Medical Image Analysis 91 (2024) 102996

Contents lists available at ScienceDirect

Medical Image Analysis

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

On the challenges and perspectives of foundation models for medical image analysis

Shaoting Zhang [a,b,*], Dimitris Metaxas [c]

[a] University of Electronic Science and Technology of China, Chengdu, Sichuan, China
[b] Shanghai Artificial Intelligence Laboratory, Shanghai, China
[c] Rutgers University, New Brunswick, NJ, USA

ARTICLE INFO

Keywords:
Foundation models

ABSTRACT

This article discusses the opportunities, applications and future directions of large-scale pretrained models, i.e., foundation models, which promise to significantly improve the analysis of medical images. Medical foundation models have immense potential in solving a wide range of downstream tasks, as they can help to accelerate the development of accurate and robust models, reduce the dependence on large amounts of labeled data, preserve the privacy and confidentiality of patient data. Specifically, we illustrate the "spectrum" of medical foundation models, ranging from general imaging models, modality-specific models, to organ/task-specific models, and highlight their challenges, opportunities and applications. We also discuss how foundation models can be leveraged in downstream medical tasks to enhance the accuracy and efficiency of medical image analysis, leading to more precise diagnosis and treatment decisions.

1. Introduction

A salient distinction exists between traditional pretrained models and contemporary foundation models. The former (Deng et al., 2009; Vaswani et al., 2017; Devlin et al., 2018; Radford et al., 2018, 2019; Dosovitskiy et al., 2021) typically require extensive supervised fine-tuning to address specific downstream tasks, whereas the latter are capable of employing few-shot learning, zero-shot learning, or prompt engineering to manage a wide variety of tasks with a singular set of model weights (Brown et al., 2020; Chowdhery et al., 2022; Ouyang et al., 2022; Touvron et al., 2023a,b; Google, 2023). Consequently, foundation models exhibit considerable generalizability and adaptability, positioning them as a focal point of recent machine learning (ML) research.

In the field of medical image analysis, however, task-specific ML models are still the main methods used, especially for clinical applications such as computer-aided disease diagnosis. Developing medical foundation models presents a significant challenge due to the diverse imaging modalities used in medicine, which differ greatly from natural images and are based on a spectrum of physics-based properties and energy sources. These modalities are based on the use of light, electrons, lasers, X-rays, ultrasound, nuclear physics, and magnetic resonance. The images produced span multiple scales, ranging from molecules and cells to organ systems and the full body. Therefore, it may be infeasible to develop a unified multi-scale foundation model trained from a combination of these multi-modality images.

In the following, we will investigate and present our vision for the "spectrum" of foundation models and their uses in medical image analysis, ranging from general vision models, modality-specific models, to organ and task-specific models (Fig. 1). Fortunately, the growing availability of high-quality publicly available annotated medical data has led to the gradual emergence of specialized foundational models with an innate capacity for generating more generalized representations of medical data. Therefore, foundation models trained with medical images and/or natural images in a self-supervised manner may serve as an improved solution basis for important clinical problems, will result in advances in the field of medical imaging, and will improve the efficacy and efficiency of disease diagnosis and treatment.

2. The spectrum of foundation models

Vision Foundation Models

A straightforward approach is to employ foundation models trained from natural images (Carion et al., 2020; Dosovitskiy et al., 2021; Zhai et al., 2022; Radford et al., 2021; He et al., 2022; Wang et al., 2022; Oquab et al., 2023), and then design sophisticated algorithms to solve downstream medical tasks.

* Corresponding author at: Shanghai Artificial Intelligence Laboratory, Shanghai, China.
E-mail address: zhangshaoting@uestc.edu.cn (S. Zhang).

https://doi.org/10.1016/j.media.2023.102996
Received 24 May 2023; Received in revised form 24 September 2023; Accepted 4 October 2023
Available online 12 October 2023

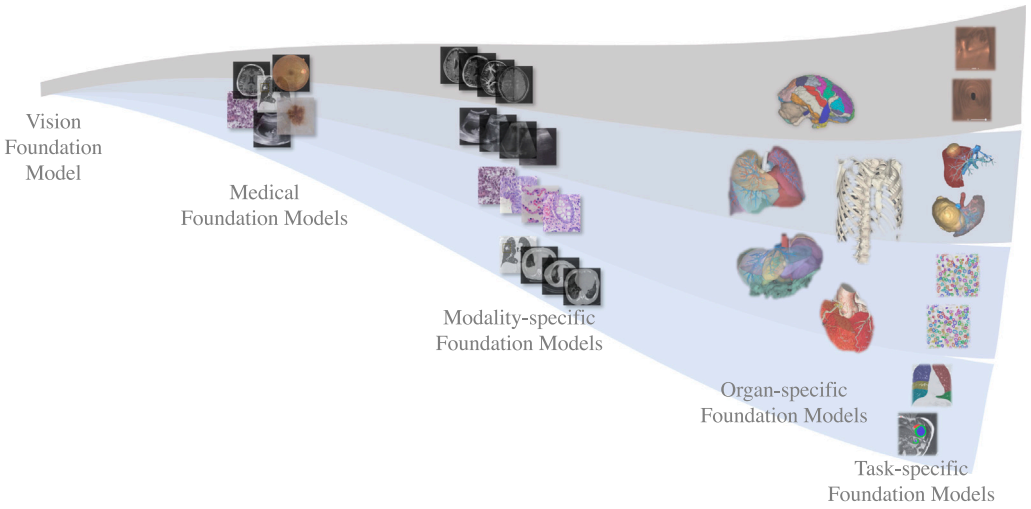

Fig. 1. The spectrum of foundation models in medical image analysis.

However, the lack of publicly available quality annotations in medical imaging has been the bottleneck for training large-scale deep learning models for many downstream clinical applications. It remains a tedious and time-consuming job for medical professionals to hand-label image data repeatedly, while providing a few differentiable sample cases is feasible and complies with the training process of medical residents. Vision foundation models, often trained on large-scale visual images of various modalities, could serve as the basis for building medical applications.

While vision foundation models can learn general representations, medical images have unique characteristics whose features and patterns differ significantly from those in natural images. Therefore, carefully designed algorithms are necessary to adapt them to domain-specific medical problems. Fine-tuning, additional adapters, prompting strategies, and specialized architectures or modifications are potential solutions to achieve optimal performance in medical problems.

For instance, the recent Segmentation Anything Model (SAM) (Kirillov et al., 2023), a promptable segmentation system with zero-shot generalization to unfamiliar objects and images, has demonstrated its impressive performance on natural images. However, its out-of-the-box performance on complex medical tasks such as pancreas, spine or cell nuclei segmentation is not satisfactory (He et al., 2023; Roy et al., 2023; Deng et al., 2023; Mazurowski et al., 2023; Shi et al., 2023). SAM can be further tuned to achieve state-of-the-art performance by leveraging high-quality downstream data and performing proper fine-tuning strategies (Ma and Wang, 2023; Paranjape et al., 2023; Cui et al., 2023), adding adapters with specially designed architectures (Wu et al., 2023a; Gong et al., 2023; Chen et al., 2023b), or effective prompts (Huang et al., 2023b; Cheng et al., 2023) with manual annotations. Venturing a step further, we could explore the possibility of combining the output of localization/detection algorithms with SAM or integrating SAM with image processing and visualization software like 3D slicer (Liu et al., 2023b). This fusion would pave the way for a robust pipeline tailored for complex medical applications.

In general, a unified foundation model approach cannot achieve state-of-the-art performance in many medical image analysis tasks due to large variations present in organs and important structures, texture, shape, size and topology (e.g., blood vessels), and imaging modalities. Furthermore, it is noteworthy to mention that there are parameter/data efficient tuning methods to adapt vision foundation models to address image analysis challenges arising from long-tail medical data.

Modality-specific Foundation Models

Depending on the pathology, various types of imaging modalities are employed for diagnostic and therapeutic purposes. They include X-rays, Computed Tomography (CT), Magnetic Resonance Image (MRI), Ultrasound imaging, and Positron Emission Tomography (PET). A modality-specific foundation model is specifically designed for a group of imaging modalities such as radiology images (including X-ray, CT, MR, and Ultrasound) (Ghesu et al., 2022), 3D images (stacks of 2D CT and MR images) (Chen et al., 2019), or a particular medical imaging modality which includes X-ray (Tiu et al., 2022), CT (Huang et al., 2023a; Wang et al., 2023d), endoscopy (Wang et al., 2023b), and pathology (Chen et al., 2022, 2023a; Vorontsov et al., 2023) images. It is then used to learn image-based features that are relevant to the intended use of the particular modality. For example, a CT-specific model may learn to identify features related to bone density and tissue contrast, while an MRI-specific model may learn to identify features related to soft tissue contrast and motion.

Vision foundation models trained on large-scale natural image datasets, can provide a strong starting point for a wide range of medical imaging analysis tasks. Using these vision foundation models, modality-specific foundation models can leverage the unique characteristics of each imaging modality and can result in models optimized for specific modalities. While they can then lead to higher accuracy and efficiency for the analysis tasks specific to that modality, they may not generalize well to other modalities.

Organ/Task-specific Foundation Models

More specifically, the foundation models could be tailored to a particular medical organ (Li et al., 2020b; Luo et al., 2022; Zhou et al., 2023) or diagnostic task, such as segmentation (Antonelli et al., 2022; Tang et al., 2022; Butoi et al., 2023). This use aims to address the challenges posed by the variability in organ appearance across medical images, as well as the diverse range of clinical tasks that are based on image analysis (Fig. 2).

Collecting data for training organ/task-specific foundation models can be challenging due to the need for large amounts of labeled data. Nevertheless, a well-trained organ/task-specific foundation model can provide better accuracy and interpretability as well as significantly reduce the amount of labeled data required for new tasks, as it has already learned relevant features from the previous training.

General vs. Specialized Foundation Models

By definition, specialists possess an in-depth understanding of a particular subject matter, whereas generalists have a broader purview, either within a single field or across multiple disciplines.

In the context of medical image analysis, a general AI system is characterized as a multitask and multimodal platform capable of performing a diverse range of tasks on multimodal images of different organs and diseases. They include classification, detection, segmentation and

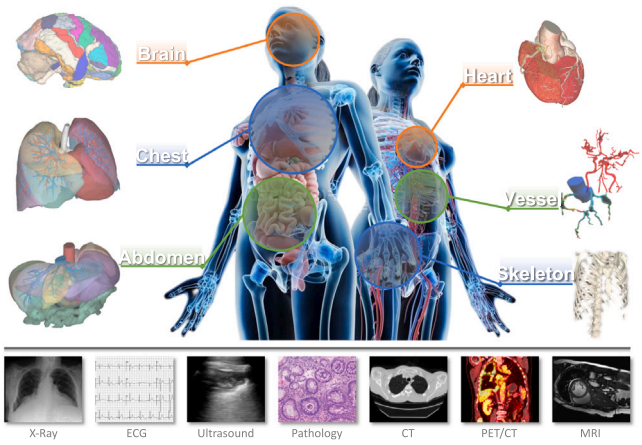

Fig. 2. Different image modalities have large image-level variations, which may cause difficulties when training a unified foundation model. Similarly each modality given the imaging formation differences, results in images with significant variations in organ appearance and related structures which determine which modality needs to be used given a target organ pathology. Given modalities, leveraging fine-grained models for learning organ appearance and pathology, will enable important clinical methods and tools such as robust computer-aided diagnosis and surgical planning.

registration and utilize a single set of model weights (Moor et al., 2023; Tu et al., 2023; Wu et al., 2023c). This approach and related models are shown in the left part of Fig. 1, which shows the spectrum of foundation models.

Conversely, specialized AI systems are designed to perform discrete clinical tasks, such as the detection of pulmonary nodules, the reconstruction of coronary arteries, or the diagnosis of hepatocellular carcinoma. These systems are generally confined to a particular organ and imaging modality, aligning more closely with the right side of the foundation model spectrum (see Fig. 1).

Within the computer science community, there is an emerging focus on the development of general AI frameworks. This shift is driven by the technical innovation inherent in exploring large, multimodal generative models capable of processing diverse types of medical data. However, research in both academia, medical institutions, and industry, largely remains focused on the development of specialized AI systems. This focus is attributable to several factors. Firstly, most existing state-of-the-art medical image analysis systems use a single type of imaging modality (unimodal) and are trained on a single task such as segmentation or classification. Secondly, AI currently serves primarily as an assistant to medical professionals who require targeted support, which is consistent with their medical training. Consequently, there is a pragmatic inclination toward designing specialized systems that excel in terms of performance and accuracy in particular tasks. Moreover, general type of AI systems tend to consume significantly greater computational resources and often lack the required accuracy.

Both specialized and general AI system approaches offer distinct advantages and are suited for different applications. Accordingly, we advocate for a comprehensive exploration of the foundation model spectrum to ascertain the optimal trade-off between developmental effort and practical efficacy.

3. Data requirements for foundation models

Data to Pretrain the Foundation Models

Data is the cornerstone for training all foundation models. Preparing data for medical foundation models has unique challenges and requires domain knowledge since medical data are expensive to collect, annotate interpret, and their quality varies significantly across hospitals and clinical studies.

Real-world images are 2D projections of the 3D world and therefore it is relatively easy to collect a large number of images that cover the variability of object(s)/scenes in terms of viewpoint, angles, scales, appearance and locations. To the contrary, medical images are acquired for a particular clinical purpose, through certain protocols and scanners that require use by an expert who controls the machine settings, including the view angles and scales. The image complexity and variability typically come from scanner differences, scanning protocols, and, most importantly, anatomical and other variations among individuals, disease appearance, location, and stage of the disease.

These unique properties of medical images manifest during the development of public datasets. Classical public datasets were collected for a specific purpose using certain protocols and scanners such as EyePACS (De Vente et al., 2023), SUN-SEG (Ji et al., 2022b), ISIC (Cassidy et al., 2022),Chestx-xay8 (Wang et al., 2017) CAMELYON (Litjens et al., 2018), and EndoVis (Allan et al., 2020), thus usually limited to a single modality and a specific anatomical area for a particular task. Recently, general-purpose datasets have been obtained using multiple protocols and scanners, such as Totalseg (Wasserthal et al., 2023), AMOS (Ji et al., 2022a), FLARE (Ma et al., 2022, 2023), autoPET (Gatidis et al., 2022), BraTS (Menze et al., 2014), ISLES (Hernandez Petzsche et al., 2022), DigestPath (Da et al., 2022), and MIMICS (Johnson et al., 2019). Correspondingly, the size of datasets has changed from small scale to large scale, and data variability is increasing. When preparing datasets to train medical foundation models, one should select the most representative cases to train generalizable models and cover corner cases instead of simply collecting large amounts of data.

In addition, instead of building a unified dataset covering all image modalities (e.g., an "ImageNet" Deng et al., 2009 type of database for medical data), a more feasible solution is to begin with modality-specific datasets and then attempt to merge those which are complementary.

Data Adaptation for Downstream Tasks

The emergence of foundation models signals a notable reversal from the previously encouraging trend of model openness and accessibility within the scientific community. Although pretrained instances of certain models like LLaMA (Touvron et al., 2023a,b) and SAM (Kirillov et al., 2023) are publicly accessible, models such as GPT-3 (Brown et al., 2020), and GPT-4 (OpenAI, 2023) are not publicly available; only API access is possible which is restricted to a limited number of users. Furthermore, the datasets employed for training foundation models are not made available to the wider research community. The computational and engineering resources required to train foundation models from the ground up are cost prohibitive for academia and most companies. This creates a barrier that prevents the vast majority of artificial intelligence researchers from participating in this pivotal ML research and methodology.

Recent advancements in ML have facilitated the efficient adaptation of foundation models for downstream tasks, requiring only a minimal number of training samples (e.g., prompt engineering and efficient methods for retraining). Recent research on the use of ML for medical image analysis has been establishing benchmarks and releasing datasets (Wang et al., 2023c; Yi et al., 2023). These efforts hold significant promise for enabling the effective deployment of large-scale foundation models to tackle an array of clinical challenges.

4. Applications and benefits of foundation models

Leveraging foundation models trained on large datasets to address specific medical needs is crucial for achieving accurate and reliable image analysis and disease diagnosis and prognosis, minimizing the need for data collection, reducing the time and cost associated with data labeling, and upholding patient data privacy and confidentiality.

Long-tailed Problems

Medical image analysis methods often face the challenging long-tail data scenario, caused by often heavily imbalanced datasets in

which many common disease cases coexist with relatively few rare disease cases. Consequently, the scarcity of data for training models to accurately identify these rare cases can lead to significant performance degradation issues. The few-shot setting aligns perfectly with the long-tailed scenario, which frequently arises in medical imaging when only a few rare disease cases/high-quality annotations are available.

The initial stage involves selecting the appropriate foundation model (including general, modality-specific, and organ/task-specific models), depending on the application and available resources. Then, data augmentation techniques are employed to augment the few annotated samples and make full use of the available supervised information. These techniques include image augmentation methods like rotation, cropping, color transformation, noise injection, and random erasing, as well as image generation techniques, e.g., GAN and diffusion-based models (Chambon et al., 2022; Pinaya et al., 2022; Ding et al., 2023).

Utilizing medical foundation models that have been trained on vast amounts of data can reduce the labeled data required for training, which minimizes the need for manual annotation by medical professionals. Additionally, these models can lead to more reliable diagnoses and treatment decisions.

Explainable and Generalizable Models

The lack of explainability in deep learning models can lead to distrust issues when clinicians are accustomed to making explainable clinical inferences. Similar to explainability, the generalizability of a model (a model trained on data from one medical center applied to data from other medical centers with significant variations or domain shifts) is also necessary due to the previously mentioned dataset limitations. Therefore, innovative methodologies are needed to improve the explainability and generalizability of the current models in order to be used effectively in clinical practice (Wang et al., 2023e).

Foundation models provide a unified framework that can support detection, segmentation, and classification tasks, which is essential for evidence-based decision-making. Moreover, these models are typically trained on large-scale datasets covering a wide range of medical centers, scanners, and protocols, resulting in promising generalizability of the learned feature representations.

Privacy Preserving Methods

While the computer vision community has an established history of open-sourcing large-scale datasets, such as the ImageNet, making publicly available large amounts of medical data is currently not possible due to regulatory and privacy issues. Foundation models offer an alternative way for knowledge sharing, while protecting patient privacy. Transfer learning techniques can be used to adapt the foundation model using a smaller dataset of interest, avoiding directly accessing massive raw data.

Furthermore, sharing multi-cohort knowledge in foundation models is also feasible through the use of federated learning (Li et al., 2020a; Kaissis et al., 2021), which enables training on data distributed across multiple institutions or devices without the data ever leaving the local machines. This paradigm ensures data privacy when training the foundation models on large distributed datasets.

Foundation models can also be used to create synthetic data (Ding et al., 2023) which can ensure data privacy preservation. Using generative models to create synthetic medical images which are statistically similar to real medical images, researchers can train models on these synthetic images instead of using real patient data.

Integration with Large Language Models

The study of vision-language models is gaining prominence due to their capacity for extracting nuanced information and learning superior representations. However, the majority of existing research predominantly concentrates on the analysis of X-ray images in conjunction with their corresponding reports (Zhang et al., 2022; Tiu et al., 2022; Zhou et al., 2022; Lee et al., 2023). The recent advances in large language models (LLMs) which are trained on vast amounts of text data have significantly improved natural language processing capabilities (Brown et al., 2020; Ouyang et al., 2022; Touvron et al., 2023a,b; Chowdhery et al., 2022; Google, 2023). The applications are further expanded beyond textual domains with the integration of vision models (OpenAI, 2023; Wu et al., 2023b; Alayrac et al., 2022; Li et al., 2023a; Driess et al., 2023; Wang et al., 2023a; Liu et al., 2023a). By combining language and vision data, these large-scale vision-language models have unlocked exciting possibilities for the future use of medical foundation models.

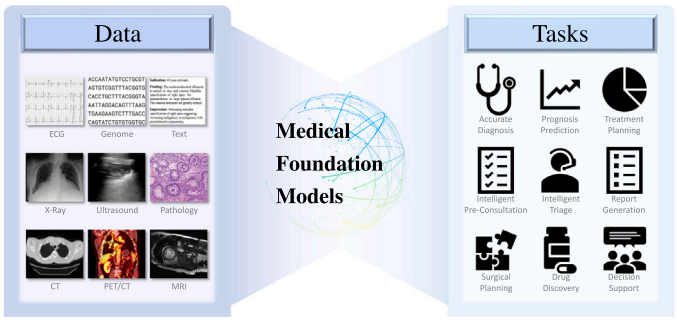

Fig. 3. Models trained on multiple medical data modalities can enable comprehensive clinical solutions.

Integrating medical image analysis systems with general domain LLMs or medical domain LLMs (Singhal et al., 2023a,b) holds immense potential for healthcare applications. For instance, these models can be trained to generate descriptive captions for medical images, facilitating automated radiology reports or succinct summaries of complex visuals. Furthermore, decision support systems can also benefit from associating visual features from medical images with text from patient records, providing accurate disease diagnosis and prognosis (Li et al., 2023b; Zhang et al., 2023; Wang et al., 2023f). However, these frameworks are still preliminary, as they usually integrate existing LLMs as a module by prompting, without fine-tuning and/or consolidating data modalities such as medical images to these models. The efforts to open-source foundation models and the ability to fine-tune them will be essential in healthcare.

5. Future directions of medical foundation models

We have discussed the challenges and opportunities of foundation models for medical image analysis. These insights can help us design more effective and generalizable foundation models. Since foundation models have only just begun to transform the way medical image systems are built and deployed worldwide, issues and challenges are still difficult to predict. We advocate that researchers from different institutions and disciplines need to collaborate to investigate and explore the spectrum of foundation models for medical image analysis, and contribute to the open-source community by releasing a family of pretrained models which will work on various imaging modalities.

Future directions include multi-modality foundation models, combining various data types (text, image, video, database, molecule) and scales (molecule, gene, cell, tissue, patient, population). Foundation models hold enormous promise as they can assimilate data from various imaging modalities and can incorporate non-imaging modalities to provide a more comprehensive understanding of a patient's condition and its assessment. By leveraging multi-modality foundation models, medical professionals can achieve a more accurate disease diagnosis and develop personalized treatment plans and disease prognosis. These models can potentially improve the overall quality of medical care by facilitating data sharing among different institutions, leading to more efficient and effective patient care and healthcare.

Advances in multi-modality foundation models can contribute to the development of clinical use cases targeting patients with different background and different diseases. A straightforward application is to support radiologists throughout their workflow, such as drafting structured radiology reports automatically and describing possible abnormalities, disease diagnosis and prognosis, as well as proposed treatment. Another possible use case is to assist surgeons. Integrating image, language, and audio modalities, surgeons can communicate with models to make real-time decisions in the operating room by detecting and identifying the anatomical location of important target structures often not clearly visible. The quest for comprehensive solutions to these and other medical problems is expected to intensify and the near future and produce the desired results (Fig. 3).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al., 2022. Flamingo: a visual language model for few-shot learning. Adv. Neural Inf. Process. Syst. 35, 23716–23736.

Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., Fuentes, F., Flouty, E., Mohammed, A., Pedersen, M., et al., 2020. 2018 Robotic scene segmentation challenge. arXiv preprint arXiv:2001.11190.

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al., 2022. The medical segmentation decathlon. Nature Commun. 13 (1), 4128.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 33, 1877–1901.

Butoi, V.I., Ortiz, J.J.G., Ma, T., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2023. UniverSeg: Universal medical image segmentation. arXiv preprint arXiv:2304.06131.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. In: European Conference on Computer Vision. Springer, pp. 213–229.

Cassidy, B., Kendrick, C., Brodzicki, A., Jaworek-Korjakowska, J., Yap, M.H., 2022. Analysis of the ISIC image datasets: Usage, benchmarks and recommendations. Med. Image Anal. 75, 102305.

Chambon, P., Bluethgen, C., Delbrouck, J.B., Van der Sluijs, R., Połacin, M., Chaves, J.M.Z., Abraham, T.M., Purohit, S., Langlotz, C.P., Chaudhari, A., 2022. RoentGen: vision-language foundation model for chest x-ray generation. arXiv preprint arXiv:2211.12737.

Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F., 2022. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16144–16155.

Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A.H., Shaban, M., et al., 2023a. A general-purpose self-supervised model for computational pathology. arXiv preprint arXiv:2308.15474.

Chen, S., Ma, K., Zheng, Y., 2019. Med3d: Transfer learning for 3d medical image analysis. arXiv preprint arXiv:1904.00625.

Chen, C., Miao, J., Wu, D., Yan, Z., Kim, S., Hu, J., Zhong, A., Liu, Z., Sun, L., Li, X., et al., 2023b. MA-SAM: Modality-agnostic SAM adaptation for 3D medical image segmentation. arXiv preprint arXiv:2309.08842.

Cheng, D., Qin, Z., Jiang, Z., Zhang, S., Lao, Q., Li, K., 2023. SAM on medical images: A comprehensive study on three prompt modes. arXiv preprint arXiv:2305.00035.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al., 2022. Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311.

Cui, C., Deng, R., Liu, Q., Yao, T., Bao, S., Remedios, L.W., Tang, Y., Huo, Y., 2023. All-in-sam: from weak annotation to pixel-wise nuclei segmentation with prompt-based finetuning. arXiv preprint arXiv:2307.00290.

Da, Q., Huang, X., Li, Z., Zuo, Y., Zhang, C., Liu, J., Chen, W., Li, J., Xu, D., Hu, Z., et al., 2022. DigestPath: A benchmark dataset with challenge review for the pathological detection and segmentation of digestive-system. Med. Image Anal. 80, 102485.

De Vente, C., Vermeer, K.A., Jaccard, N., Wang, H., Sun, H., Khader, F., Truhn, D., Aimyshev, T., Zhanibekuly, Y., Le, T.-D., et al., 2023. AIROGS: Artificial intelligence for robust glaucoma screening challenge. IEEE Trans. Med. Imaging.

Deng, R., Cui, C., Liu, Q., Yao, T., Remedios, L.W., Bao, S., Landman, B.A., Wheless, L.E., Coburn, L.A., Wilson, K.T., et al., 2023. Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging. arXiv preprint arXiv:2304.04155.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Ieee, pp. 248–255.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Ding, K., Zhou, M., Wang, H., Gevaert, O., Metaxas, D., Zhang, S., 2023. A large-scale synthetic pathological dataset for deep learning-enabled segmentation of breast cancer. Sci. Data 10 (1), 231.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations. URL: https://openreview.net/forum?id=YicbFdNTTy.

Driess, D., Xia, F., Sajjadi, M.S.M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., Zeng, A., Mordatch, I., Florence, P., 2023. PaLM-E: An embodied multimodal language model. arXiv Preprint arXiv:2303.03378.

Gatidis, S., Hepp, T., Früh, M., La Fougère, C., Nikolaou, K., Pfannenberg, C., Schölkopf, B., Küstner, T., Cyran, C., Rubin, D., 2022. A whole-body FDG-PET/CT dataset with manually annotated Tumor Lesions. Sci. Data 9 (1), 601.

Ghesu, F.C., Georgescu, B., Mansoor, A., Yoo, Y., Neumann, D., Patel, P., Vishwanath, R., Balter, J.M., Cao, Y., Grbic, S., et al., 2022. Self-supervised learning from 100 million medical images. arXiv preprint arXiv:2201.01283.

Gong, S., Zhong, Y., Ma, W., Li, J., Wang, Z., Zhang, J., Heng, P.A., Dou, Q., 2023. 3DSAM-adapter: Holistic adaptation of SAM from 2D to 3D for promptable medical image segmentation. arXiv preprint arXiv:2306.13465.

Google, 2023. PaLM 2 technical report. URL: https://ai.google/static/documents/palm2techreport.pdf.

He, S., Bao, R., Li, J., Grant, P.E., Ou, Y., 2023. Accuracy of segment-anything model (SAM) in medical image segmentation tasks. arXiv preprint arXiv:2304.09324.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009.

Hernandez Petzsche, M.R., de la Rosa, E., Hanning, U., Wiest, R., Valenzuela, W., Reyes, M., Meyer, M., Liew, S.L., Kofler, F., Ezhov, I., et al., 2022. ISLES 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. Scientific data 9 (1), 762.

Huang, Z., Wang, H., Deng, Z., Ye, J., Su, Y., Sun, H., He, J., Gu, Y., Gu, L., Zhang, S., et al., 2023a. STU-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. arXiv preprint arXiv:2304.06716.

Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., Yu, J., Chen, J., Chen, C., et al., 2023b. Segment anything model for medical images? arXiv preprint arXiv:2304.14660.

Ji, Y., Bai, H., Yang, J., Ge, C., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., Wan, X., et al., 2022a. AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. arXiv preprint arXiv:2206.08023.

Ji, G.P., Xiao, G., Chou, Y.C., Fan, D.P., Zhao, K., Chen, G., Van Gool, L., 2022b. Video polyp segmentation: A deep learning perspective. Mach. Intell. Res. 19 (6), 531–549.

Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S., 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data 6 (1), 317.

Kaissis, G., Ziller, A., Passerat Palmbach, J., Ryffel, T., Usynin, D., Trask, A., Lima, Jr., I., Mancuso, J., Jungmann, F., Steinborn, M.M., et al., 2021. End-to-end privacy preserving deep learning on multi-institutional medical imaging. Nat. Mach. Intell. 3 (6), 473–484.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R., 2023. Segment anything. arXiv:2304.02643.

Lee, H., Kim, W., Kim, J.H., Kim, T., Kim, J., Sunwoo, L., Choi, E., 2023. Unified chest X-ray and radiology report generation model with multi-view chest X-rays. arXiv preprint arXiv:2302.12172.

Li, X., Gu, Y., Dvornek, N., Staib, L.H., Ventola, P., Duncan, J.S., 2020a. Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. Med. Image Anal. 65, 101765.

Li, X., Jia, M., Islam, M.T., Yu, L., Xing, L., 2020b. Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis. IEEE Trans. Med. Imaging 39 (12), 4023–4033.

Li, J., Li, D., Savarese, S., Hoi, S., 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597.

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J., 2023b. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. arXiv preprint arXiv:2306.00890.

Litjens, G., Bandi, P., Ehteshami Bejnordi, B., Geessink, O., Balkenhol, M., Bult, P., Halilovic, A., Hermsen, M., van de Loo, R., Vogels, R., et al., 2018. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. GigaScience 7 (6), giy065.

Liu, H., Li, C., Wu, Q., Lee, Y.J., 2023a. Visual instruction tuning. arXiv preprint arXiv:2304.08485.

Liu, Y., Zhang, J., She, Z., Kheradmand, A., Armand, M., 2023b. Samm (segment any medical model): A 3d slicer integration to sam. arXiv preprint arXiv:2304.05622.

Luo, X., Liao, W., Xiao, J., Chen, J., Song, T., Zhang, X., Li, K., Metaxas, D.N., Wang, G., Zhang, S., 2022. WORD: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from CT image. Med. Image Anal. 82, 102642.

Ma, J., Wang, B., 2023. Segment anything in medical images. arXiv preprint arXiv:2304.12306.

Ma, J., Zhang, Y., Gu, S., An, X., Wang, Z., Ge, C., Wang, C., Zhang, F., Wang, Y., Xu, Y., Gou, S., Thaler, F., Payer, C., Štern, D., Henderson, E.G., McSweeney, D.M., Green, A., Jackson, P., McIntosh, L., Nguyen, Q.-C., Qayyum, A., Conze, P.-H., Huang, Z., Zhou, Z., Fan, D.P., Xiong, H., Dong, G., Zhu, Q., He, J., Yang, X., 2022. Fast and low-GPU-memory abdomen CT organ segmentation: The FLARE challenge. Med. Image Anal. 82, 102616.

Ma, J., Zhang, Y., Gu, S., Ge, C., Ma, S., Young, A., Zhu, C., Meng, K., Yang, X., Huang, Z., Zhang, F., Liu, W., Pan, Y., Huang, S., Wang, J., Sun, M., Xu, W., Jia, D., Choi, J.W., Alves, N., de Wilde, B., Koehler, G., Wu, Y., Wiesenfarth, M., Zhu, Q., Dong, G., He, J., the FLARE Challenge Consortium, Wang, B., 2023. Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the FLARE22 challenge. arXiv preprint arXiv:2308.05862.

Mazurowski, M.A., Dong, H., Gu, H., Yang, J., Konz, N., Zhang, Y., 2023. Segment anything model for medical image analysis: an experimental study. arXiv preprint arXiv:2304.10517.

Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al., 2014. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging 34 (10), 1993–2024.

Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P., 2023. Foundation models for generalist medical artificial intelligence. Nature 616 (7956), 259–265.

OpenAI, 2023. GPT-4 technical report. arXiv preprint arXiv:2303.08774.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al., 2023. DINOv2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al., 2022. Training language models to follow instructions with human feedback. Adv. Neural Inf. Process. Syst. 35, 27730–27744.

Paranjape, J.N., Nair, N.G., Sikder, S., Vedula, S.S., Patel, V.M., 2023. Adaptivesam: Towards efficient tuning of sam for surgical scene segmentation. arXiv preprint arXiv:2308.03726.

Pinaya, W.H., Tudosiu, P.D., Dafflon, J., Da Costa, P.F., Fernandez, V., Nachev, P., Ourselin, S., Cardoso, M.J., 2022. Brain imaging generation with latent diffusion models. In: MICCAI Workshop on Deep Generative Models. Springer, pp. 117–126.

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: International Conference on Machine Learning. PMLR, pp. 8748–8763.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al., 2018. Improving Language Understanding by Generative Pre-Training. OpenAI.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al., 2019. Language models are unsupervised multitask learners. OpenAI blog 1 (8), 9.

Roy, S., Wald, T., Koehler, G., Rokuss, M.R., Disch, N., Holzschuh, J., Zimmerer, D., Maier-Hein, K.H., 2023. Sam. md: Zero-shot medical image segmentation capabilities of the segment anything model. arXiv preprint arXiv:2304.05396.

Shi, P., Qiu, J., Abaxi, S.M.D., Wei, H., Lo, F.P.W., Yuan, W., 2023. Generalist vision foundation models for medical imaging: A case study of segment anything model on zero-shot medical segmentation. Diagnostics 13 (11), 1947.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al., 2023a. Large language models encode clinical knowledge. Nature 1–9.

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., et al., 2023b. Towards expert-level medical question answering with large language models. arXiv preprint arXiv:2305.09617.

Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A., 2022. Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20730–20740.

Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y., Rajpurkar, P., 2022. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. Nat. Biomed. Eng. 1–8.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al., 2023a. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al., 2023b. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.C., Carroll, A., Lau, C., Tanno, R., Ktena, I., et al., 2023. Towards generalist biomedical ai. arXiv preprint arXiv:2307.14334.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Liu, S., Mathieu, P., Eck, A.v., Lee, D., Viret, J., et al., 2023. Virchow: A million-slide digital pathology foundation model. arXiv preprint arXiv:2309.07778.

Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., Dai, J., 2023a. VisionLLM: Large language model is also an open-ended decoder for vision-centric tasks. https://arxiv.org/abs/2305.11175.

Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., Hu, X., Lu, T., Lu, L., Li, H., et al., 2022. Internimage: Exploring large-scale vision foundation models with deformable convolutions. arXiv preprint arXiv:2211.05778.

Wang, Z., Liu, C., Zhang, S., Dou, Q., 2023b. Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M., 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2097–2106.

Wang, D., Wang, X., Wang, L., Li, M., Da, Q., Liu, X., Gao, X., Shen, J., He, J., Shen, T., Duan, Q., Zhao, J., Li, K., Qiao, Y., Zhang, S., 2023c. A real-world dataset and benchmark for foundation model adaptation in medical image classification. Nat. Scientific Data.

Wang, G., Wu, J., Luo, X., Liu, X., Li, K., Zhang, S., 2023d. MIS-FM: 3D medical image segmentation using foundation models pretrained on a large-scale unannotated dataset. arXiv preprint arXiv:2306.16925.

Wang, G., Zhang, S., Huang, X., Vercauteren, T., Metaxas, D., 2023e. Editorial for special issue on explainable and generalizable deep learning methods for medical image computing. Med. Image Anal. 84, 102727.

Wang, S., Zhao, Z., Ouyang, X., Wang, Q., Shen, D., 2023f. Chatcad: Interactive computer-aided diagnosis on medical image using large language models. arXiv preprint arXiv:2302.07257.

Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., Bach, M., Segeroth, M., 2023. TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images. Radiology: Artif. Intell. 5 (5), e230024. http://dx.doi.org/10.1148/ryai.230024.

Wu, J., Fu, R., Fang, H., Liu, Y., Wang, Z., Xu, Y., Jin, Y., Arbel, T., 2023a. Medical SAM adapter: Adapting segment anything model for medical image segmentation. arXiv preprint arXiv:2304.12620.

Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., Duan, N., 2023b. Visual chatgpt: Talking, drawing and editing with visual foundation models. arXiv preprint arXiv:2303.04671.

Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W., 2023c. Towards generalist foundation model for radiology. arXiv preprint arXiv:2308.02463.

Yi, H., Qin, Z., Lao, Q., Xu, W., Jiang, Z., Wang, D., Zhang, S., Li, K., 2023. Towards general purpose medical AI: Continual learning medical foundation model. arXiv preprint arXiv:2303.06580.

Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L., 2022. Scaling vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12104–12113.

Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P., 2022. Contrastive learning of medical visual representations from paired images and text. In: Machine Learning for Healthcare Conference. PMLR, pp. 2–25.

Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., et al., 2023. Large-scale domain-specific pretraining for biomedical vision-language processing. arXiv preprint arXiv:2303.00915.

Zhou, H.-Y., Chen, X., Zhang, Y., Luo, R., Wang, L., Yu, Y., 2022. Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. Nat. Mach. Intell. 4 (1), 32–40.

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 1–8.