AutoLand: A Data-Efficient Automated Agentic Workflow for Universal Medical Landmark Detection

Ziyi Li[1], Jiarui Li[1], Yulong Shi[1], Xue Jiang[2], Shouliang Qi[1], Ting Liu[2]✉, Lin Qi[1]✉, and Wei Qian[1]

[1] Northeastern University, Shenyang, China
[2] The First Hospital of China Medical University, Shenyang, China

Abstract. Medical landmark detection varies significantly across modalities and clinical protocols, posing unique challenges for developing universally applicable models. Conventional approaches typically entail training bespoke architectures per dataset or relying on general-purpose heuristics with fixed rules. However, these methods suffer from limited scalability and lack the semantic reasoning required to adapt to complex anatomical variations. To address these limitations, we propose AutoLand, a data-efficient autonomous workflow for landmark detection in medical images, leveraging a Vision Language Model (VLM) as a medical analysis agent. Specifically, AutoLand integrates a Physics Fingerprint Extraction module to extract precise physical characteristics, ensuring the agent aligns with modality-specific data properties. Subsequently, a Medical Deep Learning Library is queried during reasoning to retrieve specialized network architectures and optimization protocols, enabling the formulation of expert-level training strategies with only a few samples. This workflow enhances robustness and adaptability while reducing manual intervention. Experiments on three datasets show that AutoLand significantly outperforms state-of-the-art methods. Moreover, ablation studies confirm that our agentic paradigm effectively overcomes the limitations of fixed baselines through semantic reasoning. The source code is available at https://github.com/ChiZai123456/AutoLand.

Keywords: Medical landmark detection · Vision language model · Automated agentic workflow · Data-Efficient learning

1 Introduction

Medical landmark detection aims to localize anatomically meaningful key points in medical images and serves as a fundamental step in many clinical applications, such as treatment planning and surgical interventions [25]. Manual landmark annotation is time-consuming and subject to inter-observer variability, motivating automated methods.

✉ Co-corresponding authors: qilin@bmie.neu.edu.cn (L. Qi), 20082180@cmu.edu.cn (T. Liu)

Current approaches can be categorized into task-specific methods [15,17,23] and general-purpose methods [5,6,9,11,24,26]. Task-specific methods achieve strong performance by designing bespoke architectures for a particular modality or protocol, but they often require non-trivial redesign of the model when transferred to new scenarios. In contrast, general-purpose methods attempt to infer configurations from dataset statistics or heuristic rules, yet such fixed heuristics lack semantic reasoning, making them suboptimal under complex anatomical variations or few-sample scenarios.

To address these limitations, we introduce AutoLand, which leverages a Vision Language Model (VLM) as a medical analysis agent to formulate dataset-specific training strategies automatically. By interpreting both the visual context and physics fingerprints of the input data, the agent can adapt its decisions to modality-specific properties. Moreover, during reasoning, it queries a specialized medical Deep Learning Library to retrieve valid network architectures and training strategies, enabling robust strategy instantiation without manual tuning.

The contributions of our work are summarized as follows: (1) We propose AutoLand, a data-efficient automated agentic workflow that covers model configuration, preprocessing, training, and validation for landmark detection on new datasets. (2) We leverage the agent's semantic reasoning capability to reduce data dependency, generating effective training strategies from only a few annotated samples. (3) We demonstrate robust generalizability of the proposed workflow and validate it across diverse anatomical targets and imaging modalities.

2 Method

Fig. 1 illustrates our data-efficient autonomous landmark detection workflow named AutoLand, which employs a VLM as a medical analysis agent. By leveraging the agent's semantic reasoning capabilities to interpret the physics fingerprints and visual contexts of input datasets, our workflow can formulate the training strategy autonomously. Consequently, our approach achieves universal landmark detection across diverse datasets without requiring manual tuning or redesigning network architecture.

2.1 Physics Fingerprint Extraction

The first step of our approach is to enable the medical analysis agent to understand the input medical images. However, feeding raw medical images to the agent directly limits information fidelity due to resolution constraints. Meanwhile, the determination of training strategies relies on precise physical parameters of input medical images, which are obscured in the visual input of agent [21]. To address these challenges, we extract physics parameters as fingerprints from the input medical images and feed them to the agent alongside the raw images. This ensures that the agent can perceive the dataset comprehensively, combining

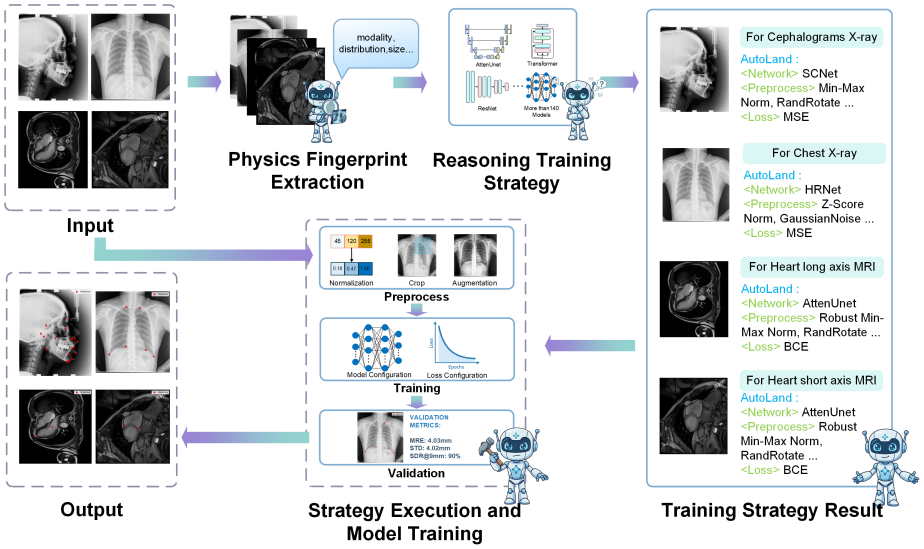

Fig. 1. Overview of our approach. Physics fingerprints are extracted from input data to represent dataset characteristics. Then, the medical analysis agent generates optimal training strategies. Finally, the automated strategy execution module builds the workflow for landmark detection.

visual context with precise physics fingerprints, thereby supporting data-efficient strategy reasoning under limited annotations.

Specifically, we define the input medical images as $\mathbf{D} = \{I_n\}_{n=1}^{K}$, where I_n denotes the n-th medical image and K denotes the number of input images. For each image, we define an extraction function as $E(I_n) = \{\mathbf{v}_n, \mathbf{s}_n, \mu_n, m_n\}$, where $\mathbf{v}_n$ denotes voxel spacing, $\mathbf{s}_n$ denotes spatial dimensions, μ_n denotes mean intensity, and m_n denotes max intensity. To eliminate the impact of outliers, we define the median voxel spacing $\tilde{\mathbf{v}}$, median spatial dimension $\tilde{\mathbf{s}}$, and mean intensity $\bar{\mu}$ as $\tilde{\mathbf{v}} = \text{median}(\{\mathbf{v}_n\}_{n=1}^{K})$, $\tilde{\mathbf{s}} = \text{median}(\{\mathbf{s}_n\}_{n=1}^{K})$, and $\bar{\mu} = \frac{1}{K}\sum_{n=1}^{K}\mu_n$. Then, to infer the bit depth of the input images, we define $\omega_{bit} = 16$ for high bit-depth images if $m_n > 255$, and $\omega_{bit} = 8$ for low bit-depth images otherwise. Finally, the physics fingerprints can be described as $\mathbf{f}_{phy} = [\tilde{\mathbf{v}}, \tilde{\mathbf{s}}, \bar{\mu}, \omega_{bit}]$.

2.2 Reasoning Training Strategy from Medical DL Library

The second step of our approach is to enable the medical analysis agent to formulate executable training strategies for different datasets. However, general-purpose agents lack specialized knowledge about domain-specific medical deep learning methods and training protocols. Consequently, the agent may generate strategies that are theoretically reasonable but actually invalid, leading to inefficient trial-and-error under limited annotations [13,18]. To address these challenges, we constructed a specialized medical Deep Learning Library (DL Library)

as $\mathbf{M}_{lib}$, which integrates over 140 network architectures from MONAI and preprocessing protocols from SOTA methods [2,3,4,7,9,16]. Serving as a knowledge base, this DL Library enables the agent to determine training strategies for specific datasets, improving data efficiency by constraining reasoning to executable options.

Specifically, to effectively utilize the DL Library for formulating training strategies, we employ a multi-modal prompt strategy. We define $\mathbf{V}$ as visual tokens encoded from the input image. Then, we define a tokenization mapping function $\Phi(\cdot)$ that serializes textual descriptions into a sequence of text tokens. We apply the mapping function $\Phi(\cdot)$ to transform the physics fingerprint $\mathbf{f}_{phy}$ and the DL Library $\mathbf{M}_{lib}$ into their respective text tokens: $\mathbf{T}_{phy} = \Phi(\mathbf{f}_{phy})$ and $\mathbf{T}_{lib} = \Phi(\mathbf{M}_{lib})$, where $\mathbf{T}_{phy}$ denotes text tokens that describe the physics fingerprint and $\mathbf{T}_{lib}$ denotes text tokens that describe the DL Library. Finally, the multi-modal prompt $\mathbf{P}$ can be described as

$$\mathbf{P} = \begin{bmatrix} \texttt{<System Role>} & : \mathbf{I}_{sys} \\ \texttt{<Visual Input>} & : \mathbf{V} \\ \texttt{<Physics Fingerprint>} & : \mathbf{T}_{phy} \\ \texttt{<Medical DL Library>} & : \mathbf{T}_{lib} \\ \texttt{<Instruction>} & : \mathbf{I}_{task} \end{bmatrix}, \tag{1}$$

Under the guidance of this prompt, the agent aligns the input images with the physics fingerprints and performs reasoning based on the DL Library to formulate the training strategy. The training strategy $\mathbf{S}$ is defined as

$$\mathbf{S} = \begin{bmatrix} \texttt{<Net Arch>} & : f_{net} \in \mathbf{M}_{lib} \\ \texttt{<Pre-process>} & : \rho_{trans} \\ \texttt{<Loss Func>} & : \mathcal{L} \\ \texttt{<Hyper-param>} & : \mathbf{h} \end{bmatrix}, \tag{2}$$

where f_{net} denotes the selected network architecture from the medical DL Library, ρ_{trans} denotes the preprocessing protocol, $\mathcal{L}$ denotes the loss function, and $\mathbf{h}$ denotes the optimization hyperparameters such as learning rate and batch size.

2.3 Strategy Execution and Model Training

The final step of our approach is to instantiate the training strategy into an executable training routine. In this section, we present model training as preprocessing, training, and validation.

In the preprocessing phase, we instantiate the preprocessing stage based on the ρ_{trans} in $\mathbf{S}$. Specifically, ρ_{trans} contains transformation identifiers such as "normalization", "crop", and "augmentation". We map these identifiers to an ordered sequence of executable transformations. In the training phase, we instantiate the network architecture by mapping the identifier f_{net} to its corresponding class definition in the medical DL Library. The executable model is constructed as $\mathbf{M} = \text{Instantiate}(f_{net}, \mathbf{M}_{lib})$. The model parameters are optimized according to the hyperparameters $\mathbf{h}$ and the loss function $\mathcal{L}$. To address invalid strategies

generated by the agent, we design a fallback mechanism that defaults to a standard UNet with basic preprocessing upon instantiation failure. In the validation phase, we evaluate the instantiated model on a hold-out validation set to verify the effectiveness of the instantiated strategy. The instantiated model performs inference to predict landmark coordinates, and the validation results are used to assess generalization.

3 Experiments

3.1 Datasets and Evaluation Metrics

We validated our approach on multimodal medical image datasets to demonstrate its effectiveness and generalizability. The Cephalometric X-ray Dataset from the ISBI 2015 Challenge includes 400 images from 400 patients [20], with 19 expert-annotated landmarks per image. The Chest X-ray Dataset from a public Kaggle repository includes 279 images from 279 patients [19], with 6 expert-annotated landmarks per image. The Cardiac MRI dataset consists of 5,293 LAX and 15,024 SAX slices collected from 144 patients at a hospital, with 2 expert-annotated landmarks per image. This study was approved by the local Institutional Review Board (IRB). To evaluate the data-efficient performance of our workflow in few-sample scenarios, we partitioned all datasets into training, validation, and test sets in a 1:1:8 ratio at the patient level for all experiments.

We used standard metrics to evaluate the detection performance: Mean Radial Error (MRE) [16] to quantify accuracy, Standard Deviation (STD) [22] to assess stability, and Success Detection Rate (SDR) [12] at thresholds of $1--10mm$ to measure the percentage of successful detections.

3.2 Implementations

All experiments leveraged a single RTX 4090 GPU system with 24 GB memory, running PyTorch 2.9.1, Python 3.10, Ubuntu 22.04, and CUDA 12.8. We used Qwen/Qwen3-VL-235B-A22B-Thinking as the medical analysis agent throughout all experiments.

3.3 Comparisons with SOTA

Table 1 and Fig. 2 present the quantitative comparison and visual results, respectively. As shown, conventional methods such as Unet [1], Shape-Model [8], and RFRV [14] exhibit limited performance due to their fixed network architectures and lack of priors. Advanced methods such as NFDP [7] attempt to improve performance by leveraging specifically designed normalizing flow models. NFDP achieves competitive performance on the Chest X-rays dataset but suffers significant degradation on the Cardiac MRI dataset, indicating poor generalization across modalities. Similarly, universal methods like GU2Net [27] and UniverDetect [16] typically require extensive supervision, leading to suboptimal

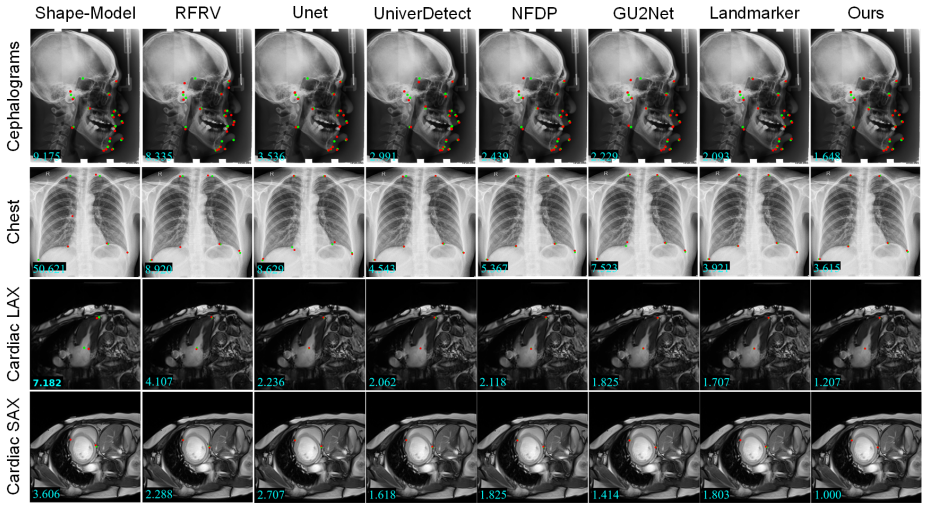

Fig. 2. Visualization of the comparative experiments. The predicted landmarks are red dots, and the ground-truth are green dots. The MRE for each image is displayed in the bottom-left corner.

adaptation in our few-sample scenarios. For instance, both GU2Net and UniverDetect lag significantly behind our method, yielding errors of 2.90 mm and 3.70 mm compared to our results, respectively. These inconsistencies highlight the limitations of static training strategies.

In contrast, our approach achieves the best performance across different datasets by using agents to tailor customized training strategies aligned with domain specificity. Notably, our method reduces the MRE to 1.05 mm compared to 2.06 mm for the second-best GU2Net in the Cardiac MRI dataset, highlighting its efficacy in MRI data. Furthermore, our approach attains a 75.82% SDR within 2 mm in the Cephalograms dataset, exceeding the 45.29% achieved by Landmarker [10]. This performance is attributed to the agent's capability to formulate specific training strategies for different datasets based on physics fingerprints rather than relying on fixed protocols. While UniverDetect aims for universality, its performance plateau implies potential difficulties in adapting to the distinct physical properties of MRI versus X-ray. Conversely, our approach addresses these limitations and achieves the best performance across all datasets.

3.4 Ablation Studies

To analyze the influence of the Physics Fingerprint and the DL Library on the agent's strategy reasoning, we conduct an ablation study by removing each component. We summarize the resulting strategy configurations in Table 2 and report the quantitative performance in Table 3.

The Physics Fingerprint is essential for enabling the agent to perceive modality-specific physical characteristics and to derive appropriate preprocessing policies

Table 1. Comparisons of our approach against SOTAs. All dataset was partitioned into training, validation, and test in 1:1:8 to simulate few-sample scenarios. The best results are red, and the second-best results are blue.

Models	Cephalograms					Chest X-ray				
	MRE↓ (mm)	STD↓ (mm)	SDR_d(%)↑ 2 mm	3 mm	4 mm	MRE↓ (mm)	STD↓ (mm)	SDR_d(%)↑ 3 mm	6 mm	9 mm
Shape-Model[8]	5.96	17.55	36.18	55.65	69.07	16.91	21.72	11.39	31.23	41.77
RFRV[14]	5.73	11.96	38.91	56.17	70.31	14.3	22.34	16.93	36.41	49.97
Unet[1]	5.05	12.59	32.86	51.82	65.24	13.81	22.05	18.88	38.96	53.61
UniverDetect[16]	3.7	7.76	41.26	58.52	70.21	7.71	25.61	47.05	78.92	88.23
NFDP[7]	3.66	4.95	47.34	55.26	70.5	5.57	6.08	56.61	78.36	86.73
GU2Net[27]	2.9	3.99	46.94	64.21	76.21	5.61	22.62	50.81	80.08	87.39
Landmarker[10]	2.55	3.67	45.29	70.36	85.58	7.75	20.94	43.92	72.35	84.14
Ours	1.41	1.36	75.82	93.51	98.34	4.03	4.02	59.04	81.12	90.56
Models	Cardiac LAX					Cardiac SAX				
	MRE↓ (mm)	STD↓ (mm)	SDR_d(%)↑ 2 mm	2.5 mm	3 mm	MRE↓ (mm)	STD↓ (mm)	SDR_d(%)↑ 2 mm	2.5 mm	3 mm
Shape-Model[8]	14.86	18.3	20.82	32.01	40.17	15.71	17.11	20.11	30.17	39.81
RFRV[14]	4.16	4.11	51.34	65.13	75.31	4.11	4.6	51.45	65.03	74.9
Unet[1]	6.31	17.2	33.9	43.75	51.79	7.11	15.84	31.36	42.14	51.34
UniverDetect[16]	3.08	7.34	56.92	70.48	79.68	3.27	5.6	56.11	69.77	79.11
NFDP[7]	6.11	7.5	35.17	45.61	52.37	5.12	11.01	44.2	57.56	66.39
GU2Net[27]	2.06	2.11	84.12	89.98	92.62	2.64	3.31	82.38	89.98	91.11
Landmarker[10]	2.18	2.75	82.34	88.9	91.92	3.17	10.16	66.84	77.45	84.2
Ours	1.05	1.97	94.49	97.08	98.54	1.18	2.48	92.52	94.89	96.13

from fewer samples. When the fingerprint is removed, the agent tends to adopt conservative, less adaptive normalization choices, which limit robustness to intensity variability. This effect is consistently reflected in performance drops across datasets. For example, on the Cardiac LAX dataset, removing the fingerprint increases MRE from 1.05 mm to 2.56 mm and reduces SDR at 2 mm from 94.49% to 43.25%. On the Chest X-ray dataset, the same ablation increases MRE from 4.03 mm to 6.44 mm and reduces SDR at 3 mm from 59.04% to 21.48%. These results indicate that physical fingerprints play a key role in guiding the agent toward preprocessing strategies that align with each dataset.

The DL Library is critical for constraining the agent's reasoning within valid medical deep learning practices, especially for network selection and optimization design. When the DL Library is removed, the agent falls back to generic configurations that are easier to justify but less suited to the anatomical and optimization characteristics of each dataset. This ablation yields the most severe degradation among all variants. On the Cephalogram dataset, removing the DL Library increases MRE from 1.41 mm to 3.89 mm and reduces SDR at 2 mm from 75.82% to 25.05%. On the Cardiac LAX dataset, the same ablation increases MRE from 1.05 mm to 3.26 mm, indicating that domain knowledge is indispensable for identifying effective optimization targets and loss formulations under limited supervision.

Overall, the ablation results confirm the complementary roles of the two components. The Physics Fingerprint improves data perception and steers preprocessing toward modality-aware decisions, while the DL Library provides domain

Table 2. Ablation study on the reasoning training strategy. The table summarizes the specific training strategy formulated by the agent under different Variants.

Dataset	Variant	Network	Preprocessing	Optimization & Loss
Cephalogram	**w/o** fingerprint	SCNet	Resize, Rot, Standard Min-Max Norm, CLAHE	AdamW, MSE, Cosine, lr = 1×10^{-3}
	w/o DL Library	UNet	Resize, Rot, Flip, Robust Min-Max Norm, Bright, CLAHE	Adam, MSE, lr = 1×10^{-4}
	Ours(AutoLand)	SCNet	Resize, Rot, Flip, Robust Min-Max Norm, Bright, CLAHE	AdamW, MSE, Cosine, lr = 1×10^{-3}
Chest	**w/o** fingerprint	HRNet	Resize, Rot, Standard Min-Max Norm, CLAHE	AdamW, MSE, Cosine, lr = 1×10^{-3}
	w/o DL Library	UNet	Resize, Flip, Rot, Z-Score Norm, CLAHE	Adam, MSE, lr = 1×10^{-4}
	Ours(AutoLand)	HRNet	Resize, Rot, Flip, Z-Score Norm, Gaussian Noise, CLAHE	AdamW, MSE, Cosine, lr = 1×10^{-3}
Cardiac	**w/o** fingerprint	AttenUnet	Resize, Rot, Standard Min-Max, CLAHE	AdamW, BCE, Cosine, lr = 5×10^{-4}
	w/o DL Library	UNet	Resize, Rot, Flip, Robust Min-Max Norm, CLAHE	Adam, MSE, lr = 1×10^{-4}
	Ours(AutoLand)	AttenUnet	Resize, Rot, Flip, Robust Min-Max Norm, CLAHE	AdamW, BCE, Cosine, lr = 5×10^{-4}

Table 3. Quantitative ablation results.

Variant	Cephalograms					Chest X-ray				
	MRE↓ (mm)	STD↓ (mm)	SDR_d(%)↑ 2 mm	3 mm	4 mm	MRE↓ (mm)	STD↓ (mm)	SDR_d(%)↑ 3 mm	6 mm	9 mm
w/o fingerprint	3.07	1.91	32.1	55.84	74.36	6.44	7.31	21.48	56.82	80.92
w/o DL Library	3.89	2.23	25.05	46.82	66.55	7.4	12.35	20.68	54.21	76.1
Ours(AutoLand)	1.41	1.36	75.82	93.51	98.34	4.03	4.02	59.04	81.12	90.56
Variant	Cardiac LAX					Cardiac SAX				
	MRE↓ (mm)	STD↓ (mm)	SDR_d(%)↑ 2 mm	2.5 mm	3 mm	MRE↓ (mm)	STD↓ (mm)	SDR_d(%)↑ 2 mm	2.5 mm	3 mm
w/o fingerprint	2.56	4.43	43.25	67.79	81.88	2.77	5.39	41.73	67.11	82.08
w/o DL Library	3.26	4.72	39.19	63.78	79.35	3.34	6.93	37.83	66.66	75.03
Ours(AutoLand)	1.05	1.97	94.49	97.08	98.54	1.18	2.48	92.52	94.89	96.13

priors that guide strategy reasoning for architecture and optimization. Together, they enable the agent to generate executable strategies that generalize across modalities in few-sample scenarios.

4 Conclusion

The proposed AutoLand presents a novel autonomous framework for data-efficient, agentic workflows in universal medical anatomical landmark detection. By coupling the Physics Fingerprint for data perception with the Medical DL Library for strategy reasoning, AutoLand demonstrates robust and adaptable detection across diverse anatomical targets and imaging modalities. Replacing static heuristics with dynamic agentic reasoning offers a promising path for reducing reliance on manual expertise when deploying models to new clinical scenarios. Future work will extend AutoLand to large-scale 3D volumetric datasets and incorporate interactive physician feedback to facilitate clinical integration.

Acknowledgements

This work was supported by the Key Research and Development Program of Liaoning Province (2024JH2/102500076) and Fundamental Research Funds for the Central Universities (N25BJD013).

Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. pp. 205–218. Springer (2022)
2. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., et al.: Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)
3. Choi, M., Jang, J.S.: Heatmap-based active shape model for landmark detection in lumbar x-ray images. Journal of Imaging Informatics in Medicine **38**(1), 291–308 (2025)
4. Goyal, S., Rastogi, E., Rajagopal, S.P., Yuan, D., Zhao, F., Chintagunta, J., Naik, G., Ward, J.: Healai: A healthcare llm for effective medical documentation. In: Proceedings of the 17th ACM International Conference on Web Search and Data Mining. pp. 1167–1168 (2024)
5. He, Y., Yang, D., Roth, H., Zhao, C., Xu, D.: Dints: Differentiable neural network topology search for 3d medical image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5841–5850 (2021)
6. Huang, Z., Li, H., Shao, S., Zhu, H., Hu, H., Cheng, Z., Wang, J., Kevin Zhou, S.: Pele scores: pelvic x-ray landmark detection with pelvis extraction and enhancement. International Journal of Computer Assisted Radiology and Surgery **19**(5), 939–950 (2024)
7. Huang, Z., Zhao, R., Leung, F.H., Banerjee, S., Lam, K.M., Zheng, Y.P., Ling, S.H.: Landmark localization from medical images with generative distribution prior. IEEE transactions on medical imaging **43**(7), 2679–2692 (2024)
8. Ibragimov, B., Likar, B., Pernuš, F., Vrtovec, T.: Shape representation for efficient landmark-based segmentation in 3-d. IEEE transactions on medical imaging **33**(4), 861–874 (2014)
9. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
10. Jonkers, J., Duchateau, L., Van Wallendael, G., Van Hoecke, S.: landmarker: A toolkit for anatomical landmark localization in 2d/3d images. SoftwareX **30**, 102165 (2025)
11. Kevin, Z., et al.: Learning to localize cross-anatomy landmarks in x-ray images with a universal model. BME frontiers (2022)

12. Lee, M., Chung, M., Shin, Y.G.: Cephalometric landmark detection via global and local encoders and patch-wise attentions. Neurocomputing **470**, 182–189 (2022)
13. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
14. Lindner, C., Bromiley, P.A., Ionita, M.C., Cootes, T.F.: Robust and accurate shape model matching using random forest regression-voting. IEEE transactions on pattern analysis and machine intelligence **37**(9), 1862–1874 (2014)
15. Liu, C., Xie, H., Zhang, S., Mao, Z., Sun, J., Zhang, Y.: Misshapen pelvis landmark detection with local-global feature learning for diagnosing developmental dysplasia of the hip. IEEE Transactions on Medical Imaging **39**(12), 3944–3954 (2020)
16. Lu, C., Yang, G., Qiao, X., Chen, W., Zeng, Q.: Univerdetect: Universal landmark detection method for multidomain x-ray images. Neurocomputing **600**, 128157 (2024)
17. Nie, X., Feng, J., Zhang, J., Yan, S.: Single-stage multi-person pose machines. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6951–6960 (2019)
18. Pal, A., Umapathi, L.K., Sankarasubbu, M.: Med-halt: Medical domain hallucination test for large language models. arXiv preprint arXiv:2307.15343 (2023)
19. Pandey, N.: Chest x-ray masks and labels. https://www.kaggle.com/datasets/nikhilpandey360/chest-xray-masks-and-labels (2018)
20. Rakshit, S.: Isbi challenge dataset (2015). https://www.kaggle.com/datasets/soumikrakshit/isbi-challenge-dataset (2018)
21. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 (2019)
22. Wang, C.W., Huang, C.T., Lee, J.H., Li, C.H., Chang, S.W., Siao, M.J., Lai, T.M., Ibragimov, B., Vrtovec, T., Ronneberger, O., et al.: A benchmark for comparison of dental radiography analysis algorithms. Medical image analysis **31**, 63–76 (2016)
23. Wang, L., Xu, Q., Leung, S., Chung, J., Chen, B., Li, S.: Accurate automated cobb angles estimation using multi-view extrapolation net. Medical image analysis **58**, 101542 (2019)
24. Yan, X., Jiang, W., Shi, Y., Zhuo, C.: Ms-nas: Multi-scale neural architecture search for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 388–397. Springer (2020)
25. Zhou, S.K., Greenspan, H., Davatzikos, C., Duncan, J.S., Van Ginneken, B., Madabhushi, A., Prince, J.L., Rueckert, D., Summers, R.M.: A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proceedings of the IEEE **109**(5), 820–838 (2021)
26. Zhu, H., Quan, Q., Yao, Q., Liu, Z., Zhou, S.K.: Uod: Universal one-shot detection of anatomical landmarks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 24–34. Springer (2023)
27. Zhu, H., Yao, Q., Xiao, L., Zhou, S.K.: You only learn once: Universal anatomical landmark detection. In: International conference on medical image computing and computer-assisted intervention. pp. 85–95. Springer (2021)