Review

A Survey of Generalization and Adaptation in Medical Imaging Foundation Models

Haoran Ying , Yijun Lia * , Zedong Fu

Posted Date: 11 July 2025

doi: 10.20944/preprints202507.0942.v1

Keywords: foundation models; domain adaptation; domain generalization; medical imaging; transfer learning; self-supervised learning; multi-modal learning; vision-language models; robustness; generalizability; zero-shot learning; few-shot learning; radiology AI; out-of-distribution detection; representation learning

Review

A Survey of Generalization and Adaptation in Medical Imaging Foundation Models

Haoran Ying, Yijun Lia * and Zedong Fu

Nanjing University, China
* Correspondence: yijun.lia@nju.edu.cn

Abstract

Medical imaging is a cornerstone of modern healthcare, playing a critical role in diagnosis, treatment planning, and disease monitoring across a wide range of clinical applications. However, the development and deployment of machine learning models in this domain face persistent challenges stemming from domain shift—the divergence in data distributions across imaging centers, patient populations, devices, and acquisition protocols. These shifts degrade model performance when applied outside the training domain, thereby undermining the generalizability and reliability of traditional supervised learning approaches in real-world clinical environments. Historically, research in domain adaptation and generalization has sought to mitigate this issue through techniques such as adversarial training, domain-invariant feature learning, and data augmentation. While these methods have shown moderate success, they often rely on access to labeled data, tailored adaptation procedures, or knowledge of the target domain, limiting their scalability and practicality. The recent advent of foundation models—large-scale, pre-trained models capable of zero-shot and few-shot inference—has introduced a paradigm shift in addressing these challenges. Leveraging vast and heterogeneous datasets, often in a self-supervised or weakly supervised manner, foundation models in medical imaging exhibit emergent properties that enable superior transferability and robustness across unseen domains and tasks. These models, including vision-language models (e.g., MedCLIP, GLoRIA, CheXzero) and large-scale unimodal encoders (e.g., Swin UNETR, TransUNet), encode rich semantic representations that are less sensitive to superficial domain-specific artifacts. As a result, they show state-of-the-art performance in tasks such as classification, segmentation, and report generation, even under substantial domain shifts and with minimal supervision. In this survey, we provide a comprehensive review of domain adaptation and generalization in the era of foundation models for medical imaging. We begin by tracing the historical evolution of domain robustness techniques, outlining their theoretical foundations and empirical limitations. We then detail the architectural and training paradigms that underpin foundation models, highlighting their ability to learn transferable, multi-modal, and clinically aligned representations. Through a comparative analysis, we evaluate the performance of foundation models against traditional domain adaptation techniques across a range of benchmarks, imaging modalities, and clinical settings. Furthermore, we explore the interpretability, label efficiency, and deployment implications of this emerging class of models. We also identify and discuss key open challenges—including ethical considerations, computational constraints, interpretability, and the lack of standardized benchmarks—and propose future research directions to ensure responsible and equitable advancement. These include developing efficient training strategies, designing clinically meaningful evaluation protocols, supporting multilingual and multimodal understanding, and integrating continual learning frameworks to adapt to evolving clinical practices. Overall, this survey aims to bridge the gap between traditional domain adaptation techniques and the emerging capabilities of foundation models, offering a unified perspective for researchers, practitioners, and policymakers seeking to develop robust, generalizable, and clinically trustworthy AI systems in medical imaging.

Keywords: foundation models; domain adaptation; domain generalization; medical imaging; transfer learning; self-supervised learning; multi-modal learning; vision-language models; robustness;

generalizability; zero-shot learning; few-shot learning; radiology AI; out-of-distribution detection; representation learning

1. Introduction

The convergence of deep learning and medical imaging has yielded groundbreaking advances in automated disease diagnosis, image segmentation, registration, and prognosis modeling. These data-driven approaches, often powered by convolutional neural networks (CNNs) and, more recently, transformers, have shown potential to augment clinical workflows and improve patient outcomes. Despite their promise, a persistent and critical challenge undermines the reliability and generalizability of these systems: the susceptibility of deep learning models to distributional shifts across domains [1]. In real-world deployment, models trained on data from a single institution or imaging protocol frequently exhibit severe performance degradation when applied to new settings characterized by different scanners, patient populations, imaging modalities, or acquisition parameters [2]. This lack of robustness is particularly concerning in safety-critical applications such as cancer detection, lesion segmentation, or surgical planning. Domain adaptation and generalization aim to address this challenge by enabling models to maintain high performance when encountering previously unseen domains [3]. Domain adaptation techniques explicitly leverage unlabeled or partially labeled data from a target domain during training, seeking to align feature representations across source and target distributions [4]. In contrast, domain generalization seeks to develop models that are inherently robust to domain shifts, even without access to target data during training. Over the last decade, a diverse landscape of algorithms has emerged in both areas, ranging from adversarial learning and meta-learning to style transfer and self-supervised representation learning. Yet, these methods are typically tailored to narrow, task-specific settings and often lack scalability across heterogeneous datasets and tasks common in medical imaging [5]. Simultaneously, a paradigm shift is underway with the advent of *foundation models*—massive pre-trained models trained on large-scale datasets using self-supervised or weakly-supervised learning objectives [6]. In the broader machine learning community, foundation models such as GPT-4 in NLP and CLIP, DINO, and SAM in vision have demonstrated unprecedented capabilities in zero-shot and few-shot learning, transferability, compositional reasoning, and multi-modal understanding. These capabilities are particularly enticing for the medical domain, where labeled data is expensive, domain shifts are pervasive, and generalization is paramount. The emergence of biomedical foundation models like BioGPT, MedCLIP, and ViT-based encoders pre-trained on public medical image repositories introduces new opportunities to rethink how we approach domain adaptation and generalization in medical imaging. This survey examines the intersection of foundation models and domain adaptation/generalization in medical imaging. Our central thesis is that foundation models offer a unifying and scalable framework to mitigate domain shifts, foster generalization across diverse clinical environments, and lower the barriers to developing robust AI systems in healthcare. We begin by providing a detailed overview of the domain shift problem in medical imaging, including sources of variability (e.g., scanner types, imaging protocols, patient demographics), common failure cases, and the implications for clinical deployment. We then review classical and contemporary approaches to domain adaptation and generalization, categorizing them into supervised, semi-supervised, unsupervised, and self-supervised paradigms. These techniques are discussed in the context of various imaging modalities, including MRI, CT, ultrasound, histopathology, and ophthalmic imaging [7**?**]. Next, we shift focus to the foundation model paradigm, presenting a taxonomy of existing models and pretraining strategies relevant to medical imaging [8]. This includes supervised pretraining on labeled radiology corpora, self-supervised learning using contrastive and generative objectives, and cross-modal models that align image and text representations. We analyze how such models can be fine-tuned, adapted, or prompted to enable robust generalization across domains. Key challenges such as catastrophic forgetting, negative transfer, and domain collapse are discussed alongside mitigation strategies including adapter tuning, prompt engineering, and continual learning. Furthermore, we explore the limitations of current foundation models in medical imaging, including data curation biases, lack of transparency in model development, limited evaluation across global clinical settings, and concerns around fairness and trustworthiness [9]. We propose benchmark criteria and evaluation protocols to assess generalization and adaptation performance in a principled manner. In addition, we highlight underexplored research directions such as federated foundation model training, multi-modal generalization, causal domain shifts, and the role of synthetic data in foundation model pretraining [10]. In summary, this survey aims to provide a cohesive, structured, and critical synthesis of the emerging landscape at the intersection of foundation models, domain adaptation, and generalization in medical imaging. As the field transitions toward the next generation of medical AI systems, we believe that foundation models—when developed with domain-specific considerations and robust evaluation—hold the key to unlocking reliable, adaptable, and globally relevant imaging solutions [11].

2. Background and Motivation

The application of deep learning in medical imaging has ushered in a new era of diagnostic automation, with systems capable of performing tasks such as tumor detection, organ segmentation, and anatomical landmark identification at or above human-level accuracy—when evaluated under controlled conditions. However, despite impressive performance on benchmark datasets, these models often falter when confronted with real-world variability. This discrepancy highlights a fundamental issue: the brittleness of machine learning models to domain shifts, which arise from discrepancies between the data distributions in training (source domain) and testing (target domain) environments. In the context of medical imaging, domain shifts are particularly pronounced and multifaceted. These shifts may stem from a variety of factors, including but not limited to:

- **Scanner and hardware differences:** Variations in imaging hardware, such as differences in MRI or CT scanner manufacturers, magnetic field strengths, and imaging protocols, can lead to divergent data distributions even for the same anatomical regions [12].
- **Institutional and demographic diversity:** Patient populations differ across institutions in terms of age, ethnicity, disease prevalence, and comorbidities, influencing both image appearance and underlying pathology [13].
- **Annotation practices and label noise:** Disparities in labeling conventions, radiologist interpretations, and inter-observer variability introduce label noise and inconsistencies that exacerbate the challenge of robust learning.
- **Temporal shifts:** Changes in clinical practice, imaging protocols, or population health over time can introduce subtle or abrupt shifts in data distributions, complicating longitudinal generalization.

These domain shifts significantly impair model performance when systems trained on data from a limited number of institutions are deployed in broader clinical settings. For instance, a model trained to detect intracranial hemorrhage on CT scans from one hospital may perform poorly on scans from another institution due to differences in contrast, resolution, or patient demographics [14]. As a result, the promise of deep learning to democratize medical diagnostics and reduce healthcare disparities remains largely unrealized [15]. **Domain adaptation** and **domain generalization** have emerged as two key research directions to address this gap. Domain adaptation methods aim to bridge the gap between source and target distributions by learning domain-invariant features or aligning distributions through adversarial learning, discrepancy minimization, or reconstruction-based approaches. Depending on the availability of target labels, these methods are categorized as supervised, semi-supervised, or unsupervised domain adaptation [16]. In contrast, domain generalization seeks to train models on multiple source domains in such a way that they can generalize to unseen target domains without requiring any target domain data during training [17]. This paradigm is particularly compelling for medical imaging applications, where privacy concerns and data sharing restrictions often prevent access to target domain data. Approaches in domain generalization include data augmentation strategies, episodic training, invariant risk minimization, and meta-learning frameworks that simulate domain shifts during training. However, both domain adaptation and generalization face intrinsic limitations [18]. First, traditional methods are often data- and compute-intensive, requiring careful design and tuning for each new task or modality. Second, these methods typically operate in low-capacity regimes, where models are sensitive to hyperparameters and initialization [19]. Third, the assumptions underlying these methods—such as covariate shift or shared label space—may not hold in practical, heterogeneous medical imaging scenarios [20]. The recent rise of **foundation models** offers a potential paradigm shift. These models are trained at scale on diverse datasets using self-supervised or weakly-supervised objectives that capture broad, transferable knowledge across tasks and domains [21]. In natural image domains, foundation models such as Vision Transformers (ViT), CLIP, and DINO have demonstrated remarkable zero-shot and few-shot generalization [22]. In the biomedical domain, efforts like RadImageNet, Med3D, CheXzero, GatorTron, and BioViL aim to replicate this success by training large-scale models on medical-specific corpora, spanning imaging, reports, and structured clinical data [23]. Foundation models can be viewed as a data-centric and model-centric generalization strategy that inherently mitigates domain shifts [24]. Their ability to learn rich, general-purpose representations reduces the reliance on hand-crafted adaptation strategies and improves performance across diverse test distributions [25]. Moreover, their flexibility—via fine-tuning, adapter modules, or prompt-based conditioning—facilitates efficient adaptation to new tasks with minimal additional data, a crucial requirement in medical settings. This convergence of foundation models with domain adaptation and generalization introduces a new set of research questions: Can pretraining on heterogeneous medical datasets substitute for traditional adaptation techniques [26]? How should foundation models be evaluated for robustness across domains? What are the trade-offs between specialization and generalization in large models trained across multiple imaging modalities? In the subsequent sections, we systematically explore the taxonomy of existing domain adaptation and generalization methods in medical imaging, and critically examine how foundation models are beginning to reshape the assumptions, methodologies, and performance bounds of these paradigms [27]. Our goal is not only to survey current progress but also to chart a path forward for robust, trustworthy, and clinically viable AI systems in medical imaging [28].

3. Taxonomy of Domain Adaptation and Generalization in Medical Imaging

The study of domain adaptation and generalization in medical imaging can be framed within the formalism of distributional learning theory. Let $\mathcal{X}$ denote the input space (e.g., the space of medical images) and $\mathcal{Y}$ the output space (e.g., segmentation masks, classification labels). A domain $\mathcal{D} = \{\mathcal{P}(x,y)\}$ consists of a joint probability distribution over $\mathcal{X} \times \mathcal{Y}$, where $(x,y) \sim \mathcal{P}$ represents image-label pairs [29]. The primary goal in domain adaptation is to learn a function $f_\theta : \mathcal{X} \to \mathcal{Y}$ parameterized by θ such that the expected risk on a *target domain* $\mathcal{D}_T$ is minimized, given labeled data from a *source domain* $\mathcal{D}_S$ [30]. In the standard supervised learning paradigm, we minimize the empirical risk over labeled samples from $\mathcal{D}_S$:

$$\mathcal{L}_{\text{sup}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(f_\theta(x_i^S), y_i^S),$$

where $\ell(\cdot,\cdot)$ is a task-specific loss function, such as cross-entropy for classification or Dice loss for segmentation. However, under domain shift, the marginal distributions $P_S(x)$ and $P_T(x)$ may differ substantially ($P_S(x) \neq P_T(x)$), leading to poor generalization due to the model's overfitting to source-specific characteristics. To address this, domain adaptation methods introduce additional objectives to align the source and target distributions in a shared representation space. A general strategy is to learn a domain-invariant feature extractor $g_\phi : \mathcal{X} \to \mathcal{Z}$ and a task-specific head $h_\psi : \mathcal{Z} \to \mathcal{Y}$, such that

$f_\theta(x) = h_\psi(g_\phi(x))$, where $\theta = \{\phi, \psi\}$ [31]. A popular alignment criterion is the minimization of some divergence measure $\mathcal{D}_{\text{div}}$ between $P_S(z)$ and $P_T(z)$:

$$\min_\phi \mathcal{D}_{\text{div}}(P_S(g_\phi(x^S)), P_T(g_\phi(x^T))),$$

which may take the form of adversarial loss (e.g., domain discriminators), statistical divergences (e.g., Maximum Mean Discrepancy), or reconstruction losses (e.g., variational autoencoders) [32].

3.1. Supervised and Semi-Supervised Domain Adaptation

In **supervised domain adaptation**, the learner has access to a small set of labeled examples from the target domain: $\mathcal{D}_T = \{(x_i^T, y_i^T)\}_{i=1}^m$ [33]. This setting enables direct minimization of the target risk:

$$\mathcal{L}_{\text{target}}(\theta) = \frac{1}{m} \sum_{i=1}^m \ell(f_\theta(x_i^T), y_i^T),$$

often in combination with domain alignment losses to prevent overfitting to the limited target labels [34]. In contrast, **semi-supervised domain adaptation** assumes that labeled data is available only from the source domain, while the target domain contains only unlabeled examples [35]. The key idea is to leverage unlabeled target data via pseudo-labeling, consistency regularization, or entropy minimization. Let $\mathcal{D}_T = \{x_j^T\}_{j=1}^m$, then the loss function may be augmented with unsupervised terms:

$$\mathcal{L}_{\text{total}}(\theta) = \mathcal{L}_{\text{sup}}(\theta) + \lambda \cdot \mathcal{L}_{\text{unsup}}(\theta),$$

where $\mathcal{L}_{\text{unsup}}$ encourages confident and consistent predictions on target inputs, and λ is a balancing hyperparameter [36].

3.2. Unsupervised Domain Adaptation

In **unsupervised domain adaptation** (UDA), the target domain remains completely unlabeled, making the learning problem even more ill-posed [37]. To cope with this, many methods exploit adversarial training frameworks where a domain discriminator D is trained to distinguish between source and target features, while the feature encoder g_ϕ tries to fool the discriminator:

$$\min_\phi \max_D \mathbb{E}_{x\sim P_S}[\log D(g_\phi(x))] + \mathbb{E}_{x\sim P_T}[\log(1 - D(g_\phi(x)))],$$

leading to domain confusion and theoretically promoting domain-invariant representations [38]. Despite their effectiveness, adversarial methods may suffer from mode collapse or training instability, particularly in high-resolution 3D medical imaging tasks such as brain MRI segmentation or chest CT detection.

3.3. Domain Generalization

Domain generalization (DG) differs from adaptation in that the model is trained on multiple source domains $\{\mathcal{D}_S^{(i)}\}_{i=1}^N$ and expected to generalize to unseen target domains without access to target data during training [39]. The generalization objective seeks to minimize the expected risk across a family of source distributions:

$$\mathbb{E}_{\mathcal{D}_S\sim\mathcal{P}}[\mathcal{L}_{\text{sup}}(\theta; \mathcal{D}_S)],$$

where $\mathcal{P}$ denotes the meta-distribution over domains. DG methods include invariant risk minimization (IRM), which enforces the optimal classifier h_ψ to be invariant across domains by optimizing:

$$\min_{\phi,\psi} \sum_{i=1}^N \mathcal{L}_{\text{sup}}^{(i)}(\phi, \psi) + \gamma \cdot \left\| \nabla_\psi \mathcal{L}_{\text{sup}}^{(i)}(\phi, \psi) \right\|^2,$$

encouraging the model to find domain-invariant mechanisms. Other strategies involve data augmentation, synthetic data generation, meta-learning (e.g., model-agnostic meta-learning, MAML), or ensemble learning techniques that combine models trained on different source domains.

3.4. Self-Supervised and Multi-Modal Learning Approaches

Recent years have witnessed the emergence of **self-supervised learning** (SSL) methods, which train models using proxy tasks that do not require manual labels [40]. Contrastive learning frameworks such as SimCLR, MoCo, and BYOL learn representations by pulling positive pairs (e.g., different augmentations of the same image) together in embedding space while pushing apart negative pairs:

$$\mathcal{L}_{\text{contrast}} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{K} \exp(\text{sim}(z_i, z_k)/\tau)},$$

where $\text{sim}(\cdot, \cdot)$ is the cosine similarity, z_i and z_j are embeddings of positive pairs, and τ is a temperature parameter [41]. SSL-pretrained models can be fine-tuned on downstream tasks and often exhibit improved generalization to out-of-domain distributions due to their exposure to diverse, unlabeled data [42]. Moreover, **multi-modal learning**, where images are paired with corresponding clinical reports, radiology text, or pathology findings, introduces cross-domain alignment capabilities [43]. Models such as MedCLIP, BioViL, and GLoRIA align visual and textual modalities in a shared latent space, enabling zero-shot reasoning and improved robustness [44]. The training objective is often a bi-modal contrastive loss:

$$\mathcal{L}_{\text{multi-modal}} = -\log \frac{\exp(\text{sim}(g_I(x), g_T(t))/\tau)}{\sum_{t'} \exp(\text{sim}(g_I(x), g_T(t'))/\tau)},$$

where g_I and g_T are encoders for image and text, respectively, and t is the paired text with image x [45]. Such models are inherently more generalizable, as they exploit complementary information across modalities to ground features in clinically meaningful semantics [46].

3.5. Summary of Taxonomy

To summarize, the landscape of domain adaptation and generalization techniques in medical imaging can be taxonomized along three principal axes: (1) the availability of labeled/unlabeled data in source and target domains; (2) the methodological backbone (e.g., adversarial training, meta-learning, contrastive pretraining); and (3) the architectural paradigm (e.g., single-modal CNNs vs. multi-modal transformers) [47]. This taxonomy lays the foundation for understanding how foundation models can further unify and enhance these strategies by offering a general-purpose representation that is adaptable, scalable, and resilient to distributional shifts [48]. In the next section, we delve into the architectural and training design of foundation models that are particularly suited for medical imaging applications [49].

4. Foundation Models for Medical Imaging

The rise of foundation models marks a significant shift in how representation learning is conceptualized and operationalized in artificial intelligence [50]. In contrast to conventional task-specific models, foundation models are trained on massive, heterogeneous datasets using general-purpose objectives—often in a self-supervised or weakly supervised manner—thereby acquiring broad and transferable capabilities that can be adapted to a wide array of downstream tasks with minimal supervision [51]. Within the medical imaging domain, the adaptation of this paradigm has been driven by a confluence of factors, including the increasing availability of large-scale imaging datasets, advancements in transformer architectures, and the pressing need for models that generalize robustly across clinical institutions, imaging devices, and patient demographics [52].

4.1. Architectural Paradigms

Most foundation models in medical imaging draw inspiration from natural image models such as Vision Transformers (ViT), convolutional neural networks (CNNs), and, more recently, hybrid architectures that combine both modalities [53]. The transformer-based models, in particular, have gained prominence due to their scalability and ability to model long-range dependencies. Unlike CNNs, which operate with strong inductive biases such as locality and translation invariance, transformers learn these properties from data, making them well-suited for highly heterogeneous domains like medical imaging [54]. In the case of multi-modal foundation models, architectures are designed to integrate information from diverse sources such as radiology images, clinical notes, and structured metadata. These models typically employ dual-stream encoders—one for each modality—and align their latent spaces via contrastive learning or cross-attention mechanisms [55]. This alignment facilitates cross-modal retrieval, report generation, and enables the use of textual supervision to guide visual representation learning [56].

4.2. Training Objectives and Strategies

A hallmark of foundation models is their use of general-purpose, label-agnostic training objectives. In the context of medical imaging, the following self-supervised or weakly supervised objectives are widely employed:

- **Contrastive Learning:** Inspired by methods like SimCLR and CLIP, medical contrastive models learn to bring together representations of images and their corresponding textual descriptions (e.g., radiology reports) while pushing apart unpaired examples. This framework has proven effective in capturing semantically rich and transferable features.
- **Masked Image Modeling (MIM):** Analogous to masked language modeling in NLP (e.g., BERT), MIM involves corrupting input images by masking out patches and training the model to reconstruct the missing information [57]. This task encourages the model to learn contextual features and global anatomical priors [58].
- **Image-Text Alignment:** In multi-modal setups, models are trained to align paired image-text embeddings using cosine similarity or contrastive loss [59]. This approach, exemplified by models like MedCLIP and GLoRIA, enables zero-shot transfer to classification and retrieval tasks.
- **Multi-Task Supervision:** Some foundation models are trained using weak labels derived from DICOM metadata, PACS reports, or large-scale automated labeling tools [60]. This strategy allows supervision at scale without the cost of manual annotation [61].

These objectives are typically optimized on large and diverse corpora that span multiple medical institutions, imaging modalities (e.g., X-ray, CT, MRI, ultrasound), anatomical sites, and disease categories [62]. The resulting models are characterized by their broad applicability, semantic richness, and robustness to distributional shifts—properties that are highly desirable for domain adaptation and generalization.

4.3. Representative Foundation Models in Medical Imaging

Several foundation models have been proposed to harness these ideas in the medical imaging domain. A non-exhaustive list of prominent efforts includes:

- **MedCLIP:** A contrastive learning-based model that aligns medical images and radiology reports using transformer-based encoders. It enables zero-shot classification and report generation without requiring fine-tuning on downstream tasks [63].
- **BioViL:** The Biomedical Vision-Language Pretraining model leverages paired image-text datasets and contrastive losses to build general-purpose medical embeddings. It has demonstrated strong performance on retrieval, classification, and segmentation tasks.
- **GLoRIA:** The Global and Local Representation Alignment model learns fine-grained correspondence between regions in medical images and textual descriptions [64]. It enhances interpretability and supports few-shot learning across tasks.

- **RadImageNet:** A CNN-based foundation model pretrained on a large collection of labeled radiological images using supervised learning [65]. Despite lacking multi-modality, it provides a strong initialization for transfer learning in X-ray and CT imaging [66].
- **CheXzero:** Built upon the CLIP framework, CheXzero trains a vision-language model using a large dataset of chest X-rays and free-text radiology reports [67]. It enables zero-shot classification and has shown performance competitive with radiologists on specific tasks.
- **UNITER and Med-UniT:** Inspired by image-text fusion models in general vision, these approaches integrate vision-language modeling into a single transformer architecture, supporting unified learning across multiple imaging modalities and clinical narratives.

4.4. Foundation Models as Implicit Domain Generalizers

One of the most compelling aspects of foundation models is their ability to implicitly perform domain adaptation and generalization without explicit alignment losses or domain-specific regularization. By pretraining on data from a wide array of sources, foundation models internalize a diverse range of imaging styles, anatomical variations, and pathological manifestations [68]. Consequently, they are less susceptible to performance degradation when transferred to novel domains [69]. Empirical studies have shown that foundation models pretrained on heterogeneous corpora exhibit strong zero-shot and few-shot transferability to previously unseen target distributions, outperforming traditional models trained with explicit domain adaptation pipelines. Furthermore, the modularity of foundation models—e.g., through prompt tuning, adapters, or low-rank fine-tuning—enables rapid adaptation to new clinical settings with minimal retraining. This flexibility is particularly crucial in healthcare, where institutional data silos, privacy constraints, and continuous shifts in clinical practice necessitate adaptive and resilient AI systems.

4.5. Challenges and Limitations

Despite their promise, foundation models in medical imaging face several open challenges:

- **Data bias and representativeness:** Even large-scale pretraining datasets may suffer from institutional or demographic biases that hinder equitable generalization [70].
- **Compute and memory costs:** Training and deploying large models requires substantial hardware resources, which may not be accessible to all research or clinical institutions [71].
- **Lack of transparency:** The interpretability and trustworthiness of foundation models remain underexplored, particularly in high-stakes clinical decision-making.
- **Evaluation complexity:** Robust, standardized benchmarks that assess generalization across diverse target domains are still emerging, making performance comparisons difficult [72].

4.6. Conclusion

In summary, foundation models offer a powerful new paradigm for building generalizable and adaptive AI systems in medical imaging [73]. Their ability to learn rich, transferable representations from diverse data and modalities fundamentally alters the landscape of domain adaptation and generalization. However, their success hinges on careful architectural design, thoughtful dataset curation, and rigorous evaluation [74]. In the following section, we provide a comparative analysis of these models against traditional domain adaptation techniques, highlighting performance metrics, data requirements, and generalization behavior across various benchmarks.

5. Comparative Analysis of Foundation Models and Traditional Adaptation Methods

The advent of foundation models in medical imaging has prompted a reevaluation of the efficacy and efficiency of traditional domain adaptation and generalization techniques [75]. To systematically assess the relative merits of these approaches, this section presents a multi-dimensional comparative analysis that spans performance metrics, data requirements, generalization behavior, interpretability, computational cost, and scalability [76]. Our analysis is guided by empirical findings from recent studies as well as theoretical insights into the representational capacity and transferability of learned features.

5.1. Quantitative Evaluation Across Domain Shifts

A central axis of comparison is the performance of models under explicit domain shifts. Traditional methods typically rely on hand-engineered strategies to reduce domain gaps (e.g., adversarial alignment, normalization-based harmonization, or domain-specific batch statistics), and are evaluated using supervised or semi-supervised benchmarks where labeled data from source and target domains are available. In contrast, foundation models often operate in zero-shot or few-shot regimes, and their robustness is attributed to large-scale, diverse pretraining.

From Table 1, we observe that foundation models such as MedCLIP and CheXzero outperform traditional domain adaptation approaches on zero-shot classification tasks, often without requiring any access to the target domain during training. This suggests a fundamental shift in the efficacy of pretraining-based transfer relative to classical alignment techniques [77]. Furthermore, foundation models tend to generalize better when evaluated on rare pathologies or under shifts in acquisition protocol, hospital system, or scanner manufacturer.

Table 1. Performance comparison between traditional domain adaptation methods and foundation models on representative benchmarks. Metrics include classification accuracy (ACC), Dice similarity coefficient (DSC), and area under the curve (AUC).

Model	Domain	Task	Performance (AUC/ACC/DSC)
DANN (Ganin et al.)	ChestX-ray14 → CheXpert	Classification	0.803 AUC
CycleGAN (Zhu et al.)	MRI (GE) → MRI (Siemens)	Segmentation	0.726 DSC
IRM (Arjovsky et al.)	Multi-site CT	Classification	0.781 ACC
MedCLIP (Zhang et al.)	ChestX-ray14, MIMIC-CXR	Zero-shot Classification	0.868 AUC
BioViL (Boecking et al.)	RSNA, MIMIC, OpenI	Retrieval + Classification	0.842 AUC
CheXzero (Tiu et al.)	ChestX-ray14 → CheXpert	Zero-shot Classification	0.881 AUC

5.2. Data Requirements and Label Efficiency

Traditional domain adaptation pipelines are heavily reliant on supervised source data and often require labeled target samples to fine-tune effectively. This dependency presents a significant bottleneck in real-world settings where annotated data is sparse, costly, or institutionally siloed [78]. Foundation models, by contrast, are typically trained using weak or self-supervised signals derived from large unlabeled datasets or radiology reports, which are often readily available in hospital information systems. This drastically reduces the need for expert annotation and makes foundation models particularly attractive for low-resource settings and emerging diseases. Moreover, their few-shot learning capabilities allow rapid adaptation with only minimal supervision, effectively bypassing the full fine-tuning stage required by traditional models.

5.3. Generalization Behavior and Robustness

An important axis of comparison is the generalization behavior of these models when applied to out-of-distribution (OOD) data [79]. Traditional methods often fail to generalize beyond the specific source-target domain pairs they were trained on, leading to brittle performance in real-world clinical deployments. In contrast, foundation models trained on a wide diversity of data sources develop internal representations that are more semantically grounded and less domain-specific. Recent work demonstrates that foundation models are more robust to domain-specific artifacts such as contrast variation, scanner noise, image resolution discrepancies, and anatomical variance. Furthermore, multimodal foundation models that incorporate text (e.g., MedCLIP, GLoRIA) tend to capture higher-level clinical semantics that are invariant to purely visual distortions, further enhancing their robustness to domain shifts.

5.4. Interpretability and Clinical Alignment

Interpretability is crucial in medical imaging applications where clinical accountability and decision traceability are non-negotiable. Traditional adaptation models often lack intrinsic mechanisms for model explanation, relying instead on post hoc visualization techniques such as Grad-CAM or feature attribution maps. In contrast, multi-modal foundation models are naturally more interpretable due to their alignment with clinical language. For example, models such as GLoRIA and CheXzero can highlight regions of interest based on textual queries (e.g., "right upper lobe nodule") and provide sentence-level justifications for predictions [80]. This alignment with radiology reports and clinical terminology significantly enhances trust and usability among practitioners.

5.5. Computational Cost and Scalability

A major criticism of foundation models pertains to their computational demands [81]. Training large-scale models requires substantial GPU/TPU resources, high-bandwidth data pipelines, and weeks of computation time. These requirements may be prohibitive for small research labs or low-resource clinical environments [82]. However, once pretrained, foundation models offer considerable flexibility in deployment. Modular fine-tuning strategies such as adapter layers, prompt tuning, and low-rank updates (LoRA) allow lightweight specialization without retraining the entire model. Furthermore, ongoing research into distillation, pruning, and quantization is making it increasingly feasible to deploy foundation models on edge devices and in real-time clinical settings.

5.6. Summary of Comparative Insights

The comparative analysis in this section suggests that while traditional domain adaptation methods remain effective in well-defined, closed-domain settings, foundation models offer a more scalable, robust, and label-efficient alternative that is well-suited to the demands of modern clinical AI. Table 2 summarizes the key trade-offs between the two paradigms [83].

Table 2. Comparison summary between traditional domain adaptation and foundation model paradigms.

Attribute	Traditional Methods	Foundation Models
Label Requirements	High (manual labeling)	Low (self/weak supervision)
Data Modality	Single (e.g., image only)	Multi-modal (image + text)
Domain Generalization	Limited (target-specific)	Strong (OOD robustness)
Interpretability	Low to moderate	High (text-guided attention)
Deployment Overhead	Low (smaller models)	High (initial training), low (adaptation)
Adaptability	Low (full retraining needed)	High (prompt/adapters)

In the following section, we discuss open research questions and future directions in domain adaptation and generalization using foundation models, highlighting challenges related to evaluation, ethics, and model scaling in clinical environments.

6. Open Challenges and Future Directions

Despite the transformative potential of foundation models in medical imaging, their widespread adoption and safe deployment remain fraught with open challenges [84]. These span technical, ethical, and infrastructural dimensions, each of which must be carefully navigated to realize the vision of robust, generalizable, and trustworthy AI systems in healthcare. In this section, we identify and elaborate on several key directions for future research and development.

6.1. Standardized Evaluation Protocols for Generalization

One of the foremost barriers to progress in domain generalization and adaptation is the lack of standardized, consensus-driven evaluation protocols [85]. While several benchmark datasets (e.g., CheXpert, MIMIC-CXR, RSNA Pneumonia) are widely used, inconsistencies in train-test splits, label harmonization, and pre-processing pipelines limit cross-study comparability. Moreover, many evaluations are conducted on in-distribution data or marginally shifted datasets, failing to capture the spectrum of real-world domain variation. Future work should prioritize the creation of large-scale, multi-institutional, and multi-modal benchmark suites specifically designed to test domain robustness. These should include intentionally designed OOD (out-of-distribution) settings, cross-modality evaluations (e.g., X-ray → CT), and fairness assessments across demographic subgroups [86]. To facilitate reproducibility, evaluation pipelines should be open-sourced and paired with comprehensive metadata [87].

6.2. Efficient and Ethical Model Scaling

The performance of foundation models is strongly correlated with model size, data scale, and compute budget. However, indiscriminate scaling is both environmentally costly and ethically fraught. Many medical institutions, particularly those in low-resource settings, lack the infrastructure to train or even deploy such models. Furthermore, large-scale data aggregation often raises privacy concerns, especially under regulations such as HIPAA and GDPR. A promising direction is the development of compute-efficient training paradigms such as parameter-efficient fine-tuning (PEFT), modular pretraining, and federated learning frameworks that allow for decentralized training across data silos without raw data exchange. In parallel, efforts should be made to rigorously document the environmental and ethical costs of model training, with emphasis on transparency, accountability, and equitable resource allocation [88].

6.3. Bias, Fairness, and Equity in Foundation Models

As foundation models absorb knowledge from massive and diverse datasets, they also inherit and potentially amplify the biases present within them [89]. This includes demographic biases (e.g., race, gender, age), institutional biases (e.g., protocol differences), and socio-cultural biases (e.g., language or practice patterns in reports) [90]. These can lead to systemic disparities in model performance, diagnosis accuracy, and clinical recommendations [91]. Detecting and mitigating such biases is particularly challenging in medical imaging, where ground truth labels may themselves be noisy, subjective, or biased [92]. Future research should explore bias quantification methods tailored to medical contexts, such as stratified evaluation and counterfactual analysis. Moreover, models should be developed and audited with fairness constraints or adversarial debiasing strategies, and evaluated across clinically relevant subgroups [93].

6.4. Interpretable and Clinically Aligned Foundation Models

Although some foundation models (e.g., MedCLIP, GLoRIA) offer interpretable outputs via image-text alignment, their inner decision-making processes often remain opaque [94]. Interpretability in medical imaging must go beyond visual saliency to provide semantically meaningful, clinically verifiable rationales for model predictions. Clinicians require not only an understanding of what the model sees but also why a particular prediction is made, especially in life-critical scenarios. Future work should integrate symbolic reasoning, medical ontologies (e.g., RadLex, SNOMED CT), and structured knowledge graphs to ground predictions in clinical concepts [95]. Additionally, human-in-the-loop evaluation protocols involving domain experts can ensure that explanations are not only technically valid but clinically actionable.

6.5. Cross-Modal and Multilingual Generalization

Medical data is inherently multi-modal and multi-lingual [96]. Imaging data is often accompanied by text reports, EHR entries, lab values, and demographic metadata, each offering complementary information [97]. Moreover, in global health contexts, radiology reports and clinical documentation may be written in diverse languages with varying terminologies [98]. While some progress has been made in vision-language models for English-language radiology, generalization across languages and modalities remains an underexplored frontier [99]. Future foundation models should support cross-modal reasoning (e.g., correlating lab tests with imaging findings) and be robust to language variation.

Approaches such as multilingual tokenization, cross-lingual contrastive learning, and meta-learning could prove instrumental in achieving this goal [100].

6.6. Lifelong Learning and Continual Domain Adaptation

Clinical practice and medical imaging protocols are continuously evolving [101]. New disease variants (e.g., COVID-19), changing scanning technologies, and updated diagnostic guidelines create a moving target for AI systems [102]. Static models—even large foundation ones—may rapidly become outdated or brittle in dynamic environments. To address this, future systems should incorporate mechanisms for continual learning, where models are updated incrementally as new data arrives without catastrophic forgetting of prior knowledge [103]. Additionally, they should be able to perform dynamic domain adaptation in real-time or near-real-time, adapting to new imaging distributions with minimal supervision and without the need for complete retraining. Techniques such as rehearsal-based memory, knowledge distillation, and elastic weight consolidation may offer promising solutions [104].

6.7. Regulatory and Clinical Integration Challenges

Finally, integrating foundation models into clinical workflows raises questions about regulation, accountability, and usability. Unlike traditional AI systems designed for narrow tasks, foundation models are general-purpose and often serve multiple roles (e.g., classification, report generation, triage). This blurs the lines between software-as-a-medical-device (SaMD) and general computing tools, creating a gray area in current regulatory frameworks [105]. Regulatory bodies such as the FDA and EMA must update guidelines to account for the complexities of foundation model behavior, versioning, validation, and risk mitigation. At the same time, clinical integration must be guided by human-centered design principles, ensuring that models augment rather than replace clinical decision-making. Comprehensive user studies, clinician-in-the-loop trials, and post-deployment monitoring will be essential in achieving trustworthy adoption [106].

6.8. Conclusion

The journey toward generalizable, adaptable, and clinically integrated AI in medical imaging is both promising and complex [107]. Foundation models offer a paradigm shift, enabling robust performance across diverse domains with minimal supervision [108]. Yet, significant challenges remain across the axes of fairness, efficiency, interpretability, and governance. Addressing these challenges will require interdisciplinary collaboration spanning machine learning, radiology, ethics, and policy. The future of domain adaptation in medical imaging will not be defined solely by algorithmic innovation, but by our collective ability to align these models with the nuanced, dynamic, and deeply human realities of clinical care.

7. Conclusion

The landscape of domain adaptation and generalization in medical imaging is undergoing a profound transformation with the advent of foundation models. Historically, the field has relied heavily on narrowly focused models trained for specific tasks and domains, often requiring meticulous tuning, handcrafted features, and large quantities of labeled data. These traditional approaches, while effective in constrained settings, have struggled to scale and generalize in the face of real-world variability, where data distributions differ across institutions, scanners, patient populations, and imaging protocols.

Foundation models offer a compelling alternative. Trained on massive, diverse datasets—often spanning multiple modalities and leveraging weak or self-supervised objectives—they exhibit remarkable robustness and adaptability. Their ability to perform zero-shot or few-shot inference, their alignment with clinical language, and their modularity for downstream task adaptation have shifted the paradigm away from bespoke models toward general-purpose representations. As demonstrated in this survey, such models consistently outperform traditional domain adaptation methods in both qualitative and quantitative metrics, particularly under out-of-distribution and low-data scenarios.

However, this new paradigm is not without its limitations. Foundation models are computationally intensive to train and often operate as opaque black boxes, raising concerns about interpretability, fairness, and regulatory oversight. They may also embed and amplify latent biases present in their training data, leading to ethical risks in clinical applications. Moreover, their evaluation remains inconsistent, lacking standardized benchmarks that fully capture the complexity and variability of real-world deployment conditions.

To responsibly harness the promise of foundation models in medical imaging, the research community must embrace a holistic approach—one that integrates rigorous empirical validation, transparent documentation, fairness auditing, and human-in-the-loop design. Future work must also prioritize the development of scalable and ethical training paradigms, efficient adaptation mechanisms, and interpretable decision-making pipelines. Only through interdisciplinary collaboration and conscientious innovation can we ensure that these models serve the ultimate goal of improving patient care equitably and effectively.

In conclusion, foundation models represent not just a technological advancement, but a conceptual shift in how we approach generalization and domain adaptation in medical AI. By unifying diverse sources of knowledge, supporting versatile applications, and enabling flexible deployment, they offer a new foundation upon which robust and trustworthy clinical decision support systems can be built. As we stand at the crossroads of innovation and implementation, the challenge ahead lies not in building larger models, but in building smarter, fairer, and more human-centered systems that truly enhance the practice of medicine.

References

1. Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Wu, Z.; Chang, B.; Sun, X.; Xu, J.; Sui, Z. A survey on in-context learning. *arXiv preprint arXiv:2301.00234* **2022**.
2. Yi, H.; Qin, Z.; Lao, Q.; Xu, W.; Jiang, Z.; Wang, D.; Zhang, S.; Li, K. Towards general purpose medical ai: Continual learning medical foundation model. *arXiv preprint arXiv:2303.06580* **2023**.
3. Naseem, U.; Khushi, M.; Kim, J. Vision-Language Transformer for Interpretable Pathology Visual Question Answering. *IEEE Journal of Biomedical and Health Informatics* **2023**, *27*, 1681–1690. https://doi.org/10.1109/JBHI.2022.3163751.
4. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked Autoencoders Are Scalable Vision Learners, 2021, http://arxiv.org/abs/2111.06377.
5. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning, 2020. https://doi.org/10.48550/arXiv.1911.05722.
6. Dippel, J.; Feulner, B.; Winterhoff, T.; Schallenberg, S.; Dernbach, G.; Kunft, A.; Tietz, S.; Jurmeister, P.; Horst, D.; Ruff, L.; et al. RudolfV: A Foundation Model by Pathologists for Pathologists. *arXiv preprint arXiv:2401.04079* **2024**.
7. Pham, V.T.; Nguyen, T.P. Identification and localization covid-19 abnormalities on chest radiographs. In Proceedings of the The International Conference on Artificial Intelligence and Computer Vision. Springer, 2023, pp. 251–261.
8. Lu, M.Y.; Chen, R.J.; Wang, J.; Dillon, D.; Mahmood, F. Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding. *arXiv preprint arXiv:1910.10825* **2019**.
9. Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; Joulin, A. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments, 2021. https://doi.org/10.48550/arXiv.2006.09882.
10. Lu, M.Y.; Chen, B.; Zhang, A.; Williamson, D.F.K.; Chen, R.J.; Ding, T.; Le, L.P.; Chuang, Y.S.; Mahmood, F. Visual Language Pretrained Multiple Instance Zero-Shot Transfer for Histopathology Images. 2023, pp. 19764–19775.
11. Weinstein, J.N.; Collisson, E.A.; Mills, G.B.; Shaw, K.R.; Ozenberger, B.A.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J.M. The cancer genome atlas pan-cancer analysis project. *Nature genetics* **2013**, *45*, 1113–1120.
12. Boecking, B.; Usuyama, N.; Bannur, S.; Castro, D.C.; Schwaighofer, A.; Hyland, S.; Wetscherek, M.; Naumann, T.; Nori, A.; Alvarez-Valle, J.; et al. Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing. In Proceedings of the Computer Vision – ECCV 2022; Avidan, S.; Brostow, G.; Cissé, M.; Farinella, G.M.; Hassner, T., Eds., Cham, 2022; pp. 1–21. https://doi.org/10.1007/978-3-031-20059-5_1.

13. Ochi, M.; Komura, D.; Onoyama, T.; Shinbo, K.; Endo, H.; Odaka, H.; Kakiuchi, M.; Katoh, H.; Ushiku, T.; Ishikawa, S. Registered multi-device/staining histology image dataset for domain-agnostic machine learning models. *Scientific Data* **2024**, *11*, 330.
14. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models, 2021, [arXiv:cs.CL/2106.09685].
15. Dippel, J.; Feulner, B.; Winterhoff, T.; Schallenberg, S.; Dernbach, G.; Kunft, A.; Tietz, S.; Milbich, T.; Heinke, S.; Eich, M.L.; et al. RudolfV: A Foundation Model by Pathologists for Pathologists, 2024. https://doi.org/10.48550/arXiv.2401.04079.
16. Zhou, H.; Gu, B.; Zou, X.; Li, Y.; Chen, S.S.; Zhou, P.; Liu, J.; Hua, Y.; Mao, C.; Wu, X.; et al. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112* **2023**.
17. Wang, Z.; Wu, Z.; Agarwal, D.; Sun, J. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163* **2022**.
18. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations, 2020, .
19. Huang, Z.; Bianchi, F.; Yuksekgonul, M.; Montine, T.J.; Zou, J. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine* **2023**, *29*, 2307–2316.
20. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions, 2021, .
21. Lin, W.; Zhao, Z.; Zhang, X.; Wu, C.; Zhang, Y.; Wang, Y.; Xie, W. PMC-CLIP: Contrastive Language-Image Pre-training Using Biomedical Documents. In Proceedings of the Medical Image Computing and Computer Assisted Intervention – MICCAI 2023; Greenspan, H.; Madabhushi, A.; Mousavi, P.; Salcudean, S.; Duncan, J.; Syeda-Mahmood, T.; Taylor, R., Eds., Cham, 2023; pp. 525–536. https://doi.org/10.1007/978-3-031-43993-3_51.
22. Schaumberg, A.J.; Juarez-Nicanor, W.C.; Choudhury, S.J.; Pastrián, L.G.; Pritt, B.S.; Prieto Pozuelo, M.; Sotillo Sánchez, R.; Ho, K.; Zahra, N.; Sener, B.D.; et al. Interpretable multimodal deep learning for real-time pan-tissue pan-disease pathology search on social media. *Modern Pathology* **2020**, *33*, 2169–2185. https://doi.org/10.1038/s41379-020-0540-1.
23. Song, X.; Xu, X.; Yan, P. General Purpose Image Encoder DINOv2 for Medical Image Registration. *arXiv preprint arXiv:2402.15687* **2024**.
24. Men, Y.; Fhima, J.; Celi, L.A.; Ribeiro, L.Z.; Nakayama, L.F.; Behar, J.A. DRStageNet: Deep Learning for Diabetic Retinopathy Staging from Fundus Images, 2023, [arXiv:eess.IV/2312.14891].
25. Gu, Y.; Yang, J.; Usuyama, N.; Li, C.; Zhang, S.; Lungren, M.P.; Gao, J.; Poon, H. BiomedJourney: Counterfactual Biomedical Image Generation by Instruction-Learning from Multimodal Patient Journeys, 2023. https://doi.org/10.48550/arXiv.2310.10765.
26. Schirris, Y.; Gavves, E.; Nederlof, I.; Horlings, H.M.; Teuwen, J. DeepSMILE: Contrastive self-supervised pre-training benefits MSI and HRD classification directly from H&E whole-slide images in colorectal and breast cancer. *Medical Image Analysis* **2022**, *79*, 102464.
27. Gao, Y.; Xia, W.; Hu, D.; Gao, X. DeSAM: Decoupling Segment Anything Model for Generalizable Medical Image Segmentation, 2023. https://doi.org/10.48550/arXiv.2306.00499.
28. Xie, Y.; Zhang, J.; Xia, Y.; Shen, C. Learning From Partially Labeled Data for Multi-Organ and Tumor Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2023**, *45*, 14905–14919. https://doi.org/10.1109/TPAMI.2023.3312587.
29. ai, k.; Aben, N.; de Jong, E.D.; Gatopoulos, I.; Känzig, N.; Karasikov, M.; Lagré, A.; Moser, R.; van Doorn, J.; Tang, F. Towards Large-Scale Training of Pathology Foundation Models, 2024. https://doi.org/10.48550/arXiv.2404.15217.
30. Sambara, S.; Zhang, S.; Banerjee, O.; Acosta, J.; Fahrner, J.; Rajpurkar, P. RadFlag: A Black-Box Hallucination Detection Method for Medical Vision Language Models. *arXiv preprint arXiv:2411.00299* **2024**.
31. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual instruction tuning. *Advances in neural information processing systems* **2023**, *36*, 34892–34916.
32. Roth, B.; Koch, V.; Wagner, S.J.; Schnabel, J.A.; Marr, C.; Peng, T. Low-resource finetuning of foundation models beats state-of-the-art in histopathology. *arXiv preprint arXiv:2401.04720* **2024**.
33. Khan, W.; Leem, S.; See, K.B.; Wong, J.K.; Zhang, S.; Fang, R. A Comprehensive Survey of Foundation Models in Medicine. *arXiv preprint arXiv:2406.10729* **2024**.
34. Glocker, B.; Jones, C.; Roschewitz, M.; Winzeck, S. Risk of bias in chest radiography deep learning foundation models. *Radiology: Artificial Intelligence* **2023**, *5*, e230060.

35. Chen, X.; Fan, H.; Girshick, R.; He, K. Improved Baselines with Momentum Contrastive Learning, 2020. https://doi.org/10.48550/arXiv.2003.04297.
36. Zhu, W.; Chen, Y.; Nie, S.; Yang, H. SAMMS: Multi-modality Deep Learning with the Foundation Model for the Prediction of Cancer Patient Survival. In Proceedings of the 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2023, pp. 3662–3668.
37. Gutiérrez, J.D.; Rodriguez-Echeverria, R.; Delgado, E.; Rodrigo, M.S.; Sánchez-Figueroa, F. No More Training: SAM's Zero-Shot Transfer Capabilities for Cost-Efficient Medical Image Segmentation. *IEEE Access* **2024**, *12*, 24205–24216. https://doi.org/10.1109/ACCESS.2024.3353142.
38. Ma, D.; Taher, M.R.H.; Pang, J.; Islam, N.U.; Haghighi, F.; Gotway, M.B.; Liang, J. Benchmarking and Boosting Transformers for Medical Image Classification. *Domain adaptation and representation transfer: 4th MICCAI Workshop, DART 2022, held in conjunction with MICCAI 2022, Singapore, September 22, 2022, proceedings. Domain Adaptation and Representation Transfer (Workshop) (4th: 2022: Sin...* **2022**, *13542*, 12–22. https://doi.org/10.1007/978-3-031-16852-9_2.
39. Yun, J.; Hu, Y.; Kim, J.; Jang, J.; Lee, S. Enhancing Whole Slide Pathology Foundation Models through Stain Normalization. *arXiv preprint arXiv:2408.00380* **2024**.
40. Shi, D.; Zhang, W.; Chen, X.; Liu, Y.; Yang, J.; Huang, S.; Tham, Y.C.; Zheng, Y.; He, M. EyeFound: A Multimodal Generalist Foundation Model for Ophthalmic Imaging, 2024, [arXiv:cs.CV/2405.11338].
41. Jun, E.; Jeong, S.; Heo, D.W.; Suk, H.I. Medical Transformer: Universal Encoder for 3-D Brain MRI Analysis. *IEEE transactions on neural networks and learning systems* **2023**, *PP*. https://doi.org/10.1109/TNNLS.2023.3308712.
42. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.T.; Parekh, Z.; Pham, H.; Le, Q.V.; Sung, Y.; Li, Z.; Duerig, T. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision, 2021, .
43. Zhang, X.; Wu, C.; Zhao, Z.; Lin, W.; Zhang, Y.; Wang, Y.; Xie, W. PMC-VQA: Visual Instruction Tuning for Medical Visual Question Answering, 2023. https://doi.org/10.48550/arXiv.2305.10415.
44. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s, 2022, .
45. Rios-Martinez, C.; Bhattacharya, N.; Amini, A.P.; Crawford, L.; Yang, K.K. Deep self-supervised learning for biosynthetic gene cluster detection and product classification. *PLOS Computational Biology* **2023**, *19*, e1011162. https://doi.org/10.1371/journal.pcbi.1011162.
46. Cheng, J.; Ye, J.; Deng, Z.; Chen, J.; Li, T.; Wang, H.; Su, Y.; Huang, Z.; Chen, J.; Jiang, L.; et al. SAM-Med2D, 2023.
47. Ma, J.; Guo, Z.; Zhou, F.; Wang, Y.; Xu, Y.; Cai, Y.; Zhu, Z.; Jin, C.; Jiang, Y.L.X.; Han, A.; et al. Towards A Generalizable Pathology Foundation Model via Unified Knowledge Distillation. *arXiv preprint arXiv:2407.18449* **2024**.
48. Bao, H.; Dong, L.; Piao, S.; Wei, F. BEiT: BERT Pre-Training of Image Transformers, 2022. https://doi.org/10.48550/arXiv.2106.08254.
49. Shi, X.; Chai, S.; Li, Y.; Cheng, J.; Bai, J.; Zhao, G.; Chen, Y.W. Cross-modality Attention Adapter: A Glioma Segmentation Fine-tuning Method for SAM Using Multimodal Brain MR Images. *arXiv preprint arXiv:2307.01124* **2023**.
50. Ouyang, C.; Biffi, C.; Chen, C.; Kart, T.; Qiu, H.; Rueckert, D. Self-supervised learning for few-shot medical image segmentation. *IEEE Transactions on Medical Imaging* **2022**, *41*, 1837–1848.
51. Oord, A.v.d.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* **2018**.
52. Wu, X.; Jiang, Y.; Xing, H.; Song, W.; Wu, P.; Cui, X.W.; Xu, G. ULS4US: universal lesion segmentation framework for 2D ultrasound images. *Physics in Medicine and Biology* **2023**, *68*. https://doi.org/10.1088/1361-6560/ace09b.
53. Pan, H.; Guo, Y.; Deng, Q.; Yang, H.; Chen, Y.; Chen, J. Improving Fine-tuning of Self-supervised Models with Contrastive Initialization, 2022. https://doi.org/10.48550/arXiv.2208.00238.
54. Brooks, T.; Holynski, A.; Efros, A.A. Instructpix2pix: Learning to follow image editing instructions. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 18392–18402.
55. Campanella, G.; Kwan, R.; Fluder, E.; Zeng, J.; Stock, A.; Veremis, B.; Polydorides, A.D.; Hedvat, C.; Schoenfeld, A.; Vanderbilt, C.; et al. Computational Pathology at Health System Scale–Self-Supervised Foundation Models from Three Billion Images. *arXiv preprint arXiv:2310.07033* **2023**.

56. Hamamci, I.E.; Er, S.; Almas, F.; Simsek, A.G.; Esirgun, S.N.; Dogan, I.; Dasdelen, M.F.; Wittmann, B.; Simsar, E.; Simsar, M.; et al. A foundation model utilizing chest CT volumes and radiology reports for supervised-level zero-shot detection of abnormalities. *arXiv preprint arXiv:2403.17834* **2024**.
57. Zimmermann, E.; Vorontsov, E.; Viret, J.; Casson, A.; Zelechowski, M.; Shaikovski, G.; Tenenholtz, N.; Hall, J.; Fuchs, T.; Fusi, N.; et al. Virchow 2: Scaling Self-Supervised Mixed Magnification Models in Pathology. *arXiv preprint arXiv:2408.00738* **2024**.
58. Wilson, P.F.; Gilany, M.; Jamzad, A.; Fooladgar, F.; To, M.N.N.; Wodlinger, B.; Abolmaesumi, P.; Mousavi, P. Self-supervised learning with limited labeled data for prostate cancer detection in high frequency ultrasound. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* **2023**.
59. de Jong, E.D.; Marcus, E.; Teuwen, J. Current Pathology Foundation Models are unrobust to Medical Center Differences. *arXiv preprint arXiv:2501.18055* **2025**.
60. Ma, D.; Pang, J.; Gotway, M.B.; Liang, J. Foundation Ark: Accruing and Reusing Knowledge for Superior and Robust Performance. In Proceedings of the Medical Image Computing and Computer Assisted Intervention – MICCAI 2023; Greenspan, H.; Madabhushi, A.; Mousavi, P.; Salcudean, S.; Duncan, J.; Syeda-Mahmood, T.; Taylor, R., Eds., Cham, 2023; pp. 651–662. https://doi.org/10.1007/978-3-031-43907-0_62.
61. Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; Xie, W. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 21372–21383.
62. Zhang, Y.; Gao, J.; Tan, Z.; Zhou, L.; Ding, K.; Zhou, M.; Zhang, S.; Wang, D. Data-centric foundation models in computational healthcare: A survey. *arXiv preprint arXiv:2401.02458* **2024**.
63. Felfeliyan, B.; Hareendranathan, A.; Kuntze, G.; Cornell, D.; Forkert, N.D.; Jaremko, J.L.; Ronsky, J.L. Self-Supervised-RCNN for Medical Image Segmentation with Limited Data Annotation, 2022. https://doi.org/10.48550/arXiv.2207.11191.
64. Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Hou, L.; Clark, K.; Pfohl, S.; Cole-Lewis, H.; Neal, D.; et al. Towards Expert-Level Medical Question Answering with Large Language Models, 2023.
65. Aben, N.; de Jong, E.D.; Gatopoulos, I.; Känzig, N.; Karasikov, M.; Lagré, A.; Moser, R.; van Doorn, J.; Tang, F.; et al. Towards Large-Scale Training of Pathology Foundation Models. *arXiv preprint arXiv:2404.15217* **2024**.
66. Wood, D.A.; Townend, M.; Guilhem, E.; Kafiabadi, S.; Hammam, A.; Wei, Y.; Al Busaidi, A.; Mazumder, A.; Sasieni, P.; Barker, G.J.; et al. Optimising brain age estimation through transfer learning: A suite of pre-trained foundation models for improved performance and generalisability in a clinical setting. *Human Brain Mapping* **2024**, *45*, e26625.
67. Vaidya, A.; Zhang, A.; Jaume, G.; Song, A.H.; Ding, T.; Wagner, S.J.; Lu, M.Y.; Doucet, P.; Robertson, H.; Almagro-Perez, C.; et al. Molecular-driven Foundation Model for Oncologic Pathology. *arXiv preprint arXiv:2501.16652* **2025**.
68. Lin, W.; Zhao, Z.; Zhang, X.; Wu, C.; Zhang, Y.; Wang, Y.; Xie, W. PMC-CLIP: Contrastive Language-Image Pre-training using Biomedical Documents, 2023. https://doi.org/10.48550/arXiv.2303.07240.
69. El-Nouby, A.; Touvron, H.; Caron, M.; Bojanowski, P.; Douze, M.; Joulin, A.; Laptev, I.; Neverova, N.; Synnaeve, G.; Verbeek, J.; et al. XCiT: Cross-Covariance Image Transformers, 2021, [arXiv:cs.CV/2106.09681].
70. Niu, C.; Wang, G. Unsupervised contrastive learning based transformer for lung nodule detection. *Physics in Medicine & Biology* **2022**, *67*, 204001.
71. Yan, K.; Cai, J.; Jin, D.; Miao, S.; Guo, D.; Harrison, A.P.; Tang, Y.; Xiao, J.; Lu, J.; Lu, L. SAM: Self-Supervised Learning of Pixel-Wise Anatomical Embeddings in Radiological Images. *IEEE transactions on medical imaging* **2022**, *41*, 2658–2669. https://doi.org/10.1109/TMI.2022.3169003.
72. Mazurowski, M.A.; Dong, H.; Gu, H.; Yang, J.; Konz, N.; Zhang, Y. Segment Anything Model for Medical Image Analysis: an Experimental Study. *Medical Image Analysis* **2023**, *89*, 102918. https://doi.org/10.1016/j.media.2023.102918.
73. Chen, Z.; Xu, Q.; Liu, X.; Yuan, Y. UN-SAM: Universal Prompt-Free Segmentation for Generalized Nuclei Images. *arXiv preprint arXiv:2402.16663* **2024**.
74. Hu, X.; Xu, X.; Shi, Y. How to Efficiently Adapt Large Segmentation Model(SAM) to Medical Images, 2023.
75. Lu, M.Y.; Chen, B.; Williamson, D.F.K.; Chen, R.J.; Ikamura, K.; Gerber, G.; Liang, I.; Le, L.P.; Ding, T.; Parwani, A.V.; et al. A Foundational Multimodal Vision Language AI Assistant for Human Pathology, 2023. https://doi.org/10.48550/arXiv.2312.07814.
76. Nechaev, D.; Pchelnikov, A.; Ivanova, E. Hibou: A Family of Foundational Vision Transformers for Pathology. *arXiv preprint arXiv:2406.05074* **2024**.

77. Shweikh, Y.; Sekimitsu, S.; Boland, M.V.; Zebardast, N. The Growing Need for Ophthalmic Data Standardization. *Ophthalmology Science* **2023**, *3*.
78. Zhang, Y.; Jiang, H.; Miura, Y.; Manning, C.D.; Langlotz, C.P. Contrastive learning of medical visual representations from paired images and text. In Proceedings of the Machine Learning for Healthcare Conference. PMLR, 2022, pp. 2–25.
79. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. DINOv2: Learning Robust Visual Features without Supervision, 2024. https://doi.org/10.48550/arXiv.2304.07193.
80. Yuan, H.; Hong, C. Foundation Model Makes Clustering A Better Initialization For Cold-Start Active Learning, 2024. https://doi.org/10.48550/arXiv.2402.02561.
81. Zhang, Y.; Liu, H.; Hu, Q. TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation, 2021. https://doi.org/10.48550/arXiv.2102.08005.
82. Jun, E.; Jeong, S.; Heo, D.W.; Suk, H.I. Medical Transformer: Universal Brain Encoder for 3D MRI Analysis, 2021, [arXiv:cs.CV/2104.13633].
83. Dai, D.; Zhang, Y.; Xu, L.; Yang, Q.; Shen, X.; Xia, S.; Wang, G. Pa-llava: A large language-vision assistant for human pathology image understanding. *arXiv preprint arXiv:2408.09530* **2024**.
84. Hua, S.; Yan, F.; Shen, T.; Zhang, X. PathoDuet: Foundation Models for Pathological Slide Analysis of H&E and IHC Stains, 2023. https://doi.org/10.48550/arXiv.2312.09894.
85. Zhou, G.; Mosadegh, B. Distilling Knowledge From an Ensemble of Vision Transformers for Improved Classification of Breast Ultrasound. *Academic Radiology* **2024**, *31*, 104–120. https://doi.org/10.1016/j.acra.2023.08.006.
86. Koohbanani, N.A.; Unnikrishnan, B.; Khurram, S.A.; Krishnaswamy, P.; Rajpoot, N. Self-path: Self-supervision for classification of pathology images with limited annotations. *IEEE Transactions on Medical Imaging* **2021**, *40*, 2845–2856.
87. Ye, Y.; Zhang, J.; Chen, Z.; Xia, Y. DeSD: Self-Supervised Learning with Deep Self-Distillation for 3D Medical Image Segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention – MICCAI 2022; Wang, L.; Dou, Q.; Fletcher, P.T.; Speidel, S.; Li, S., Eds., Cham, 2022; pp. 545–555. https://doi.org/10.1007/978-3-031-16440-8_52.
88. Dominic, J.; Bhaskhar, N.; Desai, A.D.; Schmidt, A.; Rubin, E.; Gunel, B.; Gold, G.E.; Hargreaves, B.A.; Lenchik, L.; Boutin, R.; et al. Improving Data-Efficiency and Robustness of Medical Imaging Segmentation Using Inpainting-Based Self-Supervised Learning. *Bioengineering* **2023**, *10*, 207. https://doi.org/10.3390/bioengineering10020207.
89. Ellis, D.; Srigley, J. Does standardised structured reporting contribute to quality in diagnostic pathology? The importance of evidence-based datasets. *Virchows Archiv* **2016**, *468*, 51–59.
90. Zhang, Y.; Jiang, H.; Miura, Y.; Manning, C.D.; Langlotz, C.P. Contrastive Learning of Medical Visual Representations from Paired Images and Text. In Proceedings of the Proceedings of the 7th Machine Learning for Healthcare Conference. PMLR, 2022, pp. 2–25.
91. Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; Gelly, S. Parameter-efficient transfer learning for NLP. In Proceedings of the International conference on machine learning. PMLR, 2019, pp. 2790–2799.
92. Li, G.; Togo, R.; Ogawa, T.; Haseyama, M. Self-supervised learning for gastritis detection with gastric x-ray images. *International Journal of Computer Assisted Radiology and Surgery* **2023**, *18*, 1841–1848.
93. Thawkar, O.; Shaker, A.; Mullappilly, S.S.; Cholakkal, H.; Anwer, R.M.; Khan, S.; Laaksonen, J.; Khan, F.S. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971* **2023**.
94. Tay, Y.; Dehghani, M.; Bahri, D.; Metzler, D. Efficient Transformers: A Survey. *ACM Comput. Surv.* **2022**, *55*. https://doi.org/10.1145/3530811.
95. Deng, G.; Zou, K.; Ren, K.; Wang, M.; Yuan, X.; Ying, S.; Fu, H. SAM-U: Multi-box Prompts Triggered Uncertainty Estimation for Reliable SAM in Medical Image. In Proceedings of the Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 Workshops; Woo, J.; Hering, A.; Silva, W.; Li, X.; Fu, H.; Liu, X.; Xing, F.; Purushotham, S.; Mathai, T.S.; Mukherjee, P.; et al., Eds., Cham, 2023; pp. 368–377. https://doi.org/10.1007/978-3-031-47425-5_33.
96. Xie, Y.; Zhang, J.; Shen, C.; Xia, Y. CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation, 2021. https://doi.org/10.48550/arXiv.2103.03024.

97. Felfeliyan, B.; Forkert, N.D.; Hareendranathan, A.; Cornel, D.; Zhou, Y.; Kuntze, G.; Jaremko, J.L.; Ronsky, J.L. Self-supervised-RCNN for medical image segmentation with limited data annotation. *Computerized Medical Imaging and Graphics* **2023**, *109*, 102297.
98. Lu, M.Y.; Chen, B.; Zhang, A.; Williamson, D.F.; Chen, R.J.; Ding, T.; Le, L.P.; Chuang, Y.S.; Mahmood, F. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 19764–19775.
99. Zhu, W.; Chen, Y.; Nie, S.; Yang, H. SAMMS: Multi-modality Deep Learning with the Foundation Model for the Prediction of Cancer Patient Survival. IEEE Computer Society, 2023, pp. 3662–3668. https://doi.org/10.1109/BIBM58861.2023.10385661.
100. Shaikovski, G.; Casson, A.; Severson, K.; Zimmermann, E.; Wang, Y.K.; Kunz, J.D.; Retamero, J.A.; Oakley, G.; Klimstra, D.; Kanan, C.; et al. PRISM: A Multi-Modal Generative Foundation Model for Slide-Level Histopathology, 2024, [arXiv:eess.IV/2405.10254].
101. Lin, W.; Zhao, Z.; Zhang, X.; Wu, C.; Zhang, Y.; Wang, Y.; Xie, W. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2023, pp. 525–536.
102. Walsh, J.; Othmani, A.; Jain, M.; Dev, S. Using U-Net network for efficient brain tumor segmentation in MRI images. *Healthcare Analytics* **2022**, *2*, 100098. https://doi.org/10.1016/j.health.2022.100098.
103. Gutiérrez, J.D.; Rodriguez-Echeverria, R.; Delgado, E.; Rodrigo, M.Á.S.; Sánchez-Figueroa, F. No More Training: SAM's Zero-Shot Transfer Capabilities for Cost-Efficient Medical Image Segmentation. *IEEE Access* **2024**.
104. Yan, K.; Cai, J.; Jin, D.; Miao, S.; Guo, D.; Harrison, A.P.; Tang, Y.; Xiao, J.; Lu, J.; Lu, L. SAM: Self-supervised learning of pixel-wise anatomical embeddings in radiological images. *IEEE Transactions on Medical Imaging* **2022**, *41*, 2658–2669.
105. Khan, M.O.; Afzal, M.M.; Mirza, S.; Fang, Y. How Fair are Medical Imaging Foundation Models? In Proceedings of the Proceedings of the 3rd Machine Learning for Health Symposium. PMLR, 2023, pp. 217–231.
106. Balestriero, R.; Ibrahim, M.; Sobal, V.; Morcos, A.; Shekhar, S.; Goldstein, T.; Bordes, F.; Bardes, A.; Mialon, G.; Tian, Y.; et al. A Cookbook of Self-Supervised Learning, 2023. https://doi.org/10.48550/arXiv.2304.12210.
107. Hua, Y.; Yan, Z.; Kuang, Z.; Zhang, H.; Deng, X.; Yu, L. Symmetry-Aware Deep Learning for Cerebral Ventricle Segmentation With Intra-Ventricular Hemorrhage. *IEEE Journal of Biomedical and Health Informatics* **2022**, *26*, 5165–5176.
108. Lu, J.; Yan, F.; Zhang, X.; Gao, Y.; Zhang, S. Pathotune: Adapting visual foundation model to pathological specialists. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2024, pp. 395–406.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.