC	1.70 $\pm$ 1.13
Left lateral aspect of frontal horn at AC	1.95 $\pm$ 1.19
Right lateral aspect of frontal horn at PC	1.86 $\pm$ 0.92
Left lateral aspect of frontal horn at PC	1.71 $\pm$ 0.93
Genu of corpus callosum [midline]	1.10 $\pm$ 0.39
Splenium of the corpus callosum [midline]	1.22 $\pm$ 0.40
Right anterolateral temporal horn	1.16 $\pm$ 0.70
Left anterolateral temporal horn	1.45 $\pm$ 0.52
Right superior AM temporal horn	1.45 $\pm$ 0.62
Left superior AM temporal horn	1.95 $\pm$ 0.93
Right inferior AM temporal horn	2.14 $\pm$ 0.91
Left inferior AM temporal horn	2.20 $\pm$ 1.08
Right indusium griseum origin	1.42 $\pm$ 0.78
Left indusium griseum origin	1.75 $\pm$ 0.64
Right ventral occipital horn	1.80 $\pm$ 1.60
Left ventral occipital horn	2.46 $\pm$ 3.03
Right olfactory sulcal fundus	1.43 $\pm$ 0.65
Left olfactory sulcal fundus	1.20 $\pm$ 0.45



Table 7: Landmark localization results of nnLandmark ResEncM for MML dataset with #samples 60, #landmarks 14.

Landmark Class	MRE $\pm$ Std
Left cuspid cusp	1.29 $\pm$ 1.06
Left 2nd molar crown	0.96 $\pm$ 0.56
Left 2nd molar mesial root	1.26 $\pm$ 0.65
Left 2nd molar distal root	1.33 $\pm$ 0.75
Left 3rd molar crown	1.20 $\pm$ 0.68
Left 3rd molar mesial root	1.63 $\pm$ 1.31
Left 3rd molar distal root	1.90 $\pm$ 0.89
Right cuspid cusp	1.29 $\pm$ 0.77
Right 2nd molar crown	1.17 $\pm$ 0.61
Right 2nd molar mesial root	1.25 $\pm$ 0.61
Right 2nd molar distal root	1.41 $\pm$ 0.79
Right 3rd molar crown	1.04 $\pm$ 0.46
Right 3rd molar mesial root	1.50 $\pm$ 0.81
Right 3rd molar distal root	1.74 $\pm$ 1.28

Table 8: Landmark localization results of nnLandmark ResEncM for PDDCA dataset with #samples 7, #landmarks 5.

Landmark Class	MRE $\pm$ Std
Chin	2.03 $\pm$ 1.02
Left mandibular	2.51 $\pm$ 1.65
Right mandibular	1.89 $\pm$ 0.91
Occipital bone	6.17 $\pm$ 5.74
Odontoid process	1.51 $\pm$ 1.08

Table 9: Landmark localization results of nnLandmark ResEncM for fetal pose estimation dataset with #samples 200, #landmarks 22.

<b>Landmark Class</b>	<b>MRE<math>\pm</math>Std</b>
Cranial crest	5.15 $\pm$ 8.43
Diencephalon	2.56 $\pm$ 4.74
Thalamus	1.84 $\pm$ 1.32
Nasal bone	1.93 $\pm$ 2.57
Lower alveolar	1.47 $\pm$ 1.09
Hind neck	3.78 $\pm$ 3.70
Chest wall	3.85 $\pm$ 3.04
Diaphragm lumbar	3.71 $\pm$ 4.10
Buttocks	3.98 $\pm$ 5.91
Umbilical	4.09 $\pm$ 4.41
Left shoulder	2.15 $\pm$ 1.99
Left elbow	2.37 $\pm$ 3.35
Left wrist	3.00 $\pm$ 4.86
Right shoulder	2.99 $\pm$ 5.39
Right elbow	2.54 $\pm$ 4.00
Right wrist	3.13 $\pm$ 4.95
Left hip	2.37 $\pm$ 3.90
Left knee	2.60 $\pm$ 4.10
Left ankle	3.96 $\pm$ 5.24
Right hip	2.75 $\pm$ 5.22
Right knee	2.55 $\pm$ 2.99
Right ankle	4.51 $\pm$ 5.26

Table 10: Landmark localization results of nnLandmark ResEncM for FeTA dataset with #samples 15, #landmarks 10.

Landmark Class	MRE $\pm$ Std
Brain biparietal diameter 1 (bBIP1)	0.82 $\pm$ 0.56
Brain biparietal diameter 2 (bBIP2)	1.41 $\pm$ 0.53
Skull biparietal diameter 1 (sBIP1)	4.67 $\pm$ 5.78
Skull biparietal diameter 2 (sBIP2)	4.46 $\pm$ 5.35
Height of vermis 1 (HV1)	3.50 $\pm$ 2.90
Height of vermis 2 (HV2)	10.04 $\pm$ 20.14
Length of the corpus callosum 1 (LCC2)	2.86 $\pm$ 1.31
Length of the corpus callosum 2 (LCC2)	9.47 $\pm$ 21.23
Transverse cerebellar diameter 1 (TCD1)	1.31 $\pm$ 0.78
Transverse cerebellar diameter 2 (TCD2)	1.75 $\pm$ 1.28

Table 11: Biometry measurements results of nnLandmark ResEncM for FeTA with #samples 15, #measurements 10.

Landmark Class	MRE $\pm$ Std
Brain biparietal diameter (bBIP) in the axial plane	1.71 $\pm$ 1.05
Skull biparietal diameter (sBIP) in the axial plane	2.22 $\pm$ 3.25
Height of the vermis (HV) in the sagittal plane	2.47 $\pm$ 2.67
Length of the corpus callosum (LCC) in the sagittal plane	1.66 $\pm$ 1.94
Maximum transverse cerebellar diameter (TCD) in the coronal plane	2.06 $\pm$ 3.37

Table 12: Landmark localization results of nnLandmark ResEncM for LFC dataset with #samples 60, #landmarks 12.

Landmark Class	MRE $\pm$ Std
Brain biparietal diameter 1 (bBIP1)	6.09 $\pm$ 5.76
Brain biparietal diameter 2 (bBIP2)	6.15 $\pm$ 5.74
Skull biparietal diameter 1 (sBIP1)	6.11 $\pm$ 4.77
Skull biparietal diameter 2 (sBIP2)	5.99 $\pm$ 4.97
Transverse cerebellar diameter 1 (TCD1)	1.30 $\pm$ 0.64
Transverse cerebellar diameter 2 (TCD2)	1.21 $\pm$ 0.69
Occipitofrontal diameter 1 (OFD1)	6.52 $\pm$ 7.32
Occipitofrontal diameter 2 (OFD2)	6.09 $\pm$ 5.83
Height of vermis 1 (HDV1)	1.41 $\pm$ 0.66
Height of vermis 2 (HDV2)	1.45 $\pm$ 0.58
Anteroposterior diameter of vermis 1 (ADV1)	1.57 $\pm$ 0.68
Anteroposterior diameter of vermis 2 (ADV2)	1.14 $\pm$ 0.55

Table 13: Biometry measurements results of nnLandmark ResEncM for LFC with #samples 60, #measurements 6.

Landmark Class	MRE $\pm$ Std
Brain biparietal diameter (bBIP)	1.49 $\pm$ 1.23
Skull biparietal diameter (sBIP)	1.24 $\pm$ 0.94
Transverse cerebellar diameter (TCD)	1.07 $\pm$ 0.76
Occipitofrontal diameter (OFD)	1.25 $\pm$ 0.94
Height of vermis (HDV)	1.30 $\pm$ 0.84
Anteroposterior diameter of vermis (ADV)	0.85 $\pm$ 0.65

## Appendix E. Reproducibility Results

Table 14: Reproducibility results on the complete subset of the Mandibular Molar Landmark (MML) dataset.

Method	MRE±Std [mm]
H3DE reported ( <a href="#">Huang et al., 2025</a> )	1.68±0.45
H3DE reproduced ( <a href="#">Huang et al., 2025</a> )	1.81±1.15
SR-UNet reported ( <a href="#">Zhang et al., 2024</a> )	2.01±4.33
SR-UNet reproduced ( <a href="#">Zhang et al., 2024</a> )	10.01±10.37
landmarker reported (different split) ( <a href="#">Jonkers et al., 2025b</a> )	1.39
landmarker reproduced ( <a href="#">Jonkers et al., 2025b</a> )	10.58±13.92

## Appendix F. Qualitative Examples per Dataset

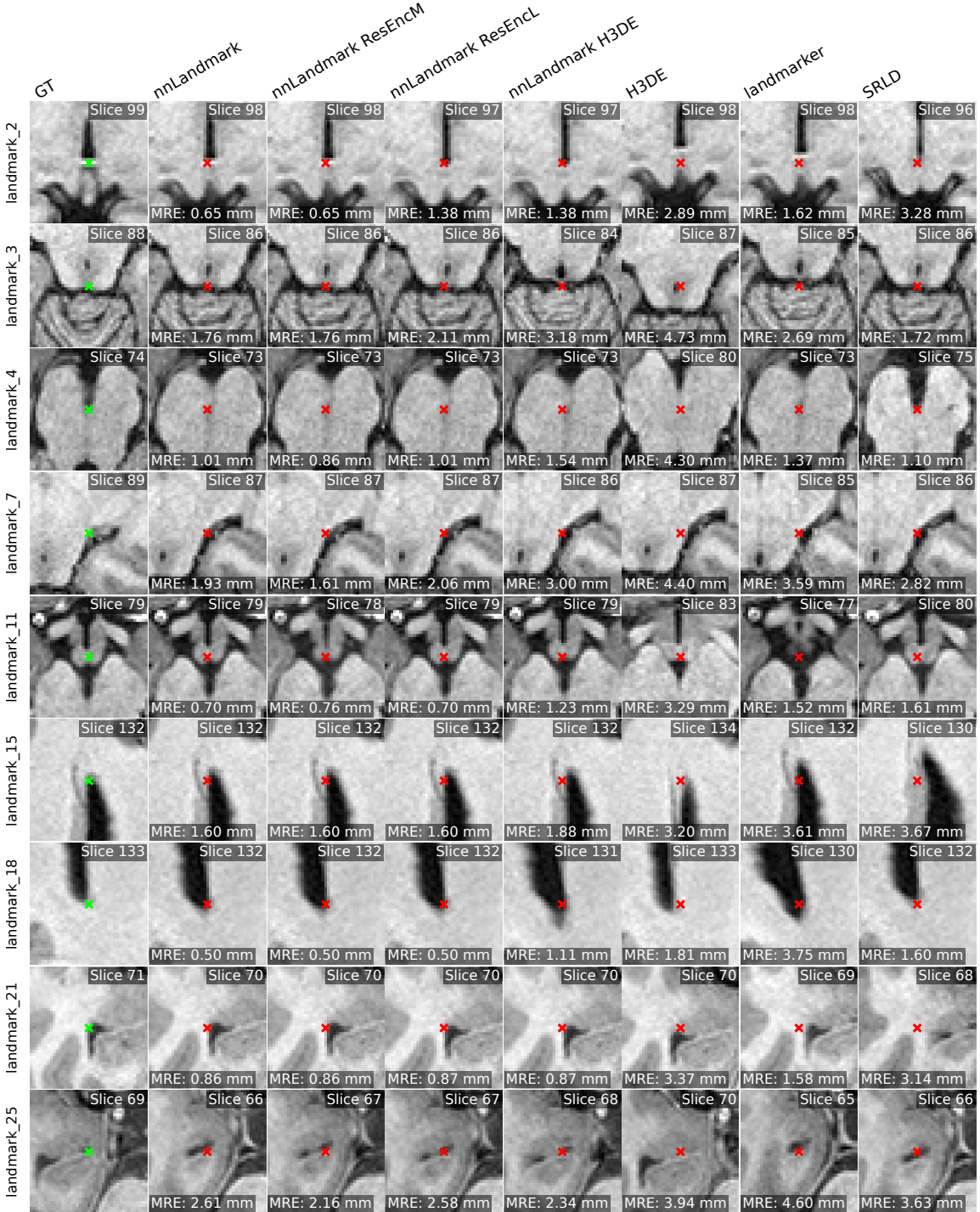


Figure 4: Qualitative examples of selected landmarks on Afids dataset.

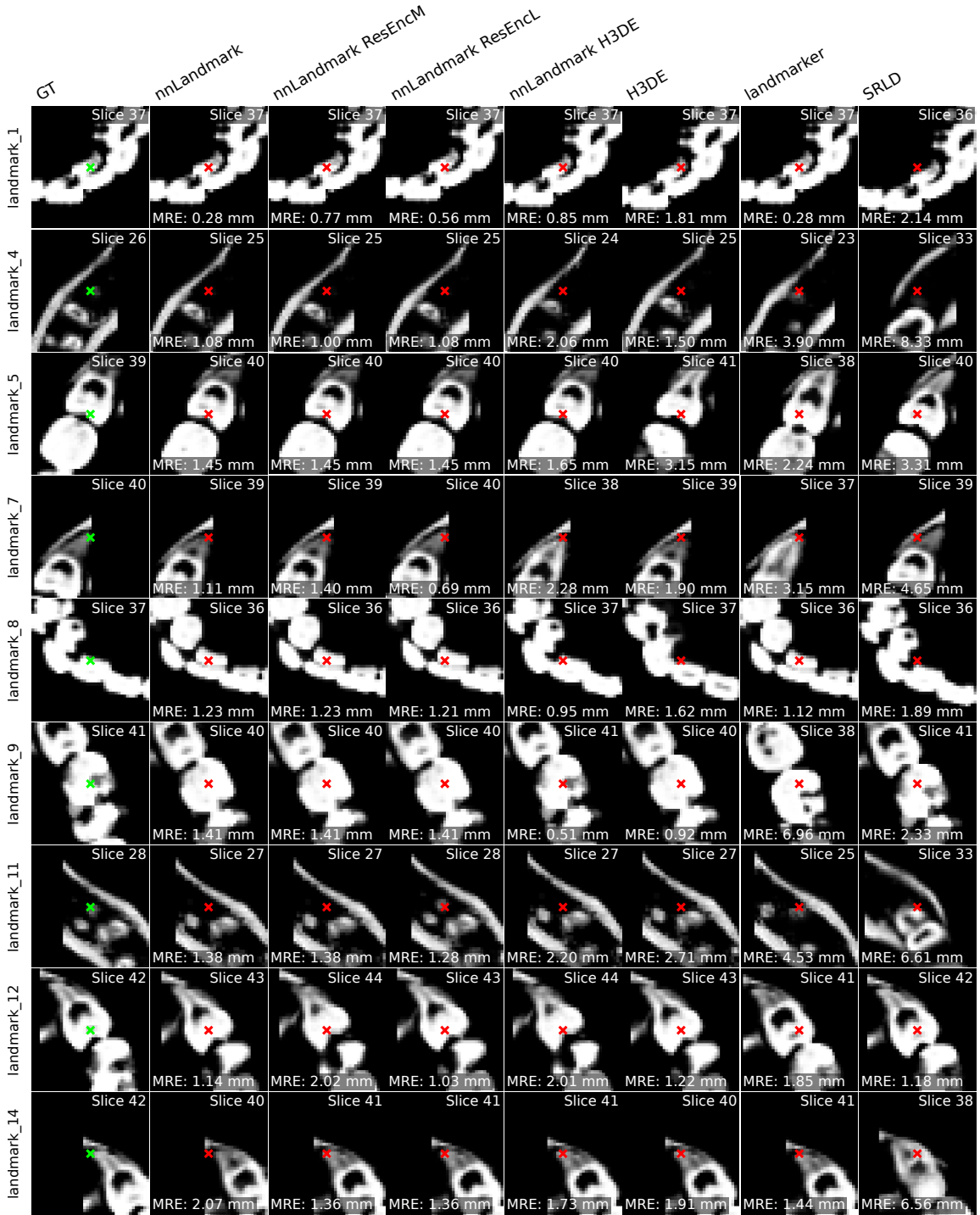


Figure 5: Qualitative examples of selected landmarks on MML dataset.



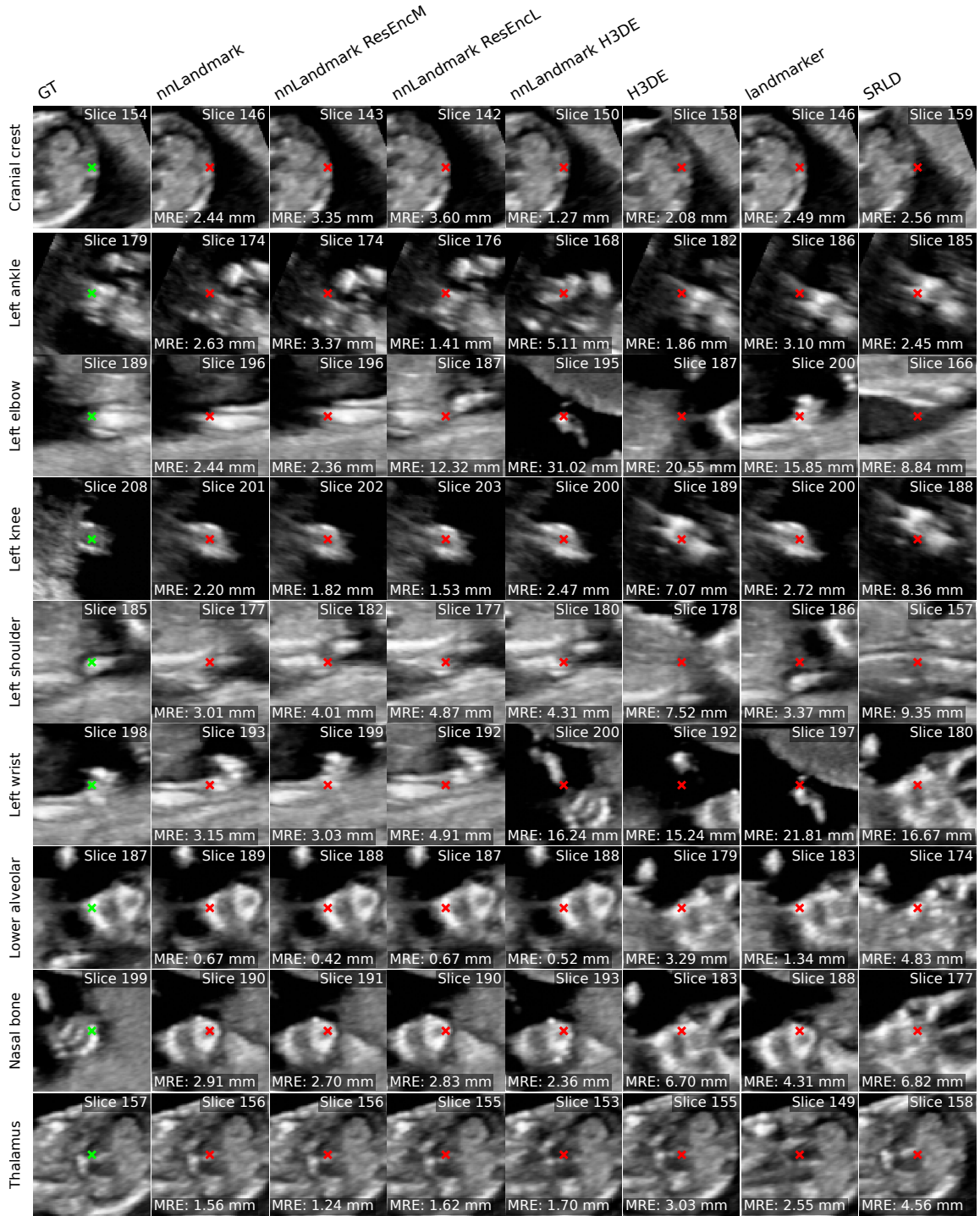


Figure 6: Qualitative examples of selected landmarks on Fetal pose dataset.

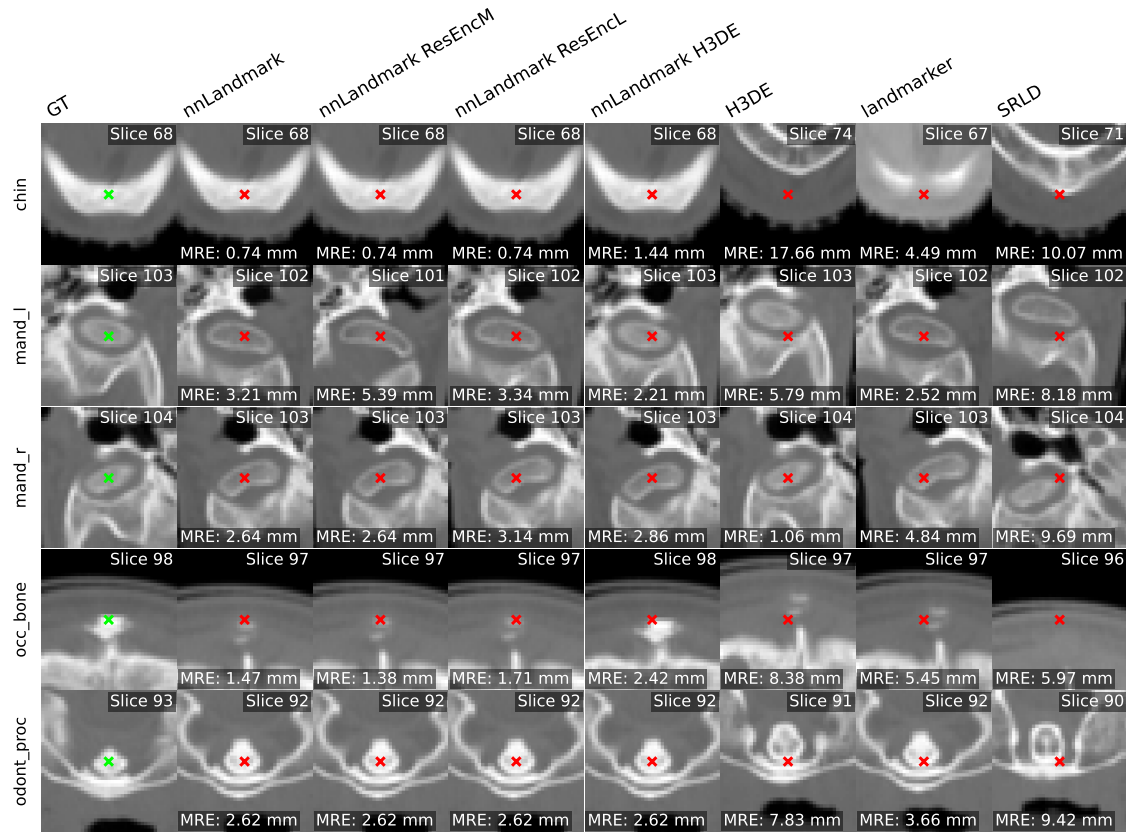


Figure 7: Qualitative examples of selected landmarks on PDDCA dataset.

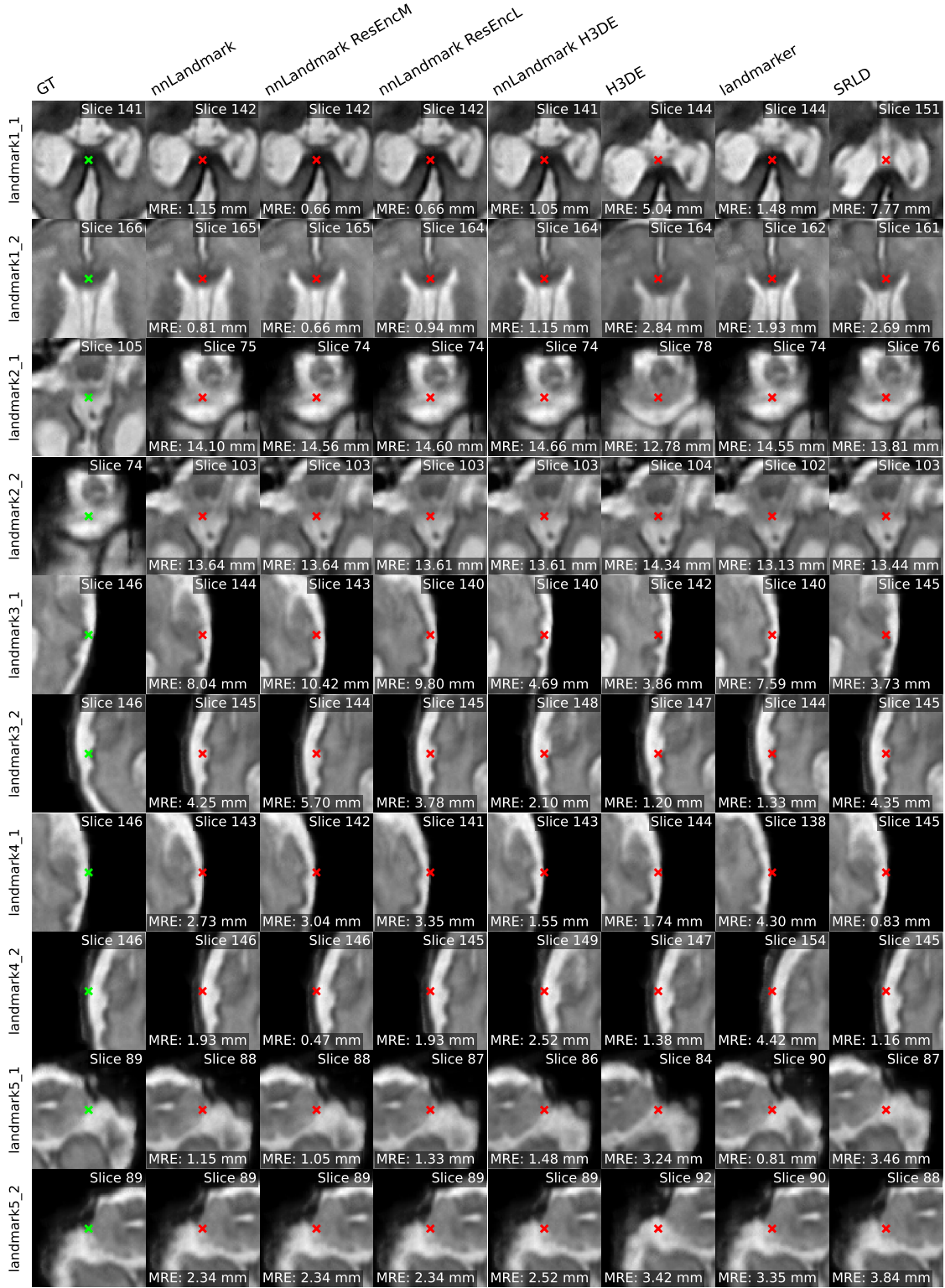


Figure 8: Qualitative examples of selected landmarks on FeTA dataset.

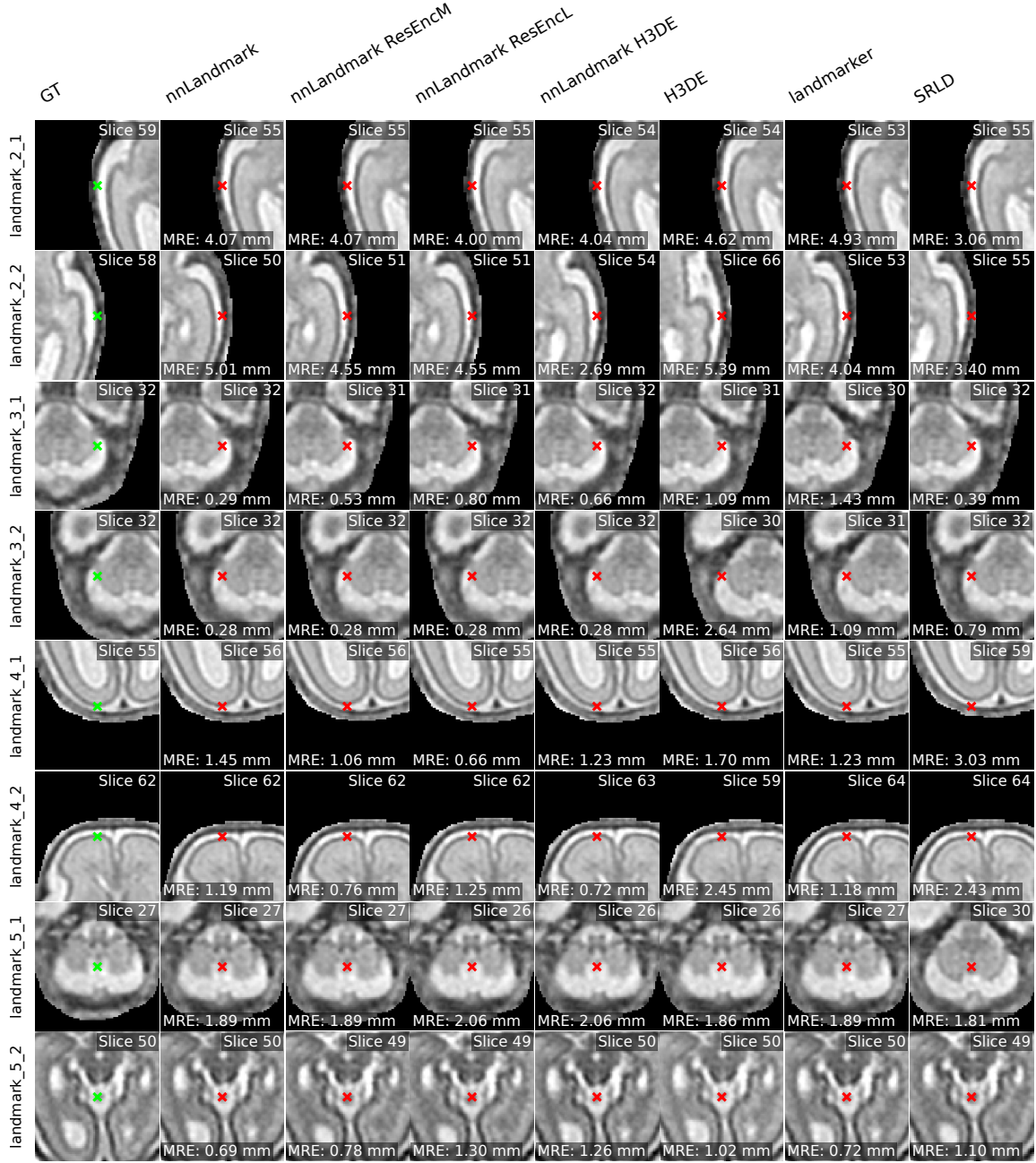


Figure 9: Qualitative examples of selected landmarks on LFC dataset.

