Medical Image Analysis 100 (2025) 103389

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Editorial for Special Issue on Foundation Models for Medical Image Analysis

The need for machine learning methods that work on applications with massive and diverse data, has resulted recently in a noticeable increase in the interest surrounding pre-training foundation models, particularly in the domains of natural language processing and computer vision. This trend has led to the emergence of significant works, such as Vision Transformers (ViT) (Dosovitskiy et al., 2021), Contrastive Language-Image Pretraining (CLIP) (Radford et al., 2021), and Segment Anything (SAM) (Kirillov et al.). They all derive from the Transformer (Vaswani et al., 2017) architecture, also primarily utilized in Medical applications (Chen et al., 2024). The expansion in the size (number of parameters) of these transformers comes with an increase in their capacity, simultaneously requiring larger volumes of training data following the scaling law. Despite these advancements, the scarcity of publicly available and well-annotated data, particularly in sectors like medical imaging, has presented a challenge for the training of universal foundation models to facilitate the vast number of applications in such domains. Instead, building a spectrum of foundation models is initiated to facilitate a relatively zoomed-in scope of downstream clinical tasks (Zhang and Metaxas, 2024). This pioneering perspective has been reverberated by a group of recently released foundation models dedicated to particular sub-domains of medical imaging, e.g., foundation models for retinal images (Zhou et al., 2023), dermatological images (Kim et al., 2024), pathology images (Huang et al., 2023, Lu et al., 2024, Chen et al., 2024, Xu et al., 2024), radiography (Huang et al., 2024), and multimodality images (Wang et al., 2024). Moreover, this topic of foundation models also attract more attention from researchers in medical image analysis, which results in a series of workshops (Deng et al., 2024).

The conventional method of fine-tuning pre-trained models necessitates a substantial quantity of domain-specific labeled data. The scarcity of such well-annotated data, particularly for domains like medical imaging, has presented a significant challenge for model tuning for each specific task. One alternative and feasible approach involves developing efficient model adaptation methodologies for medical applications suitable for the aforementioned foundation models. Particularly, the advent of prompt engineering methods and the accessibility of the foundation models have rendered this approach increasingly viable. In this context, a foundation model, often trained on a vast amount of diverse data, can function as the basis for developing medical applications using a limited number of cases presented as prompts. Consequently, presenting a small set of distinct, differentiable sample cases derived from real-world medical practice scenarios to be used for model adaptation is tenable. It also aligns with the educational process for professionals in the medical and biomedical fields, such as medical residents and bioengineers.

This is the first special issue addressing advances in foundational models for medical image analysis and it comprises fifteen peer-reviewed original articles dedicated to topics pertinent to foundational models for medical image analysis. These articles feature innovative solutions for developing and adapting foundational models for medical image analysis, as well as dataset composition techniques crucial for their training. The focus is on advancing new approaches to fundamental clinical and research tasks in medical imaging, with the goal of achieving higher efficiency and improved generalizability.

The accepted papers can be classified into three main categories. Seven papers primarily focused on the development of foundational models tailored to specific subdomains and scenarios within a clinical context, such as foundational models for ultrasound images. Five papers were centered on the adaptation or fine-tuning of existing foundational models to enhance the efficiency of analyzing specialized medical image tasks in a cost-effective and data-efficient manner. Finally, three papers introduced original methodologies for the construction of extensive datasets to facilitate the training of domain-aware foundational models.

In the first group of papers, seven articles focused on developing methods for learning foundation models that use imaging modalities or other clinical data for clinical applications. These foundation models often utilize large amounts of imaging and other associated data via self-supervised learning, which are then generalized to a variety of specific downstream tasks. (Jiao et al., 2024) introduce USFM, a universal foundation model for ultrasound (US) images optimized for label efficiency across diverse tasks and organs, based on a large-scale, diverse US database with over two million images. USFM was pre-trained in a self-supervised manner using organ-balanced sampled data. To handle low-quality images, they developed a spatial-frequency dual-masked modeling approach, incorporating spatial noise addition and frequency band-stop masking techniques. Extensive experiments demonstrate USFM's versatility and effectiveness in segmentation, classification, and enhancement tasks. (Hua et al., 2024) present PathoDuet, a self-supervised learning framework for histopathological images featuring pre-trained models. It introduces a pretext token and tasks to leverage image relationships, such as magnification and staining variations. Two tasks, cross-scale positioning and cross-stain transferring are used for pretraining on Hematoxylin and Eosin (H&E) images and adapting to immunohistochemistry (IHC) images. The models are evaluated across various tasks, including colorectal cancer subtyping and IHC marker expression prediction. Results show that PathoDuet outperforms existing methods, demonstrating the effectiveness of its pretext tasks. (Kang et al., 2024) introduce a deblurring masked image modeling (MIM) approach tailored for ultrasound (US) images, addressing their high noise-to-signal ratio. By integrating deblurring into pre-training,

https://doi.org/10.1016/j.media.2024.103389

Available online 6 November 2024

the model enhances detail recovery crucial for downstream tasks. Utilizing a multi-scale hierarchical encoder, the method improves performance on pixel-wise tasks such as segmentation. Experiments with 280,000 US images show that this approach achieves state-of-the-art results across various diagnostic and segmentation tasks, demonstrating its efficacy and potential as a specialized US image analysis model. (Cox et al., 2024) present BrainSegFounder, a novel 3D medical foundation model for multimodal neuroimage segmentation using self-supervised training. The first stage learns anatomical features from a large dataset of healthy brain MRIs, while the second stage focuses on disease-specific attributes. This method reduces data needs and adapts to various imaging modalities. Evaluations on the BraTS and ATLAS v2.0 datasets show BrainSegFounder significantly outperforms previous supervised models, demonstrating the benefits of using extensive unlabeled data and complex models for improved segmentation accuracy.

Another set of articles in the first category involves modalities other than medical imaging, which are either employed as a form of supervision or as additional information for the model training. (Li et al., 2024) present an iterative vision-language framework that refines radiology reports by emphasizing key information using a clinical dictionary and knowledge enhancement metrics. It improves the relevance of reports for fine-grained tasks through progressive learning. The framework excels in medical image analysis tasks, surpassing seven state-of-the-art methods in both fine-tuning and zero-shot settings, highlighting its potential for clinical applications. (Liu et al., 2024) introduce the Universal Model, a flexible framework that adapts to multiple datasets and new classes like organs and tumors. It uses a language-driven parameter generator for improved semantic encoding and lightweight, class-specific output heads. Tested on 3,410 CT volumes from 14 datasets and 6,173 volumes from four external sources, the model achieves top rankings in six Medical Segmentation Decathlon tasks and excels on the Beyond The Cranial Vault dataset. It is highly efficient, generalizes well, and supports easy addition of new classes without forgetting old ones. (Xie et al., 2024) introduce Masked Medical Image Modelling (MedIM), which uses radiological reports to guide masking and enhance image representation. MedIM employs Knowledge-Driven Masking (KDM) and Sentence-Driven Masking (SDM) to focus on clinically relevant image regions. Experiments show that MedIM achieves superior performance compared to traditional MIM and pre-training methods.

The articles in the second category investigate a range of techniques for effective model adaptation in specific image analysis problems. Prompt learning is efficient when leveraging pre-trained knowledge for new tasks with minimal additional training. This technique facilitates rapid model adaptation to diverse contexts and specific requirements by providing targeted instructions or examples, thus improving performance and efficiency in various applications. (Peng et al., 2024) present a novel prompt learning method for multimodal models diagnosing neurological disorders by using GPT-4 to identify relevant disease concepts and evaluate the similarity between these concepts and image patches. It then reduces the impact of irrelevant patches and constructs a graph to extract structural information, which prompts pre-trained models. This approach outperforms existing methods and is validated by clinicians. (Zu et al., 2024) explore Embedded Prompt Tuning (EPT) for adapting foundation models to medical image classification. EPT embeds prompt tokens into expanded channels, addressing limitations of existing methods and mitigating feature distribution anomalies during pre-training. Results show EPT significantly outperforms state-of-the-art techniques in few-shot medical image classification and is highly efficient in fine-tuning.

The other three papers in the second category focus on the model adaptation from other domains due to the complexity of medical image data. (Gong et al., 2024) present a method for adapting the Segment Anything Model (SAM) from 2D to 3D medical image segmentation. The approach involves modifying the architecture to handle 3D data while keeping most pre-trained parameters fixed and adding only a few spatial adapters. The method effectively bridges the domain gap and outperforms existing models on three of four tumor segmentation tasks, with significant improvements in kidney, pancreas, and colon cancer segmentation, and performs similarly for liver tumors. Comparisons with other adapters show substantial performance gains. (Song et al., 2024) enhance the diagnostic use of large language models (LLMs) by replacing the text branch with a classification head, reducing parameters. It introduces a contextual multi-token engine for adaptive diagnostic token generation and an information emitter module to transfer data from the image to diagnostic tokens. Experiments demonstrate the effectiveness of these innovations. (Chen et al., 2024) present MA-SAM, a modality-agnostic framework adapting SAM for volumetric and video medical data. It uses parameter-efficient fine-tuning and 3D adapters to extract third-dimensional info. Evaluations on 11 datasets show that MA-SAM outperforms state-of-the-art 3D methods without prompts and excels in tumor segmentation with prompts.

In the last paper category, three articles deal with the root problem of building large-scale datasets for model training. The advanced training scheme for foundation models goes beyond the conventional image-label pairs, involving richer and various formats of supervision. (Holste et al., 2024) present a publicly available benchmark dataset comprising over 350,000 chest X-rays (CXRs), each annotated with one or more of 26 clinical findings characterized by a long-tailed distribution. The authors identify key themes from successful methods in multilabel medical image classification and offer practical recommendations for addressing long-tailed data. Additionally, they propose leveraging vision-language foundation models to advance few- and zero-shot disease classification. (Hu et al., 2024) present an adapted LLM for label extraction, achieving 62% higher accuracy, and refine labels with expert feedback. It introduces Medical-CXR-VQA, a dataset for chest X-ray VQA involving detailed clinical questions. A novel VQA method is proposed using spatial, semantic, and implicit relationship graphs with graph attention learning logical reasoning paths. It demonstrates LLM prompt engineering and exhibits strong evidence and faithfulness for clinical use. (Li et al., 2024) introduce AbdomenAtlas, the largest abdominal CT dataset, comprising 20,460 3D CT volumes from 112 hospitals with diverse populations and facilities. This dataset includes 673,000 high-quality anatomical masks annotated by a team of 10 radiologists using AI algorithms. Initial manual annotations were done for 22 anatomical structures in 5,246 volumes, followed by a semiautomatic process where radiologists refined AI-predicted annotations. AbdomenAtlas is crucial for advancing AI development by providing extensive pre-trained models and alleviating radiologists' annotation workload.

The fifteen articles featured in this special issue comprise the latest advancements in the utilization of foundation models in medical image analysis. The papers demonstrate a diverse range of foundation model learning and adaptation approaches applied to various types of medical imaging data (Zhang and Metaxas, 2024). Furthermore, several works in this collection introduce novel learning techniques and promising avenues for adapting existing large-scale pre-training models to enhance their applicability in clinical settings. All contributions to this special issue were required to make the models and datasets they introduced open source. The significant clinically relevant advances highlighted in the accepted papers underscore the need to enhance further research efforts to advance techniques related to foundational models for medical image analytics.

References

Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., et al., 2024. Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. Medical Image Analysis 97, 103280.

Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al., 2024. Towards a general-purpose foundation model for computational pathology. Nature Medicine 30 (3), 850–862.

Chen, C., Miao, J., Wu, D., Yan, Z., Kim, S., Hu, J., Zhong, A., Liu, Z., Sun, L., Li, X., et al., 2024. Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. Medical Image Analysis, 103310.
Cox, J., Liu, P., Stolte, S.E., Yang, Y., Liu, K., See, K.B., Ju, H., Fang, R., 2024. Brainsegfounder: Towards 3d foundation models for neuroimage segmentation. Medical Image Analysis, 103301.
Deng, Z., Shen, Y., Kim, H.J., Jeong, W., Aviles-Rivero, A.I., He, J., S. Foundation Models for General Medical AI. MedAGI, 2024. Lecture Notes in Computer Science, vol 15184. Springer.
Dosovitskiy, A., Beyer, L., Kolesnikov, A., D.Weissenborn, X.Zhai, Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2021. An image is worth 16×16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations.
Gong, S., Zhong, Y., Ma, W., Li, J., Wang, Z., Zhang, J., Heng, P.-A., Dou, Q., 2024. 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable medical image segmentation. Medical Image Analysis.
Holste, G., Zhou, Y., Wang, S., Jaiswal, A., Lin, M., Zhuge, S., Yang, Y., Kim, D., Nguyen-Mau, T.-H., Tran, M.-T., et al., 2024. Towards long-tailed, multi-label disease classification from chest x-ray: Overview of the cxr-lt challenge. Medical Image Analysis, 103224.
Hu, X., Gu, L., Kobayashi, K., Liu, L., Zhang, M., Harada, T., Summers, R., Zhu, Y., 2024. Interpretable medical image visual question answering via multi-modal relationship graph learning. Medical Image Analysis, 103279.
Hua, S., Yan, F., Shen, T., Zhang, X., 2024. Pathoduet: Foundation models for pathological slide analysis of h&e and ihc stains. Medical Image Analysis, 103289.
Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J., 2023. A visual–language foundation model for pathology image analysis using medical Twitter. Nature Medicine 29 (9), 2307–2316.
Huang, W., Li, C., Zhou, H.-Y., Yang, H., Liu, J., Liang, Y., Zheng, H., Zhang, S., Wang, S., 2024. Enhancing representation in radiography-reports foundation model: A granular alignment algorithm using masked contrastive learning. Nature Communications.
Jiao, J., Zhou, J., Li, X., Xia, M., Huang, Y., Huang, L., Wang, N., Zhang, X., Zhou, S., Wang, Y., et al., 2024. Usfm: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. Medical Image Analysis 96, 103202.
Kang, Q., Lao, Q., Gao, J., Liu, J., Yi, H., Ma, B., Zhang, X., Li, K., 2024. Deblurring masked image modeling for ultrasound image analysis. Medical Image Analysis 97, 103256.
Kim, C., Gadgil, S.U., DeGrave, A.J., Omiye, J.A., Cai, Z.R., Daneshjou, R., Lee, S.-I., 2024. Transparent medical image AI via an image–text foundation model grounded in medical literature. Nature Medicine 30 (4), 1154–1165.
A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, R. Girshick, Segment anything, In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4015-4026).
Li, C., Huang, W., Yang, H., Liu, J., Wang, S., 2024. Enhancing the vision-language foundation model with key semantic knowledge-emphasized report refinement. Medical Image Analysis.
Li, W., Qu, C., Chen, X., Bassi, P.R., Shi, Y., Lai, Y., Yu, Q., Xue, H., Chen, Y., Lin, X., et al., 2024. Abdomenatlas: A large-scale, detailed-annotated, & multicenter dataset for efficient transfer learning and open algorithmic benchmarking. Medical Image Analysis, 103285.
Liu, J., Zhang, Y., Wang, K., Yavuz, M.C., Chen, X., Yuan, Y., Li, H., Yang, Y., Yuille, A., Tang, Y., et al., 2024. Universal and extensible language-vision models for organ segmentation and tumor detection from abdominal computed tomography. Medical Image Analysis, 103226.
Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al., 2024. A visual-language foundation model for computational pathology. Nature Medicine 30 (3), 863–874.
Peng, L., Cai, S., Wu, Z., Shang, H., Zhu, X., Li, X., 2024. Mmgpl: Multimodal medical data analysis with graph prompt learning. Medical Image Analysis, 103225.
Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: Proceedings of the International conference on machine learning, pp. 8748–8763.
Song, M., Wang, J., Yu, Z., Wang, J., Yang, L., Lu, Y., Li, B., Wang, X., Wang, X., Huang, Q., et al., 2024. Pneumollm: Harnessing the power of large language model for pneumoconiosis diagnosis. Medical Image Analysis, 103248.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., et al., 2017. Attention is all you need [j]. Advances in neural information processing systems 30 (1), 261–272.
X. Wang, X. Zhang, G. Wang, J. He, Z. Li, W. Zhu, Y. Guo, Q. Dou, X. Li, D. Wang, et al., Openmedlab: An open-source platform for multi-modality foundation models in medicine, arXiv preprint arXiv:2402.18028 (2024).
Xie, Y., Gu, L., Harada, T., Zhang, J., Xia, Y., Wu, Q., 2024. Rethinking masked image modeling for medical image representation. Medical Image Analysis, 103304.
Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al., 2024. A whole-slide foundation model for digital pathology from real-world data. Nature 1–8.
Zhang, S., Metaxas, D., 2024. On the challenges and perspectives of foundation models for medical image analysis. Medical Image Analysis 91, 102996.
Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., Liu, T., Xu, M., Lozano, M.G., Woodward-Court, P., et al., 2023. A foundation model for generalizable disease detection from retinal images. Nature 622 (7981), 156–163.
Zu, W., Xie, S., Zhao, Q., Li, G., Ma, L., 2024. Embedded prompt tuning: Towards enhanced calibration of pretrained models for medical images. Medical Image Analysis 97, 103258.

Xiaosong Wang[a], Dequan Wang[a,b], Xiaoxiao Li[c], Jens Rittscher[d], Dimitris Metaxas[e], Shaoting Zhang[a,*]
[a] *Shanghai AI Laboratory, Shanghai 200232, China*
[b] *Qing Yuan Research Institute, Shanghai Jiao Tong University, Shanghai 200240, China*
[c] *School of Electrical and Computer Engineering, University of British Columbia, Vancouver BC V6T 1Z4, Canada*
[d] *Department of Engineering Science, University of Oxford, Oxford OX3 7DQ, UK*
[e] *Department of Computer Science, Rutgers University, Piscataway, NJ 08854, USA*

[*] Corresponding author.
E-mail address: zhangshaoting@pjlab.org.cn (S. Zhang).