MedSapiens: Taking a Pose to Rethink Medical Imaging Landmark Detection

Marawan Elbatel[1]*, Anbang Wang[1]*, Keyuan Liu[2], Kaouther Mouheb[3], Enrique Almar-Munoz[4],
Lizhuo Lin[2], Yanqi Yang[2], Karim Lekadir[5], Xiaomeng Li[1]

[1]The Hong Kong University of Science and Technology, Hong Kong
[2]The University of Hong Kong, Hong Kong
[3]Erasmus MC, Netherlands
[4]Medical University of Innsbruck, Austria
[5]University of Barcelona, Spain

Abstract

This paper does not introduce a novel architecture; instead, it revisits a fundamental yet overlooked baseline: adapting human-centric foundation models for anatomical landmark detection in medical imaging. While landmark detection has traditionally relied on domain-specific models, the emergence of large-scale pre-trained vision models presents new opportunities. In this study, we investigate the adaptation of **Sapiens**, a human-centric foundation model designed for pose estimation, for medical imaging through multi-dataset pretraining, establishing a new state-of-the-art across multiple datasets. Our proposed model, **MedSapiens**, demonstrates that human-centric foundation models, inherently optimized for spatial pose localization, provide strong priors for anatomical landmark detection, yet this potential has remained largely untapped. We benchmark **MedSapiens** against existing state-of-the-art models, achieving up to **5.26%** improvement over generalist models and up to **21.81%** improvement over specialist models in the average success detection rate (SDR). To further assess **MedSapiens'** adaptability to novel downstream tasks with few annotations, we evaluate its performance in limited-data settings, achieving **2.69%** improvement over the few-shot state-of-the-art in the SDR metric. Code and model weights are available at https://github.com/xmed-lab/MedSapiens.

1 Introduction

Anatomical landmark detection refers to identifying specific points on anatomical structures, essential for spatial understanding and guiding clinical tasks such as diagnos-

tics, treatment planning, and surgical navigation. Accurate landmark detection is crucial in clinical practice for applications such as brain tumor resection [1], infant hip dysplasia diagnosis [8, 12], and cephalometric analysis in orthodontics [2]. Current methods for anatomical landmark detection are predominantly single-task oriented, constrained by limited dataset sizes, and exhibit poor generalization to novel tasks. While foundation models have demonstrated improvements in accuracy and generalization for segmentation [14, 27, 29], classification [28], and registration [3, 20], their progress in anatomical landmark detection remains constrained by the lack of a unified framework that jointly optimizes multiple downstream tasks. Therefore, developing an anatomical landmark foundation model is crucial to improving generalization, enabling cross-task adaptability, and achieving broader clinical impact in medical imaging.

Existing approaches for anatomical landmark detection leverage a variety of priors, including geometric priors [18], generative priors [9], and anatomical priors [12]. These methods, applied in fully supervised scenarios, employ strategies such as prototypical learning [2], contrastive learning [1], generative modeling [6], multi-resolution learning [19], as well as regularization techniques [10]. Additionally, few-shot settings have been explored through multi-domain pre-training [30], or leveraging existing ImageNet-pre-trained foundation models [16]. A few works aimed to achieve broader generalization abilities, as demonstrated by GU2Net [31, 32], and UniverDetect [15]. Nevertheless, these approaches face significant limitations due to pre-training on the constrained size of landmark detection datasets, which typically comprise only a few hundred images. Furthermore, these methods fail to fully exploit the potential of publicly available foundation models, leading to suboptimal generalization performance on novel tasks (See Table 3).

Foundation models offer strong generalization for seg-
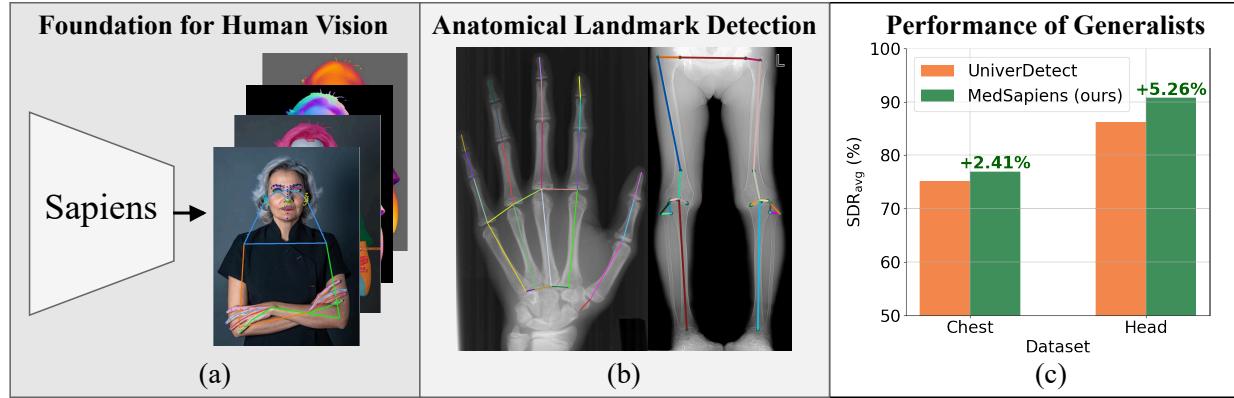
---

[1]Equal contribution.

Figure 1: (a) Sapiens [13], a foundation model devised for vision-centric tasks such as pose estimation. (b) Anatomical landmark detection shares a hierarchical structure with human-centric tasks. (c) By adapting Sapiens for anatomical landmark detection, MedSapiens surpasses the existing SOTA generalist model UniverDetect [15].

mentation and classification and have gained widespread adoption in both generic computer vision and medical imaging [4, 11, 14, 24, 27–29]. However, adapting foundation models for anatomical landmark detection is not as straightforward. Unlike segmentation and classification, which have well-established pre-trained backbones, landmark detection lacks direct counterparts, making transfer learning less effective.

Sapiens [13], a recent foundation model pre-trained on over *300 million* in-the-wild images, sets a new scale benchmark for pose estimation. Pose estimation shares conceptual similarities with anatomical landmark detection, as they aim to capture spatial hierarchies and contextual relationships between key points.

To this end, we demonstrate that *efficiently fine-tuning foundation models devised for pose estimation can achieve state-of-the-art (SOTA) performance across diverse medical imaging landmark detection tasks*. Building on this insight, we introduce **MedSapiens**, a foundation model trained on diverse anatomical landmarks from multiple medical imaging datasets, achieving a new SOTA.

To demonstrate the generalization capability of **MedSapiens**, we evaluate it on few-shot samples from an unseen novel task, where *MedSapiens outperforms the SOTA, achieving a 2.69% improvement in the average success detection rate*. While this may not be entirely surprising given that existing foundation models devised for pose estimation are inherently designed to localize anatomical landmarks in human figures, it has been largely overlooked in the literature, with no prior work explicitly exploring this direction. Our contributions can be summarized as follows:

- We highlight the previously unexplored potential of leveraging human-centric foundation models for anatomical landmark detection in medical imaging.

- We introduce **MedSapiens**, a foundation model for anatomical landmark detection, which outperforms the existing SOTA generalist model with improvements of up to **5.26%** in the average success detection rate. When further specialized for a specific task, MedSapiens surpasses the SOTA specialist model, achieving improvements of up to **21.81%** in the average success detection rate.

- Finally, to assess the generalization of our model, we evaluate **MedSapiens** on a **novel unseen task**, achieving improvements of up to **2.69%** over the few-shot SOTA, demonstrating superior cross-task adaptability and generalization capabilities.

2 Methodology

2.1 Preliminaries

Problem Formulation. Anatomical landmark detection is fundamental in medical imaging, playing a crucial role in diagnosis, surgical planning, and treatment monitoring. The objective is to accurately localize a predefined set of anatomical landmarks within a given medical image. Given an image $I \in \mathbb{R}^{H \times W}$, the task is formulated as predicting N landmarks, represented as a set of coordinate pairs $\{(x_i, y_i)\}_{i=1}^{N}$, where (x_i, y_i) denote the spatial location of the i-th landmark. The problem is inherently challenging due to the limited availability of annotated medical datasets, substantial anatomical variability across imaging tasks, and the need to model complex spatial relationships among landmarks.
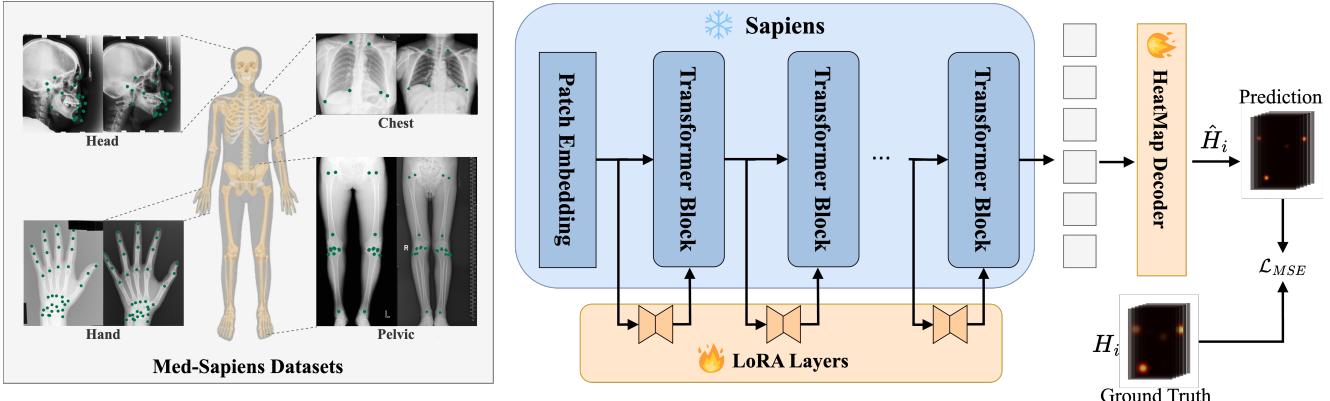
Figure 2: Overall framework for MedSapiens.

MedSapiens Landmark Datasets. To facilitate robust pre-training of foundation models, we have harmonized a diverse set of publicly available medical landmark datasets spanning different anatomical regions, comprising four datasets with 47,847 annotated landmark points across 1,778 images. The Head X-ray dataset [22, 23] consists of 400 lateral cephalograms with 19 annotated landmarks and a resolution of 0.1 mm, commonly used in orthodontics. The Hand X-ray dataset includes 909 radiographs with 37 annotated landmarks, normalized to a wrist width of 50 mm introduced in [5] and annotated by [17]. The Chest X-ray dataset [31] contains 279 images from the China Set, excluding abnormal lungs, with six manually annotated landmarks defining lung boundaries. The BMPLE dataset consists of 190 leg X-rays with 26 annotated landmarks for lower extremity analysis, introduced in CC2D [26].

2.2 Overall Framework

We propose MedSapiens, a framework designed to leverage large-scale human-centric pretraining for anatomical landmark detection in medical imaging. Built upon the Sapiens model [13], a vision transformer (ViT) trained for human-centric tasks, we adapt it to localize anatomical landmarks across diverse medical imaging modalities. To bridge the gap between human-centric tasks and domain-specific anatomical structures, we pre-train the model on a harmonized collection of public medical landmark datasets, ensuring broad anatomical representation. Given the disparity in dataset scales and the limited availability of annotated landmark detection datasets, full fine-tuning is computationally expensive and prone to overfitting. To allow the model to retain the spatial hierarchies and contextual relationships from large-scale human-centric pretraining, while effectively adapting to the constraints of medical imaging, we employ Low-Rank Adaptation (LoRA) [7], a parameter-efficient fine-tuning approach that injects trainable low-rank updates into the transformer's self-attention and projection

layers while preserving the pre-trained backbone. Finally, we incorporate a heatmap-based prediction head to refine the extracted features and generate spatial confidence maps, enabling precise and robust landmark localization.

2.3 Heatmap-Based Decoding

The Heatmap Head in MedSapiens serves as the decoding mechanism that translates feature representations from the backbone into spatial confidence maps, enabling precise localization of anatomical landmarks.

Given feature maps $F \in \mathbb{R}^{h \times w \times C}$ from the transformer backbone, the Heatmap Head applies a series of transposed convolutional layers to progressively upsample the features, followed by 1×1 convolutions for refinement. The output is a set of heatmaps $\hat{H} \in \mathbb{R}^{h' \times w' \times N}$, where N is the number of anatomical landmarks. Each heatmap represents the spatial confidence of a specific landmark.

To supervise the model, *Keypoint Mean Squared Error (MSE)* loss is employed, which compares the predicted heatmaps $\hat{H}$ to the ground-truth heatmaps H. Each ground-truth heatmap is generated by placing a Gaussian kernel centered at the true landmark location. The MSE loss is then defined as:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} \|\hat{H}_i - H_i\|_2^2,$$

where $\hat{H}_i$ and H_i represent the predicted and ground-truth heatmaps for the i-th landmark, respectively.

3 Experiments

Baselines. Given the extensive methods proposed over the past years for anatomical landmark detection, we focus on comparing MedSapiens with the two most recent SOTA approaches: NFDP [9] and UniverDetect [15]. NFDP

Table 1: Comparisons with SOTA methods on the Head and Hand datasets.

	Hand Dataset					Head Dataset				
	SDR (%)↑			SDR$_{avg}$	MRE(mm)↓	SDR (%)↑			SDR$_{avg}$	MRE(mm)↓
	2 mm	4 mm	10 mm			2 mm	3 mm	4 mm		
Generalist model										
UniverDetect [15]	95.76	99.30	99.91	98.32	$0.71_{\pm 1.78}$	75.87	88.35	94.59	86.27	$1.55_{\pm 1.74}$
MedSapiens (ours)	95.87	99.70	100.0	**98.52**	$\mathbf{0.664_{\pm .110}}$	82.29	92.80	97.33	**90.81**	$\mathbf{1.275_{\pm .285}}$
Specialist model										
NFDP [9]	95.11	99.44	99.97	98.17	$0.673_{\pm .152}$	82.00	92.44	96.99	90.48	$1.245_{\pm .276}$
MedSapiens w/ LoRA (ours)	96.32	99.74	100.0	**98.68**	$\mathbf{0.638_{\pm .106}}$	83.14	92.88	97.09	**91.04**	$\mathbf{1.244_{\pm .276}}$

employs generative priors to model anatomical distributions [9], while UniverDetect leverages a category-agnostic framework to ensure cross-domain generalizability [15]. Since UniverDetect [15] is not publicly available, we follow its training, validation, and testing dataset splits, which are predefined for each dataset, and reproduce NFDP and MedSapiens to ensure a fair comparison. To evaluate generalization, we test MedSapiens in few-shot settings on a novel unseen dataset during training, LDTeeth [21], comparing it to diverse baselines, including GU2Net [31], FM-OSD [16], and GeoSapiens [21]. For GU2Net [31], we expand training by incorporating the unseen task dataset alongside its original head, hand, and chest datasets. For FM-OSD [16], we adapt its one-shot framework for few-shot learning by increasing the training data with additional samples and using a multi-sample approach for template matching.

Implementation Details. The **generalist** MedSapiens leverages the Sapiens 0.3B Vision Transformer as the backbone, pre-trained on over 300 million in the wild–images [13]. Fine-tuning is achieved using Low-Rank Adaptation, applied to the qkv and projection layers of the transformer with a rank of 4. The model is trained using the AdamW optimizer with an initial learning rate of 5×10^{-4} and a layer-wise decay rate of 0.85. To ensure robustness, random flips, photometric distortions, and coarse dropout are incorporated as data augmentations. MedSapiens employs a top-down pose estimation pipeline for training and evaluation, assuming the bounding box encompasses the full image. To adapt MedSapiens to different datasets as a **specialist**, we further fine-tune it on each specific dataset using LoRA [7], referred to as **MedSapiens w/ LoRA** in our analysis. LoRA is selected because it imposes minimal architectural changes, provides parameter-efficient adaptation, and preserves the generalist model's representations while enabling effective specialization.

Evaluation Metrics. We strictly follow existing baselines [9, 15] for the evaluation protocol. Performance is evaluated using Mean Radial Error (MRE) and Successful Detection Rate (SDR) at various thresholds on the original scale. We follow prior work's normalization strategies for datasets without physical spacing information (e.g., hand dataset wrist width is assumed to be 50 mm) [9, 15].
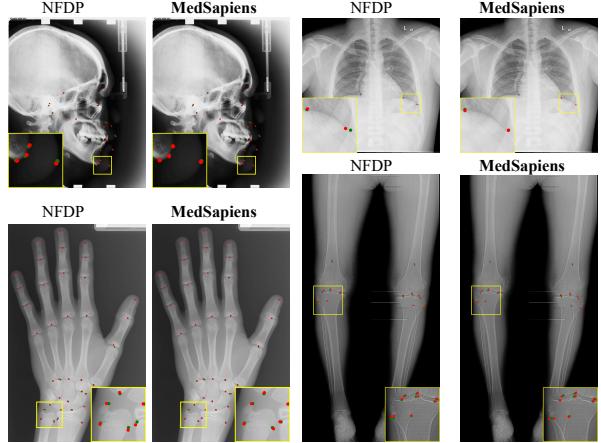
3.1 Quantitative Results

Figure 3: Qualitative results of NFDP and MedSapiens on the four testing datasets, where red points indicate predicted landmarks and green points represent ground-truth labels.

Table 1 and Table 2 show comparisons of MedSapiens with SOTA methods on the Hand, Head, Legs, and Chest datasets. The results demonstrate that MedSapiens outperforms existing methods. Compared to the generalist model, UniverDetect [15], MedSapiens achieves relative **improvements** of 0.20%, **5.26%**, and **2.41%** on the average SDR across datasets, respectively for the hand, head, and chest datasets. Specifically, MedSapiens surpasses UniverDetect with relative **improvements** of 0.11%, **8.46%**, and **6.28%** on the strict metric of 2mm, respectively for the hand, head, and chest datasets, offering precise anatomical landmark detection. Given that UniverDetect [15] does not release its model weights by the time of submission and is trained and evaluated on privately held datasets, we were unable to compare it to the publicly available Legs dataset.

While a generalist model offers broad transferability across diverse downstream tasks, task-level specialization remains desirable. To obtain such **specialist** models, we fine-tune the generalist MedSapiens on each target dataset using LoRA [7], denoted as MedSapiens w/ LoRA. We demonstrate that MedSapiens w/ LoRA can achieve SOTA

Table 2: Comparisons with SOTA methods on the Chest and Legs datasets. ‡ Results were not reported by the authors, and the model is not publicly available. We follow UniverDetect [15] testing splits, which are predefined for each dataset, and reproduce NFDP [9] to ensure a fair comparison.

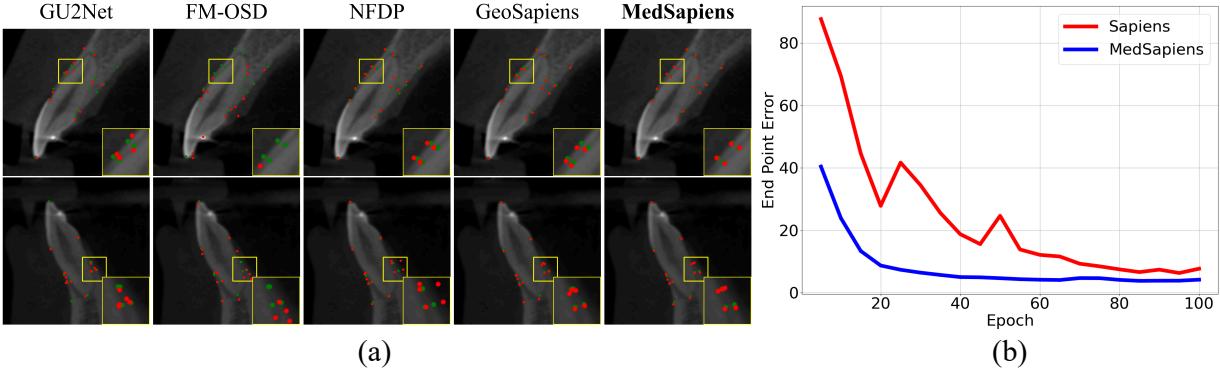|	Legs Dataset					Chest Dataset				
	SDR (%)↑			SDR_{avg}	MRE(mm)↓	SDR (%)↑			SDR_{avg}	MRE(px)↓
	2 mm	4 mm	10 mm			3 px	6 px	9 px		
					Generalist model					
UniverDetect [15]	‡	‡	‡	‡	‡	50.81	82.52	92.27	75.19	$4.06_{\pm 3.73}$
MedSapiens (ours)	48.08	85.08	97.15	76.77	$2.69_{\pm .555}$	54.00	83.67	93.33	**77.00**	$\mathbf{3.715_{\pm 1.31}}$
					Specialist model					
NFDP [9]	50.69	84.62	96.92	75.89	$2.685_{\pm .617}$	31.00	70.00	89.67	63.55	$5.13_{\pm 1.44}$
MedSapiens w/ LoRA (ours)	53.54	87.77	97.54	**79.61**	$\mathbf{2.509_{\pm .556}}$	51.67	82.33	93.67	**77.41**	$\mathbf{3.734_{\pm 1.24}}$

(a)   (b)

Figure 4: Qualitative comparison of our MedSapiens method with existing baselines: (a) Results on dental images, where red points indicate predicted landmarks and green points represent ground-truth labels. Our MedSapiens achieves superior alignment with the ground truth compared to other methods. (b) Convergence analysis of MedSapiens vs. Sapiens in terms of end-point error across epochs. Our method exhibits faster and more stable convergence with lower error.

Table 3: Comparison with SOTA Few-shot methods on the Downstream Task.

	SDR (%)↑			SDR_{avg}	MRE(px)↓
	0.5mm	1mm	2mm		
GU2Net [31]	45.21	64.12	81.90	63.74	1.312
FM-OSD [16]	32.95	55.79	77.62	55.45	1.520
NFDP [9]	55.01	80.40	92.09	75.83	0.825
		Taking a Pose			
Sapiens (Baseline)	60.07	82.72	93.13	78.64	0.766
GeoSapiens [21]	63.19	84.14	93.36	80.23	0.747
Med-Sapiens (ours)	65.66	87.31	94.21	**82.39**	**0.724**

performance compared to the specialist model, NFDP [9], without requiring any prior assumptions or distribution knowledge over the dataset. Specifically, MedSapiens w/ LoRA surpasses NFDP [9] with relative **improvements** of 0.52%, 0.62%, **4.90%**, and **21.81%** on the average SDR across datasets, respectively for the hand, head, legs, and chest datasets. Additionally, it achieves relative **improvements** of 1.27%, 1.39%, **5.62%**, and **66.68%** on the strict metric of 2mm, respectively for the hand, head, legs, and chest datasets, offering precise anatomical landmark detec-

tion. The substantial improvements observed in the chest X-ray dataset may be attributed to its high anatomical variability, encompassing differences across male and female subjects. Traditional models may struggle due to limited distributional and generative priors, whereas our model demonstrates robustness in adapting to these variations, ensuring more reliable anatomical landmark detection. We present qualitative results in Fig. 3 across the four datasets, comparing our method to NFDP.

3.2 Few-Shot Novel Task: Teeth Landmarks

We evaluate MedSapiens on a novel task of detecting dental landmarks in Cone-Beam Computed Tomography (CBCT) images under few-shot conditions. Accurate teeth landmark detection, targeting key features such as the cementoenamel junction, physiological crest, and root apex, is vital for precise clinical measurements, which play a crucial role in diagnosis and risk assessment [25]. Dentists typically employ strict criteria for detecting landmarks, using a threshold of 0.5 mm for assessing patients with dental diseases due to the necessity for high precision in clinical eval-

uations [25]. We adopt the LDTeeth dataset, introduced in GeoSapiens [21], and follow the same experimental protocol, including its official patient-wise train/test split. Specifically, we use 3 patients for training and 19 patients for testing.

Experimental results on LDTeeth demonstrate that MedSapiens establishes a new SOTA in few-shot dental landmark detection. Compared to NFDP [9], MedSapiens improves SDR_{avg} by **8.65%**. To ensure a fair capacity-matched comparison, we additionally evaluate GeoSapiens [21], which—similar to MedSapiens—employs a parameter-efficient tuning strategy with the same number of trainable parameters (**24M**). Without relying on complex task-specific geometric loss functions or additional optimization strategies, MedSapiens further achieves a **+2.69%** relative gain in SDR_{avg} over GeoSapiens. Moreover, MedSapiens attains higher precision under the strict 0.5 mm clinical threshold (**65.66% vs. 63.19%**), highlighting its stronger anatomical localization capability in clinically relevant scenarios.

4 Conclusion

MedSapiens highlights the potential of leveraging large-scale pre-trained models for anatomical landmark detection. By integrating LoRA fine-tuning with a diverse set of anatomical landmark datasets, it outperforms both generalist and specialist benchmarks, achieving state-of-the-art performance. Future work will explore multimodal adaptation and scaling model parameters as larger datasets become available.

References

[1] Amirhossein, R. Hassan, X. Y. S. Soorena, and Rasoulian. Towards multi-modal anatomical landmark detection for ultrasound-guided brain tumor resection with contrastive learning. In Anant, M. Parvin, S. Septimiu, D. James, S.-M. Tanveer, T. R. G. Hayit, and Madabhushi, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 668–678. Springer Nature Switzerland, 2023.

[2] Chong, M. Lanzhuju, Y. Tong, Z. Min, S. Dinggang, C. Z. W. Han, and Wang. Cephalometric landmark detection across ages with prototypical network. In Qi, F. Aasa, G. Stamatia, G. Ben, L. Karim, S. J. A. L. M. George, and Dou, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 155–165. Springer Nature Switzerland, 2024.

[3] B. Demir, L. Tian, T. H. Greer, R. Kwitt, F.-X. Vialard, R. S. J. Estepar, S. Bouix, R. J. Rushmore, E. Ebrahim, and M. Niethammer. multigradicon: A foundation model for multimodal medical image registration. *arXiv preprint arXiv:2408.00221*, 2024.

[4] M. Elbatel, K. Liu, Y. Yang, and X. Li. Fd-sos: Vision-language open-set detectors for bone fenestration and dehiscence detection from intraoral images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 629–639. Springer, 2024.

[5] A. Gertych, A. Zhang, J. Sayre, S. Pospiech-Kurkowska, and H. Huang. Bone age assessment of children using a digital hand atlas. *Computerized Medical Imaging and Graphics*, 31(4):322–331, 2007. Computer-aided Diagnosis (CAD) and Image-guided Decision Support.

[6] A. Hadzic, L. Bogensperger, S. J. Joham, and M. Urschler. Synthetic augmentation for anatomical landmark localization using ddpms. In V. Fernandez, J. M. Wolterink, D. Wiesner, S. Remedios, L. Zuo, and A. Casamitjana, editors, *Simulation and Synthesis in Medical Imaging*, pages 1–12, Cham, 2025. Springer Nature Switzerland.

[7] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

[8] T. Huang, J. Shi, J. Li, J. Wang, J. Du, and J. Shi. Involution transformer based u-net for landmark detection in ultrasound images for diagnosis of infantile ddh. *IEEE Journal of Biomedical and Health Informatics*, 28:4797–4809, 2024.

[9] Z. Huang, R. Zhao, F. H. Leung, S. Banerjee, K. M. Lam, Y. P. Zheng, and S. H. Ling. Landmark localization from medical images with generative distribution prior. *IEEE Transactions on Medical Imaging*, 43:2679–2692, 2024.

[10] Jiahao, H. Wenjian, D. Pei, Q. Z. W. Yao, and Chen. Learnable skeleton-based medical landmark estimation with graph sparsity and fiedler regularizations. In Qi, F. Aasa, G. Stamatia, G. Ben, L. Karim, S. J. A. L. M. George, and Dou, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 599–609. Springer Nature Switzerland, 2024.

[11] Q. Jiang, J. Huo, X. Chen, Y. Xiong, Z. Zeng, Y. Chen, T. Ren, J. Yu, and L. Zhang. Detect anything via next point prediction, 2025.

[12] Jing, J. Ge, L. Juncheng, W. Jun, D. Jun, S. J. H. Tianxiang, and Shi. Topological gcn for improving detection of hip landmarks from b-mode ultrasound images. In Qi, F. Aasa, G. Stamatia, G. Ben, L. Karim, S. J. A. L. M. George, and Dou, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 692–701. Springer Nature Switzerland, 2024.

[13] R. Khirodkar, T. Bagautdinov, J. Martinez, S. Zhaoen, A. James, P. Selednik, S. Anderson, and S. Saito. Sapiens: Foundation for human vision models. *ECCV*, 2024.

[14] T. Koleilat, H. Asgariandehkordi, H. Rivaz, and Y. Xiao. Medclip-sam: Bridging text and image towards universal medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 643–653, Cham, 2024. Springer Nature Switzerland.

[15] C. Lu, G. Yang, X. Qiao, W. Chen, and Q. Zeng. Univerdetect: Universal landmark detection method for multidomain x-ray images. *Neurocomputing*, 600, 10 2024.

[16] J. Miao, C. Chen, K. Zhang, J. Chuai, Q. Li, and P.-A. Heng. FM-OSD: Foundation Model-Enabled One-Shot Detection of Anatomical Landmarks . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15011. Springer Nature Switzerland, October 2024.

[17] C. Payer, D. Štern, H. Bischof, and M. Urschler. Integrating spatial configuration into heatmap regression based cnns for landmark localization. *Medical Image Analysis*, 54:207–219, 2019.

[18] Ruize, L. Yaoqian, S. Weixin, Q. Jing, H. P.-A. P. Jialun, and Cui. Depth-driven geometric prompt learning for laparoscopic liver landmark detection. In Qi, F. Aasa, G. Stamatia, G. Ben, L. Karim, S. J. A. L. M. George, and Dou, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 154–164. Springer Nature Switzerland, 2024.

[19] Serie, C. J.-H. V. Thanaporn, and Ma. Anatomical landmark detection using a multiresolution learning approach with a hybrid transformer-cnn model. In Anant, M. Parvin, S. Septimiu, D. James, S.-M. Tanveer, T. R. G. Hayit, and Madabhushi, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 433–443. Springer Nature Switzerland, 2023.

[20] L. Tian, H. Greer, R. Kwitt, F.-X. Vialard, R. S. J. Estepar, S. Bouix, R. Rushmore, and M. Niethammer. unigradicon: A foundation model for medical image registration. *arXiv preprint arXiv:2403.05780*, 2024.

[21] A. Wang, M. Elbatel, K. Liu, L. Lin, M. Lan, Y. Yang, and X. Li. Geometric-Guided Few-Shot Dental Landmark Detection with Human-Centric Foundation Model . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15964. Springer Nature Switzerland, September 2025.

[22] C.-W. Wang, C.-T. Huang, et al. Evaluation and comparison of anatomical landmark detection methods for cephalometric x-ray images: A grand challenge. *IEEE Transactions on Medical Imaging*, 34(9):1890–1900, 2015.

[23] C.-W. Wang, C.-T. Huang, J.-H. Lee, C.-H. Li, S.-W. Chang, M.-J. Siao, T.-M. Lai, B. Ibragimov, T. Vrtovec, O. Ronneberger, P. Fischer, T. F. Cootes, and C. Lindner. A benchmark for comparison of dental radiography analysis algorithms. *Medical Image Analysis*, 31:63–76, 2016.

[24] Q. Wu, Y. Zhang, and M. Elbatel. Self-prompting large vision models for few-shot medical image segmentation. In *MICCAI Workshop on Domain Adaptation and Representation Transfer*, pages 156–167. Springer, 2023.

[25] A. Yagci, I. Veli, T. Uysal, F. I. Ucar, T. Ozer, and S. Enhos. Dehiscence and fenestration in skeletal class i, ii, and iii malocclusions assessed with cone-beam computed tomography. *The Angle Orthodontist*, 82(1):67–74, Jan. 2012. Epub 2011 Jun 22.

[26] Q. Yao, J. Wang, Y. Sun, Q. Quan, H. Zhu, and S. K. Zhou. Relative distance matters for one-shot landmark detection, 2022.

[27] K. Zhang and D. Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023.

[28] Y. Zhang, J. Gao, et al. Text-guided foundation model adaptation for pathological image classification. In *MICCAI*, 2023.

[29] T. Zhao, Y. Gu, et al. A foundation model for joint segmentation, detection, and recognition of biomedical objects across nine modalities. *Nature Methods*, 2024.

[30] H. Zhu, Q. Quan, Q. Yao, Z. Liu, and S. K. Zhou. Uod: Universal one-shot detection of anatomical landmarks. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, volume 14220 LNCS, pages 24–34. Springer Science and Business Media Deutschland GmbH, 2023.

[31] H. Zhu, Q. Yao, L. Xiao, and S. K. Zhou. *You only Learn Once: Universal Anatomical Landmark Detection*, page 85–95. Springer International Publishing, 2021.

[32] H. Zhu, Q. Yao, L. Xiao, and S. K. Zhou. Learning to localize cross-anatomy landmarks in x-ray images with a universal model. *BME frontiers*, 2022:9765095, 2022.