Medical Image Analysis 111 (2026) 104062



Contents lists available at ScienceDirect

Medical Image Analysis

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



Review



Learning with less supervision: A survey of label-efficient learning for medical image analysis

Cheng Jin [a,1], Zhengrui Guo [a,1], Yi Lin [b], Luyang Luo [c], Hao Chen [a,d,e,f,g,*]

[a] Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Kowloon Hong Kong SAR, China
[b] Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA
[c] Department of Biomedical Informatics, Harvard University, Cambridge, MA, USA
[d] Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Kowloon Hong Kong SAR, China
[e] Division of Life Science, The Hong Kong University of Science and Technology, Hong Kong SAR, China
[f] HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Futian Shenzhen, China
[g] State Key Laboratory of Nervous System Disorders, The Hong Kong University of Science and Technology, Hong Kong SAR, China

ARTICLE INFO

Keywords:
Medical image analysis
Label-efficient learning
Health foundation models

ABSTRACT

Deep learning has significantly advanced medical imaging analysis (MIA), achieving state-of-the-art performance across diverse clinical tasks. However, its success largely depends on large-scale, high-quality labeled datasets, which are costly and time-consuming to obtain due to the need for expert annotation. To mitigate this limitation, label-efficient deep learning methods have emerged to improve model performance under limited supervision by leveraging labeled, unlabeled, and weakly labeled data. In this survey, we systematically review relevant peer-reviewed studies as well as influential preprints and present a comprehensive taxonomy of label-efficient learning methods in MIA. These methods are categorized into four labeling paradigms: *no label*, *insufficient label*, *inexact label*, and *label refinement*. For each scenario, we analyze representative techniques across imaging modalities and clinical tasks, and highlight shared methodological principles as well as adaptations. Crucially, we emphasize how the advent of health foundation models (HFMs) has fundamentally transformed label-efficient learning in medical imaging. Finally, we discuss ongoing challenges and outline future research directions spanning the research-to-deployment continuum. By synthesizing recent advances and open questions, this survey aims to provide a unified perspective to guide the development and clinical translation of robust, label-efficient solutions for medical image analysis.

1. Introduction

Deep learning (DL) has revolutionized medical image analysis (MIA), significantly improving the efficiency and accuracy of disease detection, diagnosis, and treatment (De Fauw et al., 2018b; Cao et al., 2023; Vorontsov et al., 2024). By providing a data-driven framework for interpreting large and diverse medical image datasets, DL models have achieved unprecedented performance.

Despite these advancements, the success of DL models remains heavily dependent on large volumes of precisely annotated data, which are costly and time-consuming to obtain due to the need for expert input (Yu et al., 2021b). This rising demand for annotations contrasts sharply with the limited availability of medical experts (Rosenkrantz et al., 2016; Lu and Chen, 2020), exacerbating the disparity between the rapidly expanding volume of medical images and the constrained capacity for their annotation. Consequently, reducing annotation costs, accelerating annotation workflows, and alleviating the burden on annotators have become critical challenges in DL-based MIA.

To address the annotation bottleneck in medical imaging, researchers have proposed a variety of learning paradigms, including self-supervised, semi-supervised, weakly supervised, and active learning. These approaches are designed to handle scenarios where annotations are missing, limited, imprecise, or require refinement. By leveraging different levels of supervision, ranging from pixel-level labels to weaker forms such as points, scribbles, bounding boxes, or even unlabeled data, they provide flexibility across diverse labeling conditions. The emergence of health foundation models (HFMs) has further strengthened these strategies by pretraining on large-scale medical datasets to extract generalizable features. These features can be effectively transferred to downstream tasks such as classification, segmentation, or detection, which improves performance and reduces the need for extensive labeled data during fine-tuning (He et al., 2024b). In this paper, we refer to

* Corresponding author.
E-mail address: jhc@cse.ust.hk (H. Chen).
[1] Equal contribution.

https://doi.org/10.1016/j.media.2026.104062
Received 2 June 2025; Received in revised form 28 February 2026; Accepted 30 March 2026
Available online 2 April 2026
1361-8415/© 2026 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

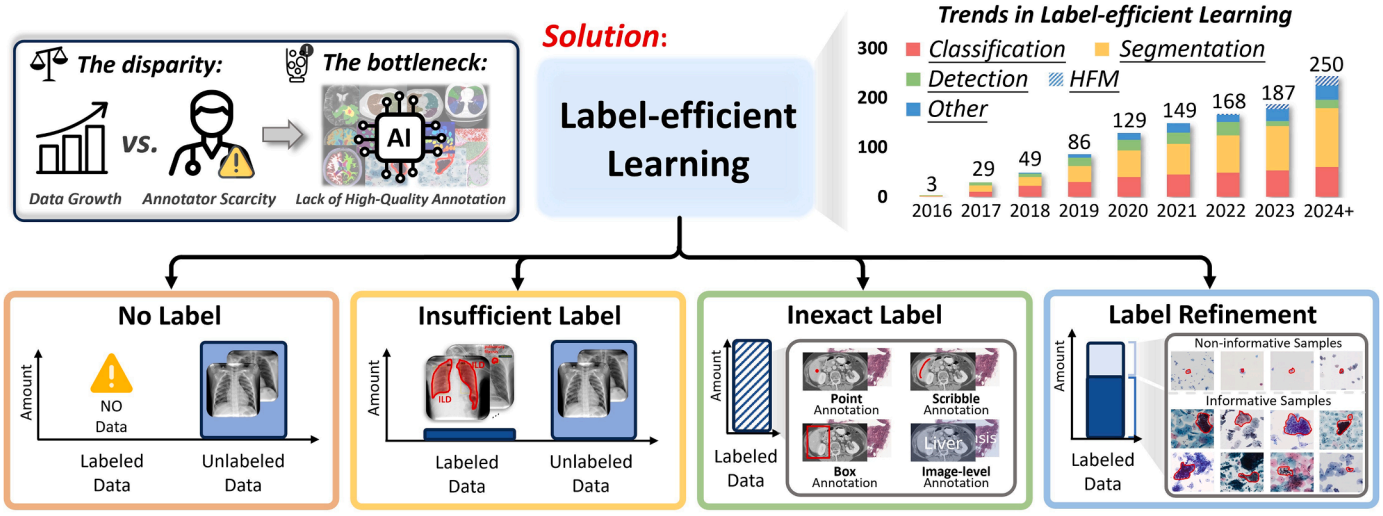


Fig. 1. Motivation for label-efficient learning in medical image analysis. The expanding volume of imaging data and limited expert annotations result in an annotation bottleneck. Common scenarios of label scarcity encountered in clinical practice are illustrated, including no label, insufficient labels, inexact labels, and label refinement.

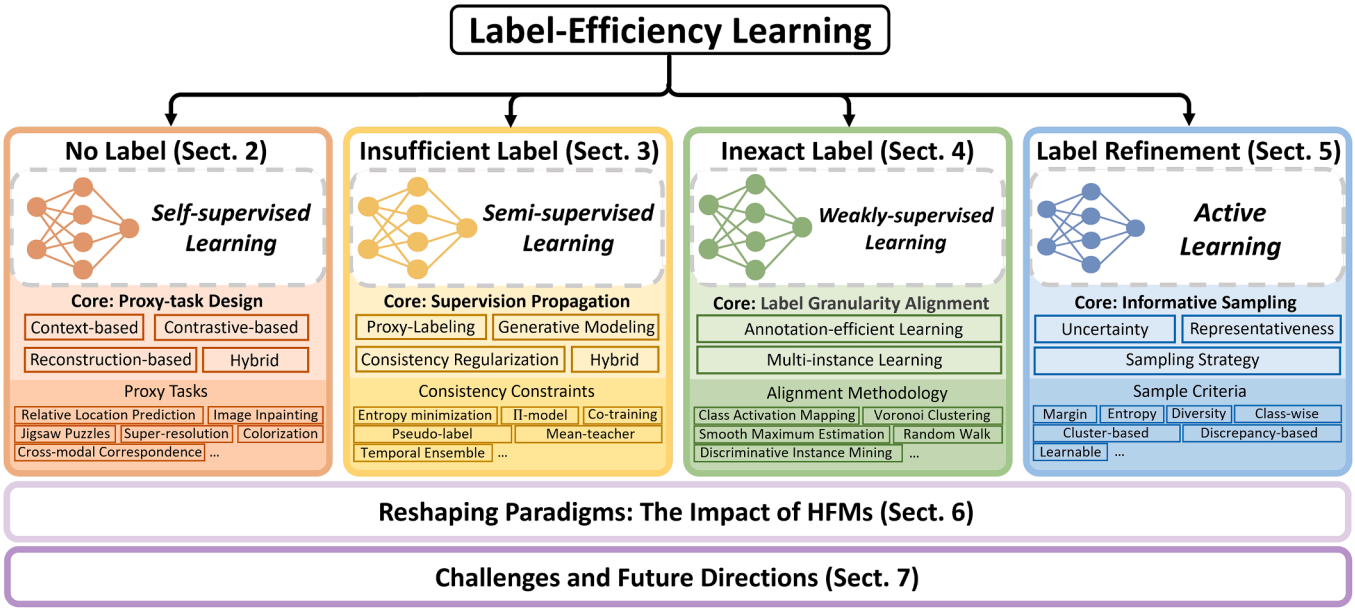


Fig. 2. Overview of this survey. This survey categorizes approaches based on four labeling scenarios: No label (Section 2), insufficient label (Section 3), inexact label (Section 4), and label refinement (Section 5). The survey further examines how health foundation models (HFMs) are reshaping these approaches (Section 6) and concludes with a discussion of current challenges and future research directions (Section 7).

the full spectrum of these methods, both traditional and HFM-based, as **label-efficient learning**. As illustrated in Fig. 1, label-efficient learning methods have rapidly expanded in recent years. High-level tasks such as classification, segmentation, and detection remain their primary focus, while applications to low-level tasks, including denoising, image registration, and super-resolution, are also gaining momentum. This growing versatility underscores the increasing impact of label-efficient learning across the MIA pipeline, supporting its broader integration into both research and clinical workflows.

Several surveys have previously addressed label-efficient learning in MIA, each offering valuable insights yet exhibiting certain limitations. Cheplygina et al. (2019) introduced the term "not-so-supervised" learning and categorized methods into supervised, semi-supervised, multi-instance, and transfer learning. However, their work focused on theoretical concepts and lacked practical relevance to clinical MIA applications. Budd et al. (2021) emphasized human-in-the-loop strategies, while Wang et al. (2024c) provided a more recent survey, but covered only a subset of label-efficient methods, limiting its scope. In contrast, our taxonomy is organized around specific annotation scenarios, offering a more intuitive and application-oriented framework. Furthermore, we analyze how HFMs are reshaping traditional paradigms within each scenario, highlighting open research questions that merit future exploration, as shown in Fig. 2.

In an effort to provide a comprehensive overview of label-efficient learning in MIA, we review relevant peer-reviewed studies as well as influential preprints and categorize them into four annotation scenarios: *no label*, where data lacks annotations; *insufficient label*, where labeled data is limited; *inexact label*, where the annotations are noisy or coarse; and *label refinement*, where existing labels require improvement. The literature included in this survey was identified through a systematic search of Google Scholar and the arXiv preprint server for relevant publications up to March 2025. In addition, we examined major conference

proceedings, such as CVPR, ICCV, ECCV, NeurIPS, AAAI, and MICCAI, as well as leading journals including *Medical Image Analysis* (MIA), *IEEE Transactions on Medical Imaging* (TMI), and *Nature Biomedical Engineering*, by screening paper titles and abstracts. Reference lists of selected papers were further reviewed to ensure comprehensive coverage, and in cases where similar work appeared in multiple venues, only the most impactful or comprehensive publication was included. Appendix A describes the search process in more detail. To the best of our knowledge, this is the first extensive review to systematically cover all four scenarios. For each, we define the core challenges, provide essential background, and examine the role of HFMs in enhancing performance. Our analysis not only synthesizes recent progress but also identifies key limitations and outlines future research directions, offering a roadmap for advancing label-efficient learning in MIA. The remainder of this paper is organized as follows. Sections 2–5 introduce the four annotation scenarios: *no label* in Section 2, *insufficient label* in Section 3, *inexact label* in Section 4, and *label refinement* in Section 5. Section 6 reviews the role and impact of HFM in label-efficient learning. Section 7 discusses current challenges and explores potential solutions and research opportunities. Finally, we conclude this survey in Section 8.

2. No label

The *no label* scenario represents one of the most challenging yet prevalent settings in MIA, where large volumes of imaging data are available but no annotations exist. This commonly arises in applications involving novel modalities, rare diseases, or legacy datasets lacking corresponding labels. In the absence of ground truth, conventional supervised learning becomes inapplicable. To address this, **self-supervised learning (Self-SL)** has emerged as an effective solution. Self-SL derives supervisory signals directly from the data itself by designing pretext tasks that exploit the inherent structural and semantic patterns within medical images. As illustrated in Fig. 3, Self-SL enables models to learn meaningful and transferable representations from unlabeled data through automatically generated supervision signals. Typically, the workflow involves first pretraining a model on large-scale unlabeled data using self-supervised objectives, followed by fine-tuning on a labeled dataset. This strategy not only facilitates data-specific feature learning but also helps reduce overfitting by leveraging the abundance of unlabeled data. Moreover, recent advances have scaled up both the size of datasets and the capacity of models, giving rise to HFMs. These models have demonstrated strong diagnostic and prognostic performance, along with robust generalization across diverse medical imaging tasks. Based on the design of proxy tasks, we categorize Self-SL methods in MIA as **reconstruction-based**, **context-based**, and **contrastive-based** approaches. Representative approaches within each category are summarized in Table 1.

2.1. Reconstruction-based methods

Reconstruction-based methods aim to learn semantic features by reconstructing input data, enabling representation learning without manual annotations. The core principle involves training an encoder—decoder architecture in which the encoder learns to extract meaningful representations by attempting to reconstruct the original input—or a deliberately corrupted version—through the decoder. This reconstruction objective forces the model to capture essential structural and semantic information required to faithfully reproduce the input, thereby encouraging the acquisition of rich and transferable features. The quality of reconstruction itself acts as a self-supervised signal, where better representations naturally lead to more accurate reconstructions. In medical image analysis, such reconstruction-based objectives have been instantiated in diverse forms, including *super-resolution* (Zhao et al., 2020a; Li et al., 2021a), which emphasizes recovering fine-grained anatomical details from low-resolution scans; *inpainting* (Zhao et al., 2021a), which promotes contextual reasoning by restoring missing regions; *colorization* (Abbet et al., 2020), which links structural cues to appropriate color or staining patterns; and *multi-modal reconstruction* (Hervella et al., 2018; Cao et al., 2020a), which leverages complementary information across imaging modalities. In the following, we examine these task formulations in turn, beginning with super-resolution as one of the most clinically relevant forms of self-supervised reconstruction in MIA.

A basic form of reconstruction was introduced by Li et al. (2020a), who trained an auto-encoder to reconstruct normal-dose CT images by minimizing mean squared error (MSE) loss. The encoder, pre-trained via self-supervision, was later reused for downstream tasks. However, reliance on pixel-wise losses may bias models toward low-level appearance features while overlooking structural semantics (Abbet et al., 2020), prompting the development of more informative proxy tasks.

Super-resolution tasks generate high-resolution outputs from low-resolution inputs, forcing the model to learn fine-grained semantic structures. Zhao et al. (2020a) proposed an anti-aliasing super-resolution method for MRIs, while Li et al. (2023a) leveraged frequency information for fundus image enhancement. GAN-based super-resolution has also been applied to gigapixel histopathology WSIs (Li et al., 2021a). These super-resolution approaches are particularly effective in domains where high-resolution details are critical for diagnosis, and they encourage the model to focus on subtle anatomical structures. However, they may also lead to overfitting to texture or noise if not properly regularized, and their effectiveness can be limited in cases where semantic content is not directly tied to resolution.

Colorization tasks predict RGB images from grayscale inputs, encouraging the model to capture structural and contextual cues. Abbet et al. (2020) applied this to colorectal cancer survival analysis by converting images into hematoxylin and eosin channels. Fan et al. (2023) extended this approach with a cross-channel task that reconstructs lightness from color channels. Compared to pure pixel reconstruction, colorization requires the model to infer semantic correspondences between structure and color, thus promoting higher-level feature learning. Nonetheless, the colorization task may introduce ambiguity, as multiple plausible color mappings can exist for the same structure, potentially limiting the precision of the learned representations.

Inpainting tasks require models to restore missing regions, promoting understanding of object continuity and context. Zhao et al. (2021a) applied this strategy to OCT and chest X-rays, showing improved anatomical feature learning. Inpainting is especially useful for encouraging global context reasoning, as the model must integrate information from surrounding regions to plausibly reconstruct missing parts. However, if the occluded regions are too small or too easily inferred from context, the task may become trivial and fail to drive the model towards learning meaningful semantics.

Multi-modal reconstruction tasks aim to reconstruct one modality from another, facilitating joint representation learning. Hervella et al. (2018) combined retinography and angiography for retinal analysis, while Cao et al. (2020a) proposed a collaborative learning framework using auto-encoders and GANs to synthesize missing modalities. These approaches are well-suited for leveraging complementary information across imaging modalities, which is common in clinical practice. Yet, the challenge remains in aligning heterogeneous data distributions and ensuring that the shared representations capture clinically relevant features rather than modality-specific artifacts.

Recent advances in diffusion models (Ho et al., 2020; Yang et al., 2022a; Rombach et al., 2022) have further improved self-supervised reconstruction. These models, characterized by stability and generative fidelity, have been applied to medical imaging by Korkmaz et al. (2023), Cho et al. (2023), and Purma et al. (2024), enabling robust feature learning from complex anatomical data. While diffusion models offer superior generative performance and can model complex data distributions, they typically require significant computational resources and their interpretability in clinical contexts remains an open question.

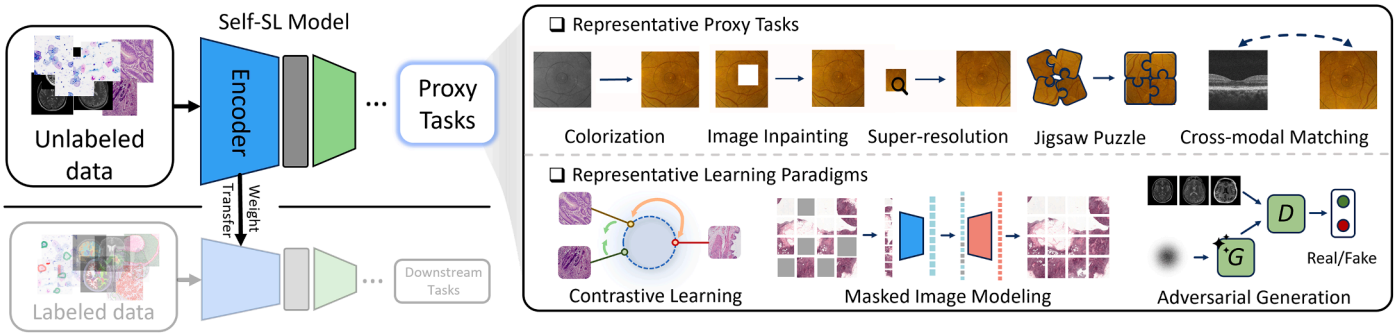


Fig. 3. Overview of self-supervised learning (Self-SL) paradigm which addresses the *no label* scenario. Self-SL aims to learn a pre-trained model by developing various proxy tasks based solely on unlabeled data. Then the pre-trained model can be fine-tuned on different downstream tasks with labeled datasets.

In summary, reconstruction-based self-supervised methods provide a flexible framework for learning rich representations from unlabeled medical images. Each proxy task emphasizes different aspects of image semantics—whether fine-grained structure, global context, or cross-modal relationships. However, these methods also face challenges, such as balancing low-level appearance with high-level semantics, handling ambiguous or ill-posed reconstruction targets, and ensuring clinical relevance of learned features. Going forward, integrating more task-aware proxy objectives and leveraging advances in generative modeling hold promise for further improving the effectiveness and applicability of reconstruction-based self-supervised learning in medical imaging.

2.2. Context-based methods

While reconstruction-based methods focus on pixel-level restoration, **context-based methods** take a different approach by exploiting the spatial and structural relationships within medical images. These methods typically formulate self-supervised learning as predictive objectives, where models learn meaningful representations by predicting various contextual properties such as spatial arrangements, anatomical relationships, or structural characteristics inherent in the data. Researchers have explored novel predictive tasks for specific MIA tasks by training the network for prediction of the output class or the localization of objects with the original image as the supervision signal (Bai et al., 2019; Spitzer et al., 2018; Xu et al., 2020; Srinidhi et al., 2022). Bai et al. (2019) propose a proxy task to predict the anatomical positions from cardiac chamber view planes by applying an encoder-decoder structure. This proxy task properly employs the chamber view plane information, which is available from cardiac MR scans easily. Zheng et al. (2023) enhance medical image segmentation through multi-scale consistency objectives for finer-grained representation across different target scales. For 3D medical images, He et al. (2023) introduce Geometric Visual Similarity Learning, incorporating topological invariance prior to ensure consistent representation of semantic regions. Meanwhile, Srinidhi et al. (2022) propose Resolution Sequence Prediction for whole slide images (WSIs), training networks to predict the order of multi-resolution patches, thereby capturing both contextual structure at lower magnifications and local details at higher magnifications.

Other efforts have been made to explore the spatial context structure of input data, such as the order of different patches constituting an image, or the relative position of several patches in the same image, which can provide useful semantic features for the network. Chen et al. (2019) introduce context restoration, where patch positions are randomly switched and then restored, enabling straightforward semantic feature learning. Li et al. (2021e) employ rotation angle prediction with augmented retinal images, encouraging the model to predict rotation angles while clustering similar features. Advanced spatial reasoning tasks such as jigsaw puzzles and Rubik's Cube have gained traction in medical imaging. Taleb et al. (2021) enhance Jigsaw Puzzles with multi-modal data, requiring models to restore original images from out-of-order patches of different modalities. For 3D medical data, Zhuang et al. (2019) and Tao et al. (2020) adapt Rubik's Cube by dividing volumes into cube grids, applying random rotations, and training networks to recover the volume.

However, for WSI images, common proxy tasks such as prediction of the rotation or relative position of objects may only provide minor improvements to the model in histopathology due to the lack of a sense of global orientation in WSIs (Graham et al., 2020; Koohbanani et al., 2021). To address these limitations, Koohbanani et al. (2021) propose proxy tasks targeted at histopathology, namely, magnification prediction, solving the magnification puzzle, and hematoxylin channel prediction. In this way, their model can significantly integrate and learn the contextual, multi-resolution, and semantic features inside the WSIs.

In summary, context-based self-supervised methods provide a versatile framework for capturing semantic and structural information in medical images. Their success largely depends on the careful design of proxy tasks that align with the intrinsic properties of the data and the requirements of downstream applications. Nonetheless, challenges remain in selecting context cues that are both informative and robust across diverse imaging modalities. Future work may focus on dynamic or adaptive task selection, as well as integrating multiple context signals to further enhance feature learning in MIA.

2.3. Contrastive-based methods

Beyond the pixel-space focus of reconstruction approaches and the spatial reasoning of context-based methods, **contrastive-based methods** operate on the principle that representations of different views of the same image should be similar, while those of different images should be distinguishable. The learning mechanism involves creating positive pairs (different augmented views of the same image) and negative pairs (views from different images), then training the model to maximize similarity between positive pairs while minimizing similarity between negative pairs in the learned representation space. This contrastive objective encourages the model to learn invariant features that capture semantic content while remaining robust to various transformations and augmentations. The resulting representations effectively encode semantic similarities and differences across the dataset.

Several high-performance contrastive learning algorithms originally developed for natural images, such as SimCLR (Chen et al., 2020) and BYOL (Grill et al., 2020), have been successfully adapted to medical image analysis (Hou et al., 2021; Azizi et al., 2021; Wang et al., 2021d). Azizi et al. (2021) developed multi-instance contrastive learning (MICLe), extending SimCLR by minimizing disagreement between views from multiple images of the same patient, creating richer positive pair relationships. In parallel, Wang et al. (2021d) applied the BYOL architecture to histopathology image classification, making a substantial contribution by assembling the largest WSI dataset for Self-SL pre-training at the time—comprising 2.7 million patches from 32,529 WSIs spanning over 25 anatomic sites and 32 cancer subtypes.

Table 1
Surveyed representative self-supervised learning-based studies for the *No Label* scenario, organized by downstream task type, with corresponding target organs, proxy tasks, datasets, and publication venues.

	Reference	Organ	Proxy Task Design	Dataset	Publication
Classification	Li et al. (2020d)	Retina	Multi-modal Contrastive Learning	ADAM; PALM	TMI 2020
	Koohbanani et al. (2021)	Breast; Cervix; Colon	Magnification Prediction; Solving Magnification Puzzle; Hematoxylin Channel Prediction	CAMELYON 2016; KATHER; Private Dataset: 217 Images	TMI 2021
	Azizi et al. (2021)	Skin; Lung	Multi-Instance Contrastive Learning	Priavte Dermatology Dataset; CheXpert	CVPR 2021
	Chen et al. (2022b)	Breast; Lung; Kidney	Contrastive Learning	TCGA-BRCA; TCGA-NSCLS; TCGA-RCC	CVPR 2022
	Mahapatra et al. (2022)	Lymph; Lung; Retina; Prostate	Contrastive Learning Variant	CAMELYON 2017; DRD; GGC	TMI 2022
	Wang et al. (2023a)	Skin	Self-supervised Knowledge Distillation	ISIC 2019	MIA 2023
	Huang et al. (2024)	Multi-Organ	SimCLR, MOCOv2, SwAV, BYOL, SimSiam, DINO, BarlowTw	TissueMNIST; PathMNIST; TMED-2; AIROGS	CVPR 2024
	Tang et al. (2024a)	Lung	Self-Supervised Representation Distribution Learning	TCGA-EGFR; TCGA-Lung; Private Lung Dataset	TMI 2024
Segmentation	Hervella et al. (2018)	Retina	Multi-modal Reconstruction	Isfahan MISP	MICCAI 2018
	Spitzer et al. (2018)	Brain	Patch Distance Prediction	BigBrain	MICCAI 2018
	Bai et al. (2019)	Heart	Anatomical Position Prediction	Private Dataset: 3825 Subjects	MICCAI 2019
	Sahasrabudhe et al. (2020)	Multi-Organ	WSI Patch Magnification Identification	MoNuSeg	MICCAI 2020
	Tao et al. (2020)	Pancreas	Rubik's Cube Recovery	NIH PCT; MRBrainS18	MICCAI 2020
	Lu et al. (2021b)	Brain	Fiber Streamlines Density Map Prediction; Registration-based Segmentation Imitation	dHCP	MIA 2021
	Tang et al. (2022)	Abdomen; Liver; Prostate	Contrastive Learning; Masked Volume Inpainting; 3D Rotation Prediction	DECATHLON; BTCV	CVPR 2022
	Jiang et al. (2023)	Multi-organ	Anatomical-invariant Contrastive Learning	FLARE 2022; BTCV	CVPR 2023
	He et al. (2023)	Heart; Artery; Brain	Geometric Visual Similarity Learning	MM-WHS-CT; ASOCA; CANDI; STOIC	CVPR 2023
	Liu et al. (2023c)	Tooth	Hierarchical Global-local Contrastive Learning	Private Dataset: 13,000 Scans	TMI 2023
	Zheng et al. (2023)	Multi-Organ	Multi-scale Visual Representation Self-supervised Learning	BCV; MSD; KiTS	TMI 2023
	Peng et al. (2024)	Heart; Prostate	Contrastive Learning	ACDC; PROMISE12	MIA 2024
	Purma et al. (2024)	Multi-Organ	Diffusion-based Reconstruction	Head and Neck Cancer; GlaS; MoNuSeg	TMI 2024
Regression	Abbet et al. (2020)	Gland	Image Colorization	Private Dataset: 660 Images	MICCAI 2020
	Srinidhi et al. (2022)	Breast; Colon	WSI Patch Resolution Sequence Prediction	BreastPathQ; CAMELYON 2016; KATHER	MIA 2022
	Fan et al. (2023)	Brain; Lung	Image Colorization; Cross-channel	GBM; TCGA-LUSC; NLST	TMI 2023
Others	Zhuang et al. (2019)	Brain	Rubik's Cube Recovery	BraTS 2018; Private Dataset: 1,486 Images	MICCAI 2019
	Chen et al. (2019)	Multi-Organ	Disturbed Image Context Restoration	Private Fetus Dataset: 2,694 Images; Private Multi-organ Dataset: 150 Images; BraTS 2017	MIA 2019
	Zhao et al. (2020a)	Brain	Super-resolution Reconstruction	Private Dataset: 47 Images	TMI 2020
	Li et al. (2020a)	Abdomen	CT Reconstruction	LDCTGC	TMI 2020
	Cao et al. (2020a)	Brain	Missing Modality Synthesis	BraTS 2015; ADNI	AAAI 2020

Table 1
Continue

Reference	Organ	Proxy Task Design	Dataset	Publication
Haghighi et al. (2020)	Lung	Self-Discovery + Self-Classification +Self-Restoration	LUNA; LiTS; CAD-PE; BraTS 2018; ChestX-ray14; LIDC-IDRI; SIIM-ACR	MICCAI 2020
Taleb et al. (2020)	Brain; Retina; Pancreas	3D Contrastive Predictive Coding; 3D Jigsaw Puzzles; 3D Rotation Prediction; 3D Exemplar Networks Relative 3D Patch Location;	BraTS 2018; DECATHLON; DRD	NeurIPS 2020
Li et al. (2021a)	Breast; Pancreas; Kidney	Super-resolution Reconstruction; Color Normalization	WTS; Private Dataset: 533 Images	MIA 2021
Wang et al. (2021d)	Multi-Organ	Contrastive Learning	TCGA; KATHER; MHIST PAIP; PatchCAMELYON	MICCAI 2021
Zhou et al. (2021a)	Lung; Brain; Liver	Contrastive Learning + Image Reconstruction	ChestX-ray14; CheXpert; LUNA BraTS 2018; LiTS;	CVPR 2021
Yan et al. (2022)	Multi-Organ	Global and Local Contrastive Learning	DeepLesion; NIH LN; Private Dataset: 94 Patients	TMI 2022
Haghighi et al. (2022)	Lung	Contrastive Learning + Reconstruction + Adversarial Learning	ChestX-ray14; CheXpert; Montgomery	CVPR 2022
Cai et al. (2023)	Lung; Brain; Retina	Dual-Distribution Reconstruction	RSNA-Lung; LAG; VinDr-CXR; Brain Tumor MRI; Private Lung Dataset: 5,000 Images	MIA 2023
Li et al. (2023a)	Retina	Frequency-boosted Image Enhancement	EyePACS; Private Dataset: more than 10,000 Images	MIA 2023

Large-scale dataset utilization has emerged as a critical trend in contrastive-based methods. Ghesu et al. (2022) developed a contrastive learning and online clustering algorithm leveraging over 100 million radiography, CT, MRI, and ultrasound images. This extensive pre-training yielded significant improvements in both performance and convergence rates compared to previous state-of-the-art approaches. Similarly, Nguyen et al. (2023) integrated over 1.3 million multi-modal images from 55 publicly available datasets to enhance representation learning. Beyond multi-view approaches, Jiang et al. (2023) introduced a specialized contrastive objective for learning anatomically invariant features, designed to exploit inherent similarities in anatomical structures across diverse medical imaging volumes.

Subsequent research has further refined contrastive learning by incorporating both global and local contrast for more comprehensive representation learning, typically using InfoNCE loss (van den et al., 2018). Yan et al. (2022) apply this at the pixel level to generate embeddings that accurately describe anatomical locations, creating representations at both global and local scales. Building on this multi-level approach, Liu et al. (2023c) develop hierarchical contrastive learning for intra-oral mesh scans, capturing unsupervised representations across point-level, region-level, and cross-level interactions.

Compared to reconstruction- and context-based methods, contrastive learning is less reliant on proxy task design and better suited to leveraging large unlabeled datasets, often resulting in more robust and transferable representations. However, it faces challenges such as constructing meaningful positive/negative pairs and mitigating spurious correlations, especially in settings with limited patient diversity or multi-modal data. Looking ahead, advances in pair mining and domain-adaptive objectives are expected to further improve the effectiveness and clinical relevance of contrastive-based self-supervised learning in medical imaging.

2.4. Discussion

Self-SL methods aim to leverage unlabeled data to learn rich, transferable representations by designing effective proxy tasks. While many existing approaches directly adapt proxy tasks from natural image domains, **unique characteristics of medical images derived from various imaging modalities, such as CT, MRI, and WSI,** *necessitate tailored proxy task designs* that account for domain-specific semantics and structures. Building on this general perspective, it is important to consider how the suitability of a given Self-SL approach can vary according to the properties of specific imaging modalities. **Modality-Specific Strategy Selection.** The choice of Self-SL strategy often depends on the imaging modality: *reconstruction-based methods* generally perform well for modalities with clear anatomical structure (i.e., CT and MRI), *contrastive-based approaches* excel with high-dimensional histopathology data that benefits from instance discrimination, while *context-based methods* are particularly effective when spatial relationships are diagnostically significant. Researchers have also explored *hybrid frameworks* that combine multiple types of Self-SL tasks to capture diverse aspects of medical data (Tang et al., 2022; Zhou et al., 2021a; Haghighi et al., 2020; Yang et al., 2022c). Despite these advances, systematic guidelines for strategy selection remain elusive, and most approaches rely on empirical validation rather than principled design principles.

Real-World Clinical Impact and Performance Gains. The methodological advances have led to significant clinical benefits. Self-SL methods have demonstrated substantial clinical value in practical applications. For instance, CheXzero (Tiu et al., 2022) uses contrastive learning on chest X-ray and radiology report pairs to achieve radiologist-level pathology detection without manual annotations, while enabling zero-shot classification of unseen conditions. Similarly, self-supervised Vision Transformers trained on histopathology data (Cisternino et al., 2024) achieve 92% accuracy in tissue classification and can identify missed pathological features not annotated in original reports. These examples illustrate how Self-SL reduces annotation burden while improving diagnostic accuracy and enabling the discovery of unannotated pathologies—capabilities essential for real-world medical applications where expert-level time and labor are precious. However, critical limitations persist in this field. The lack of standardized evaluation protocols

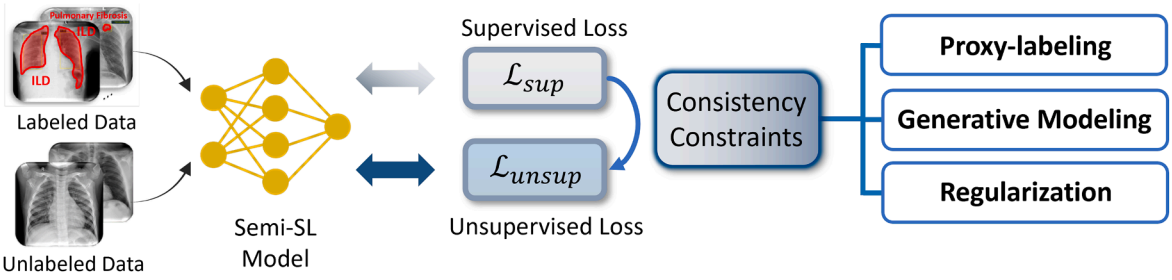


Fig. 4. Overview of Semi-supervised learning (Semi-SL) paradigm which addresses the *insufficient label* scenario. Semi-SL leverages a small amount of labeled data and a large amount of unlabeled data to jointly train a model. The goal is to ensure prediction consistency between labeled and unlabeled examples, typically by optimizing a combination of supervised ($\mathcal{L}_{sup}$) and unsupervised ($\mathcal{L}_{unsup}$) loss functions.

across studies makes fair comparison difficult, and the interpretability requirements essential for clinical adoption remain largely unaddressed. Moreover, the robustness of these methods under distribution shifts, which is common in multi-center clinical deployment, requires more rigorous investigation.

3. Insufficient label

The *insufficient label* scenario arises when only a small portion of available medical imaging data is annotated, while the majority remains unlabeled. This scenario is prevalent in clinical settings, where expert annotations are costly and time-consuming, yet raw images are readily accessible. In such cases, supervised learning alone proves inadequate due to limited labeled data.

As illustrated in Figs. 4–6, **semi-supervised learning (Semi-SL)** addresses this challenge by leveraging both labeled and unlabeled data during training. A core principle of Semi-SL is supervision propagation, which assumes that unlabeled data should have predictions consistent with labeled data. In this way, Semi-SL enables better generalization while reducing the need for manual annotation.

In the following, we categorize existing Semi-SL methods in MIA based on how each method enforces or approximates supervision consistency between labeled and unlabeled data into three categories: **proxy-labeling**, **generative modeling**, and **regularization**. Representative works are summarized in Appendix Table 2.

3.1. Proxy-labeling methods

Proxy-labeling methods utilize the idea of supervision consistency propagation by assigning pseudo labels to unlabeled samples and incorporate high-confidence examples into the training process through an iterative approach. These methods can be divided into two principal subcategories: *Self-training methods* and *multi-view learning methods*.

3.1.1. Self-training methods

Self-training methods operate through the bootstrapping mechanism. Initially, a prediction function f_θ with parameters θ is trained using available labeled data samples $x \in X_L$. Subsequently, this trained model generates predictions for unlabeled data samples $x \in X_U$. A confidence threshold τ is established, and sample-label pairs $(x, \text{argmax} f_\theta(x))$ whose prediction confidence exceeds τ are added to the labeled dataset X_L. This augmented labeled dataset is then used to retrain the prediction function, creating an iterative cycle that continues until the model can no longer make sufficiently confident predictions on remaining unlabeled data.

Entropy minimization (Grandvalet and Bengio, 2004) represents a foundational approach in this category, regularizing models based on the low-density assumption by encouraging low-entropy predictions for unlabeled data. Building on this concept, **Pseudo-label** (Lee et al., 2013) provides a straightforward yet effective self-training mechanism that applies entropy minimization principles in prediction space. While labeled samples undergo supervised training, unlabeled data receive labels corresponding to the model's most confident predictions. In medical image analysis, Pseudo-label has been widely employed as an auxiliary component to enhance model performance across various applications (Fan et al., 2020; Zhang et al., 2021b, 2022b; Chaitanya et al., 2023).

A significant challenge with proxy labels is their inherent noise and potential deviation from ground truth. To address this limitation, researchers have developed quality assurance mechanisms including uncertainty-aware confidence evaluation (Wang et al., 2021a), conditional random field-based proxy label refinement (Bai et al., 2017), and adversarial training-based methods (Zhou et al., 2019a). These approaches help ensure that proxy labels provide reliable supervisory signals during training. Pseudo-label has also been used in MIA to refine a given annotation with the assistance of unlabeled data. Qu et al. (2020) introduce pseudo-label into nuclei segmentation and design an iterative learning algorithm to refine the background of weakly labeled images where only nuclei are annotated, leaving large areas ignored. Similar ideas can also be seen in Nie et al. (2018).

3.1.2. Multi-view learning methods

Multi-view learning methods assume that each sample has two or multiple complementary views and features of the same sample extracted with different views are supposed to be consistent. Therefore, the key idea of multi-view learning methods is to train the model with multiple views of the sample or train multiple learners and minimize the disagreement between them, thus learning the underlying features of the data from multiple aspects. **Co-training** is a method that falls into this category. It assumes that data sample x can be represented by two views, $\mathbf{v}_1(x)$ and $\mathbf{v}_2(x)$, and each of them are capable of solely training a good learner, respectively. Consequently, the two learners are set to make predictions of each view's unlabeled data, and iteratively choose the candidates with the highest confidence for the other model (Yang et al., 2021). Another variation of multi-view learning methods is Tri-training (Zhou and Li, 2005), which is proposed to tackle the lack of multi-view data and mistaken labels of unlabeled data produced by self-training methods. Tri-training aims to learn three models from three different training sets obtained with bootstrap sampling. A deep learning version of Tri-training, i.e. Tri-Net, has been further proposed in Chen et al. (2018).

Co-training, or deep co-training, is dominant in multi-view learning in MIA, with a steady flow of publications (Zhao et al., 2019; Zhou et al., 2019b; Xia et al., 2020; Wang et al., 2021b; Fang and Li, 2020; Hou et al., 2022b; Zeng et al., 2023). To conduct whole brain segmentation, Zhao et al. (2019) implements co-training by obtaining different views of data with data augmentation. A similar idea can be seen for 3D medical image segmentation in Xia et al. (2020) and Zhou et al. (2019b). These two works both utilize co-training by learning individual models from different views of 3D volumes such as the sagittal, coronal, and axial planes. Further works have been proposed to refine co-training. To produce reliable and confident predictions, Wang et al. (2021b) develops a self-paced learning strategy for co-training, forcing the network

Table 2
Surveyed representative semi-supervised learning-based studies for the *insufficient label* scenario, organized by downstream task type with corresponding target organs, consistency enforcement method, datasets, and publication venues.

	Reference	Organ	Consistency Enforcement Design	Dataset	Publication
Classification	Madani et al. (2018)	Lung	Semi-superivsed GAN	NIH PLCO; NIH Chest X-Ray	ISBI 2018
	Diaz-Pinto et al. (2019)	Retina	Semi-supervised DCGAN	ORIGA-light; DRISHTI-GS; RIM-ONE; HRF; DRD; sjchoi86-HRF; ACRIMA; DRIVE; Messidor	TMI 2019
	Shi et al. (2020)	Lung; Breast	Graph Temporal Ensembling	TCGA-Lung; TCGA-Breast	MIA 2020
	Yu et al. (2021a)	Colon	Mean Teacher	Private Dataset: 13,111 Images	Nat. Commun. 2021
	Wang et al. (2021c)	Breast; Retina	Virtual Adversarial Training + Self-training	RetinalOCT; Private Dataset: 39,904 Images	MIA 2021
	Liu et al. (2022)	Lung; Skin	Anti-curriculum Self-training	ChestX-Ray14; ISIC 2018	CVPR 2022
	Zhang et al. (2022b)	Spinal cord	Consistency Regularization + Pseudo-labeling + Active Learning	Private Dataset: 7,295 Images;	CVPR 2022
	Gao et al. (2023)	Multi-Organ	Dual-task Consistency	TGCA-RCC; TCGA-BR; TGCA-LU	MIA 2023
	Zeng et al. (2023)	Colon; Skin; Chest	Self-training + Feature Adversarial Training	NCT-CRC-HE; ISIC 2018; Chest X-Ray14	CVPR 2023
	Xie et al. (2023)	Retina	Semi-supervised GAN	iChallenge; ODIR	TMI 2023
	Yang et al. (2023)	Multi-Organ	Self-training	KC Dataset; ISIC 2018; RSNA Dataset	TMI 2023
	Huang et al. (2024)	Multi-Organ	Pseudo Label; Mean Teacher; MixMatch; FixMatch; FlexMatch; CoMatch	TissueMNIST; PathMNIST; TMED-2; AIROGS	CVPR 2024
	Berenguer et al. (2024)	Lung; Skin; Knee	Consistency Regularization; Graph Attention	NIH-14 Chest X-ray; ISIC 2018; COV19-CT; Knee MRNet	MIA 2024
Segmentation	Bai et al. (2017)	Heart	CRF-based Self-training	Private Dataset: 8050 Images	MICCAI 2017
	Li et al. (2018)	Skin	Π-model	ISIC 2017	ArXiv 2018
	Nie et al. (2018)	Prostate	Self-training	Private Dataset: 70 Images	MICCAI 2018
	Yu et al. (2019)	Heart	Uncertainty-aware Mean Teacher	ASG	MICCAI 2019
	Zhou et al. (2019b)	Multi-Organ	Multi-planar Co-training	Private Dataset: 310 Volumes	WACV 2019
	Li et al. (2020e)	Liver; Retina; Skin	Transformation-consistent Mean Teacher	ISIC 2017; REFUGE; LiTS	TNNLS 2020
	Liu et al. (2020)	Skin; Lung	Mean Teacher + Sample Relation Consistency	ISIC 2018; ChestX-ray14	TMI 2020
	Li et al. (2020b)	Heart	Shape-aware Consistency Regularization	ASG	MICCAI 2020
	Fan et al. (2020)	Lung	Attention Self-training	IS-COVID	TMI 2020
	Chaitanya et al. (2021)	Heart; Prostate; Pancreas	Semi-supervised GAN + Deformation and Additive Intensity Field	ACDC; DECATHLON	MIA 2021
	Luo et al. (2021b)	Nasopharynx	Uncertainty Rectified Pyramid Consistency	Private Dataset: 258 MR Images	MICCAI 2021
	Luo et al. (2021a)	Heart;	Dual-task Consistency	ASG; NIH PCT	AAAI 2021
	Li et al. (2021c)	Lung; Skin; Liver	StyleGAN2	ChestX-ray14; JSRT Database; ISIC 2018; LiTS; CHAOS	CVPR 2021
	You et al. (2022)	Heart; Pancreas	Mean Teacher + Contrastive Learning	ASG; NIH PCT	TMI 2022
	Wang et al. (2022b)	Heart; Prostate	Mean Teacher + Contrastive Learning	ACDC; ProMRI	MIA 2022
	Wu et al. (2022)	Heart; Pancreas	Uncertainty-based Mutual Consistency	ASG; NIH PCT; ACDC	MIA 2022
	Luo et al. (2022)	Heart	Co-training Variant	ACDC	MIDL 2022
	Lei et al. (2022)	Liver; Skin	Adversarial Consistency + Dynamic Convolution Network	LiTS; ISIC 2018	TMI 2022
	Meng et al. (2022)	Multi-Organ	Consistency Regularization + Adaptive Graph Neural Network	SEG (Combined by Refuge; Drishti-GS; ORIGA; RIGA;	TMI 2022
	Shi et al. (2023c)	Multi-Organ	Consistency Regularization + Teacher-student Model	CVC; ETIS-Larib Polyp; Private Dataset: 1,100 images	MIA 2023

Table 2
Continue

	Reference	Organ	Consistency Enforcement Design	Dataset	Publication
	Xu et al. (2023)	Multi-Organ	Mean Teacher	Left Atrium (LA) Dataset; BraTS 2019	MIA 2023
	Bai et al. (2023b)	Multi-Organ	Mean Teacher	Left Atrium (LA) Dataset; Pancreas-NIH; ACDC	CVPR 2023
	Miao et al. (2023)	Multi-Organ	Causality Co-training	ACDC; Pancreas-CT; BraTS 2019	ICCV 2023
	Chaitanya et al. (2023)	Heart; Prostate	Self-training	ACDC; MICCAI 2019; MMWHS	MIA 2023
	Wang et al. (2023b)	Heart; Pancreas	Co-training	Left Atrial (LA) Dataset; NIH-Pancreas	CVPR 2023
	Basak and Yin (2023)	Heart; Kidney; Gland	Contrastive Self-training	ACDC; KiTS19; CRAG	CVPR 2023
	Zhang et al. (2023a)	Heart; Pancreas	Co-training	Left Atrial (LA) Dataset; NIH-Pancreas	MIA 2023
	Chen et al. (2022a)	Heart; Brain	Task-specific Consistency Regularization	MICCAI 2018; DECATHLON	TMI 2022
	Bashir et al. (2024)	Multi-Organ	Consistency Regularization	MoNuSeg; BCSS	MIA 2024
	Su et al. (2024)	Heart; Stomach; Brain	Mutual Learning; Pseudo Label	Left Atrium (LA) Dataset; Pancreas-CT Dataset; Brats-2019	MIA 2024
	Chi et al. (2024)	Heart; Prostate	Consistency Regularization	ACDC; PROMISE12	CVPR 2024
	Ma et al. (2024b)	Eye; Prostate; Heart	Mean Teacher with Domain Knowledge Transfer	Fundus dataset; Prostate dataset; M&Ms dataset	CVPR 2024
	Miao et al. (2024)	Breast; Heart	Prompting Segment Anything Model	BUSI; ACDC	MICCAI 2024
Detection	Wang et al. (2020a)	Lung	MixMatch + Focal Loss	LUNA; NLST	CVPR 2020
	Zhou et al. (2021b)	Multi-Organ	Teacher-student Model + Adaptive Consistency Loss	DSB; DeepLesion	MIA 2021
	Zhang et al. (2024)	Lung; Eye	GAN with Spatial Attention	RSNA Dataset; VinDr-CXR	MICCAI 2024

to start with the easier-to-segment regions and transition to the difficult areas gradually. Rather than discarding samples with low-quality pseudo-labels, Zeng et al. (2023) introduces a novel regularization approach, which focuses on extracting discriminative information from such samples by injecting adversarial noise at the feature level, thereby smoothing the decision boundary. Meanwhile, to avoid the errors of different model components accumulating and causing deviation, Fang and Li (2020) develop an end-to-end model called difference minimization network for medical image segmentation by conducting co-training with an encoder shared by two decoders.

3.2. Generative modeling methods

While proxy-labeling methods directly assign labels to unlabeled data, **generative modeling methods** propagate supervision from labeled to unlabeled data by assuming that both labeled and unlabeled data are sampled from a shared latent distribution. By learning this underlying distribution with the help of unlabeled data, these methods enable the model to transfer information across the entire dataset. The learned latent representation is then combined with supervised information from labeled examples to further improve performance. For readers interested in the theoretical foundations and mathematical details of generative modeling, we provide an overview in Appendix B.

Variational autoencoders (VAEs) were among the first successful approaches for utilizing unlabeled data in generative modeling. Based on Bayesian inference theory (Kingma and Welling, 2013), VAEs encode data into latent variables and reconstruct inputs by maximizing the variational lower bound. In medical image analysis, VAEs primarily learn feature similarities from large unlabeled datasets, creating well-constrained latent spaces that reduce dependence on labeled data (Sedai et al., 2017; Wang and Lukasiewicz, 2022). Sedai et al. (2017) proposed a dual-VAE framework for semi-supervised optic cup segmentation in retinal images, where one VAE learns data distribution from unlabeled data and transfers this knowledge to a second VAE performing segmentation with labeled data. In this study, 600 labeled and 11,400 unlabeled retinal fundus images are included. Wang and Lukasiewicz (2022) adapted VAEs for 3D medical image segmentation by replacing the conventional mean vector and variance vector with a mean vector and covariance matrix, accounting for correlations between different slices of an input volume. However, VAEs often suffer from blurry reconstructions due to the KL divergence regularization term (Di Folco et al., 2024), and their posterior approximation can limit the expressiveness of learned representations (Li et al., 2022a; Di Folco et al., 2024), particularly for complex medical image structures with fine anatomical details.

Generative adversarial networks (GANs) subsequently revolutionized generative modeling by effectively leveraging both labeled and unlabeled data through a two-player minimax game between a generator G and discriminator D (Goodfellow et al., 2014). In medical image analysis, several semi-supervised approaches incorporate unlabeled data during adversarial training. Chaitanya et al. (2021) and Hou et al. (2022a) utilize unlabeled samples to introduce greater variation in shape and intensity, enhancing model robustness. Zhou et al. (2019a) generate pseudo lesion masks for unlabeled data with quality facilitated by the discriminator. The study utilizes 54 pixel-level annotated images for supervised lesion segmentation and around 35k image-level annotated (but pixel-level unlabeled) images as unlabeled data for semi-supervised segmentation and disease grading. Other researchers modify the discriminator's objective beyond binary classification: Odena (2016) extend it to predict K classes plus an additional real/fake class, allowing unlabeled data to contribute to multi-class discrimination. This architecture has been successfully applied to retinal image synthesis (Kamran et al., 2021; Diaz-Pinto et al., 2019; Xie et al., 2023), glaucoma assessment (Diaz-Pinto et al., 2019), chest X-ray classification (Madani et al., 2018), and other medical imaging tasks (Hou et al., 2022a). Despite their effectiveness, GANs face significant challenges, including training

instability (Wiatrak et al., 2019), mode collapse where the generator fails to capture the full data distribution (Barsha and Eberle, 2025), and difficulty in convergence (Mescheder et al., 2018).

Diffusion models are a class of generative models that learn to synthesize data by reversing a gradual noising process. Inspired by nonequilibrium thermodynamics, they iteratively denoise samples from a prior distribution (typically Gaussian noise) to generate realistic outputs (Ho et al., 2020; Yang et al., 2022a; Rombach et al., 2022). Unlike adversarial approaches, diffusion models optimize a stable likelihood-based objective, offering improved convergence and sample diversity. More recently, they have emerged as the dominant generative paradigm, largely supplanting GANs in computer vision applications. This shift has been driven by several key advantages: superior training stability without adversarial dynamics, elimination of mode collapse issues that plagued GANs, and enhanced controllability through classifier-free guidance. The decisive turning point came when Dhariwal and Nichol (2021) demonstrated that diffusion models could achieve superior FID scores compared to GANs while maintaining better sample diversity. In the Semi-SL domain, diffusion models offer particular promise through their iterative denoising processes, enabling robust uncertainty quantification and enhanced sample quality in label-scarce scenarios. Recent applications demonstrate their effectiveness in midline shift (MLS) quantification (Gong et al., 2023) and medical image segmentation (Liu et al., 2024b), with their ability to model complex anatomical structures, making them especially valuable for medical imaging tasks where training stability and sample quality are critical. For instance, Gong et al. (2023) utilized diffusion models to formulate the MLS measurement task as a deformation estimation problem and solve it using 124 labeled and 2669 unlabeled MLS CT volumes.

3.3. Regularization-based methods

In contrast to the explicit labeling approach of proxy methods and the distribution modeling of generative techniques, regularization-based methods enforce consistency through direct constraints on the model's behavior by assuming that the perturbation of data points does not change the prediction of the model, allowing these consistency constraints to be applied to unlabeled data while leveraging available labels for supervision.

Π-model (Sajjadi et al., 2016) effectively implements consistency regularization by using a shared encoder to process differently augmented views of the same input and enforcing consistent predictions across these views, while incorporating label information to improve classifier performance. Inspired by this line of work, Li et al. (2018) introduced a transformation-consistent self-ensembling model for semi-supervised skin lesion segmentation, which adapts consistency regularization to the segmentation setting by enforcing equivariance under transformations. Remarkably, their approach achieved state-of-the-art performance on the ISIC 2017 challenge using only 300 labeled images plus 1700 unlabeled ones, surpassing previous fully-supervised methods trained on 2000 labeled images. Similar consistency-based approaches appear in Bortsova et al. (2019), who enforce prediction consistency across transformations for chest X-ray segmentation, and Meng et al. (2023), who employ graph convolution networks to maintain regional and marginal consistency for semi-supervised optic disc and cup segmentation.

Temporal ensembling (Laine and Aila, 2016) improves the Π-model's prediction stability by incorporating exponential moving averages, an approach widely adopted in medical image analysis (Cao et al., 2020b; Gyawali et al., 2019; Shi et al., 2020; Luo et al., 2020). For breast mass segmentation, Cao et al. (2020b) integrate uncertainty maps as guidance to ensure prediction reliability. Similarly, Luo et al. (2020) propose uncertainty-aware temporal ensembling for chest X-ray screening with partially labeled data. Gyawali et al. (2019) enhance the method by first using a VAE to extract disentangled latent space representations as stochastic embeddings, improving chest X-ray classification performance. A key characteristic of temporal ensembling is that each training sample's activation is updated only once per epoch.

Mean teacher (Tarvainen and Valpola, 2017) apply exponentially moving average to model parameters rather than network activations, addressing the limitations of temporal ensembling and finding various applications in medical imaging (Li et al., 2020e; Yu et al., 2019; Wang et al., 2020d; Xu et al., 2023; Adiga et al., 2023). Li et al. (2020e) apply this approach to transformation-consistent medical image segmentation. Since teacher models can generate inaccurate targets for unlabeled data, Yu et al. (2019) and Adiga et al. (2023) incorporate uncertainty maps to ensure target reliability. Wang et al. (2020d) further propose a double-uncertainty-weighted method for left atrium and kidney segmentation, extending uncertainty from segmentation to feature level. Xu et al. (2023) focus on selecting productive unsupervised consistency targets through an ambiguity-consensus mean-teacher model that better exploits complementary information from unlabeled data.

3.4. Discussion

Semi-supervised learning addresses the scarcity of labeled data by exploiting large amounts of unlabeled samples and enforcing supervision consistency across the dataset. However, the fundamental assumption that labeled and unlabeled data share similar distributions often fails in medical imaging, where patient populations, imaging protocols, and disease prevalence can vary significantly across institutions, potentially leading to degraded performance when distribution mismatch occurs.

Methodological Considerations. Given these considerations, the choice of Semi-SL strategy should be guided by data characteristics and task requirements. For instance, *proxy-labeling methods* like self-training tend to perform well when high-confidence predictions can be reliably identified; *multi-view learning approaches* appear particularly suited for volumetric data where different perspectives provide complementary information; *generative modeling* shows promise with complex anatomical structures that benefit from learned prior distributions; while *regularization-based methods* often demonstrate robustness across diverse imaging modalities. Despite these apparent strengths, systematic guidelines for method selection remain largely empirical. For example, proxy-labeling approaches suffer from confirmation bias and error accumulation, particularly in medical domains where false positives can have severe consequences. Meanwhile, the claimed robustness of regularization methods has primarily been demonstrated on relatively homogeneous datasets, raising questions about their generalizability to real-world clinical scenarios.

Key Challenges. A persistent challenge in Semi-SL lies in the utilization of noisy or imperfect unlabeled data. *The generation and selection of reliable pseudo labels are critical*, as label noise can easily propagate through the training process and undermine model performance. Moreover, the theoretical understanding of how different Semi-SL techniques interact within hybrid systems remains limited, especially when dealing with heterogeneous data sources (Wang et al., 2021c; Zhang et al., 2022b; Wang et al., 2020a; Gyawali et al., 2020; Hou et al., 2023b; Miao et al., 2023).

Real-World Clinical Impact and Performance Gains. The practical value of Semi-SL is increasingly evident in recent large-scale clinical studies. In colorectal cancer diagnosis across 13 centers, Semi-SL achieved comparable patient-level performance (AUC: 0.974 ± 0.013) to fully supervised learning (AUC: 0.980 ± 0.010) while using only 15% of labeled data (Yu et al., 2021a). Similarly, Co-BioNet showed that Semi-SL with 20% labeled data matched fully supervised performance across CT and MRI modalities (Peiris et al., 2023). *These findings translate to significant practical benefits*: 80–85% reduction in annotation costs, diagnostic accuracy comparable to experienced pathologists, and enhanced robustness across multi-center settings. Such performance gains make Semi-SL particularly valuable for rare diseases or emerging imaging modalities where extensive annotations are prohibitively expensive, ul-

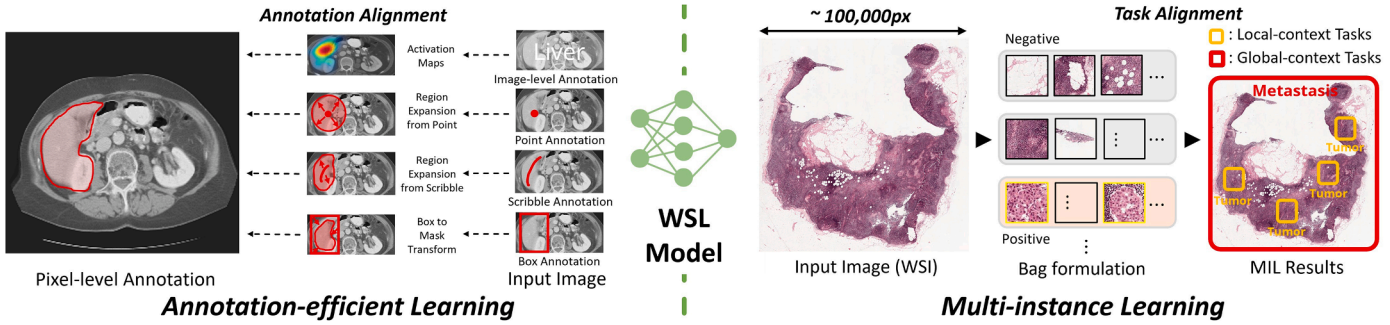


Fig. 5. Overview of weakly-supervised learning paradigm which addresses the *inexact label* scenario. On the annotation side, annotation-efficient learning strategy aims to maximize the utility of limited or partially annotated data while minimizing annotation burden. On the task side, multi-instance learning strategy aims to bridge the label granularity gap by inferring fine-grained predictions from coarse supervision.

timately facilitating broader AI adoption in resource-constrained clinical environments.

4. Inexact label

Obtaining precise, pixel-level annotations in MIA is often prohibitively expensive or time-consuming. In practice, clinicians frequently provide only incomplete or coarse annotations—such as sparse points or image-level labels—resulting in the so-called **inexact label** setting, where the available supervision does not match the granularity required for the target task. **Weakly supervised learning (WSL)** provides a general framework for addressing the challenges posed by inexact labels. The core issue in this context is label granularity alignment, which refers to bridging the gap between coarse-grained annotations and the fine-grained predictions demanded by clinical applications.

Granularity alignment in WSL can be approached from two complementary perspectives. On the annotation side, **annotation-efficient learning** scheme aims to maximize the utility of limited or partially annotated data while minimizing annotation burden. These methods leverage weak or coarse labels, sometimes in combination with unlabeled data, to enhance model performance with reduced annotation cost. On the task side, the challenge is to infer fine-grained information, such as pixel-level segmentation or lesion localization, from supervision that is only available at a coarser level, such as the whole image. **Multiple instance learning (MIL)** provides a representative learning scheme for this scenario by associating global image-level labels with collections of instances, such as patches or regions, and enabling the model to identify relevant instances based on weak supervision.

In the following, we elaborate on the principles and representative applications of these strategies, with a focus on how label granularity alignment enables robust learning from inexact labels in MIA.

4.1. Annotation-efficient learning

To address label granularity alignment from the annotation perspective, **annotation-efficient learning** leverages deep learning techniques with partially labeled data to enhance labeling efficiency for dense predictions. A strategic approach to boost annotation efficiency is to utilize markings other than complete dense annotations. While there may be some overlap with other methodologies, annotation-efficient learning methods are specifically tailored to harness the unique attributes of various annotation forms to improve efficiency and bridge the granularity disparity between annotation and prediction. In this section, we review representative annotation-efficient learning strategies that address label granularity alignment through a coarse-to-fine approach, focusing on techniques related to **image-level**, **point**, **scribble**, and **box** annotations. Appendix Tables 3, C.1 provides an overview of notable publications in this domain.

4.1.1. Image-level annotation

Image-level annotation, or tag annotation, is a concise text or binary label assigned to each image, stands out as the most efficient form of annotation. Many image-level annotation-based methods draw inspiration from the concept of class activation mapping (CAM) (Zhou et al., 2016). Several approaches utilize CAM to generate object localization proposals or achieve pixel-wise segmentation of entire objects. For **detection** tasks, Hwang and Kim (2016) introduced a dual-branch network that concurrently optimizes classification and lesion detection. This approach involves supervising the CAM-based lesion detection network solely with image-level annotations. The two branches are guided in tandem through a weight-sharing technique, employing a weighting parameter to regulate the learning focus between classification and detection tasks. In the realm of lesion detection, Dubost et al. (2020) proposed a weakly supervised regression network, validated on both 2D and 3D medical images. For the **segmentation** task, Li et al. (2022b) proposed a breast tumor segmentation method with only image-level annotations based on CAM and deep-level set (CAM-DLS). It integrates domain-specific anatomical information from breast ultrasound to reduce the search space for breast tumor segmentation. Similarly, Li et al. (2024b) proposed a self-guided multiple information aggregation network for skin lesion segmentation using multiple spatial perceptron solely using classification information as guidance to discriminate the key classification features of lesion areas. Meanwhile, Chen et al. (2022c) proposes a causal CAM method for organ segmentation, which is based on the idea of causal inference with a category-causality chain and an anatomy-causality chain. In addition, several studies (Lin et al., 2019, 2021) demonstrate that bridging the classification task and dense prediction task (e.g., detection and segmentation) via CAM-based methods is beneficial for both tasks. Compared to natural images, medical images typically exhibit low contrast, limited texture, and are subject to varying acquisition protocols (Zhang et al., 2021a), which makes directly applying CAM-based methods less effective. Fortunately, incorporating the clinical priors (e.g., objects' size Früh et al., 2021 and lesion priors Hou et al., 2023a) into the weakly supervised task is promising to improve the performance.

4.1.2. Point annotation

Point annotation refers to the annotation of a single point of an object. Several studies (Roth et al., 2021; Dorent et al., 2021; Khan et al., 2019) focus on using extreme points as the annotation to perform pixel-level segmentation. These methods typically consist of three steps: 1) extreme point selection; 2) initial segmentation with a random walk algorithm; 3) training of the segmentation model with the initial segmentation results. The last two steps can be iterated until the segmentation results are stable. However, since extreme points must lie on the extremities of the target, these methods inherently require the annotators to locate the boundary of the objects, which is still laborious in practice. In contrast, other studies (Yoo et al., 2019; Zhao and Yin, 2020; Qu et al.,

2020, 2019; Tian et al., 2020; Belharbi et al., 2021; Lin et al., 2023; Valvano et al., 2021) use center point annotation, which merely requires indicating the object's location without delineating its precise boundaries, to perform pixel-level segmentation for the task of cell/nuclear segmentation. These methods typically adopt the Voronoi diagram (Kise et al., 1998) and clustering algorithms to perform coarse segmentation. Then different methods are used to refine the segmentation results, such as iterative optimization (Qu et al., 2019, 2020), self-training (Zhao and Yin, 2020), and co-training (Lin et al., 2023).

Compared with full annotation, point annotation can reduce the annotation time by around 80% (Qu et al., 2020). However, some issues have not been addressed. First, existing methods typically derived pseudo labels from the point annotation, which are based on strong constraints or assumptions (e.g., Voronoi) from the data, restricting the generalization of these methods to other datasets (Lin et al., 2023). Second, due to the lack of explicit boundary supervision, there is a non-negligible performance gap between the weakly supervised methods with points and the fully supervised methods.

4.1.3. Scribble annotation

Scribble annotation, a set of scribbles drawn on an image by the annotators, has been recognized as a user-friendly alternative to bounding box annotation (Tajbakhsh et al., 2020). Compared with point annotation, it provides the rough shape and size information of the objects, which is promising to improve the segmentation performance, especially for objects with complex shapes. Wang et al. (2018a) propose a self-training framework with differences in model predictions and user-provided scribbles. Can et al. (2018) develop a random walk algorithm that incrementally performs region growing method around the scribble ground truth, while Lee and Jeong (2020) introduce Scribble2Label, a method that integrates the supervision signals from both scribble annotations and pseudo labels with the exponential moving average. Furthermore, Dorent et al. (2020) extend the Scribble-Pixel method to the domain adaptation scenario, where a new formulation of domain adaptation is proposed based on CRF and co-segmentation with the scribble annotation. Zhang and Zhuang (2022) adopt mix augmentation and cycle consistency for the Scribble-Pixel method, demonstrating the improvement of both weakly and fully supervised segmentation methods. Zhou et al. (2023a) proposed a scribble-supervised approach that combines superpixel-guided scribble walking with class-wise contrastive regularization to augment the structural priors into the weak annotations. Building upon the concept of structural priors, anatomical priors can also serve as a complementary family of weak annotations alongside explicit geometric markings. By directly incorporating domain knowledge, such as expected object volume or boundary characteristics, into the learning process, these methods reduce the reliance on dense manual delineation. A prominent example is the boundary loss introduced by Kervadec et al. (2021) and its variants, which effectively leverage anatomical constraints to guide highly unbalanced segmentation tasks.

4.1.4. Box annotation

Box annotation encloses the segmented region within a rectangle, and various studies have focused on this Box-Pixel scenario. Rajchl et al. (2016) employ a densely-connected random field (DCRF) with an iterative optimization method for MRI segmentation. Wang and Xia (2021a,b) adopt smooth maximum approximation based on the bounding box tightness prior (Hsu et al., 2019), that is, an object instance should touch all four sides of its bounding box. Thus, a vertical or horizontal crossing line within a box yields a positive bag because it covers at least one foreground pixel. Studies (Wang and Xia, 2021b) demonstrate that the Box-Pixel method yields promising performance, being only 1–2% inferior to the fully supervised methods. In a recent work, Wong et al. (2025) proposed ScribblePrompt, which is flexible to different annotation styles, including bounding boxes, points, and scribbles. Overall, box annotation offers a strong balance between efficiency and accuracy, but may struggle with irregular or overlapping shapes.

4.2. Multi-instance learning

To address label granularity alignment from the task perspective, **multi-instance learning (MIL)** enables deep learning models to make fine-grained predictions using only coarse, image-level supervision. MIL organizes each image or specimen as a *bag* of multiple *instances* (e.g., patches, regions, or cells), with only the bag-level label observed during training. Under the standard MIL assumption, a bag is positive if at least one of its instances is positive. This framework supports weak supervision and allows for fine-grained analysis—such as localizing disease regions—even when only image-level labels are available.

Distinct from approaches that primarily calibrate the form of supervision, MIL is specifically tailored to bridge the granularity gap between global supervision and dense prediction requirements. By modeling the instance collections within each image, MIL enables inference of instance-level relevance from bag labels, thereby supporting fine-grained pattern recognition using only bag-level supervision.

In this survey, we categorize MIL approaches by their prediction target: **local-context MIL** focuses on detecting and classifying patterns at the instance level, while **global-context MIL** aims for bag-level predictions. Representative works are summarized in Table 4.

4.2.1. Local-context MIL

Within the MIL paradigm, local-context MIL approaches focus on the fine-grained identification and localization of specific disease patterns within medical images. These methods are designed to infer instance-level labels for individual patches, enabling precise delineation of pathological regions. By determining the status of each constituent patch, local-context MIL naturally encompasses global-context capabilities—the image-level classification emerges as a function of the detected local annotations. This hierarchical inference process allows clinicians not only to receive diagnostic outcomes but also to visualize the spatial distribution of disease manifestations, enhancing interpretability and clinical utility. The ability to simultaneously perform localization and diagnosis makes local-context MIL particularly valuable in applications requiring both detection sensitivity and anatomical precision.

Schwab et al. (2020) apply MIL to chest X-ray classification and localization by processing image patches through a CNN to assess their probability of containing critical findings. While traditional MIL approaches integrate patch-level predictions using max-pooling or average-pooling, Couture et al. (2018) enhance this by implementing a quantile function for pooling, providing better characterization of sample heterogeneity. The field has evolved toward learning-based aggregation methods, notably the attention-based MIL developed by Ilse et al. (2018), which better captures and interprets regions of interest. Wang et al. (2021e) extend this concept with an inductive attention guidance network for pancreatic ductal adenocarcinoma, where the attention mechanism bridges global classification and local segmentation by identifying relevant regions.

Other intriguing improvements in local-context MIL are springing up as well. Researchers have tried many different ways to facilitate instance prediction (Dov et al., 2021; Manivannan et al., 2017; Jia et al., 2017; Xu et al., 2019). Dov et al. (2021) address cytopathology challenges—where informative instances are sparse and exhibit varied abnormalities—by implementing maximum likelihood estimation to simultaneously predict bag-level labels, diagnostic scores, and instance-level labels. For retinal nerve fiber layer classification, Manivannan et al. (2017) overcome strong intra-class variation by mapping instances into a discriminative subspace that enhances feature disentanglement. Jia et al. (2017) incorporate multi-scale image features to extract more latent information from histopathology images. Addressing the limitation of image-level-only labels in MIL, Xu et al. (2019) develop an automatic

Table 3
Surveyed representative annotation-efficient learning studies for the *inexact label* scenario, organized by downstream task type, with corresponding target organs, label granularity alignment strategy, datasets, and publication venues.

	Reference	Organ	Granularity Alignment Strategy	Dataset	Publication
Detection	Hwang and Kim (2016)	Lung; Breast	CAM + Self-Transfer Learning	Private Dataset: 11K X-rays; DDSM; MIAS	MICCAI 2016
	Gondal et al. (2017)	Eye	CAM	DRD; DiaretDB1	ICIP 2017
	Wang et al. (2018b)	Eye	Expectation-Maximization	DRD; Messidor	TMI 2018
	Wang et al. (2020b)	Lung	CAM + Unsupervised Segmentation	Private Dataset: 540 CT Images	TMI 2020
	Shen et al. (2021)	Breast	Globally-aware Multiple Instance Classifier	NYUBCS; CBIS-DDSM	MIA 2021
	Liu et al. (2023b)	Eye	contrastive learning; knowledge distillation	Private: 7,000 OCT	CBM 2023
	Wang et al. (2024d)	Breast	Two-Stage Detection and Diagnosis Network	BUSI; BLUI	ISBI 2024
	Wang and Yin (2021)	Cell	Multi-task Learning	VGG Cell; MBM; ADI; DCC	MICCAI 2021
Segmentation	Gong et al. (2024)	Multi-organ	Segment Anything Model (SAM)	KiTS21; MSD-Pancreas; LiTS17; MSD-Colon	MIA 2024
	Cui et al. (2024)	Multi-organ	All–in–SAM	MoNuSeg	ACCC 2023
	Hu et al. (2023b)	Brain; Kidney	Conditional Diffusion Models	BraTS; CHAOS	MICCAI 2023
	Du et al. (2024)	Brain; Skin; Lung	CAM	BraTs; ISIC; COVID-19	CBME 2024
	Shen et al. (2024)	Abdominal; Brain	SAM; ViT	AMOS; TotalSegmentator; BraTS	MICCAI 2024
	Zhu et al. (2023)	Prostate	Collaborative Annotation Calibration	PROMISE12; CVC-ClinicDB; NCI-ISBI	TMI 2023
	Chen et al. (2024a)	Abdominal; Prostate	Modality-agnostic SAM Adaptation	BTCV; NIC-ISBI; MSDPancreas	MIA 2024
	Lei et al. (2023)	Abdominal	Propagation-Reconstruction Network	TCIA; StructSeg19; CHAOS; BTCV	TMI 2023
	Wang et al. (2024a)	Brain	SAM; Mixture of Experts (MoE)	SPPIN	MICCAI 2024
	Wong et al. (2025)	Multi-organ	Prompt Learning	77 datasets	ECCV 2024
	Li et al. (2024b)	Skin	Self-Guided Multiple Information Aggregation Network	HAM10000; PH2	CBME 2024
	Yue et al. (2024)	Surgical Instruments	Contrastive Prototype Learning; SAM	EndoVis2018; EndoVis2017	AAAI 2024
	Zhou et al. (2023a)	Cardiac	Superpixel; Contrastive Regularization	ACDC	MICCAI 2023
	Ying et al. (2023)	Uterus	Exponential Geodesic Distance Loss	Private: 135 MRI	CBME 2023
	Nie et al. (2023)	Brain; Liver	Weakly Interactive Mixed Learning	BraTS; LiTS	JBHI 2023
	Kervadec et al. (2021)	Brain	Anatomical Prior; Boundary Loss	ISLES; WMH	MIA 2021
	Du et al. (2023)	Liver; Kidney	Geometric Prior; Point Cloud	LiTS; KiTS	TMI 2023
	Nguyen et al. (2019)	Eye	CAM + CRF + Active Shape Model	Private Dataset: 40 MRI Images	ICMIDL 2019
	Chen et al. (2022c)	prostate; Cardiac;	Causal Inference; CAM	ACDC; ProMRI; CHAOS	CVPR 2022
	Khan et al. (2019)	Multi-organ	Confidence Map Supervision	SegTHOR	MICCAI 2019
	Zhao and Yin (2020)	Cell	Self-/Co-/Hybrid-Training	PHC; Phase100	TMI 2020
	Dorent et al. (2021)	Brain	CNN + CRF	Vestibular-Schwannoma-SEG	MICCAI 2021
	Guo et al. (2023)	Multi-organ	Superpixel; Confident learning	MoNuSeg	MIA 2023
	Xia et al. (2023)	Multi-organ	Multi-task	MoNuSeg	TIM 2023
	Wang et al. (2018a)	Body	Image-Specific Fine-Tuning	Private Dataset: 18 MRI Images; BRATS	TMI 2018
	Lee and Jeong (2020)	Cell	Exponential Moving Average	MoNuSeg	MICCAI 2020
	Zhang and Zhuang (2022)	Heart	Mixup + Consistency	ACDC; MSCMRseg	CVPR 2022

Table 3
Continue

Reference	Organ	Granularity Alignment Strategy	Dataset	Publication
Rajchl et al. (2016)	Brain; Lung	DenseCRF	Private Dataset: 55 MRI Images	TMI 2016
Wang et al. (2022a)	lymph; Lung; Skin	RECIST measurement propagation algorithm;	TCIA;	TMI 2022
Zhu et al. (2023)	Prostate	Annotation calibration; Gradient de-conflicting	PROMISE12	TMI 2023

instance-level label generation method, creating a promising direction for local-context MIL algorithm development.

In parallel, there has been significant progress in related domains such as phenotype categorization (Yao et al., 2020; Hashimoto et al., 2020; Yao et al., 2019) and multi-label classification (Mercan et al., 2017). These investigations have further exemplified the versatility and potential of the MIL algorithm in addressing complex challenges across various subfields.

4.2.2. Global-context MIL

In contrast to the localization focus of local-context approaches, global-context MIL aims to detect whether or not target patterns exist. For example, for the COVID-19 screening problem, researchers (Li et al., 2021f) have designed MIL algorithms to classify an input sample as severe or not instead of locating every abnormal patch.

To facilitate the prediction of image-level labels (*e.g.* WSI-level label in the context of computational pathology), researchers normally start from one of two aspects, namely instance- and bag-level. Most existing MIL algorithms (Tomita et al., 2019; Hashimoto et al., 2020; Naik et al., 2020; Lu et al., 2021a) are based on the basic assumption that instances of the same bag are independent and identically distributed. Consequently, the correlations among instances are neglected, which is not realistic. Several subsequent works have taken the correlation among instances or tissues into consideration (Shao et al., 2021; Wang et al., 2022c, 2019; Raju et al., 2020; Han et al., 2020). Shao et al. (2021) introduce Vision Transformer (ViT) into MIL for gigapixel WSIs due to its great advantage in capturing the long-distance information and correlation among instances in a sequence. Meanwhile, to conduct precise lymph node metastasis prediction, Wang et al. (2022c) not only incorporate a pruned Transformer into MIL but also develop a knowledge distillation mechanism based on other similar datasets, such as a papillary thyroid carcinoma dataset, effectively avoiding the overfitting problem caused by the insufficient number of samples in the original dataset. Similarly, Raju et al. (2020) design a graph attention MIL algorithm for colorectal cancer staging, which utilizes different tissues as nodes to construct graphs for instance relation learning. Further, in order to utilize the multi-resolution characteristics of WSIs, Shi et al. (2023b) consider WSIs as multi-scale graphs and utilize attention mechanism to integrate their information for primary tumor stage prediction. Similar idea can be found in Yan et al. (2023), Shi et al. (2023a), Xiang et al. (2023). Besides, Liu et al. (2024a) firstly propose an integration of GAN with MIL mechanism for robust and interpretable WSI survival analysis by more accurately estimating target distribution.

For bag-level improvement, recent years have witnessed two feasible approaches, namely, improved pooling methods and pseudo bags. On the one hand, in order to aggregate the instances with the most information, some researchers have developed novel aggregation methods in MIL algorithms instead of the traditional max pooling (Chikontwe et al., 2020; Das et al., 2018; Jin et al., 2024a; Guo et al., 2024b). For example, in Chikontwe et al. (2020), the authors design a pyramid feature aggregation method to directly obtain a bag-level feature vector. On the other hand, however, there is an inherent problem for MIA, especially for histopathology — the number of WSIs (bags) is usually small, while in contrast, one WSI has numerous patches, leading to an imbalance in the number of bags and instances. To address this problem, Zhang et al. (2022a) randomly split the instances of a bag into several smaller bags, called “pseudo bags”, with labels that are consistent with the original bag. A similar idea can also be seen in Li et al. (2021f). Moreover, Jin et al. (2024a) introduce a hierarchical multi-instance learning (HMIL) framework that enhances WSI classification through explicit modeling of the hierarchical relationships between instance-level and bag-level label distributions, creating a more cohesive cross-granularity learning paradigm.

Other improvements in MIL algorithms are also worth mentioning (Su et al., 2022; Tennakoon et al., 2019; Wang et al., 2020c). In Su et al. (2022), a novel sampling method is developed to collect instances with high confidence. This method excludes patches shared among different classes and tends to select the patches that match with the bag-level label. In Tennakoon et al. (2019), the authors utilize the extreme value theory to measure the maximum feature deviations and consequently leverage them to recognize the positive instances, while in Wang et al. (2020c), Wang et al. introduce an uncertainty evaluation mechanism into MIL for the first time, and train a robust classifier based on this mechanism to cope with OCT image classification problem.

4.3. Discussion

Weakly supervised learning addresses the challenges of limited and inexact annotations in MIA by *calibrating both the annotation form and granularity of supervision*. Within this paradigm, **annotation-efficient learning and MIL represent two complementary facets** for coping with weak supervision.

Annotation-Efficient Learning for Annotation Form Calibration. Annotation-efficient learning leverages *diverse annotation types*—points, scribbles, boxes, and image tags—each suited to different object characteristics and annotation budgets. *Points and scribbles are efficient for objects with uniform shape*, while *boxes handle morphological variation*, and *image tags offer the lowest annotation cost* at the expense of spatial detail. However, the cost-performance trade-offs are rarely systematically evaluated across different clinical scenarios, making it difficult to provide evidence-based recommendations for annotation strategy selection.

Multi-Instance Learning (MIL) for Granularity Calibration. Multi-instance learning (MIL) addresses *the calibration of annotation granularity*, particularly for tasks such as gigapixel-sized WSI analysis, where only bag-level labels are available. MIL enables fine-grained, instance-level inference by *modeling bag-instance and instance-instance relationships*, effectively aligning global supervision with the need for precise localization or characterization.

Real-World Clinical Impact and Performance Gains. Weakly supervised learning delivers substantial clinical value through reduced annotation burden while maintaining diagnostic accuracy. *In computational pathology*, CLAM (Lu et al., 2021a) achieves AUC > 0.95 for cancer subtyping using only slide-level labels, representing a **10-100× reduction in annotation time** compared to pixel-level annotation while enabling population-scale cancer screening. *In radiology*, RECIST-Sup (Wang et al., 2022a) leverages routine RECIST measurements to achieve lesion segmentation performance comparable to full supervision with **80-92% less annotation cost**. These approaches address critical deployment bottlenecks: they enable AI development for rare diseases where extensive expert annotations are unavailable, facilitate rapid clin-

Table 4
Surveyed representative multi-instance learning-based studies for the *inexact abel* scenario, organized by downstream task type, with corresponding target organs, label granularity calibration strategy, datasets, and publication venues.

	Reference	Organ	Granularity Calibration Strategy	Dataset	Publication
Classification	Manivannan et al. (2017)	Retina; Breast	Discriminative Subspace Transformation + Margin-based Loss	Messidor; TMA-UCSB; DR Dataset; Private Dataset: 884 Images	TMI 2017
	Ilse et al. (2018)	Breast; Colon	Attention-based MIL	TMA-UCSB; CRCHistoPhenotypes	ICML 2018
	Couture et al. (2018)	Breast	Quantile Function-based MIL	CBCS3	MICCAI 2018
	Liu et al. (2018)	Brain	Landmark-based MIL	ADNI; MIRIAD	MIA 2018
	Campanella et al. (2019)	Prostate; Skin; Lymph	MIL + RNN	Private Dataset: 44,732 Images	Nat. Med. 2019
	Wang et al. (2019)	Breast	Instance Features Recalibration	Private Dataset: 608 Images	MIA 2019
	Yao et al. (2019)	Lung; Brain	Multiple Instance FCN	NLST; TCGA	MICCAI 2019
	Zhao et al. (2020b)	Colon	VAE-GAN Feature Extraction +	TCGA-COAD	CVPR 2020
	Chikontwe et al. (2020)	Colon	Jointly Learning of Instance- and Bag-level Feature	Private Dataset: 366 Images	MICCAI 2020
	Raju et al. (2020)	Colon	Graph Attention MIL	MCO	MICCAI 2020
	Han et al. (2020)	Lung	Automatic Instance Generation	Private Dataset: 460 Examples	TMI 2020
	Yao et al. (2020)	Lung; Colon	Siamese Multi-instance FCN + Attention MIL	NLST; MCO	MIA 2020
	Hashimoto et al. (2020)	Lymph	Domain Adversarial + Multi-scale MIL	Private Dataset: 196 Images	CVPR 2020
	Shao et al. (2021)	Breast; Lung; Kidney	Transformer-based MIL	CAMELYON 2016; TCGA-NSCLC; TCGA-RCC	NeurIPS 2021
	Li et al. (2021b)	Breast; Lung	Dual-stream MIL + Contrastive Learning	CAMELYON 2016; TCGA Lung Cancer	CVPR 2021
	Li et al. (2021f)	Lung	Virtual Bags + Self-SL Location Prediction	Private Dataset: 460 Examples	MIA 2021
	Lu et al. (2021a)	Kidney; Lung; Lymph node	Attention-based MIL + Clustering	TCGA-RCC + Private Dataset: 135 WSIs; CPTAC-NSCLC + Private Dataset: 131 WSIs; CAMELYON 201,617 + Private Dataset: 133 WSIs	Nat. Biomed. Eng. 2021
	Wang et al. (2022c)	Thyroid	Transformer-based MIL + Knowledge Distillation	Private Dataset: 595 Images	TMI 2022
	Zhang et al. (2022a)	Breast; Lung	Double-Tier Feature Distillation MIL	CAMELYON 2016; TCGA-Lung	CVPR 2022
	Schirris et al. (2022)	Breast; Colon	Heterogeneity-aware MIL + Contrastive Learning	TCGA-CRCk; TCGA-BC	MIA 2022
	Su et al. (2022)	Breast; Kidney	Intelligent Sampling Method + Attention MIL	CAMELYON 2016; Private Dataset: 112 Images	MIA 2022
	Zhu et al. (2022)	Breast; Lung; Kidney	Reinforcement Learning + Contrastive Learning + MIL	CAMELYON 2016; TCGA-Lung; TCGA-Kidney	TMI 2022
	Yang et al. (2022b)	Colon; Muscle	Curriculum Learning + MIL	CRCHistoPhenotypes; Private Muscle Dataset: 266 Images	TMI 2022
	Shi et al. (2023b)	Stomach; Bladder	Multi-scale Graph MIL	TCGA-STAD; TCGA-BLCA; Private Stomach Dataset: 574 Images	TMI 2023
	Yan et al. (2023)	Bladder	Hierarchical Deep MIL	TCGA-Bladder	MIA 2023
	Shi et al. (2023a)	Breast; Kidney	Multi-scale Transformer + MIL	BRIGHT; TCGA-BRCA; TCGA-RCC	TMI 2023
	Liu et al. (2024a)	Lung; Breast; Brain	GAN + MIL	NLST; TCGA-BRCA; TCGA-LGG	MIA 2024
	Jin et al. (2024a)	Breast; Prostate; Cervix	Coarse-to-fine Hierarchical Encoding MIL	BRACS; PANDA; Private Cervical Dataset	TMI 2024
	Deng et al. (2024)	Brain; Colon	Multi-scale Attention MIL	TCGA-GBMLGG; Private Colon Dataset	MIA 2024
	Qu et al. (2024)	Cervix	Vision-language Alignment MIL	TCGA-CESC; Private Cervical Dataset	ECCV 2024
	Li et al. (2024a)	Esophagus; Kidney; Lung	Dynamic Graph MIL	TCGA-ESCA; TCGA-Kidney; TCGA-Lung	CVPR 2024
	Guo et al. (2024a)	Multi-Organ	Hierarchical Region Encoding & Pooling	TCGA-BRCA, LUAD, Kidney, STAD; UBC-OCEAN; TUPAC16 *etc.*	MICCAI 2024

Table 4
Continue

	Reference	Organ	Granularity Calibration Strategy	Dataset	Publication
	Yang et al. (2024a)	Multi-Organ	State Space Model-based MIL	TCGA-BLCA, BRCA, Lung, Kidney, STAD; BRACS	MICCAI 2024
	Tang et al. (2024b)	Multi-Organ	Regional Transformer-based MIL	TCGA-BLCA, BRCA, LUAD, LUSC; CAMELYON16	CVPR 2024
Segmentation	Jia et al. (2017)	Colon	Multi-scale MIL + Area Constraint Regularization	Private TMA/Colon Dataset: 60 Images/910 Images	TMI 2017
	Xu et al. (2019)	Breast	Instance-level and Pixel-level Label Generation	CAMELYON 2016	ICCV 2019
	Dov et al. (2021)	Thyroid	Maximum Likelihood Estimation-based MIL	Private Dataset: 908 Images	MIA 2021
Others	Schwab et al. (2020)	Lung	Jointly Classification and Localization	RSNA-Lung; MIMIC-CXR; Private Dataset: 1,003 Images	ISBI 2020
	Wang et al. (2021e)	Pancreas	Jointly Global-level Classification and Local-level Segmentation	Private Dataset: 800 Images	TMI 2021

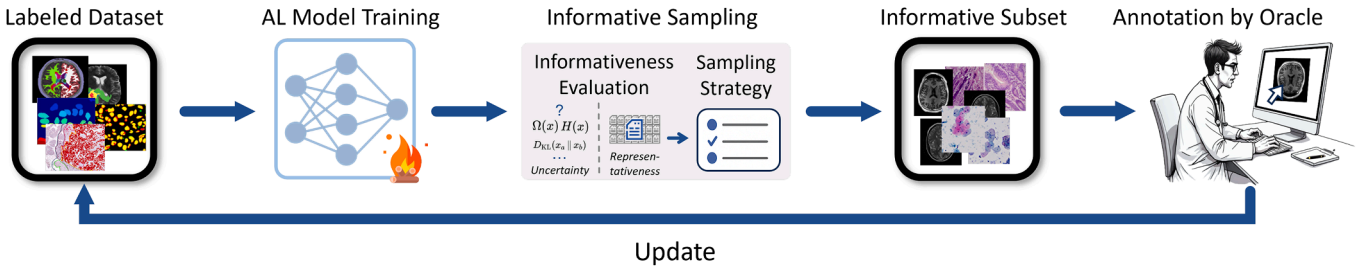


Fig. 6. Overview of the active learning (AL) paradigm, which addresses the *label refinement* scenario. AL iteratively trains a model by employing informative sampling strategies to select and annotate the most valuable unlabeled samples, thereby maximizing model performance under a limited annotation budget.

ical translation, and support resource-constrained healthcare settings where traditional supervised learning would be prohibitively expensive.

5. Label refinement

The *label refinement* scenario arises when existing annotations are limited or suboptimal, often due to annotation costs and expert scarcity in MIA. **Active learning (AL)** has emerged as a primary solution by prioritizing expert annotation on the most informative samples. Unlike prior paradigms that passively exploit available labels, AL actively involves human experts to iteratively improve data quality. Typically, AL starts with a model trained on a small, initially labeled dataset and, in each cycle, ranks the remaining unlabeled or weakly labeled samples by informativeness for expert review. The model is then retrained with the refined labels, and this process repeats until the desired performance or annotation budget is reached.

Informative sampling is central to AL for label refinement, as its effectiveness relies on selecting samples whose refined labels most benefit the model. Accordingly, AL methods are typically classified by their **informativeness evaluation** criteria, such as uncertainty-, representativeness-based, or hybrid criteria, and **sampling strategies** for efficient querying. Table 5 summarizes representative methods. The following subsections detail these criteria and strategies.

5.1. Informativeness evaluation

Informativeness evaluation is fundamental to active learning, guiding the selection of unlabeled samples that are most beneficial for model improvement. Existing methods primarily fall into three categories: uncertainty-based, representativeness-based, and hybrid strategies.

Uncertainty-based methods select samples where the model is least confident, employing measures such as entropy, margin, or Bayesian uncertainty (Yang et al., 2017; Chen et al., 2016). These techniques have evolved to incorporate feature-space similarity and loss prediction networks, especially in structured tasks like segmentation, where metrics such as the Dice coefficient are used to estimate sample informativeness (Zhao et al., 2021b; Wu et al., 2021). Recent advances further enhance robustness by integrating gradient-based uncertainty and out-of-distribution detection (Hu et al., 2023a; Schmidt et al., 2024).

Representativeness-based approaches aim to ensure that selected samples capture the diversity of the data distribution. Clustering, curriculum learning, and latent space distance metrics have been widely adopted for this purpose (Li et al., 2021d; Guo et al., 2018). In medical imaging, where class imbalance and label noise are common, mechanisms such as automatic exclusion of noisy samples and task-specific loss functions are often incorporated (Huang et al., 2019; Liu, 2017).

A growing body of work, which constitutes the third category of hybrid approaches (Budd et al., 2021), seeks to unify these perspectives through hybrid criteria that combine uncertainty, representativeness, and other informativeness signals (Qu et al., 2023; Mahapatra et al., 2024a; Li et al., 2023b). Such methods leverage, for example, Mahalanobis distance, graph-based propagation, or divergence-based measures to improve sample selection efficiency and model robustness (Tang et al., 2023; Atzeni et al., 2022). In summary, informativeness evaluation in AL has thus evolved from simple uncertainty heuristics to hybrid strategies that jointly consider uncertainty, representativeness, and deep feature representations, significantly improving annotation efficiency and model robustness.

5.2. Sampling strategy

Once informativeness has been assessed, the next challenge lies in designing effective sampling strategies that maximize annotation efficiency, particularly when labeling resources are limited.

A straightforward approach is top-k selection, where samples with the highest informativeness scores are prioritized. However, this can

Table 5
Surveyed representative active learning-based studies for the *label refinement* scenario, organized by downstream task type, with corresponding target organs, informative sampling method, datasets, and publication venues.

	Reference (Year)	Organ	Informative Sampling Method	Dataset	Publication[a]
Classification	Gal et al. (2017)	Skin	BALD + KL-divergence	ISIC 2016	ICML 2017
	Wu et al. (2021)	Lung	Loss Prediction Network	CC-CCII Dataset	MIA 2021
	Li et al. (2021d)	Prostate	CurriculumNet + O2U-Net	ISIC 2017; PANDA Dataset	TMI 2021
	Qu et al. (2023)	Colon	Mahalanobis Distance	NCT-CRC-HE-100K Dataset	MICCAI 2023
	Hu et al. (2023a)	Colon; Lung	Gradient + Uncertainty	NCT-CRC-HE-100K Dataset; LC25000	TMI 2023
	Mahapatra et al. (2024a)	Chest	Graph Informativeness + Informative Augmentation	CheXpert; ChestXray14; MedMNIST	MIA 2024
	Schmidt et al. (2024)	Prostate	Uncertainty + Out-of-Distribution Detection	PANDA Dataset	MIA 2024
Segmentation	Yang et al. (2017)	Gland; Lymph	Cosine Similarity + Bootstrapping + FCN	GlaS 2015; Private Dataset: 80 US images	MICCAI 2017
	Li and Yin (2020)	Gland; Brain	Cosine Similarity + Attention FCN	GlaS 2015; iSeg 2017	MICCAI 2020
	Shen et al. (2020)	Breast	Cosine Similarity + Entropy + Dice	Private Dataset: 2,154 IHC images	MICCAI 2020
	Nath et al. (2021)	Brain	Entropy + SVGD Optimization	MSD 2018	TMI 2021
	Ozdemir et al. (2021)	Shoulder	BNN + MMD Divergence	Private Dataset: 36 Volume of MRIs	Knowl.-Based Syst. 2021
	Zhao et al. (2021b)	Hand; Skin	Dice + U-Net	RSNA-Bone; ISIC 2017	JBHI 2021
	Atzeni et al. (2022)	Brain	Dice + FCN	MICCAI 2013 SATA Challenge; Private Dataset: 266 Histology ROIs	MIA 2022
	Tang et al. (2023)	Carotid	KL-divergence + Mean Teacher U-Net++	CUBS; Private Dataset: 350 US Images	MICCAI 2023
	Li et al. (2023b)	Breast; Chest; Liver	Entropy + Consistency + Diversity	BUSI; CXRSet; Private Dataset: 90 3D CT Images; 3200 US Images	MIA 2023
	Kadir et al. (2023)	Eye	KL-divergence + Entropy	Duke; AROI; UMN	MICCAI 2023
	Bai et al. (2023a)	Breast	Diversity + Uncertainty	Private Dataset: 941 CT Images	MICCAI 2023
	Wang et al. (2024b)	Head; Neck	Euclidean Distance	Private Dataset: 1,057 Volume of MRIs	TMI 2024
	Mahapatra et al. (2024b)	Chest; Breast	Uncertainty + Similarity + Data Augmentation	ChestXray14; CheXpert; Camelyon17	MIA 2024
Others	Mahapatra et al. (2018)$_{CS}$	Chest	Bayesian Neural Network + cGAN Data Augmentation	JSRT Database; ChestX-ray8	MICCAI 2018
	Zhou et al. (2021c)$_{CD}$	Colon	Traditional Data Augmentation Entropy + Diversity	Private Dataset: 6 colonoscopy videos 38 polyp videos + 121 CTPA datasets	MIA 2021

[a] For the sake of brevity, we denote references that contain more than one task in the following abbreviations: **C**: Classification, **S**: Segmentation, **D**: Detection.

lead to redundancy, as similar samples may be repeatedly chosen (Zhou et al., 2021c). To overcome this, many later strategies integrate additional criteria such as diversity and representativeness into the selection process (Ozdemir et al., 2021; Zhao et al., 2017). By jointly considering these factors, hybrid methods aim to ensure that selected samples are both informative and broadly representative of the underlying data distribution—a consideration especially important in domains characterized by data heterogeneity or class imbalance.

To further enhance the sampling process, some approaches incorporate data augmentation or generative models, increasing the diversity of candidate samples and reducing annotation redundancy (Mahapatra et al., 2018). Optimization-based formulations have also been proposed, dynamically balancing uncertainty, consistency, and diversity during selection (Li et al., 2023b; Qu et al., 2023). Other strategies leverage feature-space metrics, clustering, or graph-based algorithms to maximize coverage and minimize overlap among selected samples (Bai et al., 2023a; Mahapatra et al., 2024a,b).

Overall, effective sampling strategies in active learning seek to maximize annotation value by balancing informativeness, diversity, and representativeness (Schmidt et al., 2024; Wang et al., 2024b; Tang et al., 2023). These principles have been shown to improve label efficiency and model robustness, particularly in complex and data-rich applications.

5.3. Discussion

Active learning addresses label scarcity in medical imaging by iteratively selecting the most informative samples for expert annotation, a process fundamentally driven by the interplay between informativeness evaluation and sampling strategy. In the early stages, model performance largely depends on how well informativeness is quantified, with advanced criteria providing clear gains over naive uncertainty measures (Wang et al., 2024c). As the labeled set expands, the challenge shifts from simply identifying informative samples to maximizing the added value of each annotation and avoiding redundancy, which becomes a key bottleneck (Föllmer et al., 2024).

Methodology Considerations. Building on the above, practical deployment of AL requires tailoring the selection strategy to different phases of the annotation process. In the early rounds, uncertainty-driven selection often achieves the highest marginal gains per query by rapidly correcting model blind spots. Its effectiveness, however, diminishes as the labeled pool grows, when redundancy and uncertainty miscalibration become more prominent under distribution shift. In the

mid to late phases, diversity-based or hybrid criteria are more effective, as they reduce redundancy and improve coverage, particularly in long-tailed or heterogeneous datasets. For dense prediction or spatially structured tasks, structure-aware selection at the patch or region level with flexible supervision types (point, box, mask) increases the information gain per annotation compared with image-level querying. Quality- and shift-aware selection, incorporating calibrated uncertainty and out-of-distribution detection, can prevent low-value annotations, although at the cost of additional system complexity. Staged pipelines applying uncertainty-based selection in early rounds, diversity or hybrid methods in later rounds, and structure-aware strategies for segmentation across all phases have demonstrated higher annotation efficiency and robustness. Representative examples include DSAL for segmentation (Zhao et al., 2021b), PLD for prototype-aware selection (Tang et al., 2023), and OpenAL for open-set recognition (Qu et al., 2023), which outperform single-criterion methods, especially under class imbalance, domain shift, or label noise (Li et al., 2023b; Huang et al., 2025).

Key Challenges. While these methodological patterns provide a practical roadmap, several unresolved issues hinder consistent performance across settings. First, traditional AL often yields marginal gains (<1% Dice) over random sampling in segmentation under strong regularization (Munjal et al., 2022; Burmeister et al., 2022), though batch-level diversity can recover 3–5% improvements by mitigating patch redundancy (Gaillochet et al., 2023). Second, uncertainty estimation is frequently poorly calibrated; while clinical deployment requires high reliability, AL models often suffer from overconfidence that biases sample selection and undermines performance during domain shifts (Wang et al., 2024c; Dai et al., 2024). Third, many diversity-based strategies (e.g., core-set) exhibit high computational complexity, posing significant scaling bottlenecks as unlabeled pools grow to tens of thousands of cases (Budd et al., 2021). Finally, annotation efficiency remains a bottleneck; recent real-world deployments highlight the need for AL frameworks that flexibly select both the image and the optimal label granularity, integrating point, box, and mask supervision to maximize clinical impact per expert-hour (Paranjape et al., 2024; Huang et al., 2025).

Real-World Clinical Impact and Performance Gains. Addressing these challenges is not solely an academic concern, as real-world studies already demonstrate the potential benefits when AL is well aligned with clinical workflows. Evidence from deployed or near-deployment systems shows that active learning can substantially reduce annotation workload without compromising clinical accuracy. In retinal disease screening, uncertainty-based triage has been reported to reduce expert reading workload while maintaining sensitivity (De Fauw et al., 2018a). In medical image segmentation, AIDE achieved near—fully supervised Dice scores using approximately 20% of the annotations (Wang and Yin, 2021). In digital pathology, prioritizing the most uncertain slides halved the annotation workload and improved detection sensitivity (Huang et al., 2025). For reproducibility, studies should report both accuracy and human-effort metrics such as annotation time, number of interactions, and coverage of challenging cases, and adopt label-granularity control to maximize clinical impact per expert hour. Including calibration and out-of-distribution checks can further prevent low-value queries in multi-institutional deployments.

6. Reshaping paradigms: The impact of HFMs

Building on the four label-efficient scenarios introduced in Sections 2–5, this section explores how HFMs transform label-efficient learning in MIA. Specifically, we aim to bridge the gap between conventional learning paradigms and the advancements enabled by HFMs, highlighting how large-scale pretraining fundamentally redefines the interaction between annotation quality and model performance. In this section, we begin by examining the paradigm shift from task-specific training to domain-scale pretraining, followed by an analysis of the role HFMs play within each annotation scenario. Finally, we synthesize how these advancements pave the way for unified workflows that transcend traditional boundaries, enabling adaptive and efficient solutions tailored to diverse data and resource constraints. Representative studies are summarized in Table 6.

6.1. The shift to domain-specific foundation models

The most consequential change introduced by HFMs is a move from isolated task-specific models to domain-scale pretraining on heterogeneous medical data (Zhou et al., 2023b; Chen et al., 2024c). Through self-supervision at scale, these models capture transferable representations that encode anatomical structure, disease patterns, and modality-specific cues (Zhou et al., 2023b; Chen et al., 2024c). Notable examples include RETFound for ophthalmology (Zhou et al., 2023b), UNI and Virchow for pathology (Chen et al., 2024c; Vorontsov et al., 2024), RAD-DINO for radiology (Perez-Garcia et al., 2025), and vision-language models such as CONCH, which is trained on millions of image-report pairs (Lu et al., 2024). These HFMs typically yield at least a 10% performance improvement over traditional task-specific models, with some human-machine comparison studies even demonstrating performance on par with human experts. As a result, the de facto starting point for label-efficient tasks has shifted from training networks from scratch to fine-tuning powerful, generalist models that encode extensive domain knowledge, substantially reducing the data and annotation requirements needed to achieve high performance.

6.2. Enhancement of conventional paradigms

Anchored in the shift to domain-scale pretraining outlined above, HFMs enhance conventional label-efficient paradigms by enabling label-efficient initialization, improving pseudo-label fidelity, and refining uncertainty estimation. For instance, in the *no label* and *insufficient label* settings, pretrained representations provide superior feature representations and facilitate accurate pseudo-labeling, which allows compact specialist heads to adapt from small curated sets (Tiu et al., 2022; Lu et al., 2023; Zhang et al., 2023c, 2025). In the *inexact label* and *label refinement* settings, transferable features strengthen weak supervision, such as in MIL, improve localization without pixel-level annotation, and leverage calibrated uncertainty to prioritize cases for review. Similarly, AL benefits from improved informativeness evaluation, enabling more efficient annotation strategies. Interactive guidance further converts sparse inputs into dense outputs, enabling refinement in complex scenarios (Lu et al., 2023; Ma et al., 2024a, 2025). Collectively, HFMs serve as unifying backbones that elevate performance and reduce annotation burden across the four scenarios, converting conventional paradigms into data-efficient and reliable pipelines.

6.3. New paradigms and unified frameworks

Beyond enhancement, HFMs introduce capabilities that transcend conventional boundaries, including prompt-based interaction for dense prediction, zero-shot generalization grounded in vision—language pretraining, and unified frameworks that adapt strategies to available supervision. *Prompt-based interaction models* (Wang et al., 2024c), exemplified by MedSAM (Ma et al., 2024a), reconceptualize efficiency by minimizing human effort per image through sparse prompts that elicit dense predictions. MedSAM2 (Ma et al., 2025) extends this to temporal sequences, enabling efficient video annotation. *Zero-shot learning models* leverage vision-language pre-training (Radford et al., 2021) to achieve expert-level performance on downstream tasks without requiring any task-specific manual labels. For example, CheXzero (Tiu et al., 2022) attains radiologist-level chest X-ray diagnosis by employing contrastive learning on image-text pairs, demonstrating robust zero-shot transfer. Knowledge-enhanced models (Zhang et al., 2023b) further incorporate domain knowledge to improve zero-shot generalization. While prompt-based and zero-shot models enhance flexibility at inference time, establishing a robust foundation model also requires overcoming the

Table 6
Surveyed representative health foundation models, with emphasis on their respective contributions to enhancing or reshaping the four label-efficient scenarios in medical image analysis.

	Reference (Year)	Domain	Methodology	Data Size	Publication[a]
HFM Pretraining	Zhou et al. (2022)	Chest X-ray	Vision-Language Contrastive Pretraining	377,110 Chest X-ray Images with Reports	Nat. Mach. Intell. 2022
	Azizi et al. (2023)	Pathology, Radiology, Dermatology & Ophthalmology	Contrastive Self-supervised Pretraining	~52.7M Images across 6 Domains	Nat. Biomed. Eng. 2023
	Zhou et al. (2023b)	Ophthalmology	Masked Self-supervised Pretraining	1,640,612 Retinal Images	Nature 2023
Enhancement	Chen et al. (2024c)	Pathology	Masked Self-supervised Pretraining	100,426 WSIs	Nat. Med. 2024
	Lu et al. (2024)	Pathology	Vision-Language Contrastive Pretraining	1,786,362 Pathology Image-Report Pairs	Nat. Med. 2024
	Vorontsov et al. (2024)	Pathology	Masked Self-supervised Pretraining	1,488,550 WSIs	Nat. Med. 2024
	Xu et al. (2024b)	Pathology, Genomics	Multi-modal Contrastive Pretraining	26,169 WSI-Genomics-Report Pairs	arXiv 2024
	Xu et al. (2024a)	Pathology	Masked Self-supervised Pretraining	171,189 WSIs	Nature 2024
	Wu et al. (2024a)	CT	Contrastive Self-supervised Pretraining	10,881 CT Scans	CVPR 2024
	Perez-Garcia et al. (2025)	Chest X-ray	Masked Self-supervised Pretraining	838,336 Chest X-ray Images	Nat. Mach. Intell. 2025
	Park et al. (2023)	Chest X-ray	Parameter-Efficient Fine-Tuning	677 cases across 10 datasets	MIA 2024
	Duy et al. (2023)	Multiple	SAM-Enhanced Graph Matching	~1.3M Images across 5 Domains	NeurIPS 2023
	Zhang et al. (2023c)	Cardiac MRI	SAM-Assisted Semi-supervised Learning	100 3D MR Scans	BIBM 2024
	Gupte et al. (2024)	Multiple	FM-enhanced Uncertainty Estimation	62,028 Images across 4 Domains	TMLR 2024
	Zhang et al. (2025)	MRI, PET	SAM-Assisted Semi-supervised Learning	535 3D Scans	arXiv 2025
New Paradigm	Tiu et al. (2022)	Chest X-ray	Vision-Language Zero-shot Learning	377,110 Chest X-ray Images with Reports	Nat. Biomed. Eng. 2022
	Lu et al. (2023)	Pathology	Vision-Language Zero-shot Transfer	33,480 Image-Caption Pairs	CVPR 2023
	Zhang et al. (2023b)	Chest X-ray	Knowledge-Enhanced Zero-shot Learning	377,110 Chest X-ray Images with Reports	Nat. Commun. 2023
	Zhou et al. (2024)	Multiple	HFMs with Multi-task Learning	29,594,554 Images	arXiv 2024
	He et al. (2024a)	Multiple	HFMs with Context-Enhanced Inference	2,043,955 Training Samples	arXiv 2024
	Ma et al. (2024a)	Multiple	Prompt-based Interactive Segmentation	1,570,263 Image-Mask Pairs	Nat. Commun. 2024
	Ma et al. (2025)	Multiple	Prompt-based Interactive Segmentation	530,827 3D Image-Mask Pairs / Frames	arXiv 2025

[a] For brevity, **HFM Pretraining** refers to large-scale pretraining strategies for HFMs; **Enhancement** denotes works where HFMs improve existing label-efficient paradigms; **New Paradigm** describes novel workflows or capabilities enabled by HFMs.

extreme heterogeneity of medical training data. Specifically, a critical step toward building fully unified frameworks is the ability to integrate multiple partially labeled datasets with conflicting annotations. To address this, several universal segmentation models have explored unique architectural adaptations. For instance, MultiTalent (Ulrich et al., 2023) tackles label inconsistency and overlapping structures across datasets by decoupling class predictions using independent sigmoid activations rather than mutually exclusive softmax, achieving high-precision results such as a 96.19% Dice score for liver segmentation. To handle multi-task routing across diverse datasets, UniSeg (Ye et al., 2023) differentiates itself from the aforementioned interactive models (e.g., MedSAM) by employing a learnable universal prompt tensor. Injecting this task-specific prompt early into the decoding stage allows the model to capture inter-task correlations implicitly, attaining a mean Dice score of 78.2% across 11 upstream datasets. Furthermore, addressing the label orthogonality problem inherent in standard multi-dataset training, the CLIP-Driven Universal Model (Liu et al., 2023a) replaces traditional one-hot encodings with CLIP text embeddings. Rather than utilizing CLIP solely for representation pretraining, this approach leverages the language prior to encode semantic anatomical relationships directly into the segmentation optimization, enabling the joint segmentation of 25 organs and 6 types of tumors. Equipped with these robust multi-dataset foundations and versatile inference mechanisms, HFMs are beginning to blur the traditional boundaries between different label-efficient paradigms. The emergence of *unified frameworks* that can seamlessly transition between zero-shot and fine-tuning scenarios represents a paradigm shift from specialized, single-purpose methods to flexible, adaptive systems. Recent work exemplifies this convergence. For example, MedVersa (Zhou et al., 2024) uses domain-aware learning and LLM orchestration to adapt inference strategies based on available annotations, achieving radiologist-level performance and outperforming other models by over 10% on multiple tasks. Radiologist evaluations show that MedVersa-generated reports match or exceed human reports in 71% of cases overall, and in 95% of normal studies. Similarly, MedDr (He et al., 2024a) combines retrieval-augmented diagnosis with lightweight specialist collaboration to achieve strong out-of-domain performance without task-specific fine-tuning, improving accuracy by up to 20% on complex diagnostic tasks compared to previous state-of-the-art, and outperforming both generalist foundation models and specialists on all out-of-domain datasets. Together, these advances suggest that the choice of label-efficient strategy will increasingly depend on available data and computational resources, rather than being fixed, fundamentally transforming label-efficient learning in medical imaging.

7. Challenges and future directions

Despite previous sections having addressed paradigm-specific challenges, substantial barriers remain in translating label-efficient learning methods for medical image analysis into clinical practice. While our preceding discussion on HFMs offers a forward-looking perspective on a highly popular and rapidly evolving domain, the core focus of this survey remains firmly grounded in established label-efficient paradigms. Contextualizing these dynamic trends within the broader landscape, this section systematically examines key cross-cutting obstacles and outlines potential research directions across the research-to-deployment continuum, as illustrated in Fig. 7.

7.1. Human-in-the-loop interaction

The introduction of HFMs pretrained on large-scale medical data has facilitated more efficient **human-in-the-loop (HITL)** collaboration in MIA, particularly for reducing expert annotation costs (Tanno

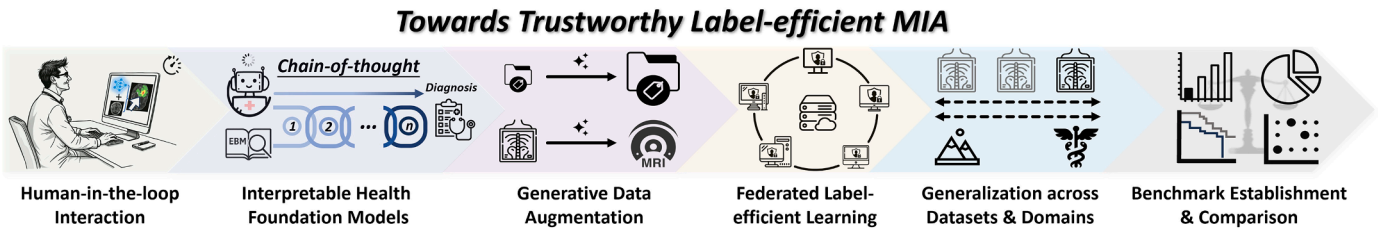


Fig. 7. Future directions for label-efficient MIA. Illustrated along the research-to-deployment continuum, encompassing the development of human-in-the-loop interaction, interpretable health foundation models, generative data augmentation, federated label-efficient learning, cross-domain generalization, and the establishment of standardized benchmarks for clinical evaluation.

et al., 2025). Unlike traditional label-efficient approaches such as active learning, HITL strategies incorporate real-time expert feedback during inference rather than just selecting informative samples for training (Wang et al., 2024c). Beyond technical benefits, effective HITL approaches aim to capture clinical reasoning patterns that transform sparse labels into knowledge-rich representations (Tanno et al., 2025). Recent advances in preference-based alignment suggest that even minimal human feedback may substantially enhance model outputs (Feng et al., 2025), indicating promising potential for reasoning-level feedback in medical imaging. Tanno et al. (2025) show that incorporating expert feedback through structured frameworks has the potential to improve clinical plausibility, while explainable interactions may enable clinicians to investigate and correct model predictions (Chanda et al., 2025), which is an important consideration for building trustworthiness in medical applications. Despite broader machine learning advances in adaptive querying and efficient verification of auto-suggested labels, these innovations remain underexplored in MIA, where annotation costs, privacy, and workflow integration pose unique challenges.

7.2. Interpretable health foundation models

Building upon HITL approaches that capture clinical reasoning, **interpretable health foundation models (HFMs)** represent the next critical step in maximizing label efficiency by making model reasoning processes transparent and aligned with medical knowledge (He et al., 2024b). This interpretability is particularly crucial for label-efficient learning, as understanding model reasoning helps identify where additional annotations would be most beneficial (Nie et al., 2025). Concept-based interpretable models that explicitly incorporate medical knowledge have demonstrated improved performance with fewer labels by focusing on clinically relevant visual features (Wen et al., 2024; Nie et al., 2025). Similarly, knowledge-guided prompt tuning approaches (Wang et al., 2025a) enable HFMs to leverage structured medical information alongside limited labeled data. While these techniques show promise, current HFMs often lack the ability to provide **evidence-based reasoning traces** necessary for clinical adoption. Recent advances in reinforcement post-training have shown potential for enhancing reasoning capabilities (Guo et al., 2025), suggesting pathways for medical HFMs to produce clinically-sound reasoning chains rather than post-hoc rationalizations. Future research may benefit from investigating how **chain-of-thought** approaches could potentially align model reasoning with clinical guidelines (Wang et al., 2025c), which could be essential for creating models that augment rather than replace clinical decision-making while minimizing annotation requirements.

7.3. Generative data augmentation

Building upon interpretable foundation models developed through expert collaboration, **generative data augmentation** directly addresses label scarcity by synthesizing rare pathological examples that would otherwise require extensive manual annotation. For medical conditions with limited available cases, acquiring sufficient labeled examples is prohibitively expensive and often infeasible (Chen et al., 2021). Conditional generation is particularly valuable for label efficiency, as text-guided diffusion models can create diverse examples of specific pathologies using just a handful of labeled reference cases (Wang et al., 2024e). The integration of expert knowledge in the generation process further enhances label efficiency by ensuring that synthetic examples maintain diagnostic features critical for model learning (Wang et al., 2025b). Moving forward, research should focus on optimizing the trade-off between synthetic data quantity and quality to maximize the label efficiency benefits, with particular attention to user-directed interfaces that allow clinicians to generate specifically needed rare cases with minimal annotation effort.

7.4. Federated label-efficient learning

As the scale of medical imaging data continues to grow, critical bottlenecks such as privacy concerns, uneven computational resources, and the need for cross-institutional collaboration have become increasingly prominent. **Federated learning (FL)** addresses these challenges by enabling institutions to collaboratively train models without sharing raw data, thus aggregating scarce annotations while safeguarding privacy. This approach facilitates label-efficient learning by maximizing the utility of limited labeled data across distributed sites (Rieke et al., 2020). This approach has yielded fruitful results in the field of MIA (Li et al., 2020c; Dayan et al., 2021; Lu et al., 2022; Che et al., 2025). However, current FL algorithms are primarily trained in a supervised manner. When applying the FL to real-world scenarios in MIA, a crucial problem, namely, label deficiency, may appear in local health datasets. Labels may be missing to varying degrees between medical centers, or the granularity of the labels will vary. Therefore, designing label-efficient federated learning methods to address this significant problem is an important research direction. For example, incorporating informative sampling strategies from active learning (Chen et al., 2024b; Wu et al., 2024b) into FL frameworks shows great potential for improving label efficiency in this setting. Nonetheless, FL faces unique challenges such as communication delays and efficiency; while recent works explore communication-efficient strategies (Yan et al., 2025), future research should investigate approaches such as adapters or low-rank updates, which can be communicated across institutions to jointly optimize both label efficiency and communication efficiency.

7.5. Generalization across domains and datasets

Building on advances in annotation efficiency and multi-center collaboration, the ultimate test of label-efficient learning in medical imaging is not merely reducing annotation costs, but ensuring that models can robustly adapt to the diverse data characteristics and clinical environments encountered across institutions. Despite progress in lowering annotation requirements, reliable clinical deployment remains challenging due to domain shifts arising from hardware heterogeneity, imaging protocols, and patient demographics (Yang et al., 2024b; He et al.,

2024b; Yoon et al., 2024). To address this, **domain generalization (DG)** must become a central design objective for future label-efficient MIA systems. Recent research also cautions that the apparent robustness of HFMs may be partially attributable to extensive data coverage or even inadvertent data leakage, rather than genuine intrinsic generalizability (Teterwak et al., 2025; Mayilvahanan et al., 2025). As a result, there is growing emphasis on systematically identifying and mitigating domain biases that persist after large-scale pretraining, with adaptive model recalibration being an active area of investigation (Zu et al., 2024; Chen et al., 2024d; Jiang et al., 2025). These advances collectively aim to ensure that models retain reliability and generalizability under label efficiency, which is essential for clinically robust and globally deployable MIA solutions.

7.6. Benchmark establishment and comparison

As the preceding sections have systematically reviewed the main challenges and promising directions in label-efficient MIA, their practical value, however, remains difficult to assess due to the lack of standardized and rigorous evaluation. The current landscape is highly fragmented, with studies differing in tasks, target organs, annotation budgets, protocols, and data partitioning (Bassi et al., 2024). Privacy constraints further restrict access to large and diverse datasets, limiting evaluation to single-center cohorts with non-standardized splits (Jin et al., 2024b). This fragmentation is further exacerbated by the inherent ambiguity in clinical "ground truth," where significant inter-observer variability among experts makes the establishment of a singular reference standard for evaluation highly complex. Beyond the subjective nature of labeling, the physical heterogeneity of imaging hardware and acquisition parameters across medical centers creates profound domain shifts that are difficult to encapsulate within a static benchmark. Simultaneously, the emergence of conflicting class definitions across multi-institutional datasets means that identical anatomical labels may follow disparate clinical inclusion criteria, further undermining the comparability of results. Ultimately, even when datasets are harmonized, a fundamental conceptual barrier remains in defining a universal metric for annotation cost, as the cognitive load and temporal investment required for different label types do not scale linearly across varied clinical specialties.

Establishing a rigorous evaluation framework that neutralizes these semantic, physical, and conceptual hurdles requires a paradigm shift toward a unified benchmarking infrastructure. Central to overcoming the conceptual cost barrier is the prioritization of curated multi-task datasets with fixed label budgets, which effectively grounds methodological comparisons in realistic clinical constraints. The influence of inter-observer variability and conflicting class definitions can be further mitigated by incorporating standardized human-in-the-loop protocols and unified curation guidelines. For the persistent challenges of data privacy and cross-hardware domain shifts, federated benchmarking offers a vital technical solution by allowing code to run securely across institutions without raw data sharing. By consolidating these strategies into cost-aware metrics and continuous, blinded leaderboards, the community can establish the reproducibility needed to translate label-efficient innovations into reliable, deployable systems.

8. Conclusion

Label-efficient learning is reshaping medical image analysis by directly addressing annotation scarcity and prompting a re-examination of how data, supervision, and clinical value interact. In this survey, we have introduced a unified taxonomy that spans scenarios of no label, insufficient label, inexact label, and label refinement. Through a critical analysis of state-of-the-art methodologies, we have highlighted both recent advances and the persistent barriers to clinical adoption. Our synthesis demonstrates that although the emergence of health foundation models has substantially reduced annotation demands and improved the transferability of learned representations, their generalization across diverse clinical settings and lack of interpretability continue to impede clinical adoption. Furthermore, the absence of standardized, cost-aware benchmarks limits the objective evaluation and comparison of label-efficient methods. By providing a unified perspective and clarifying the methodological landscape, this survey aims to offer a comprehensive reference for researchers and practitioners, and to support the development of clinically trustworthy, label-efficient solutions for medical imaging.

CRediT authorship contribution statement

Cheng Jin: Writing – review & editing, Writing – original draft, Visualization, Formal analysis, Conceptualization; **Zhengrui Guo:** Writing – review & editing, Writing – original draft, Formal analysis, Conceptualization; **Yi Lin:** Writing – original draft, Formal analysis; **Luyang Luo:** Writing – review & editing; **Hao Chen:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

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.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62202403), Hong Kong Innovation and Technology Commission (Project No. MHP/002/22 and ITCPD/17-9), and Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. R6003-22 and C4024-22GF).

Appendix A. Literature selection

To ensure the comprehensiveness of our literature corpus, we followed the PRISMA guidelines (Page et al., 2021) and implemented a four-stage search protocol covering the period from January 2016 to March 2025. We systematically queried Google Scholar and arXiv, using the Boolean search string ("medical image*" OR radiology OR "MRI" OR "CT" OR ultrasound OR "WSI" OR pathology) AND ("self-supervised" OR "semi-supervised" OR "weakly supervised" OR "active learning" OR "label-efficient" OR "foundation model"). This string was applied to both Google Scholar and the arXiv categories cs.CV and eess.IV. In addition, major conference proceedings (including CVPR, ICCV, ECCV, ICLR, NeurIPS, AAAI, WACV, MICCAI, MIDL, BIBM, and ISBI) and leading journals (including Medical Image Analysis, IEEE Transactions on Medical Imaging, IEEE Journal of Biomedical and Health Informatics, Nature Communications, Nature Machine Intelligence, Nature Biomedical Engineering, and Nature Medicine) were screened by filtering titles and abstracts using the same query. The automated search, conducted via Python scripts utilizing the scholarly and arXiv APIs, yielded 1462 records (Google Scholar = 1,183; arXiv = 279). Subsequently, full texts were reviewed to exclude non-imaging studies, purely theoretical works, and survey or editorial articles, yielding 1050 eligible publications. The yearly distribution of selected papers is visualized in Fig. 1.

Appendix B. Concepts of prior knowledge

B.1. Assumptions and details in semi-supervised learning

B.1.1. Assumptions in semi-supervised learning

Semi-SL is not universally effective, as stated in Van Engelen and Hoos (2020), Xiaojin (2008), a necessary condition for Semi-SL algorithms to work is that the marginal data distribution $p(x)$ contains underlying information about the posterior distribution $p(y|x)$, where x

and y represent the data sample over input space $\mathcal{X}$ and the associated label, respectively. Otherwise the additional unlabeled data will be useless to infer information about $p(y|x)$, which means the Semi-SL algorithms may achieve similar or even worse performance compared with supervised learning algorithms. Therefore, several assumptions over the input data distribution have been proposed to constrain the data structure and ensure the algorithms can be generalized from a limited labeled dataset to a large-scale unlabeled dataset. Following Van Engelen and Hoos (2020), Ouali et al. (2020), the assumptions in Semi-SL are introduced as follows:

Smoothness assumption. Suppose $x_1,\ x_2 \in X$ are two input data samples over input space $\mathcal{X}$. If the distance between x_1 and x_2 is very close, *i.e.*, $D(x_1, x_2) < \varepsilon$, where ε is an artificially set threshold, then the associated labels y_1 and y_2 should also be the same. Note that sometimes there is an additional constraint in the smoothness assumption. In Ouali et al. (2020), x_1 and x_2 are required to belong to the same high-density region, so as to avoid the situation that these two samples reside on the brink of different high-density regions and are misclassified as one category.

Cluster assumption. In this assumption, we assume that data points with similar underlying information are likely to form high-density regions, i.e., clusters. If the two data points x_1 and x_2 lie in the same cluster, then they are expected to have the same label. In fact, the cluster assumption can be considered as a special case of the smoothness assumption. According to Chapelle et al. (2009), if the two data points x_1 and x_2 can be connected with a line that does not pass through any low-density area, they belong to the same cluster.

Low-density assumption. The decision boundary of the classifier is assumed to lie in the low-density areas instead of high-density ones, which can be derived from the cluster assumption and smoothness assumption. On the one hand, if the decision boundary resides in the high-density regions, the two data points x_1 and x_2 located in the same cluster but opposite sided of the decision boundary will be categorized as different classes, which obviously violates the cluster assumption and smoothness assumption. On the other hand, following the cluster and smoothness assumption, data points in any high-density areas are expected to be assigned the same label, which means the decision boundary of the model can only lie in the low-density areas, thus satisfying the low-density assumption.

Manifold assumption. A manifold is a concept in geometry, that represents a geometric structure in a high-dimensional space, i.e., a collection of data points in the input space $\mathcal{X}$. For example, a curve in 2-dimensional space can be thought of as a 1-dimensional manifold, and a surface in 3-dimensional space can be seen as a 2-dimensional manifold. The manifold assumption states that there is a certain geometry of the data distribution in the high-dimensional space, namely that the data are concentrated around a certain low-dimensional manifold. Due to the fact that high-dimensional data not only poses a challenge to machine learning algorithms, but also leads to a large computational load and the problem of dimensional catastrophe, it will be much more effective to estimate the data distribution if they lie in a low-dimensional manifold.

B.1.2. Detail of key generative methods

Researchers can obtain various generative methods according to different assumptions on the latent distribution. On the one hand, it can be easy to formulate a generative method once an assumption on the distribution is made, whereas on the other, the hypothetical generative model must match the real data distribution to avoid the unlabeled data in turn degrading the generalization performance. One can formulate the modeling process of generative methods as follows:

$$\begin{aligned} y^* &= \arg\max_y p(y|x) = \arg\max_y \frac{p(x|y)p(y)}{p(x)} \\ &= \arg\max_y p(x|y)p(y), \end{aligned} \tag{B.1}$$

where the generative methods models the joint distribution $p(x, y)$. Eq. (B.1) indicates that if the correct assumption on prior $p(y)$ and conditional distribution $p(x|y)$ is made, the input data can be expected to come from the latent distribution.

Definition of Variational Autoencoder (VAE). The typical VAE consists of two objectives: one is to minimize the discrepancy between input data x and its reconstruction version $\hat{x}$ produced by the decoder, and the other is to model a latent space $p(z)$ following a simple distribution, such as a standard multivariate Gaussian distribution. Thus, the loss function for training a VAE can be formulated as follows:

$$\min_\theta \sum_{x\in X} \mathcal{L}(x,\theta) = \mathcal{L}_{MSE}(x,\hat{x}_\theta) + \mathcal{L}_{KL}(p_\theta(z|x)||p(z)),$$

where $\mathcal{L}_{MSE}$ represents the mean square error; $\hat{x}_\theta$ is the reconstruction version of input data x generated by the decoder $p_\phi(x|z)$ given parameters ϕ; $\mathcal{L}_{KL}(\cdot||\cdot)$ represents the Kullback-Leibler divergence which measures the distance between two distributions; and $p_\theta(z|x)$ denotes the posterior distribution produced by the encoder given parameters θ.

Definition of Generative Adversarial Network (GAN). The aim of generator $\mathcal{G}$ is to iteratively learn the latent distribution from real data x starting from generating data with random noise distribution $p(z)$. Meanwhile, the goal of the discriminator $\mathcal{D}$ is to correctly distinguish the fake input generated by $\mathcal{G}$ and real data x. Formally, we can formulate the optimization problem of a GAN as follows:

$$\begin{aligned} \min_{\mathcal{G}}\max_{\mathcal{D}} \mathcal{L}(\mathcal{G},\mathcal{D}) &= \mathbb{E}_{x\sim p(x)}[\log \mathcal{D}(x)] \\ &+ \mathbb{E}_{z\sim p(z)}[1-\log(\mathcal{D}(\mathcal{G}(z)))], \end{aligned}$$

where $\mathcal{L}$ represents the loss function of generator $\mathcal{G}$ and discriminator $\mathcal{D}$. Concretely, $\mathcal{G}$ aims to minimize the objective function by confusing $\mathcal{D}$ with generated data $\mathcal{G}(z)$, while $\mathcal{D}$ aims at maximizing the objective function by making correct predictions.

Definition of Diffusion Models. Diffusion models generate data by reversing a gradual noising process. They consist of a forward process that progressively adds Gaussian noise to data over timesteps and a learned reverse process that denoises it step by step. Formally, the forward process corrupts clean data x_0 into noisy data x_t following a variance schedule β_t:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t \mathbf{I}\right).$$

The model learns to approximate the reverse process $p_\theta(x_{t-1} \mid x_t)$ to recover the original data. Training minimizes a simplified denoising objective:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{x_0,\epsilon,t}\left[\left\|\epsilon - \epsilon_\theta(x_t, t)\right\|^2\right],$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is the noise added in the forward process, and ϵ_θ is the learned denoising network. Diffusion models offer stable training, strong sample diversity, and fine-grained control, making them highly effective for medical imaging tasks.

B.2. Conventional uncertainty metrics in active learning

The uncertainty measure reflects the degree of dispersion of a random input. There are many ways to measure the uncertainty of inputs. Starting with simple metrics like standard deviation and variance, current studies in MIA mainly focus on **margin sampling** (Campbell et al., 2000) and **entropy sampling** (Holub et al., 2008). Denote the probability as p, we give the definition of these metrics as follows.

Margin sampling (Campbell et al., 2000) estimates the probability difference $\mathcal{M}$ between the first and second most likely labels $\hat{y}_1, \hat{y}_2$ according to the deep model parameter θ and expect the least residual value by the following notation:

$$\mathcal{M} = \arg\min_x [p_\theta(\hat{y}_1 \mid x) - p_\theta(\hat{y}_2 \mid x)]$$

Entropy sampling (Holub et al., 2008) is another conventional metric for sampling. In a binary or multi-classification scenario, the sampled data with higher entropy can be selected as the expected annotation data. For a C-class task, entropy sampling metric $\mathcal{E}$ can be denoted as follows:

$$\mathcal{E} = \underset{x}{\operatorname{argmax}}\left(-\sum_{c=1}^{C} p(y_c \mid x) \log p(y_c \mid x) \right)$$

Appendix C. Surveyed datasets

As a supplement to the main text, we summarize representative publicly available datasets across 16 different organs such as the brain, chest, prostate, *etc.* in Table C.1. These publicly available MIA datasets can be leveraged to construct label-efficient learning algorithms for numerous purposes, including classification, detection, and segmentation.

Table C.1
Summary of publicly available databases for label-efficient learning in MIA.

Organ	Dataset (Year)	Task	Link
Brain	BraTS (2012)	Segmentation	http://www.imm.dtu.dk/projects/BRATS2012/data.html
	BraTS (2013) (Menze et al., 2014)	Segmentation	https://www.smir.ch/BRATS/Start2013{***}{---}{***}download
	BraTS (2015)	Segmentation	https://www.smir.ch/BRATS/Start2015
	BraTS (2017)	Segmentation	https://sites.google.com/site/braintumorsegmentation/
	BraTS (2018)	Segmentation	https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=37224922
	MSD (2018) (Simpson et al., 2019)	Segmentation	https://drive.google.com/drive/folders/1HqEgzS8BV2c7xYNrZdEAnrHk7osJJ--2
	dHCP (2018) (Makropoulos et al., 2018)	Segmentation	http://www.developingconnectome.org/data-release/
	JSRT Database (2000) (Shiraishi et al., 2000)	Classification	http://db.jsrt.or.jp/eng.php
	MRBrainS18 (2018)	Segmentation	https://mrbrains18.isi.uu.nl/data/
	BigBrain (2013) (Amunts et al., 2013)	Segmentation	https://bigbrainproject.org/maps-and-models.html{***}download
	MALC (2012)	Segmentation	http://www.neuromorphometrics.com/2012_MICCAI_Challenge_Data.html
	TCIA (2015) (Seff et al., 2015)	Segmentation	https://www.cancerimagingarchive.net/
	OASIS (2007)	Segmentation	https://www.oasis-brains.org/{***}data
	UKBB (2016)	Classification	https://www.ukbiobank.ac.uk/
	ADNI (2010)	Classification	https://www.adni-info.org/
	ABIDE (2016)	Classification	https://fcon_1000.projects.nitrc.org/indi/abide/
	MIRIAD (2012)	Classification	https://www.ucl.ac.uk/drc/research/research-methods/minimal-interval-resonance-imaging-alzheimers-disease-miriad
Chest	IS-COVID (2020) (Fan et al., 2020)	Segmentation	http://medicalsegmentation.com/covid19/
	CC-COVID (2020) (Zhang et al., 2020)	Segmentation	https://ncov-ai.big.ac.cn/download?lang=en
	NLST (2009)	Detection	https://cdas.cancer.gov/datasets/nlst/
	NIH Chest X-ray (2017) (Wang et al., 2017)	Classification	https://www.kaggle.com/datasets/nih-chest-xrays/data
	TCGA-Lung	Classification	https://portal.gdc.cancer.gov/repository
	LDCTGC (2016)	Detection	https://www.aapm.org/grandchallenge/lowdosect/
	ChestX (2018) (Kermany et al., 2018)	Classification	https://data.mendeley.com/datasets/rscbjbr9sj/3
	LUNA (2016)	Detection	https://luna16.grand-challenge.org/
Chest (Continued)	CAD-PE (2019)	Segmentation	https://ieee-dataport.org/open-access/cad-pe
	SIIM-ACR (2019)	Segmentation	https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation
	RSNA-Lung (2018)	Detection	https://www.kaggle.com/c/rsna-pneumonia-detection-challenge
	VinDr-CXR (2021)	Detection	https://vindr.ai/datasets/cxr
	Montgomery (2022)	Segmentation	https://www.kaggle.com/datasets/raddar/tuberculosis-chest-xrays-montgomery
	ChestXR (2021)	Classification	https://cxr-covid19.grand-challenge.org/Dataset/
	MIMIC-CXR (2019) (Johnson et al., 2019)	Detection	https://physionet.org/content/mimic-cxr/2.0.0/
	CC-CCII (2020)(Zhang et al., 2020)	Classification	http://ncov-ai.big.ac.cn/download/
	ChestX-ray8 (2017) (Wang et al., 2017)	Segmentation	https://nihcc.app.box.com/v/ChestXray-NIHCC/
	ChestX-ray14 (2019)	Classification	https://stanfordmlgroup.github.io/competitions/chexpert/
	CheXpert (2019) (Wang et al., 2017)	Segmentation	https://stanfordmlgroup.github.io/competitions/chexpert/
Gland	GlaS (2015)(Sirinukunwattana et al., 2015)	Segmentation	https://warwick.ac.uk/fac/cross_fac/tia/data/glascontest/download/
	CRAG (2017)	Segmentation	https://warwick.ac.uk/fac/sci/dcs/research/tia/data/mildnet
Prostate	Prostate-MRI-US-Biopsy (2013) (Sonn et al., 2013)	Segmentation	https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=68550661
	PANDA (2020) (Bulten et al., 2020)	Classification	https://www.kaggle.com/c/prostate-cancer-grade-assessment/data/
	ProMRI (2012) (Litjens et al., 2014; Tian et al., 2015)	Segmentation	https://promise12.grand-challenge.org/
	TMA-Zurich (2018) (Arvaniti et al., 2018)	Classification	https://www.nature.com/articles/s41598-018-30535-1?source=app****data-availability

Table C.1
Continue

Organ	Dataset (Year)	Task	Link
Heart	GGC (2019) MSCMRseg (Zhuang, 2016)	Classification Segmentation	https://gleason2019.grand-challenge.org/Register/ https://zmiclab.github.io/zxh/0/mscmrseg19/
	MM-WHS (2017)	Segmentation	https://zmiclab.github.io/zxh/0/mmwhs/
	Endocardium-MRI (2008) (Andreopoulos and Tsotsos, 2008)	Segmentation	https://www.sciencedirect.com/science/article/pii/S1361841508000029***aep-e-component-id41
	M&Ms (2020)	Segmentation	https://www.ub.edu/mnms/
	ASG (2018) (Xiong et al., 2021)	Segmentation	http://atriaseg2018.cardiacatlas.org/
Eye	DRISHTI-GS (2014) (Sivaswamy et al., 2014)	Segmentation	https://www.kaggle.com/datasets/lokeshsaipureddi/drishtigs-retina-dataset-for-onh-segmentation
Eye (Continued)	REFUGE (Orlando et al., 2020)	Segmentation	https://ieee-dataport.org/documents/refuge-retinal-fundus-glaucoma-challenge
	DRD (2015)	Detection	https://www.kaggle.com/competitions/diabetic-retinopathy-detection
	RetinalOCT (2018) (Kermany et al., 2018)	Classification	https://www.kaggle.com/datasets/paultimothymooney/kermany2018
	ReTOUCH (2017) (Bogunović et al., 2019)	Classification	https://retouch.grand-challenge.org/
	ORIGA-light (2010) (Zhang et al., 2010)	Classification	https://www.medicmind.tech/retinal-image-databases
	sjchoi86-HRF (2017)	Classification	https://github.com/cvblab/retina_dataset
	DRIVE (2004)	Classification	https://drive.grand-challenge.org/
	Isfahan MISP (2017)	Segmentation	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5394805/
	Duke-AMD (2014) (Farsiu et al., 2014)	Classification	https://people.duke.edu/~sf59/RPEDC_Ophth_2013_dataset.htm
	ADAM (2020) (Fang et al., 2022)	Segmentation	https://amd.grand-challenge.org/
	PALM (2019) (Fu et al., 2019)	Segmentation	https://palm.grand-challenge.org/
	FFA (2012) (Hajeb Mohammad Alipour et al., 2012)	Classification	http://misp.mui.ac.ir/data/eye-images.html
	OCTA-500 (2022)	Classification	https://ieee-dataport.org/open-access/octa-500
	GAMMA (2021)	Classification	https://aistudio.baidu.com/aistudio/competition/detail/119/0/introduction
	Messidor (2014) (Decencière et al., 2014)	Classification	https://www.adcis.net/en/third-party/messidor/
Kidney	KiTS (2019)	Segmentation	https://kits21.kits-challenge.org/
Skin	HAM10000 (2018) (Tschandl et al., 2018)	Segmentation	https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T
	ISIC (2016) (Gutman et al., 2016)	Classification	https://challenge.isic-archive.com/data/***2016
	ISIC (2017) (Codella et al., 2018)	Segmentation	https://challenge.isic-archive.com/data/***2017/
	ISIC (2018) (Milton, 2019)	Segmentation	https://challenge.isic-archive.com/landing/2018/
Hand	RSNA-Bone (2017)(Halabi et al., 2019)	Segmentation	https://www.rsna.org/education/ai-resources-and-training/ai-image-challenge/rsna-pediatric-bone-age-challenge-2017/
Colon	KATHER (2018) (Kather et al., 2019)	Classification	https://zenodo.org/record/1214456***.Y8fgV-zP1hE
	MHIST (2021)	Classification	https://bmirds.github.io/MHIST/
	CRCHistoPhenotypes (2016) (Sirinukunwattana et al., 2016)	Classification	https://warwick.ac.uk/fac/cross_fac/tia/data/crchistolabelednucleihe
Abdomen	ACDC (2018) (Bernard et al., 2018)	Segmentation	https://www.creatis.insa-lyon.fr/Challenge/acdc/databases.html
	CHAOS (2021) (Kavur et al., 2021)	Segmentation	https://chaos.grand-challenge.org/
Breast	BACH (2018) (Aresta et al., 2019)	Classification	https://iciar2018-challenge.grand-challenge.org/Dataset/
	NYUBCS (2019) (Wu et al., 2019)	Segmentation	https://datacatalog.med.nyu.edu/dataset/10518
	CBIS-DDSM (2017) (Lee et al., 2017)	Segmentation	https://www.kaggle.com/datasets/awsaf49/cbis-ddsm-breast-cancer-image-dataset
	MIAS (2015) (Suckling et al., 2015)	Detection	https://www.kaggle.com/datasets/kmader/mias-mammography
	TCGA-Breast	Classification	https://portal.gdc.cancer.gov/repository
	INBreast (2012)	Classification	https://biokeanos.com/source/INBreast
	BreastPathQ (2019)	Classification	https://breastpathq.grand-challenge.org/Overview/

Table C.1
Continue

Organ	Dataset (Year)	Task	Link
	CAMELYON (2016)	Classification	https://camelyon16.grand-challenge.org/Data/
	CAMELYON (2017)	Classification	https://camelyon17.grand-challenge.org/Data/
	BreakHis (2016)	Classification	https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/
	CBCS3 (2018) (Troester et al., 2018)	Classification	https://unclineberger.org/cbcs/for-researchers/
	TNBC (2018) (Naylor et al., 2018)	Segmentation	https://ega-archive.org/datasets/EGAD00001000063
	TUPAC (2016) (Veta et al., 2019)	Segmentation	https://github.com/CODAIT/deep-histopath
	MITOS12 (Ludovic et al., 2013)	Segmentation	http://ludo17.free.fr/mitos_2012/dataset.html
	MITOS14	Segmentation	https://mitos-atypia-14.grand-challenge.org/Dataset/
	TMA-UCSB (2014) (Kandemir et al., 2014)	Classification	https://bioimage.ucsb.edu/research/biosegmentation
Cell	PHC (2013) (Maška et al., 2014)	Segmentation	http://celltrackingchallenge.net/
	CPM (2017) (Vu et al., 2019)	Segmentation	http://simbad.u-strasbg.fr/simbad/sim-id?Ident=CPM+17
Liver	LiTS (2017)	Segmentation	https://competitions.codalab.org/competitions/17094
	PAIP (2019)	Segmentation	https://paip2019.grand-challenge.org/Dataset/
	TCGA-Liver	Classification	https://portal.gdc.cancer.gov/repository
Lymph Node	PatchCAMELYON (2017)	Classification	https://patchcamelyon.grand-challenge.org/Download/
	NIH LN (2016)	Classification	https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=19726546
Pancreas	NIH PCT	Segmentation	https://wiki.cancerimagingarchive.net/display/Public/Pancreas-CT
Multi-organ	DSB (2018)	Segmentation	https://www.kaggle.com/competitions/data-science-bowl-2018/data
	DeepLesion (2018) (Yan et al., 2018)	Detection	https://nihcc.app.box.com/v/DeepLesion
	WTS (2020) (Keikhosravi et al., 2020)	Super-resolution	https://www.nature.com/articles/s42003-020-01151-5***data-availability
	DECATHLON (2019) (Simpson et al., 2019)	Segmentation	http://medicaldecathlon.com/
	MoNuSeg (2017) (Kumar et al., 2017)	Segmentation	https://monuseg.grand-challenge.org/
	MoCTSeg (2018) (Gibson et al., 2018a)	Segmentation	https://www.synapse.org/{***}{---}Synapse:syn3376386
	BTCV (2017) (Gibson et al., 2018b)	Segmentation	https://zenodo.org/record/1169361{***}.Y8Ud-OxBwUE
	CT-ORG (Roth et al., 2015; Rister et al., 2020)	Segmentation	https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=61080890
	NIH PLCO (2011) (Oken et al., 2011)	Classification	https://cdas.cancer.gov/datasets/plco/
	BCV (2017)	Segmentation	https://www.synapse.org/{***}{---}Synapse:syn3193805/files/
	MIDOG (Aubreville et al., 2021)	Segmentation	https://imig.science/midog/the-dataset/

References

Abbet, C., Zlobec, I., Bozorgtabar, B., Thiran, J.-P., 2020. Divide-and-rule: self-supervised learning for survival analysis in colorectal cancer. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 480–489.

Adiga, S., Dolz, J., Lombaert, H., 2024. Anatomically-aware uncertainty for semi-supervised image segmentation. Med. Image Anal. 91 103011.

Amunts, K., Lepage, C., Borgeat, L., Mohlberg, H., Dickscheid, T., Rousseau, M.-É., Bludau, S., Bazin, P.-L., Lewis, L.B., Oros-Peusquens, A.-M., et al., 2013. Bigbrain: an ultrahigh-resolution 3D human brain model. Science 340 (6139), 1472–1475.

Andreopoulos, A., Tsotsos, J.K., 2008. Efficient and generalizable statistical models of shape and appearance for analysis of cardiac MRI. Med. Image Anal. 12 (3), 335–357.

Aresta, G., Araújo, T., Kwok, S., Chennamsetty, S.S., Safwan, M., Alex, V., Marami, B., Prastawa, M., Chan, M., Donovan, M., et al., 2019. Bach: grand challenge on breast cancer histology images. Med. Image Anal. 56, 122–139.

Arvaniti, E., Fricker, K.S., Moret, M., Rupp, N., Hermanns, T., Fankhauser, C., Wey, N., Wild, P.J., Rueschoff, J.H., Claassen, M., 2018. Automated gleason grading of prostate cancer tissue microarrays via deep learning. Sci. Rep. 8 (1), 1–11.

Atzeni, A., Peter, L., Robinson, E., Blackburn, E., Althonayan, J., Alexander, D.C., Iglesias, J.E., 2022. Deep active learning for suggestive segmentation of biomedical image stacks via optimisation of dice scores and traced boundary length. Med. Image Anal. 81, 102549.

Aubreville, M., Bertram, C., Veta, M., Klopfleisch, R., Stathonikos, N., Breininger, K., ter Hoeve, N., Ciompi, F., Maier, A., 2021. Mitosis domain generalization challenge. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention, pp. 1–15.

Azizi, S., Culp, L., Freyberg, J., Mustafa, B., Baur, S., Kornblith, S., Chen, T., Tomasev, N., Mitrović, J., Strachan, P., et al., 2023. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. Nat. Biomed. Eng. 7 (6), 756–779.

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., et al., 2021. Big self-supervised models advance medical image classification. In: Proc. IEEE Int. Conf. Comput. Vis., pp. 3478–3488.

Bai, F., Yan, K., Bai, X., Mao, X., Yin, X., Zhou, J., Shi, Y., Lu, L., Meng, M. Q.-H., 2023a. SLPT: selective labeling meets prompt tuning on label-limited lesion segmentation. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 14–24.

Bai, W., Chen, C., Tarroni, G., Duan, J., Guitton, F., Petersen, S.E., Guo, Y., Matthews, P.M., Rueckert, D., 2019. Self-supervised learning for cardiac mr image segmentation by anatomical position prediction. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 541–549.

Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G., Glocker, B., King, A., Matthews, P.M., Rueckert, D., 2017. Semi-supervised learning for network-based cardiac MR image segmentation. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 253–260.

Bai, Y., Chen, D., Li, Q., Shen, W., Wang, Y., 2023b. Bidirectional copy-paste for semi-supervised medical image segmentation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 11514–11524.

Barsha, F.L., Eberle, W., 2025. An in-depth review and analysis of mode collapse in generative adversarial networks. Mach. Learn. 114 (6), 141.

Basak, H., Yin, Z., 2023. Pseudo-label guided contrastive learning for semi-supervised medical image segmentation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 19786–19797.
Bashir, R. M.S., Qaiser, T., Raza, S. E.A., Rajpoot, N.M., 2024. Consistency regularisation in varying contexts and feature perturbations for semi-supervised semantic segmentation of histology images. Med. Image Anal. 91, 102997.
Bassi, P.R., Li, W., Tang, Y., Isensee, F., Wang, Z., Chen, J., Chou, Y.-C., Kirchhoff, Y., Rokuss, M.R., Huang, Z., et al., 2024. Touchstone benchmark: are we on the right way for evaluating AI algorithms for medical segmentation? Proc. Adv. Neural Inf. Process. Syst. 37, 15184–15201.
Belharbi, S., Ben Ayed, I., McCaffrey, L., Granger, E., 2021. Deep active learning for joint classification & segmentation with weak annotator. In: Proc. IEEE Winter Conf. App. Comput. Vis., pp. 3338–3347.
Berenguer, A.D., Kvasnytsia, M., Bossa, M.N., Mukherjee, T., Deligiannis, N., Sahli, H., 2024. Semi-supervised medical image classification via distance correlation minimization and graph attention regularization. Med. Image Anal. 94, 103107.
Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.-A., Cetin, I., Lekadir, K., Camara, O., Ballester, M. A.G., et al., 2018. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging 37 (11), 2514–2525.
Bogunović, H., Venhuizen, F., Klimscha, S., Apostolopoulos, S., Bab-Hadiashar, A., Bagci, U., Beg, M.F., Bekalo, L., Chen, Q., Ciller, C., et al., 2019. Retouch: the retinal oct fluid detection and segmentation benchmark and challenge. IEEE Trans. Med. Imaging 38 (8), 1858–1874.
Bortsova, G., Dubost, F., Hogeweg, L., Katramados, I., Bruijne, M.d., 2019. Semi-supervised medical image segmentation via learning consistency under transformations. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 810–818.
Budd, S., Robinson, E.C., Kainz, B., 2021. A survey on active learning and human-in-the-loop deep learning for medical image analysis. Med. Image Anal. 71, 102062.
Bulten, W., Litjens, G., Pinckaers, H., Ström, P., Eklund, M., Kartasalo, K., Demkin, M., Dane, S., 2020. The PANDA challenge: prostate cANcer grade assessment using the gleason grading system. MICCAI Challenge. 28 (1), 154–163.
Burmeister, J.-M., Rosas, M.F., Hagemann, J., Kordt, J., Blum, J., Shabo, S., Bergner, B., Lippert, C., 2022. Less is more: a comparison of active learning strategies for 3D medical image segmentation. arXiv:2207.00845.
Cai, Y., Chen, H., Yang, X., Zhou, Y., Cheng, K.-T., 2023. Dual-distribution discrepancy with self-supervised refinement for anomaly detection in medical images. Med. Image Anal. 86, 102794.
Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J., 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. 25 (8), 1301–1309.
Campbell, C., Cristianini, N., Smola, A., et al., 2000. Query learning with large margin classifiers. In: ICML. Vol. 20, p. 0.
Can, Y.B., Chaitanya, K., Mustafa, B., Koch, L.M., Konukoglu, E., Baumgartner, C.F., 2018. Learning to segment medical images with scribble-supervision alone. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 236–244.
Cao, B., Zhang, H., Wang, N., Gao, X., Shen, D., 2020a. Auto-GAN: self-supervised collaborative learning for medical image synthesis. In: AAAI Conf. Artif. Intell. Vol. 34, pp. 10486–10493.
Cao, K., Xia, Y., Yao, J., Han, X., Lambert, L., Zhang, T., Tang, W., Jin, G., Jiang, H., Fang, X., et al., 2023. Large-scale pancreatic cancer detection via non-contrast CT and deep learning. Nat. Med. 29 (12), 3033–3043.
Cao, X., Chen, H., Li, Y., Peng, Y., Wang, S., Cheng, L., 2020b. Uncertainty aware temporal-ensembling model for semi-supervised abus mass segmentation. IEEE Trans. Med. Imaging 40 (1), 431–443.
Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E., 2023. Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation. Med. Image Anal. 87, 102792.
Chaitanya, K., Karani, N., Baumgartner, C.F., Erdil, E., Becker, A., Donati, O., Konukoglu, E., 2021. Semi-supervised task-driven data augmentation for medical image segmentation. Med. Image Anal. 68, 101934.
Chanda, T., Haggenmueller, S., Bucher, T.-C., Holland-Letz, T., Kittler, H., Tschandl, P., Heppt, M.V., Berking, C., Utikal, J.S., Schilling, B., et al., 2025. Dermatologist-like explainable AI enhances melanoma diagnosis accuracy: eye-tracking study. Nat. Commun. 16 (1), 4739.
Chapelle, O., Scholkopf, B., Zien, A., et al., 2009. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. IEEE Trans. Neural Netw. 20 (3), 542–542.
Che, H., Jin, H., Guo, Z., Lin, Y., Jin, C., Chen, H., 2025. LLM-driven medical report generation via communication-efficient heterogeneous federated learning. arXiv:2506.17562.
Chen, C., Miao, J., Wu, D., Zhong, A., Yan, Z., Kim, S., Hu, J., Liu, Z., Sun, L., Li, X., Liu, T., Heng, P.-A., Li, Q., 2024a. MA-SAM: modality-agnostic SAM adaptation for 3D medical image segmentation. Med. Image Anal. 98, 103310.
Chen, D.-D., Wang, W., Gao, W., Zhou, Z.-H., 2018. Tri-net for semi-supervised deep learning. In: Proceedings of Twenty-Seventh International Joint Conference on Artificial Intelligence, pp. 2014–2020.
Chen, H., Qi, X., Yu, L., Heng, P.-A., 2016. Dcan: deep contour-aware networks for accurate gland segmentation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 2487–2496.
Chen, J., Ma, B., Cui, H., Xia, Y., 2024b. Think twice before selection: federated evidential active learning for medical image analysis with domain shifts. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 11439–11449.
Chen, J., Zhang, J., Debattista, K., Han, J., 2022a. Semi-supervised unpaired medical image segmentation through task-affinity consistency. IEEE Trans. Med. Imaging 42 (3), 594–605.
Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., Rueckert, D., 2019. Self-supervised learning for medical image analysis using image context restoration. Med. Image Anal. 58, 101539.
Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F., 2022b. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 16144–16155.
Chen, R.J., Ding, T., Lu, M.Y., Williamson, D. F.K., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al., 2024c. Towards a general-purpose foundation model for computational pathology. Nat. Med. 30 (3), 850–862.
Chen, R.J., Lu, M.Y., Chen, T.Y., Williamson, D. F.K., Mahmood, F., 2021. Synthetic data in machine learning for medicine and healthcare. Nat. Biomed. Eng. 5 (6), 493–497.
Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations. In: Proc. Int. Conf. Mach. Learn. PMLR, pp. 1597–1607.
Chen, Z., Pan, Y., Ye, Y., Lu, M., Xia, Y., 2024d. Each test image deserves a specific prompt: continual test-time adaptation for 2D medical image segmentation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 11184–11193.
Chen, Z., Tian, Z., Zhu, J., Li, C., Du, S., 2022c. C-CAM: causal CAM for weakly supervised semantic segmentation on medical image. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 11676–11685.
Cheplygina, V., de Bruijne, M., Pluim, J. P.W., 2019. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Med. Image Anal. 54, 280–296.
Chi, H., Pang, J., Zhang, B., Liu, W., 2024. Adaptive bidirectional displacement for semi-supervised medical image segmentation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 4070–4080.
Chikontwe, P., Kim, M., Nam, S.J., Go, H., Park, S.H., 2020. Multiple instance learning with center embeddings for histopathology classification. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 519–528.
Cho, J., Jun, Y., Wang, X., Kobayashi, C., Bilgic, B., 2023. Improved multi-shot diffusion-weighted MRI with zero-shot self-supervised learning reconstruction. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 457–466.
Cisternino, F., Ometto, S., Chatterjee, S., Giacopuzzi, E., Levine, A.P., Glastonbury, C.A., 2024. Self-supervised learning for characterising histomorphological diversity and spatial RNA expression prediction across 23 human tissue types. Nat. Commun. 15 (1), 5906.
Codella, N. C.F., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al., 2018. Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging (ISBI). In: Proc. IEEE Int. Symp. Biomed. Imaging. IEEE, pp. 168–172.
Couture, H.D., Marron, J.S., Perou, C.M., Troester, M.A., Niethammer, M., 2018. Multiple instance learning for heterogeneous images: training a cnn for histopathology. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 254–262.
Cui, C., Deng, R., Liu, Q., Yao, T., Bao, S., Remedios, L.W., Landman, B.A., Tang, Y., Huo, Y., 2024. All-in-SAM: from weak annotation to pixel-wise nuclei segmentation with prompt-based finetuning. J. Phys. Conf. Ser. 2722 (1), 012012.
Dai, L., Sheng, B., Chen, T., Wu, Q., Liu, R., Cai, C., Wu, L., Yang, D., Hamzah, H., Liu, Y., et al., 2024. A deep learning system for predicting time to progression of diabetic retinopathy. Nat. Med. 30 (2), 584–594.
Das, K., Conjeti, S., Roy, A.G., Chatterjee, J., Sheet, D., 2018. Multiple instance learning of deep convolutional neural networks for breast histopathology whole slide classification. In: Proc. IEEE Int. Symp. Biomed. Imaging. IEEE, pp. 578–581.
Dayan, I., Roth, H.R., Zhong, A., Harouni, A., Gentili, A., Abidin, A.Z., Liu, A., Costa, A.B., Wood, B.J., Tsai, C.-S., et al., 2021. Federated learning for predicting clinical outcomes in patients with COVID-19. Nat. Med. 27 (10), 1735–1743.
De Fauw, J., Ledsam, J.R., Romera-Paredes, B., et al., 2018a. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat. Med. 25, 70–76.
De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., et al., 2018b. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat. Med. 24 (9), 1342–1350.
Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., Gain, P., Ordonez, R., Massin, P., Erginay, A., et al., 2014. Feedback on a publicly distributed image database: the messidor database. Image Anal. Stereol. 33 (3), 231–234.
Deng, R., Cui, C., Remedios, L.W., Bao, S., Womick, R.M., Chiron, S., Li, J., Roland, J.T., Lau, K.S., Liu, Q., et al., 2024. Cross-scale multi-instance learning for pathological image diagnosis. Med. Image Anal. 94, 103124.
Dhariwal, P., Nichol, A., 2021. Diffusion models beat gans on image synthesis. Adv. Neural Inf. Process. Syst. 34, 8780–8794.
Di Folco, M., Bercea, C.I., Chan, E., Schnabel, J.A., 2024. Interpretable representation learning of cardiac MRI via attribute regularization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 492–501.
Diaz-Pinto, A., Colomer, A., Naranjo, V., Morales, S., Xu, Y., Frangi, A.F., 2019. Retinal image synthesis and semi-supervised learning for glaucoma assessment. IEEE Trans. Med. Imaging 38 (9), 2211–2218.
Dorent, R., Joutard, S., Shapey, J., Bisdas, S., Kitchen, N., Bradford, R., Saeed, S., Modat, M., Ourselin, S., Vercauteren, T., 2020. Scribble-based domain adaptation via co-segmentation. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 479–489.
Dorent, R., Joutard, S., Shapey, J., Kujawa, A., Modat, M., Ourselin, S., Vercauteren, T., 2021. Inter extreme points geodesics for end-to-end weakly supervised image seg-

mentation. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 615–624.
Dov, D., Kovalsky, S.Z., Assaad, S., Cohen, J., Range, D.E., Pendse, A.A., Henao, R., Carin, L., 2021. Weakly supervised instance learning for thyroid malignancy prediction from whole slide cytopathology images. Med. Image Anal. 67, 101814.
Du, H., Dong, Q., Xu, Y., Liao, J., 2023. Weakly-supervised 3D medical image segmentation using geometric prior and contrastive similarity. IEEE Trans. Med. Imaging 42, 2936–2947.
Du, W., Huo, Y., Zhou, R., Sun, Y., Tang, S., Zhao, X., Li, Y., Li, G., 2024. Consistency label-activated region generating network for weakly supervised medical image segmentation. Comput. Biol. Med. 173, 108380.
Dubost, F., Adams, H., Yilmaz, P., Bortsova, G., van Tulder, G., Ikram, M.A., Niessen, W., Vernooij, M.W., de Bruijne, M., 2020. Weakly supervised object detection with 2D and 3D regression neural networks. Med. Image Anal. 65, 101767.
Duy, M.H.N., Nguyen, H., Diep, N., Pham, T.N., Cao, T., Nguyen, B., Swoboda, P., Ho, N., Albarqouni, S., Xie, P., et al., 2023. LVM-MED: learning large-scale self-supervised vision models for medical imaging via second-order graph matching. Adv. Neural Inf. Process. Syst. 36, 27922–27950.
Fan, D.-P., Zhou, T., Ji, G.-P., Zhou, Y., Chen, G., Fu, H., Shen, J., Shao, L., 2020. INF-Net: automatic Covid-19 lung infection segmentation from CT images. IEEE Trans. Med. Imaging 39 (8), 2626–2637.
Fan, L., Sowmya, A., Meijering, E., Song, Y., 2023. Cancer survival prediction from whole slide images with self-supervised learning and slide consistency. IEEE Trans. Med. Imaging. 42 (5), 1401–1412.
Fang, H., Li, F., Fu, H., Sun, X., Cao, X., Lin, F., Son, J., Kim, S., Quellec, G., Matta, S., et al., 2022. Adam challenge: detecting age-related macular degeneration from fundus images. IEEE Trans. Med. Imaging. 41 (10), 2828–2847.
Fang, K., Li, W.-J., 2020. DMNet: difference minimization network for semi-supervised segmentation in medical images. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 532–541.
Farsiu, S., Chiu, S.J., O'Connell, R.V., Folgar, F.A., Yuan, E., Izatt, J.A., Toth, C.A., Group, A.-R. E. D. S. . A. S. D. O. C. T.S. et al., 2014. Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography. Ophthalmology 121 (1), 162–172.
Feng, X., Jiang, Z., Kaufmann, T., Xu, P., Hüllermeier, E., Weng, P., Zhu, Y., 2025. Duo: diverse, uncertain, on-policy query generation and selection for reinforcement learning from human feedback. In: AAAI Conf. Artif. Intell. Vol. 39, pp. 16604–16612.
Föllmer, B., Schulze, K., Wald, C., Stober, S., Samek, W., Dewey, M., 2024. Active learning with the nnUNet and sample selection with uncertainty-aware submodular mutual information measure. In: Proc. Int. Conf. Medical Imaging Deep Learn. PMLR, pp. 480–503.
Früh, M., Fischer, M., Schilling, A., Gatidis, S., Hepp, T., 2021. Weakly supervised segmentation of tumor lesions in PET-CT hybrid imaging. J. Med. Imaging 8 (5), 054003.
Fu, H., Li, F., Orlando, J.I., Bogunovic, H., Sun, X., Liao, J., Xu, Y., Zhang, S., Zhang, X., 2019. Palm: pathologic myopia challenge. IEEE Dataport.
Gaillochet, M., Desrosiers, C., Lombaert, H., 2023. Active learning for medical image segmentation with stochastic batches. Med. Image Anal. 90, 102958.
Gal, Y., Islam, R., Ghahramani, Z., 2017. Deep bayesian active learning with image data. In: Proc. Int. Conf. Mach. Learn. PMLR, pp. 1183–1192.
Gao, Z., Hong, B., Li, Y., Zhang, X., Wu, J., Wang, C., Zhang, X., Gong, T., Zheng, Y., Meng, D., et al., 2023. A semi-supervised multi-task learning framework for cancer classification with weak annotation in whole-slide images. Med. Image Anal. 83, 102652.
Ghesu, F.C., Georgescu, B., Mansoor, A., Yoo, Y., Neumann, D., Patel, P., Vishwanath, R.S., Balter, J.M., Cao, Y., Grbic, S., et al., 2022. Self-supervised learning from 100 million medical images. arXiv:2201.01283.
Gibson, E., Giganti, F., Hu, Y., Bonmati, E., Bandula, S., Gurusamy, K., Davidson, B., Pereira, S.P., Clarkson, M.J., Barratt, D.C., 2018a. Automatic multi-organ segmentation on abdominal CT with dense v-networks. IEEE Trans. Med. Imaging 37 (8), 1822–1834.
Gibson, E., Giganti, F., Hu, Y., Bonmati, E., Bandula, S., Gurusamy, K., Davidson, B., Pereira, S.P., Clarkson, M.J., Barratt, D.C., 2018b. Automatic multi-organ segmentation on abdominal CT with dense v-networks. IEEE Trans. Med. Imaging 37 (8), 1822–1834.
Gondal, W.M., Köhler, J.M., Grzeszick, R., Fink, G.A., Hirsch, M., 2017. Weakly-supervised localization of diabetic retinopathy lesions in retinal fundus images. In: 2017 IEEE International Conference on Image Processing (ICIP). IEEE, pp. 2069–2073.
Gong, S., Chen, C., Gong, Y., Chan, N.Y., Ma, W., Mak, C. H.-K., Abrigo, J., Dou, Q., 2023. Diffusion model based semi-supervised learning on brain hemorrhage images for efficient midline shift quantification. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 69–81.
Gong, S., Zhong, Y., Ma, W., Li, J., Wang, Z., Zhang, J., Heng, P.-A., Dou, Q., 2024. 3DSAM-adapter: holistic adaptation of SAM from 2D to 3D for promptable tumor segmentation. Med. Image Anal. 98, 103324.
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. Proc. Adv. Neural Inf. Process. Syst. 27. https://proceedings.neurips.cc/paper_files/paper/2014/file/f033ed80deb0234979a61f95710dbe25-Paper.pdf
Graham, S., Epstein, D., Rajpoot, N., 2020. Dense steerable filter cnns for exploiting rotational symmetry in histology images. IEEE Trans. Med. Imaging 39 (12), 4124–4136.
Grandvalet, Y., Bengio, Y., 2004. Semi-supervised learning by entropy minimization. Proc. Adv. Neural Inf. Process. Syst. 17. https://proceedings.neurips.cc/paper_files/paper/2004/file/96f2b50b5d3613adf9c27049b2a888c7-Paper.pdf
Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al., 2020. Bootstrap your own latent-a new approach to self-supervised learning. Proc. Adv. Neural Inf. Process. Syst. 33, 21271–21284.
Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al., 2025. Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv:2501.12948.
Guo, R., Xie, K., Pagnucco, M., Song, Y., 2023. SAC-Net: learning with weak and noisy labels in histopathology image segmentation. Med. Image Anal. 86, 102790.
Guo, S., Huang, W., Zhang, H., Zhuang, C., Dong, D., Scott, M.R., Huang, D., 2018. Curriculumnet: weakly supervised learning from large-scale web images. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 135–150.
Guo, Z., Ma, J., Xu, Y., Wang, Y., Wang, L., Chen, H., 2024a. Histgen: histopathology report generation via local-global feature encoding and cross-modal context interaction. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 189–199.
Guo, Z., Xiong, C., Ma, J., Sun, Q., Feng, L., Wang, J., Chen, H., 2024b. Focus: Knowledge-enhanced adaptive visual compression for few-shot whole slide image classification. arXiv:2411.14743.
Gupte, S.R., Aklilu, J., Nirschl, J.J., Yeung-Levy, S., 2024. Revisiting active learning in the era of vision foundation models. Trans. Mach. Learn. Res. https://openreview.net/forum?id=u8K83M9mbG.
Gutman, D., Codella, N. C.F., Celebi, E., Helba, B., Marchetti, M., Mishra, N., Halpern, A., 2016. Skin lesion analysis toward melanoma detection: a challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC). arXiv:1605.01397.
Gyawali, P.K., Ghimire, S., Bajracharya, P., Li, Z., Wang, L., 2020. Semi-supervised medical image classification with global latent mixing. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 604–613.
Gyawali, P.K., Li, Z., Ghimire, S., Wang, L., 2019. Semi-supervised learning by disentangling and self-ensembling over stochastic latent space. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 766–774.
Haghighi, F., Hosseinzadeh Taher, M.R., Zhou, Z., Gotway, M.B., Liang, J., 2020. Learning semantics-enriched representation via self-discovery, self-classification, and self-restoration. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 137–147.
Haghighi, F., Taher, M. R.H., Gotway, M.B., Liang, J., 2022. Dira: discriminative, restorative, and adversarial learning for self-supervised medical image analysis. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 20824–20834.
Hajeb Mohammad Alipour, S., Rabbani, H., Akhlaghi, M.R., 2012. Diabetic retinopathy grading by digital curvelet transform. Comput. Math. Methods Med. 2012. 12 (1), 761901.
Halabi, S.S., Prevedello, L.M., Kalpathy-Cramer, J., Mamonov, A.B., Bilbily, A., Cicero, M., Pan, I., Pereira, L.A., Sousa, R.T., Abdala, N., et al., 2019. The RSNA pediatric bone age machine learning challenge. Radiology 290 (2), 498.
Han, Z., Wei, B., Hong, Y., Li, T., Cong, J., Zhu, X., Wei, H., Zhang, W., 2020. Accurate screening of COVID-19 using attention-based deep 3D multiple instance learning. IEEE Trans. Med. Imaging 39 (8), 2584–2594.
Hashimoto, N., Fukushima, D., Koga, R., Takagi, Y., Ko, K., Kohno, K., Nakaguro, M., Nakamura, S., Hontani, H., Takeuchi, I., 2020. Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 3852–3861.
He, S., Nie, Y., Wang, H., Yang, S., Wang, Y., Cai, Z., Chen, Z., Xu, Y., Luo, L., Xiang, H., et al., 2024a. Gsco: towards generalizable AI in medicine via generalist-specialist collaboration. 18, 172–191. arXiv:2404.15127.
He, Y., Huang, F., Jiang, X., Nie, Y., Wang, M., Wang, J., Chen, H., 2024b. Foundation model for advancing healthcare: challenges, opportunities and future directions. IEEE Rev. Biomed. Eng. .
He, Y., Yang, G., Ge, R., Chen, Y., Coatrieux, J.-L., Wang, B., Li, S., 2023. Geometric visual similarity learning in 3D medical image self-supervised pre-training. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 9538–9547.
Hervella, Á.S., Rouco, J., Novo, J., Ortega, M., 2018. Retinal image understanding emerges from self-supervised multimodal reconstruction. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 321–328.
Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. Proc. Adv. Neural Inf. Process. Syst. 33, 6840–6851.
Holub, A., Perona, P., Burl, M.C., 2008. Entropy-based active learning for object recognition. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognit. Workshops. IEEE, pp. 1–8.
Hou, J., Ding, X., Deng, J.D., 2022a. Semi-supervised semantic segmentation of vessel images using leaking perturbations. In: Proc. IEEE Winter Conf. App. Comput. Vis., pp. 2625–2634.
Hou, J., Xiao, F., Xu, J., Feng, R., Zhang, Y., Zou, H., Lu, L., Xue, W., 2023a. Diabetic retinopathy grading with weakly-supervised lesion priors. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 1–5.
Hou, J., Xu, J., Jiang, L., Du, S., Feng, R., Zhang, Y., Shan, F., Xue, X., 2021. Periphery-aware COVID-19 diagnosis with contrastive representation enhancement. Pattern Recognit. 118, 108005.
Hou, J., Xu, J., Xiao, F., Zhang, B., Xu, Y., Zhang, Y., Zou, H., Feng, R., 2023b. Towards label-efficient deep learning for myopic maculopathy classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 31–45.
Hou, J., Xu, J., Xiao, F., Zhao, R.-W., Zhang, Y., Zou, H., Lu, L., Xue, W., Feng, R., 2022b. Cross-field transformer for diabetic retinopathy grading on two-field fundus images. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, pp. 985–990.
Hsu, C.-C., Hsu, K.-J., Tsai, C.-C., Lin, Y.-Y., Chuang, Y.-Y., 2019. Weakly supervised instance segmentation using the bounding box tightness prior. Proc. Adv. Neural Inf. Process. Syst. 32.

Hu, W., Cheng, L., Huang, G., Yuan, X., Zhong, G., Pun, C.-M., Zhou, J., Cai, M., 2023a. Learning from incorrectness: active learning with negative pre-training and curriculum querying for histological tissue classification. IEEE Trans. Med. Imaging. 43 (2), 625–637.

Hu, X., Chen, Y.-J., Ho, T.-Y., Shi, Y., 2023b. Conditional diffusion models for weakly supervised medical image segmentation. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention, pp. 756–765.

Huang, J., Qu, L., Jia, R., Zhao, B., 2019. O2u-Net: a simple noisy label detection approach for deep neural networks. In: Proc. IEEE Int. Conf. Comput. Vis., pp. 3326–3334.

Huang, Z., Jiang, R., Aeron, S., Hughes, M.C., 2024. Systematic comparison of semi-supervised and self-supervised learning for medical image classification. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 22282–22293.

Huang, Z., Yang, E., Shen, J., Gratzinger, D., Eyerer, F., Liang, B., Nirschl, J., Bingham, D., Dussaq, A.M., Kunder, C., et al., 2025. A pathologist–AI collaboration framework for enhancing diagnostic accuracies and efficiencies. Nat. Biomed. Eng. 9 (4), 455–470.

Hwang, S., Kim, H.-E., 2016. Self-transfer learning for weakly supervised lesion localization. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 239–246.

Ilse, M., Tomczak, J., Welling, M., 2018. Attention-based deep multiple instance learning. In: Proc. Int. Conf. Mach. Learn. PMLR, pp. 2127–2136.

Jia, Z., Huang, X., Eric, I., Chang, C., Xu, Y., 2017. Constrained deep weak supervision for histopathology image segmentation. IEEE Trans. Med. Imaging 36 (11), 2376–2388.

Jiang, H., Jin, C., Lin, H., Zhou, Y., Wang, X., Ma, J., Ding, L., Hou, J., Liu, R., Chai, Z., et al., 2025. Generalizable cervical cancer screening via large-scale pretraining and test-time adaptation. arXiv:2502.09662.

Jiang, Y., Sun, M., Guo, H., Bai, X., Yan, K., Lu, L., Xu, M., 2023. Anatomical invariance modeling and semantic alignment for self-supervised learning in 3D medical image analysis. In: Proc. IEEE Int. Conf. Comput. Vis., pp. 15859–15869.

Jin, C., Luo, L., Lin, H., Hou, J., Chen, H., 2024a. HMIL: hierarchical multi-instance learning for fine-grained whole slide image classification. IEEE Trans. Med. Imaging. 44 (4), 1796–1808.

Jin, R., Xu, Z., Zhong, Y., Yao, Q., QI, D., Zhou, S.K., Li, X., 2024b. FairMedFM: fairness benchmarking for medical imaging foundation models. Proc. Adv. Neural Inf. Process. Syst. 37, 111318–111357.

Johnson, A. E.W., 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. Sci. Data 6 (1), 1–8.

Kadir, M.A., Alam, H. M.T., Sonntag, D., 2023. EdgeAL: an edge estimation based active learning approach for OCT segmentation. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 79–89.

Kamran, S.A., Hossain, K.F., Tavakkoli, A., Zuckerbrod, S.L., Baker, S.A., 2021. VTGAN: semi-supervised retinal image synthesis and disease prediction using vision transformers. In: Proc. IEEE Int. Conf. Comput. Vis., pp. 3235–3245.

Kandemir, M., Zhang, C., Hamprecht, F.A., 2014. Empowering multiple instance histopathology cancer diagnosis by cell graphs. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 228–235.

Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.-A., Gaiser, T., Marx, A., Valous, N.A., Ferber, D., et al., 2019. Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. PLoS Med. 16 (1), e1002730.

Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.-H., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., Özkan, S., et al., 2021. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. Med. Image Anal. 69, 101950.

Keikhosravi, A., Li, B., Liu, Y., Conklin, M.W., Loeffler, A.G., Eliceiri, K.W., 2020. Non-disruptive collagen characterization in clinical histopathology using cross-modality image synthesis. Commun. Biol. 3 (1), 1–12.

Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C. C.S., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al., 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 172 (5), 1122–1131.

Kervadec, H., Bouchtiba, J., Desrosiers, C., Granger, E., Dolz, J., Ayed, I.B., 2021. Boundary loss for highly unbalanced segmentation. Med. Image Anal. 67, 101851.

Khan, S., Shahin, A.H., Villafruela, J., Shen, J., Shao, L., 2019. Extreme points derived confidence map as a cue for class-agnostic interactive segmentation using deep neural network. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 66–73.

Kingma, D.P., Welling, M., 2013. Auto-encoding variational bayes. arXiv:1312.6114.

Kise, K., Sato, A., Iwata, M., 1998. Segmentation of page images using the area voronoi diagram. Comput. Vision Image Understanding 70 (3), 370–382.

Koohbanani, N.A., Unnikrishnan, B., Khurram, S.A., Krishnaswamy, P., Rajpoot, N., 2021. Self-path: self-supervision for classification of pathology images with limited annotations. IEEE Trans. Med. Imaging 40 (10), 2845–2856.

Korkmaz, Y., Cukur, T., Patel, V.M., 2023. Self-supervised MRI reconstruction with unrolled diffusion models. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 491–501.

Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., Sethi, A., 2017. A dataset and a technique for generalized nuclear segmentation for computational pathology. IEEE Trans. Med. Imaging 36 (7), 1550–1560.

Laine, S., Aila, T., 2016. Temporal ensembling for semi-supervised learning. arXiv:1610.02242.

Lee, D.-H., et al., 2013. Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on Challenges in Representation Learning, ICML. Vol. 3, p. 896.

Lee, H., Jeong, W.-K., 2020. Scribble2label: scribble-supervised cell segmentation via self-generating pseudo-labels with consistency. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 14–23.

Lee, R.S., Gimenez, F., Hoogi, A., Miyake, K.K., Gorovoy, M., Rubin, D.L., 2017. A curated mammography data set for use in computer-aided detection and diagnosis research. Sci. Data 4 (1), 1–9.

Lei, T., Zhang, D., Du, X., Wang, X., Wan, Y., Nandi, A.K., 2022. Semi-supervised medical image segmentation using adversarial consistency learning and dynamic convolution network. IEEE Trans. Med. Imaging. 42 (5), 1265–1277.

Lei, W., Su, Q., Jiang, T., Gu, R., Wang, N., Liu, X., Wang, G., Zhang, X., Zhang, S., 2023. One-shot weakly-supervised segmentation in 3D medical images. IEEE Trans. Med. Imaging. 43 (1), 175–189.

Li, B., Keikhosravi, A., Loeffler, A.G., Eliceiri, K.W., 2021a. Single image super-resolution for whole slide image using convolutional neural networks and self-supervised color normalization. Med. Image Anal. 68, 101938.

Li, B., Li, Y., Eliceiri, K.W., 2021b. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 14318–14328.

Li, D., Yang, J., Kreis, K., Torralba, A., Fidler, S., 2021c. Semantic segmentation with generative models: semi-supervised learning and strong out-of-domain generalization. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 8300–8311.

Li, H., Liu, H., Fu, H., Xu, Y., Shu, H., Niu, K., Hu, Y., Liu, J., 2023a. A generic fundus image enhancement network boosted by frequency self-supervised representation learning. Med. Image Anal. 90, 102945.

Li, H., Yin, Z., 2020. Attention, suggestion and annotation: a deep active learning framework for biomedical image segmentation. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 3–13.

Li, J., Chen, Y., Chu, H., Sun, Q., Guan, T., Han, A., He, Y., 2024a. Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 11323–11332.

Li, J., Fragemann, J., Ahmadi, S.-A., Kleesiek, J., Egger, J., 2022a. Training β-VAE by aggregating a learned gaussian posterior with a decoupled decoder. In: MICCAI Workshop on Medical Applications with Disentanglements. Springer, pp. 70–92.

Li, M., Hsu, W., Xie, X., Cong, J., Gao, W., 2020a. SACNN: self-attention convolutional neural network for low-dose CT denoising with self-supervised perceptual loss network. IEEE Trans. Med. Imaging 39 (7), 2289–2301.

Li, S., Zhang, C., He, X., 2020b. Shape-aware semi-supervised 3D semantic segmentation for medical images. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 552–561.

Li, W., Li, J., Wang, Z., Polson, J., Sisk, A.E., Sajed, D.P., Speier, W., Arnold, C.W., 2021d. PathAL: an active learning framework for histopathology image analysis. IEEE Trans. Med. Imaging. 41 (1), 1176–1187.

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

Li, X., Hu, X., Qi, X., Yu, L., Zhao, W., Heng, P.-A., Xing, L., 2021e. Rotation-oriented collaborative self-supervised learning for retinal disease diagnosis. IEEE Trans. Med. Imaging 40 (9), 2284–2294.

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

Li, X., Xia, M., Jiao, J., Zhou, S., Chang, C., Wang, Y., Guo, Y., 2023b. HAL-IA: a hybrid active learning framework using interactive annotation for medical image segmentation. Med. Image Anal. 88, 102862.

Li, X., Yu, L., Chen, H., Fu, C.-W., Heng, P.-A., 2018. Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model. arXiv:1808.03887.

Li, X., Yu, L., Chen, H., Fu, C.-W., Xing, L., Heng, P.-A., 2020e. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. IEEE Trans. Neural Netw. Learn. Syst. 32 (2), 523–534.

Li, Y., Liu, Y., Huang, L., Wang, Z., Luo, J., 2022b. Deep weakly-supervised breast tumor segmentation in ultrasound images with explicit anatomical constraints. Med. Image Anal. 76, 102315.

Li, Z., Zhang, N., Gong, H., Qiu, R., Zhang, W., 2024b. SG-MIAN: self-guided multiple information aggregation network for image-level weakly supervised skin lesion segmentation. Comput. Biol. Med. 170, 107988.

Li, Z., Zhao, W., Shi, F., Qi, L., Xie, X., Wei, Y., Ding, Z., Gao, Y., Wu, S., Liu, J., et al., 2021f. A novel multiple instance learning framework for COVID-19 severity assessment via data augmentation and self-supervised learning. Med. Image Anal. 69, 101978.

Lin, Y., Liu, L., Ma, K., Zheng, Y., 2021. Seg4reg+: consistency learning between spine segmentation and cobb angle regression. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 490–499.

Lin, Y., Qu, Z., Chen, H., Gao, Z., Li, Y., Xia, L., Ma, K., Zheng, Y., Cheng, K.-T., 2023. Nuclei segmentation with point annotations from pathology images via self-supervised learning and co-training. Med. Image Anal. 89, 102933.

Lin, Y., Zhou, H.-Y., Ma, K., Yang, X., Zheng, Y., 2019. Seg4reg networks for automated spinal curvature estimation. In: International Workshop and Challenge on Computational Methods and Clinical Applications for Spine Imaging. Springer, pp. 69–74.

Litjens, G., Toth, R., van de Ven, W., Hoeks, C., Kerkstra, S., van Ginneken, B., Vincent, G., Guillard, G., Birbeck, N., Zhang, J., et al., 2014. Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. Med. Image Anal. 18 (2), 359–373.

Liu, F., Tian, Y., Chen, Y., Liu, Y., Belagiannis, V., Carneiro, G., 2022. ACPL: anti-curriculum pseudo-labelling for semi-supervised medical image classification. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 20697–20706.

Liu, J., Zhang, Y., Chen, J.-N., Xiao, J., Lu, Y., Bennett, A.L., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z., 2023a. Clip-driven universal model for organ segmentation and tumor detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 21152–21164.

Liu, M., Zhang, J., Adeli, E., Shen, D., 2018. Landmark-based deep multi-instance learning for brain disease diagnosis. Med. Image Anal. 43, 157–168.
Liu, P., Ji, L., Ye, F., Fu, B., 2024a. AdvMIL: adversarial multiple instance learning for the survival analysis on whole-slide images. Med. Image Anal. 91, 103020.
Liu, Q., 2017. Stein variational gradient descent as gradient flow. Proc. Adv. Neural Inf. Process. Syst. 30. https://proceedings.neurips.cc/paper_files/paper/2017/file/17ed8abedc255908be746d245e50263a-Paper.pdf
Liu, Q., Yu, L., Luo, L., Dou, Q., Heng, P.A., 2020. Semi-supervised medical image classification with relation-driven self-ensembling model. IEEE Trans. Med. Imaging 39 (11), 3429–3440.
Liu, X., Li, W., Yuan, Y., 2024b. Diffrect: latent diffusion label rectification for semi-supervised medical image segmentation. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 56–66.
Liu, X., Liu, Q., Zhang, Y., Wang, M., Tang, J., 2023b. TSSK-Net: weakly supervised biomarker localization and segmentation with image-level annotation in retinal OCT images. Comput. Biol. Med. 153, 106467.
Liu, Z., He, X., Wang, H., Xiong, H., Zhang, Y., Wang, G., Hao, J., Feng, Y., Zhu, F., Hu, H., 2023c. Hierarchical self-supervised learning for 3D tooth segmentation in intra-oral mesh scans. IEEE Trans. Med. Imaging 42 (2), 467–480.
Lu, M.Y., Chen, B., Williamson, D. F.K., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al., 2024. A visual-language foundation model for computational pathology. Nat. Med. 30 (3), 863–874.
Lu, M.Y., Chen, B., Zhang, A., Williamson, D. F.K., Chen, R.J., Ding, T., Le, L.P., Chuang, Y.-S., Mahmood, F., 2023. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 19764–19775.
Lu, M.Y., Chen, R.J., Kong, D., Lipkova, J., Singh, R., Williamson, D. F.K., Chen, T.Y., Mahmood, F., 2022. Federated learning for computational pathology on gigapixel whole slide images. Med. Image Anal. 76, 102298.
Lu, M.Y., Williamson, D. F.K., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F., 2021a. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat. Biomed. Eng. 5 (6), 555–570.
Lu, Q., Li, Y., Ye, C., 2021b. Volumetric white matter tract segmentation with nested self-supervised learning using sequential pretext tasks. Med. Image Anal. 72, 102094.
Lu, Z.H., Chen, J., 2020. National pathology quality report in 2019. Chin. J. Pathol. 49 (7), 667–669.
Ludovic, R., Daniel, R., Nicolas, L., Maria, K., Humayun, I., Jacques, K., Frédérique, C., Catherine, G., et al., 2013. Mitosis detection in breast cancer histological images an ICPR 2012 contest. J. Pathol. Inf. 4 (1), 8.
Luo, L., Yu, L., Chen, H., Liu, Q., Wang, X., Xu, J., Heng, P.-A., 2020. Deep mining external imperfect data for chest x-ray disease screening. TMI 39 (11), 3583–3594.
Luo, X., Chen, J., Song, T., Wang, G., 2021a. Semi-supervised medical image segmentation through dual-task consistency. In: AAAI Conf. Artif. Intell. Vol. 35, pp. 8801–8809.
Luo, X., Hu, M., Song, T., Wang, G., Zhang, S., 2022. Semi-supervised medical image segmentation via cross teaching between cnn and transformer. In: Proc. Int. Conf. Medical Imaging Deep Learn. PMLR, pp. 820–833.
Luo, X., Liao, W., Chen, J., Song, T., Chen, Y., Zhang, S., Chen, N., Wang, G., Zhang, S., 2021b. Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 318–329.
Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B., 2024a. Segment anything in medical images. Nat. Commun. 15 (1), 654.
Ma, J., Yang, Z., Kim, S., Chen, B., Baharoon, M., Fallahpour, A., Asakereh, R., Lyu, H., Wang, B., 2025. MedSAM2: segment anything in 3D medical images and videos. arXiv:2504.03600.
Ma, Q., Zhang, J., Qi, L., Yu, Q., Shi, Y., Gao, Y., 2024b. Constructing and exploring intermediate domains in mixed domain semi-supervised medical image segmentation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 11642–11651.
Madani, A., Moradi, M., Karargyris, A., Syeda-Mahmood, T., 2018. Semi-supervised learning with generative adversarial networks for chest x-ray classification with ability of data domain adaptation. In: Proc. IEEE Int. Symp. Biomed. Imaging. IEEE, pp. 1038–1042.
Mahapatra, D., Bozorgtabar, B., Ge, Z., Reyes, M., 2024a. Gandalf: graph-based transformer and data augmentation active learning framework with interpretable features for multi-label chest xray classification. Med. Image Anal. 93, 103075.
Mahapatra, D., Bozorgtabar, B., Thiran, J.-P., Reyes, M., 2018. Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 580–588.
Mahapatra, D., Ge, Z., Reyes, M., 2022. Self-supervised generalized zero shot learning for medical image classification using novel interpretable saliency maps. IEEE Trans. Med. Imaging. 41 (9), 2443–2456.
Mahapatra, D., Tennakoon, R., George, Y., Roy, S., Bozorgtabar, B., Ge, Z., Reyes, M., 2024b. Alfredo: active learning with feature disentangelement and domain adaptation for medical image classification. Med. Image Anal. 97, 103261.
Makropoulos, A., Robinson, E.C., Schuh, A., Wright, R., Fitzgibbon, S., Bozek, J., Counsell, S.J., Steinweg, J., Vecchiato, K., Passerat-Palmbach, J., et al., 2018. The developing human connectome project: a minimal processing pipeline for neonatal cortical surface reconstruction. Neuroimage 173, 88–112.
Manivannan, S., Cobb, C., Burgess, S., Trucco, E., 2017. Subcategory classifiers for multiple-instance learning and its application to retinal nerve fiber layer visibility classification. IEEE Trans. Med. Imaging 36 (5), 1140–1150.
Maška, M., Ulman, V., Svoboda, D., Matula, P., Matula, P., Ederra, C., Urbiola, A., España, T., Venkatesan, S., Balak, D. M.W., et al., 2014. A benchmark for comparison of cell tracking algorithms. Bioinformatics 30 (11), 1609–1617.
Mayilvahanan, P., Zimmermann, R.S., Wiedemer, T., Rusak, E., Juhos, A., Bethge, M., Brendel, W., 2025. In search of forgotten domain generalization. The Thirteenth International Conference on Learning Representations https://openreview.net/forum?id=Fk3eod9aaD
Meng, Y., Zhang, H., Zhao, Y., Gao, D., Hamill, B., Patri, G., Peto, T., Madhusudhan, S., Zheng, Y., 2022. Dual consistency enabled weakly and semi-supervised optic disc and cup segmentation with dual adaptive graph convolutional networks. IEEE Trans. Med. Imaging 42 (2), 416–429.
Meng, Y., Zhang, H., Zhao, Y., Gao, D., Hamill, B., Patri, G., Peto, T., Madhusudhan, S., Zheng, Y., 2023. Dual consistency enabled weakly and semi-supervised optic disc and cup segmentation with dual adaptive graph convolutional networks. IEEE Trans. Med. Imaging 42 (2), 416–429.
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.
Mercan, C., Aksoy, S., Mercan, E., Shapiro, L.G., Weaver, D.L., Elmore, J.G., 2017. Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images. IEEE Trans. Med. Imaging 37 (1), 316–325.
Mescheder, L., Geiger, A., Nowozin, S., 2018. Which training methods for GANs do actually converge? In: International Conference on Machine Learning. PMLR, pp. 3481–3490.
Miao, J., Chen, C., Liu, F., Wei, H., Heng, P.-A., 2023. Caussl: causality-inspired semi-supervised learning for medical image segmentation. In: Proc. IEEE Int. Conf. Comput. Vis., pp. 21426–21437.
Miao, J., Chen, C., Zhang, K., Chuai, J., Li, Q., Heng, P.-A., 2024. Cross prompting consistency with segment anything model for semi-supervised medical image segmentation. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 167–177.
Milton, M. A.A., 2019. Automated skin lesion classification using ensemble of deep neural networks in ISIC 2018: skin lesion analysis towards melanoma detection challenge. arXiv:1901.10802.
Munjal, P., Hayat, N., Hayat, M., Sourati, J., Khan, S., 2022. Towards robust and reproducible active learning using neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 223–232.
Naik, N., Madani, A., Esteva, A., Keskar, N.S., Press, M.F., Ruderman, D., Agus, D.B., Socher, R., 2020. Deep learning-enabled breast cancer hormonal receptor status determination from base-level h&e stains. Nat. Commun. 11 (1), 1–8.
Nath, V., Yang, D., Landman, B.A., Xu, D., Roth, H.R., 2021. Diminishing uncertainty within the training pool: active learning for medical image segmentation. IEEE Trans. Med. Imaging 40 (10), 2534–2547.
Naylor, P., Laé, M., Reyal, F., Walter, T., 2018. Segmentation of nuclei in histopathology images by deep regression of the distance map. IEEE Trans. Med. Imaging 38 (2), 448–459.
Nguyen, D. M.H., Nguyen, H., Diep, N.T., Pham, T.N., Cao, T., Nguyen, B.T., Swoboda, P., Ho, N., Albarqouni, S., Xie, P., et al., 2023. LVM-MED: Learning large-scale self-supervised vision models for medical imaging via second-order graph matching. arXiv:2306.11925.
Nguyen, H.-G., Pica, A., Hrbacek, J., Weber, D.C., La Rosa, F., Schalenbourg, A., Sznitman, R., Cuadra, M.B., 2019. A novel segmentation framework for uveal melanoma in magnetic resonance imaging based on class activation maps. In: Proc. Int. Conf. Medical Imaging Deep Learn. PMLR, pp. 370–379.
Nie, D., Gao, Y., Wang, L., Shen, D., 2018. ASDNet: attention based semi-supervised deep networks for medical image segmentation. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 370–378.
Nie, X., Liu, L., He, L., Zhao, L., Lu, H., Lou, S., Xiong, R., Wang, Y., 2023. Weakly-interactive-mixed learning: less labelling cost for better medical image segmentation. IEEE J. Biomed. Health. Inf. 27, 3270–3281.
Nie, Y., He, S., Bie, Y., Wang, Y., Chen, Z., Yang, S., Cai, Z., Wang, H., Wang, X., Luo, L., et al., 2025. An explainable biomedical foundation model via large-scale concept-enhanced vision-language pre-training. arXiv:2501.15579.
Odena, A., 2016. Semi-supervised learning with generative adversarial networks. arXiv:1606.01583.
Oken, M.M., Hocking, W.G., Kvale, P.A., Andriole, G.L., Buys, S.S., Church, T.R., Crawford, E.D., Fouad, M.N., Isaacs, C., Reding, D.J., et al., 2011. Screening by chest radiograph and lung cancer mortality: the prostate, lung, colorectal, and ovarian (PLCO) randomized trial. Jama 306 (17), 1865–1873.
Orlando, J.I., Fu, H., Breda, J.B., van Keer, K., Bathula, D.R., Diaz-Pinto, A., Fang, R., Heng, P.-A., Kim, J., Lee, J., et al., 2020. Refuge challenge: a unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Med. Image Anal. 59, 101570.
Ouali, Y., Hudelot, C., Tami, M., 2020. An overview of deep semi-supervised learning. arXiv:2006.05278.
Ozdemir, F., Peng, Z., Fuernstahl, P., Tanner, C., Goksel, O., 2021. Active learning for segmentation based on Bayesian sample queries. Knowl. Based Syst. 214, 106531.
Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., et al., 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. bmj 372.
Paranjape, J.N., Sikder, S., Vedula, S.S., Patel, V.M., 2024. S-Sam: SVD-based fine-tuning of segment anything model for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 720–730.
Park, S., Lee, E.S., Shin, K.S., Lee, J.E., Ye, J.C., 2023. Self-supervised multi-modal training from uncurated images and reports enables monitoring AI in radiology. Med. Image Anal. 91, 103021.
Peiris, H., Hayat, M., Chen, Z., Egan, G., Harandi, M., 2023. Uncertainty-guided dual-views for semi-supervised volumetric medical image segmentation. Nat. Mach. Intell. 5 (7), 724–738.

Peng, J., Wang, P., Pedersoli, M., Desrosiers, C., 2024. Boundary-aware information maximization for self-supervised medical image segmentation. Med. Image Anal. 94, 103150.
Perez-Garcia, F., Sharma, H., Bond-Taylor, S., Bouzid, K., Salvatelli, V., Ilse, M., Bannur, S., Castro, D.C., Schwaighofer, A., Lungren, M.P., et al., 2025. Exploring scalable medical image encoders beyond text supervision. Nat. Mach. Intell. 7 (1), 119–130.
Purma, V., Srinath, S., Srirangarajan, S., Kakkar, A., Prathosh, A.P., 2024. Genselfdiff-HIS: generative self-supervision using diffusion for histopathological image segmentation. IEEE Trans. Med. Imaging. 44 (2), 618–631.
Qu, H., Wu, P., Huang, Q., Yi, J., Riedlinger, G.M., De, S., Metaxas, D.N., 2019. Weakly supervised deep nuclei segmentation using points annotation in histopathology images. In: Proc. Int. Conf. Medical Imaging Deep Learn. PMLR, pp. 390–400.
Qu, H., Wu, P., Huang, Q., Yi, J., Yan, Z., Li, K., Riedlinger, G.M., De, S., Zhang, S., Metaxas, D.N., 2020. Weakly supervised deep nuclei segmentation using partial points annotation in histopathology images. IEEE Trans. Med. Imaging 39 (11), 3655–3666.
Qu, L., Ma, Y., Yang, Z., Wang, M., Song, Z., 2023. Openal: an efficient deep active learning framework for open-set pathology image classification. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 3–13.
Qu, L., Yang, D., Huang, D., Guo, Q., Luo, R., Zhang, S., Wang, X., 2024. Pathology-knowledge enhanced multi-instance prompt learning for few-shot whole slide image classification. In: Proc. Eur. Conf. Comput. Vis. Springer, pp. 196–212.
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.
Rajchl, M., Lee, M. C.H., Oktay, O., Kamnitsas, K., Passerat-Palmbach, J., Bai, W., Damodaram, M., Rutherford, M.A., Hajnal, J.V., Kainz, B., et al., 2016. Deepcut: object segmentation from bounding box annotations using convolutional neural networks. IEEE Trans. Med. Imaging 36 (2), 674–683.
Raju, A., Yao, J., Haq, M.M., Jonnagaddala, J., Huang, J., 2020. Graph attention multi-instance learning for accurate colorectal cancer staging. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 529–539.
Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H.R., Albarqouni, S., Bakas, S., Galtier, M.N., Landman, B.A., Maier-Hein, K., et al., 2020. The future of digital health with federated learning. NPJ Digit. Med. 3 (1), 1–7.
Rister, B., Yi, D., Shivakumar, K., Nobashi, T., Rubin, D.L., 2020. CT-org, a new dataset for multiple organ segmentation in computed tomography. Sci. Data 7 (1), 1–9.
Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 10684–10695.
Rosenkrantz, A.B., Hughes, D.R., Duszak, , R.Jr, 2016. The US radiologist workforce: an analysis of temporal and geographic variation by using large national datasets. Radiology 279 (1), 175–184.
Roth, H.R., Lu, L., Farag, A., Shin, H.-C., Liu, J., Turkbey, E.B., Summers, R.M., 2015. DeeporGAN: multi-level deep convolutional networks for automated pancreas segmentation. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 556–564.
Roth, H.R., Yang, D., Xu, Z., Wang, X., Xu, D., 2021. Going to extremes: weakly supervised medical image segmentation. Mach. Learn. Knowl. Extr. 3 (2), 507–524.
Sahasrabudhe, M., Christodoulidis, S., Salgado, R., Michiels, S., Loi, S., André, F., Paragios, N., Vakalopoulou, M., 2020. Self-supervised nuclei segmentation in histopathological images using attention. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 393–402.
Sajjadi, M., Javanmardi, M., Tasdizen, T., 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. Proc. Adv. Neural Inf. Process. Syst. 29.
Schirris, Y., Gavves, E., Nederlof, I., Horlings, H.M., Teuwen, J., 2022. DeepSMILE: contrastive self-supervised pre-training benefits MSI and HRD classification directly from h&e whole-slide images in colorectal and breast cancer. Med. Image Anal. 79, 102464.
Schmidt, A., Morales-Álvarez, P., Cooper, L. A.D., Newberg, L.A., Enquobahrie, A., Molina, R., Katsaggelos, A.K., 2024. Focused active learning for histopathological image classification. Med. Image Anal. 95, 103162.
Schwab, E., Gooßen, A., Deshpande, H., Saalbach, A., 2020. Localization of critical findings in chest x-ray without local annotations using multi-instance learning. In: Proc. IEEE Int. Symp. Biomed. Imaging. IEEE, pp. 1879–1882.
Sedai, S., Mahapatra, D., Hewavitharanage, S., Maetschke, S., Garnavi, R., 2017. Semi-supervised segmentation of optic cup in retinal fundus images using variational autoencoder. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 75–82.
Seff, A., Lu, L., Barbu, A., Roth, H., Shin, H.-C., Summers, R.M., 2015. Leveraging mid-level semantic boundary cues for automated lymph node detection. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 53–61.
Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al., 2021. TransMIL: transformer based correlated multiple instance learning for whole slide image classification. Proc. Adv. Neural Inf. Process. Syst. 34, 2136–2147.
Shen, H., Tian, K., Dong, P., Zhang, J., Yan, K., Che, S., Yao, J., Luo, P., Han, X., 2020. Deep active learning for breast cancer segmentation on immunohistochemistry images. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 509–518.
Shen, Y., Li, J., Shao, X., Inigo Romillo, B., Jindal, A., Dreizin, D., Unberath, M., 2024. FastSAM3D: an efficient segment anything model for 3D volumetric medical images. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention, pp. 542–552.
Shen, Y., Wu, N., Phang, J., Park, J., Liu, K., Tyagi, S., Heacock, L., Kim, S.G., Moy, L., Cho, K., et al., 2021. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. Med. Image Anal. 68, 101908.
Shi, J., Tang, L., Gao, Z., Li, Y., Wang, C., Gong, T., Li, C., Fu, H., 2023a. MG-trans: multi-scale graph transformer with information bottleneck for whole slide image classification. IEEE Trans. Med. Imaging. 46 (12), 3871–3883.
Shi, J., Tang, L., Li, Y., Zhang, X., Gao, Z., Zheng, Y., Wang, C., Gong, T., Li, C., 2023b. A structure-aware hierarchical graph-based multiple instance learning framework for PT staging in histopathological image. IEEE Trans. Med. Imaging. 42 (10), 3000–3011.
Shi, X., Su, H., Xing, F., Liang, Y., Qu, G., Yang, L., 2020. Graph temporal ensembling based semi-supervised convolutional neural network with noisy labels for histopathology image analysis. Med. Image Anal. 60, 101624.
Shi, Y., Wang, H., Ji, H., Liu, H., Li, Y., He, N., Wei, D., Huang, Y., Dai, Q., Wu, J., et al., 2023c. A deep weakly semi-supervised framework for endoscopic lesion segmentation. Med. Image Anal. 90, 102973.
Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K.-i., Matsui, M., Fujita, H., Kodera, Y., Doi, K., 2000. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. AJR Am. J. Roentgenol. 174 (1), 71–74.
Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al., 2019. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv:1902.09063.
Sirinukunwattana, K., Raza, S. E.A., Tsang, Y.-W., Snead, D. R.J., Cree, I.A., Rajpoot, N.M., 2016. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. IEEE Trans. Med. Imaging 35 (5), 1196–1206.
Sirinukunwattana, K., Snead, D. R.J., Rajpoot, N.M., 2015. A stochastic polygons model for glandular structures in colon histology images. IEEE Trans. Med. Imaging 34 (11), 2366–2378.
Sivaswamy, J., Krishnadas, S.R., Joshi, G.D., Jain, M., Tabish, A. U.S., 2014. Drishti-GS: retinal image dataset for optic nerve head (onh) segmentation. In: Proc. IEEE Int. Symp. Biomed. Imaging. IEEE, pp. 53–56.
Sonn, G.A., Natarajan, S., Margolis, D. J.A., MacAiran, M., Lieu, P., Huang, J., Dorey, F.J., Marks, L.S., 2013. Targeted biopsy in the detection of prostate cancer using an office based magnetic resonance ultrasound fusion device. J. Urol. 189 (1), 86–92.
Spitzer, H., Kiwitz, K., Amunts, K., Harmeling, S., Dickscheid, T., 2018. Improving cytoarchitectonic segmentation of human brain areas with self-supervised siamese networks. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 663–671.
Srinidhi, C.L., Kim, S.W., Chen, F.-D., Martel, A.L., 2022. Self-supervised driven consistency training for annotation efficient histopathology image analysis. Med. Image Anal. 75, 102256.
Su, J., Luo, Z., Lian, S., Lin, D., Li, S., 2024. Mutual learning with reliable pseudo label for semi-supervised medical image segmentation. Med. Image Anal. 94, 103111.
Su, Z., Tavolara, T.E., Carreno-Galeano, G., Lee, S.J., Gurcan, M.N., Niazi, M., 2022. Attention2majority: weak multiple instance learning for regenerative kidney grading on whole slide images. Med. Image Anal. 79, 102462.
Suckling, J., Parker, J., Dance, D., Astley, S., Hutt, I., Boggis, C., Ricketts, I., Stamatakis, E., Cerneaz, N., Kok, S., et al., 2015. Mammographic image analysis society (mias) database v1. 21. https://cir.nii.ac.jp/crid/1883680965782533376
Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J.N., Wu, Z., Ding, X., 2020. Embracing imperfect datasets: a review of deep learning solutions for medical image segmentation. Med. Image Anal. 63, 101693.
Taleb, A., Lippert, C., Klein, T., Nabi, M., 2021. Multimodal self-supervised learning for medical image analysis. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 661–673.
Taleb, A., Loetzsch, W., Danz, N., Severin, J., Gaertner, T., Bergner, B., Lippert, C., 2020. 3D self-supervised methods for medical imaging. Proc. Adv. Neural Inf. Process. Syst. 33, 18158–18172.
Tang, K., Jiang, Z., Wu, K., Shi, J., Xie, F., Wang, W., Wu, H., Zheng, Y., 2024a. Self-supervised representation distribution learning for reliable data augmentation in histopathology WSI classification. IEEE Trans. Med. Imaging. 44 (1), 462–474.
Tang, W., Zhou, F., Huang, S., Zhu, X., Zhang, Y., Liu, B., 2024b. Feature re-embedding: towards foundation model-level performance in computational pathology. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 11343–11352.
Tang, Y., Hu, Y., Li, J., Lin, H., Xu, X., Huang, K., Lin, H., 2023. PLD-AL: pseudo-label divergence-based active learning in carotid intima-media segmentation for ultrasound images. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 57–67.
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: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 20730–20740.
Tanno, R., Barrett, D. G.T., Sellergren, A., Ghaisas, S., Dathathri, S., See, A., Welbl, J., Lau, C., Tu, T., Azizi, S., et al., 2025. Collaboration between clinicians and vision–language models in radiology report generation. Nat. Med. 31 (2), 599–608.
Tao, X., Li, Y., Zhou, W., Ma, K., Zheng, Y., 2020. Revisiting Rubik's cube: self-supervised learning with volume-wise transformation for 3D medical image segmentation. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 238–248.
Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. Proc. Adv. Neural Inf. Process. Syst. 30.
Tennakoon, R., Bortsova, G., Ørting, S., Gostar, A.K., Wille, M. M.W., Saghir, Z., Hoseinnezhad, R., de Bruijne, M., Bab-Hadiashar, A., 2019. Classification of volumetric images using multi-instance learning and extreme value theorem. IEEE Trans. Med. Imaging 39 (4), 854–865.

Teterwak, P., Saito, K., Tsiligkaridis, T., Plummer, B.A., Saenko, K., 2025. Is large-scale pretraining the secret to good domain generalization? https://openreview.net/forum?id=wCOJpXm0Me

Tian, K., Zhang, J., Shen, H., Yan, K., Dong, P., Yao, J., Che, S., Luo, P., Han, X., 2020. Weakly-supervised nucleus segmentation based on point annotations: a coarse-to-fine self-stimulated learning strategy. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 299–308.

Tian, Z., Liu, L., Zhang, Z., Fei, B., 2015. Superpixel-based segmentation for 3D prostate MR images. IEEE Trans. Med. Imaging 35 (3), 791–801.

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. 6 (12), 1399–1406.

Tomita, N., Abdollahi, B., Wei, J., Ren, B., Suriawinata, A., Hassanpour, S., 2019. Attention-based deep neural networks for detection of cancerous and precancerous esophagus tissue on histopathological slides. JAMA Netw. Open 2 (11), e1914645–e1914645.

Troester, M.A., Sun, X., Allott, E.H., Geradts, J., Cohen, S.M., Tse, C.-K., Kirk, E.L., Thorne, L.B., Mathews, M., Li, Y., et al., 2018. Racial differences in PAM50 subtypes in the carolina breast cancer study. J. Natl. Cancer Inst. 110 (2), 176–182.

Tschandl, P., Rosendahl, C., Kittler, H., 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data 5 (1), 1–9.

Ulrich, C., Isensee, F., Wald, T., Zenk, M., Baumgartner, M., Maier-Hein, K.H., 2023. Multitalent: a multi-dataset approach to medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 648–658.

Valvano, G., Leo, A., Tsaftaris, S.A., 2021. Learning to segment from scribbles using multi-scale adversarial attention gates. IEEE Trans. Med. Imaging 40 (8), 1990–2001.

van den, O.A., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. arXiv:1807.03748.

Van Engelen, J.E., Hoos, H.H., 2020. A survey on semi-supervised learning. Mach. Learn. 109 (2), 373–440.

Veta, M., Heng, Y.J., Stathonikos, N., Bejnordi, B.E., Beca, F., Wollmann, T., Rohr, K., Shah, M.A., Wang, D., Rousson, M., et al., 2019. Predicting breast tumor proliferation from whole-slide images: the TUPAC16 challenge. Med. Image Anal. 54, 111–121.

Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Severson, K., Zimmermann, E., Hall, J., Tenenholtz, N., Fusi, N., et al., 2024. A foundation model for clinical-grade computational pathology and rare cancers detection. Nat. Med. 30 (10), 2924–2935.

Vu, Q.D., Graham, S., Kurc, T., To, M. N.N., Shaban, M., Qaiser, T., Koohbanani, N.A., Khurram, S.A., Kalpathy-Cramer, J., Zhao, T., et al., 2019. Methods for segmentation and classification of digital microscopy tissue images. Front. Bioeng. Biotechnol. 7, 53.

Wang, D., Zhang, Y., Zhang, K., Wang, L., 2020a. FocalMix: semi-supervised learning for 3D medical image detection. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 3951–3960.

Wang, G., Li, W., Zuluaga, M.A., Pratt, R., Patel, P.A., Aertsen, M., Doel, T., David, A.L., Deprest, J., Ourselin, S., et al., 2018a. Interactive medical image segmentation using deep learning with image-specific fine tuning. IEEE Trans. Med. Imaging 37 (7), 1562–1573.

Wang, G., Ye, J., Cheng, J., Li, T., Chen, Z., Cai, J., He, J., Zhuang, B., 2024a. SAM-MED3D-MOE: towards a non-forgetting segment anything model via mixture of experts for 3D medical image segmentation. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer Nature Switzerland, pp. 552–561.

Wang, G., Zhai, S., Lasio, G., Zhang, B., Yi, B., Chen, S., Macvittie, T.J., Metaxas, D., Zhou, J., Zhang, S., 2021a. Semi-supervised segmentation of radiation-induced pulmonary fibrosis from lung CT scans with multi-scale guided dense attention. IEEE Trans. Med. Imaging 41 (3), 531–542.

Wang, H., Chen, J., Zhang, S., He, Y., Xu, J., Wu, M., He, J., Liao, W., Luo, X., 2024b. Dual-reference source-free active domain adaptation for nasopharyngeal carcinoma tumor segmentation across multiple hospitals. IEEE Trans. Med. Imaging. 43 (12), 4078–4090.

Wang, H., Jin, Q., Li, S., Liu, S., Wang, M., Song, Z., 2024c. A comprehensive survey on deep active learning in medical image analysis. Med. Image Anal. 95, 103201.

Wang, H., Yi, F., Wang, J., Yi, Z., Zhang, H., 2022a. RECISTSup: weakly-supervised lesion volume segmentation using recist measurement. IEEE Trans. Med. Imaging. 41 (7), 1849–1861.

Wang, H., Zhao, S., Qiang, Z., Li, Z., Liu, C., Xi, N., Du, Y., Qin, B., Liu, T., 2025a. Knowledge-tuning large language models with structured medical knowledge bases for trustworthy response generation in chinese. ACM Trans. Knowl. Discov. Data 19 (2), 1–17.

Wang, J., Lukasiewicz, T., 2022. Rethinking bayesian deep learning methods for semi-supervised volumetric medical image segmentation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 182–190.

Wang, J., Qiao, L., Zhou, S., Zhou, J., Wang, J., Li, J., Ying, S., Chang, C., Shi, J., 2024d. Weakly supervised lesion detection and diagnosis for breast cancers with partially annotated ultrasound images. IEEE Trans. Med. Imaging 43 (7), 2509–2521.

Wang, J., Wang, K., Yu, Y., Lu, Y., Xiao, W., Sun, Z., Liu, F., Zou, Z., Gao, Y., Yang, L., et al., 2025b. Self-improving generative foundation model for synthetic medical image generation and clinical applications. Nat. Med. 31 (2), 609–617.

Wang, J., Xia, B., 2021a. Accurate cup-to-disc ratio measurement with tight bounding box supervision in fundus photography. arXiv:2110.00943.

Wang, J., Xia, B., 2021b. Bounding box tightness prior for weakly supervised image segmentation. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 526–536.

Wang, K., Zhan, B., Zu, C., Wu, X., Zhou, J., Zhou, L., Wang, Y., 2022b. Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning. Med. Image Anal. 79, 102447.

Wang, P., Peng, J., Pedersoli, M., Zhou, Y., Zhang, C., Desrosiers, C., 2021b. Self-paced and self-consistent co-training for semi-supervised image segmentation. Med. Image Anal. 73, 102146.

Wang, R., Chen, B., Meng, D., Wang, L., 2018b. Weakly supervised lesion detection from fundus images. IEEE Trans. Med. Imaging 38 (6), 1501–1512.

Wang, S., Zhu, Y., Yu, L., Chen, H., Lin, H., Wan, X., Fan, X., Heng, P.-A., 2019. RMDL: recalibrated multi-instance deep learning for whole slide gastric image classification. Med. Image Anal. 58, 101549.

Wang, X., Chen, H., Xiang, H., Lin, H., Lin, X., Heng, P.-A., 2021c. Deep virtual adversarial self-training with consistency regularization for semi-supervised medical image classification. Med. Image Anal. 70, 102010.

Wang, X., Deng, X., Fu, Q., Zhou, Q., Feng, J., Ma, H., Liu, W., Zheng, C., 2020b. A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT. IEEE Trans. Med. Imaging 39 (8), 2615–2625.

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: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 2097–2106.

Wang, X., Tang, F., Chen, H., Luo, L., Tang, Z., Ran, A.-R., Cheung, C.Y., Heng, P.-A., 2020c. UD-MIL: uncertainty-driven deep multiple instance learning for OCT image classification. IEEE J. Biomed. Health. Inf. 24 (12), 3431–3442.

Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Huang, J., Yang, W., Han, X., 2021d. Transpath: transformer-based self-supervised learning for histopathological image classification. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 186–195.

Wang, Y., Li, W., Zhao, M., Huang, L., Xiao, J., 2024e. Enhancing label-efficient medical image segmentation with text-guided diffusion models. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention.

Wang, Y., Tang, P., Zhou, Y., Shen, W., Fishman, E.K., Yuille, A.L., 2021e. Learning inductive attention guidance for partially supervised pancreatic ductal adenocarcinoma prediction. IEEE Trans. Med. Imaging 40 (10), 2723–2735.

Wang, Y., Wang, Y., Cai, J., Lee, T.K., Miao, C., Wang, Z.J., 2023a. SSD-KD: a self-supervised diverse knowledge distillation method for lightweight skin lesion classification using dermoscopic images. Med. Image Anal. 84, 102693.

Wang, Y., Xiao, B., Bi, X., Li, W., Gao, X., 2023b. MCF: mutual correction framework for semi-supervised medical image segmentation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 15651–15660.

Wang, Y., Zhang, Y., Tian, J., Zhong, C., Shi, Z., Zhang, Y., He, Z., 2020d. Double-uncertainty weighted method for semi-supervised learning. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 542–551.

Wang, Z., Wu, J., Cai, L., Low, C.H., Yang, X., Li, Q., Jin, Y., 2025c. Medagent-pro: towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. arXiv:2503.18968.

Wang, Z., Yin, Z., 2021. Annotation-efficient cell counting. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2021. Springer International Publishing, pp. 405–414.

Wang, Z., Yu, L., Ding, X., Liao, X., Wang, L., 2022c. Lymph node metastasis prediction from whole slide images with transformer-guided multi-instance learning and knowledge transfer. IEEE Trans. Med. Imaging. 41 (10), 2777–2787.

Wen, C., Ye, M., Li, H., Chen, T., Xiao, X., 2024. Concept-based lesion aware transformer for interpretable retinal disease diagnosis. IEEE Trans. Med. Imaging. 44 (1), 57–68.

Wiatrak, M., Albrecht, S.V., Nystrom, A., 2019. Stabilizing generative adversarial networks: a survey. arXiv:1910.00927.

Wong, H.E., Rakic, M., Guttag, J., Dalca, A.V., 2025. ScribblePrompt: fast and flexible interactive segmentation for any biomedical image. In: Proc. Eur. Conf. Comput. Vis. Springer Nature Switzerland, pp. 207–229.

Wu, L., Zhuang, J., Chen, H., 2024a. Voco: a simple-yet-effective volume contrastive learning framework for 3D medical image analysis. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 22873–22882.

Wu, N., Phang, J., Park, J., Shen, Y., Kim, S.G., Heacock, L., Moy, L., Cho, K., Geras, K.J., 2019. The NYU breast cancer screening dataset v1. 0. New York Univ., New York, NY, USA, Tech. Rep.

Wu, N., Sun, Z., Yan, Z., Yu, L., 2024b. Feda3I: annotation quality-aware aggregation for federated medical image segmentation against heterogeneous annotation noise. In: AAAI Conf. Artif. Intell. Vol. 38, pp. 15943–15951.

Wu, X., Chen, C., Zhong, M., Wang, J., Shi, J., 2021. Covid-AL: the diagnosis of covid-19 with deep active learning. Med. Image Anal. 68, 101913.

Wu, Y., Ge, Z., Zhang, D., Xu, M., Zhang, L., Xia, Y., Cai, J., 2022. Mutual consistency learning for semi-supervised medical image segmentation. Med. Image Anal. 81, 102530.

Xia, L., Qu, Z., An, J., Gao, Z., 2023. A weakly supervised method with colorization for nuclei segmentation using point annotations. IEEE Trans. Instrum. Meas. 72, 1–11.

Xia, Y., Yang, D., Yu, Z., Liu, F., Cai, J., Yu, L., Zhu, Z., Xu, D., Yuille, A., Roth, H., 2020. Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. Med. Image Anal. 65, 101766.

Xiang, H., Shen, J., Yan, Q., Xu, M., Shi, X., Zhu, X., 2023. Multi-scale representation attention based deep multiple instance learning for gigapixel whole slide image analysis. Med. Image Anal. 89, 102890.

Xiaojin, Z., 2008. Semi-supervised learning literature survey. Comput. Sci. TR. 1530, 60.

Xie, Y., Wan, Q., Xie, H., Xu, Y., Wang, T., Wang, S., Lei, B., 2023. Fundus image-label pairs synthesis and retinopathy screening via GANs with class-imbalanced semi-supervised learning. IEEE Trans. Med. Imaging. 42 (9), 2714–2725.

Xiong, Z., Xia, Q., Hu, Z., Huang, N., Bian, C., Zheng, Y., Vesal, S., Ravikumar, N., Maier, A., Yang, X., et al., 2021. A global benchmark of algorithms for segmenting the left

atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. Med. Image Anal. 67, 101832.
Xu, G., Song, Z., Sun, Z., Ku, C., Yang, Z., Liu, C., Wang, S., Ma, J., Xu, W., 2019. Camel: a weakly supervised learning framework for histopathology image segmentation. In: Proc. IEEE Int. Conf. Comput. Vis., pp. 10682–10691.
Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al., 2024a. A whole-slide foundation model for digital pathology from real-world data. Nature 630 (8015), 181–188.
Xu, J., Hou, J., Zhang, Y., Feng, R., Ruan, C., Zhang, T., Fan, W., 2020. Data-efficient histopathology image analysis with deformation representation learning. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, pp. 857–864.
Xu, Y., Wang, Y., Zhou, F., Ma, J., Yang, S., Lin, H., Wang, X., Wang, J., Liang, L., Han, A., et al., 2024b. A multimodal knowledge-enhanced whole-slide pathology foundation model. arXiv:2407.15362.
Xu, Z., Wang, Y., Lu, D., Luo, X., Yan, J., Zheng, Y., Tong, R. K.-y., 2023. Ambiguity-selective consistency regularization for mean-teacher semi-supervised medical image segmentation. Med. Image Anal. 88, 102880.
Yan, G., Xie, L., Gao, X., Zhang, W., Shen, Q., Fang, Y., Wu, Z., 2025. FedVCK: non-IID robust and communication-efficient federated learning via valuable condensed knowledge for medical image analysis. In: AAAI Conf. Artif. Intell. Vol. 39, pp. 21904–21912.
Yan, K., Cai, J., Jin, D., Miao, S., Guo, D., Harrison, A.P., Tang, Y., Xiao, J., Lu, J., Lu, L., 2022. SAM: self-supervised learning of pixel-wise anatomical embeddings in radiological images. IEEE Trans. Med. Imaging. 41 (10), 2658–2669.
Yan, K., Wang, X., Lu, L., Summers, R.M., 2018. Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. J. Med. Imaging 5 (3), 036501.
Yan, R., Shen, Y., Zhang, X., Xu, P., Wang, J., Li, J., Ren, F., Ye, D., Zhou, S.K., 2023. Histopathological bladder cancer gene mutation prediction with hierarchical deep multiple-instance learning. Med. Image Anal. 87, 102824.
Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z., 2017. Suggestive annotation: a deep active learning framework for biomedical image segmentation. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 399–407.
Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Shao, Y., Zhang, W., Cui, B., Yang, M.-H., 2022a. Diffusion models: a comprehensive survey of methods and applications. arxiv 2022. arXiv:2209.00796.
Yang, M., Xie, Z., Wang, Z., Yuan, Y., Zhang, J., 2022b. Su-MICL: severity-guided multiple instance curriculum learning for histopathology image interpretable classification. IEEE Trans. Med. Imaging 41 (12), 3533–3543.
Yang, P., Yin, X., Lu, H., Hu, Z., Zhang, X., Jiang, R., Lv, H., 2022c. CS-CO: a hybrid self-supervised visual representation learning method for h&e-stained histopathological images. Med. Image Anal. 81, 102539.
Yang, Q., Chen, Z., Yuan, Y., 2023. Hierarchical bias mitigation for semi-supervised medical image classification. IEEE Trans. Med. Imaging. 42 (8), 2200–2210.
Yang, S., Wang, Y., Chen, H., 2024a. MambaMIL: enhancing long sequence modeling with sequence reordering in computational pathology. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 296–306.
Yang, X., Song, Z., King, I., Xu, Z., 2021. A survey on deep semi-supervised learning. arXiv:2103.00550.
Yang, Y., Zhang, H., Gichoya, J.W., Katabi, D., Ghassemi, M., 2024b. The limits of fair medical imaging AI in real-world generalization. Nat. Med. 30 (10), 2838–2848.
Yao, J., Zhu, X., Huang, J., 2019. Deep multi-instance learning for survival prediction from whole slide images. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 496–504.
Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N., Huang, J., 2020. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. Med. Image Anal. 65, 101789.
Ye, Y., Xie, Y., Zhang, J., Chen, Z., Xia, Y., 2023. UniSeg: a prompt-driven universal segmentation model as well as a strong representation learner. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 508–518.
Ying, J., Huang, W., Fu, L., Yang, H., Cheng, J., 2023. Weakly supervised segmentation of uterus by scribble labeling on endometrial cancer MR images. Comput. Biol. Med. 167, 107582.
Yoo, I., Yoo, D., Paeng, K., 2019. PseudoEdgenet: nuclei segmentation only with point annotations. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 731–739.
Yoon, J.S., Oh, K., Shin, Y., Mazurowski, M.A., Suk, H.-I., 2024. Domain generalization for medical image analysis: a review. Proc. IEEE. 112 (10), 1583–1609.
You, C., Zhou, Y., Zhao, R., Staib, L., Duncan, J.S., 2022. SimCVD: simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. IEEE Trans. Med. Imaging. 41 (9), 2228–2237.
Yu, G., Sun, K., Xu, C., Shi, X.-H., Wu, C., Xie, T., Meng, R.-Q., Meng, X.-H., Wang, K.-S., Xiao, H.-M., et al., 2021a. Accurate recognition of colorectal cancer with semi-supervised deep learning on pathological images. Nat. Commun. 12 (1), 1–13.
Yu, H., Yang, L.T., Zhang, Q., Armstrong, D., Deen, M.J., 2021b. Convolutional neural networks for medical image analysis: state-of-the-art, comparisons, improvement and perspectives. Neurocomputing 444, 92–110.
Yu, L., Wang, S., Li, X., Fu, C.-W., Heng, P.-A., 2019. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 605–613.
Yue, W., Zhang, J., Hu, K., Xia, Y., Luo, J., Wang, Z., 2024. SurgicalSAM: efficient class promptable surgical instrument segmentation. AAAI Conf. Artif. Intell. 38 (7), 6890–6898.
Zeng, Q., Xie, Y., Lu, Z., Xia, Y., 2023. Pefat: boosting semi-supervised medical image classification via pseudo-loss estimation and feature adversarial training. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 15671–15680.
Zhang, D., Chen, B., Chong, J., Li, S., 2021a. Weakly-supervised teacher-student network for liver tumor segmentation from non-enhanced images. Med. Image Anal. 70, 102005.
Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y., 2022a. DTFD-MIL: double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 18802–18812.
Zhang, K., Liu, X., Shen, J., Li, Z., Sang, Y., Wu, X., Zha, Y., Liang, W., Wang, C., Wang, K., et al., 2020. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. Cell 181 (6), 1423–1433.
Zhang, K., Zhuang, X., 2022. CycleMix: a holistic strategy for medical image segmentation from scribble supervision. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 11656–11665.
Zhang, N., Hou, J., Zhao, R.-W., Feng, R., Zhang, Y., 2021b. Semi-supervised medical image segmentation with distribution calibration and non-local semantic constraint. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, pp. 1171–1178.
Zhang, S., Zhang, J., Tian, B., Lukasiewicz, T., Xu, Z., 2023a. Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation. Med. Image Anal. 83, 102656.
Zhang, W., Zhu, L., Hallinan, J., Zhang, S., Makmur, A., Cai, Q., Ooi, B.C., 2022b. BoostMIS: boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 20666–20676.
Zhang, X., Wu, C., Zhang, Y., Xie, W., Wang, Y., 2023b. Knowledge-enhanced visual-language pre-training on chest radiology images. Nat. Commun. 14 (1), 4542.
Zhang, Y., Cheng, Y., Qi, Y., 2023c. Semisam: Exploring sam for enhancing semi-supervised medical image segmentation with extremely limited annotations. arXiv:2312.06316.
Zhang, Y., Lv, B., Xue, L., Zhang, W., Liu, Y., Fu, Y., Cheng, Y., Qi, Y., 2025. SemiSAM+: rethinking semi-supervised medical image segmentation in the era of foundation models. arXiv:2502.20749.
Zhang, Z., Sun, Z., Liu, Z., Zhao, Z., Yu, R., Du, B., Xu, Y., 2024. Spatial-aware attention generative adversarial network for semi-supervised anomaly detection in medical image. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 638–648.
Zhang, Z., Yin, F.S., Liu, J., Wong, W.K., Tan, N.M., Lee, B.H., Cheng, J., Wong, T.Y., 2010. Origa-light: an online retinal fundus image database for glaucoma analysis and research. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology. IEEE, pp. 3065–3068.
Zhao, C., Dewey, B.E., Pham, D.L., Calabresi, P.A., Reich, D.S., Prince, J.L., 2020a. Smore: a self-supervised anti-aliasing and super-resolution algorithm for mri using deep learning. IEEE Trans. Med. Imaging 40 (3), 805–817.
Zhao, H., Li, Y., He, N., Ma, K., Fang, L., Li, H., Zheng, Y., 2021a. Anomaly detection for medical images using self-supervised and translation-consistent features. IEEE Trans. Med. Imaging 40 (12), 3641–3651.
Zhao, S., Song, J., Ermon, S., 2017. Infovae: information maximizing variational autoencoders. arXiv:1706.02262.
Zhao, T., Yin, Z., 2020. Weakly supervised cell segmentation by point annotation. IEEE Trans. Med. Imaging 40 (10), 2736–2747.
Zhao, Y., Yang, F., Fang, Y., Liu, H., Zhou, N., Zhang, J., Sun, J., Yang, S., Menze, B., Fan, X., et al., 2020b. Predicting Lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 4837–4846.
Zhao, Y.-X., Zhang, Y.-M., Song, M., Liu, C.-L., 2019. Multi-view semi-supervised 3D whole brain segmentation with a self-ensemble network. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 256–265.
Zhao, Z., Zeng, Z., Xu, K., Chen, C., Guan, C., 2021b. DSAL: deeply supervised active learning from strong and weak labelers for biomedical image segmentation. IEEE J. Biomed. Health. Inf. 25 (10), 3744–3751.
Zheng, R., Zhong, Y., Yan, S., Sun, H., Shen, H., Huang, K., 2023. MsVRL: self-supervised multiscale visual representation learning via cross-level consistency for medical image segmentation. IEEE Trans. Med. Imaging 42 (1), 91–102.
Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 2921–2929.
Zhou, H.-Y., Acosta, J.N., Adithan, S., Datta, S., Topol, E.J., Rajpurkar, P., 2024. Medversa: a generalist foundation model for medical image interpretation. arXiv:2405.07988.
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, H.-Y., Lu, C., Yang, S., Han, X., Yu, Y., 2021a. Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In: Proc. IEEE Int. Conf. Comput. Vis., pp. 3499–3509.
Zhou, H.-Y., Wang, C., Li, H., Wang, G., Zhang, S., Li, W., Yu, Y., 2021b. SSMD: semi-supervised medical image detection with adaptive consistency and heterogeneous perturbation. Med. Image Anal. 72, 102117.
Zhou, M., Xu, Z., Zhou, K., Tong, R. K.-y., 2023a. Weakly supervised medical image segmentation via superpixel-guided scribble walking and class-wise contrastive regularization. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer Nature Switzerland, pp. 137–147.

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., 2023b. A foundation model for generalizable disease detection from retinal images. Nature 622 (7981), 156–163.

Zhou, Y., He, X., Huang, L., Liu, L., Zhu, F., Cui, S., Shao, L., 2019a. Collaborative learning of semi-supervised segmentation and classification for medical images. In: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 2079–2088.

Zhou, Y., Wang, Y., Tang, P., Bai, S., Shen, W., Fishman, E., Yuille, A., 2019b. Semi-supervised 3D abdominal multi-organ segmentation via deep multi-planar co-training. In: Proc. IEEE Winter Conf. App. Comput. Vis., pp. 121–140.

Zhou, Z., Shin, J.Y., Gurudu, S.R., Gotway, M.B., Liang, J., 2021c. Active, continual fine tuning of convolutional neural networks for reducing annotation efforts. Med. Image Anal. 71, 101997.

Zhou, Z.-H., Li, M., 2005. Tri-training: exploiting unlabeled data using three classifiers. IEEE Trans. Knowl. Data Eng. 17 (11), 1529–1541.

Zhu, M., Chen, Z., Yuan, Y., 2023. FedDM: federated weakly supervised segmentation via annotation calibration and gradient de-conflicting. IEEE Trans. Med. Imaging 42 (6), 1632–1643.

Zhu, Z., Yu, L., Wu, W., Yu, R., Zhang, D., Wang, L., 2022. MuRCL: multi-instance reinforcement contrastive learning for whole slide image classification. IEEE Trans. Med. Imaging. 42 (5), 1337–1348.

Zhuang, X., 2016. Multivariate mixture model for cardiac segmentation from multi-sequence MRI. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 581–588.

Zhuang, X., Li, Y., Hu, Y., Ma, K., Yang, Y., Zheng, Y., 2019. Self-supervised feature learning for 3D medical images by playing a Rubik's cube. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. Springer, pp. 420–428.

Zu, W., Xie, S., Zhao, Q., Li, G., Ma, L., 2024. Embedded prompt tuning: towards enhanced calibration of pretrained models for medical images. Med. Image Anal. 97, 103258.