DEER: A Foundation Model with Dual-Branch CT-Enhanced Embeddings for Thoracoabdominal Radiographs

Yihua Sun[1,2†], Jia Guo[2†], Qijing Wang[2,3†], Jingzhou Ouyang[3], Ge Li[3], Nan Zhang[4(✉)], Fang Chen[1(✉)], and Hongen Liao[1,2(✉)]

[1] School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China chen-fang@sjtu.edu.cn, liaohongen@sjtu.edu.cn

[2] School of Biomedical Engineering, Tsinghua Medicine, Tsinghua University, Beijing, China

[3] Zhongshi Kangkai Technology Co., Ltd., Beijing, China

[4] State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China zhangnan@ia.ac.cn

Abstract. The chest and abdomen are critical, anatomically complex regions with diverse pathologies, with digital radiography (DR) serving as the primary modality for screening and diagnosis. Developing an effective DR foundation model with strong disease discrimination through self-supervised learning remains a significant challenge. Moreover, prior research has largely overlooked abdominal DR and the potential of computed tomography (CT) localizers as a bridge to fine-grained CT knowledge. In this work, we pioneer the incorporation of large-scale, previously neglected CT localizers and CT reports to train a unified DR foundation model on both chest and abdominal regions. We propose a dual-branch contrastive pretraining approach using 1.1 million DR and 1 million CT localizers, each paired with radiologist reports. By integrating comprehensive knowledge from CT reports, we substantially enhance the model's capability in DR interpretation. Extensive evaluations on 10 downstream tasks demonstrate the superiority of our model. To the best of our knowledge, this is the first foundational pretraining framework that jointly leverages DR, CT localizers, and reports, scaling to an unprecedented 2.1 million image-report pairs. It offers valuable insights into CT-enhanced intelligent DR diagnosis and scalable multimodal medical foundation models, with demonstrated robustness across diverse clinical tasks. The model is available at https://github.com/sunyh1/DEER.

Keywords: Contrastive learning · Foundation model · Localizer · Radiograph · Thoracoabdominal region.

1 Introduction

The chest and abdomen are anatomically complex regions associated with a broad spectrum of pathologies. Digital radiography (DR) is the primary imaging

† Y. Sun, J. Guo and Q. Wang are the co-first authors.

modality for screening and diagnosis [18]. However, current research has predominantly focused on chest DR, while abdominal DR remains underexplored. This highlights the need for more generalizable foundation models for interpreting a wide range of pathologies across trunk DR images.

However, scaling up foundation models for robust disease interpretation remains challenging. Current methods include image-only self-supervised learning [13,21], which lacks the ability to extract disease-relevant semantics. Alternatively, image-label supervision [12] relies on labor-intensive manual annotations or on automatically generated labels that often oversimplify the rich information in radiology reports. In contrast, multimodal self-supervised approaches combining images and radiology reports [15] offer a promising solution. They allow foundation models to scale up disease interpretation capabilities directly from routine clinical data, without additional human labeling.

On the other hand, radiologists have limited ability to interpret diseases from projective 2D DR, which limits the diagnostic information in DR reports. In contrast, 3D computed tomography (CT) provides much richer disease insights. Therefore, leveraging CT to enhance DR-based models could be highly beneficial. Here, we identify the previously underexplored potential of CT localizers (LOC) as a bridge between the detailed disease information in CT reports and DR images, enhancing disease discriminative representations from DR.

In this study, we jointly leverage DR, LOC, and reports, scaling to an unprecedented 2.1 million image-report pairs for multimodal self-supervised pretraining. Building on this, we present a foundation model (Fig. 1a) with **D**ual-branch CT-**E**nhanced **E**mbeddings for thoracoabdominal **R**adiographs (**DEER**). Our key novelties and contributions are: 1) **Thoracoabdominal DR foundation model**: The model enables chest and abdominal disease interpretation and generalizes well to pelvic diseases. 2) **CT-enhanced DR embeddings**: We improve disease discriminative representations from DR using LOC and CT reports. 3) **Large-scale multimodal pretraining**: DEER is pretrained on 2.1 million image-report pairs, the largest multimodal radiograph dataset to date.

Evaluations on 10 downstream tasks demonstrate DEER's superiority. Our results highlight the potential of CT-enhanced embeddings for advancing DR intelligence. Notably, by leveraging routinely collected image-report pairs for self-supervised pretraining, DEER surpasses models trained with large-scale label supervision. This demonstrates the feasibility of scaling self-supervised learning to develop more generalizable and high-performing medical foundation models.

2 Methods

2.1 Pretraining Datasets

We use only image-report pairs from routine radiology workflows, without extra manual labels, enabling scalable self-supervised foundation model development.

DR image and report. The **chest DR** primarily covers the lungs and may include part of the upper abdomen. Similarly, **abdominal DR** mainly encompasses the abdomen and typically includes the pelvis and part of the lower

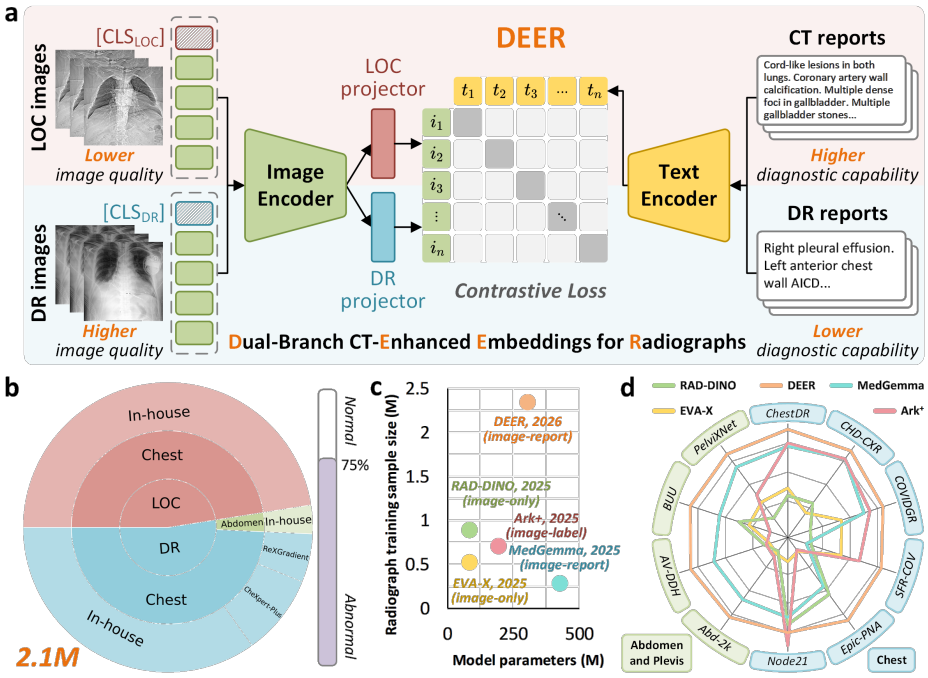

Fig. 1. (a) DEER trains on routinely collected clinical image-report data without additional manual annotations. We propose a dual-branch contrastive learning approach that uses CT localizers and CT reports to enhance DR image representations. (b) DEER jointly learns from both chest and abdominal DRs, as well as CT localizers. (c) DEER uses a larger model trained on more data than SOTA methods, enabling further foundation model scaling. (d) DEER achieves superior performance in downstream evaluations (fine-tuning), showing strong robustness across diverse clinical applications.

lungs. Therefore, jointly learning from chest and abdominal DR facilitates the integration of both overlapping and region-specific information across different DR types, enabling a more comprehensive understanding of trunk DR. We use public chest DR datasets for training: CheXpert-Plus [1] and ReXGradient [22]. In addition, we incorporate a large-scale private chest and abdominal DR dataset, resulting in a total of 1.1 million DR image-report pairs (Fig. 1b).

CT localizer and report. Radiologists face inherent limitations in interpreting diseases from 2D DR, leading to relatively limited diagnostic information in DR reports. To address this gap, we investigate the largely unexplored role of CT localizers as a bridge between diagnostic-rich CT reports and DR images to enhance disease-discriminative DR representations. We utilize 1 million private LOC images paired with CT reports to enhance DR training (Fig. 1b).

To the best of our knowledge, we construct the first multimodal dataset combining DR, LOC, and reports, totaling an unprecedented 2.1 million image-report pairs. We define a report-image pair as abnormal if the report contains any

positive finding or impression. In our dataset, 75% of the pairs are abnormal (Fig. 1b), highlighting a diversity that enables robust foundation model development.

2.2 Model Framework

DR images exhibit higher image quality, yet DR reports provide relatively limited diagnostic information. In contrast, LOC has lower image quality, while CT reports offer richer and more diagnostically informative content. DEER is designed to leverage the strengths of both modalities, which combine high-quality DR image understanding with strong diagnostic capability (Fig. 1a).

Dual-Branch Visual Encoding. To incorporate LOC while preserving the ability to interpret high-quality DR images, we design a dual-branch visual encoder (E_I) based on a shared ViT [4] backbone (Fig. 1a). For DR and LOC input images (I_{DR}, I_{LOC}), we employ two independent classification tokens ([$\mathrm{CLS}_{\mathrm{DR}}$], [$\mathrm{CLS}_{\mathrm{LOC}}$]) and separate projection layers (f_{DR}, f_{LOC}) to extract visual features: $i_{mod} = f_{mod}\left(E_I\left([\mathrm{CLS}_{mod}], I_{mod}\right)\right)$, where $mod \in \{\mathrm{DR}, \mathrm{LOC}\}$. The resulting features i_{DR} and i_{LOC} are the extracted visual features, respectively.

Image-Report Contrastive Learning. As shown in Fig. 1a, to fully leverage radiology report knowledge for DR image representation learning, we adopt the contrastive language-image pretraining (CLIP) framework [14,5]. Specifically, both DR and CT reports (T) are encoded into textual feature representations t using CLIP's Transformer-based text encoder E_T: $t = E_T(T)$.

A contrastive learning objective is then applied to pull matched image-text pairs closer in the shared embedding space while pushing unmatched pairs farther apart. Given a batch of size N, let p, q denote the index of an element within the batch. The contrastive learning loss can be written as:

$$\mathcal{L}_{i2t} = -\frac{1}{N}\sum_{p}^{N}\log\frac{\exp\left(\langle i_p, t_p\rangle/\tau\right)}{\sum_{q}^{N}\exp\left(\langle i_p, t_q\rangle/\tau\right)},\ \mathcal{L}_{t2i} = -\frac{1}{N}\sum_{q}^{N}\log\frac{\exp\left(\langle i_q, t_q\rangle/\tau\right)}{\sum_{p}^{N}\exp\left(\langle i_p, t_q\rangle/\tau\right)},$$

where $\langle\cdot,\cdot\rangle$ denotes the cosine similarity, and τ is the temperature parameter.

The final training loss is $\mathcal{L} = (\mathcal{L}_{i2t} + \mathcal{L}_{t2i})/2$. CT reports are introduced to enhance diagnostic discriminability and refine textual representations, thereby optimizing image features in the shared embedding space via contrastive loss.

3 Results

3.1 Implementation Details and Evaluation Metrics

Radiology reports were standardized for positive findings and impressions by directly prompting a large language model. DEER was pretrained using 2 × 96 GB RTX PRO 6000 GPUs. The model is implemented based on OpenCLIP [3,9]. We continued pretraining on our datasets from OpenAI's ViT-L/14-336 checkpoint [14] for 3 epochs, using an increased input resolution of 518 × 518. The private datasets were curated by Zhongshi Kangkai Technology Co., Ltd. following rigorous de-identification and anonymization protocols.

Table 1. 10 downstream datasets for comprehensive evaluation across anatomies.

Dataset	Anatomy	Labels
Abd-2k (in-house)	Abdomen	Intestinal obstruction, Dilated bowel, Aortic calcification, Kidney calculi, Urinary bladder calculi, Fracture, Hyperostosis, Scoliosis
AV-DDH [7]	Pelvis	Left/Right/Bilateral developmental dysplasia of the hip
BUU-LSPINE [10]	Spine	Left/Right laterolisthesis, Anterolisthesis, Retrolisthesis
ChestDR [19]	Chest	19 thoracic diseases
CHD-CXR [23]	Chest	Atrial septal defect, Ventricular septal, Patent ductus arteriosus
COVIDGR [17]	Chest	COVID-19 severity: Normal-PCR+, Mild, Moderate, Severe
Epic-PNA [8]	Chest	Pneumonia
Node21 [16]	Chest	Nodule
PelviXNet [2]	Pelvis	Hip and pelvic fracture
SFR-COVID-19 [11]	Chest	Typical, indeterminate, or atypical for COVID-19

For evaluation, we fine-tune DEER and the following state-of-the-art (SOTA) approaches on downstream tasks: Ark$^+$ [12], MedGemma (we only evaluate the MedSigLIP encoder) [15], RAD-DINO [13], and EVA-X [21]. We evaluate the above models on the 10 downstream datasets, including chest, abdominal, pelvic, and spinal DR scans, as shown in Table 1. Model performances are assessed using both the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC). We follow the official splits and report the mean and std. over five runs. If the official split does not provide a validation set, we construct one from the training set. In the absence of an official train/test split, we adopt five-fold cross-validation and report the mean and std. across folds. Statistical significance is evaluated using t-test.

3.2 Chest Disease Diagnosis

Reports and Labels Strengthen Model Performance. We conducted extensive experiments on chest DR images (Table 1) covering conventional pulmonary diseases (ChestDR), pediatric congenital heart diseases (CHD-CXR), COVID-19 (COVIDGR, SFR-COVID-19), pneumonia (Epic-PNA), and nodules (Node21). As shown in Fig. 2, models trained with image-report (DEER, MedGemma) or image-label (Ark$^+$) supervision consistently achieve higher performance across 6 downstream chest datasets. In contrast, image-only supervision (EVA-X, RAD-DINO) provides no explicit diagnostic guidance, making it difficult for models to disentangle and encode clinically meaningful semantics.

Image-Report Training Improves Generalization and Scalability. Notably, DEER, based on image-report pretraining, achieves leading performance on 5 out of 6 downstream chest datasets (Fig. 2). This training paradigm not only leverages image-report pairs routinely generated in clinical radiology practice without requiring additional human annotation, but also yields stronger generalization across diverse tasks. In contrast, image-label training requires either manual annotation or automated extraction from reports, which oversimplifies the rich findings information contained in the reports. Notably, the self-supervised image-report pretrained DEER outperforms Ark$^+$ trained on large-

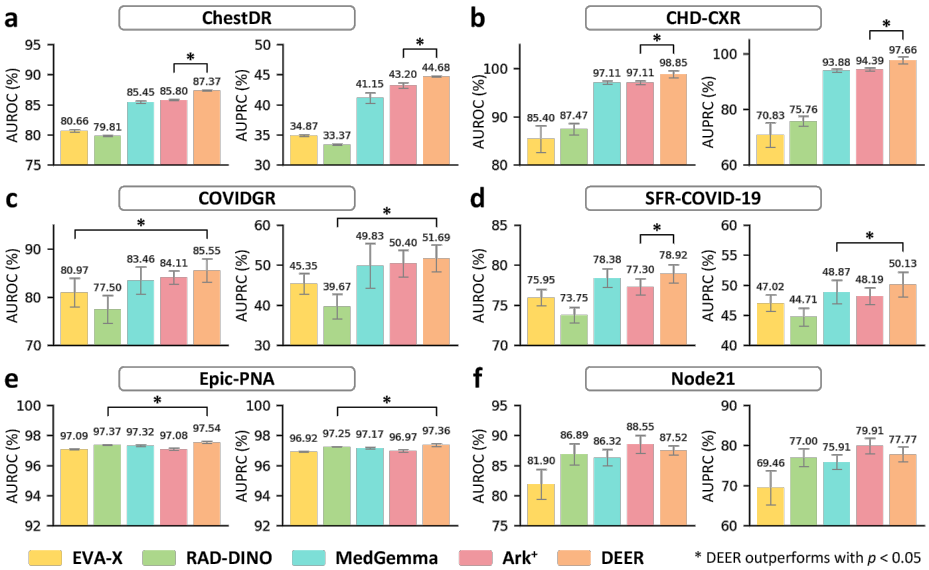

Fig. 2. DEER achieves superior performance in chest disease diagnosis. Image-report pretraining (DEER, MedGemma) learns more diagnostically discriminative representations, outperforming image-only training (EVA-X, RAD-DINO). Moreover, DEER, pretrained via self-supervised image-report learning, outperforms models trained with large-scale label supervision (Ark$^+$), demonstrating the feasibility of scaling self-supervised medical foundation models with superior diagnostic performance.

scale chest DR labels. This result highlights the effectiveness of self-supervised learning and its scalability toward larger medical foundation models.

3.3 Abdominal Disease Diagnosis

To evaluate the model's performance on the abdomen, we collected a dataset comprising 2,000 abdominal DR images across eight diseases (Abd-2k, Table 1). DEER demonstrates strong capability in abdominal DR analysis (Fig. 3a). Notably, this dataset exhibits a long-tailed, class-imbalanced distribution. Therefore, AUPRC better reflects the improvements of our model. In contrast, although abdominal diseases are clinically important, they remain largely underexplored in prior work. Existing DR foundation models show noticeable performance degradation when applied to abdominal cases (Fig. 3a).

3.4 Transferability to Pelvic and Spinal Scans

Chest and abdominal DR scans share overlapping anatomical regions, and abdominal scans frequently include the pelvic area. Learning from both types of DR therefore facilitates a more comprehensive understanding of trunk DR. We conducted extensive experiments on pelvic DR images (Table 1), including hip

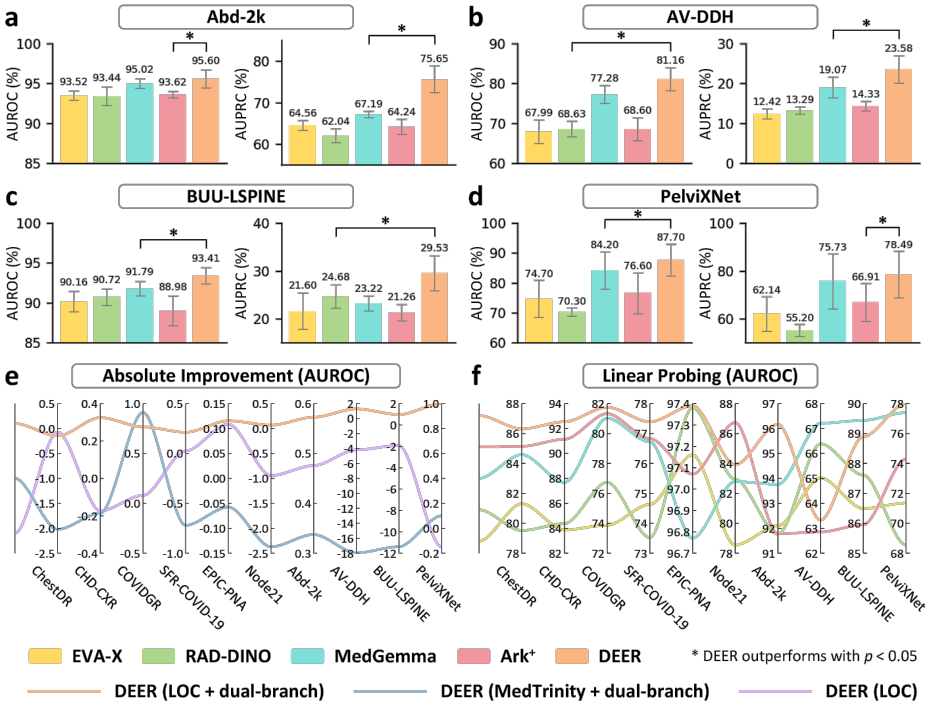

Fig. 3. (a) DEER demonstrates strong performance on abdominal DR. (b-d) DEER shows superior transferability on pelvic DR, while MedGemma performs better among baseline methods. This finding supports the advantage of image-report pretraining for improved generalization. (e) Incorporating LOC via dual-branch encoding further enhances performance, with greater gains in the abdomen and pelvis than in the chest. (f) DEER demonstrates superior linear probing performance on 7 out of 10 datasets.

developmental dysplasia (AV-DDH), lumbar spondylolisthesis (BUU-LSPINE), and fractures (PelviXNet). As shown in Fig. 3c-d, DEER demonstrates strong transferability to pelvic and spinal scans, with consistently superior performance.

In contrast, current DR foundation models exhibit limited generalization to the pelvic region (Fig. 3c-d). Among the baseline methods, MedGemma achieves the best cross-region transfer performance, whereas label-supervised approaches (Ark$^+$) achieve relatively lower performance. This observation further supports the effectiveness of image-report self-supervised pretraining in building foundation models with stronger generalization.

3.5 CT Localizers Further Boost the Performance

We further enhanced the embeddings by incorporating LOC and CT reports through the dual-branch encoding framework, leading to additional performance gains (Fig. 3e). Without dual-branch encoding, the entanglement of DR and LOC

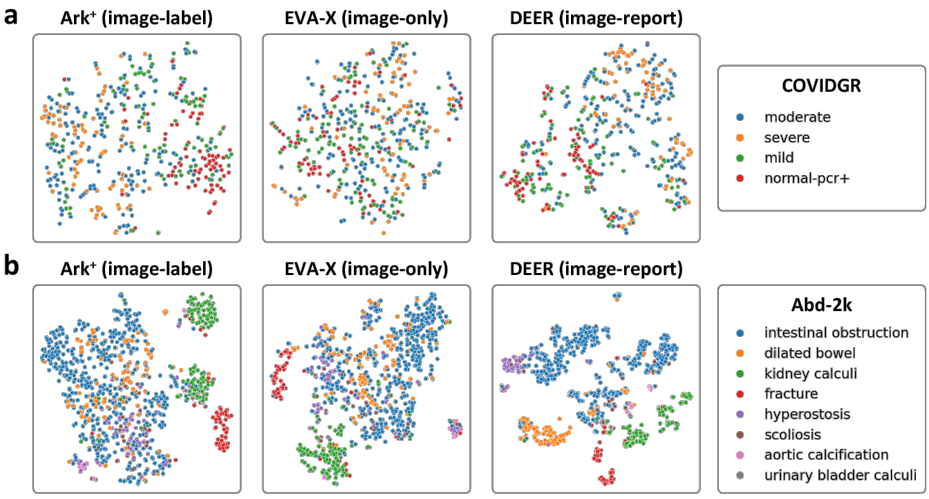

Fig. 4. (a) The label-supervised model (Ark$^+$) and the image-report model (DEER) show clearer feature separation on chest DR, whereas the image-only model yields more entangled clusters. (b) On abdominal tasks, DEER maintains stronger discriminability, while the current DR foundation model shows overlapping representations.

information often degrades representation quality and downstream performance (Fig. 3e). Notably, the improvement in the abdomen and pelvis settings was greater than that observed for chest DR. This difference may be attributed to the broader anatomical coverage of LOC compared with conventional chest DR, which enriches the image distribution of abdominal and pelvic regions.

To further demonstrate that LOC contributes to effective training benefits, we sampled 1 million non-DR image-text pairs from MedTrinity-25M [20], including CT, magnetic resonance, ultrasound, etc. These samples were combined with the DR data to train the model. As shown in Fig. 3e, compared with an equal amount of LOC data, incorporating other data sources does not yield performance gains and may even substantially degrade model performance. This finding further suggests that LOC and CT reports are essential for improving DR foundation model training, rather than simply increasing data volume.

3.6 Qualitative *t*-SNE Visualizations

The *t*-SNE (Fig. 4a) shows that both the label-supervised model (Ark$^+$) and the image-report self-supervised model (DEER) achieve clearer feature separation on chest DR, particularly between "severe" (orange) and "normal-pcr+" (red) cases. In contrast, the image-only self-supervised model (EVA-X) produces more entangled features with overlapping diagnostic clusters. On abdominal tasks, DEER maintains stronger feature discriminability, while the current DR foundation model exhibits mixed and less structured representations (Fig. 4b). This demonstrates the feasibility of leveraging self-supervised image-text learning to train

foundation models with strong diagnostic discriminative representations. Also, DEER achieves superior linear probing performance on 7 out of 10 datasets, further validating the effectiveness of the learned representations (Fig. 3f).

4 Discussion

DR is the primary tool for screening and diagnosis [18], especially in under-resourced regions [6]. DEER demonstrates strong performance across diverse chest, abdominal, and pelvic tasks, covering underexplored abdominal DR. This offers significant clinical value for intelligent DR diagnosis and could be deployed in remote areas. By leveraging previously overlooked CT localizers and reports, our image-text self-supervised approach outperforms large-scale label-supervised methods. These findings provide insights for scaling DR foundation models with multimodal CT integration. Future directions include evaluating intelligent DR models for complex diseases that typically require CT for diagnosis, as well as developing a full-body DR foundation model to enhance clinical coverage.

Acknowledgments. The authors acknowledge supports from Science and Technology Commission of Shanghai Municipality (24511104100), National Natural Science Foundation of China (82572314, U22A2051, 62271246, 62501594), National Key Research and Development Program of China (2022YFC2405200).

Disclosure of Interests. The authors declare no competing interests.

References

1. Chambon, P., Delbrouck, J.B., Sounack, T., Huang, S.C., et al.: Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats (2024), https://arxiv.org/abs/2405.19538
2. Cheng, C.T., et al.: A scalable physician-level deep learning algorithm detects universal trauma on pelvic radiographs. Nature Communications **12**(1), 1066 (2021)
3. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., et al.: Reproducible scaling laws for contrastive language-image learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2818–2829 (2023)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
5. Du, J., et al.: Ret-clip: A retinal image foundation model pre-trained with clinical diagnostic reports. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. pp. 709–719. Springer Nature Switzerland, Cham (2024)
6. Frija, G., Blažić, I., Frush, D.P., Hierath, M., et al.: How to improve access to medical imaging in low- and middle-income countries ? eClinicalMedicine **38** (2021)
7. Haddad, B., AlHosanie, T., et al.: Acetabular-vision hip developmental dysplasia's (av-ddh). Mendeley Data (2025). https://doi.org/10.17632/4gvcb6gmh2.1
8. Hira, M.I.K., Bithee, M.M.A., Ahmed, S., Akter, L., Anonna, M.J.M.: A primary chest x-ray dataset of normal and pneumonia cases from epic chittagong, bangladesh. Mendeley Data, V2 (2025). https://doi.org/10.17632/wndbd5r26y.2
9. Ilharco, G., et al.: OpenCLIP (2021), https://doi.org/10.5281/zenodo.5143773

10. Klinwichit, P., Yookwan, W., Limchareon, S., et al.: Buu-lspine: A thai open lumbar spine dataset for spondylolisthesis detection. Applied Sciences **13**(15) (2023)
11. Lakhani, P., Mongan, J., Singhal, C., Zhou, Q., et al.: The 2021 siim-fisabio-rsna machine learning covid-19 challenge: Annotation and standard exam classification of covid-19 chest radiographs. Journal of Digital Imaging **36**(1), 365–372 (2023)
12. Ma, D., Pang, J., Gotway, M.B., Liang, J.: A fully open AI foundation model applied to chest radiography. Nature **643**(8071), 488–498 (2025)
13. Pérez-García, F., Sharma, H., et al.: Exploring scalable medical image encoders beyond text supervision. Nature Machine Intelligence **7**(1), 119–130 (2025)
14. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021)
15. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., et al.: Medgemma technical report (2025), https://arxiv.org/abs/2507.05201
16. Sogancioglu, E., et al.: Node21 (2021). https://doi.org/10.5281/zenodo.5548363
17. Tabik, S., Gómez-Ríos, A., Martín-Rodríguez, J.L., et al.: Covidgr dataset and covid-sdnet methodology for predicting covid-19 based on chest x-ray images. IEEE Journal of Biomedical and Health Informatics **24**(12), 3595–3605 (2020)
18. United Nations Scientific Committee on the Effects of Atomic Radiation: Sources, Effects and Risks of Ionizing Radiation, United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) 2020/2021 Report. United Nations, 2021 edn. (2022)
19. Wang, D., Wang, X., et al.: A real-world dataset and benchmark for foundation model adaptation in medical image classification. Scientific Data **10**(1), 574 (2023)
20. Xie, Y., Zhou, C., Gao, L., Wu, J., Li, X., Zhou, H.Y., Liu, S., Xing, L., Zou, J., Xie, C., Zhou, Y.: Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine (2025), https://arxiv.org/abs/2408.02900
21. Yao, J., Wang, X., Song, Y., et al.: EVA-X: a foundation model for general chest x-ray analysis with self-supervised learning. npj Digital Medicine **8**(1), 678 (2025)
22. Zhang, X., et al.: Rexgradient-160k: A large-scale publicly available dataset of chest radiographs with free-text reports (2025), https://arxiv.org/abs/2505.00228
23. Zhixin, L., Gang, L., Zhixian, J., Sibao, W., Silin, P.: Chd-cxr: a de-identified publicly available dataset of chest x-ray for congenital heart disease. Frontiers in Cardiovascular Medicine **11** (2024). https://doi.org/10.3389/fcvm.2024.1351965