Dwarf: Disease-weighted network for attention map refinement

Haozhe Luo^{1,†}, Aurélie Pahud de Mortanges¹, Oana Inel^{2,*}, Abraham Bernstein^{2,*}, and Mauricio Reyes^{1,*}

¹ ARTORG Center for Biomedical Engineering Research, University of Bern, Bern, Switzerland
² University of Zurich, Switzerland

Abstract. The interpretability of deep learning is crucial for evaluating the reliability of medical imaging models and reducing the risks of inaccurate patient recommendations. This study addresses the "human out of the loop" and "trustworthiness" issues in medical image analysis by integrating medical professionals into the interpretability process. We propose a disease-weighted attention map refinement network (DWARF) that leverages expert feedback to enhance model relevance and accuracy. Our method employs cyclic training [13] to iteratively improve diagnostic performance, generating precise and interpretable feature maps. Experimental results demonstrate significant improvements in interpretability and diagnostic accuracy across multiple medical imaging datasets. This approach fosters effective collaboration between AI systems and healthcare professionals, ultimately aiming to improve patient outcomes. The code is available on *censored for review*.

1 Introduction

Machine learning (ML) techniques, especially deep learning (DL) have significantly expanded in both research and industrial sectors, particularly with the advancements in deep neural networks (DNN). The impact and potential repercussions of these technologies have become too significant to overlook. In certain applications, failure is unacceptable; for example, a temporary malfunction in a computer vision algorithm for autonomous vehicles can result in fatalities. In the medical field, the stakes are even higher as human lives are directly affected. Early detection of diseases is crucial for patient recovery and for preventing the progression of illnesses to more severe stages. Despite recent promising results from machine learning methods [18,3,11,27,29,28,12], current methods are not without imperfections [14,7,23]. Specifically, many medical AI systems still struggle with issues such as short-cut learning [5] and misattribution [6], which can hamper the reliability of medical AI system.

The significance of interpretability in medical imaging arises from the crucial need for transparency and trust in healthcare applications of AI. Traditionally, medical imaging analysis prioritized accuracy, but with the increasing integration of AI, the emphasis has shifted towards creating understandable and explainable AI systems. The goal of explainable AI (XAI) is to make AI decision-making processes in medical imaging more comprehensible, thereby enhancing reliability and enabling healthcare professionals to effectively integrate AI tools into clinical practice [24,1]. Current XAI methods interpret model outputs through various means, but due to the inherent uncertainty and complexity of deep learning patterns, translating these into intuitive interpretations for users remains difficult. This situation underscores the necessity for trustworthy ML systems in healthcare, which demand transparency and active involvement of medical professionals to ensure accuracy and relevance [4,17].

Saliency map-based techniques are extensively utilized in medical explainable AI (XAI) due to their capability to highlight critical regions in medical images that influence model predictions. These techniques enhance transparency and trust in AI-driven diagnostic tools by visually representing areas of interest, such as tumors or lesions [2]. Methods like Grad-CAM [20], Integrated Gradients [22], and SmoothGrad [21] provide reliable explanations by generating class-specific localization maps and reducing noise. Consequently, saliency maps play a crucial role in making AI decisions interpretable and justifiable in clinical settings, promoting their adoption in healthcare.

In this work, we tackle the "human out of the loop" and "trustworthiness" issues in medical image analysis by incorporating medical professionals into the interpretability process [16]. By leveraging their insights, we enhance interpretability maps, aligning deep learning explanations more closely with medical intuition. This approach enhances the relevance and utility of deep learning interpretations in medical diagnostics through expert feedback. Achieving effective human-machine teaming, where human decision-making and ML system performance are integrated, is essential for improving patient outcomes. To be specific, in this work, we are aiming to address the imperfect alignment between medical items and corresponding visual regions [26]. We take the clinicians' attention annotations as visual guidance during the classification model training to simultaneously optimise attention maps and classification performance. Through extensive experiments, DWARF outperforms other baselines on both classification performance and attention map performance across different datasets. Further feedback from clinicians demonstrates that DWARF enhances clinicians' confidence in AI assisted disease classification.

2 Method

The object of our method is to introducing disease specific attention as a guidance during the classification model training. In this section, we introduce our DWARF from three aspects: architecture and training strategy, losses and network initialization.

2.1 Architecture and training strategy

To refine the saliency map with finding-related prior knowledge, we introduce our DWARF module, as shown in Fig.1. The overall structure of DWARF consists of a pretrained Vision-Language Model (VLM), denoted as $f_{\rm vlm}$, and expert heads f_{heads} . For a multi-modality model, it is hard to directly optimize cross attention map because the attention difference between human and the model [25] as well as the scale changes (the cross attention value is not bounded within 0-1 according to experiments). To address this, we utilize the finding-specific heads to project the origin attention map denoted as $M_c \in \mathbb{R}^{h \times w}$ of class c where c is in a collection of different findings' labels N from its origin embedding space to the visualization space for clinicians. Finally we get the segmentation map M'_c = $f_{\text{head}}(M_c)$. To accumulate finding-specific knowledge effectively, we introduce a cyclic training process. The cyclic training mechanism is designed to iteratively refine the network's understanding and segmentation of specific findings. By incorporating cyclic training, the network can effectively refine its ability to identify and segment specific medical findings, leading to improved diagnostic performance. The overall training pseudo code is shown in Algorithm.1



Fig. 1: Flow chart of finetuning the classification model. Our method only trains single disease each epoch with disease name as prompt. For each disease, we add an additional head to mapping origin attention to refined segmentation map.

2.2 Losses and network initialization

Loss Function In our framework, we employ a cross-entropy loss, denoted as \mathcal{L}_{cls} , for multi-label classification tasks. Additionally, we use a modified Dice loss, \mathcal{L}_{seg} , optimized for attention maps. Attention maps in medical image analysis are critical for detecting disease-related markers. Training and validation typically focus on positive samples, which may cause models to overestimate certain

Algorithm 1 Training Process for DWARF

Input: Multi-label dataset D_{multi} , Segmentation head f_{head} , Ground truth G **Output:** Optimized network parameters θ 1: Data Collection: 2: Decompose D_{multi} into multiple single-label datasets D_{single} 3: for each finding f in D_{multi} do $D_{single}[f] \leftarrow \text{createSingleLabelDataset}(D_{multi}, f)$ 4: 5: end for 6: Segmentation: 7: for each single-label dataset $D_{single}[f]$ do 8: for each image I in $D_{single}[f]$ do 9: $M_c[I] \leftarrow f_{head}(I)$ end for 10: 11: end for 12: Classification and Segmentation Feedback Loop: 13: for each single-label dataset $D_{single}[f]$ do 14: for each image I in $D_{single}[f]$ do 15: $C_{output}[I] \leftarrow \text{classify}(I, M_c[I])$ 16: $L_{seg} \leftarrow \text{calculateLoss}(M_c[I], G[I]_{seg})$ 17: $L_{cls} \leftarrow \text{calculateLoss}(C_{output}[I], G[I]_{cls})$ $L_{total} \leftarrow \lambda L_{seg} + (1 - \lambda) L_{cls}$ 18: $\theta \leftarrow updateNetworkParams(\theta, L_{total})$ 19:20: end for 21: end for 22: Iterative Refinement: 23: for each epoch do for each single-label dataset $D_{single}[f]$ do 24: $M_c \leftarrow \text{segmentation}(D_{single}[f], f_{head})$ 25:26: $\theta \leftarrow \text{feedbackLoop}(M_c, G, \theta)$ end for 27:28: end for 29: Return θ

features, leading to false positives. Our False Positive Suppression technique mitigates this by adjusting the Dice score to penalize false positives.

The standard metric, the Soft Dice Score, is mathematically represented as:

$$\mathcal{L}_{\text{Dice}} = \frac{2 \cdot |X \cap Y| + \alpha + \varepsilon}{|X| + |Y| + \alpha + \varepsilon} \tag{1}$$

where X and Y are sets representing the predicted and true regions, respectively, α is a smoothing constant to prevent division by zero, and ε ensures numerical stability.

To specifically address false positives, we define:

$$\mathcal{L}_{\text{seg}} = \frac{2 \cdot |X \cap Y| + \alpha + \varepsilon}{|X| + \text{adjusted}|Y| + \alpha + \varepsilon}$$
(2)

$$adjusted|Y| = |Y| + (w_{FP} - 1) \cdot FP \tag{3}$$

Here, FP is the count of false positives, and w_{FP} is the weighting factor penalizing each false positive.

The combined loss function, aimed at minimizing, is expressed as:

$$\mathcal{L} = \lambda \mathcal{L}_{seq} + (1 - \lambda) \mathcal{L}_{cls} \tag{4}$$

where α adjusts the emphasis on attention annotations.

Model Initialization We initialize the text encoder using the weights from Med-KEBERT[28] (An advanced text encoder pretrained on medical knowledge graph). For the image encoder and cross-attention layers, we adopt the architecture from DeViDe. Additionally, we introduce disease-specific segmentation heads for targeted diseases.

We propose the **Identity Enhanced Initialization (IEI)** technique to address the limitations associated with random or simplistic initializations, which can often lead to suboptimal learning trajectories. Our observations indicate that random initialization of segmentation expert heads tends to encourage the model to learn shortcuts, as depicted in Figure 2(a). Conversely, the pretrained Visual Language Model (VLM) already offers robust image-text correspondence[12], which can serve as an effective foundation for initialization. The IEI method involves initializing the weights of the segmentation heads with an identity matrix, focusing on enhancing the model's sensitivity to structures pertinent to specific diseases. This approach directs the learning process towards more precise feature recognition from the start. By avoiding reliance on the simplest or most obvious features (referred to as the "shortcut path"). A qualitative comparison is illustrated in Figure 2.

3 Experiments and Results

To fully assess the properties of our framework, we con- duct extensive experiments across quantitative metrics and qualitative indices.

3.1 Dataset

We used three different publicly available datasets: ChestX-Det [10], CheXlocalize [19], and Vindr-CXR [15]. These datasets contain between 1,000 to 10,000 chest X-rays (CXRs). Each dataset includes multi-label classification labels as well as segmentation labels, provided at the bounding box or polygonal levels.

 $\mathbf{5}$

6 H. LUO ET AL.



Fig. 2: With random initialization, the model tends to directly learn shortcut results which always highlight the same area. While using IEI initialization, the model can start from pretrained VLM's attention to refine its focus.

The ChestXDet dataset is segmented into three versions based on the segmentation difficulty of the findings. The four-findings version includes common findings such as Atelectasis, Cardiomegaly, Consolidation, and Effusion, which are prevalent across various datasets. The expanded seven-findings version adds Diffuse Nodule, Emphysema, and Mass, which show relatively high performance. The full version encompasses the original ChestXDet dataset with 13 findings.

3.2 Baselines

To ascertain the efficacy of the DWARF method for modeling, we established several baselines for comparison. These include a pretrained vision language model without fine-tuning including DeViDe[12] and KAD[28], a finetuned VLM employing only multi-label classification loss, and a finetuned VLM training with classification loss and multi-label segmentation loss (extra supervision strategy of GAIN [8]).

3.3 Training Details

With ViT-B as the visual backbone and Med-KEBERT as the textual backbone, we finetune on the ChestX-Det dataset [9] on an image size of 224. We utilize the AdamW optimizer with learning rates $lr = 5 \times 10^{-5}$. We optimize on V100 16G GPUS with a total batch size of 32 for a total of 500 epochs.

3.4 Quantitative results

DWARF achieves SoTA results compared to other pretrained/finetuned VLM baselines. In the Tab 1, we compared the performance of our DWARF with various state-of-the-art models, including pretrained DeViDe, KAD and finetuned GAIN which trained with direct Cross-Entropy Loss and Dice Loss. These models were evaluated based on different metrics such as Max AUC, Max Dice, F1 Score, and MCC. Our analysis extends to various datasets including ChestX-Det, cheXlocalize and Vindr-CXR, highlighting the models' adaptability and effectiveness across different medical imaging contexts. DWARF yields stably better results than the other aforementioned methods.

DWARF achieves enhanced Stability and Scalability To explore the scalability and stability of DWARF, we firstly compared DWARF with GAIN across different numbers of diseases and observed significant and consistent improvements on ChestXDet dataset. For 4 diseases (defined in sec.3.1), the Dice score improved from 0.1438 to **0.3854**, and the max AUC score increased from 0.8660 to **0.8871**. Similarly, for 7 diseases, the Dice score increased from 0.1903 to **0.3492**, and the max AUC score rose from 0.8519 to **0.8717**. These results, as shown in Fig.3, highlight the consistent enhancement provided by DWARF across different numbers of diseases. Additionally, since our model is only trained once per epoch, it results in insufficient training within the same number of epochs. Therefore, we extended the training from 500 epochs to 1000 epochs to explore the scalability of the model's performance. We found that the Dice score improved from 0.1805 to 0.2302, as shown in Fig. 3.

DWARF's Independence from Extensive Annotation To address the challenges associated with obtaining dense and consistently high-quality annotations for medical imaging, we are exploring the feasibility of substituting human annotations with pseudo labels generated by disease-specific models. This approach leverages the expertise encapsulated in pre-trained segmentation models for various diseases. The results of these experiments, as detailed in Tab 3, indicate that the DWARF system continues to demonstrate substantial improvements, achieving an impressive enhancement in performance by **0.1271**.



Fig. 3: DWARF demonstrates sustained learning capacity, benefiting from extended training epochs, whereas the baseline model suffers from overfitting with additional training.

8 H. LUO ET AL.

DWARF Enhances Clinician Confidence in Classification Models To validate our approach's effectiveness in enhancing clinicians' confidence in classification models, we conducted a double-blind experiment. We randomly selected 5 samples each of four diseases (4 common diseases across datasets: Atelectasis, Cardiomegaly, Consolidation, and Effusion) from the ChestXDet dataset, totaling 20 samples. Using DWARF and DeViDe, we generated attention maps for each sample, creating 20 anonymized sets for clinician preference evaluation. Two clinicians assessed the maps based on: 1) Accuracy (whether the map accurately pinpointed the finding with high confidence), and 2) Specificity (whether the map was focused and intuitive). DWARF was preferred in 15 out of 20 and 18 out of 20 cases, with an average preference rate of 82.5

3.5 Ablation results

Disease-specific head makes attention map trainable According to the analysis presented in sec.2.1, the segmentation expert heads are pivotal for projecting the cross-attention values effectively. To explore the contributions of these heads, we conducted an ablation study, the results of which are detailed in Tab 4. The inclusion of expert heads significantly enhanced the attention performance, with the metric improving from 0.2288 to a robust **0.3559**.

3.6 Qualitative Result

To enhance the clarity of the explanation regarding the visual explainability representation of our method, we illustrate the attention map using the test set of the CheX-Det dataset. As depicted in Fig.4, our DWARF method significantly improves the focus of the classification model's attention. This enhancement allows the model to more precisely highlight the relevant areas that form the basis for its classification decisions.



Fig. 4: Qualitative results of training with and without the DWARF architecture demonstrate that utilizing our DWARF framework consistently enhances the aggregation of feature maps and provides prior region information.

4 Conclusion

In this research, we have developed a two-stage saliency map revision strategy. This approach effectively integrates disease-related knowledge and clinicians' preferences into the generation of saliency maps. By incorporating this methodology, we are also introducing clinicians into the AI training loop. This strategy not only improves the accuracy of the AI but also makes it more userfriendly for clinicians, ensuring that their expertise and insights are reflected in the AI's learning process. There are still some unvalidated capabilities including the transferability and few-shot ability of the model. We will conduct further experiments to address them.

Table 1: DWARF outperforms other finetuned/pretrained VLM on classification performance and attention accuracy on 4 metrics: AUC, Dice, F1 score and MCC. All models take the same transformer architecture as encoder. The best methods are bolded

Method	Dataset	AUC (%)	F1 Score (%)	MCC (%)	Max Dice (%)	Model Type
DeViDe [12]	ChestX-Det	74.24	42.46	34.29	13.66	Pretrained VLM
KAD [28]	ChestX-Det	73.81	40.04	31.84	13.89	Pretrained VLM
GAIN [8]	ChestX-Det	80.90	48.57	42.65	13.90	Finetuned VLM
DWARF	$\mathbf{ChestX}\textbf{-}\mathbf{Det}$	81.94 ± 0.37	53.73 ± 0.29	49.87 ± 0.06	18.24 ± 0.18	Finetuned VLM
DeViDe	cheXlocalize	78.26	41.66	59.83	11.93	Pretrained VLM
KAD	cheXlocalize	74.22	58.01	41.53	11.59	Pretrained VLM
GAIN	cheXlocalize	83.64	62.86	50.18	11.91	Finetuned VLM
DWARF	cheXlocalize	84.93 ± 0.05	63.44 ± 0.21	50.79 ± 0.32	13.40 ± 0.51	Finetuned VLM
DeViDe	Vindr-CXR	72.92	41.28	31.43	7.19	Pretrained VLM
KAD	Vindr-CXR	73.19	40.22	30.78	7.06	Pretrained VLM
GAIN	Vindr-CXR	78.51	45.20	36.48	7.23	Finetuned VLM
DWARF	$\operatorname{Vindr-CXR}$	80.01 ± 0.23	47.05 ± 1.14	39.55 ± 1.07	10.21 ± 0.42	Finetuned VLM

Table 2: Comparison of DWARF and GAIN models on different numbers of selected diseases from the ChestX-Det dataset.

Backbone	Disease number	• AUC (%)	F1 (%)	MCC (%)	Max Dice $(\%)$
GAIN	$\frac{4}{7}$	86.80 85.10	-	-	14.38
	13	80.90	48.57	42.65	19.03 13.90
DWARF	4	88.71	-	-	41.47
	$7\\13$	$\begin{array}{c} 87.17\\ 81.94\end{array}$	60.17 53.73	52.01 49.87	$\begin{array}{c} 39.11 \\ 18.24 \end{array}$

Table 3: Ablations of expert supervision. DWARF taking the trained segmentation models' generated pseudo label could also achieve impressive improvement. Seven usual findings in ChestX-Det are used for evaluation including: Atelectasis, Cardiomegaly, Consolidation, Effusion, Diffuse Nodule, Emphysema, and Mass.

Method	Disease $\#$	AUC (%)	DICE (%)	Max DICE (%)
GaIN(cls)	7	86.80	14.38	14.38
DWARF (expert)	7	84.73	31.71	36.94
DWARF	7	87.17	38.56	39.11

Table 4: Ablations of diseasespecific head. Seven usual findings are selected for evaluation (see Table 3). Compared to directly optimizing cross attention value, introducing additional segmentation expert head improves both the classification and attention performance.

Method	AUC	Max Dice
Directly optimize	86.63	22.88
Disease-specific head	87.32	35.59

References

- Band, S.S., Yarahmadi, A., Hsu, C.C., Biyari, M., Sookhak, M., Ameri, R., Dehzangi, I., Chronopoulos, A.T., Liang, H.W.: Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. Informatics in Medicine Unlocked p. 101286 (2023)
- Banerjee, S., Mitra, S., Shankar, B.U.: Automated 3d segmentation of brain tumor using visual saliency. Information Sciences 424, 337–353 (2018)
- Cao, K., Xia, Y., Yao, J., Han, X., Lambert, L., Zhang, T., Tang, W., Jin, G., Jiang, H., Fang, X., et al.: Large-scale pancreatic cancer detection via non-contrast ct and deep learning. Nature medicine 29(12), 3033–3043 (2023)
- Chen, H., Gomez, C., Huang, C.M., Unberath, M.: Explainable medical imaging ai needs human-centered design: guidelines and evidence from a systematic review. NPJ digital medicine 5(1), 156 (2022)
- Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence 2(11), 665–673 (2020)
- Hatherley, J.J.: Limits of trust in medical ai. Journal of medical ethics 46(7), 478–481 (2020)
- Kaviani, S., Han, K.J., Sohn, I.: Adversarial attacks and defenses on ai in medical imaging informatics: A survey. Expert Systems with Applications 198, 116815 (2022)
- Li, K., Wu, Z., Peng, K.C., Ernst, J., Fu, Y.: Tell me where to look: Guided attention inference network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9215–9223 (2018)
- Lian, J., Liu, J., Zhang, S., Gao, K., Liu, X., Zhang, D., Yu, Y.: A structureaware relation network for thoracic diseases detection and segmentation. IEEE Transactions on Medical Imaging 40(8), 2042–2052 (2021)
- Liu, J., Lian, J., Yu, Y.: Chestx-det10: Chest x-ray dataset on detection of thoracic abnormalities (2020)
- Luo, H., Changdong, Y., Selvan, R.: Hybrid ladder transformers with efficient parallel-cross attention for medical image segmentation. In: International conference on medical imaging with deep learning. pp. 808–819. PMLR (2022)
- Luo, H., Zhou, Z., Royer, C., Sekuboyina, A., Menze, B.: Devide: Faceted medical knowledge for improved medical vision-language pre-training. arXiv preprint arXiv:2404.03618 (2024)
- Ma, D., Pang, J., Gotway, M.B., Liang, J.: Foundation ark: Accruing and reusing knowledge for superior and robust performance. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 651–662. Springer (2023)
- 14. Maier-Hein, L., Menze, B., et al.: Metrics reloaded: Pitfalls and recommendations for image analysis validation. arXiv. org (2206.01653) (2022)
- Nguyen, H.Q., Lam, K., Le, L.T., Pham, H.H., Tran, D.Q., Nguyen, D.B., Le, D.D., Pham, C.M., Tong, H.T., Dinh, D.H., et al.: Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. Scientific Data 9(1), 429 (2022)
- 16. Patrício, C., Neves, J.C., Teixeira, L.F.: Explainable deep learning methods in medical image classification: A survey. ACM Computing Surveys **56**(4), 1–41 (2023)
- 17. Prentzas, N., Kakas, A., Pattichis, C.S.: Explainable ai applications in the medical domain: a systematic review. arXiv preprint arXiv:2308.05411 (2023)

- 12 H. LUO ET AL.
- Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention-MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
- Saporta, A., Gui, X., Agrawal, A., Pareek, A., Truong, S.Q., Nguyen, C.D., Ngo, V.D., Seekins, J., Blankenberg, F.G., Ng, A.Y., et al.: Benchmarking saliency methods for chest x-ray interpretation. Nature Machine Intelligence 4(10), 867– 878 (2022)
- Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Gradcam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
- Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 (2017)
- 22. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International conference on machine learning. pp. 3319–3328. PMLR (2017)
- Topol, E.J.: High-performance medicine: the convergence of human and artificial intelligence. Nature medicine 25(1), 44–56 (2019)
- Van der Velden, B.H., Kuijf, H.J., Gilhuijs, K.G., Viergever, M.A.: Explainable artificial intelligence (xai) in deep learning-based medical image analysis. Medical Image Analysis 79, 102470 (2022)
- 25. Yan, K., Ji, L., Wang, Z., Wang, Y., Duan, N., Ma, S.: Voila-a: Aligning visionlanguage models with user's gaze attention. arXiv preprint arXiv:2401.09454 (2023)
- You, D., Liu, F., Ge, S., Xie, X., Zhang, J., Wu, X.: Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. pp. 72–82. Springer (2021)
- You, S., Wiest, R., Reyes, M.: Sarf: Saliency regularized feature learning improves mri sequence classification. Computer methods and programs in biomedicine 243, 107867 (2024)
- Zhang, X., Wu, C., Zhang, Y., Xie, W., Wang, Y.: Knowledge-enhanced visuallanguage pre-training on chest radiology images. Nature Communications 14(1), 4542 (2023)
- Zhou, Z., Luo, H., Pang, J., Ding, X., Gotway, M., Liang, J.: Learning anatomically consistent embedding for chest radiography. arXiv preprint arXiv:2312.00335 (2023)